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 Algorithm: Efficient UCB Variance

2.1 Notations: We denote the set of arms by A, with the individual arms labeled $i, i = 1, \ldots, K$. We denote an arbitrary round of EClusUCB 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 true 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 the rewards of all arms are bounded in $[0, 1]$.

2.2 The algorithm. 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 $n_m = \left\lceil \frac{2\log(T\epsilon_m)}{\epsilon_m} \right\rceil$ 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 pulls n_m allocated to each arm per round in EClusUCB is lower than that of UCB-Improved and also that of Median-Elimination,

Algorithm 1 EUCBV

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

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

Pull each arm once

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

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

$t := t + 1$

Arm Elimination

For each arm $i \in B_m$ if

$$\hat{r}_i + \sqrt{\frac{\rho \log (\psi T \epsilon_m^2)}{4 n_m} + \frac{\log (\psi T \epsilon_m^2)}{4 n_m}} < \max_{j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho \log (\psi T \epsilon_m^2)}{4 n_m} + \frac{\log (\psi T \epsilon_m^2)}{4 n_m}} \right\}$$

$|B_m| := |B_m| - 1$

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

Reset Parameters

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

$B_{m+1} := B_m$

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

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

$m := m + 1$

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

end if

end for

which used $n_m = \frac{4}{\epsilon} \log \left(\frac{3}{\delta} \right)$, where ϵ, δ are confidence parameters. To handle the second problem mentioned above, EClusUCB 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 **??**, EClusUCB begins with an 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 timestep of EClusUCB 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. In EClusUCB 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. We further

modify the idea by introducing clustering and arm elimination parameters. 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. The exploration regulatory factor ψ governing the arm and cluster elimination conditions in EClusUCB 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. Also we use different exploration regulatory factor than Liu and Tsuruoka (2016) and we come up with a cumulative regret bound whereas Liu and Tsuruoka (2016) only gives simple regret bound for the CCB algorithm.

In Liu and Tsuruoka (2016), the authors recommend incorporating a factor of d_i inside the log-term of the UCB values, i.e., $\max\{\hat{r}_i + \sqrt{\frac{d_i \log T\epsilon_m}{2n_m}}\}$. 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 and 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 EClusUCB as compared to UCB-Improved and MOSS (Corollary **??**).

2 Main Results

2.1 Proof of 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* $\dfrac{\rho m \log(2)}{\log(\psi T) - 2m\log(2)} \leq 1$.

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

$$
\begin{aligned}
&\frac{\rho m \log(2)}{\log(\psi T) - 2m\log(2)} \geq 1 \\
&\Rightarrow \rho m \log(2) \geq \log(\psi T) - 2m\log(2) \\
&\Rightarrow \rho m \log(2) \geq 2\log(\frac{T}{9K}) - 2m\log(2) \text{ , as } \psi = \frac{T}{81K^2} \\
&\Rightarrow 2.5m\log(2) + 2\log(9K) \geq 2\log(T) \text{ , as } \rho = \frac{1}{2} \\
&\Rightarrow 1.25\log(2)\log_2(\frac{T}{e}) + 2\log(9K) \geq 2\log(T) \text{ , as } m \leq \frac{1}{2}\log_2(\frac{T}{e}) \\
&\Rightarrow \frac{1.25\log(2)\log(\frac{T}{e})}{\log(2)} + 2\log(9K) \geq 2\log(T) \\
&\Rightarrow 1.25\log(T) + 2\log 9K - 1.25 \geq 2\log(T) \\
&\Rightarrow 2\log K \geq 0.75\log T + 1.25 - 2\log 9
\end{aligned}
$$

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

2.2 Proof of Theorem 1

Theorem 1. *The regret R_T for EClusUCB-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 $b \geq \sqrt{\frac{e}{T}}$. In the above, C_1, C_2 are as defined in Theorem **??**.

Proof. Let, for each sub-optimal arm i, $m_i = \min\{m|\sqrt{\rho} < \frac{\Delta_i}{3}\}$. 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 α: *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 $n_{m_i} = \frac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ and putting this in c_{m_i} we get,

$$\begin{aligned}
c_{m_i} = \sqrt{\frac{\rho\epsilon_{m_i}\log(\psi T\epsilon_{m_i})}{\log(\psi T\epsilon_{m_i}^2)}} = \sqrt{\frac{\rho\epsilon_{m_i}\log(\frac{\psi T\epsilon_{m_i}^2}{\epsilon_{m_i}})}{\log(\psi T\epsilon_{m_i}^2)}} &= \sqrt{\frac{\rho\epsilon_{m_i}\log(\psi T\epsilon_{m_i}^2) - \rho\epsilon_{m_i}\log(\epsilon_{m_i})}{\log(\psi T\epsilon_{m_i}^2)}} \\
&\leq \sqrt{\rho\epsilon_{m_i} - \frac{\rho\epsilon_{m_i}\log(\frac{1}{2^{m_i}})}{\log(\psi T\frac{1}{2^{2m_i}})}} \leq \sqrt{\rho\epsilon_{m_i} + \frac{\rho\epsilon_{m_i}\log(2^{m_i})}{\log(\psi T) - \log(2^{2m_i})}} \\
&\leq \sqrt{\rho\epsilon_{m_i} + \frac{\rho\epsilon_{m_i} m_i\log(2)}{\log(\psi T) - 2m_i\log(2)}} \\
&\leq \sqrt{\rho\epsilon_{m_i} + \epsilon_{m_i}}, \text{ applying Lemma 1} \\
&< \sqrt{2\epsilon_{m_i}}, \text{ as } \rho \in (0, \frac{1}{2}] \\
&< \frac{\Delta_i}{3}
\end{aligned}$$

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

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

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

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

Similarly, $\mathbb{P}\{\hat{r}^* \leq r^* - c_{m_i}\} \leq \frac{1}{(\psi T \epsilon_{m_i})^\rho}$. 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

$$\sum_{i \in A'} \left(\frac{2T\Delta_i}{(\psi T \epsilon_{m_i})^\rho}\right) \leq \sum_{i \in A'} \left(\frac{2T\Delta_i}{(\psi T \frac{\Delta_i^2}{2.9})^\rho}\right) \leq \sum_{i \in A'} \left(\frac{2^{1+\rho}.9^\rho}{\psi^\rho \Delta_i^{2\rho-1}}\right)$$

$$= \sum_{i \in A'} \left(\frac{C_1(\rho) T^{1-\rho}}{\Delta_i^{2\rho-1}}\right), \text{where } C_1(x) = \frac{2^{1+x}.9^x}{\psi^x}$$

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 *b*1: $* \in B_{m_i}$ *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 *b*2: *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,

$$
\begin{aligned}
&\sum_{m_*=0}^{max_{j\in A^{'}} m_j} \sum_{i\in A^{''}:m_i>m_*} \left(\frac{2}{(\psi T\epsilon_{m_*})^\rho}\right).T \max_{j\in A^{''}:m_j\geq m_*} \Delta_j \\
&\leq \sum_{m_*=0}^{max_{j\in A^{'}} m_j} \sum_{i\in A^{''}:m_i>m_*} \left(\frac{2\sqrt{2}}{(\psi T\epsilon_{m_*})^\rho}\right).T.3\sqrt{\epsilon_{m_*}} \\
&\leq \sum_{m_*=0}^{max_{j\in A^{'}} m_j} \sum_{i\in A^{''}:m_i>m_*} 6\sqrt{2}\left(\frac{T^{1-\rho}}{\psi^\rho \epsilon_{m_*}^{\rho-\frac{1}{2}}}\right) \\
&\leq \sum_{i\in A^{''}:m_i>m_*} \sum_{m_*=0}^{\min\{m_i,m_b\}} \left(\frac{6\sqrt{2}T^{1-\rho}}{\psi^\rho 2^{-(\rho-\frac{1}{2})m_*}}\right) \\
&\leq \sum_{i\in A^{'}} \left(\frac{6\sqrt{2}T^{1-\rho}}{\psi^\rho 2^{-(\rho-\frac{1}{2})m_*}}\right) + \sum_{i\in A^{''}\setminus A^{'}} \left(\frac{6\sqrt{2}T^{1-\rho}}{\psi^\rho 2^{-(\rho-\frac{1}{2})m_b}}\right) \\
&\leq \sum_{i\in A^{'}} \left(\frac{6\sqrt{2}T^{1-\rho} * 2^{\frac{\rho}{2}-\frac{1}{4}} * 3^{\rho-\frac{1}{2}}}{\psi^\rho \Delta_i^{\rho-\frac{1}{2}}}\right) + \sum_{i\in A^{''}\setminus A^{'}} \left(\frac{6\sqrt{2}T^{1-\rho_a} * 3^{\rho-\frac{1}{2}}}{\psi^\rho b^{\rho-\frac{1}{2}}}\right) \\
&\leq \sum_{i\in A^{'}} \left(\frac{2^{\frac{\rho}{2}+\frac{3}{4}+\frac{1}{2}}.3^{\rho_a+\frac{1}{2}}.T^{1-\rho}}{\psi^\rho \Delta_i^{2\rho-1}}\right) + \sum_{i\in A^{''}\setminus A^{'}} \left(\frac{2^{\frac{\rho}{2}+\frac{3}{4}+\frac{1}{2}}.3^{2\rho+\frac{1}{2}}.T^{1-\rho}}{\psi^\rho b^{2\rho_a-1}}\right) \\
&= \sum_{i\in A^{'}} \left(\frac{C_2(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}}\right) + \sum_{i\in A^{''}\setminus A^{'}} \left(\frac{C_2(\rho)T^{1-\rho}}{b^{2\rho-1}}\right), \text{where } C_2(x) = \frac{2^{\frac{\rho}{2}+\frac{5}{4}}.3^{x+\frac{1}{2}}}{\psi^x}
\end{aligned}
$$

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

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

2.3 Proof of Corollary 1

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,

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

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

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

2.4 Proof of Corollary 2

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 Exp4**?** algorithm. Also we know from **?** that the function $x \in [0,1] \mapsto x \exp(-Cx^2)$ is decreasing on $\left[\dfrac{1}{\sqrt{2C}}, 1 \right]$ for any $C > 0$. So, taking $C = \left\lfloor \dfrac{T}{e} \right\rfloor$ and similarly, by choosing $\Delta_i = \Delta = \sqrt{\dfrac{K \log K}{T}} > \sqrt{\dfrac{e}{T}}$ for all $i : i \neq * \in A$, in the bound of UCB1**?** 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 Tsitsiklis**?**.

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,

$$\sum_{i\in A:\Delta_i>b}\frac{41\log(\psi T\Delta_i^4)}{\Delta_i} = \frac{41K\sqrt{T}\log(T^2\frac{K^2(\log K)^2}{T^2K^2})}{\sqrt{K\log K}} \leq \frac{82\sqrt{KT}\log(\log K)}{\sqrt{\log K}}$$
$$\leq 82\sqrt{KT}\text{ , as }\frac{\log(\log K)}{\sqrt{\log K}} \leq 1 \text{ for } K \geq 3$$

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,

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

Bibliography

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

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.

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

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*.