UCB with clustering and improved exploration

Anonymous Author(s)

Abstract

In this paper, we present a novel algorithm for the stochastic multi-armed bandit (MAB) problem. Our proposed Clustered UCB method, referred to as ClusUCB partitions the arms into clusters and then follows the UCB-Improved strategy with aggressive exploration factors to eliminate sub-optimal arms, as well as entire clusters. Through a theoretical analysis, we establish that ClusUCB achieves a better gap-dependent regret upper bound than UCB1 (Auer et al., 2002) and UCB-Improved (Auer and Ortner, 2010) and in the worst case matches the gap-dependent bound of MOSS (Audibert and Bubeck, 2009) and OCUCB (Lattimore, 2015) algorithms. ClusUCB also achieves a gap-independent regret bound of $O\left(\sqrt{KT}\right)$ which is better than UCB1 and UCB-Improved, is also comparable to MOSS and OCUCB and is order optimal. Further, numerical experiments on test-cases with small gaps between optimal and sub-optimal mean rewards show that ClusUCB results in lower cumulative regret than several popular UCB variants as well as MOSS, OCUCB, Thompson sampling and Bayes-UCB.

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 timesteps, $N_i(T) = \sum_{m=1}^{T} I(I_m = i)$ is the number of times the algorithm has chosen arm i up to timestep T. The expected regret of an algorithm after T timesteps 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 with the problem of choosing between two treatments to administer on patients who come in sequentially. Following

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

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 and 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., 2002), 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 and 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 $O\big(\frac{K^2\log\left(T\Delta^2/K\right)}{\Delta}\big)$ and in certain regimes, this can be worse than even UCB1 (see (Audibert and Bubeck, 2009; Lattimore, 2015)). The UCB-Improved algorithm, proposed in Auer and 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 $O\big(\sqrt{KT}\big)$. This is the best known bound for the $1-$sub-Gaussian distributions in the bandit literature. Moreover, certain powerful algorithms have also been proposed which we will not discuss in detail here for the sake of brevity. These algorithms, like KL-UCB (Garivier and Cappé, 2011), Bayes-UCB (Kaufmann et al., 2012) and Thompson Sampling (Thompson, 1933; Agrawal and Goyal, 2011) are known to perform well empirically and have strong gap-dependent regret guarantees. However, we show that all the aforementioned algorithms fail to take advantage of certain reward structures that our algorithm, by virtue of its implementation, is able to leverage. This discussion is deferred to the contribution section.

The idea of clustering in the bandit framework is not entirely new. In particular, the idea of clustering has been extensively studied in the contextual bandit setup, an extension of the MAB where side information or features are attached to each arm (see Auer (2002); Langford and Zhang (2008); Li et al. (2010); Beygelzimer et al. (2011); Slivkins (2014)) . The clustering in this case is typically done over the feature space Bui et al. (2012); Cesa-Bianchi et al. (2013); Gentile et al. (2014), however, in our work we cluster or group the arms.

1.1 Our Contribution

We propose a variant of UCB algorithm, called Clustered UCB, henceforth referred to as ClusUCB, that incorporates clustering and an improved exploration scheme. ClusUCB starts with partitioning of arms into small clusters, each having same number of arms. The clustering is done at the start with a prespecified number of clusters. At the end of every round ClusUCB conducts both (individual) arm elimination as well as cluster elimination. This is the first algorithm in bandit literature which uses two simultaneous arm elimination conditions and shows both theoretically and empirically that such an approach is indeed helpful.

The clustering of arms provides two benefits. First, it creates a context where a 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.

We will also show that in certain environments ClusUCB is able to take advantage of the underlying structure of the reward distribution of arms that other algorithms fail to take advantage of. We will briefly discuss two of these examples here.

[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.

1.Bernoulli Distribution with small gaps: In this environment there are 20 arms with means $r_{1:12} = 0.01$, $r_{13:19} = 0.07$ and $r^*_{20} = 0.1$. Here, ClusUCB because of random partitioning of arms into clusters, will create clusters where there are atleast one arm with means 0.07 and a significant number of arms with 0.01 means. These clusters behave like independent UCB-Improved algorithms with improved exploration factors and the arms with means 0.01 are quickly eliminated. Note that since gaps are very small and the gaps of arms with means 0.07 are very close to the optimal arm, comparing all arms to the single best performing arm at every timestep will result is fewer arm eliminations. Hence utilizing the clusters as in ClusUCB results in faster elimination of arms. This is shown in Experiment 1.

2.Gaussian Distribution with different variances: In this environment there are 100 arms with means $r_{1:66} = 0.1, \sigma^2_{1:66} = 0.7, r_{67:99} = 0.8, \sigma^2_{67:99} = 0.1$ and $r^*_{100} = 0.9, \sigma^2_{100} = 0.7$. Here, the variance of the optimal arm and arms with mean farthest from the optimal arm are the highest. Whereas, the arms having mean closest to the optimal arm have lowest variances. In these type of cases, due to clustering ClusUCB is able to eliminate the arms with means 0.7 quickly because clusters containing atleast one arm with 0.8 mean behaves as independent UCB-Improved algorithms with improved exploration factors. This is shown in Experiment 2. Again, note that due to high variance of the optimal arm, comparing only with the best performing arm at every timestep results in fewer arm eliminations.

Theoretically, 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 same as that of MOSS (see Table 1. Moreover, the gap-independent bound of ClusUCB is of the same order as of MOSS and OCUCB, i.e., $O\left(\sqrt{KT}\right)$.

Table 1: Regret upper bound of different algorithms

Algorithm	Gap-Dependent	Gap-Independent
ClusUCB	$O\left(\dfrac{K\log(T\Delta^2/K)}{\Delta}\right)$	$O\left(\sqrt{KT}\right)$
UCB1	$O\left(\dfrac{K\log T}{\Delta}\right)$	$O\left(\sqrt{KT\log T}\right)$
UCB-Imp	$O\left(\dfrac{K\log(T\Delta^2)}{\Delta}\right)$	$O\left(\sqrt{KT\log K}\right)$
MOSS	$O\left(\dfrac{K\log(T\Delta^2/K)}{\Delta}\right)$	$O\left(\sqrt{KT}\right)$
OCUCB	$O\left(\dfrac{K\log(T/H)}{\Delta}\right)$	$O\left(\sqrt{KT}\right)$

On two synthetic setups (as discussed before) with small gaps, we observe empirically that ClusUCB outperforms UCB-ImprovedAuer and Ortner (2010), MOSSAudibert and Bubeck (2009) and OCUCBLattimore (2015) as well as other popular stochastic bandit algorithms such as UCB-VAudibert et al. (2009), Median EliminationEven-Dar et al. (2006), Thompson SamplingAgrawal and Goyal (2011), Bayes-UCBKaufmann et al. (2012) and KL-UCBGarivier and Cappé (2011).

The rest of the paper is organized as follows: In Section 2 we introduce ClusUCB. In Section 4, we present the associated regret bounds. In Section 5, we present the numerical experiments and provide concluding remarks in Section 6. Further proofs of lemmas, corollaries, theorems and propositions presented in Section 4 are provided in the appendices.

2 Algorithm: 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 pre-specified 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

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 $m = 0, 1, ..\lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$ **do**
  Pull each arm in B_m so that the total number of times it has been pulled is $n_m = \left\lceil \frac{\log(\psi T \epsilon_m^2)}{2\epsilon_m} \right\rceil$.
  Arm Elimination
    For each cluster $s_k \in S_m$, delete arm $i \in s_k$ from B_m if

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

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

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

  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

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 (ClusUCB): As mentioned in a recent work Liu and 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 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 ($r_1 = r_2 = .. = r_{K-1} < r^*$) and small.
To reduce early exploration, the number n_m of times 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 now we 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 choice of ψ and ρ_a and ρ_s we can achieve aggressive elimination even when the gaps Δ_i are small and K is large.

In Liu and 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: $\frac{T}{t_i}$, $\frac{\sqrt{T}}{t_i}$ and $\frac{\log T}{t_i}$, where t_i is the number of times an arm i has been sampled. Unlike Liu and Tsuruoka (2016), we employ cluster as well as arm elimination and establish from a theoretical analysis that the choice $\psi = \frac{T}{K^2)}$ helps in achieving a better gap-dependent regret upper bound for ClusUCB as compared to UCB-Improved and MOSS (see Corollary 1 in the next section).

3 Algorithm: Efficient Clustered UCB

Algorithm 2 EClusUCB

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

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

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

Arm Elimination

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

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

Cluster Elimination

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

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

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

Reset Parameters

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

$B_{m+1} := B_m$

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

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

$m := m+1$

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

end if

end for

The algorithm (EClusUCB): One of the principal problems suffered by ClusUCB is that in every round it pulls all the arms equal number of time. EClusUCB remedies this by implementing optimistic greedy sampling, as done for CCB algorithm (see Liu and Tsuruoka (2016). As described in the pseudocode in Algorithm 2, EClusUCB is almost similar to ClusUCB. It starts with an initial uniform clustering of arms and the total number of rounds, $m = 0, 1, 2, \ldots, M$ is also same. Each round of EClusUCB consists a total of $|B_m|n_m$ timesteps and parameters are updated at the end of each round. The exploration parameters are also same for both the algorithms. The first major difference with ClusUCB is that because of optimistic greedy sampling, EClusUCB only pulls the arm that has the highest confidence interval at every timestep. Also EClusUCB conducts both individual arm as well as cluster elimination conditions at every timestep.

4 Main results

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

Theorem 1 (*Gap dependent regret bound*) *For $T \geq K^{2.4}$, $\rho_a = \frac{1}{2}$, $\rho_s = \frac{1}{2}$ and $\psi = \frac{T}{K^2}$ the regret R_T of ClusUCB satisfies*

$$\mathbb{E}[R_T] \leq \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} \left\{ \Delta_i + 12K + \frac{32 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right\} + \sum_{\substack{i \in A, \\ \Delta_i > b}} \left\{ 2\Delta_i + 12K + \frac{64 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right\}$$
$$+ \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} 16K + \sum_{\substack{i \in A_{s^*}, \\ 0 < \Delta_i \leq b}} 16K + \sum_{\substack{i \in A \setminus A_{s^*}: \\ \Delta_i > b}} 32K + \sum_{\substack{i \in A \setminus A_{s^*}: \\ 0 < \Delta_i \leq b}} 32K + \max_{i:\Delta_i \leq b} \Delta_i T,$$

where $b \geq \sqrt{\frac{e}{T}}$, and A_{s^*} is the subset of arms in cluster s^* containing optimal arm a^*.

Proof 1 *The proof of this theorem is given in Appendix C.*

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

We now show the the gap-independent regret bound of ClusUCB in Corollary 1.

Corollary 1 (*Gap-independent bound*) *Considering the same gap of $\Delta_i = \Delta = \sqrt{\frac{K \log K}{T}}$ for all $i : i \neq *$ and with $\psi = \frac{T}{K^2}$, $p = \left\lceil \frac{K}{\log K} \right\rceil$, $\rho_a = \frac{1}{2}$ and $\rho_s = \frac{1}{2}$ and for $T \geq K^{2.4}$, we have the following gap-independent bound for the regret of ClusUCB:*

$$\mathbb{E}[R_T] \leq 96\sqrt{KT} + 12K^2 + 44K \log K + \frac{64K^3}{K + \log K}$$

Proof 2 *The proof of this corollary is given in Appendix D*

Remarks: From the above result, we observe that the order of the regret upper bound of ClusUCB is $O(\sqrt{KT})$, and this matches the order of MOSS and OUCUCB and is order optimal. This bound is also better than UCB1 and UCB-Improved.

Next, we state the regret upper bound for the special case of ClusUCB when $p = 1$, i.e there is a single cluster and there are no cluster elimination condition but only arm elimination condition. We name this algorithm ClusUCB-AE.

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

$$\mathbb{E}[R_T] \leq \sum_{i \in A:\Delta_i > b} \left\{ 12K + \left(\Delta_i + \frac{32 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right) + 16K \right\} + \sum_{i \in A:0 < \Delta_i \leq b} 16K + \max_{i \in A:\Delta_i \leq b} \Delta_i T,$$

for all $b \geq \sqrt{\frac{e}{T}}$.

Proof 3 *The proof of this proposition is given in Appendix E*

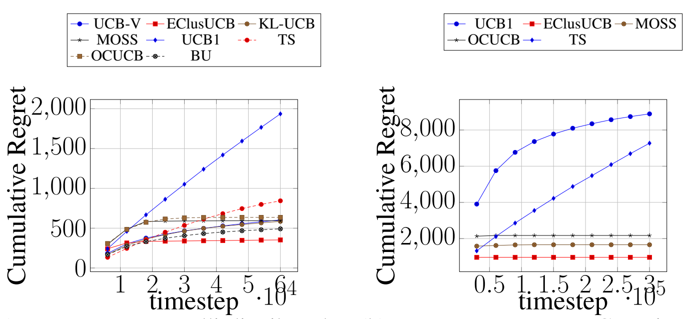
Analysis of elimination error (Why Clustering?)

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 C). As stated before 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. The regret upper bound for ClusUCB-AE is given in Proposition 1.

For ClusUCB-AE, the quantity $\widetilde{R}_T$ can be extracted from the proofs (in particular, case b2 in Appendix E) and simplified to obtain $\widetilde{R}_T = 32K^2$. Finally, for ClusUCB, the relevant terms from Theorem 1 that corresponds to $\widetilde{R}_T$ can be simplified with $\rho_a = \frac{1}{2}$, $\rho_s = \frac{1}{2}, p = \lceil \frac{K}{\log K} \rceil$ and $\psi = \frac{T}{K^2}$ (as in Corollary 1 to obtain $\tilde{R}_T = 32K \log K + \frac{64K^3}{K + \log K}$. Hence, in comparison to ClusUCB-AE which has an elimination regret bound of $O(K^2)$, the elimination error regret bound of ClusUCB is lower and of the order $O\left(\frac{K^3}{K+\log K}\right)$. 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 K^2 for the gap-independent error regret bound for ClusUCB. Also in section 5, in experiment 4 we show that ClusUCB outperforms ClusUCB-AE.

5 Simulation experiments

We conduct an empirical performance using cumulative regret as the metric. We implement the following algorithms: KL-UCBGarivier and Cappé (2011), MOSSAudibert and Bubeck (2009), UCB1Auer et al. (2002), UCB-ImprovedAuer and Ortner (2010), Median EliminationEven-Dar et al. (2006), Thompson Sampling(TS)Agrawal and Goyal (2011), OCUCBLattimore (2015), Bayes-UCB(BU)Kaufmann et al. (2012) and UCB-VAudibert et al. (2009)[2]. The parameters of EClusUCB algorithm for all the experiments are set as follows: $\psi = \frac{T}{K^2}$, $\rho_s = 0.5$, $\rho_a = 0.5$ and $p = \lceil \frac{K}{\log K} \rceil$ (as in Corollary 1).



(a) Expt-1: 20 Bernoulli-distributed arms.

(b) Expt-2: 100 Gaussian-distributed arms.

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

First experiment (Bernoulli with small gaps) : This is conducted over a testbed of 20 arms in an environment involving Bernoulli reward distributions 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, BU 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.

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

Second experiment (Gaussian with different variances): This is conducted over a testbed of 100 arms involving Gaussian reward distributions with expected rewards of the arms $r_{1:33} = 0.7$, $r_{34:99} = 0.8$ and $r^*_{100} = 0.9$ with variance set at $\sigma^2_{1:33} = 0.7, \sigma^2_{34:99} = 0.1$ and $\sigma^2_* = 0.7$. The horizon T is set for a large duration of 3×10^5 and the regret is averaged over 100 independent runs and is shown in Figure 1(b). From the results in Figure 1(b), we observe that EClusUCB outperforms MOSS, UCB1, UCB-Improved and Median-Elimination($\epsilon = 0.1, \delta = 0.1$). 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. Note that the performance of TS is poor and this is in line with the observation in Lattimore (2015) that the worst case regret of TS in Gaussian distributions is $\Omega\left(\sqrt{KT\log T}\right)$.

6 Conclusions and future work

From a theoretical viewpoint, we conclude that the gap-dependent regret bound of ClusUCB is lower than UCB1 and UCB-Improved and its gap-independent regret bound is of the same order as MOSS and OCUCB and is also order optimal. From the numerical experiments in specific environments, we observed that EClusUCB outperforms several popular bandit algorithms, including OCUCB, TS and BU which fail to leverage the structure of the rewards. Also ClusUCB is remarkably stable for a large horizon and large number of arms and performs well across different types of distributions. 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-dependent regret bound of OCUCB. This is also one of the first papers to apply clustering in stochastic MAB and another future direction is to use this in contextual or in distributed bandits.

References

Agrawal, S. and Goyal, N. (2011). Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*.

Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

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

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

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandit algorithms with supervised learning guarantees. In *AISTATS*, pages 19–26.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2012). Bandits with heavy tail. *arXiv preprint arXiv:1209.1727*.

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

Bui, L., Johari, R., and Mannor, S. (2012). Clustered bandits. *arXiv preprint arXiv:1206.4169*.

Cappe, O., Garivier, A., and Kaufmann, E. (2012). pymabandits. `http://mloss.org/software/view/415/`.

Cesa-Bianchi, N., Gentile, C., and Zappella, G. (2013). A gang of bandits. In *Advances in Neural Information Processing Systems*, pages 737–745.

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

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. *arXiv preprint arXiv:1102.2490*.

Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *ICML*, pages 757–765.

Kaufmann, E., Cappé, O., and Garivier, A. (2012). On bayesian upper confidence bounds for bandit problems. In *AISTATS*, pages 592–600.

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

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824.

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

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM.

Liu, Y.-C. and Tsuruoka, Y. (2016). Modification of improved upper confidence bounds for regulating exploration in monte-carlo tree search. *Theoretical Computer Science*.

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

Slivkins, A. (2014). Contextual bandits with similarity information. *Journal of Machine Learning Research*, 15(1):2533–2568.

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

Appendix

The Appendix is organized as follows. First we prove some technical lemmas in Appendix A and Appendix B. Next we prove the main theorem in Appendix C. In Appendix D we prove Corollary 1. In Appendix E we prove Proposition 1.

A Proof of Lemma 1

Lemma 1 *If $T \geq K^{2.4}$, $\psi = \frac{T}{K^2}$, $\rho_a = \frac{1}{2}$ and $m \leq \frac{1}{2}\log_2\left(\frac{T}{e}\right)$, then,*

$$\frac{\rho_a m \log(2)}{\log(\psi T) - 2m\log(2)} \leq \frac{3}{2}$$

Proof 4 *The proof is based on contradiction. Suppose*

$$\frac{\rho_a m \log(2)}{\log(\psi T) - 2m\log(2)} > \frac{3}{2}.$$

Then, with $\psi = \frac{T}{K^2}$ and $\rho_a = \frac{1}{2}$, we obtain

$$\frac{\rho_a m \log(2)}{\log(\frac{T^2}{K^2}) - 2m\log(2)} > \frac{3}{2}$$
$$\Rightarrow 2\rho_a m \log(2) > 6\log(\frac{T}{K}) - 6m\log(2)$$

This can be further reduced to,

$$\begin{aligned} 6\log(K) &> 6\log(T) - 7m\log(2) \\ &\overset{(a)}{\geq} 6\log(T) - \frac{7}{2}\log_2\left(\frac{T}{e}\right)\log(2) \\ &= 2.5\log(T) + 3.5\log_2(e)\log(2) \\ &\overset{(b)}{=} 2.5\log(T) + 3.5 \end{aligned}$$

where (a) is obtained using $m \leq \frac{1}{2}\log_2\left(\frac{T}{e}\right)$, while (b) follows from the identity $\log_2(e)\log(2) = 1$. Finally, for $T \geq K^{2.4}$ we obtain, $6\log(K) > 6\log(K) + 3.5$, which is a contradiction. Hence, for $T \geq K^{2.4}$, $\psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$ and $m \leq \frac{1}{2}\log_2\left(\frac{T}{e}\right)$ we have,

$$\frac{\rho m \log(2)}{\log(\psi T) - 2m\log(2)} \leq \frac{3}{2}$$

B Proof of Lemma 2

Lemma 2 *If $T \geq K^{2.4}$, $\psi = \frac{T}{K^2}$, $\rho_a = \frac{1}{2}$, $m_i = min\{m|\sqrt{2\epsilon_m} < \frac{\Delta_i}{4}\}$ and $c_{m_i} = \sqrt{\frac{\rho_a \log(\psi T \epsilon_{m_i})}{2n_{m_i}}}$, then, $c_{m_i} < \frac{\Delta_i}{4}$.*

Proof 5 *In the m_i-th round $c_{m_i} = \sqrt{\frac{\rho_a \log(\psi T \epsilon_{m_i})}{2n_{m_i}}}$. Substituting the value of $n_{m_i} = \frac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ in c_{m_i} we get,*

$$c_{m_i} \leq \sqrt{\frac{\rho_a \epsilon_{m_i} \log(\psi T \epsilon_{m_i})}{\log(\psi T \epsilon_{m_i}^2)}} \leq \sqrt{\frac{\rho_a \epsilon_{m_i} \log(\frac{\psi T \epsilon_{m_i}^2}{\epsilon_{m_i}})}{\log(\psi T \epsilon_{m_i}^2)}}$$

$$= \sqrt{\frac{\rho_a \epsilon_{m_i} \log(\psi T \epsilon_{m_i}^2) - \rho_a \epsilon_{m_i} \log(\epsilon_{m_i})}{\log(\psi T \epsilon_{m_i}^2)}} \leq \sqrt{\rho_a \epsilon_{m_i} - \frac{\rho_a \epsilon_{m_i} \log(\frac{1}{2^{m_i}})}{\log(\psi T \frac{1}{2^{2m_i}})}}$$
$$\leq \sqrt{\rho_a \epsilon_{m_i} + \frac{\rho_a \epsilon_{m_i} \log(2^{m_i})}{\log(\psi T) - \log(2^{2m_i})}} \leq \sqrt{\rho_a \epsilon_{m_i} + \frac{\rho_a \epsilon_{m_i} m_i \log(2)}{\log(\psi T) - 2m_i \log(2)}}$$
$$\overset{(a)}{\leq} \sqrt{\rho_a \epsilon_{m_i} + \frac{3}{2}\epsilon_{m_i}} < \sqrt{2\epsilon_{m_i}} < \frac{\Delta_i}{4}$$

In the above simplification, (a) is obtained using Lemma 1.

C Proof of Theorem 1

Proof 6 *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 true highest payoff in a cluster s_k is denote by $a_{\max_{s_k}}$. Let for each sub-optimal arm $i \in A$, $m_i = \min\{m | \sqrt{2\epsilon_m} < \frac{\Delta_i}{4}\}$ and let for each cluster $s_k \in S$, $g_{s_k} = \min\{g | \sqrt{2\epsilon_g} < \frac{\Delta_{a_{\max_{s_k}}}}{4}\}$. 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 and 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 $c_{m_i} = \sqrt{\frac{\rho_a \log(\psi T \epsilon_{m_i})}{2n_{m_i}}}$. The arm i gets eliminated because

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

In the above, we have used the fact that $c_{m_i} = \sqrt{\epsilon_{m_i+1}} < \frac{\Delta_i}{4}$, from Lemma 2. 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:

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

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

$$\sum_{i \in A^{'}_{s^*}} \left(\frac{2T\Delta_i}{(\psi T \epsilon_{m_i})^{\rho_a}}\right) \leq \sum_{i \in A^{'}_{s^*}} \left(\frac{2T\Delta_i}{(\psi T \frac{\Delta_i^2}{32})^{\rho_a}}\right) \overset{(a)}{\leq} \sum_{i \in A^{'}_{s^*}} \left(\frac{2T\Delta_i}{(\frac{T^2}{K^2}\frac{\Delta_i^2}{32})^{\frac{1}{2}}}\right) \leq 8\sqrt{2} \sum_{i \in A^{'}_{s^*}} K$$

Here, in (a) we substituted the value ρ_a and ψ.

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,

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

We can prove using Chernoff-Hoeffding bounds and considering independence of events mentioned above, that for $c_{g_{s_k}} = \sqrt{\frac{\rho_s \log(\psi T \epsilon_{g_{s_k}})}{2n_{g_{s_k}}}}$ and $n_{g_{s_k}} = \frac{\log(\psi T \epsilon^2_{g_{s_k}})}{2\epsilon_{g_{s_k}}}$ the probability of the above two events is bounded by $\left(\frac{2}{(\psi T \epsilon_{g_{s_k}})^{\rho_s}}\right)$.

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 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 $T\Delta_i$,

$$\sum_{k=1}^{p} \sum_{i \in A^{'}_{s_k}} \left(\frac{2T\Delta_i}{(\psi T \frac{\Delta_i^2}{16})^{\rho_s}}\right) = \sum_{i \in A'} \left(\frac{2T\Delta_i}{(\psi T \frac{\Delta_i^2}{16})^{\rho_s}}\right)$$

$$\overset{(a)}{\leq} \sum_{i \in A'} \left(\frac{2T\Delta_i}{(\frac{T^2}{K^2}\frac{\Delta_i^2}{32})^{\frac{1}{2}}}\right) = \sum_{i \in A'} \left(8\sqrt{2}K\right)$$

Again we obtain (a) by substituting the value of ρ_s and ψ.

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:

$$\sum_{i \in A^{'}_{s^*}} 8\sqrt{2}K + \sum_{i \in A'} 8\sqrt{2}K \leq \sum_{i \in A^{'}_{s^*}} 12K + \sum_{i \in A'} 12K$$

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 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 $n_{g_{s_k}}$ 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:

$$\sum_{i \in A'} \Delta_i \left\lceil \frac{\log(\psi T \epsilon^2_{m_i})}{2\epsilon_{m_i}} \right\rceil + \sum_{k=1}^{p} \sum_{i \in A^{'}_{s_k}} \Delta_i \left\lceil \frac{\log(\psi T \epsilon^2_{g_{s_k}})}{2\epsilon_{g_{s_k}}} \right\rceil$$

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

$$\overset{b}{\leq} \sum_{i \in A'} \left[2\Delta_i + \frac{16(\log(\frac{T^2}{K^2}\frac{\Delta_i^4}{1024}) + \log(\frac{T^2}{K^2}\frac{\Delta_i^4}{1024}))}{\Delta_i}\right] \leq \sum_{i \in A'} \left[2\Delta_i + \frac{32\left(\log(\frac{T\Delta_i^2}{K}) + \log(\frac{T\Delta_i^2}{K})\right)}{\Delta_i}\right]$$

In the above, the (a) follows since $\sqrt{2\epsilon_{m_i}} < \frac{\Delta_i}{4}$ and $\sqrt{2\epsilon_{n_{g_{s_k}}}} < \frac{\Delta_{a_{\max_{s_k}}}}{4}$ and (b) is obtained by substituting the values of ρ_a, ρ_s and ψ.

Case b2: *$*$ is eliminated by some sub-optimal arm in s^**

Optimal arm $*$ 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, $c_{m_i} = \sqrt{\frac{\rho_a \log(\psi T \epsilon_{m_i})}{2n_{m_i}}}$. 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 $\frac{2}{(\psi T \epsilon_{m_*})^{\rho_a}}$. 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 $m_b = \min\{m | \sqrt{2\epsilon_m} < \frac{\Delta_b}{4}\}$. Hence, the maximal regret after eliminating the arm $*$ is upper bounded by,*

$$
\begin{aligned}
&\sum_{m_*=0}^{max_{j\in A^{'}_{s^*}} m_j} \sum_{\substack{i\in A^{''}_{s^*}:\\ m_i\geq m_*}} \left(\frac{2}{(\psi T\epsilon_{m_*})^{\rho_a}}\right).T \max_{\substack{j\in A^{''}_{s^*}:\\ m_j\geq m_*}} \Delta_j \\
&\leq \sum_{m_*=0}^{max_{j\in A^{'}_{s^*}} m_j} \sum_{i\in A^{''}_{s^*}:m_i\geq m_*} \left(\frac{2}{(\psi T\epsilon_{m_*})^{\rho_a}}\right).T.4\sqrt{2\epsilon_{m_*}} \\
&\leq \sum_{m_*=0}^{max_{j\in A^{'}_{s^*}} m_j} \sum_{i\in A^{''}_{s^*}:m_i\geq m_*} 8\sqrt{2}\left(\frac{T^{1-\rho_a}}{\psi^{\rho_a}\epsilon_{m_*}^{\rho_a-\frac{1}{2}}}\right) \\
&\leq \sum_{i\in A^{''}_{s^*}:m_i\geq m_*} \sum_{m_*=0}^{\min\{m_i,m_b\}} \left(\frac{8\sqrt{2}T^{1-\rho_a}}{\psi^{\rho_a}2^{-(\rho_a-\frac{1}{2})m_*}}\right) \\
&\leq \sum_{i\in A^{'}_{s^*}} \frac{8\sqrt{2}T^{1-\rho_a}}{\psi^{\rho_a}2^{-(\rho_a-\frac{1}{2})m_*}} + \sum_{i\in A^{''}_{s^*}\setminus A^{'}_{s^*}} \frac{8\sqrt{2}T^{1-\rho_a}}{\psi^{\rho_a}2^{-(\rho_a-\frac{1}{2})m_b}} \\
&\leq \sum_{i\in A^{'}_{s^*}} \frac{T^{1-\rho_a}2^{\rho_a+\frac{7}{2}}}{\psi^{\rho_a}\Delta_i^{2\rho_a-1}} + \sum_{i\in A^{''}_{s^*}\setminus A^{'}_{s^*}} \frac{T^{1-\rho_a}2^{\rho_a+\frac{7}{2}}}{\psi^{\rho_a}b^{2\rho_a-1}} \\
&\leq \sum_{i\in A^{'}_{s^*}} 16K + \sum_{i\in A^{''}_{s^*}\setminus A^{'}_{s^*}} 16K
\end{aligned}
$$

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 $\frac{2}{(\psi T\epsilon_{g_{s^*}})^{\rho_s}}$ 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,*

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

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:

$$\sum_{i \in A' \setminus A'_{s*}} \frac{T^{1-\rho_s} 2^{\rho_s + \frac{5}{2}}}{\psi^{\rho_s} \Delta_i^{2\rho_s - 1}} + \sum_{i \in A'' \setminus A' \cup A'_{s*}} \frac{T^{1-\rho_s} 2^{\rho_s + \frac{5}{2}}}{\psi^{\rho_s} b^{2\rho_s - 1}}$$
$$= \sum_{i \in A' \setminus A'_{s*}} 16K + \sum_{i \in A'' \setminus A' \cup A'_{s*}} 16K$$

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 more than $\left\lceil \frac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}} \right\rceil$ number of times, which leads to the following bound the contribution to the expected regret, as in Case $b1$:

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

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,

$$\sum_{i \in A' \setminus A'_{s*}} 16K + \sum_{i \in A'' \setminus A' \cup A'_{s*}} 16K$$

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

D Proof of Corollary 1

Proof 7 *First we recall the definition of Theorem 1 below,*

$$\mathbb{E}[R_T] \le \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} \left\{ \Delta_i + 12K + \frac{32 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right\} + \sum_{\substack{i \in A, \\ \Delta_i > b}} \left\{ 2\Delta_i + 12K + \frac{64 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right\}$$
$$+ \sum_{\substack{i \in A_{s^*}, \\ \Delta_i > b}} 16K + \sum_{\substack{i \in A_{s^*}, \\ 0 < \Delta_i \le b}} 16K + \sum_{\substack{i \in A \setminus A_{s^*}: \\ \Delta_i > b}} 32K + \sum_{\substack{i \in A \setminus A_{s^*}: \\ 0 < \Delta_i \le b}} 32K + \max_{i: \Delta_i \le b} \Delta_i T$$

Now 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 $C = \left\lfloor \frac{T}{e} \right\rfloor$ and by choosing $\Delta_i = \Delta = \sqrt{\frac{K \log K}{T}} > \sqrt{\frac{e}{T}}$ for all $i : i \neq * \in A$ and substituting $p = \left\lceil \frac{K}{\log K} \right\rceil$ in the bound of ClusUCB we get,

$$\sum_{i \in A_{s^*}: \Delta_i > b} 12K = 12 \frac{K^2}{p}$$

Similarly, for the term,

$$\sum_{i \in A: \Delta_i > b} 12K = 12K^2$$

For the term regarding number of pulls,

$$\sum_{i \in A : \Delta_i > b} \frac{64 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \leq \frac{64K\sqrt{T} \log\left(T \dfrac{K \log K}{TK}\right)}{\sqrt{K \log K}} \leq \frac{64\sqrt{KT} \log\left(\log K\right)}{\sqrt{\log K}}$$
$$\stackrel{(a)}{\leq} 64\sqrt{KT}$$

Here (a) is obtained by the identity $\dfrac{\log \log K}{\sqrt{\log K}} < 1$ for $K \geq 2$. Lastly we can bound the error terms as,

$$\sum_{i \in A_{s^*} : 0 \leq \Delta_i \leq b} 16K = \frac{16K^2}{p} \stackrel{<}{(a)} 16K \log K$$

Here we obtain (a) by substituting the value of p. Similarly for the term,

$$\sum_{i \in A \setminus A_{s^*} : \Delta_i > b} 16K = \frac{16K^2}{p} < 16K \log K$$

Also, for all $b \geq \sqrt{\dfrac{e}{T}}$,

$$\sum_{i \in A \setminus A_{s^*} : 0 < \Delta_i \leq b} 32K = \left(K - \frac{K}{p}\right) 32K$$

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 substituting the value of $p = \left\lceil \dfrac{K}{\log K} \right\rceil$, we get,

$$\sum_{i \in A \setminus A_{s^*} : 0 < \Delta_i \leq b} 32K = \left(K - \frac{K}{p}\right) 32K < \frac{32K^3}{K + \log K}$$

Summing up all the contribution from the individual cases as shown above, the total gap-independent regret is given by,

$$\mathbb{E}[R_T] \leq 12K \log K + 32\sqrt{KT} + 12K^2 + 64\sqrt{KT} + 32K \log K + \frac{64K^3}{K + \log K}$$

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

$$\mathbb{E}[R_T] \leq 96\sqrt{KT} + 12K^2 + 44K \log K + \frac{64K^3}{K + \log K}$$

E Proof of Proposition 1

Proof 8 *Let $p = 1$ such that all the arms in A belongs to a single cluster. Hence, in ClusUCB-AE there is only arm elimination and no cluster elimination. Let, for each sub-optimal arm i, $m_i = \min\{m | \sqrt{\epsilon_m} < \dfrac{\Delta_i}{2}\}$. Also $\rho_a = \frac{1}{2}$ 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. Note that $c_{m_i} = \sqrt{\frac{\rho_a \log(\psi T \epsilon_{m_i})}{2n_{m_i}}}$*. A sub-optimal arm* i *will get eliminated in the* m_i*-th round because* $n_{m_i} = \dfrac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ *and substituting this in* c_{m_i} *and applying Lemma 2 we get,* $c_{m_i} < \dfrac{\Delta_i}{4}$*. Hence, for a sub-optimal arm* $i \in A^{'}$*,*

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

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

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

Similarly, $\mathbb{P}\{\hat{r}^* \leq r^* - c_{m_i}\} \leq \dfrac{1}{(\psi T \epsilon_{m_i})^{\rho_a}}$*. Summing the two up, the probability that a sub-optimal arm* i *is not eliminated on or before* m_i*-th round is* $\left(\dfrac{2}{(\psi T \epsilon_{m_i})^{\rho_a}}\right)$*.*

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

$$\sum_{i \in A'} \left(\frac{2T\Delta_i}{(\psi T \epsilon_{m_i})^{\rho_a}}\right) \leq \sum_{i \in A'} \left(\frac{2T\Delta_i}{(\psi T \frac{\Delta_i^2}{32})^{\rho_a}}\right) \leq \sum_{i \in A'} \left(\frac{2^{1+5\rho_a} T^{1-\rho_a} \Delta_i}{\psi^{\rho_a} \Delta_i^{2\rho_a}}\right) \leq \sum_{i \in A'} \left(\frac{2^{1+5\rho_a} T^{1-\rho_a}}{\psi^{\rho_a} \Delta_i^{2\rho_a - 1}}\right)$$
$$\stackrel{(a)}{\leq} \sum_{i \in A'} \leq 8\sqrt{2}K$$

Here, (a) *is obtained by substituting the values of* ψ *and* ρ_a*.*

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,*

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

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

$$\Delta_i \left\lceil \frac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}} \right\rceil \leq \Delta_i \left\lceil \frac{\log(\psi T (\frac{\Delta_i}{4\sqrt{2}})^4)}{(\frac{\Delta_i}{4\sqrt{2}})^2} \right\rceil, \text{ since } \sqrt{2\epsilon_{m_i}} < \frac{\Delta_i}{4}$$
$$\stackrel{(a)}{\leq} \Delta_i \left(1 + \frac{32 \log(\frac{T}{K^2} T (\Delta_i)^4)}{\Delta_i^2}\right) \leq \Delta_i \left(1 + \frac{32 \log(\frac{T\Delta_i^2}{K})}{\Delta_i^2}\right)$$

In the above case, (a) is obtained by substituting the values of ψ and ρ_a. Summing over all arms in $A^{'}$ the total regret is given by,

$$\sum_{i\in A'} \Delta_i \bigg(1 + \frac{32\log\big(\frac{T\Delta_i^2}{K}\big)}{\Delta_i^2}\bigg)$$

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 $m_b = \min\{m|\sqrt{2\epsilon_m} < \dfrac{\Delta_b}{4}\}$. Hence, the maximal regret after eliminating the arm $*$ is upper bounded by,

$$\begin{aligned}
&\sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i\in A'':m_i>m_*} \bigg(\frac{2}{(\psi T\epsilon_{m_*})^{\rho_a}}\bigg).T \max_{j\in A'':m_j\geq m_*} \Delta_j \\
&\leq \sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i\in A'':m_i>m_*} \bigg(\frac{2}{(\psi T\epsilon_{m_*})^{\rho_a}}\bigg).T.4\sqrt{2}\sqrt{\epsilon_{m_*}} \\
&\leq \sum_{m_*=0}^{max_{j\in A'} m_j} \sum_{i\in A'':m_i>m_*} 8\sqrt{2}\bigg(\frac{T^{1-\rho_a}}{\psi^{\rho_a}\epsilon_{m_*}^{\rho_a-\frac{1}{2}}}\bigg) \\
&\leq \sum_{i\in A'':m_i>m_*} \sum_{m_*=0}^{\min\{m_i,m_b\}} \bigg(\frac{8\sqrt{2}T^{1-\rho_a}}{\psi^{\rho_a}2^{-(\rho_a-\frac{1}{2})m_*}}\bigg) \\
&\leq \sum_{i\in A'} \bigg(\frac{8\sqrt{2}T^{1-\rho_a}}{\psi^{\rho_a}2^{-(\rho_a-\frac{1}{2})m_*}}\bigg) + \sum_{i\in A''\setminus A'} \bigg(\frac{8\sqrt{2}T^{1-\rho_a}}{\psi^{\rho_a}2^{-(\rho_a-\frac{1}{2})m_b}}\bigg) \\
&\leq \sum_{i\in A'} \bigg(\frac{4T^{1-\rho_a} * 2^{\rho_a-\frac{1}{2}}}{\psi^{\rho_a}\Delta_i^{8\sqrt{2}\rho_a-1}}\bigg) + \sum_{i\in A''\setminus A'} \bigg(\frac{8\sqrt{2}T^{1-\rho_a} * 2^{\rho_a-\frac{1}{2}}}{\psi^{\rho_a}b^{2\rho_a-1}}\bigg) \\
&\leq \sum_{i\in A'} \bigg(\frac{T^{1-\rho_a}2^{\rho_a+\frac{7}{2}}}{\psi^{\rho_a}\Delta_i^{2\rho_a-1}}\bigg) + \sum_{i\in A''\setminus A'} \bigg(\frac{T^{1-\rho_a}2^{\rho_a+\frac{7}{2}}}{\psi^{\rho_a}b^{2\rho_a-1}}\bigg) \\
&\overset{(a)}{\leq} \sum_{i\in A'} 16K + \sum_{i\in A''\setminus A'} 16K
\end{aligned}$$

*Again (a) is obtained by substituting the values of ψ and ρ_a. Summing up **Case a** and **Case b**, the total regret till round m is given by,*

$$\mathbb{E}[R_T] \leq \sum_{i\in A:\Delta_i>b} \bigg\{12K + \bigg(\Delta_i + \frac{32\log\big(\frac{T\Delta_i^2}{K}\big)}{\Delta_i}\bigg) + 16K\bigg\} + \sum_{i\in A:0<\Delta_i\leq b} 16K + \max_{i\in A:\Delta_i\leq b} \Delta_i T$$