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 than UCB-Improved and MOSS algorithm and thereby try to answer an open problem which has been raised before. Our proposed method partitions the arms into clusters and then following the UCB-Improved strategy eliminates sub-optimal arms and weak clusters. We corroborate our findings both theoretically and empirically and show that by sufficient tuning of parameters, 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

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. In Lai and Robbins (1985) it's shown 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^*.

There has been a large number of algorithms with strong regret guarantees, the foremost among them as mentioned in Auer et al. (2002a) called UCB1 has a regret upper bound of $O\left(\frac{K \log T}{\Delta}\right)$. This result is asymptotically order-optimal with respect to the class of algorithms mentioned in Lai and Robbins (1985), but its worst case regret as shown in Audibert and Bubeck (2009) can be as bad as $\Omega\left(\sqrt{TK \log T}\right)$ in certain regimes. In the same paper(Audibert and Bubeck (2009)) the authors come up with an algorithm called MOSS(Minimax Optimal Strategy in Stochastic case) which suffers a worst case regret of $O\left(\sqrt{TK}\right)$ which improves upon UCB1 by a factor of order $\sqrt{\log T}$ and a gap-dependent regret of $\sum_{i:\Delta_i>0} \frac{K \log(2 + T\Delta_i^2/K)}{\Delta_i}$. But MOSS also has its limitations whereby in certain regimes its gap-dependent regret bound is even worse than UCB1. The UCB-Improved algorithm bridged many of its gap by coming up with a gap-dependent regret bound of the order of $O\left(\frac{K \log T\Delta^2}{\Delta}\right)$ which improves upon UCB1 and a worst case regret of $O\left(\sqrt{TK \log K}\right)$. But this algorithm performs worse empirically as stated in Lattimore (2015). *In this work we try to address one of the open questions as raised in Bubeck and Cesa-Bianchi (2012) that is to find an algorithm with regret always better than MOSS and UCB-Improved.*

In this context of Multi-Armed Stochastic Bandit, we will also like to mention some of the variance based algorithm such as UCB-Normal(Auer et al. (2002a)) , UCB-Tuned(Auer et al. (2002a)) and UCB-Variance(Audibert et al. (2009)) which shows that variance-aware algorithms tend to perform better than the algorithms that don't(UCB1, UCB-Improved, MOSS). It can be shown that when the variance of some sub-optimal arm is lower, a variance-aware algorithm detects it quickly thereby reducing regret. *We test our algorithm against such variance-aware techniques and do an empirical analysis.*

A discussion on some recent algorithms employing divergence based techniques must also be done. Firstly, the algorithm proposed 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(they only give asymptotic guarantees). 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. Secondly, the more recent algorithm called KL-UCB(Garivier and Cappé (2011)) using KL-Divergence comes up with 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. *We empirically test against this algorithm and show that in certain regimes our algorithm performs better than KL-UCB and DMED.*

We also make a distinction between round-based algorithms like UCB-Improved, Successive Reject(Audibert and Bubeck (2010)), Successive Elimination(Even-Dar et al. (2006)) and Median-Elimination(Even-Dar et al. (2006)) whereby the algorithm pulls all the arms equal number of times in each round/phase, then proceeds to eliminate some number of arms it identifies to be sub-optimal with high probability and this continues till you are left with one arm. *Our algorithm is also a round based algorithm.* Finally, we must point out that our setup is strictly limited to stochastic scenario as opposed to adversarial setup as studied in the paper Auer et al. (2002b) where an adversary sets the reward for the arms for every timestep.

Contribution

We first analyze what are the main drawbacks of UCB-Improved as discussed in Liu and Tsuruoka (2016) before proposing our improvements.

- **Early Exploration:** A significant number of pulls are spend in initial exploration, since UCB-Improved is pulling all the arms $n_m = \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$ for the m-th round where $\tilde{\Delta}_m$ is initialized at 1 and halved after every round.
- **Not an anytime algorithm:** Since the horizon T has to be pre-specified to the algorithm, the algorithm cannot be stopped anytime because various properties(on which arms are eliminated and pulls calculated) might not hold when it is stopped prematurely.
- **Conservative arm elimination:** In UCB-Improved an arm is only eliminated after $\tilde{\Delta}_m < \frac{\Delta_i}{2}$, for some sub-optimal arm a_i. When K is large and in the critical case when $r_1 = r_2 = .. = r_{K-1} < r^*$ and Δ_i's are small this has significant disadvantage as it will wait till later rounds to eliminate the sub-optimal arms.

To counter early exploration in Liu and Tsuruoka (2016) as well as in our algorithm we propose an exploration regulatory factor to control exploration. *Our algorithm is also not an anytime algorithm*(neither is MOSS, UCB-Improved) and in this context we point out that knowledge of the horizon actually facilitates learning as it can exploit more information(as stated in Lattimore (2015)). We also employ a couple of more strategies to bring down our regret as summarized below:-

- **Clustering and local exploration:** Partition the action space(arms) into small clusters, each having uniform set of arms. Perform limited local exploration in this partitions thereby reducing expected regret. This is a one-time clustering done at the start before beginning the rounds and the number of clusters is pre-specified to the algorithm.
- **Arm Elimination:** Run an independent version of UCB-Improved inside each such clusters formed and eliminate sub-optimal arms within such clusters. We call this arm elimination condition.
- **Cluster Elimination:** Select the best payoff arm from each cluster, consider them as representative of their clusters and run cluster elimination condition on them to remove the entire cluster.
- **Exploration Regulatory Factor:** Introduce an exploration regulatory factor to control exploration so that in the later rounds we taper our exploration.
- **Aggressive Elimination:** Since we have reduced the amount of exploration by dividing the larger problem into smaller sub-problems and decreasing the number of pulls, we need aggressive elimination to remove sub-optimal arms so that within the same number of rounds as UCB-Improved we can eliminate all the sub-optimal arms with high probability. We introduce parameters ρ_a and ρ_s which helps in aggressive arm elimination and cluster elimination respectively. We do a theoretical analysis of it and show that the introduction of this parameter reduces expected regret but increases risk. We also show how various choices of these two parameters affect the expected regret.

Summarizing our contributions below:-

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

2. We achieved a lower regret upper bound(Theorem1,table in Appendix F) than UCB-Improved(Auer and Ortner (2010)), UCB1(Auer et al. (2002a)) and MOSS (Audibert and Bubeck (2009) which we verify theoretically and empirically.

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

4. In the critical case when $r_1 = r_2 = .. = r_{K-1} < r^*$ and Δ_i's are small and K is large which is encountered frequently in web-advertising domain (as stated in Garivier and Cappé (2011)) this approach has a significant advantage over other methods.

5. The only parameter that needs to be pre-specified to the algorithm is the number of clusters to be formed but unlike KL-UCB our algorithm parameter p 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 denotes the set of all clusters $s_k \in S$ in the m-th round and $|S| = p \leq K$. p is the number of clusters pre-specified at the start of algorithm. The variable ℓ is the cluster size limit. $\hat{r}_i$ 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 $s_k \in S$, 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. A_{s_k} denotes the arm set in the cluster s_k. The exploration regulatory factor is denoted by ψ_m. We also assume that all rewards are bounded in $[0, 1]$. For simplicity we will assume that there is only one optimal arm a^*. The cluster containing a^* is denoted by s^*.

3 Clustered UCB algorithm

Algorithm

The steps are presented in Algorithm 1.

Algorithm 1: ClusUCB(p)

1: Let A be the set which contains all the arms
2: Call SubroutineCluster(p)
3: $B_0 := A$
4: $\epsilon_0 := 1, w_0 := 1$
5: For rounds $m = 0, 1, .. \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$

6: Pull each arm in s_k,
$n_m = \left\lceil \frac{2\log(T\epsilon_m^2)}{\epsilon_m} \right\rceil$ number of times, $\forall s_k \in S$

7: **Arm Elimination:**
Delete all arms $a_i \in s_k$ for which
$\left\{ \hat{r}_i + \sqrt{\frac{\rho_a \log(T\epsilon_m^2)}{2n_m}} \right\} < \max_{a_j \in s_k} \left\{ \hat{r}_j - \sqrt{\frac{\rho_a \log(T\epsilon_m^2)}{2n_m}} \right\}$, $\forall s_k \in S$ where $\rho_a = \frac{1}{w_m}$ and remove all such arms from B_m.

8: **Cluster Elimination:**
Delete all clusters $s_k \in S$ for which $\max_{a_i \in s_k} \left\{ \hat{r}_i + \sqrt{\frac{\rho_s \log(T\epsilon_m^2)}{2n_m}} \right\}$
$< \max_{a_j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho_s \log(T\epsilon_m^2)}{2n_m}} \right\}$, where $\rho_s = \frac{1}{w_m}$ and remove all such arms in the cluster s_k from B_m to obtain B_{m+1}.

9: **Reset Parameters:**
$\epsilon_{m+1} := \frac{\epsilon_m}{2}, w_{m+1} := 2w_m$

10: **Stopping Condition:**
Stop if $|B_m| = 1$ and pull $\max_{a_i \in B_m} \hat{r}_i$ till T is reached.

Subroutine: SubroutineCluster(p)[Random Uniform Allocation]

1: $\ell = \left\lceil \frac{K}{p} \right\rceil$
2: Permute arms in A randomly to create A_{random}.
3: Create clusters $s_i = \{a_i\}, \forall i \in A_{random}$
4: For $i = 1$ to K:
5: For $j = i + 1$ to K:
6: Merge s_i, s_j into s_i if $\exists a_i \in s_i$ and $\exists a_j \in s_j$, such that $|s_i| + |s_j| \leq \ell$
7: Rename all partitions that have been formed as s_1 to s_p, where $|S| = p$ after merging

---

In the above algorithm we are dividing the arm set A into smaller clusters which belong to the cluster set S, where after clustering $|S| = p$. We are bounding the cluster size by ℓ such that $\ell = \left\lceil \frac{K}{p} \right\rceil$ where p is the number of clusters pre-specified. We randomly uniformly allocate arms in each of the clusters and this is shown in SubroutineCluster(p).

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-Improved $n_m = \left\lceil \frac{2log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$, where $\tilde{\Delta}_m$ is initialized at 1 and halved after every round or Median-Elimination $\frac{4}{\epsilon^2}\log\left(\frac{3}{\delta}\right)$, where ϵ, δ are the parameters for PAC guarantee.

Next, in each of the round we eliminate arms like UCB-Improved(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 that is p. 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-Improved 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. The parameters $\rho_a \in$

$(0, 1]$ and $\rho_s \in (0, 1]$ for arm and cluster elimination helps in aggressive elimination of arms and clusters respectively. At the end of each round we update all the parameters.

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 and $p > 1$, the total regret till T is upper bounded by $R_T \leq \sum_{i \in A:\Delta_i \geq b} \left\{ 2\Delta_i + \left(\frac{32}{\Delta_i^3}\right) + \left(\frac{2^{2+4\rho_s}\rho_s^{2\rho_s}T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right) + \left(\frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}\right) + \left(\frac{32\rho_s\log(T\frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}\right)\right\} + \sum_{i \in A_{s^*}:0 \leq \Delta_i \leq b} \left\{ \left(\frac{12}{\Delta_i^3}\right) + \left(\frac{12}{b^3}\right)\right\} + \sum_{i \in A:0 \leq \Delta_i \leq b} \left\{ \left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right) + \left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{b^{4\rho_s-1}}\right)\right\} + max_{i:\Delta_i \leq b}\Delta_i T$, where $\rho_s \in (0, 1]$ is the cluster elimination parameter, A_{s^*} is the number of arms in optimal cluster s^* such that $|A_{s^*}| \leq \lceil \frac{K}{p} \rceil$, p is the number of clusters and T is the horizon.*

Proof. The proof of this is given in **Appendix C**.(Supplementary Material) □

Remark 1. *By setting $b = \sqrt{\frac{e}{T}}$, the term $max_{i:\Delta_i \leq b}\Delta_i T$ gets trivially bounded by $\sqrt{KT}$ as mentioned in UCB-Improved.*

Remark 2. *From the statement of Theorem 1 we can see that taking various values of ρ_s and ρ_a and how these values are reduced after every round significantly affects our regret bound.A discussion on them is deferred to **Remark 11, Appendix A**, **Remark 13,14, Appendix B** and **Remark 16, Appendix C** and its various definitions are given in **Appendix E**. A discussion on exploration regulatory factor(ψ_m) and its effect is deferred to **Appendix D**. In the analysis of proof(**Remark 3-7**) we will use both ψ_m and ρ_s. The full regret bound which includes the factors ρ_s, ρ_a and ψ_m can be found in **Remark 17, Appendix D**.*

Remark 3. *So, for the term $\sum_{i \in A:\Delta_i \geq b} \frac{32\rho_s\log(\psi_m T\frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}$ by setting $b \approx \sqrt{\frac{K\log K}{T}}$ in the term we get, $K\frac{32\rho_s\log(\psi_m T\frac{K^2(\log K)^2}{16\rho_s^2T^2})}{\sqrt{\frac{K\log K}{T}}} = \frac{32K\sqrt{T}\rho_s\log(\psi_m\frac{K^2(\log K)^2}{16\rho_s^2T})}{\sqrt{K\log K}}$. Again, since there are $\lfloor \frac{1}{2}\log_2\frac{T}{e} \rfloor$ rounds then $\rho_s \geq \frac{1}{4}$(from Definition 2, **Appendix E**) and taking $\psi_m = K^2T$(from Definition 3, **Appendix D**). So we get, $\frac{32\sqrt{KT}\log(16\frac{K^4(\log K)^2}{16*T})}{4\sqrt{\log K}} \leq \frac{8\sqrt{KT}\log(\frac{K^4(\log K)^2}{T})}{\sqrt{\log K}}$.*

Remark 4. *Again, of the several terms in the regret bound, the terms like $\sum_{i \in A:\Delta_i \geq b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{\psi_m^{\rho_s}\Delta_i^{4\rho_s-1}}\right)$ can become very large given a small Δ_i and large T. Drawing from the conclusion of our exploration regulatory factor in **Appendix D**, we can introduce $\psi_m = K^2T$ in our confidence interval thereby when*

$\rho_s = \frac{1}{4}$*(from Definition* 2, ***Appendix E****) then the term essentially becomes* $K\left(\frac{T^{\frac{3}{4}}(\frac{1}{4})^{\frac{1}{2}}2^{\frac{1}{2}+3}}{(K^2T)^{\frac{1}{4}}(\Delta_i)^{4*\frac{1}{4}-1}}\right) = 5.6\sqrt{KT}$*. A similar result can also be shown for* $\sum_{i\in A:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{b^{4\rho_s-1}}\right) = 5.6\sqrt{KT}$.

Remark 5. *So, we see that putting* $b = \sqrt{\frac{K\log K}{T}}$*, the corresponding terms of* ρ_s *and* ψ_m *and then combining* ***remark 3 and 4*** *we get the worst case gap-independent regret bound of* $\left\{\frac{\sqrt{KT}\log(\frac{K^4(\log K)^2}{T})}{\sqrt{\log K}}\right\}$ *which coincides with the term* bT *of the regret bound in* ***Theorem 1****, except the term* $\frac{\log(\frac{K^4(\log K)^2}{T})}{\sqrt{\log K}}$*. When we consider that* $T \approx K^4(\log K)^2$*, this term becomes negligible and so the regret bound becomes better than UCB-Improved non-gap based regret upper bound of* $\sqrt{KT}\frac{\log(K\log K)}{\sqrt{\log K}}$ *and slightly better than the MOSS bound of* $49\sqrt{KT}$*. This shows that for large number of arms* K *and moderately high time horizon* T *we have a very strong regret guarantee.*

Remark 6. *A similar result can be shown for the term* $\frac{32\rho_a\log(\psi_m T\frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i}$ *and for the terms like* $\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right)$ *which come from Arm Elimination method. A clear distinction between using only arm elimination method(as proved in* ***Proposition 1*** *or using only cluster elimination method(as proved in* ***Proposition 2*** *or both arm and cluster elimination(as proved in* ***Theorem 1*** *is shown in* ***Remark 16****.*

Remark 7. *When we consider gap-dependent bound, by choosing appropriate values of* ρ_s, ψ_m *and* p *we see that the most significant term in the regret is* $\sum_{i\in A:\Delta_i\geq b}\frac{32\rho_s\log(\psi_m T\frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} = \frac{8K\log(T\log T\frac{\Delta_i^4}{32})}{\Delta_i}$ *(when* $\rho_s \geq \frac{1}{4}$ *and* $\psi_m = \frac{\log T}{2^m}$*), which is significantly lower than UCB1, MOSS and UCB-Improved. A table for the various regret upper bounds is given in* ***Appendix F****. From the* ***Remarks 3-6*** *we see that* ρ_s *actually helps in reducing the constant associated with the factor* $\sqrt{KT}$ *and the* $\frac{\log(\psi_m\frac{T\Delta_i^4}{\rho_s^2})}{\Delta_i}$ *term and also both* ψ_m *and* ρ_s *helps in stabilizing the term* $\log(\psi_m\frac{T\Delta_i^4}{\rho_s^2})$*.*

Remark 8. *In the experimental setup we use* ψ_m *as stated in* ***Appendix D, Definition 5*** *and* ρ_s *and* rho_a *as stated in* ***Appendix E, Definition 5****. The corresponding bound can be found in* ***Remark 18****.*

Remark 9. *A sketch of the proof for* ***Theorem*** 1 *is given here. In the first step, we try to bound the probability of arm elimination of any sub-optimal arm without cluster elimination. This is shown in Proposition* 1*. In second step, we try to bound the probability of cluster elimination(without any arm elimination) with all arms within it. This is shown in Proposition* 2*. Finally, in the proof of Theorem 1 we combine all these to get the regret bound.*

Proposition 1. *Considering only the arm elimination condition and* $\rho_a = 1$ *and* $p = 1$ *the total regret till* T *is upper bounded by* $R_T \leq \sum_{i\in A:\Delta_i\geq b}\left\{\left(\frac{44}{(\Delta_i)^3}\right) + \left(\Delta_i + \frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}\right)\right\} + \sum_{i\in A:0\leq\Delta_i\leq b}\frac{12}{b^3} +$

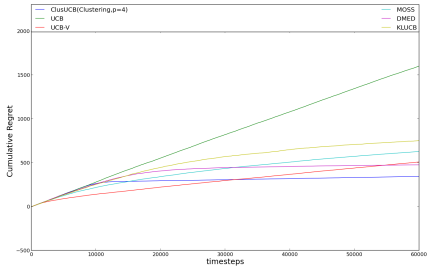

Figure 1: Experiment 1: Regret for various Algorithms. $T = 60000$

$max_{i:\Delta_i \leq b} \Delta_i T$ *for all* $b \geq \sqrt{\frac{e}{T}}$, *where* ρ_a *is the arm elimination parameter,* p *is the number of clusters and* T *is the horizon.*

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

Remark 10. *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-Improved 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 can also point out out here intuitively that when* $c_m = \sqrt{\frac{\rho_a \log(T\epsilon_m^2)}{2n_m}}$ *and* ρ_a *is decreased after every round, then an arbitrary arm* a_i *is removed as soon as* $\sqrt{\rho_a \epsilon_m} < \frac{\Delta_i}{2}$ *which essentially makes it a faster elimination procedure than UCB-Improved if* $\rho_a \leq \epsilon_m$. *The proof of a similar approach regarding cluster elimination is shown in proposition 2.*

Proposition 2. *Considering only the cluster elimination condition and* $p > 1$, *the total regret till* T *is upper bounded by* $R_T \leq \sum_{i \in A:\Delta_i \geq b} \left\{ \left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} \right) + \left(\Delta_i + \frac{32\rho_s \log\left(T \frac{\Delta_i^4}{16\rho_s^2}\right)}{\Delta_i} \right) + \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\Delta_i^{4\rho_s - 1}} \right) \right\} +$ $\sum_{i \in A:0 \leq \Delta_i \leq b} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{b^{4\rho_s - 1}} \right) + max_{i:\Delta_i \leq b} \Delta_i T$ *for all* $b \geq \sqrt{\frac{e}{T}}$, *where* $\rho_s \in (0, 1]$ *is the cluster elimination parameter,* p *is the number of clusters and* T *is the horizon.*

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

5 Simulation experiments

In the stochastic bandit literature there are several powerful algorithms with and without proven regret bounds. Algorithms like ϵ-greedy(Sutton and Barto (1998)) or softmax(Sutton and Barto (1998)) or UCB-Tuned(Auer et al. (2002a)) has no proven regret bounds. Again algorithms like UCB-δ(Abbasi-Yadkori et al. (2011)) with proven regret bound better than UCB1 falls within the realm of fixed confidence setting whereas one has to provide the probability of error δ. We also make a distinction between frequentist based

approach like the UCB algorithms and the Bayesian approach like the Thompson Sampling(Agrawal and Goyal (2011)). We will not be experimenting against this algorithms but focus in the more recent, algorithms like KL-UCB, DMED, MOSS, UCB1, UCB-V, etc.

The first experiment is conducted over a testbed of 20 arms for the test-cases involving Bernoulli reward distribution with expected rewards of the arms $r_{i_{a_i \neq a^*}} = 0.07$ and $r^* = 0.1$. These type of cases are frequently encountered in web-advertising domain. The horizon is set for $T = 60000$ and the parameters of ClusUCB are $p = 4, \psi_m = K^{\frac{1}{\rho_s}}$(Appendix D, Definition 5) and $\rho_s = \frac{1}{2^{2m+1}}, \rho_a = \frac{1}{2^{4m+1}}$(Appendix E, Definition 5), where m is the round number. The regret is averaged over 100 independent runs over each testbed and is shown in **Figure 1**. 6 algorithms, ClusUCB, MOSS, UCB1, UCB-V, KL-UCB, DMED are run over this testbed and shown in this figure. Here, we see that ClusUCB performs better than all the algorithms mentioned above.

6 Conclusions and future work

Our study concludes that theoretically the regret of ClusUCB is lower than UCB1, MOSS and UCB-Improved while empirically we compared against UCB2, KL-UCB, UCB-Tuned, DMED and Median Elimination under certain cases when the Δ_i's are small 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. For the critical case when $\Delta_{i:r_i<r*}, \forall i \in A$ are equal then ClusUCB performs better than UCB-Improved, 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 every round. 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 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. 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(advertisements) 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.

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.

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.

Tor Lattimore. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

Yun-Ching Liu and Yoshimasa Tsuruoka. Modification of improved upper confidence bounds for regulating exploration in monte-carlo tree search. *Theoretical Computer Science*, 2016.

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 3. *Considering only the arm elimination condition and $\rho_a = 1$ and $p = 1$ the total regret till T is upper bounded by $R_T \leq \sum_{i \in A:\Delta_i \geq b} \left\{ \left(\frac{44}{(\Delta_i)^3} \right) + \left(\Delta_i + \frac{32 \log (T \frac{\Delta_i^4}{16})}{\Delta_i} \right) \right\} + \sum_{i \in A: 0 \leq \Delta_i \leq b} \frac{12}{b^3} + max_{i:\Delta_i \leq b} \Delta_i T$ for all $b \geq \sqrt{\frac{e}{T}}$, where ρ_a is the arm elimination parameter, p is the number of clusters and T is the horizon.*

Proof. of Proposition 3:

Let, $\rho_a = 1$. Also let $p = 1$ whereby we partition the arm set A into equal clusters each containing one arm which gives rise to the worst case scenario, that is we have one UCB-Improved running throughout. So, for each sub-optimal arm a_i, $m_i = \min \{m | \sqrt{\epsilon_m} \leq \frac{\Delta_i}{2}\}$ be the first round when $\sqrt{\epsilon_m} \leq \frac{\Delta_i}{2}$. Also in this proof, since the clusters are fixed, so throughout the rounds each m_i is tied to a single arm.

A.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 $\hat{r}_i + c_{m_i} \leq \hat{r}^* - c_{m_i}$.
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,

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

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,

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

$$\leq \frac{1}{T\epsilon_{m_i}^2}$$

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 $T\Delta_i$,

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

A.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,

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

A.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 $\sqrt{\epsilon_m} < \frac{\Delta_i}{2}$

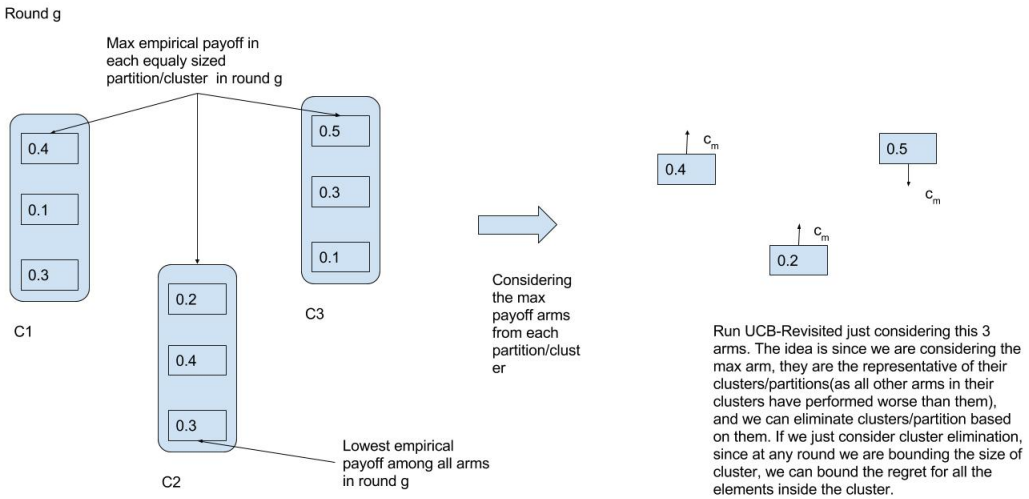

Figure 2: Cluster Elimination

$$\leq \sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i\in A'':m_i>m_*} 4\left(\frac{1}{\epsilon_m^{3/2}}\right)$$
$$\leq \sum_{i\in A'':m_i>m_*} \sum_{m_*=0}^{\min\{m_i,m_b\}} \left(\frac{4}{2^{-(3/2)m_*}}\right)$$
$$\leq \sum_{i\in A'} \left(\frac{4}{2^{-(3/2)m_*}}\right) + \sum_{i\in A''\setminus A'} \left(\frac{4}{2^{-(3/2)m_b}}\right)$$
$$\leq \sum_{i\in A'} \left(\frac{4*2^{3/2}}{\Delta_i^3}\right) + \sum_{i\in A''\setminus A'} \left(\frac{4*2^{3/2}}{b^3}\right)$$
$$\leq \sum_{i\in A'} \left(\frac{12}{\Delta_i^3}\right) + \sum_{i\in A''\setminus A'} \left(\frac{12}{b^3}\right)$$

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

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

Remark 11. *In this proof for simplicity we take $\rho_a = 1$. From the result we can see that for small Δ_i and large K, the terms like $\sum_{i\in A:\Delta_i\geq b}\left(\frac{44}{(\Delta_i)^3}\right) + \sum_{i\in A:0\leq\Delta_i\leq b}\frac{12}{b^3}$ can become the dominant term in the regret rather than $\sum_{i\in A:\Delta_i\geq b}\frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}$. Intuitively, this actually suggests that the algorithm is actually trying to eliminate arms with too low exploration and so the probability of elimination is low and error(risk) is high. For this essentially we introduce ρ_a, ρ_s and ψ_m which enables us to eliminate arms and clusters aggressively and thereby reduce the first two terms. After proving proposition 4 we give an alternate regret bound for arm elimination, which closely follows from the proof of both proposition 3 and 4.*

B Appendix B

Proposition 4. *Considering only the cluster elimination condition and $p > 1$, the total regret till T is upper bounded by* $R_T \leq \sum_{i \in A:\Delta_i \geq b} \left\{ \left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} \right) + \left(\Delta_i + \frac{32\rho_s \log (T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right) + \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\Delta_i^{4\rho_s - 1}} \right) \right\} + \sum_{i \in A:0 \leq \Delta_i \leq b} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{b^{4\rho_s - 1}} \right) + max_{i:\Delta_i \leq b} \Delta_i T$ *for all* $b \geq \sqrt{\frac{e}{T}}$*, where* $\rho_s \in (0, 1]$ *is the cluster elimination parameter, p is the number of clusters and T is the horizon.*

Remark 12. *A sketch of the proof is given below, An illustrative diagram explaining Cluster Elimination is given in Figure 2.*

- *Define C_g as the active cluster set which contains the max payoff arm from all the clusters active in the g-th round. So, the maximum cardinality of C_g is p and also this is always a non-increasing set.*
- *Now, the elements in C_g which are the max-payoff arm from each cluster(if there are more than 1 max-payoff arm in a cluster then choose randomly between them) are not fixed and hence like B_m in proposition 3 cannot be tied to an arm a_i throughout the rounds. For this purpose the i-th element in C_g we tie to a cluster, that is $a_{s_k} \in C_g$ is the max payoff arm in the cluster s_k if the cluster s_k still surviving in round g and call a_{s_k} as a cluster arm.*
- *Define g-th round(in proposition 3) as the same way as the m-th round signifying the first/minimum round when a cluster gets eliminated.*
- *For regret bound proof according to proposition 3, consider only this C_g and proof following the same way as in proposition 3 with some changes. In any round g, atleast one of these 5 events must occur:-*
  - ***Case a1:** In the g-th round, $a^* \in C_g$ and bound the maximum probability that a sub-optimal cluster arm $a_{s_k} \in C_g$ gets eliminated.*
  - ***Case a2:** In the g-th round, $a^* \notin C_g$ and bound the maximum probability that a sub-optimal cluster arm a_{s_k} gets eliminated by $a_{max_{s^*}} \in C_g$, such that $a_{max_{s^*}}, a^* \in s^*$ and $a_{max_{s^*}} \neq a^*$. So, a^* survives by this condition.*
  - ***Case b1:** In the g-th round, $a^* \notin C_g$ and no cluster arm gets eliminated and calculate the regret for all the surviving arms/clusters(clusters are fixed).*
  - ***Case b2:** In the g-th round $a^* \in C_g$ and it gets eliminated by another sub-optimal cluster arm and bound this probability.*
  - ***Case b3:** In the g-th round $a^* \notin C_g$ and it gets eliminated by another sub-optimal cluster arm and bound this probability.*

  For bounding the probability we use Chernoff bound and use it on the current set of cluster arms $a_{s_k} \in C_g, \forall s_k \in S$. This will work because of the algorithm, we are pulling all the surviving arms equal number of times in each round and applying the same confidence interval for cluster elimination to all of the elements of C_g and also we fix the clusters from beginning of the rounds.
- *Introduce a bounded parameter $\rho_s \in (0, 1]$ for cluster elimination to mimic the arm elimination condition in proposition 3, but with the condition that whenever $\sqrt{\rho_s \epsilon_{g_{s_k}}} \leq \frac{\Delta_{a_{s_k}}}{2}, a_{s_k} \in C_g$, then the*

cluster s_k(where $\hat{r}_{a_{s_k}}$ 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 $\ell = \left\lceil \frac{K}{p} \right\rceil$.

Proof. of Proposition 4:

Let $C_g = \{\hat{r}_{max_{s_k}} | \forall s_k \in S\}$, 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 cluster arm $a_{s_k} \in C_g$, $g_{s_k} = \min\{g | \sqrt{\rho_s \epsilon_{g_{s_k}}} \leq \frac{\Delta_{a_{s_k}}}{2}\}$. So, g_{s_k} be the first round when $\sqrt{\rho_s \epsilon_{g_{s_k}}} \leq \frac{\Delta_{a_{s_k}}}{2}$ where $a_{s_k} \in C_g$ is the maximum payoff arm in cluster s_k and then s_k gets eliminated. Here, a_{s_k} is called cluster arm. We will also consider for Case $a1, b1$ and $b2$ that $\max \hat{r}_i \in C_g$ is a^* and it has not still been eliminated. For case $a2$ and $b2$ we will consider a^* not in C_g. A_{s_k} denotes the arm set in the cluster s_k. So, for cluster elimination we will only be considering the arms in C_g called cluster arms, and follow the proof of proposition 3,

B.1 Case a1:

Some sub-optimal cluster arm a_{s_k} is not eliminated in round g_{s_k} or before and the optimal cluster arm $a^* \in C_{g_{s_k}} \subset B_m$

In arm elimination condition, given the choice of confidence interval $c_{g_{s_k}}$, we want to bound the event $\hat{r}_{s_k} + c_{g_{s_k}} \leq \hat{r}^* - c_{g_{s_k}}$.

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

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

$$c_{g_{s_k}} = \sqrt{\frac{\rho_s * \epsilon_{g_{s_k}} \log(T\epsilon_{g_{s_k}}^2)}{2 * 2\log(T\epsilon_{g_{s_k}}^2)}} = \sqrt{\frac{\rho_s \epsilon_{g_{s_k}}}{2}} = \sqrt{\rho_s \epsilon_{g_{s_k}+1}} < \frac{\sqrt{\rho_s}\Delta_{a_{s_k}}}{4} < \frac{\Delta_{a_{s_k}}}{4}$$

Again, $\exists a_{s_k} \in C_g$ such that,
$$\begin{aligned} \hat{r}_{a_{s_k}} + c_{g_k} &\leq r_{a_{s_k}} + 2c_{g_k} \\ &= \hat{r}_{a_{s_k}} + 4c_{g_k} - 2c_{g_k} \\ &\leq r_{a_{s_k}} + \Delta_{a_{s_k}} - 2c_{g_k} \\ &< r^* - 2c_{g_k} \\ &\leq \hat{r}^* - c_{g_k} \end{aligned}$$

Hence, we get that as soon as $\sqrt{\rho_s \epsilon_{g_{s_k}}} < \frac{\Delta_{a_{s_k}}}{2}$, $\exists a_{s_k} \in C_g$ which gets eliminated.

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

So, we need to bound the probability,

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

Applying Chernoff-Hoeffding bound and considering independence of events,

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

$$\leq \frac{1}{(T\epsilon_{g_k}^2)^{\rho_s}}$$

Similarly, $\mathbb{P}\{\hat{r}_{a_{s_k}} \geq r_{a_{s_k}} + c_{g_{s_k}}\} \leq \frac{1}{(T\epsilon_{g_{s_k}}^2)^{\rho_s}}$

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

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 since clusters are fixed so we can bound the maximum probability that a sub-optimal arm $a_j \in A$ and $a_j \in s_k$ such that $a_{s_k} \in C_g$ is not eliminated in the g-th round by the same probability of $\left(\frac{2}{(T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right)$.

Summing up over all arms in s_k and bounding trivially by $T\Delta_i$, $\sum_{i \in A_{s_k}} \left(\frac{2T\Delta_i}{(T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right)$

Summing up over all clusters p and bounding trivially by $T\Delta_i$,

$$\sum_{k=1}^{p} \sum_{i \in A_{s_k}} \left(\frac{2T\Delta_i}{(T\frac{\Delta_i^4}{16\rho_s^2})^{\rho_s}}\right)$$

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

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

Thus, we see that putting a value of $\rho_s = 1$, nicely brings out the result of case a of proposition 3.

B.2 Case a2:

In the above case we considered that $a^* \in C_g$ being the max-payoff arm from a cluster s^*(say). Now, if that is not the case and if in s^* both $a_{max_{s^*}}, a^* \in s^*$ such that $\hat{r}_{a_{max_{s^*}}} > \hat{r}^*$, so $a_{max_{s^*}} \in C_g$ in the g-th round then for some sub-optimal arm $a_{s_k} \in C_g$, $\hat{r}_{a_{s_k}} + c_{g_{s_k}} < \hat{r}_{a_{max_{s^*}}} - c_{g_{s_k}}$. But, this probability can be no worse than case $a1$ since $r_{max_{s^*}} < r^*$ and all arms get pulled $n_{g_{s_k}}$ number of times. So the regret can be no worse than $\left(\frac{2}{(T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right)$. After summing over all arms in A and bounding trivially by $T\Delta_i$ we get the same result as above. The regret can be no more than,

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

B.3 Case b1:

Either an arm $a_{s_k} \in C_g$ is eliminated along with all the arms in the cluster s_k in round g_{s_k}(or before) or else there is no optimal arm $a^* \in C_g$. Again, in the g-th round, the maximum total elements in the cluster can be no more than $\ell = \left\lceil \frac{K}{p} \right\rceil$.

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

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

So, the total contribution of a_{s_k} along with all the other arms in the cluster till round g_{s_k} is given by,

$$\sum_{i\in A_{s_k}} \Delta_i \left\lceil \frac{2\log(T\epsilon_{g_{s_k}}^2)}{\epsilon_{g_{s_k}}} \right\rceil \leq \sum_{i\in A_{s_k}} \Delta_i \left\lceil \frac{2\log(T(\frac{\Delta_i}{2\sqrt{\rho_s}})^4)}{(\frac{\Delta_i}{2\sqrt{\rho_s}})^2} \right\rceil, \text{ since } \sqrt{\rho_s\epsilon_{g_{s_k}}} \leq \frac{\Delta_{a_{s_k}}}{2} \leq \frac{\Delta_i}{2} \text{ as}$$

$\hat{r}_{a_{s_k}} > \hat{r}_i, \forall i \in s_k$

$$\leq \sum_{i\in A_{s_k}} \Delta_i \left(1 + \frac{32 * \rho_s * \log(T(\frac{\Delta_i}{2\sqrt{\rho_s}})^4)}{\Delta_i^2}\right)$$

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

Summing over all clusters p ,

$$\leq \sum_{k=1}^{p} \sum_{i\in A_{s_k}} \Delta_i \left(1 + \frac{32\rho_s \log(T\frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i^2}\right)$$

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

B.4 Case b2:

In this case we will consider that the cluster s^* 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 $g_{s_k} = g_*$ or it will lead to the contradiction that $r_{a_{s_k}} > r^*$ where $a_{s_k}, a^* \in C_g$. In any round g_*, if the optimal arm a^* gets eliminated then for any round from 1 to g_{s_j} all arms $a_{s_j} \in C_g, \forall s_j \neq s^*$ such that $\sqrt{\rho_s\epsilon_{g_{s_k}}} < \frac{\Delta_{a_{s_j}}}{2}$ 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_{s_b} such that $\sqrt{\rho_s\epsilon_{g_{s_b}}} \geq \frac{\Delta_{a_{s_b}}}{2}$ 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 $a_j \in B_m$ with $g_{s_j} \geq g_*$. Let g_{s_b} be the round when $\sqrt{\rho_s\epsilon_{g_{s_b}}} < \frac{\Delta_{s_b}}{2}$ that is $g_b = min\{g|\sqrt{\rho_s\epsilon_{g_{s_b}}} < \frac{\Delta_b}{2}\}$ and the cluster s_b gets eliminated. Hence, the maximal regret after eliminating the arm a^* is upper bounded by,

$\sum_{g_*=0}^{max_{j\in C_g^{'}} g_{s_j}} \sum_{i\in C_g^{''}:g_{s_k}>g_*} \left(\frac{2}{(T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right).T \max_{j\in C_g^{''}:g_{s_j}\geq g_*} \Delta_{a_{s_j}}$

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 $A^{''} \supset C_g^{''}$ where $A^{'}$ is the set of all the arms across clusters surviving till g_* round and $A^{''}$ be the set of all arms across clusters to still survive and eliminate arm a^* in round g_* respectively.

$\leq \sum_{g_*=0}^{max_{j\in A^{'}} g_{s_j}} \sum_{i\in A^{''}:g_{s_k}>g_*} \left(\frac{2}{(T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right).T \max_{j\in A^{''}:g_{s_j}\geq g_*} \Delta_{a_{s_j}}$

$\leq \sum_{g_*=0}^{max_{j\in A^{'}} g_{s_j}} \sum_{i\in A^{''}:g_{s_k}>g_*} \left(\frac{2}{(T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right).T.2\sqrt{\rho_s\epsilon_{g_{s_j}}}$, since $\sqrt{\rho_s\epsilon_{g_{s_j}}} \leq \frac{\Delta_{a_{s_j}}}{2} \leq \frac{\Delta_j}{2}$ as $\hat{r}_{a_{s_j}} > \hat{r}_j, \forall j \in$

s_j

$$\leq \sum_{g_*=0}^{max_{j\in A'} g_{s_j}} \sum_{i\in A'':g_{s_k}>g_*} \left(\frac{4\rho_s^{2\rho_s}T^{1-\rho_s}}{\epsilon_{g_{s_k}}^{2\rho_s-\frac{1}{2}}}\right)$$

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

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

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

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

B.5 Case b3:

Now, in the above case again we considered that a^* is in C_g. In this case we will consider that the cluster s^* containing the optimal arm a^* was eliminated by another sub-optimal cluster and $a^* \notin C_g$. This will be the mirror case of Case $b2$ and since s^* gets eliminated by a_{s_b} such that $\sqrt{\rho_s\epsilon_{g_{s_b}}} \geq \frac{\Delta_{a_{s_b}}}{2}$ round g_*. Also maximal regret per step after eliminating a^* is the maximal Δ_j among the remaining arms $a_j \in B_m$ with $g_{s_j} \geq g_*$. Following the same way above we can bound the regret as,

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

Summing up **Case a1**, **Case a2**, **Case b1**,**Case b2** and **Case b3**, the total regret till round g is given by,

$$R_T \leq \sum_{i\in A:\Delta_i\geq b} \Bigg\{ \underbrace{\left(\frac{2^{1+4\rho_s}\rho_s^{2\rho_s}T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case a1}} + \underbrace{\left(\frac{2^{1+4\rho_s}\rho_s^{2\rho_s}T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case a2}} + \underbrace{\left(\Delta_i + \frac{32\rho_s\log(T\frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}\right)}_{\text{case b1}} +$$

$$\underbrace{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+\frac{3}{2}}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case b2}} + \underbrace{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+\frac{3}{2}}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case b3}} \Bigg\} + \underbrace{\sum_{i\in A:0\leq\Delta_i\leq b} \left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+\frac{3}{2}}}{b^{4\rho_s-1}}\right)}_{\text{case b2}} + \underbrace{\sum_{i\in A:0\leq\Delta_i\leq b} \left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+\frac{3}{2}}}{b^{4\rho_s-1}}\right)}_{\text{case b3}}$$

$$+ max_{i:\Delta_i\leq b}\Delta_i T$$

$$= \sum_{i\in A:\Delta_i\geq b} \Bigg\{ \underbrace{\left(\frac{2^{2+4\rho_s}\rho_s^{2\rho_s}T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case a1+a2}} + \underbrace{\left(\Delta_i + \frac{32\rho_s\log(T\frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}\right)}_{\text{case b1}} + \underbrace{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case b2+b3}} \Bigg\}$$

$$+ \underbrace{\sum_{i\in A:0\leq\Delta_i\leq b} \left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{b^{4\rho_s-1}}\right)}_{\text{case b2+b3}} + max_{i:\Delta_i\leq b}\Delta_i T \qquad \square$$

Remark 13. *Again, we see that $\rho_s = 1$ and $p = K$(that is each arm is in a cluster of its own and so $\left\lceil\frac{K}{p}\right\rceil = 1$) brings out a near equivalent result as the result of proposition 3. That is, for $\rho_s = 1, p = K$,*

$$R_T \leq \sum_{i\in A:\Delta_i\geq b} \left\{ \left(\frac{96}{(\Delta_i)^3}\right) + \left(\Delta_i + \frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}\right)\right\} + \sum_{i\in A:0\leq\Delta_i\leq b} \frac{32}{b^3} + max_{i:\Delta\leq b}\Delta_i T$$

We also see that both the term $\sum_{i \in A:\Delta_i \geq b} \left(\frac{96}{(\Delta_i)^3} \right)$ *which results from sub-optimal arms not being eliminated and* $\sum_{i \in A:0 \leq \Delta_i \leq b} \frac{32}{b^3}$ *which is the error bound more than proposition* 3*. For the last two terms choosing appropriate exploration regulatory factor* ψ_m *and* ρ_s *will reduce these terms considerably. So simply running Cluster elimination on the arms will result in more regret than running Arm elimination on all the arms without clustering.*

Another takeaway from this result is that a lower value of $\rho_s \in (0, 1]$ *makes the algorithm risky(by increasing the error bound), but reduces expected regret whereas a higher value of* $\rho_s > 1$ *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 not getting eliminated increases without proper exploration. So, there is always this trade-off between exploration, elimination and risk. So, in our algorithm we decrease the* ρ_s *and* ρ_a *in a graded fashion halving after every round but* ρ_s *and* ρ_a *is always bounded that is,* $\rho_s, \rho_a \in (0, 1]$*.*

Remark 14. *Considering,* $\rho_a \in (0, 1]$ *and* $p = 1$*, the regret upper bound for arm elimination can be modified in the same way giving us a bound of*

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

Putting $\rho_a = 1$ *gives us directly the result of proposition* 3*. Here we will only consider case* a*, case* $b1$ *and case* $b2$ *of proposition* 3*.*

C Appendix C

Theorem 2. *Considering both the arm elimination and cluster elimination condition and* $p > 1$*, the total regret till* T *is upper bounded by* $R_T \leq \sum_{i \in A:\Delta_i \geq b} \left\{ 2\Delta_i + \left(\frac{32}{\Delta_i^3} \right) + \left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} \right) + \right.$
$\left. \left(\frac{32 \log (T \frac{\Delta_i^4}{16})}{\Delta_i} \right) + \left(\frac{32\rho_s \log (T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right) \right\} + \sum_{i \in A_{s^*}:0 \leq \Delta_i \leq b} \left\{ \left(\frac{12}{\Delta_i^3} \right) + \left(\frac{12}{b^3} \right) \right\} + \sum_{i \in A:0 \leq \Delta_i \leq b} \left\{ \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + 3}}{\Delta_i^{4\rho_s - 1}} \right) + \right.$
$\left. \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + 3}}{b^{4\rho_s - 1}} \right) \right\} + max_{i:\Delta_i \leq b} \Delta_i T$*, where* $\rho_s \in (0, 1]$ *is the cluster elimination parameter,* A_{s^*} *is the number of arms in optimal cluster* s^* *such that* $|A_{s^*}| \leq \lceil \frac{K}{p} \rceil$ *,* p *is the number of clusters and* T *is the horizon.*

Proof. The optimal cluster which contains a^* is denoted by s^*. The arms belonging to cluster s_k is denoted by A_{s_k}. Combining both the cases of Proposition 3 and Proposition 4 we can see that a sub-optimal arm a_i can only be eliminated given that either m_i or $g_{s_k} : a_i \in s_k$ happens with $a^* \in s^*$ still surviving. In Proposition 3 we consider only arm elimination and in Proposition 4 we consider only cluster elimination. One vital point we point out is that, ϵ_m(in proposition 3) = ϵ_g(in proposition 4). Also this proof we will consider $p > 1$. So there will be slight modification from what we proved in proposition 3 with $p = 1$. For random uniform allocation we will assume that each cluster $s_k, \forall s_k \in S$, gets such an arm such that $a_{max_{s_k}} > a_i, \forall i \in s_k$. Again, $a^* >= a_{max_{s_k}}, \forall s_k \in S$.

C.1 Case a:

In this case, a sub-optimal arm a_i can get eliminated in 4 ways,

1 a_i in s^* and m_i happens which is Proposition 3(case $a1$)

2 a_i in s_k, where $r_{max_{s_k}} < r^*$ and m_i happens. This case was not dealt in Proposition 3 as there we took $p = 1$. Since, now $p > 1$ and $r_{max_{s_k}} < r^*$, following the same way as case a, Proposition 3 we can show that the probability of a_i not getting eliminated cannot be worse than Proposition 3(case $a1$) given that $\sqrt{\epsilon_m} < \frac{\Delta_i^{'}}{2}$ where $\Delta_i^{'} = r_{max_{s_k}} - r_i$ such that $r_i \in s_k$. Plugging in this $\Delta_i^{'}$ in Proposition 3, case a we can derive a similar bound where $\Delta_i^{'} >= \Delta_{a_{max_{s_k}}}$ because otherwise case 3(next case) will happen before as $\sqrt{\epsilon_m} < \frac{\Delta_{a_{max_{s_k}}}}{2}$ and the cluster s_k gets eliminated or $a_{max_{s_k}}$ will eliminate a^* which is dealt later.

3 $a_i \in s_k, a^* \in C_g$ and g_{s_k} happens which is Proposition 4(case $a1$)

4 $a_i \in s_k, a^* \notin C_g$ and g_{s_k} happens which is Proposition 4(case $a2$)

Taking summation of the events mentioned above($a1$-$a4$) gives us an upper bound on the regret given that the optimal arm a^* is still surviving,

$$\underbrace{\sum_{i \in A_{s^*}} \frac{32}{\Delta_i^3}}_{\text{case a1}} + \underbrace{\sum_{i \in A \setminus A_{s^*}} \frac{32}{\Delta_i^3}}_{\text{case a2}} + \sum_{i \in A} \underbrace{\left\{ \left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} \right) \right\}}_{\text{case a3+a4}}$$

$$= \sum_{i \in A} \left\{ \left(\frac{32}{\Delta_i^3} \right) + \left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} \right) \right\}$$

C.2 Case b1:

For any sub-optimal arm still surviving given m_i or g_{s_k} have not happened and $a^* \in s^*$ still surviving then they get pulled $n_{m_i} = n_{g_{s_k}}$ number of times(combining the result of Proposition 3(case $b1$) and Proposition 4(case $b1$)). Hence, 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,

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

C.3 Case b2:

Lastly we have to take into consideration the error bound, that the optimal arm a^* or the optimal cluster s^* gets eliminated by any sub-optimal arm or sub-optimal cluster. This, can happen in 3 ways,

1. In s^*, a^* got eliminated by other arms surviving till m_*. Let, the arms surviving till m_* round be denoted by $A_{s^*}^{'}$. 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 $A_{s^*}^{''}$. As proved in Proposition 3, case $b2$ this regret can be no more than,
$\sum_{i \in A_{s^*}^{'}} \left(\frac{12}{\Delta_i^3} \right) + \sum_{i \in A_{s^*}^{''} \setminus A_{s^*}^{'}} \left(\frac{12}{b^3} \right)$

We also see that here, we are concerned only within s^* because of our assumption that there is only one $a^* \in A$.

2. $a^* \in C_g$ and s^* gets eliminated by some other cluster. This is equivalent to Proposition 4, case $b2$

3. $a^* \notin C_g$ and s^* gets eliminated by some other cluster. This is equivalent to Proposition 4, case $b3$

Combining cases $b21, b22$ and $b23$ as mentioned above we can show,

$$\underbrace{\sum_{i \in A'_{s^*}} \left(\frac{12}{\Delta_i^3}\right) + \sum_{i \in A''_{s^*} \setminus A'_{s^*}} \left(\frac{12}{b^3}\right)}_{\text{case b21}} + \underbrace{\sum_{i \in A'} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case b22}} + \underbrace{\sum_{i \in A'' \setminus A'} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{b^{4\rho_s-1}}\right)}_{\text{case b23}}$$

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

$$R_T \leq \sum_{i \in A} \Bigg\{ \underbrace{\left(\frac{32}{\Delta_i^3}\right)}_{\text{case a1+a2}} + \underbrace{\left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right)}_{\text{case a3}} + \underbrace{\left(\Delta_i + \frac{32 \log (T \frac{\Delta_i^4}{16})}{\Delta_i}\right) + \left(\Delta_i + \frac{32\rho_s \log (T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}\right)}_{\text{case b1}} \Bigg\} +$$

$$\underbrace{\sum_{i \in A_{s^*}} \left\{ \left(\frac{12}{\Delta_i^3}\right) + \left(\frac{12}{b^3}\right) \right\} + \sum_{i \in A} \left\{ \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right) + \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{b^{4\rho_s-1}}\right) \right\}}_{\text{case b2}} + max_{i:\Delta_i \leq b} \Delta_i T$$

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

$$\sum_{i \in A_{s^*}: 0 \leq \Delta_i \leq b} \left\{ \left(\frac{12}{\Delta_i^3}\right) + \left(\frac{12}{b^3}\right) \right\} + \sum_{i \in A: 0 \leq \Delta_i \leq b} \left\{ \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right) + \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{b^{4\rho_s-1}}\right) \right\} +$$

$$max_{i:\Delta_i \leq b} \Delta_i T \qquad \square$$

Remark 15. *Considering the case when $\rho_a \in (0, 1]$ and $\rho_s \in (0, 1]$, the total regret bound can be rewritten as*

$$R_T \leq \sum_{i \in A: \Delta_i \geq b} \Bigg\{ \underbrace{\left(\frac{2^{1+4\rho_a} \rho_a^{2\rho_a} T^{1-\rho_a}}{\Delta_i^{4\rho_a-1}}\right)}_{\textit{case a, Arm Elim}} + \underbrace{\left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right)}_{\textit{case a, Clus Elim}} + \underbrace{\left(\Delta_i + \frac{32\rho_a \log (T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i}\right)}_{\textit{case b1, Arm Elim}} +$$

$$+ \underbrace{\left(\Delta_i + \frac{32\rho_s \log (T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}\right)}_{\textit{case b1, Clus Elim}} \Bigg\} + \sum_{i \in A_{s^*}: 0 \leq \Delta_i \leq b} \left\{ \underbrace{\left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a+\frac{3}{2}}}{\Delta_i^{4\rho_a-1}}\right) + \left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a+\frac{3}{2}}}{b^{4\rho_a-1}}\right)}_{\textit{case b2, Arm Elim}} \right\}$$

$$+ \sum_{i \in A: 0 \leq \Delta_i \leq b} \left\{ \underbrace{\left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right) + \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{b^{4\rho_s-1}}\right)}_{\textit{case b2, Clus Elim}} \right\} + max_{i:\Delta_i \leq b} \Delta_i T$$

So we see that two terms can become most significant terms in the regret bound. One is the term $\left(K \frac{32\rho_s \log (T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}\right)$ and second is the terms like $\left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right)$. Hence, its evident from the result that different values of ρ_a and ρ_s affects the regret bound differently. When K is large and p is small it is advantageous to run $\rho_a \leq \rho_s$ for any round m. This will aggressively eliminate arms within cluster while cluster elimination will be more conservative since each cluster will contain a large number of arms

it is a good idea to eliminate clusters in the later rounds when sufficient exploration has been done. In the algorithm ρ_a and ρ_s are decreased in a graded manner so that the risk is low in the initial rounds when sufficient exploration has not been done.

Remark 16. *Another important distinction we want to make is the apparent use of clustering. So when UCB-Clustered is run with $p = 1$, that is just one cluster(so no cluster elimination, only arm elimination) and when UCB-Clustered is run with only Cluster Elimination(Proposition 4) the regret for using just arm Elimination (the regret term in **Remark 14**) turns out to be better for the same value of $\rho_a = \rho_s$. So, all the observations from **Remark 3-7** holds for **Remark 14**. But one important distinction we make here for the error term which makes the algorithms safe. So comparing **Remark 14 and Theorem 2**, we see that when we use just the arm elimination and do not do any clustering, the error term is $\sum_{i\in A:0\leq\Delta_i\leq b}\left\{\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\Delta_i^{4\rho_a-1}}\right)+\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{b^{4\rho_a-1}}\right)\right\}$ while when we do both clustering and arm elimination the error term is $\sum_{i\in A_{s^*}:0\leq\Delta_i\leq b}\underbrace{\left\{\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\Delta_i^{4\rho_a-1}}\right)+\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{b^{4\rho_a-1}}\right)\right\}}_{\text{case b2, Arm Elim}}+$ $\sum_{i\in A:0\leq\Delta_i\leq b}\underbrace{\left\{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right)+\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{b^{4\rho_s-1}}\right)\right\}}_{\text{case b2, Clus Elim}}$. While looking at the error term for both the cases we see that, $\sum_{i\in A:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\Delta_i^{4\rho_a-1}}\right)+\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{b^{4\rho_a-1}}\right)$ is more than $\sum_{i\in A_{s^*}:0\leq\Delta_i\leq b}\underbrace{\left\{\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\Delta_i^{4\rho_a-1}}\right)+\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{b^{4\rho_a-1}}\right)\right\}}_{\text{case b2, Arm Elim}}$ (depending on how we choose p since $|A_{s^*}|\leq\lceil\frac{A}{p}\rceil$) whereas we can make the term $\sum_{i\in A:0\leq\Delta_i\leq b}\underbrace{\left\{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{\Delta_i^{4\rho_s-1}}\right)+\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{b^{4\rho_s-1}}\right)\right\}}_{\text{case b2, Clus Elim}}$ sufficiently less by choosing a sufficient $\rho_s >> \rho_a$(**Definition 4, Appendix E**) thereby reducing the risk of eliminating a^*. An intuitive justification is that since we are reducing ρ_a and ρ_s very fast, and if we partition the action space into clusters and for arm elimination only consider each cluster individually, the probability of eliminating a^* reduces. This is shown in the experiment in **Figure 3**. In this experiment we show the regret for 20 arms, over $p = \{1, 4, 10\}$ and $T = 60000$ timesteps, averaged over 100 independent runs for Bernoulli distributions, where $r_{i:a_i\neq a^*} = 0.06, \forall i \in A$ and $r^* = 0.1$. Here, $\psi_m = 1$ and ρ_s, ρ_a are taken from **Definition 4, Appendix E**. In this case we see that, since ρ_a is decreased very fast, the optimal arm a^* gets eliminated most of the time for no clustering $p = 1$. While a balance of p, ρ_a and ρ_s gives a much better result. ClusUCB with $p = 4$ and 10 perform better than MOSS, while $p = 1$ with just arm elimination does not converge and $p = 5$ with just Cluster elimination and no arm elimination also does not converge. As proved in **Proposition** 4, regret for using just cluster elimination is higher than using just arm elimination. A balance of cluster and arm elimination works best. For using just cluster elimination in ClusUCB($1 < p \leq \frac{K}{2}$) we stop when we are left with one cluster and output the max payoff arm of that cluster.*

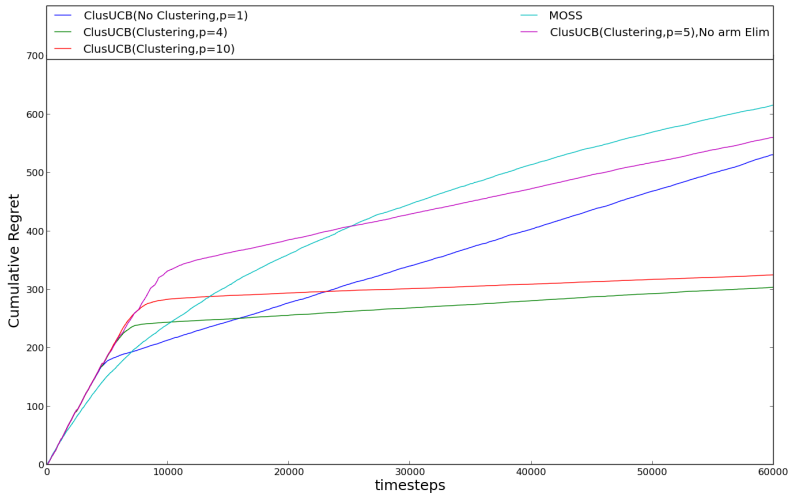

Figure 3: ClusUCB for various p

D Appendix D

In this section we discuss about the various exploration regulatory function that we can use to control exploration. A similar topic has already been handled in Liu and Tsuruoka (2016) where they have introduced mainly three types of regulatory factors(d_i) to the term $\left\{\hat{r}_i + \sqrt{\frac{d_i \log T \tilde{\Delta}_m^2}{2n_m}}\right\}$ which is used for selecting an arbitrary arm a_i in the t-th timestep which maximizes the above term. This regulatory term d_i can be of the form as $\frac{T}{t_i}$, $\frac{\sqrt{T}}{t_i}$ and $\frac{\log T}{t_i}$, where t_i is the number of times an arm a_i has been sampled. One has to choose a regulatory factor based on how fast the algorithm should taper its exploration in the later rounds since with time, t_i increases and only the numerator decides how fast the exploration must decrease.

We also deploy a similar exploration regulatory factor called ψ_m for two different purposes. The first reason comes from the fact that because of limited exploration in the initial rounds the probability for sub-optimal arm elimination is low. Secondly, the increased risk(error bound) stemming from the fact that the optimal arm might get eliminated by any sub-optimal arm because exploration is low. m signifies the round number and $m = 0, 1, ..., \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$.

Number	Definition	Upper Bound	Lower Bound	Remarks
1	$\psi_m = \frac{\sqrt{\log T}}{2^m}$	$\sqrt{\log T}$	$\frac{\sqrt{e \log T}}{\sqrt{T}}$	Used in the algorithm.
2	$\psi_m = \frac{\log T}{2^m}$	$\log T$	$\frac{\sqrt{e} \log T}{\sqrt{T}}$	Decreases at a lesser rate than Definition 1.
3	$\psi_m = K^2 T$	$K^2 T$	$K^2 T$	Constant factor. Used in **Remark 3**.
4	$\psi_m = 2^m \sqrt{T}$	$\frac{T}{\sqrt{e}}$	$\sqrt{T}$	Increasing factor. Can be used in cases where the variance is high. Highly efficient for decreasing risk but also increases regret.
5	$\psi_m = K^{\frac{1}{\rho_s}}$	$K\sqrt{\frac{T}{e}}$	K	Used in **Remark 18**. ρ_s from **Definition 1, Appendix E**.

Depending on how fast the exploration must be reduced/increased we might choose any one of the above. So ψ_m is dependent on round as opposed to each timestep or the number of times an individual arm is sampled as done in Liu and Tsuruoka (2016). Also as m increases, ψ_m decreases/increases and we have a tapered/increased exploration(for definition $1-2$, definition 4 respectively). We also point out that now our number of pulls becomes $n_m = \left\lceil \frac{2\log(\psi_m T \epsilon_m^2)}{\epsilon_m} \right\rceil$ and both the arm elimination and cluster elimination confidence interval becomes $\sqrt{\frac{\rho_a \log(\psi_m T \epsilon_m^2)}{2n_m}}$ and $\sqrt{\frac{\rho_s \log(\psi_m T \epsilon_m^2)}{2n_m}}$ respectively. The rest of the theoretical analysis remains unchanged.

But this $\psi_m = K^2 T$ significantly affects two factors, the maximum regret suffered for not eliminating arms and clusters and the error bound. Firstly(in **Theorem 1, case** $b1$, it becomes $\sum_{i \in A:\Delta_i \geq b} \left\{ \left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{(K^2T)^{\rho_s} (\Delta_i)^{4\rho_s - 1}} \right) \right\}$ from case a) in Theorem 1 and $\sum_{i \in A:0 \leq \Delta_i \leq b} \left(\frac{T^{1-\rho_s} 2^{2\rho_s+3}}{(K^2T)^{\rho_s} (b)^{4\rho_s-1}} \right)$ from **case** $b2$**, Theorem 1**. So, in both the cases we see that the exploratory regulatory factor actually decreases our risk by decreasing the error bound and at the same time increases the probability of arm or cluster elimination.

For the regret contributing term $\left(\frac{32\rho_s \log(\psi_m T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right) = \left(\frac{32\rho_s \log(K^2T^2 \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right)$, the term $\rho_s \log(K^2T^2 \frac{\Delta_i^4}{16\rho_s^2})$ can be sufficiently balanced by choosing appropriate $\rho_s = max\left\{ \frac{1}{2^m}, \frac{1}{4} \right\}$ as shown in analysis of **Theorem** 1**, Remark 3**.

Remark 17. *Considering the case when $\rho_a \in (0,1]$, $\rho_s \in (0,1]$ and taking ψ_m , the total regret bound can be rewritten as*

$$R_T \leq \sum_{i \in A:\Delta_i \geq b} \left\{ \left(\frac{2^{1+4\rho_a} \rho_a^{2\rho_a} T^{1-\rho_a}}{\psi_m^{\rho_a} \Delta_i^{4\rho_a - 1}} \right) + \left(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\psi_m^{\rho_s} \Delta_i^{4\rho_s - 1}} \right) + \left(\Delta_i + \frac{32\rho_a \log(\psi_m T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} \right) + \right.$$
$$\left. + \left(\Delta_i + \frac{32\rho_s \log(\psi_m T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right) \right\} + \sum_{i \in A_{s^*}:0 \leq \Delta_i \leq b} \left\{ \left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi_m^{\rho_a} \Delta_i^{4\rho_a - 1}} \right) + \left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi_m^{\rho_a} b^{4\rho_a - 1}} \right) \right\} +$$
$$\sum_{i \in A:0 \leq \Delta_i \leq b} \left\{ \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\psi_m^{\rho_s} \Delta_i^{4\rho_s - 1}} \right) + \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\psi_m^{\rho_s} b^{4\rho_s - 1}} \right) \right\} + max_{i:\Delta_i \leq b} \Delta_i T$$

E Appendix E

In this section we mention the various definitions of ρ_a and ρ_s. All of this definitions bounds the value of ρ_a and ρ_s between $(0, 1]$ with some of the bounds explicitly mentioning the lower and upper limits of ρ_a and ρ_s. m signifies the round number and $m = 0, 1, ..., \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$.

Number	Definition(ρ_s, ρ_a)	Upper Bound(ρ_s, ρ_a)	Lower Bound(ρ_s, ρ_a)	Remarks
1	$\rho_s = \frac{1}{2^m}, \rho_a = \frac{1}{2^m}$	$1, 1$	$\sqrt{\frac{e}{T}}, \sqrt{\frac{e}{T}}$	Used in algorithm.
2	$\rho_s, \rho_a = max\left\{\frac{1}{2^m}, \frac{1}{4}\right\}$	$1, 1$	$\frac{1}{4}, \frac{1}{4}$	A less risky definition having a higher lower bound than Definition 1 since $\sqrt{\frac{e}{T}} < \frac{1}{4}$ if $T > 16e$
3	$\rho_s = \frac{1}{2^m}, \rho_a = \frac{1}{2^{2m}}$	$1, 1$	$\sqrt{\frac{e}{T}}, \frac{e}{T}$	A more aggressive arm elimination and conservative cluster elimination. Increases risk but reduces expected regret if the optimal arm survives.
4	$\rho_s = \frac{1}{2^{2m+1}}, \rho_a = \frac{1}{2^{4m+1}}$	$\frac{1}{2}, \frac{1}{2}$	$\frac{e}{2T}, \frac{e^2}{2T^2}$	A more aggressive arm elimination and conservative cluster elimination definition than Definition 3. Used in Experiment.

Remark 18. *Putting the value of $\psi_m = (K)^{\frac{1}{\rho_s}}$, using **Appendix E, Definition 1** for ρ_s and $b = \sqrt{\frac{K \log K}{T}}$ in $\sum_{i \in A: \Delta_i \geq b} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{\psi_m^{\rho_s} \Delta_i^{4\rho_s-1}} \right) = K \left(\frac{T^{1-\sqrt{\frac{e}{T}}} (\sqrt{\frac{e}{T}}^{2\sqrt{\frac{e}{T}}}) 2^{2\sqrt{\frac{e}{T}}+3}}{(K)^{\sqrt{\frac{T}{e}} * \sqrt{\frac{e}{T}}} (\Delta_i)^{4 * \sqrt{\frac{e}{T}} - 1}} \right) \approx \frac{8KT\Delta_i}{K} = \frac{8T\sqrt{K \log K}}{\sqrt{T}} = 8\sqrt{KT \log K}$. A similar result can also be shown for $\sum_{i \in A: 0 \leq \Delta_i \leq b} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s+3}}{b^{4\rho_s-1}} \right) \approx \frac{8KT\Delta_i}{K} = \sqrt{eTK}$ for $b = \sqrt{\frac{e}{T}}$.*

For the term $\sum_{i \in A: \Delta_i \geq b} \frac{32\rho_s \log(\psi_m T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}$ by setting $b \approx \sqrt{\frac{K \log K}{T}}$ and the above parameter values for ψ_m, ρ_s in the term we get, $\frac{32K\rho_s \log(\psi_m T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \leq \frac{32K\rho_s \log(\psi_m)}{\Delta_i} + \frac{32K\rho_s \log(T \frac{\Delta_i^4}{\rho_s^2})}{\Delta_i}$

$$= \frac{32K\sqrt{\frac{e}{T}}\log(K)\sqrt{\frac{T}{e}}}{\sqrt{\frac{K\log K}{T}}} + \frac{32K\sqrt{\frac{e}{T}}\log(T^2\frac{K^2(\log K)^2}{eT^2})}{\sqrt{\frac{K\log K}{T}}} = 32\sqrt{KT} + \frac{32\sqrt{eK}\log(K^2(\log K)^2)}{\sqrt{\log K}}$$

So, the total bound comes of as $\max\left\{32\sqrt{KT} + \frac{32\sqrt{eK}\log(K^2(\log K)^2)}{\sqrt{\log K}}, 8\sqrt{KT\log K}\right\}$

F Appendix F

Algorithm	Cumulative Regret Upper Bound
UCB1	$\min\left\{O(\sqrt{KT\log T}), O\left(\frac{K\log T}{\Delta}\right)\right\}$
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)	$O\left(\frac{16K}{\Delta}\log\left(\frac{2K}{\Delta\delta}\right)\right)$, where δ is the error probability
UCB-Improved	$\min\left\{O\left(\sqrt{KT}\frac{log(K\log K)}{\sqrt{\log K}}\right), O\left(\frac{K\log(T\Delta^2)}{\Delta}\right)\right\}$
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	$\min\left\{O\left(\frac{\sqrt{KT}\log(\frac{K^4(\log K)^2}{T})}{\sqrt{\log K}}\right), O\left(\frac{K\rho_s\log(\frac{T\Delta^4}{\rho_s^2})}{\Delta}\right)\right\}$, where $\rho_s \in (0, 1]$