Efficient UCBV: An Almost Optimal Algorithm using Variance Estimates

Subhojyoti Mukherjee, KP Naveen, 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 UCB Variance method, referred to as EUCBV is an arm elimination algorithm based on UCB-Improved and UCBV strategy which takes into account the empirical variance of the arms and along with aggressive exploration factors eliminate sub-optimal arms. Through a theoretical analysis, we establish that EUCBV achieves a better gap-dependent regret upper bound than UCB-Improved Auer and Ortner (2010), MOSS Audibert and Bubeck (2009), UCB1 Auer et al. (2002a) and UCBV Audibert et al. (2009) algorithms. EUCBV enjoys an order optimal gap-independent regret bound same as that of OCUCB Lattimore (2015) and MOSS and better than UCB-Improved, UCB1 and UCBV. Further, numerical experiments on test-cases with small gaps between optimal and sub-optimal mean rewards show that EUCBV owing to its utilization of the variance estimates of the arms results in lower cumulative regret than several popular UCB variants like MOSS, OCUCB, Thompson sampling and Bayes-UCB Kaufmann et al. (2012).

Keywords: Multi-armed Bandits, Cumulative Regret, UCBV, UCB-Improved

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 rewards 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 $r_i < r^*, \forall i \in \mathcal{A}$. So at every timestep the learner faces the task of balancing exploitation and exploration, that is whether to pull the arm for which it has seen the best estimates or to explore more arms in the hope of finding better performing arms. 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:

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

where T is the number of timesteps, $z_i(T) = \sum_{j=1}^{T} I(I_j = i)$ 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,

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

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

In recent years the MAB setting has garnered extensive popularity because of its simple learning model and its practical applications in a wide-range of industry defined problems. From the area of mobile channel allocations to active learning or computer simulation games, MABs have become an increasingly useful tool which has been implemented successfully in a wide range of applications.

1.1 Related Works

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). In this section we will exclusively focus on upper confidence bound algorithms that balance the exploration-exploitation dilemma by carefully tracking the uncertainty in estimates. One of 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. But, 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

[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.

has a gap-dependent regret bound of $O\left(\frac{K\sigma^2_{\max}\log T}{\Delta}\right)$, where $\sigma^2_{\max}$ denotes the maximum variance among all the arms $i \in \mathcal{A}$. Its gap-independent regret bound can be inferred to be same as that of UCB1 i.e $O\left(\sqrt{KT\log T}\right)$. 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 combines the approach of UCB-Improved, CCB Liu and Tsuruoka (2016) and UCBV algorithms. EUCBV by virtue of taking into account the empirical variance of the arms performs significantly better than the existing algorithms in the stochastic MAB setting. EUCBV outperforms UCBV Audibert et al. (2009) which also 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. Through our theoretical analysis we establish the exact values of the exploration parameters for the best performance of EUCBV. Our proof technique is highly generic and can be easily extended to other MAB settings. An illustrative table containing the bounds is provided in Table 1.

Table 1: Regret upper bound of different algorithms

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

Empirically we show that EUCBV owing to its estimating the variance of the arms performs significantly better than MOSS, OUCUB, UCB-Improved, UCB1, UCBV, Thompson Sampling, Bayes-UCB, DMED, KL-UCB and Median Elimination algorithms. Please note that except UCBV all the afore-mentioned algorithms does not take

into account the empirical variance estimates of the arms. 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 $O\left(\sqrt{KT\log K}\right)$.

The rest of the paper is organized as follows. In section 2 we state the main algorithm EUCBV and in the next section 3 we state all the main results of the paper. In section 4 we establish the proofs of all the Lemma, Theorem and Corollaries and section 5 contains the numerical experiments. We conclude in section 6 and discuss about future works.

2 Algorithm: Efficient UCB Variance

Algorithm 1 EUCBV

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

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

Pull each arm once

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

Pull arm $i \in \arg\max_{j\in B_m}\left\{\hat{r}_j + \sqrt{\frac{\rho\hat{v}_j\log(\psi T\epsilon_m^2)}{4z_j} + \frac{\rho\log(\psi T\epsilon_m)}{4z_j}}\right\}$, 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,

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

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

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 $\hat{r}_i(t) = \frac{1}{z_i(t)} \sum_{\ell=1}^{z_i(t)} X_{i,\ell}$, where $X_{i,\ell}$ 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., $\hat{v}_i(t) = \frac{1}{z_i(t)} \sum_{\ell=1}^{z_i(t)} (X_{i,\ell} - \hat{r}_i)^2$. 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 z_i to denote the respective quantities. We assume the rewards of all arms are bounded in $[0, 1]$.

2.2 The algorithm: Earlier arm elimination algorithms like Median Elimination Even-Dar et al. (2006) and UCB-Improved Auer and Ortner (2010) mainly suffered from two basic problems:

(i) *Initial exploration:* Both of these algorithms pull each arm equal number of times in each round, and hence waste a significant number of pulls in initial explorations.

(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 uniform ($r_1 = r_2 = \cdots = r_{K-1} < r^*$) and small this results in very high regret.

UCBEV algorithm which is mainly based on the arm elimination technique of the UCB-Improved algorithm remedies these by employing exploration regulatory factor ψ and arm elimination parameter ρ for aggressive elimination of sub-optimal arms. Along with these, like CCB Liu and Tsuruoka (2016) algorithm, EUCBV uses optimistic greedy sampling whereby at every timestep it only pulls the arm with the highest upper confidence bound rather than pulling all the arms equal number of times in each round. Also, unlike the UCB-Improved, UCB1. MOSS and OCUCB algorithm (which are based on mean estimation) EUCBV employs variance estimates (as in Audibert et al. (2009)) for arm elimination. Further, we allow for arm-elimination at every time-step, which is in contrast to the earlier work (e.g., Auer and Ortner (2010); Even-Dar et al. (2006)) where the arm elimination takes place only at the end of the respective exploration rounds.

3 Main Results

3.1 Lemma 1

A technical lemma used to prove Theorem 1 is presented below.

Lemma 1. *If* $T \geq K^{2.7}$, $\psi = \dfrac{T}{K^2}$, $\rho = \dfrac{1}{2}$ *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. The proof is given in Section 4.1.

We present below the main theorem of the paper which establishes the regret upper bound for the EUCBV algorithm.

3.2 Main Theorem

Theorem 1. *For $T \geq K^{2.7}$, the regret R_T for EUCBV satisfies*

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

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

Proof. The proof is given in Section 4.2.

Next, we specialize the result of Theorem 1 in Corollary 1 and Corollary 2.

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:*

$$\mathbb{E}[R_T] \sum_{i \in \mathcal{A}: \Delta_i > b} \left\{ 34K + \frac{82 \log(\frac{T\Delta_i^2}{K})}{\Delta_i} \right\} + \max_{i \in \mathcal{A}: \Delta_i \leq b} \Delta_i T$$

Proof. The proof is given in Section 4.3.

Thus, we clearly see that the most significant term in the gap-dependent bound is $\frac{82K \log(T\Delta^2/K)}{\Delta}$ and it is better than UCB1, UCBV, MOSS and UCB-Improved. In Audibert and Bubeck (2010) the authors define the term $H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$ as the hardness of a problem and in Bubeck and Cesa-Bianchi (2012) the authors conjectured that the gap-dependent regret upper bound can match the quantity of $O\left(\frac{K \log(T/H_1)}{\Delta_i}\right)$. But Lattimore (2015) proved that the gap-dependent regret bound cannot be lower than $O\left(\sum_{i=2}^{K} \frac{\log(T/H_i)}{\Delta_i}\right)$, where $H_i = \sum_{j=1}^{K} \min\{\frac{1}{\Delta_i^2}, \frac{1}{\Delta_j^2}\}$ and only in the worst case scenario, when all the gaps are equal then $H_1 = H_i = \sum_{i=1}^{K} \frac{1}{\Delta^2}$. In such a scenario the EUCBV gap-dependent bound of $O\left(\frac{K \log(T\Delta^2/K)}{\Delta}\right)$ reduces to $O\left(\frac{K \log(T/H_1)}{\Delta}\right)$ and hence matches the gap-dependent bound of OCUCB.

3.4 Corollary 2

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

$$\mathbb{E}[R_T] \leq 51K^2 + 82\sqrt{KT}$$

Proof. The proof is given in Section 4.4.

Here, in the gap-independent bound of EUCBV the most significant term is $O\left(\sqrt{KT}\right)$ which exactly matches the upper bound of MOSS and OCUCB and is better than UCB-Improved, UCB1 and UCBV.

4 Proofs

4.1 Proof of Lemma 1

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

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

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

4.2 Proof of Theorem 1

Proof. Let, for each sub-optimal arm i, $m_i = \min\{m | \sqrt{2\epsilon_{m_i}} < \frac{\Delta_i}{3}\}$. Also $\rho \in (0, 1]$ is a constant in this proof. Let $\mathcal{A}^{'} = \{i \in \mathcal{A} : \Delta_i > b\}$ and $\mathcal{A}^{''} = \{i \in \mathcal{A} : \Delta_i > 0\}$. 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 $* \in B_{m_i}$* and $c_i \leq c^*$

An arbitrary sub-optimal arm $i \in \mathcal{A}^{'}$ can get eliminated only when the event,

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

takes place. So to bound the regret we need to bound the probability of the complementary event of these two conditions. We denote $c_i = \sqrt{\frac{\rho(\hat{v}_i+1)\log(\psi T\epsilon_{m_i})}{4z_i}}$ for the i-th arm. Note that as arm elimination condition is being checked in every timestep, in the m_i-th round whenever $z_i \geq n_{m_i} = \dfrac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ we have,

$$
\begin{aligned}
c_i &\leq \sqrt{\frac{\rho(\hat{v}_i+1)\epsilon_{m_i}\log(\psi T\epsilon_{m_i})}{2\log(\psi T\epsilon_{m_i}^2)}} \overset{(a)}{\leq} \sqrt{\frac{\rho\epsilon_{m_i}\log(\frac{\psi T\epsilon_{m_i}^2}{\epsilon_{m_i}})}{\log(\psi T\epsilon_{m_i}^2)}} \\
&= \sqrt{\frac{\rho\epsilon_{m_i}\log(\psi T\epsilon_{m_i}^2) - \rho\epsilon_{m_i}\log(\epsilon_{m_i})}{\log(\psi T\epsilon_{m_i}^2)}} \leq \sqrt{\rho\epsilon_{m_i} - \frac{\rho\epsilon_{m_i}\log(\frac{1}{2^{m_i}})}{\log(\psi T\frac{1}{2^{2m_i}})}} \\
&\leq \sqrt{\rho\epsilon_{m_i} + \frac{\rho\epsilon_{m_i}\log(2^{m_i})}{\log(\psi T) - \log(2^{2m_i})}} \leq \sqrt{\rho\epsilon_{m_i} + \frac{\rho\epsilon_{m_i}m_i\log(2)}{\log(\psi T) - 2m_i\log(2)}} \\
&\overset{(b)}{\leq} \sqrt{\rho\epsilon_{m_i} + \epsilon_{m_i}} < \sqrt{2\epsilon_{m_i}} < \frac{\Delta_i}{3}
\end{aligned}
$$

In the above (a) happens because $\hat{v}_i \in [0,1]$, $\rho \in (0, \frac{1}{2}]$ and (b) occurs by applying the result from Lemma 1. Again, in the m_i-th round a sub-optimal arm $i \in \mathcal{A}^{'}$ gets eliminated as,

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

Thus, the probability that a bad arm is not eliminated correctly in the m_i-th round (or before) is given by ,

$$
\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right) \tag{2}
$$

where

$$
\bar{c}_i = \sqrt{\frac{\rho(\sigma_i^2 + \sqrt{\epsilon_{m_i}} + 1)\log(\psi T\epsilon_{m_i})}{4z_i}}
$$

Note that, substituting $z_i = n_{m_i} \geq \frac{\log(\psi T\epsilon_{m_i})}{2\epsilon_{m_i}}$, $\bar{c}_i$ can be simplified to obtain,

$$
\bar{c}_i \leq \sqrt{\frac{\rho\epsilon_{m_i}(\sigma_i^2 + \sqrt{\epsilon_{m_i}} + 1)}{2}} \leq \sqrt{\epsilon_{m_i}}. \tag{3}
$$

The first term in the LHS of (2) can be bounded using the Chernoff-Hoeffding bound as below:

$$
\begin{aligned}
&\mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) \leq \exp\left(-(\bar{c}_i)^2 z_i\right) \\
&\leq \exp\left(-\rho(\sigma_i^2 + \sqrt{\epsilon_{m_i}} + 1)\log(\psi T\epsilon_{m_i})\right)
\end{aligned}
$$

$$\overset{(a)}{\leq} \exp\left(-\rho \log(\psi T \epsilon_{m_i})\right) = \frac{1}{(\psi T \epsilon_{m_i})^{\rho}} \tag{4}$$

where, (a) occurs because $(\sigma_i^2 + \sqrt{\rho \epsilon_{m_i}} + 1) \geq 1$.

The second term in the LHS of (2) can be simplified as follows:

$$\begin{aligned}
&\mathbb{P}\left\{ \hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}} \right\} \\
&\leq \mathbb{P}\left\{ \frac{1}{n_i} \sum_{t=1}^{n_i} (X_{i,t} - r_i)^2 - (\hat{r}_i - r_i)^2 \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}} \right\} \\
&\leq \mathbb{P}\left\{ \frac{\sum_{t=1}^{n_i} (X_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}} \right\} \\
&\overset{(a)}{\leq} \mathbb{P}\left\{ \frac{\sum_{t=1}^{n_i} (X_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \bar{c}_i \right\} \\
&\overset{(b)}{\leq} \exp\left(-\rho(\sigma_i^2 + \sqrt{\epsilon_{m_i}} + 1) \log(\psi T \epsilon_{m_i})\right) = \frac{1}{(\psi T \epsilon_{m_i})^{\rho}}
\end{aligned} \tag{5}$$

where inequality (a) is obtained using (3), while (b) follows from the Chernoff-Hoeffding bound.

Thus, using (4) and (5) in (2) we obtain $\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \frac{2}{(\psi T \epsilon_{m_i})^{\rho}}$. Similarly, $\mathbb{P}\{\hat{r}^* \leq r^* - c^*\} \leq \frac{2}{(\psi T \epsilon_{m_i})^{\rho}}$. Summing the two up, the probability that a sub-optimal arm i is not eliminated on or before m_i-th round is $\left(\frac{4}{(\psi T \epsilon_{m_i})^{\rho}} \right)$.

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

$$\begin{aligned}
&\sum_{i \in \mathcal{A}'} \left(\frac{4T\Delta_i}{(\psi T \epsilon_{m_i})^{\rho}} \right) \leq \sum_{i \in \mathcal{A}'} \left(\frac{4T\Delta_i}{(\psi T \frac{\Delta_i^2}{2.9})^{\rho}} \right) \leq \sum_{i \in \mathcal{A}'} \left(\frac{2^{2+\rho}.9^{\rho}}{\psi^{\rho} \Delta_i^{2\rho-1}} \right) \\
&= \sum_{i \in \mathcal{A}'} \left(\frac{C_1(\rho) T^{1-\rho}}{\Delta_i^{2\rho-1}} \right), \text{where } C_1(x) = \frac{2^{2+x}.9^{x}}{\psi^{x}}
\end{aligned}$$

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 *b*1: *$* \in B_{m_i}$ and each $i \in \mathcal{A}^{'}$ 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,

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

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 $\mathcal{A}^{'}$ the total regret is given by,

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

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 $\mathcal{A}^{'}$. 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 $\mathcal{A}^{''}$. 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,

$$\sum_{m_*=0}^{max_{j \in \mathcal{A}'} m_j} \sum_{i \in \mathcal{A}'' : m_i > m_*} \left(\frac{4}{(\psi T \epsilon_{m_*})^\rho}\right) . T \max_{j \in \mathcal{A}'' : m_j \geq m_*} \Delta_j$$

$$\leq \sum_{m_*=0}^{max_{j \in \mathcal{A}'} m_j} \sum_{i \in \mathcal{A}'' : m_i > m_*} \left(\frac{4\sqrt{2}}{(\psi T \epsilon_{m_*})^\rho}\right) . T . 3\sqrt{\epsilon_{m_*}}$$

$$\leq \sum_{m_*=0}^{max_{j \in \mathcal{A}'} m_j} \sum_{i \in \mathcal{A}'' : m_i > m_*} 12\sqrt{2} \left(\frac{T^{1-\rho}}{\psi^\rho \epsilon_{m_*}^{\rho - \frac{1}{2}}}\right)$$

$$\leq \sum_{i \in \mathcal{A}'' : m_i > m_*} \sum_{m_*=0}^{\min\{m_i, m_b\}} \left(\frac{12\sqrt{2} T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_*}}\right)$$

$$\leq \sum_{i \in \mathcal{A}'} \left(\frac{12\sqrt{2} T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_*}}\right) + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} \left(\frac{12\sqrt{2} T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_b}}\right)$$

$$\leq \sum_{i \in \mathcal{A}'} \left(\frac{12\sqrt{2} T^{1-\rho} * 2^{\frac{\rho}{2} - \frac{1}{4}} * 3^{\rho - \frac{1}{2}}}{\psi^\rho \Delta_i^{\rho - \frac{1}{2}}}\right) + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} \left(\frac{12\sqrt{2} T^{1-\rho_a} * 3^{\rho - \frac{1}{2}}}{\psi^\rho b^{\rho - \frac{1}{2}}}\right)$$

$$\leq \sum_{i \in \mathcal{A}'} \left(\frac{2^{\frac{\rho}{2}+\frac{7}{4}+\frac{1}{2}}.3^{\rho_a+\frac{1}{2}}.T^{1-\rho}}{\psi^{\rho}\Delta_i^{2\rho-1}} \right) + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} \left(\frac{2^{\frac{\rho}{2}+\frac{7}{4}+\frac{1}{2}}.3^{2\rho+\frac{1}{2}}.T^{1-\rho}}{\psi^{\rho} b^{2\rho_a-1}} \right)$$

$$= \sum_{i \in \mathcal{A}'} \left(\frac{C_2(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}} \right) + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} \left(\frac{C_2(\rho)T^{1-\rho}}{b^{2\rho-1}} \right), \text{ where } C_2(x) = \frac{2^{\frac{\rho}{2}+\frac{9}{4}}.3^{x+\frac{1}{2}}}{\psi^x}$$

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

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

4.3 Proof of Corollary 1

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

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

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

$$\sum_{i \in \mathcal{A}:\Delta_i > b} \left(\frac{T^{1-\rho}}{\psi^{\rho}\Delta_i^{2\rho-1}} \right) = \sum_{i \in \mathcal{A}:\Delta_i > b} \left(\frac{T^{1-\frac{1}{2}}2^{2+\frac{1}{2}}.9^{\frac{1}{2}}}{(\frac{T}{(K)^2})^{\frac{1}{2}}\Delta_i^{2.\frac{1}{2}-1}} \right) = \sum_{i \in \mathcal{A}:\Delta_i > b} 17K$$

For the term involving arm pulls,

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

Lastly we can bound the error terms as,

$$\sum_{i \in \mathcal{A}:0<\Delta_i \leq b} \left(\frac{T^{1-\rho}2^{\frac{\rho}{2}+\frac{9}{4}}.3^{\rho+\frac{1}{2}}}{\psi^{\rho}\Delta_i^{2\rho-1}} \right) = \sum_{i \in \mathcal{A}:0<\Delta_i \leq b} 17K$$

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

$$\sum_{i \in \mathcal{A}:\Delta_i > b} \left\{ 34K + \frac{82 \log(\frac{T\Delta_i^2}{K})}{\Delta_i} \right\} + \max_{i \in \mathcal{A}:\Delta_i \leq b} \Delta_i T$$

4.4 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 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 $i : i \neq * \in \mathcal{A}$, 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 Tsitsiklis Mannor 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 1 below for all $b \geq \sqrt{\frac{e}{T}}$,

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

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

$$\sum_{i\in\mathcal{A}:\Delta_i>b}\left(\frac{C_1(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}}\right) = \sum_{i\in\mathcal{A}:\Delta_i>b}\left(\frac{T^{1-\frac{1}{2}}2^{2+\frac{1}{2}}.9^{\frac{1}{2}}}{(\frac{T}{(K)^2})^{\frac{1}{2}}\Delta_i^{2.\frac{1}{2}-1}}\right) \leq 17K^2$$

For the term regarding number of pulls,

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

Lastly we can bound the error terms as,

$$\sum_{i\in\mathcal{A}:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho}2^{\frac{\rho}{2}+\frac{9}{4}}}{\psi^{\rho}\Delta_i^{2\rho-1}}\right)=K\left(\frac{T^{1-\frac{1}{2}}2^{\frac{1}{4}+\frac{9}{4}}}{(\frac{T}{K^2})^{\frac{1}{2}}(\Delta_i)^{2*\frac{1}{2}-1}}\right)<17K^2$$

Similarly for the term,

$$\left(\frac{C_2(\rho)T^{1-\rho}}{\Delta_i^{2\rho-1}}\right)\leq\sum_{i\in\mathcal{A}:\Delta_i>b}\left(\frac{T^{1-\rho}2^{\frac{\rho}{2}+\frac{9}{4}}}{(\psi^{\rho})\Delta_i^{2\rho-1}}\right)<17K^2$$

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

$$\mathbb{E}[R_T]\leq 51K^2+82\sqrt{KT}$$

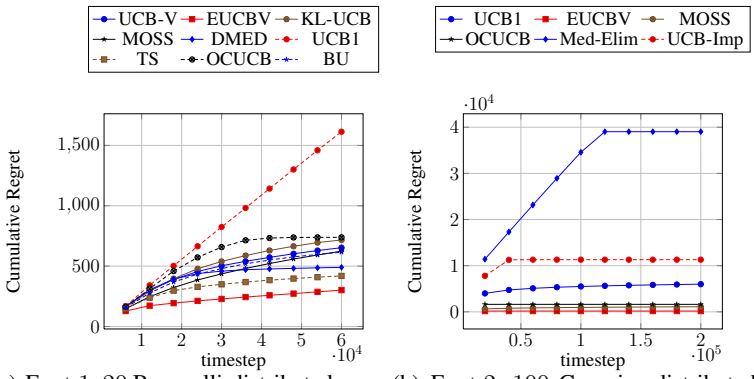
5 Experimental Section

In this section we conduct an extensive empirical evaluation of EUCBV against several other popular bandit algorithms. We use cumulative regret as the metric of comparison. 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 EUCBV algorithm for all the experiments are set as follows: $\psi=\frac{T}{K^2}$ and $\rho=0.5$ (as in Corollary 2).

First experiment: This experiment is conducted to observe the performance of EUCBV over a short horizon. The horizon T is set to 60000. These type of cases are frequently encountered in web-advertising domain. The testbed comprises of 20 Bernoulli distributed arms with expected rewards of the arms as $r_{i_{i\neq *}}=0.07$ and $r^*=0.1$. The regret is averaged over 100 independent runs and is shown in Figure 1(a). EUCBV, MOSS, UCB1, UCB-V, KL-UCB, TS, BU and DMED are run in this experimental setup. Here not only we observe that EUCBV performs better than all the non-variance based algorithms like MOSS, OCUCB, UCB-Improved and UCB1 but it also outperforms UCBV because of the choice of the exploration parameters. 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 experiment is conducted on a large horizon and over a large set of arms. The horizon T is set for a large duration of 2×10^5. This testbed comprises of 100 arms involving Gaussian reward distributions with expected rewards of the arms $r_{i_{i\neq *:1-33}}=0.1$, $r_{i_{i\neq *:34-99}}=0.6$, $r^*_{i=100}=0.9$ and variance set at $\sigma_i^2=0.3,\forall i\in\mathcal{A}$. 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 EUCBV outperforms all the non-variance based algorithms MOSS, OCUCB, UCB1, UCB-Improved and Median-Elimination($\epsilon=0.1,\delta=0.1$).

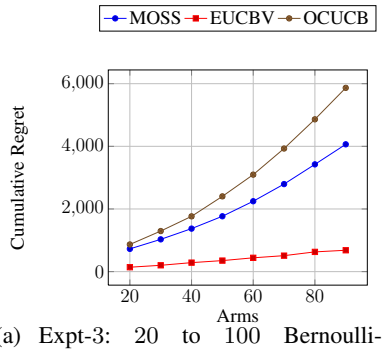
[2] The implementation for KL-UCB, Bayes-UCB and DMED were taken from Cappe et al. (2012)



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

(b) Expt-2: 100 Gaussian-distributed arms with $r_{i_{i\neq *:1-33}} = 0.1$, $r_{i_{i\neq *:34-99}} = 0.6$, $r^*_{i=100} = 0.9$ and $\sigma^2_{i=1:100} = 0.3$

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



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

Fig. 2: Further Experiments with EUCBV

Third experiment: This experiment is conducted to show the stability and performance of EUCBV over a very large horizon and over a large number of arms. This testbed comprises 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 averaged over 100 independent runs. The horizon is set at $T = 10^5 + K^3_{20:100}$ timesteps. Please note that algorithms like Thompson Sampling or Bayes-UCB are too slow to be run for such large arms (see Lattimore (2015)) and over such large horizon. We report the performance of MOSS, OCUCB and EUCBV only over this uniform gap setup. From the results in Figure 2(a), it is evident that the growth of regret for EUCBV is much lower

than that of OCUCB and MOSS. This also 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 EUCBV outperforms OCUCB.

6 Conclusion and Future Works

In this paper, we studied the EUCBV algorithm which takes into account the empirical variance of the arms and employs aggressive exploration parameters to eliminate sub-optimal arms. Our theoretical analysis conclusively established that EUCBV enjoys an order-optimal gap-independent regret bound of $O\left(\sqrt{KT}\right)$ for $T \geq O\left(K^{2.7}\right)$. Empirically we showed that EUCBV performs superbly across a diverse experimental setting and outperforms most of the bandit algorithms in stochastic MAB setup. Our experiments showed that EUCBV is extremely stable for larger horizons and performs superbly across different types of distributions. One future work is to remove the constraint of $T \geq O\left(K^{2.7}\right)$ required for EUCBV to reach the order optimal regret bound. Another future direction is to come up with an anytime version of EUCBV.

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. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p.

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. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

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.

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.