UCB with clustering and improved exploration

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 both MOSS and UCB-Improved. 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. We also empirically test our algorithm against many recent algorithms like UCB-V, KL-UCB and DMED and obtain favourable results.

1 Introduction

In this paper, we consider the stochastic multi-armed bandit problem, a classical problem in sequential decision making. In this setting, a learning algorithm is provided with a set arms with reward distributions unknown to the algorithm. The learning proceeds in an iterative fashion, where in each round, the algorithm chooses an arm and receives a stochastic reward that is drawn from a stationary distribution specific to the arm selected. Given the goal of maximizing the cumulative reward, the learning algorithm faces the exploration-exploitation dilemma, i.e., in each round should the algorithm select the arm which has the highest observed mean reward so far (*exploitation*), or should the algorithm choose a new arm to gain more knowledge of the true mean reward of the arms and thereby avert a sub-optimal greedy decision (*exploration*).

Formally, let r_i, $i = 1, \ldots, K$ denote the mean rewards of the K arms and $r^* = \max_i r_i$ the optimal mean reward. The objective in the stochastic bandit problem is to minimize the cumulative regret, which is defined as follows:

$$R_T = r^*T - \sum_{i \in A} r_i N_i,$$

where T is the number of rounds, $N_i = \sum_{m=1}^{n} I(I_m = i)$ is the number of times the algorithm chose arm i up to round T. The expected regret of an algorithm after T rounds can be written as

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \mathbb{E}[N_i]\Delta_i,$$

where $\Delta_i = r^* - r_i$ denotes the gap between the means of the optimal arm and of the i-th arm.

An early work involving a bandit setup is [14], where the author deals the problem of choosing between two treatments to administer on patients who come in sequentially. Following the seminal work of Robbins [12], bandit algorithms have been extensively studied in a variety of applications. From a theoretical standpoint, an asymptotic lower bound for the regret was established in [10]. In particular, it was shown there that for any consistent allocation strategy, we have

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

where $D(p_i||p^*)$ is the Kullback-Leibler divergence between the reward densities p_i and p^*, corresponding to arms with mean r_i and r^*, respectively.

There have been several algorithm with strong regret guarantees. The foremost among them is UCB1 [6], which has a regret upper bound of $O\left(\frac{K \log T}{\Delta}\right)$, where $\Delta = \min_{i:\Delta_i > 0} \Delta_i$. This result is asymptotically order-optimal for the class of distributions considered. However, the worst case regret of UCB1 can be as bad as $\Omega\left(\sqrt{TK \log T}\right)$ - see [3]. In the latter reference, the authors propose the MOSS algorithm and establish that

Preliminary work. Under review by AISTATS 2017. Do not distribute.

the worst case regret of MOSS is $O\left(\sqrt{TK}\right)$ which improves upon UCB1 by a factor of order $\sqrt{\log T}$ and a gap-dependent regret of $\sum_{i:\Delta_i>0} \frac{K\log(2+T\Delta_i^2/K)}{\Delta_i}$. However, in certain regimes, the gap-dependent regret bound of MOSS can be worse than UCB1 . The UCB-Improved algorithm, proposed in [5], is a round-based algorithm variant of UCB1 that has a gap-dependent regret bound of $O\left(\frac{K\log T\Delta^2}{\Delta}\right)$ and a worst case regret of $O\left(\sqrt{TK\log K}\right)$. An algorithm is *round-based* if it pulls all the arms equal number of times in each round and then proceeds to eliminate one or more arms that it identifies to be sub-optimal.

Can you give a refer-

Our Work As mentioned in a recent work [11], UCB-Improved has two shortcomings: **(i)** a significant number of pulls are spent in early exploration, since each round m of UCB-Improved involves pulling every arm an identical $n_m = \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$ number of times. The quantity $\tilde{\Delta}_m$ is initialized to 1 and halved after every round. **(ii)** In UCB-Improved, arms are eliminated conservatively, i.e, only after $\tilde{\Delta}_m < \frac{\Delta_i}{2}$, the sub-optimal arm i is discarded. This is disadvantageous when K is large and the gaps are identical ($r_1 = r_2 = .. = r_{K-1} < r^*$) and small.

We propose a variant of UCB algorithm (henceforth referred to as ClusUCB) that incorporates clustering and an improved exploration scheme. ClusUCB is a round-based algorithm that starts with a partition of the arms into small clusters, each having uniform set of arms. The clustering is done at the start with a pre-specified the number of clusters. Each round of ClusUCB involves both (individual) arm elimination as well as cluster elimiation. While the conditions governing the arm and cluster eliminations are inspired by UCB-Improved, though the exploration factors governing these conditions are relatively more aggressive than that in UCB-Improved.

From a theoretical analysis that involves carefully setting the exploration factors for arm and cluster elimination conditions, we observe that the gap-dependent regret of ClusUCB is better than that of UCB-Improved (see Table 1). On the other hand, the gap-independent bound of ClusUCB is of the same order as UCB-Improved, i.e., $O\left(\sqrt{KT\log K}\right)$. While ClusUCB is not able to match the gap-independent bound of $O(\sqrt{KT})$ for MOSS, the gap-dependent bound of ClusUCB is better than MOSS when $\sqrt{\log(KT)} > K$.

On a setup with $r_1 = r_2 = .. = r_{K-1} < r^*$, small Δ_i's and large K, we observe empirically that ClusUCB outperforms UCB-Improved[5] and MOSS[3] as well as other popular stochastic bandit algorithms such as DMED[9], UCB-V[4] and KL-UCB[8].

Table 1: Gap-dependent regret bounds for different bandit algorithms

Algorithm	Upper bound
UCB1	$O\left(\frac{K\log T}{\Delta}\right)$
UCB-Improved	$O\left(\frac{K\log(T\Delta^2)}{\Delta}\right)$
MOSS	$O\left(\frac{K\log\left(T\Delta^2/K\right)}{\Delta}\right)$
ClusUCB	$O\left(\frac{K\log\left(\frac{T\Delta^2}{\sqrt{\log(KT)}}\right)}{\Delta}\right)$

The rest of the paper is organized as follows: In Section 2, we present the ClusUCB algorithm. In Section 3, we present the associated regret bounds and sketch the proofs in Section 4. In Section 5, we present the numerical experiments and provide concluding remarks in Section 6.

2 Clustered UCB

Notation. We denote the set of arms by A, with the individual arms labeled $a_i, i = 1, \ldots, K$. We denote an arbitrary round of ClusUCB by m. We denote an arbitrary cluster by s_k, the subset of arms within the cluster s_k by A_{s_k} and the set of clusters by S with $|S| = p \leq K$. Here p is a pre-specified limit for the number of clusters. We denote the sample mean of the rewards seen so far for arm i by $\hat{r}_i$, for the best arm within a cluster s_i by $\hat{r}_{max_{s_k}} \in s_k$ and the worst arm within a cluster s_i by $\hat{r}_{min_{s_k}} \in s_k$ We assume that all arms' rewards are bounded in $[0, 1]$. For simplicity, we assume that the optimal arm is unique and denote it by a^*, with s^* denoting the corresponding cluster.

The algorithm

Algorithm 1 presents the pseudocode of ClusUCB.

ClusUCB begins with a initial clustering of arms that is performed by random uniform allocation - see Algorithm 2. The set of clusters S thus obtained satisfies $|S| = p$, with individual clusters having a size that is bounded above by $\ell = \left\lceil \frac{K}{p} \right\rceil$.

Each round of ClusUCB involves both individual arm as

Algorithm 1 ClusUCB

```
Input: A partition of $A$ into clusters from Algorithm 2, exploration parameters $\rho_a$, $\rho_s$, time horizon $T$.
Initialization: Set $B_0 := A$, $\psi_m = \log T$, $\epsilon_0 := 1$ and $w_0 := 1$.
for $m = 0, 1, .. \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$ do
    Pull each arm in $s_k$ $n_m = \left\lceil \frac{2 \log (\psi_m T \epsilon_m^2)}{\epsilon_m} \right\rceil$ number of times, $\forall s_k \in S$
    Arm Elimination
        For each cluster $s_k \in S$, delete arm $a_i \in s_k$ if
        $$\hat{r}_i + \sqrt{\frac{\rho_a \log (\psi_m T \epsilon_m^2)}{2n_m}} < \max_{a_j \in s_k} \left\{ \hat{r}_j - \sqrt{\frac{\rho_a \log (T \epsilon_m^2)}{2n_m}} \right\}.$$
    Cluster Elimination
        Delete cluster $s_k \in S$ if
        $$\max_{a_i \in s_k} \left\{ \hat{r}_i + \sqrt{\frac{\rho_s \log (\psi_m T \epsilon_m^2)}{2n_m}} \right\} < \max_{a_j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho_s \log (T \epsilon_m^2)}{2n_m}} \right\}.$$
    Set $\epsilon_{m+1} := \frac{\epsilon_m}{2}$, $w_{m+1} := 2w_m$
    Stop if $|B_m| = 1$ and pull $\max_{a_i \in B_m} \hat{r}_i$ till $T$ is reached.
end for
```

Algorithm 2 Clustering by random uniform allocation

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

well as cluster elimination conditions that are inspired by UCB-Improved. Notice that, unlike UCB-Improved, there is no longer a single point of reference based on which we are eliminating arms. Instead now we have as many reference points to eliminate arms as number of clusters formed. Further, the exploration factors $\rho_a \in (0, 1]$ and $\rho_s \in (0, 1]$ governing the arm and cluster elimination conditions, respectively, are relatively more aggressive than that in UCB-Improved. The number n_m of times each arm is pulled per round in ClusUCB is lower than that of UCB-Improved, which used $n_m = \left\lceil \frac{2log(T\epsilon_m^2)}{\epsilon_m^2} \right\rceil$ and Median-Elimination, which used $n_m = \frac{4}{\epsilon^2} \log \left(\frac{3}{\delta} \right)$, where ϵ, δ are confidence parameters.

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 probability. In particular, when K is large and the gaps Δ_i are small, it is efficient to remove sub-optimal arms quickly rather getting tied down in hopeless exploration. The theoretical analysis confirms this intuition, as choosing $\rho_a = \rho_s = 1/2^m$ and $\psi_m = K^2 T$, ClusUCB is able to achive a better gap-dependent regret upper bound as compared to UCB-Improved and MOSS.

3 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* $\mathbb{E}[R_T] \leq \sum_{i \in A : \Delta_i \geq b} \Bigg\{ \Bigg(2\Delta_i + \frac{2^{1+4\rho_a} \rho_a^{2\rho_a} T^{1-\rho_a}}{\psi_m^{\rho_a} \Delta_i^{4\rho_a - 1}} \Bigg) + \Bigg(\frac{2^{2+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\psi_m^{\rho_s} \Delta_i^{4\rho_s - 1}} \Bigg) + \Bigg(\frac{32\rho_a \log (\psi_m T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} \Bigg) + \Bigg(\frac{32\rho_s \log (\psi_m T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \Bigg) \Bigg\} + \sum_{i \in A_{s*} : 0 \leq \Delta_i \leq b} \Bigg\{ \Bigg(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi_m^{\rho_a} \Delta_i^{4\rho_a - 1}} \Bigg) + \Bigg(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi_m^{\rho_a} b^{4\rho_a - 1}} \Bigg) \Bigg\} + \sum_{i \in A : 0 \leq \Delta_i \leq b} \Bigg\{ \Bigg(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + 3}}{\psi_m^{\rho_s} \Delta_i^{4\rho_s - 1}} \Bigg) + \Bigg(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + 3}}{\psi_m^{\rho_s} b^{4\rho_s - 1}} \Bigg) \Bigg\} + max_{i : \Delta_i \leq b} \Delta_i T$ *for all* $b \geq \sqrt{\frac{e}{T}}$, *where* $\rho_a, \rho_s \in (0, 1]$ *are the arm and cluster elimination parameter respectively,* 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.*

The proof of this is given in Appendix A.(Supplementary Material)

Remark 1 *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 2(Appendix B), Remark 3(Appendix C) and their various definitions are given in Appendix G. A discussion on ex-*

ploration regulatory factor(ψ_m) and its effect is deferred to Appendix F. A discussion on the three approaches of using only arm elimination in Algorithm 1 or using only cluster elimination or using both is deferred to Remark 4(Appendix G).

Corollary 1 *Taking into account the regret bound in Theorem 1, putting the value of $\psi_m = K^2T$ and using $\rho_a = max\left\lbrace\dfrac{1}{2^m}, \dfrac{1}{4}\right\rbrace, \rho_s = max\left\lbrace\dfrac{1}{2^m}, \dfrac{1}{4}\right\rbrace$ and $b \approx \sqrt{\dfrac{K\log K}{T}}$ we can have a gap independent regret bound as $\left\lbrace 6\sqrt{KT} + 2\dfrac{\sqrt{KT}}{p} + 64\sqrt{KT\log K} + \dfrac{32\log\left(\log K\right)}{\sqrt{\log K}}\right\rbrace$.*

The proof of this corollary is given in Appendix D.(Supplementary Material)

So, the total order come of as $O(\sqrt{KT\log K})$ which matches the order of $O(\sqrt{KT})$ of MOSS except the term $\sqrt{\log K}$ and is same as the UCB-Improved gap independent bound of $O\big(\sqrt{KT\log K}\big)$.

Corollary 2 *Taking into account the regret bound in Theorem 1, putting the value of $\psi_m = \dfrac{T}{\log(KT)}$ and using $\rho_a = max\left\lbrace\dfrac{1}{2^m}, \dfrac{1}{2}\right\rbrace, \rho_s = max\left\lbrace\dfrac{1}{2^m}, \dfrac{1}{2}\right\rbrace$ and $b \approx \sqrt{\dfrac{K\log K}{T}}$ we can have a gap dependent regret bound as $\sum_{i\in A:\Delta_i\geq b}\left\lbrace\dfrac{12\sqrt{\log(KT)}}{\Delta_i} + \dfrac{64\log\left(T\dfrac{\Delta_i^2}{\sqrt{\log(KT)}}\right)}{\Delta_i}\right\rbrace + \sum\limits_{i\in A_{s^*}:0\leq\Delta_i\leq b}\dfrac{2.8\sqrt{\log(KT)}}{\Delta_i} + \sum\limits_{i\in A:0\leq\Delta_i\leq b}\dfrac{8\sqrt{\log(KT)}}{\Delta_i}$.*

The proof of this corollary is given in Appendix E.(Supplementary Material)

When we consider gap-dependent bound, by choosing appropriate values of ρ_a, ρ_s, ψ_m and p we see that the most significant term in the regret is $\sum_{i\in A:\Delta_i\geq b}\dfrac{64\log\left(T\dfrac{\Delta_i^2}{\sqrt{\log(KT)}}\right)}{\Delta_i}$. Hence we see the regret upper bound of $O\left(\dfrac{K\log\left(\dfrac{T\Delta^2}{\sqrt{\log(KT)}}\right)}{\Delta}\right)$ is lower than the gap dependent bound of UCB1 of $O\left(\dfrac{K\log T}{\Delta}\right)$ and that of UCB-Improved, $O\left(\dfrac{K\log T\Delta^2}{\Delta}\right)$. Also comparing against that of MOSS, $O\left(\dfrac{K\log(T\Delta^2/K)}{\Delta}\right)$ the regret bound of our algorithm is better if $\sqrt{\log(KT)} > K$. A table for the various regret upper bounds is given in Appendix H. Thus we see that ρ_a, ρ_s actually helps in reducing the constant associated with the factor $\sqrt{KT}$ and the $\dfrac{\log\left(\psi_m\dfrac{T\Delta_i^4}{\rho_s^2}\right)}{\Delta_i}$ term and also both ψ_m and ρ_a, ρ_s helps in stabilizing the term $\log\left(\psi_m\dfrac{T\Delta_i^4}{\rho_s^2}\right)$.

4 Proof Sketch

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. By simply taking $p = 1$ we can achieve arm elimination without any cluster elimination. In second step, we try to bound the probability of cluster elimination(without any arm elimination) with all arms within it. A slight modification to the algorithm allows us to do this. By taking $p > 1$, removing the arm elimination condition, stopping when we are just left with one cluster and pulling the $max\{\hat{r}_i\}$, where $a_i \in B_m$ we can achieve this. This is shown in Proposition 2. Finally in step three, in the proof of Theorem 1 we combine both arm elimination and cluster elimination to get the regret upper bound.

Proposition 1 *Considering only the arm elimination condition and $p = 1$ the total regret till T is upper bounded by $\mathbb{E}[R_T] \leq \sum\limits_{i\in A:\Delta_i\geq b}\left\lbrace\left(\dfrac{2^{1+4\rho_a}\rho_a^{2\rho_a}T^{1-\rho_a}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right) + \left(\Delta_i + \dfrac{32\rho_a\log\left(\psi_m T\dfrac{\Delta_i^4}{16\rho_a^2}\right)}{\Delta_i}\right) + \left(\dfrac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right)\right\rbrace + \sum\limits_{i\in A:0\leq\Delta_i\leq b}\left(\dfrac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}b^{4\rho_a-1}}\right) + max_{i:\Delta_i\leq b}\Delta_i T$ for all $b \geq \sqrt{\dfrac{e}{T}}$, where $\rho_a \in (0, 1]$ is the arm elimination parameter, ψ_m is the exploration regulatory factor, p is the number of clusters and T is the horizon.*

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

Thus, we see that the confidence interval term $c_m = \sqrt{\dfrac{\rho_a\log(\psi_m T\epsilon_m^2)}{2n_m}}$ makes the algorithm eliminate an arm a_i as soon as $\sqrt{\rho_a\epsilon_m} < \dfrac{\Delta_i}{2}$. The above result is in contrast with UCB-Improved which only deletes an arm if $\tilde{\Delta}_m < \dfrac{\Delta_i}{2}$, where $\tilde{\Delta}_m$ is initialized at 1 and is halved after every round. We also point out here that when

$c_m = \sqrt{\frac{\rho_a \log(\psi_m T \epsilon_m^2)}{2n_m}}$ 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 $\rho_a \leq \epsilon_m$.

Corollary 3 *For* $\rho_a = 1$ *in the result of proposition* 1, $\sum_{i\in A:\Delta_i \geq b}\Bigg\{\bigg(\frac{2^{1+4\rho_a}\rho_a^{2\rho_a}T^{1-\rho_a}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\bigg) + \bigg(\Delta_i + \frac{32\rho_a \log\big(\psi_m T \frac{\Delta_i^4}{16\rho_a^2}\big)}{\Delta_i}\bigg) + \bigg(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\bigg)\Bigg\} + \sum_{i\in A:0\leq\Delta_i\leq b}\bigg(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}b^{4\rho_a-1}}\bigg) + max_{i:\Delta_i\leq b}\Delta_i T$ *we get a regret bound of* $\sum_{i\in A:\Delta_i\geq b}\bigg(\Delta_i + \frac{44}{\psi_m(\Delta_i)^3} + \frac{32\log(\psi_m T\Delta_i^4)}{\Delta_i}\bigg) + \sum_{i\in A:0\leq\Delta_i\leq b}\frac{12}{\psi_m b^3}$ *for using just arm elimination with* $p = 1$.

Proof 1 *In the result of Proposition* 1 *if we take* $\rho_a = 1$ *then the regret bound becomes* $\sum_{i\in A:\Delta_i\geq b}\bigg(\Delta_i + \frac{44}{\psi_m(\Delta_i)^3} + \frac{32\log(\psi_m T\Delta_i^4)}{\Delta_i}\bigg) + \sum_{i\in A:0\leq\Delta_i\leq b}\frac{12}{\psi_m b^3}$. *From the result we can see that for small* Δ_i *and large* K, *the terms like* $\sum_{i\in A:\Delta_i\geq b}\bigg(\frac{44}{\psi_m(\Delta_i)^3}\bigg) + \sum_{i\in A:0\leq\Delta_i\leq b}\frac{12}{\psi_m b^3}$ *can become the dominant term in the regret rather than* $\sum_{i\in A:\Delta_i\geq b}\frac{32\log(\psi_m T\Delta_i^4)}{\Delta_i}$. *Intuitively, this actually suggests that the algorithm is 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 *and by carefully defining their values enables us to eliminate arms and clusters aggressively and thereby reduce those two terms.*

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

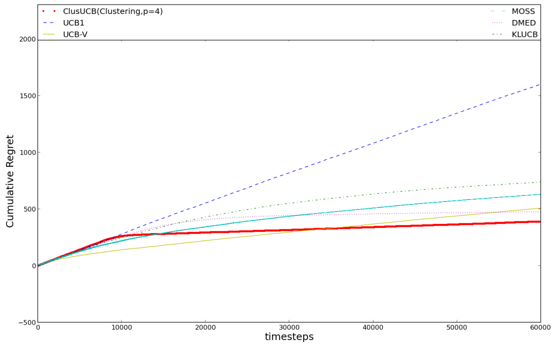


Figure 1: Experiment 1: Regret for various Algorithms. $T = 60000$

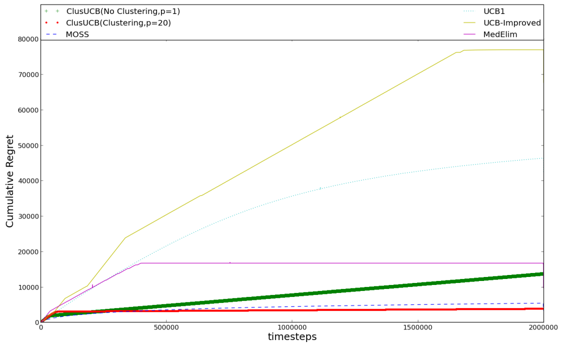


Figure 2: Experiment 2: Regret for various Algorithms. $T = 2 \times 10^6$

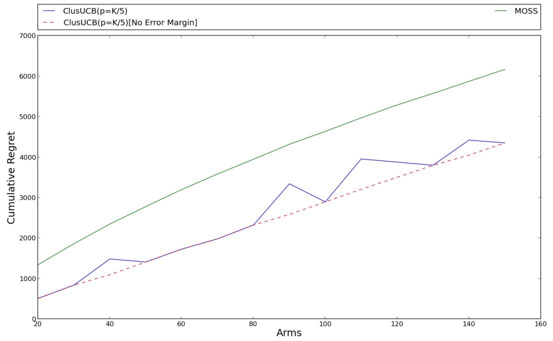


Figure 3: Experiment 3: Regret Growth for ClusUCB and MOSS . $T = 1 \times 10^6$

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([13]) or softmax([13]) or UCB-Tuned([6]) has no proven regret bounds. Again algorithms like UCB-δ([1]) 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([2]). We will not be experimenting against these algorithms but focus on the more recent, algorithms like KL-UCB, DMED, MOSS, UCB1, UCB-V, etc. The codes for KL-UCB and DMED are taken from [7].

In the experimental setup we use $\psi_m = \log T$, $\rho_s = \frac{1}{2^{2m+1}}$ and $\rho_a = \frac{1}{2^{4m+1}}$.

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 $p = 4$ for ClusUCB. The regret is averaged over 100 independent runs over each timestep 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.

The second experiment is conducted over a testbed of 100 arms for the test-cases involving Gaussian reward distribution with expected rewards of the arms $r_{i_{a_i \neq a^*:1-33}} = 0.01$, $r_{i_{a_i \neq a^*:34-99}} = 0.06$ and $r^*_{i=100} = 0.1$. The horizon is set for $T = 2 \times 10^6$ and and $p = 20$ for ClusUCB. The regret is averaged over 100 independent runs over each timestep and is shown in Figure 2. In this test we also show the no-clustering version of ClusUCB algorithm with $p = 1$ and keeping all the other parameters same. 4 other algorithms, MOSS, UCB1, UCB-Improved and Median-Elimination are run over this testbed and shown in this figure. Here, we see that ClusUCB($p = 20$) performs better than all the algorithms mentioned above. We also see that no-clustering and using only arm elimination version of algorithm performs worse than the clustering version which stems from the fact that it outputs sub-optimal arm more often than the clustering version.

The third experiment is conducted over a testbed of $20 - 200$(interval of 10) arms for the test-cases involving Bernoulli reward distribution with expected rewards of the arms $r_{i_{a_i \neq a^*}} = 0.05$ and $r^* = 0.1$. The horizon is set for $T = 1 \times 10^6$ and the parameters of ClusUCB are $p = K/5$ for $K = 20$ to 200. The regret is averaged over 100 independent runs over each timestep and is shown in Figure 3. We only check the growth of regret for the two algorithms MOSS and ClusUCB over this testbed and see that ClusUCB outperforms MOSS and its growth of regret is lesser than MOSS. The jumps in the graph for ClusUCB happens because of the error(eliminating optimal arm) and the margin of error(in red) is also shown in the graph.

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 UCB1, MOSS, UCB-Improved, UCB-V, KL-UCB, 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 Median Elimination with careful choosing of parameters as shown empirically. Also ClusUCB scales well with large K as compared with UCB1, UCB-Improved, MOSS, KL-UCB and Median Elimination. We must also remember as the number of arms increases UCB1, MOSS, KL-UCB and DMED will take more time as all of these algorithms have to build their confidence set over all the arms whereas algorithms like ClusUCB, UCB-Improved and Median Elimination will take much lesser time as they keep on removing sub-optimal arms after each round. Again 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

[1] 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.

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

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

[4] 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.

[5] 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.

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

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

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

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

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

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

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

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

[14] 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 Proof of Proposition 1

Proof 2 *:*

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

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

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

$c_{m_i} = \sqrt{\frac{\rho_a \epsilon_{m_i} \log(\psi_m T \epsilon_{m_i}^2)}{2 * 2\log(\psi_m T \epsilon_{m_i}^2)}} = \frac{\sqrt{\rho_a \epsilon_{m_i}}}{2} = \sqrt{\rho_a \epsilon_{m_i+1}} < \frac{\Delta_i}{4}$*, as* $\rho_a \in (0, 1]$

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 $\sqrt{\rho_a \epsilon_{m_i}} < \frac{\Delta_i}{2}$, $\exists a_i$ 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 $\sqrt{\rho_a \epsilon_{m_i}} < \frac{\Delta_i}{2}$ becomes true for some arm a_i after the m-th round and $c_m = \sqrt{\frac{\rho_a \log(\psi_m T \epsilon_{m_i}^2)}{2 n_{m_i}}}$.

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,

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

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

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

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

$$\begin{aligned}\sum_{i\in A}\left(\frac{2T\Delta_i}{(\psi_m T \epsilon_{m_i}^2)^{\rho_a}}\right) &\leq \sum_{i\in A}\left(\frac{2T\Delta_i}{(\psi_m T \frac{\Delta_i^4}{16\rho_a^2})^{\rho_a}}\right)\\ &\leq \sum_{i\in A}\left(\frac{2^{1+4\rho_a} T^{1-\rho_a} \rho_a^{2\rho_a} \Delta_i}{\psi_m^{\rho_a} \Delta_i^{4\rho_a}}\right)\\ &\leq \sum_{i\in A}\left(\frac{2^{1+4\rho_a} \rho_a^{2\rho_a} T^{1-\rho_a}}{\psi_m^{\rho_a} \Delta_i^{4\rho_a - 1}}\right)\end{aligned}$$

Case b1: *Either an arm a_i is eliminated in round m_i or before or else there is no optimal arm $a^* \in B_{m_i}$.*

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

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

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

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

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

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

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

Summing over all arms in A,

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

Case b2: *Optimal arm a^* is 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 $\sqrt{\rho_a \epsilon_m} < \frac{\Delta_j}{2}$ 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 $\sqrt{\rho_a \epsilon_m} \geq \frac{\Delta_b}{2}$ 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 $\sqrt{\sqrt{\rho_a}\epsilon_m} < \frac{\Delta_b}{2}$ that is $m_b = min\{m | \sqrt{\rho_a \epsilon_m} < \frac{\Delta_b}{2}\}$. Hence, the maximal regret after eliminating the arm a^* is upper bounded by,

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

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

$$\leq \sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i \in A'': m_i > m_*} 4\left(\frac{T^{1-\rho_a}}{\psi_m^{\rho_a} \epsilon_m^{2\rho_a - \frac{1}{2}}}\right)$$

$$\leq \sum_{i \in A'': m_i > m_*} \sum_{m_*=0}^{\min\{m_i, m_b\}} \left(\frac{4T^{1-\rho_a}}{\psi_m^{\rho_a} 2^{-(2\rho_a - \frac{1}{2})m_*}}\right)$$

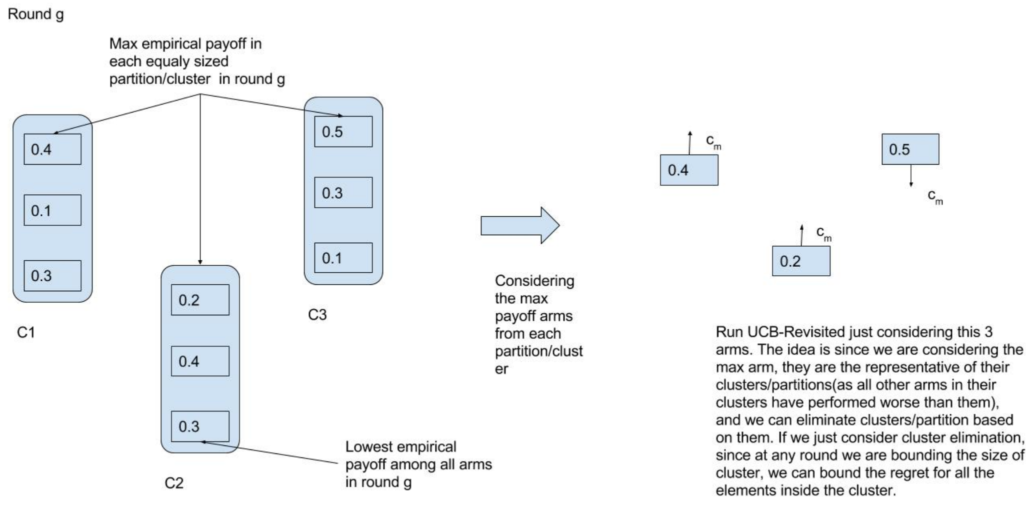
$$\leq \sum_{i \in A'} \left(\frac{4T^{1-\rho_a}}{\psi_m^{\rho_a} 2^{-(2\rho_a - \frac{1}{2})m_*}}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{4}{\psi_m^{\rho_a} 2^{-(2\rho_a - \frac{1}{2})m_b}}\right)$$

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

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

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

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


Round g
Max empirical payoff in each equaly sized partition/cluster in round g
0.4
0.1
0.3
C1
0.2
0.4
0.3
C2
0.5
0.3
0.1
C3
Lowest empirical payoff among all arms in round g
Considering the max payoff arms from each partition/clust er
0.4
c_m
0.2
c_m
0.5
c_m

Run UCB-Revisited just considering this 3 arms. The idea is since we are considering the max arm, they are the representative of their clusters/partitions(as all other arms in their clusters have performed worse than them), and we can eliminate clusters/partition based on them. If we just consider cluster elimination, since at any round we are bounding the size of cluster, we can bound the regret for all the elements inside the cluster.

Figure 4: Cluster Elimination

$$\sum_{i\in A:0\le\Delta_i\le b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}b^{4\rho_a-1}}\right)+max_{i:\Delta_i\le b}\Delta_i T$$

B Proof of Proposition 2

An illustrative diagram explaining Cluster Elimination is given in **Figure 4**. A slight modification to the algorithm allows us to do cluster elimination without any arm elimination. By taking $p>1$, removing the arm elimination condition, stopping when we are just left with one cluster and pulling the $max\{\hat{r}_i\}$, where $a_i\in B_m$ we can achieve this.

Proof 3 *:*

A sketch of the proof is given below,

- *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 $a_{max_{s_k}}\in C_g$ is the max payoff arm in the cluster s_k if the cluster s_k still surviving in round g and call $a_{max_{s_k}}$ as a cluster arm.*
- *Define g-th round(like 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 $a_{max_{s_k}}\in C_g$ gets eliminated.*
  - ***Case a2:** In the g-th round, $a^*\notin C_g$ and bound the maximum probability that a sub-optimal cluster arm $a_{max_{s_k}}$ 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 in 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 $a_{max_{s_k}} \in C_g, \forall s_k \in S$. 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 $\sqrt{\rho_s \epsilon_{g_{s_k}}} \leq \dfrac{\Delta_{a_{max_{s_k}}}}{2}$, $a_{max_{s_k}} \in C_g$, then the cluster s_k(where $\hat{r}_{a_{max_{s_k}}}$ 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 $\ell = \left\lceil \dfrac{K}{p} \right\rceil$.*

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

Case a1: *Some sub-optimal cluster arm $a_{max_{s_k}}$ 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, $c_{g_{s_k}} = \sqrt{\dfrac{\rho_s \log(\psi_m T \epsilon^2_{g_{s_k}})}{2 n_{g_{s_k}}}}$, *where* $0 < \rho_s \leq 1$

Putting the value of $n_{g_{s_k}} = \dfrac{2\log\left(\psi_m T \epsilon^2_{g_{s_k}}\right)}{\epsilon_{g_{s_k}}}$ *in* $c_{g_{s_k}}$,

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

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

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

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

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,

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

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

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

Now, for each round g*, all the elements of* $C_{g_{s_k}}$ *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* $A \supset C_{g_{s_k}}$*, we are pulling all the surviving arms equally in each round 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* $a_{max_{s_k}} \in C_{g_{s_k}}$ *is not eliminated in the* g*-th round by the same probability of* $\left(\frac{2}{(\psi_m T \epsilon_{g_{s_k}}^2)^{\rho_s}}\right)$.

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

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

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

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

Case a2: *Some sub-optimal cluster arm* $a_{max_{s_k}}$ *is not eliminated in round* g_{s_k} *or before and the optimal cluster arm* $a^* \notin C_{g_{s_k}} \subset B_m$.

In the above case we considered that $a^* \in C_{g_{s_k}}$ *being the max-payoff arm from optimal cluster* s^**. Now, if that is not the case and* $\exists a_{max_{s^*}} \in s^*$ *such that* $\hat{r}_{a_{max_{s^*}}} > \hat{r}^*$*, so* $a_{max_{s^*}}$ *will be in* $C_{g_{s_k}}$ *in the* g_{s_k}*-th round. In this case for some sub-optimal arm* $a_{max_{s_k}} \in C_{g_{s_k}}$*, we have to bound the probability* $\mathbb{P}\left\{\hat{r}_{a_{max_{s_k}}} + c_{g_{s_k}}\right\} < \mathbb{P}\left\{\hat{r}_{a_{max_{s^*}}} - c_{g_{s_k}}\right\}$*. 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 in the* g*-th round. So the regret for not eliminating a sub-optimal cluster even when* $a^* \notin C_{g_{s_k}}$*(but still surviving in* s^**) 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,*

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

Case b1: *Either an arm* $a_{max_{s_k}} \in C_{g_{s_k}}$ *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* $a^* \in C_{g_{s_k}}$.

Again, in the g_{s_k}*-th round, the maximum total elements in the cluster* s_k *can be no more than* $\ell = \left\lceil \frac{K}{p} \right\rceil$.

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

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

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

$\sum_{i \in A_{s_k}} \Delta_i \left\lceil \frac{2\log{(\psi_m T \epsilon^2_{g_{s_k}})}}{\epsilon_{g_{s_k}}} \right\rceil \leq \sum_{i \in A_{s_k}} \Delta_i \left\lceil \frac{2\log{(\psi_m T (\frac{\Delta_i}{2\sqrt{\rho_s}})^4)}}{(\frac{\Delta_i}{2\sqrt{\rho_s}})^2} \right\rceil$, *since* $\sqrt{\rho_s \epsilon_{g_{s_k}}} \leq \frac{\Delta_{a_{max_{s_k}}}}{2} \leq \frac{\Delta_i}{2}$ *as* $r_{a_{max_{s_k}}} > r_i, \forall i \in s_k$

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

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

Summing over all p clusters ,

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

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

Case b2: *Cluster s^* containing the optimal arm a^* is eliminated by a sub-optimal cluster and $a^* \in C_g$.*

Firstly, if conditions of case $b1$ holds then the optimal arm $a^* \in C_{g_{s_k}}$ will not be eliminated in round $g_{s_k} = g_*$ or it will lead to the contradiction that $r_{a_{max_{s_k}}} > r^*$ where $a_{max_{s_k}}, a^* \in C_{g_{s_k}}$. In any round g_*, if the optimal arm a^* gets eliminated then for any round from 1 to g_{s_j} all arms $a_{s_j} \in C_{g_{s_k}}, \forall s_j \neq s^*$ 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,

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

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.

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

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

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

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

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

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

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

Case b3: *Cluster s^* containing the optimal arm a^* is eliminated by a sub-optimal cluster and $a^* \notin C_g$.*

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,
$\leq \sum_{i\in A'} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + \frac{3}{2}}}{\psi_m^{\rho_s} \Delta_i^{4\rho_s - 1}} \right) + \sum_{i \in A'' \backslash A'} \left(\frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + \frac{3}{2}}}{\psi_m^{\rho_s} b^{4\rho_s - 1}} \right)$
*Summing up **Case a1**, **Case a2**, **Case b1**,**Case b2** and **Case b3**, the total regret till round g is given by,*

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

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

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

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

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

Remark 2 *Thus, we see that the regret bound for using just arm elimination is actually less than the regret bound for using just cluster elimination. Another takeaway from this result is that a lower value of $\rho_a, \rho_s \in (0, 1]$ makes the algorithm risky(by increasing the error bound), but reduces expected regret whereas a higher value of $\rho_a, \rho_s > 1$ 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 but ρ_s and ρ_a is always bounded that is, $\rho_s, \rho_a \in (0, 1]$(see Appendix E).*

C Proof of Theorem 1

Proof 4 *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}s.t $\exists a_i \in s_k$ happens with $a^* \in s^*$ still surviving. In Proposition 3 we consider only arm elimination and in Proposition 4 we consider only cluster elimination. 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 $r_{max_{s_k}} \geq r_{a_i}, \forall i \in s_k$. Again, $r_{a^*} \geq r_{max_{s_k}}, \forall s_k \in S$.*

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 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 $r_{max_{s_k}} \leq r^*$ and m_i happens. This case was not dealt in Proposition 3 as there we took $p = 1$. Since, now $p > 1$ and $r_{max_{s_k}} \leq r^*$, 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 $\sqrt{\rho_a \epsilon_m} < \frac{\Delta_i^{'}}{2}$ where $\Delta_i^{'} = r_{max_{s_k}} - r_i \geq \Delta_{a_{max_{s_k}}}$ such that $r_i \in s_k$. Plugging in this $\Delta_i^{'}$ in Proposition 3, case a we can derive a similar bound where $\Delta_i^{'} \geq \Delta_{a_{max_{s_k}}}$ because otherwise case 3(next case) will happen before as $\sqrt{\epsilon_m \rho_s} < \frac{\Delta_{a_{max_{s_k}}}}{2}$ and the cluster s_k gets eliminated or $a_{max_{s_k}}$ will eliminate a^* which is dealt later.*

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

4. *$a_i \in s_k, a^* \notin C_{g_{s_k}}$ 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,

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

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

Case b1: *Either an arm a_i is eliminated in round $m_i or g_{s_k}$ or before or else there is no optimal arm $a^* \in B_{m_i}$.*

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} or $n_{g_{s_k}}$ 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,

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

Case b2: *Optimal arm a^* is eliminated by a sub-optimal arm.*

Here, we 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 $\sqrt{\rho_a \epsilon_m} \geq \frac{\Delta_b}{2}$ 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,*
   $$\textstyle\sum_{i \in A_{s*}^{'}} \left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi_m^{\rho_a} \Delta_i^{4\rho_a - 1}} \right) + \sum_{i \in A_{s*}^{''} \setminus A_{s*}^{'}} \left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi_m^{\rho_a} b^{4\rho_a - 1}} \right)$$
   We also see that here, we are concerned only within s^* because of our assumption that there is only one $a^* \in A$ and clusters are fixed.

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,

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

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

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

$$\underbrace{\sum_{i\in A_{s^*}}\left\{\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right)+\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right)\right\}+\sum_{i\in A}\left\{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{\psi_m^{\rho_s}\Delta_i^{4\rho_s-1}}\right)+\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{\psi_m^{\rho_s}b^{4\rho_s-1}}\right)\right\}}_{\textit{case b2}} +$$

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

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

Remark 3 *If $\rho_s >> \rho_a$ we can 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 $\rho_a << \rho_s$ 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.*

D Proof of Corollary 1

Proof 5 *Here we take $b \approx \sqrt{\frac{K\log K}{T}}$, $\psi_m = K^2T$, $\rho_a = max\left\{\frac{1}{2^m}, \frac{1}{4}\right\}$ and $\rho_s = max\left\{\frac{1}{2^m}, \frac{1}{4}\right\}$*

Taking into account Theorem 1, putting the above values in $\sum_{i\in A:\Delta_i\geq b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{1+4\rho_a}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right) = \left(K\frac{T^{1-\frac{1}{4}}\frac{1}{4}^{2\frac{1}{4}}2^{1+4\frac{1}{4}}}{(T)^{\frac{1}{4}}\Delta_i^{4\frac{1}{4}-1}}\right) = 2\sqrt{KT}$. Similarly, for $\sum_{i\in A:\Delta_i\geq b}\left(\frac{T^{1-\rho_s}\rho_a^{2\rho_s}2^{2+4\rho_s}}{\psi_m^{\rho_s}\Delta_i^{4\rho_s-1}}\right) = 2\sqrt{KT}$.

For the term $\sum_{i\in A:\Delta_i\geq b}\frac{32\rho_s\log{(\psi_m T\frac{\Delta_i^4}{16\rho_s^2})}}{\Delta_i}$ we get, $\frac{32\sqrt{T}\frac{1}{4}\log{(T^2\frac{K^4(\log K)^2}{T^2})}}{\sqrt{K\log K}} \leq \frac{16\sqrt{KT}\log{(K^2(\log K))}}{\sqrt{\log K}} = 32\sqrt{KT\log K}+\frac{16\log{(\log K)}}{\sqrt{\log K}}$. For the term $\sum_{i\in A:\Delta_i\geq b}\frac{32\rho_a\log{(\psi_m T\frac{\Delta_i^4}{16\rho_a^2})}}{\Delta_i}$ we get $32\sqrt{KT\log K}+\frac{16\log{(\log K)}}{\sqrt{\log K}}$.

Similarly, $\sum_{i\in A_{s^*}:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right) = \frac{K}{p}\left(\frac{T^{1-\frac{1}{4}}\frac{1}{4}^{2\frac{1}{4}}2^{2\frac{1}{4}+\frac{3}{2}}}{(T)^{\frac{1}{4}}(\Delta_i)^{4*\frac{1}{4}-1}}\right) = \frac{2\sqrt{KT}}{p}$. For the term $\sum_{i\in A:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{(\psi_m^{\rho_s})\Delta_i^{4\rho_s-1}}\right) = 2\sqrt{KT}$.

So, the total bound for using both arm and cluster elimination comes of as $\left\{6\sqrt{KT} + 2\frac{\sqrt{KT}}{p} + 64\sqrt{KT\log K} + \frac{32\log(\log K)}{\sqrt{\log K}}\right\}$.

E Proof of Corollary 2

Proof 6 *Here we take* $b \approx \sqrt{\frac{K\log K}{T}}$, $\psi_m = \frac{T}{\log(KT)}$, $\rho_a = max\left\{\frac{1}{2^m}, \frac{1}{2}\right\}$ *and* $\rho_s = max\left\{\frac{1}{2^m}, \frac{1}{2}\right\}$

Case a: *Taking into account Theorem 1, putting the above values in* $\sum_{i\in A:\Delta_i\geq b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{1+4\rho_a}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right) = \sum_{i\in A:\Delta_i\geq b}\left(\frac{T^{1-\frac{1}{2}}\frac{1}{2}^{2*\frac{1}{2}}2^{1+4*\frac{1}{2}}}{(\frac{T}{\log(KT)})^{\frac{1}{2}}\Delta_i^{4*\frac{1}{2}-1}}\right) = \sum_{i\in A:\Delta_i\geq b}\frac{4\sqrt{\log(KT)}}{\Delta_i}$. *Similarly, for* $\sum_{i\in A:\Delta_i\geq b}\left(\frac{T^{1-\rho_S}\rho_a^{2\rho_s}2^{2+4\rho_s}}{\psi_m^{\rho_s}\Delta_i^{4\rho_s-1}}\right) = \sum_{i\in A:\Delta_i\geq b}\frac{8\sqrt{\log(KT)}}{\Delta_i}$.

For the term $\sum_{i\in A:\Delta_i\geq b}\frac{32\rho_s\log(\psi_m T\frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}$ *we get,* $\sum_{i\in A:\Delta_i\geq b}\frac{16\log(T^2\frac{\Delta_i^4}{4\log(KT)})}{\Delta_i} \approx \sum_{i\in A:\Delta_i\geq b}\frac{32\log(T\frac{\Delta_i^2}{\sqrt{\log(KT)}})}{\Delta_i}$. *Similarly the term* $\sum_{i\in A:\Delta_i\geq b}\frac{32\rho_a\log(\psi_m T\frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} \approx \sum_{i\in A:\Delta_i\geq b}\frac{32\log(T\frac{\Delta_i^2}{\sqrt{\log(KT)}})}{\Delta_i}$.

Lastly for the term $\sum_{i\in A_{s*}:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi_m^{\rho_a}\Delta_i^{4\rho_a-1}}\right) = \sum_{i\in A_{s*}:0\leq\Delta_i\leq b}\frac{2.8\sqrt{\log(KT)}}{\Delta_i}$. *For the term* $\sum_{i\in A:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{(\psi_m^{\rho_s})\Delta_i^{4\rho_s-1}}\right) = \sum_{i\in A_{s*}:0\leq\Delta_i\leq b}\frac{8\sqrt{\log(KT)}}{\Delta_i}$.

So, the total gap dependent regret bound for using both arm and cluster elimination comes of as $\sum_{i\in A:\Delta_i\geq b}\left\{\frac{12\sqrt{\log(KT)}}{\Delta_i} + \frac{64\log(T\frac{\Delta_i^2}{\sqrt{\log(KT)}})}{\Delta_i}\right\} + \sum_{i\in A_{s*}:0\leq\Delta_i\leq b}\frac{2.8\sqrt{\log(KT)}}{\Delta_i} + \sum_{i\in A:0\leq\Delta_i\leq b}\frac{8\sqrt{\log(KT)}}{\Delta_i}$.

F Exploration Regulatory Factor

In this section in Table 2 we discuss about the various exploration regulatory function that we can use to control exploration. A similar topic has already been handled in [11] 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

Table 2: Definition of ψ_m

No.	Definition	UB	LB	Remarks
1	$\psi_m = \log T$	$\log T$	$\log T$	Constant factor. Used in experiment.
2	$\psi_m = \frac{\log T}{2^m}$	$\log T$	$\frac{\sqrt{e}\log T}{\sqrt{T}}$	Decreasing factor.
3	$\psi_m = T$	T	T	Constant factor. Used in Proof of Corollary 1.
4	$\psi_m = \frac{T}{\log(KT)}$	$\frac{T}{\log(KT)}$	$\frac{T}{\log(KT)}$	Constant factor. Used in Proof of Corollary 2.

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 $m = 0, 1, ..., \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$.

Depending on how fast the exploration must be done 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 [11]. 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.

But this $\psi_m = T$ significantly affects two factors, the maximum regret suffered for not eliminating arms and clusters and the error bound. Firstly(from Theorem 1, case b1), it becomes $\sum_{i\in A:\Delta_i\geq b} \left\{ \left(\frac{2^{2+4\rho_s}\rho_s^{2\rho_s} T^{1-\rho_s}}{(T)^{\rho_s}(\Delta_i)^{4\rho_s-1}} \right) \right\}$ from case a) in Theorem 1 and $\sum_{i\in A:0\leq\Delta_i\leq b} \left(\frac{T^{1-\rho_s} 2^{2\rho_s+3}}{(T)^{\rho_s}(b)^{4\rho_s-1}} \right)$ from Theorem 1, case b2. 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 $\left(\frac{32K\rho_s \log(\psi_m T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right) = \left(\frac{32K\rho_s \log(T^2 \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right)$, the term $\rho_s \log(T^2 \frac{\Delta_i^4}{16\rho_s^2})$ can be sufficiently balanced by choosing appropriate ρ_s as shown in analysis of Corollary 1, 2.

G Arm Elimination Parameters and Why Clustering?

In this section in Table 3 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 $m = 0, 1, ..., \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$.

Remark 4 *Another important distinction we want to make is the apparent use of clustering. The regret bound for arm elimination and pulls and error bound when a sub-optimal arm or a sub-optimal cluster eliminates a^* or s^* is given in Table 4, 5respectively.*

From Table 4 we can see that from the definition of ρ_s, ρ_a and ψ_m we can have a regret bound for jointly doing the arm and cluster elimination of the same order as doing arm elimination or cluster elimination alone. While looking at the

Table 3: Definitions of ρ_a, ρ_s

No.	Definition(ρ_s, ρ_a)	UB(ρ_s, ρ_a)	LB(ρ_s, ρ_a)	Remarks
1	$\rho_s = \dfrac{1}{2^m}, \rho_a = \dfrac{1}{2^m}$	$1, 1$	$\sqrt{\frac{e}{T}}, \sqrt{\frac{e}{T}}$	Used in algorithm.
2	$\rho_s, \rho_a = max\left\{\dfrac{1}{2^m}, \dfrac{1}{2}\right\}$	$1, 1$	$\dfrac{1}{2}, \dfrac{1}{2}$	A less risky definition having a higher lower bound than Definition 1. Used in Corollary 2.
3	$\rho_s, \rho_a = max\left\{\dfrac{1}{2^m}, \dfrac{1}{4}\right\}$	$1, 1$	$\dfrac{1}{4}, \dfrac{1}{4}$	A more risky definition having a lower bound lesser than Definition 2. Used in Corollary 1.
4	$\rho_s = \dfrac{1}{2^{2m+1}}$, $\rho_a = \dfrac{1}{2^{4m+1}}$	$\dfrac{1}{2}, \dfrac{1}{2}$	$\dfrac{e}{2T}, \dfrac{e^2}{2T^2}$	A more aggressive arm elimination and conservative cluster elimination definition than Definition 3. Used in Experiment.

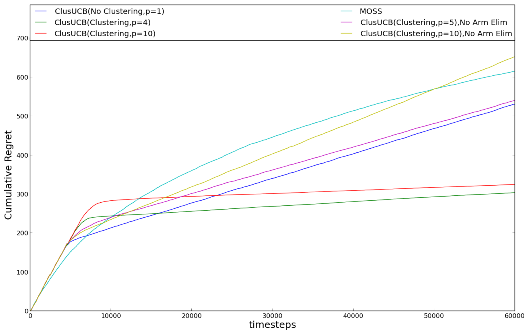


Figure 5: ClusUCB for various p

error term for the 3 *cases we see that using just Cluster elimination the error term is more than using just arm elimination and the least error term comes from using both arm and cluster elimination simultaneously and* $\rho_a << \rho_s$*. From Table 5, we can see that the error term for using both arm and cluster elimination can become low depending on how we choose* p *since* $|A_{s^*}| \leq \lceil \frac{A}{p} \rceil$*. 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 5. 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* $\rho_s = \frac{1}{2^{m+1}}, \rho_a = \frac{1}{2^{2m+1}}$*. 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 2, 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(*$\sqrt{\log K} < p \leq \dfrac{K}{2}$*) we stop when we are left with one cluster and output the max payoff arm of that cluster.*

Table 4: Regret Bound on Arm Elimination and Pulls

Elim Type	**Regret Bound on Arm Elimination and Pulls**	**Remarks**
Only Arm Elimination	$\sum_{i\in A:\Delta_i\geq b}\left\{\left(\frac{2^{1+4\rho_a}\rho_a^{2\rho_a}T^{1-\rho_a}}{(\psi_m)^{\rho_a}\Delta_i^{4\rho_a-1}}\right)+\left(\Delta_i+\frac{32\rho_a\log{(\psi_m T\frac{\Delta_i^4}{16\rho_a^2})}}{\Delta_i}\right)\right\}$	For $\rho_a = \frac{1}{2^{2m}}$ and $\psi_m = T$ this gives $2\sqrt{KT\log K} + \frac{64\sqrt{K}\log{(TK\log K)}}{\sqrt{T\log K}}$.. Hence the order is given by $O(\sqrt{KT\log K})$.
Only Cluster Elimination	$\sum_{i\in A:\Delta_i\geq b}\left\{\underbrace{\left(\frac{2^{2+4\rho_s}\rho_s^{2\rho_s}T^{1-\rho_s}}{(\psi_m)^{\rho_s}\Delta_i^{4\rho_s-1}}\right)}_{\text{case a1+a2}}+\underbrace{\left(\Delta_i+\frac{32\rho_s\log{(\psi_m T\frac{\Delta_i^4}{16\rho_s^2})}}{\Delta_i}\right)}_{\text{case b1}}\right\}$	With $\rho_s = \frac{1}{2^{2m}}$ and $\psi_m = T$ this gives $4\sqrt{KT\log K} + \frac{64\sqrt{K}\log{(TK\log K)}}{\sqrt{T\log K}}$. Hence, this is larger bound than using only arm elimination though the order is same $O(\sqrt{KT\log K})$.
Arm & Cluster Elimination	$\sum_{i\in A:\Delta_i\geq b}\left\{\underbrace{\left(\frac{2^{1+4\rho_a}\rho_a^{2\rho_a}T^{1-\rho_a}}{(\psi_m)^{\rho_a}\Delta_i^{4\rho_a-1}}\right)}_{\text{case a, Arm Elim}}+\underbrace{\left(\frac{2^{2+4\rho_s}\rho_s^{2\rho_s}T^{1-\rho_s}}{(\psi_m)^{\rho_s}\Delta_i^{4\rho_s-1}}\right)}_{\text{case a, Clus Elim}}\right.$ $\left.+\underbrace{\left(\Delta_i+\frac{32\rho_a\log{(\psi_m T\frac{\Delta_i^4}{16\rho_a^2})}}{\Delta_i}\right)}_{\text{case b1, Arm Elim}}+\underbrace{\left(\Delta_i+\frac{32\rho_s\log{(\psi_m T\frac{\Delta_i^4}{16\rho_s^2})}}{\Delta_i}\right)}_{\text{case b1, Clus Elim}}\right\}$	With $\rho_a = \frac{1}{2^{2m}} << \rho_s = \frac{1}{2^m}$ and $\psi_m = T$ this gives $\left\{2\sqrt{KT\log K} + \frac{4\sqrt{T\log K}}{K^{\sqrt{\frac{T}{e}-\frac{1}{2}}}} + \frac{64\sqrt{K}\log{(\sqrt{T}K\log K)}}{\sqrt{\log K}} + \frac{64\sqrt{K}\log{(TK\log K)}}{\sqrt{T\log K}}\right\}$. This is larger than the previous 2 bounds though the order is same $O(\sqrt{KT\log K})$.

Table 5: Error Bound

Elim Type	Error Bound	Remarks
Only Arm Elimination	$\sum_{i\in A:0\leq\Delta_i\leq b}\left\{\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{(\psi_m)^{\rho_a}\Delta_i^{4\rho_a-1}}\right)+\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{(\psi_m)^{\rho_a}b^{4\rho_a-1}}\right)\right\}$	With $\rho_a = \frac{1}{2^{2m}}$ and $\psi_m = T$ this gives $5.6\sqrt{KT\log K}$. Hence, this has an order of $O(\sqrt{KT\log K})$.
Only Cluster Elimination	$\sum_{i\in A:0\leq\Delta_i\leq b}\left\{\underbrace{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{(\psi_m)^{\rho_s}\Delta_i^{4\rho_s-1}}\right)}_{\text{case b2+b3, Proposition 4}}+\underbrace{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{(\psi_m)^{\rho_s}b^{4\rho_s-1}}\right)}_{\text{case b2+b3, Proposition 4}}\right\}$	With $\rho_s = \frac{1}{2^{2m}}$ and $\psi_m = T$ this gives $16\sqrt{KT\log K}$. This is more than the bound using only arm elimination but has an order of $O(\sqrt{KT\log K})$.
Arm & Cluster Elimination	$\sum_{i\in A_{s*}:0\leq\Delta_i\leq b}\left\{\underbrace{\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{(\psi_m)^{\rho_a}\Delta_i^{4\rho_a-1}}\right)+\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{(\psi_m)^{\rho_a}b^{4\rho_a-1}}\right)}_{\text{case b2, Arm Elim, Theorem 2}}\right\}$ $+\sum_{i\in A:0\leq\Delta_i\leq b}\left\{\underbrace{\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{(\psi_m)^{\rho_s}\Delta_i^{4\rho_s-1}}\right)+\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+3}}{(\psi_m)^{\rho_s}b^{4\rho_s-1}}\right)}_{\text{case b2, Clus Elim, Theorem 2}}\right\}$	With $\rho_a = \frac{1}{2^{2m}} << \rho_s = \frac{1}{2^m}$ and $\psi_m = T$ this gives $2.8\frac{\sqrt{KT\log K}}{p} + \frac{8\sqrt{T\log K}}{K^{\sqrt{\frac{T}{e}-\frac{1}{2}}}}$. This is less than the previous 2 bounds and has an order of $O(\sqrt{KT})$ for $p = \sqrt{\log K}$. So, the range of p is given by $\sqrt{\log K} \leq p \leq \frac{K}{2}$

Table 6: Regret Upper Bound of Algorithms

Algorithm	**Regret Upper Bound**
UCB1	$\min\left\{O(\sqrt{KT\log T}), O\left(\frac{K\log T}{\Delta}\right)\right\}$
ϵ_n-greedy	$O\left(\frac{K\Delta\log T}{d^2}\right), 0 < d < \Delta$
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	$\min\left\{O\left(\sqrt{KT}\right), O\left(\frac{K\log(T\Delta^2/K)}{\Delta}\right)\right\}$
UCB-Clustered	$\min\left\{O\left(\sqrt{KT\log K}\right), O\left(\frac{K\log\big(\frac{T\Delta^2}{\sqrt{\log(KT)}}\big)}{\Delta}\right)\right\}$

H Regret Bound Table

The regret upper bound of various algorithms are given in Table 6.