Almost Optimal Algorithm using Variance Estimates

Subhojyoti Mukherjee, Nandan Sudarsanam and Balaraman Ravindran

IIT Madras, Chennai, India,
subho@cse.iitm.ac.in

Abstract. In this paper, we present a novel algorithm for the stochastic multi-armed bandit (MAB) problem. Our proposed Efficient Clustered UCB method, referred to as EClusUCB 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 EClusUCB achieves a better gap-dependent regret upper bound than UCB-Improved Auer and Ortner (2010) and MOSS Audibert and 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), Thompson sampling and Bayes-UCBKaufmann et al. (2012).

Keywords: Multi-armed Bandits, Cumulative Regret, UCBV

1 Introduction

In this paper we deal with the stochastic multi-armed bandit (MAB) setting. Stochastic MABs are classic instances of sequential learning model where at each timestep a learner is exposed to a finite set of actions (or arms) and it has to choose one arm at a time. After choosing (or pulling) an arm the learner sees the reward for the arm as revealed by the environment. Each of these reward is an i.i.d random variable as sampled from the distribution associated with each arm. The mean of the reward distribution associated with an arm i is denoted by r_i whereas the mean of the reward distribution of the optimal arm $*$ is denoted by r^* such that $r_i < r^*, \forall i \in \mathcal{A}$. One of the fundamental assumptions in stochastic MAB is that the distribution associated with each arm does not change over the entire time horizon T. 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, $n_i(T) = \sum_{j=1}^{T} I(I_j = i)$ 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.

1.1 Related Works

There has been a significant amount of research in the area of stochastic MABs. One of the earliest work can be traced to Thompson (1933), which deals with the problem of choosing between two treatments to administer on patients who come in sequentially. Other seminal works include that of Robbins (1952) and then that of Lai and Robbins (1985) which established an asymptotic lower bound for the cumulative regret. It showed that for any consistent allocation strategy, we can have $\liminf_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log T} \geq \sum_{\{i:r_i<r^*\}} \frac{(r^*-r_i)}{D(Q_i\|Q^*)}$, where $D(Q_i\|Q^*)$ is the Kullback-Leibler divergence between the reward densities Q_i and Q^*, corresponding to arms with mean r_i and r^*, respectively.

Over the years stochastic MABs has seen several algorithms with strong regret guarantees. For further reference an interestetd reader can look into Bubeck et al. (2012). The earliest among these algorithms is UCB1 Auer et al. (2002a), which has a gap-dependent regret upper bound of $O\left(\frac{K\log T}{\Delta}\right)$, where $\Delta = \min_{i:\Delta_i>0}\Delta_i$. This result is asymptotically order-optimal for the class of distributions considered. However, the worst case gap-independent regret bound of UCB1 is found to be $O\left(\sqrt{KT\log T}\right)$. In the later work of Audibert and Bubeck (2009), the authors propose the MOSS algorithm and showed that the worst case gap-independent regret bound of MOSS is $O\left(\sqrt{KT}\right)$ which improves upon UCB1 by a factor of order $\sqrt{\log T}$. However, the gap-dependent regret of MOSS is $O\left(\frac{K^2\log\left(T\Delta^2/K\right)}{\Delta}\right)$ 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 $O\left(\frac{K\log T\Delta^2}{\Delta}\right)$, which is better than that of UCB1. On the other hand, the worst case gap-independent regret bound of UCB-Improved is $O\left(\sqrt{KT\log K}\right)$. Recently in Lattimore (2015), the authors showed that 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 $O\left(\sqrt{KT}\right)$. But OCUCB does not take into account the variance of the arms and we show that our algorithm outperforms OCUCB in all the environments considered.

We must also mention about UCBV Audibert et al. (2009) algorithm which is quite different from the above algorithms owing to its utilization of variance estimates. UCBV has a gap-dependent regret bound of $O\left(\frac{K\sigma_{\max}^2\log T}{\Delta}\right)$, where σ_i^2 denotes the variance of the arm i. Its gap-independent regret bound can be inferred to be same as that of UCB1

[1] An algorithm is *round-based* if it pulls all the arms equal number of times in each round and then eliminates one or more arms that it deems to be sub-optimal.

$O\left(\sqrt{KT\log T}\right)$. Empirically, Audibert et al. (2009) showed that UCBV outperforms UCB1 in several scenarios.

1.2 Contribution

In this paper we propose the Efficient UCB Variance (hence referred to as EUCBV) algorithm for the stochastic MAB setting. EUCBV by virtue of it taking into account the empirical variance of the arms performs significantly better than the existing algorithms in the stochastic MAB setting. EUCBV also outperforms UCBV Audibert et al. (2009) which takes into account empirical variance but is less powerful than EUCBV because of the usage of exploration regulatory factor and arm elimination parameter by EUCBV. Theoretically we prove that for $T \geq K^{2.7}$ our algorithm is order optimal and enjoys a worst case gap-independent regret bound of $O\left(\sqrt{KT}\right)$ same as that of MOSS and OCUCB and better than UCBV, UCB1 and UCB-Improved. Also the gap-dependent regret bound of EUCBV is better than UCB1, UCB-Improved and MOSS and is poorer than OCUCB. But EUCBV gap-dependent bound matches OCUCB in the worst case scenario when all the gaps are equal. An illustrative table containing the bounds is provided in Table 1.

Table 1: Regret upper bound of different algorithms

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

Empirically we show that EUCBV owing to its estimating the variance of the arms performs significantly better MOSS, OUCUB and UCB-Improved. Also EUCBV is the first arm-elimination algorithm that takes into account the variance estimates of the arm for minimizing cumulative regret and thereby answers an open question raised by Auer and Ortner (2010). Also it is first algorithm that follows the same proof technique of UCB-Improved and achieves a gap-independent regret bound of $O\left(\sqrt{KT}\right)$ thereby closing the gap of UCB-Improved Auer and Ortner (2010) which achieved a gap-independent regret bound of $O\left(\sqrt{KT\log K}\right)$.

2 Algorithm: Efficient UCB Variance

Algorithm 1 EUCBV

Input: Time horizon T, exploration parameters ρ and ψ.

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

Pull each arm once

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

Pull arm $i \in \arg\max_{j \in B_m} \left\{ \hat{r}_j + \sqrt{\frac{\rho \hat{v}_i \log(\psi T \epsilon_m^2)}{4 z_j} + \frac{\log(\psi T \epsilon_m)}{4 z_j}} \right\}$, where z_j is the number of times arm j has been pulled

$t := t + 1$

Arm Elimination

For each arm $i \in B_m$, remove arm i from B_m if,

$$\hat{r}_i + \sqrt{\frac{\rho \hat{v}_i \log(\psi T \epsilon_m)}{4 n_m} + \frac{\log(\psi T \epsilon_m)}{4 n_m}} < \max_{j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho \hat{v}_i \log(\psi T \epsilon_m)}{4 n_m} + \frac{\log(\psi T \epsilon_m)}{4 n_m}} \right\}$$

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

Reset Parameters

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

$B_{m+1} := B_m$

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

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

$m := m + 1$

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

end if

end for

2.1 Notations: We denote the set of arms by $\mathcal{A}$, with the individual arms labeled $i, i = 1, \ldots, K$. We denote an arbitrary round of EUCBV by m. For simplicity, we assume that the optimal arm is unique and denote it by $*$. We denote the sample mean of the rewards for an arm i at time instant t by $\hat{r}_i(t) = \frac{1}{n_i(t)} \sum_{z=1}^{n_i(t)} X_{i,z}$, where $X_{i,z}$ is the reward sample received when arm i is pulled for the z-th time. $z_i(t)$ is the number of times an arm i has been pulled till timestep t. We denote the variance of an arm by σ_i^2 while $\hat{v}_i(t)$ is the estimated variance, i.e., $\hat{v}_i(t) = \frac{1}{n_i(t)} \sum_{z=1}^{n_i(t)} (X_{i,z} - \hat{r}_i)^2$. Whenever there is no ambiguity about the underlaying time index t, for simplicity we neglect t from the notations and simply use $\hat{r}_i$, $\hat{v}_i$, and n_i to denote the respective quantities. We assume the rewards of all arms are bounded in $[0, 1]$.

2.2 The algorithm: The authors in a recent work Liu and Tsuruoka (2016) identified two major problems with UCB-Improved:

(i) Early exploration: Since UCB-Improved pulls each arm equal number of times in each round, a significant number of pulls is wasted in initial explorations. For any round m UCB-Improved pulls all arms $n_m = \left\lceil \frac{2\log(T\epsilon_m)}{\epsilon_m} \right\rceil$ number of times. The quantity ϵ_m is initialized to 1 and halved after every round.

(ii) Conservative arm-elimination: 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. In the worst case scenario when K is large and the gaps are identical ($r_1 = r_2 = \cdots = r_{K-1} < r^*$) and small this is very disadvantageous.

To reduce early exploration, the number of pulls n_m allocated to each arm per round in EUCBV is lower than that of UCB-Improved and also that of Median-Elimination, which used $n_m = \frac{4}{\epsilon}\log\left(\frac{3}{\delta}\right)$, where ϵ, δ are confidence parameters. As described in the pseudocode in Algorithm 1, each timestep of EUCBV involves both individual arm as well as cluster elimination conditions. In EUCBV we also introduce the idea of optimistic greedy sampling similar to Liu and Tsuruoka (2016) which they used to modify the UCB-Improved algorithm. In optimistic greedy sampling, we only sample the arm with the highest upper confidence bound in each timestep. EUCBV 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. The exploration regulatory factor ψ governing the arm and cluster elimination conditions in EUCBV is more aggressive than that in UCB-Improved. With appropriate choices of ψ and ρ, we can achieve aggressive elimination even when the gaps Δ_i are small and K is large.

In Liu and Tsuruoka (2016), the authors also recommend incorporating an exploration regulatory factor d_i inside the log-term of the UCB values, i.e., $\max\{\hat{r}_i + \sqrt{\frac{d_i \log T\epsilon_m}{2n_m}}\}$. The authors choices for d_i include $\frac{T}{z_i}$, $\frac{\sqrt{T}}{z_i}$ and $\frac{\log T}{z_i}$, where z_i is the number of times an arm i has been pulled. Unlike Liu and Tsuruoka (2016), we employ a different exploration regulatory factor $\psi = \dfrac{T}{K^2}$ and we derive a cumulative regret bound as opposed to the simple regret bound derived in Liu and Tsuruoka (2016).

3 Main Results

3.1 Lemma 1

Lemma 1. *If* $T \geq 81K^{2.7}$, $\psi = \dfrac{T}{81K^2}$, $\rho = \dfrac{1}{2}$ *and* $m \leq \dfrac{1}{2}\log_2(\dfrac{T}{e})$, *then,*

$$\frac{\rho m \log(2)}{\log(\psi T) - 2m\log(2)} \leq 1$$

Proof. We are going to prove this by contradiction. Let's say,

$$\begin{aligned}
&\frac{\rho m \log(2)}{\log(\psi T) - 2m\log(2)} \geq 1 \\
&\Rightarrow \rho m \log(2) \geq \log(\psi T) - 2m\log(2) \\
&\Rightarrow \rho m \log(2) \geq 2\log(\frac{T}{9K}) - 2m\log(2)\text{ , as } \psi = \frac{T}{81K^2}
\end{aligned}$$

$$\Rightarrow 2.5m\log(2)+2\log(9K)\geq 2\log(T)\text{ , as }\rho=\frac{1}{2}$$
$$\Rightarrow 1.25\log(2)\log_2(\frac{T}{e})+2\log(9K)\geq 2\log(T)\text{ , as }m\leq\frac{1}{2}\log_2(\frac{T}{e})$$
$$\Rightarrow \frac{1.25\log(2)\log(\frac{T}{e})}{\log(2)}+2\log(9K)\geq 2\log(T)$$
$$\Rightarrow 1.25\log(T)+2\log 9K-1.25\geq 2\log(T)$$
$$\Rightarrow 2\log K\geq 0.75\log T+1.25-2\log 9$$

But, for $T \geq 81K^{2.7}$, this is clearly not possible. Hence, $\dfrac{m\log(2)}{\log(\psi T)-2m\log(2)}\leq 1$.

3.2 Theorem 1

Theorem 1. *The regret R_T for EUCBV 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 $b\geq\sqrt{\frac{e}{T}}$. In the above, $C_1(x)=\frac{2^{1+x}.9^x}{\psi^x}$ and $C_2(x)=\frac{2^{\frac{\rho}{2}+\frac{5}{4}}.3^{x+\frac{1}{2}}}{\psi^x}$.

Proof. Let, for each sub-optimal arm i, $m_i=\min\{m|\sqrt{\rho}<\dfrac{\Delta_i}{3}\}$. Also $\rho\in(0,1]$ is a constant in this proof. Let $A^{'}=\{i\in A:\Delta_i>b\}$ and $A^{''}=\{i\in A:\Delta_i>0\}$. Also z_i denotes total number of times an arm i has been pulled. In the m-th round, n_m denotes the number of pulls allocated to the surviving arms in B_m.

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 **??** Case $a1$, an arbitrary sub-optimal arm $i\in A^{'}$ can get eliminated only when the event,

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

takes place. So to bound the regret we need to bound the probability of the complementary event of these two conditions. Note that $c_{m_i}=\sqrt{\frac{\rho\log(\psi T\epsilon_{m_i})}{2n_{m_i}}}$. As arm elimination condition is being checked in every timestep, any arm i cannot be pulled more than $z_i=n_{m_i}$ times or it will get eliminated. This is because in the m_i-th round $n_{m_i}=\dfrac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ and putting this in c_{m_i} we get,

$$c_{m_i}=\sqrt{\frac{\rho\epsilon_{m_i}\log(\psi T\epsilon_{m_i})}{\log(\psi T\epsilon_{m_i}^2)}}=\sqrt{\frac{\rho\epsilon_{m_i}\log(\frac{\psi T\epsilon_{m_i}^2}{\epsilon_{m_i}})}{\log(\psi T\epsilon_{m_i}^2)}}=\sqrt{\frac{\rho\epsilon_{m_i}\log(\psi T\epsilon_{m_i}^2)-\rho\epsilon_{m_i}\log(\epsilon_{m_i})}{\log(\psi T\epsilon_{m_i}^2)}}$$

$$\le \sqrt{\rho\epsilon_{m_i} - \frac{\rho\epsilon_{m_i}\log(\frac{1}{2^{m_i}})}{\log(\psi T\frac{1}{2^{2m_i}})}} \le \sqrt{\rho\epsilon_{m_i} + \frac{\rho\epsilon_{m_i}\log(2^{m_i})}{\log(\psi T) - \log(2^{2m_i})}}$$
$$\le \sqrt{\rho\epsilon_{m_i} + \frac{\rho\epsilon_{m_i} m_i \log(2)}{\log(\psi T) - 2m_i\log(2)}}$$
$$\le \sqrt{\rho\epsilon_{m_i} + \epsilon_{m_i}}, \text{ applying Lemma 1}$$
$$< \sqrt{2\epsilon_{m_i}}, \text{ as } \rho \in (0, \frac{1}{2}]$$
$$< \frac{\Delta_i}{3}$$

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

$$\begin{aligned}\hat{r}_i + c_{m_i} &\le r_i + 2c_{m_i} = \hat{r}_i + 3c_{m_i} - 2c_{m_i}\\ &< r_i + \Delta_i - 2c_{m_i} \le r^* - 2c_{m_i} \le \hat{r}^* - c_{m_i}\end{aligned}$$

Applying Chernoff-Hoeffding bound and considering independence of complementary of the two events in 1,

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

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

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

$$\begin{aligned}&\sum_{i\in A^{'}}\left(\frac{2T\Delta_i}{(\psi T\epsilon_{m_i})^\rho}\right) \le \sum_{i\in A^{'}}\left(\frac{2T\Delta_i}{(\psi T\frac{\Delta_i^2}{2.9})^\rho}\right) \le \sum_{i\in A^{'}}\left(\frac{2^{1+\rho}.9^\rho}{\psi^\rho\Delta_i^{2\rho-1}}\right)\\ &= \sum_{i\in A^{'}}\left(\frac{C_1(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}}\right), \text{ where } C_1(x) = \frac{2^{1+x}.9^x}{\psi^x}\end{aligned}$$

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

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,

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

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

$$\Delta_i \left\lceil \frac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}} \right\rceil \leq \Delta_i \left\lceil \frac{\log(\psi T (\frac{\Delta_i}{81})^4)}{2(\frac{\Delta_i}{81})} \right\rceil, \text{ since } \sqrt{\epsilon_{m_i}} < \frac{\Delta_i}{3}$$

$$\leq \Delta_i \left(1 + \frac{41 \log(\psi T (\frac{\Delta_i^4}{162})}{\Delta_i^2} \right) \approx \Delta_i \left(1 + \frac{41 \log(\psi T \Delta_i^4)}{\Delta_i^2} \right)$$

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

$$\sum_{i \in A'} \Delta_i \left(1 + \frac{41 \log(\psi T \Delta_i^4)}{\Delta_i^2} \right) = \sum_{i \in A'} \left(\Delta_i + \frac{41 \log(\psi T \Delta_i^4)}{\Delta_i} \right)$$

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 $m_b = \min\{m | \sqrt{2\epsilon_m} < \frac{\Delta_b}{3}\}$. Hence, the maximal regret after eliminating the arm $*$ is upper bounded by,

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

$$\leq \sum_{m_*=0}^{max_{j \in A'} m_j} \sum_{i \in A'': m_i > m_*} \left(\frac{2\sqrt{2}}{(\psi T \epsilon_{m_*})^\rho} \right) .T.3\sqrt{\epsilon_{m_*}}$$

$$\leq \sum_{m_*=0}^{max_{j \in A'} m_j} \sum_{i \in A'': m_i > m_*} 6\sqrt{2} \left(\frac{T^{1-\rho}}{\psi^\rho \epsilon_{m_*}^{\rho - \frac{1}{2}}} \right)$$

$$\leq \sum_{i \in A'': m_i > m_*} \sum_{m_*=0}^{\min\{m_i, m_b\}} \left(\frac{6\sqrt{2} T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_*}} \right)$$

$$\leq \sum_{i \in A'} \left(\frac{6\sqrt{2} T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_*}} \right) + \sum_{i \in A'' \setminus A'} \left(\frac{6\sqrt{2} T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_b}} \right)$$

$$\leq \sum_{i \in A'} \left(\frac{6\sqrt{2} T^{1-\rho} * 2^{\frac{\rho}{2} - \frac{1}{4}} * 3^{\rho - \frac{1}{2}}}{\psi^\rho \Delta_i^{\rho - \frac{1}{2}}} \right) + \sum_{i \in A'' \setminus A'} \left(\frac{6\sqrt{2} T^{1-\rho_a} * 3^{\rho - \frac{1}{2}}}{\psi^\rho b^{\rho - \frac{1}{2}}} \right)$$

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

$$= \sum_{i\in A'} \left(\frac{C_2(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) + \sum_{i\in A''\setminus A'} \left(\frac{C_2(\rho)T^{1-\rho}}{b^{2\rho-1}} \right), \text{ where } C_2(x) = \frac{2^{\frac{\rho}{2}+\frac{5}{4}}.3^{x+\frac{1}{2}}}{\psi^x}$$

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

$$\mathbb{E}[R_T] \leq \sum_{i\in A:\Delta_i>b} \left\{ \left(\frac{C_1(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) + \left(\Delta_i + \frac{41\log(\psi T\Delta_i^4)}{\Delta_i} \right) + \left(\frac{C_2(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) \right\}$$
$$+ \sum_{i\in A:0<\Delta_i\leq b} \left(\frac{C_2(\rho)T^{1-\rho}}{b^{2\rho-1}} \right) + \max_{i\in A:0<\Delta_i\leq b} \Delta_i T$$

3.3 Corollary 1

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

$$\mathbb{E}[R_T] \sum_{i\in A:\Delta_i>b} \left\{ 17K + \frac{82\log(\dfrac{T\Delta_i^2}{K})}{\Delta_i} \right\} + \max_{i\in A:\Delta_i\leq b} \Delta_i T$$

Proof. Here we take $\psi = \dfrac{T}{(K)^2}$ and $\rho = \dfrac{1}{2}$. Taking into account Theorem **??** below for all $b \geq \sqrt{\dfrac{e}{T}}$

$$\mathbb{E}[R_T] \leq \sum_{i\in A:\Delta_i>b} \left\{ \left(\frac{C_1(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) + \left(\Delta_i + \frac{41\log(\psi T\Delta_i^4)}{\Delta_i} \right) + \left(\frac{C_2(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) \right\}$$
$$+ \sum_{i\in A:0<\Delta_i\leq b} \left(\frac{C_2(\rho)T^{1-\rho}}{b^{2\rho-1}} \right) + \max_{i\in A:0<\Delta_i\leq b} \Delta_i T$$

and putting the parameter values in the above Theorem **??** result,

$$\sum_{i\in A:\Delta_i>b} \left(\frac{T^{1-\rho}}{\psi^{\rho}\Delta_i^{2\rho-1}} \right) = \sum_{i\in A:\Delta_i>b} \left(\frac{T^{1-\frac{1}{2}}2^{1+\frac{1}{2}}.9^{\frac{1}{2}}}{(\frac{T}{(K)^2})^{\frac{1}{2}}\Delta_i^{2.\frac{1}{2}-1}} \right) = \sum_{i\in A:\Delta_i>b} 8.5K$$

For the term involving arm pulls,

$$\sum_{i\in A:\Delta_i>b} \frac{41\log(\psi T\Delta_i^4)}{\Delta_i} = \sum_{i\in A:\Delta_i>b} \frac{82\log(\dfrac{T\Delta_i^2}{K})}{\Delta_i}$$

Lastly we can bound the error terms as,

$$\sum_{i \in A_{s^*}:0<\Delta_i \leq b} \left(\frac{T^{1-\rho} 2^{\frac{\rho}{2}+\frac{5}{4}} .3^{\rho+\frac{1}{2}}}{\psi^{\rho} \Delta_i^{2\rho-1}} \right) = \sum_{i \in A_{s^*}:0<\Delta_i \leq b} 8.5K$$

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

$$\sum_{i \in A:\Delta_i > b} \left\{ 17K + \frac{82 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right\} + \max_{i \in A:\Delta_i \leq b} \Delta_i T$$

3.4 Corollary 2

Corollary 2 (***Gap-dependent bound***)**.** *With* $\psi = \frac{T}{K^2}$ *and* $\rho = \frac{1}{2}$*, we have the following gap-independent bound for the regret of EUCBV:*

$$\mathbb{E}[R_T] \leq 26K^2 + 82\sqrt{KT}$$

Proof. As stated in Auer and Ortner (2010), we can have a bound on regret of the order of $\sqrt{KT \log K}$ in non-stochastic MAB setting. This is shown in Exp4Auer 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 $C = \left\lfloor \frac{T}{e} \right\rfloor$ and similarly, by choosing $\Delta_i = \Delta = \sqrt{\frac{K \log K}{T}} > \sqrt{\frac{e}{T}}$ for all $i : i \neq * \in A$, in the bound of UCB1Auer 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 and Bubeck (2009) for MOSS which is consistent with the lower bound as proposed by Mannor and TsitsiklisMannor and Tsitsiklis (2004).

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

Taking into account Theorem **??** below for all $b \geq \sqrt{\frac{e}{T}}$,

$$\mathbb{E}[R_T] \leq \sum_{i \in A:\Delta_i > b} \left\{ \left(\frac{C_1(\rho) T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) + \left(\Delta_i + \frac{41 \log(\psi T \Delta_i^4)}{\Delta_i} \right) + \left(\frac{C_2(\rho) T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) \right\}$$

$$+ \sum_{i \in A: 0 < \Delta_i \leq b} \left(\frac{C_2(\rho) T^{1-\rho}}{b^{2\rho-1}} \right) + \max_{i \in A: 0 < \Delta_i \leq b} \Delta_i T$$

and putting the parameter values in the above Theorem **??** result,

$$\sum_{i \in A: \Delta_i > b} \left(\frac{C_1(\rho) T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) = \sum_{i \in A: \Delta_i > b} \left(\frac{T^{1-\frac{1}{2}} 2^{1+\frac{1}{2}} .9^{\frac{1}{2}}}{(\frac{T}{(K)^2})^{\frac{1}{2}} \Delta_i^{2.\frac{1}{2}-1}} \right) \leq 8.5K^2$$

For the term regarding number of pulls,

$$\sum_{i \in A: \Delta_i > b} \frac{41 \log(\psi T \Delta_i^4)}{\Delta_i} = \frac{41K\sqrt{T} \log(T^2 \frac{K^2 (\log K)^2}{T^2 K^2})}{\sqrt{K \log K}} \leq \frac{82\sqrt{KT} \log(\log K)}{\sqrt{\log K}}$$

$$\leq 82\sqrt{KT} \text{ , as } \frac{\log(\log K)}{\sqrt{\log K}} \leq 1 \text{ for } K \geq 3$$

Lastly we can bound the error terms as,

$$\sum_{i \in A: 0 \leq \Delta_i \leq b} \left(\frac{T^{1-\rho} 2^{\frac{\rho}{2}+\frac{5}{4}}}{\psi^{\rho} \Delta_i^{2\rho-1}} \right) = K \left(\frac{T^{1-\frac{1}{2}} 2^{\frac{1}{4}+\frac{5}{4}}}{(\frac{T}{K^2})^{\frac{1}{2}} (\Delta_i)^{2*\frac{1}{2}-1}} \right) < 8.5K^2$$

Similarly for the term,

$$\left(\frac{C_2(\rho) T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) \leq \sum_{i \in A: \Delta_i > b} \left(\frac{T^{1-\rho} 2^{\frac{\rho}{2}+\frac{5}{4}}}{(\psi^{\rho}) \Delta_i^{2\rho-1}} \right) < 8.5K^2$$

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

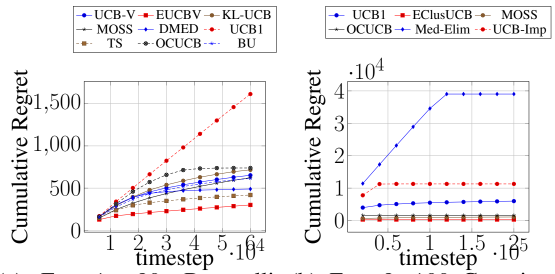
$$\mathbb{E}[R_T] \leq 26K^2 + 82\sqrt{KT}$$

4 Experimental Section

We conduct an empirical performance using cumulative regret as the metric. We implement the following algorithms: KL-UCBGarivier and Cappé (2011), DMEDHonda and Takemura (2010), MOSSAudibert and Bubeck (2009), UCB1Auer et al. (2002a), 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: $\psi = \frac{T}{196 \log K}$, $\rho_s = 0.5$, $\rho_a = 0.5$ and $p = \lceil \frac{K}{\log K} \rceil$ (as in Corollary **??**).

First experiment: This is conducted over a testbed of 20 arms in an environment involving Bernoulli reward distributions with expected rewards of the arms $r_{i_{i \neq *}} =$

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



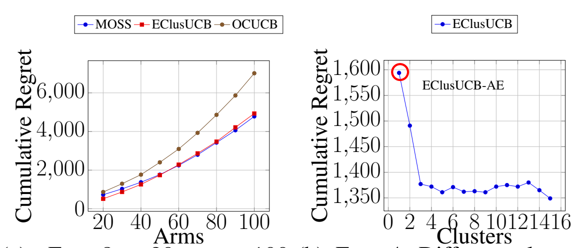
(a) Expt-1: 20 Bernoulli-distributed arms.

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

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

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.

Second experiment: This is conducted over a testbed of 100 arms involving Gaussian reward distributions 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). 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.



(a) Expt-3: 20 to 100 Bernoulli-distributed arms.

(b) Expt-4: Different cluster experiment

Fig. 2: Cumulative regret and choice of parameter p

Third experiment: This is conducted over a testbed of $20 - 100$ (interval of 10) arms with Bernoulli reward distributions, where the expected rewards of the arms are $r_{i_{i \neq *}} = 0.05$ and $r^* = 0.1$. For each of these testbeds of $20 - 100$ arms, we report the cumulative regret over a large horizon $T = 10^5 + K^3_{20:100}$ timesteps averaged over 100 independent runs. We report the performance of MOSS, OCUCB and EClusUCB only over this uniform gap setup. Algorithms like Thompson Sampling or Bayes-UCB are too slow to be run for such large K (see Lattimore (2015)). 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) which states that MOSS breaks down only when the number of arms are exceptionally large or the horizon is unreasonably high and gaps are very small. We consistently see that in uniform gap testcases EClusUCB outperforms OCUCB.

Fourth experiment: This 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 $r^* = 0.1$ averaged over 100 independent runs for each cluster. 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 have exactly two 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 $p = 1$ gives us EClusUCB-AE 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. Its poor performance stems from the fact that it eliminates optimal arm in many of the runs as opposed to EClusUCB. More experiments are shown in Appendix **??**.

5 Conclusion and Future Works

Bibliography

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 trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

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.

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.

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

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*.

Honda, J. and Takemura, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer.

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.

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

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

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

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

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.