UCB with clustering and improved exploration

Anonymous Authors[1]

Abstract

In this paper, we present three novel algorithms for the stochastic multi-armed bandit problem. Our proposed method, referred to as ClusUCB and its variants EClusUCB and AClusUCB, partitions the arms into clusters and then follows the UCB-Improved strategy with aggressive exploration factors to eliminate sub-optimal arms as well as clusters. Through a theoretical analysis, we establish that ClusUCB achieves a better gap-dependent regret upper bound than UCB-Improved (Auer & Ortner, 2010) and MOSS (Audibert & Bubeck, 2009) algorithms. Further, numerical experiments on test-cases with small gaps between optimal and sub-optimal mean rewards show that EClusUCB results in lower cumulative regret than several popular UCB variants as well as MOSS, OCUCB (Lattimore, 2015) and Thompson sampling.

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 of decisions (or 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*).

Let r_i, $i = 1, \ldots, K$ denote the mean reward of the ith arm out 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:

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

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

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

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

An early work involving a bandit setup is Thompson (1933), 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 (1952), 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 Lai & Robbins (1985). In particular, it was shown 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 algorithms with strong regret guarantees. For further reference we point the reader to Bubeck et al. (2012). The foremost among them is UCB1 (Auer et al., 2002a), which has a regret upper bound of $O\big(\frac{K \log T}{\Delta}\big)$, 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 gap independent regret bound of UCB1 can be as bad as $O\big(\sqrt{TK \log T}\big)$. In Audibert & Bubeck (2009), the authors propose the MOSS algorithm and establish that the worst case regret of MOSS is $O\big(\sqrt{TK}\big)$ which improves upon UCB1 by a factor of order $\sqrt{\log T}$. However, the gap-dependent regret of MOSS is $O\Big(\frac{K^2 \log\big(T\Delta^2/K\big)}{\Delta}\Big)$ and in certain regimes, this can be worse than even UCB1 (see (Audibert & Bubeck, 2009),(Lattimore, 2015)). The UCB-Improved algorithm, proposed in Auer & Ortner (2010), is a round-based algorithm[1] variant of UCB1 that has a gap-dependent regret bound of $O\big(\frac{K\log T\Delta^2}{\Delta}\big)$, which is better than that of UCB1. On the other hand, the worst case regret of UCB-Improved is $O\big(\sqrt{TK\log K}\big)$. Recently in Lattimore (2015), the algorithm OCUCB achieves order-optimal gap-dependent regret bound of $O\left(\sum_{i=2}^{K}\frac{\log(T/H_i)}{\Delta_i}\right)$ where $H_i = \sum_{j=1}^{K}\min\{\frac{1}{\Delta_i^2},\frac{1}{\Delta_j^2}\}$ and gap-independent regret bound of $O\big(\sqrt{KT}\big)$.

Our Work

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 arms into small clusters, each having same number of arms. The clustering is done at the start with a prespecified number of clusters. Each round of ClusUCB involves both (individual) arm elimination as well as cluster elimination.

The clustering of arms provides two benefits. First, it creates a context where UCB-Improved like algorithm can be run in parallel on smaller sets of arms with limited exploration, which could lead to fewer pulls of sub-optimal arms with the help of more aggressive elimination of sub optimal arms. Second, the cluster elimination leads to whole sets of sub-optimal arms being simultaneously eliminated when they are found to yield poor results. These two simultaneous criteria for arm elimination can be seen as borrowing the strengths of UCB-Improved as well as other popular round based approaches.

While ClusUCB does not achieve the gap-dependent regret bound of OCUCB, the theoretical analysis establishes that the gap-dependent regret of ClusUCB is always better than that of UCB-Improved and better than that of MOSS when $\sqrt{\frac{K}{14T}} \leq \Delta \leq 1$ (see Table 1). Moreover, the gap-independent bound of ClusUCB is of the same order as UCB-Improved, i.e., $O\left(\sqrt{KT\log K}\right)$. However, ClusUCB is not able to match the gap-independent bound of $O(\sqrt{KT})$ for MOSS and OCUCB. We also establish the exact values for the exploration parameters and the number of clusters required for optimal behavior in the corollaries.

While ClusUCB is a round-based algorithm, we also introduce Efficient ClusUCB or EClusUCB which has the same theoretical guarantees as ClusUCB but empirically behaves much better. On five synthetic setups with small gaps, we observe empirically that EClusUCB outperforms UCB-Improved(Auer & Ortner, 2010), MOSS(Audibert & Bubeck, 2009) and OCUCB(Lattimore, 2015) as well as other popular stochastic bandit algorithms such as DMED(Honda & Takemura, 2010), UCB-V(Audibert et al., 2009), Median Elimination(Even-Dar et al., 2006), Thompson Sampling(Agrawal & Goyal, 2011) and KL-UCB(Garivier & Cappé, 2011).

[1]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.

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	$O\left(\frac{K^2\log\left(T\Delta^2/K\right)}{\Delta}\right)$
ClusUCB/EClusUCB	$O\left(\frac{K\log\left(\frac{T\Delta^2}{\sqrt{\log(K)}}\right)}{\Delta}\right)$

The rest of the paper is organized as follows: In Section 2, we present the ClusUCB algorithm and in Section 3 we introduce EClusUCB. In Section 4, we present the associated regret bounds and prove the main theorem on the regret upper bound for ClusUCB in Section 5. In Section 6, we present the numerical experiments and provide concluding remarks in Section 7. Further proofs of corollaries, theorems and proposition presented in Section 5 are provided in the appendices. The algorithm Adaptive ClusUCB is presented in Appendix G and more experiments are presented in Appendix H.

2. Clustered UCB

Notation. We denote the set of arms by A, with the individual arms labeled $i, i = 1,\ldots,K$. 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 prespecified limit for the number of clusters. For simplicity, we assume that the optimal arm is unique and denote it by $*$, with s^* denoting the corresponding cluster. The best arm in a cluster s_k is denoted by $a_{max_{s_k}}$. We denote the sample mean of the rewards seen so far for arm i by $\hat{r}_i$ and for the true best arm within a cluster s_k by $\hat{r}_{a_{\max_{s_k}}}$. z_i is the number of times an arm i has been pulled. We assume that the rewards of all arms are bounded in $[0,1]$.

The algorithm. As mentioned in a recent work (Liu & Tsuruoka, 2016), UCB-Improved has two shortcomings:
(i) A significant number of pulls are spent in early exploration, since each round m of UCB-Improved involves

Algorithm 1 ClusUCB

Input: Number of clusters p, time horizon T, exploration parameters ρ_a, ρ_s and ψ.
Initialization: Set $B_0 := A$, $S_0 = S$ and $\epsilon_0 := 1$.
Create a partition S_0 of the arms at random into p clusters of size up to $\ell = \left\lceil \frac{K}{p} \right\rceil$ each.
for $m = 0, 1, ..\lfloor \frac{1}{2}\log_2 \frac{14T}{K} \rfloor$ **do**
Pull each arm in B_m so that the total number of times it has been pulled is $n_m = \left\lceil \frac{2\log(\psi T\epsilon_m^2)}{\epsilon_m} \right\rceil$.
Arm Elimination
For each cluster $s_k \in S_m$, delete arm $i \in s_k$ from B_m if

$$\hat{r}_i + \sqrt{\frac{\rho_a \log(\psi T\epsilon_m^2)}{2n_m}} < \max_{j\in s_k}\left\{\hat{r}_j - \sqrt{\frac{\rho_a \log(\psi T\epsilon_m^2)}{2n_m}}\right\}$$

Cluster Elimination
Delete cluster $s_k \in S_m$ and remove all arms $i \in s_k$ from B_m if

$$\max_{i\in s_k}\left\{\hat{r}_i + \sqrt{\frac{\rho_s \log(\psi T\epsilon_m^2)}{2n_m}}\right\} < \max_{j\in B_m}\left\{\hat{r}_j - \sqrt{\frac{\rho_s \log(\psi T\epsilon_m^2)}{2n_m}}\right\}.$$

Set $\epsilon_{m+1} := \frac{\epsilon_m}{2}$
Set $B_{m+1} := B_m$
Stop if $|B_m| = 1$ and pull $i \in B_m$ till T is reached.
end for

pulling every arm an identical $n_m = \left\lceil \frac{2\log(T\epsilon_m^2)}{\epsilon_m^2} \right\rceil$ number of times. The quantity ϵ_m is initialized to 1 and halved after every round.
(ii) In UCB-Improved, arms are eliminated conservatively, i.e, only after $\epsilon_m < \frac{\Delta_i}{2}$, the sub-optimal arm i is discarded with high probability. This is disadvantageous when K is large and the gaps are identical ($r_1 = r_2 = \cdots = r_{K-1} < r^*$) and small.
To reduce early exploration, the number of times n_m each arm is pulled per round in ClusUCB is lower than that of UCB-Improved and also that of Median-Elimination, which used $n_m = \frac{4}{\epsilon^2}\log\left(\frac{3}{\delta}\right)$, where ϵ, δ are confidence parameters. To handle the second problem mentioned above, ClusUCB partitions the larger problem into several small sub-problems using clustering and then performs local exploration aggressively to eliminate sub-optimal arms within each clusters with high probability.

As described in the pseudocode in Algorithm 1, ClusUCB begins with a initial clustering of arms that is performed by random uniform allocation. 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 well as cluster elimination conditions. These elimination conditions 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 we now have as many reference points to eliminate arms as number of clusters formed.

The exploration regulatory factor ψ governing the arm and cluster elimination conditions in ClusUCB is more aggressive than that in UCB-Improved. With appropriate choices of ψ, ρ_a and ρ_s, we can achieve aggressive elimination even when the gaps Δ_i are small and K is large.

In Liu & Tsuruoka (2016), the authors recommend incorporating a factor of d_i inside the log-term of the UCB values, i.e., $\max\{\hat{r}_i + \sqrt{\frac{d_i \log T\epsilon_m^2}{2n_m}}\}$. The authors there examine the following choices for d_i: $\frac{T}{z_i}$, $\frac{\sqrt{T}}{z_i}$ and $\frac{\log T}{z_i}$, where z_i is the number of times an arm i has been sampled. Unlike Liu & Tsuruoka (2016), we employ cluster as well as arm elimination and establish from a theoretical analysis that the choice $\psi = \frac{T}{196\log(K)}$ helps in achieving a better gap-dependent regret upper bound for ClusUCB as compared to UCB-Improved and MOSS (see Corollary 1 in section 4).

3. Efficient Clustered UCB

One principal disadvantage with ClusUCB is that it is a round-based algorithm. In this section, we introduce a further modification, as shown in Algorithm 2 called Efficient Clustered UCB or EClusUCB where we introduce the idea of optimistic greedy sampling similar to Liu & Tsuruoka (2016) which they used to modify the UCB-Improved algorithm. We further modify the idea by introducing clustering and arm elimination parameters. Also we use different exploration regulatory factor and we come up with a cumulative regret bound whereas Liu & Tsuruoka (2016) only gives simple regret bound. In optimistic greedy sampling, we only sample the arm with the highest upper confidence bound in each timestep. EClusUCB checks arm and cluster elimination conditions in every timestep and update parameters when a round is complete. It divides each round into $|B_m|n_m$ timesteps so that each surviving arms can be allocated atmost n_m pulls.

Adaptive Clustered UCB: One of the disadvantages of EClusUCB is that it requires the knowledge of the number of clusters p. To counter this we introduce Adaptive Clustered UCB or AClusUCB which employs hierarchical clustering and is shown in Appendix G.

Algorithm 2 EClusUCB

Input: Number of clusters p, time horizon T, exploration parameters ρ_a, ρ_s and ψ.

Initialization: Set $m := 0$, $B_0 := A$, $S_0 = S$, $\epsilon_0 := 1$, $M = \left\lfloor \frac{1}{2}\log_2 \frac{14T}{K} \right\rfloor$, $n_0 = \left\lceil \frac{2\log(\psi T \epsilon_0^2)}{\epsilon_0} \right\rceil$ and $N_0 = K n_0$.

Create a partition S_0 of the arms at random into p clusters of size up to $\ell = \left\lceil \frac{K}{p} \right\rceil$ each.

Pull each arm once

for $t = K+1, .., T$ **do**

Pull arm $i \in B_m$ such that $\mathrm{argmax}_{j \in B_m} \left\{ \hat{r}_j + \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2 z_j}} \right\}$, where z_j is the number of times arm j has been pulled

$t := t + 1$

Arm Elimination

For each cluster $s_k \in S_m$, delete arm $i \in s_k$ from B_m if

$$\hat{r}_i + \sqrt{\frac{\rho_a \log(\psi T \epsilon_m^2)}{2 z_i}} < \max_{j \in s_k} \left\{ \hat{r}_j - \sqrt{\frac{\rho_a \log(\psi T \epsilon_m^2)}{2 z_j}} \right\}$$

Cluster Elimination

Delete cluster $s_k \in S_m$ and remove all arms $i \in s_k$ from B_m if

$$\max_{i \in s_k} \left\{ \hat{r}_i + \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2 z_i}} \right\} < \max_{j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2 z_j}} \right\}.$$

if $t \geq N_m$ and $m \leq M$ **then**

$\epsilon_{m+1} := \frac{\epsilon_m}{2}$

$B_{m+1} := B_m$

$n_{m+1} := \left\lceil \frac{2\log(\psi T \epsilon_{m+1}^2)}{\epsilon_{m+1}} \right\rceil$

$N_{m+1} := t + |B_{m+1}| n_{m+1}$

$m := m + 1$

Stop if $|B_m| = 1$ and pull $i \in B_m$ till T is reached.

end if

end for

4. Main results

We now state the main result that upper bounds the expected regret of ClusUCB.

Theorem 1 (***Regret bound***)**.** *The regret R_T of ClusUCB satisfies*

$$\begin{aligned}
\mathbb{E}[R_T] \leq & \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} \left\{ \frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \Delta_i \right. \\
& \left. + \frac{32 \rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} \right\} + \sum_{\substack{i \in A, \\ \Delta_i > b}} \left\{ 2\Delta_i + \frac{C_1(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} \right. \\
& \left. + \frac{32 \rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} + \frac{32 \rho_s \log(\psi T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right\} \\
& + \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} \frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \sum_{\substack{i \in A_{s^*}, \\ 0 < \Delta_i \leq b}} \frac{C_2(\rho_a) T^{1-\rho_a}}{b^{4\rho_a - 1}} \\
& + \sum_{\substack{i \in A \setminus A_{s^*}: \\ \Delta_i > b}} \frac{2 C_2(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} + \sum_{\substack{i \in A \setminus A_{s^*}: \\ 0 < \Delta_i \leq b}} \frac{2 C_2(\rho_s) T^{1-\rho_s}}{b^{4\rho_s - 1}} \\
& + \max_{i: \Delta_i \leq b} \Delta_i T,
\end{aligned}$$

where $b \geq \sqrt{\frac{K}{14T}}$, $C_1(x) = \frac{2^{1+4x} x^{2x}}{\psi^x}$, $C_2(x) = \frac{2^{2x + \frac{3}{2}} x^{2x}}{\psi^x}$, and A_{s^*} is the subset of arms in cluster s^* containing optimal arm a^*.

Proof. See Section 5. □

We now specialize the result in the theorem above by substituting specific values for the exploration constants ρ_s, ρ_a and ψ.

Corollary 1 (***Gap-dependent bound***)**.** *With $\psi = \frac{T}{196\log(K)}$, $\rho_a = \frac{1}{2}$, and $\rho_s = \frac{1}{2}$, we have the following gap-dependent bound for the regret of ClusUCB:*

$$\begin{aligned}
\mathbb{E}[R_T] \leq & \sum_{\substack{i \in A_{s^*}: \\ \Delta_i > b}} \left\{ \frac{96\sqrt{\log(K)}}{\Delta_i} + \Delta_i + \right. \\
& \left. \frac{32 \log(T \frac{\Delta_i^2}{\sqrt{\log(K)}})}{\Delta_i} \right\} + \sum_{i \in A: \Delta_i > b} \left\{ \frac{56\sqrt{\log(K)}}{\Delta_i} + 2\Delta_i \right. \\
& \left. + \frac{64 \log(T \frac{\Delta_i^2}{\sqrt{\log(K)}})}{\Delta_i} \right\} + \sum_{\substack{i \in A_{s^*}: \\ 0 < \Delta_i \leq b}} \frac{40\sqrt{\log(K)}}{\Delta_i} \\
& + \sum_{\substack{i \in A \setminus A_{s^*}: \\ \Delta_i > b}} \frac{80\sqrt{\log(K)}}{\Delta_i} + \sum_{\substack{i \in A \setminus A \cup A_{s^*}: \\ 0 < \Delta_i \leq b}} \frac{80\sqrt{\log(K)}}{\Delta_i} \\
& + \max_{i \in A: \Delta_i \leq b} \Delta_i T, \quad \textit{for all } b \geq \sqrt{\frac{K}{14T}}.
\end{aligned}$$

Proof. See Appendix B. □

The most significant term in the bound above is $\sum_{i\in A:\Delta_i\geq b}\frac{64\log\left(T\frac{\Delta_i^2}{\sqrt{\log(K)}}\right)}{\Delta_i}$ and hence, the regret upper bound for ClusUCB is of the order $O\left(\frac{K\log\left(\frac{T\Delta^2}{\sqrt{\log(K)}}\right)}{\Delta}\right)$. Since Corollary 1 holds for all $\Delta\geq\sqrt{\frac{K}{14T}}$, it can be clearly seen that for all $\sqrt{\frac{K}{14T}}\leq\Delta\leq 1$ and $K\geq 2$, the gap-dependent bound is better than that of UCB1, UCB-Improved and MOSS (see Table 1).

Corollary 2 (***Gap-independent bound***)**.** *Considering the same gap of $\Delta_i=\Delta=\sqrt{\frac{K\log K}{T}}$ for all $i:i\neq *$ and with $\psi=\frac{T}{196\log K}$, $p=\left\lceil\frac{K}{\log K}\right\rceil$, $\rho_a=\frac{1}{2}$ and $\rho_s=\frac{1}{2}$, we have the following gap-independent bound for the regret of ClusUCB:*

$$
\begin{aligned}
\mathbb{E}[R_T]\leq &270\frac{\sqrt{T}\log K}{\sqrt{K}}+\frac{32\sqrt{T\log K}\log(\log K)}{\sqrt{K}}\\
&+56\sqrt{KT}+128\sqrt{KT\log K}\\
&+\frac{64\sqrt{KT}\log(\log K)}{\sqrt{\log K}}+150\sqrt{\frac{T\log K}{e}}\\
&+300\sqrt{\frac{T}{e}}(\log K)^{\frac{3}{2}}+300\frac{K}{K+\log K}\sqrt{KT}
\end{aligned}
$$

Proof. See Appendix C. □

From the above result, we observe that the order of the regret upper bound of ClusUCB is $O(\sqrt{KT\log K})$, and this matches the order of UCB-Improved. However, this is not as low as the order $O(\sqrt{KT})$ of MOSS or OCUCB. Also, the gap-independent bound of UCB-Improved holds for $\sqrt{\frac{e}{T}}\leq\Delta\leq 1$ while in our case the gap independent bound holds for $\sqrt{\frac{K}{14T}}\leq\Delta\leq 1$.

Analysis of elimination error

Let $\widetilde{R}_T$ denote the contribution to the expected regret in the case when the optimal arm $*$ gets eliminated during one of the rounds of ClusUCB. This can happen if a sub-optimal arm eliminates $*$ or if a sub-optimal cluster eliminates the cluster s^* that contains $*$ – these correspond to cases b2 and b3 in the proof of Theorem 1 (see Section 5). We shall denote variant of ClusUCB that includes arm elimination condition only as ClusUCB-AE while ClusUCB corresponds to Algorithm 1, which uses both arm and cluster elimination conditions.

For ClusUCB-AE, the quantity $\widetilde{R}_T$ can be extracted from the proofs (in particular, case b2 in Appendix A) and simplified using the values $\rho_a=\frac{1}{2}$ and $\psi=\frac{T}{196\log K}$, to obtain $\widetilde{R}_T=150\sqrt{KT\log K}+150\sqrt{KT}$. Finally, for ClusUCB, the relevant terms from Theorem 1 that corresponds to $\widetilde{R}_T$ can be simplified with $\rho_a=\frac{1}{2}$, $\rho_s=\frac{1}{2}$, $p=\left\lceil\frac{K}{\log K}\right\rceil$ and $\psi=\frac{T}{\log K}$ (as in Corollary 2 to obtain $\widetilde{R}_T=\frac{150\sqrt{T}\log K^{\frac{3}{2}}}{\sqrt{K}}+\frac{150\sqrt{T}\log K}{\sqrt{K}}+300\frac{K}{K+\log K}\sqrt{KT\log K}+300\frac{K}{K+\log K}\sqrt{KT}$. Hence, in comparison to ClusUCB-AE which has an elimination regret bound of $O(\sqrt{KT\log K})$, the elimination error contribution to regret is lower in ClusUCB which has a bound of $O(\frac{K}{K+\log K}\sqrt{KT\log K})$. Thus, we observe that clustering in conjunction with improved exploration via ρ_a,ρ_s,p and ψ helps in reducing the factor associated with $\sqrt{KT\log K}$ for the gap-independent error regret bound for ClusUCB. A table containing the regret error bound is shown in Appendix D and further experiments showing that the performance of ClusUCB-AE is indeed sub-optimal and choice of p is indeed close to optimal is shown in Appendix H.

5. Proof of Theorem 1

Proof. Let $A^{'}=\{i\in A,\Delta_i>b\}$, $A^{''}=\{i\in A,\Delta_i>0\}$, $A^{'}_{s_k}=\{i\in A_{s_k},\Delta_i>b\}$ and $A^{''}_{s_k}=\{i\in A_{s_k},\Delta_i>0\}$. C_g is the cluster set containing max payoff arm from each cluster in g-th round. The arm having the highest payoff in a cluster s_k is denote by $a_{\max_{s_k}}$. Let for each sub-optimal arm $i\in A$, $m_i=\min\{m|\sqrt{\rho_a\epsilon_m}<\frac{\Delta_i}{2}\}$ and let for each cluster $s_k\in S$, $g_{s_k}=\min\{g|\sqrt{\rho_s\epsilon_g}<\frac{\Delta_{a_{\max_{s_k}}}}{2}\}$. Let $\check{A}=\{i\in A^{'}|i\in s_k,\forall s_k\in S\}$.

The analysis proceeds by considering the contribution to the regret in each of the following cases:

Case a: *Some sub-optimal arm i is not eliminated in round* $\max(m_i,g_{s_k})$ *or before, with the optimal arm $*\in$* $C_{\max(m_i,g_{s_k})}$.

We consider an arbitrary sub-optimal arm i and analyze the contribution to the regret when i is not eliminated in the following exhaustive sub-cases:
Case a1: *In round* $\max(m_i,g_{s_k})$, $i\in s^*$.

Similar to case (a) of Auer & Ortner (2010), observe that when the following two conditions hold, arm i gets eliminated:

$$\hat{r}_i\leq r_i+c_{m_i}\text{ and }\hat{r}^*\geq r^*-c_{m_i},\tag{1}$$

where $c_{m_i}=\sqrt{\frac{\rho_a\log(\psi T\epsilon_{m_i}^2)}{2n_{m_i}}}$. The arm i gets eliminated because

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

In the above, we have used the fact that

$c_{m_i} = \sqrt{\rho_a \epsilon_{m_i+1}} < \frac{\Delta_i}{4}$, since $n_{m_i} = \frac{2\log(\psi T \epsilon^2_{m_i})}{\epsilon_{m_i}}$ and $\rho_a \in (0, 1]$.

From the foregoing, we have to bound the events complementary to that in (1) for an arm i to not get eliminated. Considering Chernoff-Hoeffding bound this is done as follows:

$$
\begin{aligned}
&\mathbb{P}\left(\hat{r}^* \le r^* - c_{m_i}\right) \le \exp(-2c^2_{m_i} n_{m_i}) \\
&\le \exp(-2 * \frac{\rho_a \log(\psi T \epsilon^2_{m_i})}{2n_{m_i}} * n_{m_i}) \\
&\le \frac{1}{(\psi T \epsilon^2_{m_i})^{\rho_a}}
\end{aligned}
$$

Along similar lines, we have $\mathbb{P}\left(\hat{r}_i \ge r_i + c_{m_i}\right) \le \frac{1}{(\psi T \epsilon^2_{m_i})^{\rho_a}}$. Thus, the probability that a sub-optimal arm i is not eliminated in any round on or before m_i is bounded above by $\left(\frac{2}{(\psi T \epsilon^2_{m_i})^{\rho_a}}\right)$. Summing up over all arms in A'_{s^*} in conjunction with a simple bound of $T\Delta_i$ for each arm, we obtain

$$
\begin{aligned}
&\sum_{i \in A'_{s^*}} \left(\frac{2T\Delta_i}{(\psi T \epsilon^2_{m_i})^{\rho_a}}\right) \le \sum_{i \in A'_{s^*}} \left(\frac{2T\Delta_i}{(\psi T \dfrac{\Delta_i^4}{16\rho_a^2})^{\rho_a}}\right) \\
&= \sum_{i \in A'_{s^*}} \left(\frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}}\right), \text{where } C_1(x) = \frac{2^{1+4x} x^{2x}}{\psi^x}
\end{aligned}
$$

Case a2: *In round* $\max(m_i, g_{s_k})$, $i \in s_k$ *for some* $s_k \neq s^*$.

Following a parallel argument like in Case $a1$, we have to bound the following two events of arm $a_{\max_{s_k}}$ not getting eliminated on or before g_{s_k}-th round,

$$
\hat{r}_{a_{\max_{s_k}}} \ge r_{a_{\max_{s_k}}} + c_{g_{s_k}} \text{ and } \hat{r}^* \le r^* - c_{g_{s_k}}
$$

We can prove using Chernoff-Hoeffding bounds and considering independence of events mentioned above, that for $c_{g_{s_k}} = \sqrt{\frac{\rho_s \log(\psi T \epsilon^2_{g_{s_k}})}{2n_{g_{s_k}}}}$ and $n_{g_{s_k}} = \frac{2\log(\psi T \epsilon^2_{g_{s_k}})}{\epsilon_{g_{s_k}}}$ the probability of the above two events is bounded by $\left(\dfrac{2}{(\psi T \epsilon^2_{g_{s_k}})^{\rho_s}}\right)$. Now, for any round g_{s_k}, all the elements of $C_{\max(m_i, g_{s_k})}$ are the respective maximum payoff arms of their cluster $s_k, \forall s_k \in S$, and since all the surviving arms are pulled equally in each round and since clusters are fixed so we can bound the maximum probability that a sub-optimal arm $i \in A'$ and $i \in s_k$ such that $a_{\max_{s_k}} \in C_{g_{s_k}}$ is not eliminated on or before the g_{s_k}-th round by the same probability as above.

Summing up over all p clusters and bounding the regret for each arm $i \in A'_{s_k}$ trivially by $T\Delta_i$,

$$
\begin{aligned}
&\sum_{k=1}^{p} \sum_{i \in A'_{s_k}} \left(\frac{2T\Delta_i}{(\psi T \frac{\Delta_i^4}{16\rho_s^2})^{\rho_s}}\right) = \sum_{i \in A'} \left(\frac{2T\Delta_i}{(\psi T \frac{\Delta_i^4}{16\rho_s^2})^{\rho_s}}\right) \\
&\le \sum_{i \in A'} \left(\frac{2^{1+4\rho_s} \rho_s^{2\rho_s} T^{1-\rho_s}}{\psi^{\rho_s} \Delta_i^{4\rho_s - 1}}\right) = \sum_{i \in A'} \frac{C_1(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}}
\end{aligned}
$$

Summing the bounds in Cases $a1 - a2$ and observing that the bounds in the aforementioned cases hold for any round $C_{\max\{m_i, g_{s_k}\}}$, we obtain the following contribution to the expected regret from case a:

$$
\sum_{i \in A_{s^*}} \frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \sum_{i \in A'} \left(\frac{C_1(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}}\right)
$$

Case b: *For each arm* i*, either* i *is eliminated in round* $\max(m_i, g_{s_k})$ *or before or there is no optimal arm* $*$ *in* $C_{\max(m_i, g_{s_k})}$.

Case b1: $* \in C_{\max(m_i, g_{s_k})}$ *for each arm* $i \in A'$ *and cluster* $s_k \in \check{A}$.

The condition in the case description above implies the following:
(i) each sub-optimal arm $i \in A'$ is eliminated on or before $\max(m_i, g_{s_k})$ and hence pulled not more than pulled n_{m_i} number of times.
(ii) each sub-optimal cluster $s_k \in \check{A}$ is eliminated on or before $\max(m_i, g_{s_k})$ and hence pulled not more than pulled $n_{g_{s_k}}$ number of times.

Hence, the maximum regret suffered due to pulling of a sub-optimal arm or a sub-optimal cluster is no more than the following:

$$
\begin{aligned}
&\sum_{i \in A'} \Delta_i \left\lceil \frac{2\log(\psi T \epsilon^2_{m_i})}{\epsilon_{m_i}} \right\rceil + \sum_{k=1}^{p} \sum_{i \in A'_{s_k}} \Delta_i \left\lceil \frac{2\log(\psi T \epsilon^2_{g_{s_k}})}{\epsilon_{g_{s_k}}} \right\rceil \\
&\le \sum_{i \in A'} \Delta_i \left(1 + \frac{32\rho_a \log\left(\psi T \left(\frac{\Delta_i}{2\sqrt{\rho_a}}\right)^4\right)}{\Delta_i^2}\right) \\
&\quad + \sum_{i \in A'} \Delta_i \left(1 + \frac{32\rho_s \log\left(\psi T \left(\frac{\Delta_i}{2\sqrt{\rho_s}}\right)^4\right)}{\Delta_i^2}\right) \\
&\le \sum_{i \in A'} \left[2\Delta_i + \frac{32(\rho_a \log(\psi T \dfrac{\Delta_i^4}{16\rho_a^2}) + \rho_s \log(\psi T \dfrac{\Delta_i^4}{16\rho_s^2}))}{\Delta_i}\right]
\end{aligned}
$$

In the above, the first inequality follows since $\sqrt{\rho_a \epsilon_{m_i}} < \frac{\Delta_i}{2}$ and $\sqrt{\rho_s \epsilon_{n_{g_{s_k}}}} < \frac{\Delta_{a_{\max_{s_k}}}}{2}$.

Case b2: *$*$ is eliminated by some sub-optimal arm in s^**
Optimal arm a^* can get eliminated by some sub-optimal arm i only if arm elimination condition holds, i.e.,

$$\hat{r}_i - c_{m_i} > \hat{r}^* + c_{m_i},$$

where, as mentioned before, $c_{m_i} = \sqrt{\frac{\rho_a \log(\psi T \epsilon_{m_i}^2)}{2n_{m_i}}}$. From analysis in Case $a1$, notice that, if (1) holds in conjunction with the above, arm i gets eliminated. Also, recall from Case $a1$ that the events complementary to (1) have low-probability and can be upper bounded by $\frac{2}{(\psi T \epsilon_{m_*}^2)^{\rho_a}}$. Moreover, a sub-optimal arm that eliminates $*$ has to survive until round m_*. In other words, all arms $j \in s^*$ such that $m_j < m_*$ are eliminated on or before m_* (this corresponds to case b1). Let, the arms surviving till m_* round be denoted by A'_{s^*}. This leaves any arm a_b such that $m_b \geq m_*$ to still survive and eliminate arm $*$ in round m_*. Let, such arms that survive $*$ belong to A''_{s^*}. Also maximal regret per step after eliminating $*$ is the maximal Δ_j among the remaining arms in A''_{s^*} with $m_j \geq m_*$. Let $m_b = \min\{m | \sqrt{\rho_a \epsilon_m} < \frac{\Delta_b}{2}\}$. Let $C_2(x) = \frac{2^{2x+\frac{3}{2}} x^{2x}}{\psi^x}$. Hence, the maximal regret after eliminating the arm $*$ is upper bounded by,

$$\begin{aligned}
&\sum_{m_*=0}^{max_{j \in A'_{s^*}} m_j} \sum_{\substack{i \in A''_{s^*}: \\ m_i \geq m_*}} \left(\frac{2}{(\psi T \epsilon_{m_*}^2)^{\rho_a}} \right) .T \max_{\substack{j \in A''_{s^*}: \\ m_j \geq m_*}} \Delta_j \\
&\leq \sum_{m_*=0}^{max_{j \in A'_{s^*}} m_j} \sum_{i \in A''_{s^*}: m_i \geq m_*} \left(\frac{2}{(\psi T \epsilon_{m_*}^2)^{\rho_a}} \right) .T.2\sqrt{\rho_a \epsilon_{m_*}} \\
&\leq \sum_{m_*=0}^{max_{j \in A'_{s^*}} m_j} \sum_{i \in A''_{s^*}: m_i \geq m_*} 4 \left(\frac{T^{1-\rho_a}}{\psi^{\rho_a} \epsilon_{m_*}^{2\rho_a - \frac{1}{2}}} \right) \\
&\leq \sum_{i \in A''_{s^*}: m_i \geq m_*} \sum_{m_*=0}^{\min\{m_i, m_b\}} \left(\frac{4T^{1-\rho_a}}{\psi^{\rho_a} 2^{-(2\rho_a - \frac{1}{2}) m_*}} \right) \\
&\leq \sum_{i \in A'_{s^*}} \frac{4T^{1-\rho_a}}{\psi^{\rho_a} 2^{-(2\rho_a - \frac{1}{2}) m_*}} + \sum_{i \in A''_{s^*} \setminus A'_{s^*}} \frac{4T^{1-\rho_a}}{\psi^{\rho_a} 2^{-(2\rho_a - \frac{1}{2}) m_b}} \\
&\leq \sum_{i \in A'_{s^*}} \frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi^{\rho_a} \Delta_i^{4\rho_a - 1}} + \sum_{i \in A''_{s^*} \setminus A'_{s^*}} \frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi^{\rho_a} b^{4\rho_a - 1}} \\
&= \sum_{i \in A'_{s^*}} \frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \sum_{i \in A''_{s^*} \setminus A'_{s^*}} \frac{C_{2(\rho_a)} T^{1-\rho_a}}{b^{4\rho_a - 1}}.
\end{aligned}$$

Case b3: *s^* is eliminated by some sub-optimal cluster.*

Let $C'_g = \{a_{max_{s_k}} \in A' | \forall s_k \in S\}$ and $C''_g = \{a_{max_{s_k}} \in A'' | \forall s_k \in S\}$. A sub-optimal cluster s_k will eliminate s^* in round g_* only if the cluster elimination condition of Algorithm 1 holds, which is the following when $* \in C_{g_*}$:

$$\hat{r}_{a_{\max_{s_k}}} - c_{g_*} > \hat{r}^* + c_{g_*}. \tag{2}$$

Notice that when $* \notin C_{g_*}$, since $r_{a_{max_{s_k}}} > r^*$, the inequality in (2) has to hold for cluster s_k to eliminate s^*. As in case $b2$, the probability that a given sub-optimal cluster s_k eliminates s^* is upper bounded by $\frac{2}{(\psi T \epsilon_{g_{s^*}}^2)^{\rho_s}}$ and all sub-optimal clusters with $g_{s_j} < g_*$ are eliminated before round g_*.

This leaves any arm $a_{\max_{s_b}}$ such that $g_{s_b} \geq g_*$ to still survive and eliminate arm $*$ in round g_*. Let, such arms that survive $*$ belong to C''_g. Hence, following the same way as case $b2$, the maximal regret after eliminating $*$ is,

$$\sum_{g_*=0}^{\max_{a_{\max_{s_j}} \in C'_g} g_{s_j}} \sum_{\substack{a_{\max_{s_k}} \in C''_g: \\ g_{s_k} \geq g_*}} \left(\frac{2}{(\psi T \epsilon_{g_{s^*}}^2)^{\rho_s}} \right) T \max_{\substack{a_{\max_{s_j}} \in C''_g: \\ g_{s_j} \geq g_*}} \Delta_{a_{\max_{s_j}}}$$

Using $A' \supset C'_g$ and $A'' \supset C''_g$, we can bound the regret contribution from this case in a similar manner as Case $b2$ as follows:

$$\begin{aligned}
&\sum_{i \in A' \setminus A'_{s^*}} \frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + \frac{3}{2}}}{\psi^{\rho_s} \Delta_i^{4\rho_s - 1}} + \sum_{i \in A'' \setminus A' \cup A'_{s^*}} \frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + \frac{3}{2}}}{\psi^{\rho_s} b^{4\rho_s - 1}} \\
&= \sum_{i \in A' \setminus A'_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} + \sum_{i \in A'' \setminus A' \cup A'_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{b^{4\rho_s - 1}}
\end{aligned}$$

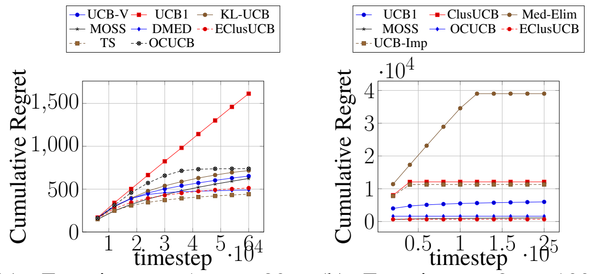
Case b4: *$*$ is not in $C_{\max(m_i, g_{s_k})}$, but belongs to $B_{\max(m_i, g_{s_k})}$.*

In this case the optimal arm $* \in s^*$ is not eliminated, also s^* is not eliminated. So, for all sub-optimal arms i in A'_{s^*} which gets eliminated on or before $\max\{m_i, g_{s_k}\}$ will get pulled no less than $\left\lceil \frac{2 \log(\psi T \epsilon_{m_i}^2)}{\epsilon_{m_i}} \right\rceil$ number of times, which leads to the following bound the contribution to the expected regret, as in Case $b1$:

$$\sum_{i \in A'_{s^*}} \left\{ \Delta_i + \frac{32\rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} \right\}$$

For arms $a_i \notin s^*$, the contribution to the regret cannot be greater than that in Case $b3$. So the regret is bounded by,

$$\sum_{i \in A' \setminus A'_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} + \sum_{i \in A'' \setminus A' \cup A'_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{b^{4\rho_s - 1}}$$



(a) Experiment 1: 20 Bernoulli-distributed arms with $r_{i_{i\neq *}} = 0.07$ and $r^* = 0.1$.

(b) Experiment 2: 100 Gaussian-distributed arms with $r_{i_{i\neq *:1-33}} = 0.1$, $r_{i_{i\neq *:34-99}} = 0.6$ and $r^*_{i=100} = 0.9$.

Figure 1: Cumulative regret for various bandit algorithms on two stochastic K-armed bandit environments.

The main claim follows by summing the contributions to the expected regret from each of the cases above. $\square$

Proposition 1. *The regret R_T for ClusUCB-AE satisfies*

$$\mathbb{E}[R_T] \leq \sum_{\substack{i\in A\\ \Delta_i > b}} \left\{ \frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \Delta_i + \frac{32\rho_a \log\left(\frac{\psi T \Delta_i^4}{16\rho_a^2}\right)}{\Delta_i} + \frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} \right\} + \sum_{\substack{i\in A\\ 0<\Delta_i\leq b}} \frac{C_2(\rho_a) T^{1-\rho_a}}{b^{4\rho_a-1}} + \max_{\substack{i\in A:\\ \Delta_i\leq b}} \Delta_i T,$$

for all $b \geq \sqrt{\frac{K}{14T}}$. In the above, C_1, C_2 are as defined in Theorem 1.

Proof. See Appendix A. $\square$

EClusucb has the same regret guarantees as ClusUCB and its regret upper bound is shown in Theorem 2 in Appendix E. Also the simple regret guarantee of EClusUCB is weaker than CCB(Liu & Tsuruoka, 2016) which is shown in Theorem 3 and Corollary 3 in Appendix F. But, this is expected as EClusUCB is geared towards minimizing cumulative regret whereas CCB is made for minimizing simple regret. Also we know from Bubeck et al. (2009) that algorithms that tend to minimize cumulative regret necessarily ends up having a poorer simple regret guarantee.

6. Simulation experiments

For the purpose of performance comparison using cumulative regret as the metric, we implement the following algorithms: KL-UCB(Garivier & Cappé, 2011), DMED(Honda & Takemura, 2010), MOSS(Audibert & Bubeck, 2009), UCB1(Auer et al., 2002a), UCB-Improved(Auer & Ortner, 2010), Median Elimination(Even-Dar et al., 2006), Thompson Sampling(TS)(Agrawal & Goyal, 2011), OCUCB(Lattimore, 2015) and UCB-V(Audibert et al., 2009)[2]. The parameters of ClusUCB/EClusUCB algorithm for all the experiments are set as follows: $\psi = \frac{T}{196 \log K}$, $\rho_s = 0.5$, $\rho_a = 0.5$ and $p = \lceil \frac{K}{\log K} \rceil$ (as in Corollary 2).

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 $r_{i_{i\neq *}} = 0.07$ and $r^* = 0.1$. These type of cases are frequently encountered in web-advertising domain. The horizon T is set to 60000. The regret is averaged over 100 independent runs and is shown in Figure 1(a). EClusUCB, MOSS, UCB1, UCB-V, KL-UCB, TS and DMED are run in this experimental setup and we observe that EClusUCB performs better than all the aforementioned algorithms except TS. Because of the small gaps and short horizon T, we do not implement UCB-Improved and Median Elimination on this test-case.

The second experiment is conducted over a testbed of 100 arms involving Gaussian reward distribution with expected rewards of the arms $r_{i_{i\neq *:1-33}} = 0.1$, $r_{i_{i\neq *:34-99}} = 0.6$ and $r^*_{i=100} = 0.9$ with variance set at $\sigma_i^2 = 0.3, \forall i \in A$. The horizon T is set for a large duration of 2×10^5 and the regret is averaged over 100 independent runs and is shown in Figure 1(b). In this case, in addition to EClusUCB, we also show the performance of ClusUCB algorithm. From the results in Figure 1(b), we observe that EClusUCB outperforms ClusUCB as well as MOSS, UCB1, UCB-Improved and Median-Elimination($\epsilon = 0.1, \delta = 0.1$). But ClusUCB and UCB-Improved behaves almost similarly in this environment. Also the performance of UCB-Improved is poor in comparison to other algorithms, which is probably because of pulls wasted in initial exploration whereas EClusUCB with the choice of ψ, ρ_a and ρ_s performs much better. More experiments are shown in Appendix H.

7. Conclusions and future work

From a theoretical viewpoint, we conclude that the gap-dependent regret bound of ClusUCB is lower than MOSS and UCB-Improved. From the numerical experiments on settings with small gaps between optimal and sub-optimal mean rewards, we observed that EClusUCB outperforms several popular bandit algorithms. While we exhibited better regret bounds for ClusUCB, it would be interesting future research to improve the theoretical analysis of ClusUCB to achieve the gap-independent regret bound of MOSS and OCUCB. This is also one of the first papers to apply clustering in stochastic bandits and another future direction is to use this in contextual bandits.

[2]The implementation for KL-UCB and DMED were taken from (Cappe et al., 2012)

References

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

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

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

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

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

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

Bubeck, Sébastien, Munos, Rémi, and Stoltz, Gilles. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pp. 23–37. Springer, 2009.

Bubeck, Sébastien, Munos, Rémi, and Stoltz, Gilles. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

Bubeck, Sébastien, Cesa-Bianchi, Nicolo, and Lugosi, Gábor. Bandits with heavy tail. *arXiv preprint arXiv:1209.1727*, 2012.

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

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

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

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

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

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

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

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

Mannor, Shie and Tsitsiklis, John N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

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

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

Tolpin, David and Shimony, Solomon Eyal. Mcts based on simple regret. In *AAAI*, 2012.

Appendix

The Appendix is organized as follows. In Appendix A we prove Proposition 1. In Appendix B we prove Corollary 1 and in Appendix C we prove Corollary 2. Appendix D deals with the idea of why we do clustering and Appendix E proves the cumulative regret upper bound of EClusUCB. The simple regret bound of EClusUCB and its associated Corollary is proved in F. Algorithm 3, Adaptive Clustered UCB is shown in Appendix G. More experiments are shown in Appendix H.

A. Proof of Proposition 1

Proof. Let $p = 1$ such that all the arms in set A belongs to a single cluster. So, for each sub-optimal arm i, $m_i = \min\{m|\sqrt{\rho_a\epsilon_m} < \frac{\Delta_i}{2}\}$. Also $\rho_a \in (0, 1]$ is a constant in this proof. Let $A^{'} = \{i \in A : \Delta_i > b\}$ and $A^{''} = \{i \in A : \Delta_i > 0\}$.

Case a: *Some sub-optimal arm i is not eliminated in round m_i or before and the optimal arm $* \in B_{m_i}$*

Following the steps of Theorem 1 Case $a1$, an arbitrary sub-optimal arm $i \in A^{'}$ can get eliminated only when the event,

$$\hat{r}_i \leq r_i + c_{m_i} \text{ and } \hat{r}^* \geq r^* - c_{m_i} \tag{3}$$

takes place. So to bound the regret we need to bound the probability of the complementary event of these two conditions.

Putting the value of $n_{m_i} = \frac{2\log(\psi T\epsilon_{m_i}^2)}{\epsilon_{m_i}}$ in c_{m_i}, $c_{m_i} = \sqrt{\frac{\rho_a\epsilon_{m_i}\log(\psi T\epsilon_{m_i}^2)}{2*2\log(\psi T\epsilon_{m_i}^2)}} = \frac{\sqrt{\rho_a\epsilon_{m_i}}}{2} = \sqrt{\rho_a\epsilon_{m_i+1}} < \frac{\Delta_i}{4}$, as $\rho_a \in (0, 1]$.

Again, for $i \in A^{'}$,

$$\hat{r}_i + c_{m_i} \leq r_i + 2c_{m_i} < r_i + \Delta_i - 2c_{m_i} = r^* - 2c_{m_i} \leq \hat{r}^* - c_{m_i}$$

Applying Chernoff-Hoeffding bound and considering independence of complementary of the two events in 3,

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

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

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

Summing up over all arms in $A^{'}$ and bounding the regret for each arm $i \in A^{'}$ trivially by $T\Delta_i$, we obtain

$$\sum_{i\in A^{'}}\left(\frac{2T\Delta_i}{(\psi T\epsilon_{m_i}^2)^{\rho_a}}\right) \leq \sum_{i\in A^{'}}\left(\frac{2T\Delta_i}{(\psi T\frac{\Delta_i^4}{16\rho_a^2})^{\rho_a}}\right) \leq \sum_{i\in A^{'}}\left(\frac{2^{1+4\rho_a}T^{1-\rho_a}\rho_a^{2\rho_a}\Delta_i}{\psi^{\rho_a}\Delta_i^{4\rho_a}}\right) \leq \sum_{i\in A^{'}}\left(\frac{2^{1+4\rho_a}\rho_a^{2\rho_a}T^{1-\rho_a}}{\psi^{\rho_a}\Delta_i^{4\rho_a-1}}\right)$$

$$= \sum_{i\in A^{'}}\left(\frac{C_1(\rho_a)T^{1-\rho_a}}{\Delta_i^{4\rho_a-1}}\right), \text{where } C_1(x) = \frac{2^{1+4x}x^{2x}}{\psi^x}$$

Case b: *Either an arm i is eliminated in round m_i or before or else there is no optimal arm $* \in B_{m_i}$*

CASE B1: *$* \in B_{m_i}$ and each $i \in A^{'}$ is eliminated on or before m_i*

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

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

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

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

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

Summing over all arms in $A^{'}$ the total regret is given by,

$$\sum_{i \in A'} \Delta_i \left(1 + \frac{32\rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i^2}\right)$$

CASE B2: *Optimal arm $*$ is eliminated by a sub-optimal arm*

Firstly, if conditions of Case a holds then the optimal arm $*$ 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 $*$ gets eliminated then for any round from 1 to m_j all arms j such that $m_j < m_*$ were eliminated according to assumption in Case a. Let, the arms surviving till m_* round be denoted by $A^{'}$. This leaves any arm a_b such that $m_b \geq m_*$ to still survive and eliminate arm $*$ in round m_*. Let, such arms that survive $*$ belong to $A^{''}$. Also maximal regret per step after eliminating $*$ is the maximal Δ_j among the remaining arms j with $m_j \geq m_*$. Let $m_b = \min\{m | \sqrt{\rho_a \epsilon_m} < \frac{\Delta_b}{2}\}$. Hence, the maximal regret after eliminating the arm $*$ is upper bounded by,

$$\begin{aligned}
&\sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i \in A'': m_i > m_*} \left(\frac{2}{(\psi T \epsilon_{m_*}^2)^{\rho_a}}\right).T \max_{j \in A'': m_j \geq m_*} \Delta_j \\
&\leq \sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i \in A'': m_i > m_*} \left(\frac{2}{(\psi T \epsilon_{m_*}^2)^{\rho_a}}\right).T.2\sqrt{\rho_a \epsilon_{m_*}} \\
&\leq \sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i \in A'': m_i > m_*} 4\left(\frac{T^{1-\rho_a}}{\psi^{\rho_a} \epsilon_{m_*}^{2\rho_a - \frac{1}{2}}}\right) \\
&\leq \sum_{i \in A'': m_i > m_*} \sum_{m_*=0}^{\min\{m_i, m_b\}} \left(\frac{4T^{1-\rho_a}}{\psi^{\rho_a} 2^{-(2\rho_a - \frac{1}{2})m_*}}\right) \\
&\leq \sum_{i \in A'} \left(\frac{4T^{1-\rho_a}}{\psi^{\rho_a} 2^{-(2\rho_a - \frac{1}{2})m_*}}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{4T^{1-\rho_a}}{\psi^{\rho_a} 2^{-(2\rho_a - \frac{1}{2})m_b}}\right) \\
&\leq \sum_{i \in A'} \left(\frac{4\rho_a^{2\rho_a} T^{1-\rho_a} * 2^{2\rho_a - \frac{1}{2}}}{\psi^{\rho_a} \Delta_i^{4\rho_a - 1}}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{4\rho_a^{2\rho_a} T^{1-\rho_a} * 2^{2\rho_a - \frac{1}{2}}}{\psi^{\rho_a} b^{4\rho_a - 1}}\right) \\
&\leq \sum_{i \in A'} \left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi^{\rho_a} \Delta_i^{4\rho_a - 1}}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{2\rho_a + \frac{3}{2}}}{\psi^{\rho_a} b^{4\rho_a - 1}}\right) \\
&= \sum_{i \in A'} \left(\frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{C_{2(\rho_a)} T^{1-\rho_a}}{b^{4\rho_a - 1}}\right), \text{where } C_2(x) = \frac{2^{2x + \frac{3}{2}} x^{2x}}{\psi^x}
\end{aligned}$$

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

$$R_T \leq \sum_{i \in A:\Delta_i > b} \Bigg\{ \Big(\frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} \Big) + \Big(\Delta_i + \frac{32\rho_a \log{(\psi T \frac{\Delta_i^4}{16\rho_a^2})}}{\Delta_i} \Big) + \Big(\frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} \Big) \Bigg\}$$
$$+ \sum_{i \in A:0 < \Delta_i \leq b} \Big(\frac{C_2(\rho_a) T^{1-\rho_a}}{\psi^{\rho_a} b^{4\rho_a - 1}} \Big) + max_{i \in A:\Delta_i \leq b} \Delta_i T$$

□

B. Proof of Corollary 1

Proof. Here we take $\psi = \frac{T}{196 \log(K)}$, $\rho_a = \frac{1}{2}$ and $\rho_s = \frac{1}{2}$. Taking into account Theorem 1 below for all $b \geq \sqrt{\frac{K}{14T}}$

$$\mathbb{E}[R_T] \leq \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} \Bigg\{ \frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \Delta_i + \frac{32\rho_a \log{(\psi T \frac{\Delta_i^4}{16\rho_a^2})}}{\Delta_i} \Bigg\} + \sum_{\substack{i \in A, \\ \Delta_i > b}} \Bigg\{ 2\Delta_i + \frac{C_1(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}}$$
$$+ \frac{32\rho_a \log{(\psi T \frac{\Delta_i^4}{16\rho_a^2})}}{\Delta_i} + \frac{32\rho_s \log{(\psi T \frac{\Delta_i^4}{16\rho_s^2})}}{\Delta_i} \Bigg\} + \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} \frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \sum_{\substack{i \in A_{s^*}, \\ 0 < \Delta_i \leq b}} \frac{C_2(\rho_a) T^{1-\rho_a}}{b^{4\rho_a - 1}}$$
$$+ \sum_{i \in A \setminus A_{s^*}:\Delta_i > b} \frac{2C_2(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} + \sum_{i \in A \setminus A_{s^*}:0 < \Delta_i \leq b} \frac{2C_2(\rho_s) T^{1-\rho_s}}{b^{4\rho_s - 1}} + + \max_{i:\Delta_i \leq b} \Delta_i T$$

and putting the parameter values in the above Theorem 1 result,

$$\sum_{i \in A_{s^*}:\Delta_i > b} \Bigg(\frac{T^{1-\rho_a} \rho_a^{2\rho_a} 2^{1+4\rho_a}}{\psi^{\rho_a} \Delta_i^{4\rho_a - 1}} \Bigg) = \sum_{i \in A_{s^*}:\Delta_i > b} \Bigg(\frac{T^{1-\frac{1}{2}} \frac{1}{2}^{2*\frac{1}{2}} 2^{1+4*\frac{1}{2}}}{(\frac{T}{196 \log(K)})^{\frac{1}{2}} \Delta_i^{4*\frac{1}{2} - 1}} \Bigg) = \sum_{i \in A_{s^*}:\Delta_i > b} \frac{56\sqrt{\log(K)}}{\Delta_i}$$

Similarly for the term,

$$\sum_{i \in A:\Delta_i > b} \Bigg(\frac{T^{1-\rho_s} \rho_a^{2\rho_s} 2^{1+4\rho_s}}{\psi^{\rho_s} \Delta_i^{4\rho_s - 1}} \Bigg) = \sum_{i \in A:\Delta_i > b} \frac{56\sqrt{\log(K)}}{\Delta_i}$$

For the term involving arm pulls,

$$\sum_{i \in A:\Delta_i > b} \frac{32\rho_s \log{(\psi T \frac{\Delta_i^4}{16\rho_s^2})}}{\Delta_i} = \sum_{i \in A:\Delta_i > b} \frac{16 \log{(T^2 \frac{\Delta_i^4}{784 \log(K)})}}{\Delta_i} \approx \sum_{i \in A:\Delta_i > b} \frac{32 \log{(T \frac{\Delta_i^2}{\sqrt{\log(K)}})}}{\Delta_i}$$

Similarly the term,

$$\sum_{i \in A:\Delta_i > b} \frac{32\rho_a \log{(\psi T \frac{\Delta_i^4}{16\rho_a^2})}}{\Delta_i} \approx \sum_{i \in A:\Delta_i > b} \frac{32 \log{(T \frac{\Delta_i^2}{\sqrt{\log(K)}})}}{\Delta_i}$$

Lastly we can bound the error terms as,

$$\sum_{i\in A_{s^*}:0<\Delta_i\le b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi^{\rho_a}\Delta_i^{4\rho_a-1}}\right)=\sum_{i\in A_{s^*}:0<\Delta_i\le b}\frac{40\sqrt{\log(K)}}{\Delta_i}$$

Similarly for the term,

$$\sum_{i\in A\setminus A_{s^*}:0<\Delta_i\le b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+\frac{3}{2}}}{(\psi^{\rho_s})\Delta_i^{4\rho_s-1}}\right)=\sum_{i\in A\setminus A_{s^*}:0<\Delta_i\le b}\frac{40\sqrt{\log(K)}}{\Delta_i}$$

So, the total gap dependent regret bound for using both arm and cluster elimination comes of as

$$\begin{aligned}&\sum_{i\in A_{s^*}:\Delta_i>b}\left\{\frac{56\sqrt{\log(K)}}{\Delta_i}+\Delta_i+\frac{32\log\left(T\frac{\Delta_i^2}{\sqrt{\log(K)}}\right)}{\Delta_i}\right\}+\sum_{i\in A:\Delta_i>b}\left\{\frac{56\sqrt{\log(K)}}{\Delta_i}+2\Delta_i+\frac{64\log\left(T\frac{\Delta_i^2}{\sqrt{\log(K)}}\right)}{\Delta_i}\right\}\\&+\sum_{i\in A_{s^*}:\Delta_i>b}\frac{40\sqrt{\log(K)}}{\Delta_i}+\sum_{i\in A_{s^*}:0<\Delta_i<\le b}\frac{40\sqrt{\log(K)}}{\Delta_i}+\sum_{i\in A\setminus A_{s^*}:\Delta_i>b}\frac{80\sqrt{\log(K)}}{\Delta_i}\\&+\sum_{i\in A\setminus A\cup A_{s^*}:0<\Delta_i\le b}\frac{80\sqrt{\log(K)}}{\Delta_i}+\max_{i\in A:\Delta_i\le b}\Delta_i T\end{aligned}$$

□

C. Proof of Corollary 2

Proof. As stated in Auer & Ortner (2010), we can have a bound on regret of the order of $\sqrt{KT\log K}$ in non-stochastic setting. This is shown in Exp4(Auer et al., 2002b) algorithm. Also we know from Bubeck et al. (2011) that the function $x\in[0,1]\mapsto x\exp(-Cx^2)$ is decreasing on $\left[\frac{1}{\sqrt{2C}},1\right]$ for any $C>0$. So, taking $C=\left\lfloor\frac{14T}{K}\right\rfloor$ and similarly, by choosing $\Delta_i=\Delta=\sqrt{\frac{K\log K}{T}}>\sqrt{\frac{K}{14T}}$ for all $i:i\neq *\in A$, in the bound of UCB1(Auer et al., 2002a) we get,

$$\sum_{i:r_i<r^*}const\frac{\log T}{\Delta_i}=\frac{\sqrt{KT}\log T}{\sqrt{\log K}}$$

So, this bound is worse than the non-stochastic setting and is clearly improvable and an upper bound regret of $\sqrt{KT}$ is possible as shown in Audibert & Bubeck (2009) for MOSS which is consistent with the lower bound as proposed by Mannor and Tsitsiklis(Mannor & Tsitsiklis, 2004).

Hence, we take $b\approx\sqrt{\frac{K\log K}{T}}>\sqrt{\frac{K}{14T}}$(the minimum value for b), $\psi=\frac{T}{196\log K}$, $\rho_a=\frac{1}{2}$ and $\rho_s=\frac{1}{2}$.

Taking into account Theorem 1 below,

$$\mathbb{E}[R_T]\le\sum_{\substack{i\in A_{s^*},\\ \Delta_i>b}}\left\{\frac{C_1(\rho_a)T^{1-\rho_a}}{\Delta_i^{4\rho_a-1}}+\Delta_i+\frac{32\rho_a\log\left(\psi T\frac{\Delta_i^4}{16\rho_a^2}\right)}{\Delta_i}\right\}+\sum_{\substack{i\in A,\\ \Delta_i>b}}\left\{2\Delta_i+\frac{C_1(\rho_s)T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right.$$

$$+\frac{32\rho_a\log(\psi T\dfrac{\Delta_i^4}{16\rho_a^2})}{\Delta_i}+\frac{32\rho_s\log(\psi T\dfrac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}\Bigg\}+\sum_{\substack{i\in A_{s^*},\\ \Delta_i>b}}\frac{C_2(\rho_a)T^{1-\rho_a}}{\Delta_i^{4\rho_a-1}}+\sum_{\substack{i\in A_{s^*},\\ 0<\Delta_i\le b}}\frac{C_2(\rho_a)T^{1-\rho_a}}{b^{4\rho_a-1}}$$

$$+\sum_{i\in A\setminus A_{s^*}:\Delta_i>b}\frac{2C_2(\rho_s)T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}+\sum_{i\in A\setminus A_{s^*}:0<\Delta_i\le b}\frac{2C_2(\rho_s)T^{1-\rho_s}}{b^{4\rho_s-1}}++\max_{i:\Delta_i\le b}\Delta_i T$$

and putting the parameter values in the above Theorem 1 result,

$$\sum_{i\in A_{s^*}:\Delta_i>b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{1+4\rho_a}}{\psi^{\rho_a}\Delta_i^{4\rho_a-1}}\right)=\left(K\frac{T^{1-\frac{1}{2}}\frac{1}{2}^{2\frac{1}{2}}2^{1+4\frac{1}{2}}}{p(\frac{T}{196\log K})^{\frac{1}{2}}\Delta_i^{4\frac{1}{2}-1}}\right)=56\frac{\sqrt{KT}}{p}$$

Similarly, for the term,

$$\sum_{i\in A:\Delta_i>b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{1+4\rho_s}}{\psi^{\rho_s}\Delta_i^{4\rho_s-1}}\right)=56\sqrt{KT}$$

For the term regarding number of pulls,

$$\sum_{i\in A:\Delta_i>b}\frac{32\rho_s\log(\psi T\dfrac{\Delta_i^4}{16\rho_s^2})}{\Delta_i}=\frac{32K\sqrt{T}\frac{1}{2}\log(T^2\dfrac{K^4(\log K)^2}{784T^2\log K})}{\sqrt{K\log K}}\le\frac{32\sqrt{KT}\log(\frac{1}{28}K^2(\sqrt{\log K}))}{\sqrt{\log K}}$$

$$\le 64\sqrt{KT\log K}+\frac{32\sqrt{KT}\log(\sqrt{\log K})}{\sqrt{\log K}}$$

Similarly for the term,

$$\sum_{i\in A:\Delta_i>b}\frac{32\rho_a\log(\psi T\dfrac{\Delta_i^4}{16\rho_a^2})}{\Delta_i}\le 64\sqrt{KT\log K}+\frac{32\sqrt{KT}\log(\sqrt{\log K})}{\sqrt{\log K}}$$

Lastly we can bound the error terms as,

$$\sum_{i\in A_{s^*}:0\le\Delta_i\le b}\left(\frac{T^{1-\rho_a}\rho_a^{2\rho_a}2^{2\rho_a+\frac{3}{2}}}{\psi^{\rho_a}\Delta_i^{4\rho_a-1}}\right)=\frac{K}{p}\left(\frac{T^{1-\frac{1}{2}}\frac{1}{2}^{2\frac{1}{2}}2^{2\frac{1}{2}+\frac{3}{2}}}{(\frac{T}{196\log K})^{\frac{1}{2}}(\Delta_i)^{4*\frac{1}{2}-1}}\right)<\frac{150\sqrt{KT\log K}}{p}$$

Similarly for the term,

$$\sum_{i\in A\setminus A_{s^*}:\Delta_i>b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+\frac{3}{2}}}{(\psi^{\rho_s})\Delta_i^{4\rho_s-1}}\right)<150(K-\frac{K}{p})\sqrt{\frac{T}{K\log K}}$$

Also, for all $b\ge\sqrt{\dfrac{K}{14T}}$,

$$\sum_{i\in A\setminus A_{s^*}:0<\Delta_i\le b}\left(\frac{T^{1-\rho_s}\rho_s^{2\rho_s}2^{2\rho_s+\frac{3}{2}}}{(\psi^{\rho_s})b^{4\rho_s-1}}\right)<150(K-\frac{K}{p})\sqrt{\frac{T\log K}{K}}$$

Now, $K-\dfrac{K}{p}=K\left(\dfrac{p-1}{p}\right)<K\left(\dfrac{\frac{K}{\log K}+1-1}{\frac{K}{\log K}+1}\right)<\dfrac{K^2}{K+\log K}$. So, after putting the value of $p=\left\lceil\dfrac{K}{\log K}\right\rceil$, we get,

Elim Type	Error Bound	Remarks
Only Arm Elimination (ClusUCB-AE)	$\underbrace{\sum_{i\in A:\Delta_i>b}\left(\frac{C_2(\rho_a)T^{1-\rho_a}}{\Delta_i^{4\rho_a-1}}\right)}_{\text{Case b2, Proposition 1}}+\underbrace{\sum_{i\in A:0<\Delta_i\leq b}\left(\frac{C_2(\rho_a)T^{1-\rho_a}}{b^{4\rho_a-1}}\right)}_{\text{Case b2, Proposition 1}}$	With $\rho_a=\frac{1}{2}$, and $\psi=\frac{T}{196\log K}$ this gives $150\sqrt{KT}+150\sqrt{KT\log K}$. Hence, this has an order of $O(\sqrt{KT\log K})$.
Arm & Cluster Elimination (ClusUCB)	$\underbrace{\sum_{i\in A_{s^*}:\Delta_i>b}\left(\frac{C_2(\rho_a)T^{1-\rho_a}}{\Delta_i^{4\rho_a-1}}\right)+\sum_{i\in A_{s^*}:0\leq\Delta_i\leq b}\left(\frac{C_2(\rho_a)T^{1-\rho_a}}{b^{4\rho_a-1}}\right)}_{\text{Case b2, Arm Elim, Theorem 1}}$ $+\underbrace{\sum_{i\in A\setminus A_{s^*}:\Delta_i>b}\left(\frac{2C_2(\rho_s)T^{1-\rho_s}}{\Delta_i^{4\rho_s-1}}\right)+\sum_{i\in A\setminus A_{s^*}:0\leq\Delta_i\leq b}\left(\frac{2C_2(\rho_s)T^{1-\rho_s}}{b^{4\rho_s-1}}\right)}_{\text{Case b3+b4, Clus Elim, Theorem 1}}$	With $\rho_a=\frac{1}{2}$, $\rho_s=\frac{1}{2}, p=\lceil\frac{K}{\log K}\rceil$ and $\psi=\frac{T}{196\log K}$ this gives $\frac{150\sqrt{T}\log K^{\frac{3}{2}}}{\sqrt{K}}+\frac{150\sqrt{T}\log K}{\sqrt{K}}+300\frac{K}{K+\log K}\sqrt{KT\log K}+300\frac{K}{K+\log K}\sqrt{KT}$. So we can reduce the error bound to $O(\frac{K}{K+\log K}\sqrt{KT\log K})$.

Table 2: Error Bound

$$\mathbb{E}[R_T]\leq 56\frac{\sqrt{T}\log K}{\sqrt{K}}+64\frac{\sqrt{T}\log K}{\sqrt{K}}+\frac{32\sqrt{T\log K}\log(\log K)}{\sqrt{K}}+56\sqrt{KT}+128\sqrt{KT\log K}$$
$$+\frac{64\sqrt{KT}\log(\log K)}{\sqrt{\log K}}+\frac{150\sqrt{T}\log K^{\frac{3}{2}}}{\sqrt{K}}+\frac{150\sqrt{T}\log K}{\sqrt{K}}+300\frac{K}{K+\log K}\sqrt{KT\log K}+300\frac{K}{K+\log K}\sqrt{KT}$$

So, the total bound for using both arm and cluster elimination cannot be worse than,

$$\mathbb{E}[R_T]\leq 270\frac{\sqrt{T}\log K}{\sqrt{K}}+\frac{32\sqrt{T\log K}\log(\log K)}{\sqrt{K}}+56\sqrt{KT}+128\sqrt{KT\log K}$$
$$+\frac{64\sqrt{KT}\log(\log K)}{\sqrt{\log K}}+\frac{150\sqrt{T}\log K^{\frac{3}{2}}}{\sqrt{K}}+300\frac{K}{K+\log K}\sqrt{KT\log K}+300\frac{K}{K+\log K}\sqrt{KT}$$

□

D. Why Clustering?

In this section we want to specify the apparent use of clustering. The error bounds are shown in Table ??.

While looking at the error terms in Table ??, we see that using just arm elimination (ClusUCB-AE) the elimination error bound is more than using both arm and cluster elimination simultaneously (ClusUCB).

E. Proof of Theorem 2

Theorem 2 (***Regret bound***). *The regret R_T of EClusUCB satisfies*

$$\mathbb{E}[R_T] \leq \sum_{\substack{i \in A_{s^*},\\ \Delta_i > b}} \left\{ \frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \Delta_i + \frac{32\rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} \right\} + \sum_{\substack{i \in A,\\ \Delta_i > b}} \left\{ 2\Delta_i + \frac{C_1(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} \right.$$

$$\left. + \frac{32\rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} + \frac{32\rho_s \log(\psi T \frac{\Delta_i^4}{16\rho_s^2})}{\Delta_i} \right\} + \sum_{\substack{i \in A_{s^*},\\ \Delta_i > b}} \frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \sum_{\substack{i \in A_{s^*},\\ 0 < \Delta_i \leq b}} \frac{C_2(\rho_a) T^{1-\rho_a}}{b^{4\rho_a - 1}}$$

$$+ \sum_{\substack{i \in A \setminus A_{s^*}:\\ \Delta_i > b}} \frac{2C_2(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} + \sum_{\substack{i \in A \setminus A_{s^*}:\\ 0 < \Delta_i \leq b}} \frac{2C_2(\rho_s) T^{1-\rho_s}}{b^{4\rho_s - 1}}$$

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

where $b \geq \sqrt{\frac{K}{14T}}$, $C_1(x) = \frac{2^{1+4x} x^{2x}}{\psi^x}$, $C_2(x) = \frac{2^{2x+\frac{3}{2}} x^{2x}}{\psi^x}$ and A_{s^*} is the subset of arms in cluster s^* containing optimal arm a^*.

We can see from the result of this theorem that the regret upper bound for EClusUCB is same as ClusUCB. Also we can specialize the result of this Theorem 2 using Corollary 1 and 2 to achieve a similar bound as for Theorem 1.

Proof. Let $A^{'} = \{i \in A, \Delta_i > b\}$, $A^{''} = \{i \in A, \Delta_i > 0\}$, $A^{'}_{s_k} = \{i \in A_{s_k}, \Delta_i > b\}$ and $A^{''}_{s_k} = \{i \in A_{s_k}, \Delta_i > 0\}$. C_g is the cluster set containing max payoff arm from each cluster in g-th round. The arm having the highest payoff in a cluster s_k is denote by $a_{\max_{s_k}}$. Let for each sub-optimal arm $i \in A$, $m_i = \min\{m | \sqrt{\rho_a \epsilon_m} < \frac{\Delta_i}{2}\}$ and let for each cluster $s_k \in S$, $g_{s_k} = \min\{g | \sqrt{\rho_s \epsilon_g} < \frac{\Delta_{a_{\max_{s_k}}}}{2}\}$. Let $\check{A} = \{i \in A^{'} | i \in s_k, \forall s_k \in S\}$. Also z_i denotes total number of times an arm i has been pulled from 1 to $(t-1)$-th timestep. In the m-th round, n_m denotes the number of pulls allocated to the surviving arms in B_m. The analysis proceeds by considering the contribution to the regret in each of the following cases:

Case a: *Some sub-optimal arm i is not eliminated in round $\max(m_i, g_{s_k})$ or before, with the optimal arm $* \in C_{\max(m_i, g_{s_k})}$.*

We consider an arbitrary sub-optimal arm i and analyze the contribution to the regret when i is not eliminated in the following exhaustive sub-cases:
Case a1: *In round $\max(m_i, g_{s_k})$, $i \in s^*$.*

This case is similar to Case a1, Theorem 1. Let $c_i = \sqrt{\frac{\rho_a \log(\psi T \epsilon_m^2)}{2z_i}}$. For an arm to get eliminated it must satisfy the following condition,

$$\hat{r}_i \leq r_i + c_i \text{ and } \hat{r}^* \geq r^* - c^*, \tag{4}$$

Since each round now consist of $|B_m| n_m$ timesteps and since at every timestep the arm elimination condition is being checked, following a parallel arguement as in Case a1, Theorem 1 we can show that when $z_i = n_{m_i}$ then
$c_i \leq c_{m_i} \leq \sqrt{\frac{\rho_a \log(\psi T \epsilon_m^2)}{2n_{m_i}}} = \sqrt{\rho_a \epsilon_{m_i+1}} < \frac{\Delta_i}{4}$, since $z_i = n_{m_i} = \frac{2\log(\psi T \epsilon_{m_i}^2)}{\epsilon_{m_i}}$ and $\rho_a \in (0, 1]$.

Subsequently following the steps of Case a1, Theorem 1 we can upper bound the probability of the complementary of the events in 4 and show that the probability that a sub-optimal arm i is not eliminated in any round on or before m_i is bounded above by $\left(\frac{2}{(\psi T \epsilon_{m_i}^2)^{\rho_a}}\right)$.

Case a2: *In round $\max(m_i, g_{s_k})$, $i \in s_k$ for some $s_k \neq s^*$.*

Let $c_{a_{\max_{s_k}}} = \sqrt{\dfrac{\rho_s \log(\psi T \epsilon_m^2)}{2 z_{a_{max_{s_k}}}}}$ and $z_{a_{max_{s_k}}}$ is the number of times $a_{max_{s_k}}$ has been pulled. Following a parallel argument like in Case $a1$, we have to bound the following two events of arm $a_{\max_{s_k}}$ not getting eliminated on or before g_{s_k}-th round,

$$\hat{r}_{a_{\max_{s_k}}} \geq r_{a_{\max_{s_k}}} + c_{a_{\max_{s_k}}} \text{ and } \hat{r}^* \leq r^* - c^*$$

As we argued in previous case, since cluster elimination condition being verified in each timestep, the above two conditions are possible when $z_{a_{max_{s_k}}} = n_{g_{s_k}}$. We can prove using Chernoff-Hoeffding bounds and considering independence of events mentioned above, that for $c_{a_{\max_{s_k}}} \leq c_{g_{s_k}} \leq \sqrt{\frac{\rho_s \log(\psi T \epsilon_{g_{s_k}}^2)}{2 n_{g_{s_k}}}}$ and $n_{g_{s_k}} = \frac{2\log(\psi T \epsilon_{g_{s_k}}^2)}{\epsilon_{g_{s_k}}}$ the probability of the above two events is bounded by $\left(\dfrac{2}{(\psi T \epsilon_{g_{s_k}}^2)^{\rho_s}}\right)$. Now, for any round g_{s_k}, all the elements of $C_{\max(m_i, g_{s_k})}$ are the respective maximum payoff arms of their cluster $s_k, \forall s_k \in S$, and since clusters are fixed so we can bound the maximum probability that a sub-optimal arm $i \in A^{'}$ and $i \in s_k$ such that $a_{\max_{s_k}} \in C_{g_{s_k}}$ is not eliminated on or before the g_{s_k}-th round by the same probability as above.

Summing up over all p clusters and bounding the regret for each arm $i \in A^{'}_{s_k}$ trivially by $T\Delta_i$ and following the steps of Case a2, Theorem 1, we can show that the regret for not eliminating clusters on or before round g_{s_k} is upper bounded by,

$$\sum_{i \in A'} \frac{C_1(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}}$$

Summing the bounds in Cases $a1 - a2$ and observing that the bounds in the aforementioned cases hold for any round $C_{\max\{m_i, g_{s_k}\}}$, we obtain the following contribution to the expected regret from Case a:

$$\sum_{i \in A_{s^*}} \frac{C_1(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \sum_{i \in A'} \left(\frac{C_1(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}}\right)$$

Case b: *For each arm i, either i is eliminated in round* $\max(m_i, g_{s_k})$ *or before or there is no optimal arm $*$ in* $C_{\max(m_i, g_{s_k})}$.

Case b1: $* \in C_{\max(m_i, g_{s_k})}$ *for each arm $i \in A'$ and cluster $s_k \in \check{A}$.*

The condition in the case description above implies the following:
(i) each sub-optimal arm $i \in A^{'}$ is eliminated on or before $\max(m_i, g_{s_k})$ and hence pulled not more than z_i number of times. But according to the condition of Case a1, $z_i < n_{m_i}$ number of times.
(ii) each sub-optimal cluster $s_k \in \check{A}$ is eliminated on or before $\max(m_i, g_{s_k})$ and hence pulled not more than $n_{a_{\max_{s_k}}}$ number of times. But again according to condition Case a2, $z_{a_{\max_{s_k}}} < n_{g_{s_k}}$ number of times.

Hence, following the same steps as in Case b1, Theorem 1, the maximum regret suffered due to pulling of a sub-optimal arm or a sub-optimal cluster is no more than the following:

$$\sum_{i \in A'} \left[2\Delta_i + \frac{32(\rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2}) + \rho_s \log(\psi T \frac{\Delta_i^4}{16\rho_s^2}))}{\Delta_i} \right]$$

Case b2: $*$ *is eliminated by some sub-optimal arm in s^**
Optimal arm a^* can get eliminated by some sub-optimal arm i only if arm elimination condition holds, i.e.,

$$\hat{r}_i - c_i > \hat{r}^* + c^*,$$

where, as mentioned before, $c_i \leq c_{m_i} \leq \sqrt{\frac{\rho_a \log(\psi T \epsilon_{m_i}^2)}{2n_{m_i}}}$. From analysis in Case $a1$, notice that, if (4) holds in conjunction with the above, arm i gets eliminated. Also, recall from Case $a1$ that the events complementary to (4) have low-probability and can be upper bounded by $\frac{2}{(\psi T \epsilon_{m_*}^2)^{\rho_a}}$. Moreover, a sub-optimal arm that eliminates $*$ has to survive until round m_*. In other words, all arms $j \in s^*$ such that $m_j < m_*$ are eliminated on or before m_* (this corresponds to case b1). Let, the arms surviving till m_* round be denoted by $A^{'}_{s^*}$. This leaves any arm a_b such that $m_b \geq m_*$ to still survive and eliminate arm $*$ in round m_*. Let, such arms that survive $*$ belong to $A^{''}_{s^*}$. Also maximal regret per step after eliminating $*$ is the maximal Δ_j among the remaining arms in $A^{''}_{s^*}$ with $m_j \geq m_*$. Let $m_b = \min\{m | \sqrt{\rho_a \epsilon_m} < \frac{\Delta_b}{2}\}$. Let $C_2(x) = \frac{2^{2x+\frac{3}{2}} x^{2x}}{\psi^x}$. Hence, the maximal regret after eliminating the arm $*$ is upper bounded by,

$$\sum_{m_*=0}^{max_{j \in A^{'}_{s^*}} m_j} \sum_{\substack{i \in A^{''}_{s^*}: \\ m_i \geq m_*}} \left(\frac{2}{(\psi T \epsilon_{m_*}^2)^{\rho_a}} \right) . T \max_{\substack{j \in A^{''}_{s^*}: \\ m_j \geq m_*}} \Delta_j$$

Therefore, following the similar steps as in Case b2, Theorem 1, we can show that the above regret is upper bounded by,

$$\sum_{i \in A^{'}_{s^*}} \frac{C_2(\rho_a) T^{1-\rho_a}}{\Delta_i^{4\rho_a - 1}} + \sum_{i \in A^{''}_{s^*} \setminus A^{'}_{s^*}} \frac{C_{2(\rho_a)} T^{1-\rho_a}}{b^{4\rho_a - 1}}$$

Case b3: *s^* is eliminated by some sub-optimal cluster.*

Let $C^{'}_g = \{a_{max_{s_k}} \in A^{'} | \forall s_k \in S\}$ and $C^{''}_g = \{a_{max_{s_k}} \in A^{''} | \forall s_k \in S\}$. A sub-optimal cluster s_k will eliminate s^* in round g_* only if the cluster elimination condition of Algorithm 2 holds, which is the following when $* \in C_{g_*}$:

$$\hat{r}_{a_{\max_{s_k}}} - c_{a_{\max_{s_k}}} > \hat{r}^* + c^*. \tag{5}$$

Notice that when $* \notin C_{g_*}$, since $r_{a_{max_{s_k}}} > r^*$, the inequality in (5) has to hold for cluster s_k to eliminate s^*. As in case $b2$, the probability that a given sub-optimal cluster s_k eliminates s^* is upper bounded by $\frac{2}{(\psi T \epsilon_{g_{s^*}}^2)^{\rho_s}}$ and all sub-optimal clusters with $g_{s_j} < g_*$ are eliminated before round g_*.

This leaves any arm $a_{\max_{s_b}}$ such that $g_{s_b} \geq g_*$ to still survive and eliminate arm $*$ in round g_*. Let, such arms that survive $*$ belong to $C^{''}_g$. Hence, following the same way as case $b2$, the maximal regret after eliminating $*$ is,

$$\sum_{g_*=0}^{\max_{a_{\max_{s_j}} \in C^{'}_g} g_{s_j}} \sum_{\substack{a_{\max_{s_k}} \in C^{''}_g: \\ g_{s_k} \geq g_*}} \left(\frac{2}{(\psi T \epsilon_{g_{s^*}}^2)^{\rho_s}} \right) T \max_{\substack{a_{\max_{s_j}} \in C^{''}_g: \\ g_{s_j} \geq g_*}} \Delta_{a_{\max_{s_j}}}$$

Using $A' \supset C^{'}_g$ and $A'' \supset C^{''}_g$, we can bound the regret contribution from this case in a similar manner as Case $b2$ as follows:

$$\sum_{i \in A' \setminus A^{'}_{s^*}} \frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + \frac{3}{2}}}{\psi^{\rho_s} \Delta_i^{4\rho_s - 1}} + \sum_{i \in A'' \setminus A' \cup A^{'}_{s^*}} \frac{T^{1-\rho_s} \rho_s^{2\rho_s} 2^{2\rho_s + \frac{3}{2}}}{\psi^{\rho_s} b^{4\rho_s - 1}}$$
$$= \sum_{i \in A' \setminus A^{'}_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} + \sum_{i \in A'' \setminus A' \cup A^{'}_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{b^{4\rho_s - 1}}$$

Case b4: *$*$ is not in $C_{\max(m_i, g_{s_k})}$, but belongs to $B_{\max(m_i, g_{s_k})}$.*

In this case the optimal arm $* \in s^*$ is not eliminated, also s^* is not eliminated. So, for all sub-optimal arms i in A'_{s^*} which gets eliminated on or before $\max\{m_i, g_{s_k}\}$ will get pulled no less than $z_i < \left\lceil \frac{2\log(\psi T\epsilon_{m_i}^2)}{\epsilon_{m_i}} \right\rceil$ number of times, which leads to the following bound the contribution to the expected regret, as in Case $b1$:

$$\sum_{i \in A'_{s^*}} \left\{ \Delta_i + \frac{32\rho_a \log(\psi T \frac{\Delta_i^4}{16\rho_a^2})}{\Delta_i} \right\}$$

For arms $a_i \notin s^*$, the contribution to the regret cannot be greater than that in Case $b3$. So the regret is bounded by,

$$\sum_{i \in A' \setminus A'_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{\Delta_i^{4\rho_s - 1}} + \sum_{i \in A'' \setminus A' \cup A'_{s^*}} \frac{C_2(\rho_s) T^{1-\rho_s}}{b^{4\rho_s - 1}}$$

The main claim follows by summing the contributions to the expected regret from each of the cases above. □

F. Proof of Theorem 3

Theorem 3. *For every $0 < \eta < 1$ and $\gamma > 1$, there exists τ such that for all $T > \tau$ the simple regret of EClusUCB is upper bounded by,*

$$SR_{EClusUCB} \leq 4\log_2 \frac{14T}{K} \gamma \sum_{i=1}^{K} \Delta_i \exp(-\frac{c_0 \sqrt{e}}{4}) \left\{ K^{\frac{3}{2}+2\rho_a} \left(\frac{\log(\psi T)}{T^{\frac{3}{2}} (\psi T^2)^{\rho_a}} \right) + K^{\frac{3}{2}+2\rho_s} \left(\frac{\log(\psi T)}{T^{\frac{3}{2}} (\psi T^2)^{\rho_s}} \right) \right\}$$

with probability at least $1 - \eta$, where $c_0 > 0$ is a constant.

Proof. We follow the same steps as in Theorem 2, (Liu & Tsuruoka, 2016). First we will state the two facts used by this proof.

1. *Fact 1:* From Theorem 2 we know that the probability of elimination of a sub-optimal arm in the $\max(m_i, g_{s_k})$ round is $\left(\frac{2}{(\psi T \epsilon_{m_i}^2)^{\rho_a}} \right)$ and of a sub-optimal cluster is $\left(\frac{2}{(\psi T \epsilon_{g_{s_k}}^2)^{\rho_s}} \right)$.
2. *Fact 2:* From Tolpin & Shimony (2012) we know that, for every $0 < \eta < 1$ and $\gamma > 1$, there exists τ such that for all $T > \tau$ the probability of a sub-optimal arm i being sampled in the m-th round is bounded by $Q_m \leq 2\gamma \exp(-c_m \frac{\sqrt{T}}{2})$, where $c_m = \frac{c_0}{2^m}$.

We start with an upper bound on the number of plays $\delta_{\max(m_i, g_{s_k})}$ in the $\max(m_i, g_{s_k})$-th round divided by the total number of plays T. We know from Fact 1 that the total number of arms surviving in the $\max(m_i, g_{s_k})$-th arm is

$$|B_{\max(m_i, g_{s_k})}| \leq \left(\frac{2K}{(\psi T \epsilon_{m_i}^2)^{\rho_a}} \right) + \left(\frac{2K}{(\psi T \epsilon_{g_{s_k}}^2)^{\rho_s}} \right)$$

Again in EClusUCB, we know that the number of pulls allocated for each surviving arm i in the $\max(m_i, g_{s_k})$-th round is $n_{\max(m_i, g_{s_k})} = \frac{2\log(\psi T \epsilon_{\max(m_i, g_{s_k})}^2)}{\epsilon_{\max(m_i, g_{s_k})}}$. Therefore, the proportion of plays $\delta_{\max(m_i, g_{s_k})}$ in the $\max(m_i, g_{s_k})$-th round can be written as,

$$\delta_{\max(m_i,g_{s_k})} = \frac{(|B_{\max(m_i,g_{s_k})}|.n_{\max(m_i,g_{s_k})})}{T} \leq \left(\frac{1}{T}.\frac{2K}{(\psi T\epsilon_{m_i}^2)^{\rho_a}}.\frac{2\log(\psi T\epsilon_{m_i}^2)}{\epsilon_{m_i}}\right) + \left(\frac{1}{T}.\frac{2K}{(\psi T\epsilon_{g_{s_k}}^2)^{\rho_s}}.\frac{2\log(\psi T\epsilon_{g_{s_k}}^2)}{\epsilon_{g_{s_k}}}\right)$$
$$\leq \left(\frac{4K\log(\psi T\epsilon_{m_i}^2)}{T\epsilon_{m_i}(\psi T\epsilon_{m_i}^2)^{\rho_a}}\right) + \left(\frac{4K\log(\psi T\epsilon_{g_{s_k}}^2)}{T\epsilon_{g_{s_k}}(\psi T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right)$$

Now, $\epsilon_{m_i} \geq \sqrt{\frac{K}{14T}}$ and $\epsilon_{g_{s_k}} \geq \sqrt{\frac{K}{14T}}$ for all rounds $m = 0, 1, 2, ..., \lfloor\frac{1}{2}\log_2\frac{14T}{K}\rfloor$.

$$\delta_{\max(m_i,g_{s_k})} \leq \left(\frac{4K\log(\psi T\epsilon_{m_i}^2)}{T\epsilon_{m_i}(\psi T\epsilon_{m_i}^2)^{\rho_a}}\right) + \left(\frac{4K\log(\psi T\epsilon_{g_{s_k}}^2)}{T\epsilon_{g_{s_k}}(\psi T\epsilon_{g_{s_k}}^2)^{\rho_s}}\right) \leq \left(\frac{4K\log(\psi T)}{T\epsilon_M(\psi T\epsilon_M^2)^{\rho_a}}\right) + \left(\frac{4K\log(\psi T)}{T\epsilon_M(\psi T\epsilon_M^2)^{\rho_s}}\right)$$
$$\leq \left(\frac{4K^{\frac{3}{2}+2\rho_a}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_a}}\right) + \left(\frac{4K^{\frac{3}{2}+2\rho_s}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_s}}\right)$$

Now, applying the bound from Fact 2, we can show that the probability of the sub-optimal arm i being pulled is bounded above by,

$$P_i = \sum_{m=0}^{M}\delta_m.Q_m \leq \sum_{m=0}^{M}\left\{\left(\frac{4K^{\frac{3}{2}+2\rho_a}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_a}}\right) + \left(\frac{4K^{\frac{3}{2}+2\rho_s}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_s}}\right)\right\}2\gamma\exp(-\frac{c_m\sqrt{T}}{4})$$
$$\leq M.\left\{\left(\frac{4K^{\frac{3}{2}+2\rho_a}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_a}}\right) + \left(\frac{4K^{\frac{3}{2}+2\rho_s}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_s}}\right)\right\}2\gamma\exp(-\frac{c_0\sqrt{T}}{2^M.4})$$
$$\leq \log_2\frac{14T}{K}\gamma\exp(\frac{c_0\sqrt{e}}{4})\left\{\left(\frac{4K^{\frac{3}{2}+2\rho_a}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_a}}\right) + \left(\frac{K^{\frac{3}{2}+2\rho_s}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_s}}\right)\right\}, \text{for } M = \lfloor\frac{1}{2}\log_2\frac{7T}{K}\rfloor$$

Hence, the simple regret of EClusUCB is upper bounded by,

$$SR_{EClusUCB} = \sum_{i=1}^{K}\Delta_i.P_i \leq \sum_{i=1}^{K}\Delta_i.\log_2\frac{14T}{K}\gamma\exp(-\frac{c_0\sqrt{e}}{4})\left\{\left(\frac{4K^{\frac{3}{2}+2\rho_a}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_a}}\right) + \left(\frac{4K^{\frac{3}{2}+2\rho_s}\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_s}}\right)\right\}$$
$$\leq 4\log_2\frac{14T}{K}\gamma\sum_{i=1}^{K}\Delta_i\exp(-\frac{c_0\sqrt{e}}{4})\left\{K^{\frac{3}{2}+2\rho_a}\left(\frac{\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_a}}\right) + K^{\frac{3}{2}+2\rho_s}\left(\frac{\log(\psi T)}{T^{\frac{3}{2}}(\psi T^2)^{\rho_s}}\right)\right\}$$

□

Corollary 3. *For $\psi = \frac{T}{196\log(K)}$, $\rho_a = \frac{1}{2}$ and $\rho_s = \frac{1}{2}$, the simple regret of EClusUCB is given by,*

$$SR_{EClusUCB} \leq 8\log_2\frac{14T}{K}K^{\frac{5}{2}}\gamma\sum_{i=1}^{K}\Delta_i\exp(-\frac{c_0\sqrt{e}}{4})\left(\frac{2\sqrt{14\log(K)}\log(\frac{T}{\sqrt{14\log(K)}})}{T^3}\right)$$

Proof. Putting $\psi = \frac{T}{196\log(K)}$, $\rho_a = \frac{1}{2}$ and $\rho_s = \frac{1}{2}$ in the simple regret obtained in Theorem 3, we get

$$SR_{EClusUCB} \leq 8\log_2\frac{7T}{K}K^{\frac{5}{2}}\gamma\sum_{i=1}^{K}\Delta_i\exp(-\frac{c_0\sqrt{e}}{4})\left(\frac{\log(\frac{T^2}{196\log(K)})}{T^{\frac{3}{2}}(\frac{T^3}{196\log(K)})^{\frac{1}{2}}}\right)$$

$$\leq 8\log_2 \frac{14T}{K} K^{\frac{5}{2}} \gamma \sum_{i=1}^{K} \Delta_i \exp(-\frac{c_0\sqrt{e}}{4}) \left(\frac{2\sqrt{14\log(K)}\log(\frac{T}{\sqrt{14\log(K)}})}{T^3} \right)$$

Thus, we see that while the simple regret of CCB decreases at the rate of $O\left(\frac{(\log T)^2}{T^4}\right)$, the simple regret of EClusUCB decreases at the rate of $O\left(\frac{\sqrt{\log K}(\log T)^2}{T^3}\right)$. A table for the simple regret of CCB and EClusUCB is given in Table 3.

□

Table 3: Simple regret upper bounds for different bandit algorithms

Algorithm	Upper bound
CCB	$O\left(\log_2 (\frac{T}{e}) K\gamma \sum_{i=1}^{K} \Delta_i \exp(2 - \frac{c_0\sqrt{e}}{4}) \frac{\log T}{T^4}\right)$
EClusUCB	$O\left(\log_2 \frac{T}{K} K^{\frac{5}{2}} \gamma \sum_{i=1}^{K} \Delta_i \exp(-\frac{c_0\sqrt{e}}{4}) \left(\frac{\sqrt{\log(K)}\log(\frac{T}{\sqrt{\log(K)}})}{T^3} \right)\right)$

G. Adaptive Clustered UCB

In Section 3, we saw that EClusUCB deals with too much early exploration through optimistic greedy sampling. This reduces the cumulative regret, but still one of the principal disadvantages that EClusUCB suffers from is the lack of knowledge of the number of clusters p. One way to handle this is to estimate the number of clusters on the fly. In Algorithm 3, named Adaptive Clustered UCB , hence referred to as AClusUCB, we explore this idea. AClusUCB uses *hierarchical clustering* (see Friedman et al. (2001)) to find the number of clusters present. AClusUCB is similar to EClusUCB with two major differences. The first difference is the call to procedure CreateClusters at every timestep. CreateClusters subroutine first creates a singleton cluster for each of the surviving arms in B_m and then clusters those singleton clusters $s_k, s_d \in S_m$ (say) into one, if any arm $i \in s_k$ and $j \in s_d$ is such that $|\hat{r}_i - \hat{r}_j| \leq \epsilon_m$. We cluster based on ϵ_m because we have no prior knowledge of the gaps and we estimate the gap by ϵ_m. Also, we destroy the clusters after every timestep and reconstruct the clusters based on the condition specified. Since, the environment is stochastic, the initial clusters will have very poor purity (arms with ϵ_m-close expected means lying in a single cluster) whereas in the later rounds the purity becomes better which leads to the optimal arm $*$ lying in a single cluster of its own which will eliminate all the other clusters based on the cluster elimination condition. The second difference is that, we limit the cluster size from start by $\ell_m = 2$ and then double it after every round. Since the environment is stochastic, if we do not limit the cluster size, then it will result in huge chains of clusters in the initial rounds because the initial estimates of $\hat{r}_i, \forall i \in A$ will be poor. This condition helps in stopping such large chains of clusters.

One of the main disadvantages of AClusUCB is that it does not come with a regret upper bound proof. We do not believe that its regret upper bound can be proved in the same way as ClusUCB or EClusUCB. The reason for this is that $a_{\max_{s_k}}$, the true best arm of a cluster is not fixed in AClusUCB as it deconstructs and then reconstructs the cluster at every timestep. This is not an issue with ClusUCB/EClusUCB as they fix the clusters from beginning and hence $a_{\max_{s_k}}$ for each cluster s_k is fixed from the start.

H. More Experiments

The third experiment is conducted over a testbed of $20-100$ (interval of 10) arms with Bernoulli reward distribution, where the expected rewards of the arms are $r_{i_{i\neq *}} = 0.05$ and $r^* = 0.1$. The horizon T is set to a very large value of $10^5 + K^3$ and the number of arms are increased from $K = 20$ to 100. The proposed algorithm EClusUCB is run with same parameters

Algorithm 3 AClusUCB

Input: Time horizon T, exploration parameters ρ_a, ρ_s and ψ.

Initialization: Set $m := 0$, $B_0 := A$, $S_0 = S$, $\epsilon_0 := 1$, $M = \lfloor \frac{1}{2}\log_2 \frac{14T}{K} \rfloor$, $n_0 = \left\lceil \frac{2\log(\psi T\epsilon_0^2)}{\epsilon_0} \right\rceil$, $\ell_0 := 2$ and $N_0 = Kn_0$.

Pull each arm once

for $t = K+1, .., T$ **do**

Pull arm $i \in B_m$ such that $\text{argmax}_{j\in B_m}\left\{\hat{r}_j + \sqrt{\frac{\rho_s \log(\psi T\epsilon_m^2)}{2z_j}}\right\}$, where z_j is the number of times arm j has been pulled

$t := t+1$

Call CreateClusters()

Arm Elimination

For each cluster $s_k \in S_m$, delete arm $i \in s_k$ from B_m if

$$\hat{r}_i + \sqrt{\frac{\rho_a \log(\psi T\epsilon_m^2)}{2z_i}} < \max_{j\in s_k}\left\{\hat{r}_j - \sqrt{\frac{\rho_a \log(\psi T\epsilon_m^2)}{2z_j}}\right\}$$

Cluster Elimination

Delete cluster $s_k \in S_m$ and remove all arms $i \in s_k$ from B_m if

$$\max_{i\in s_k}\left\{\hat{r}_i + \sqrt{\frac{\rho_s \log(\psi T\epsilon_m^2)}{2z_i}}\right\} < \max_{j\in B_m}\left\{\hat{r}_j - \sqrt{\frac{\rho_s \log(\psi T\epsilon_m^2)}{2z_j}}\right\}.$$

if $t \geq N_m$ and $m \leq M$ **then**

$\epsilon_{m+1} := \frac{\epsilon_m}{2}$

$\ell_{m+1} = 2\ell_m$

$B_{m+1} := B_m$

$n_{m+1} := \left\lceil \frac{2\log(\psi T\epsilon_{m+1}^2)}{\epsilon_{m+1}} \right\rceil$

$N_{m+1} := t + |B_{m+1}|n_{m+1}$

$m := m+1$

Stop if $|B_m| = 1$ and pull $i \in B_m$ till T is reached.

end if

end for

procedure CREATECLUSTERS

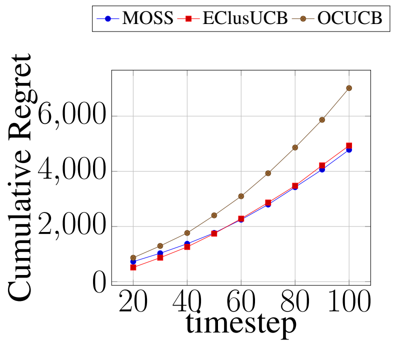
Create singleton cluster $\{i\}$ for each arm $i \in B_m$ and call this partition as S_m.

For two cluster $s_k, s_d \in S_m$, join the clusters if any $|\hat{r}_i - \hat{r_j}| \leq \epsilon_m$ and $|s_k| + |s_d| \leq \ell_m$, where $i \in s_k$ and $j \in s_d$

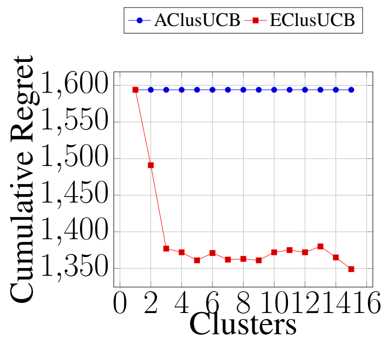
end procedure

as established in Corollary 2. The regret is averaged over 100 independent runs and is shown in Figure 2(a). We report the performance of MOSS, OCUCB and EClusUCB only over this setup. From the results in Figure 2(a), it is evident that the growth of regret for EClusUCB is lower than that of OCUCB and nearly same as MOSS. This corroborates the finding of Lattimore (2015) as well, which says that MOSS breaks down only when the number of arms are exceptionally large or the horizon is unreasonably high and gaps are very small. This experiment also proves that our algorithm EClusUCB is stable for a large horizon and large number of arms.

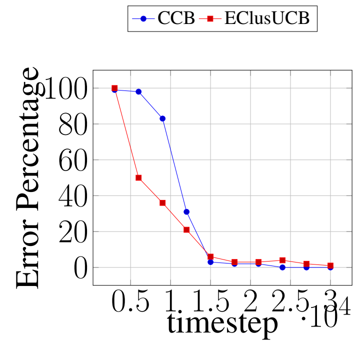
The fourth experiment is conducted to show that our choice of $p = \lceil \frac{K}{\log K} \rceil$ which we use to reduce the elimination error, is indeed close to optimal. The experiment is performed over a testbed having 30 Bernoulli-distributed arms with $r_{i:i\neq *} = 0.07, \forall i \in A$ and $r^* = 0.1$. In Figure 2(b), we report the cumulative regret over $T = 80000$ timesteps averaged over 100 independent runs plotted against the number of clusters $p = 1$ to $\frac{K}{2}$ (so that each cluster can have atleast two



(a) Experiment 3: Experiment 3: 20 to 100 Bernoulli-distributed arms with $r_{i_{i\neq *}} = 0.05$ and $r^* = 0.1$.



(b) Experiment 4: Cumulative regret for ClusUCB for various clusters



(c) Experiment 5: Error Percentage of EClusUCB and CCB

Figure 2: Cumulative regret and Error Percentage for ClusUCB variants

arms). We see that for $p = \lceil \frac{K}{\log K} \rceil = 9$, the cumulative regret of EClusuCB is almost the lowest over the entire range of clusters considered. The lowest is reached for $\frac{K}{2} = 15$, but this would increase the elimination error of EClusUCB in our theoretical analysis. So, the choice of $p = \lceil \frac{K}{\log K} \rceil$ helps to balance both theoretical and empirical performance of EClusUCB. Also, we show cumulative regret for AClusUCB (which does not have p as an input parameter) as a straight line, constant over the number of clusters. AClusUCB performs poorly as compared to EClusUCB. We conjecture that this happens because AClusUCB conducts a significant amount of initial exploration to find the number of clusters or till the optimal arm settles in its own cluster which will eliminate all the other clusters as opposed to EClusUCB which has an uniform clustering scheme from the very start. Also $p = 1$ gives us EClusUCB-AE (EClusUCB with only arm elimination) and we can clearly see that its cumulative regret is the highest among all the clusters considered showing clearly that clustering indeed has some benefits.

The fifth experiment is conducted to analyze the anytime simple regret guarantee of EClusUCB and CCB. The testbed consists of 300 Gaussian Distributed arms with $r_{i:i\neq *} = 0.6, \forall i \in A$, $r^* = 0.9$ and $\sigma_i^2 = 0.5, \forall i \in A$ (similar to the experiment in Liu & Tsuruoka (2016)). Each algorithm is run independently 100 times for 30000 timesteps and the arm with the maximum $\hat{r}_i, \forall i \in A$ as suggested by the algorithms at every timestep is recorded. The output is considered erroneous if the suggested arm is not the optimal arm. The error percentage over 100 runs is plotted against 30000 timesteps and shown in Figure 2(c). The exploration regulatory factor for CCB is chosen as $d_i = \frac{\sqrt{T}}{z_i}$ (where z_i is the number of times an arm i has been sampled) as this was found to perform the best in Liu & Tsuruoka (2016). Here we see that the performance of EClusUCB is slightly poorer than CCB towards the end of the horizon as CCB settles for a lower error percentage than EClusUCB.