Efficient-UCBV: An Almost Optimal Algorithm using Variance Estimates

Subhojyoti Mukherjee[1], K. P. Naveen[2], Nandan Sudarsanam[3], Balaraman Ravindran[1]

[1]Department of Computer Science & Engineering,
[2]Department of Electrical Engineering, [3]Department of Management Studies,
Indian Institute of Technology Madras, Chennai - 600036, India.

Abstract. We propose a novel variant of the UCB algorithm (referred to as Efficient-UCB-Variance (EUCBV)) for minimizing cumulative regret in the stochastic multi-armed bandit (MAB) setting. EUCBV incorporates the arm elimination strategy proposed in UCB-Improved [7], while taking into account the variance estimates to compute the arms' confidence bounds, similar to UCBV [4]. Through a theoretical analysis we establish that EUCBV incurs a *gap-dependent* regret bound of $O\left(\frac{K\log(T\Delta^2/K)}{\Delta}\right)$ after T trails, where Δ is the minimal gap between optimal and sub-optimal arms; the above bound is an improvement over that of existing state-of-the-art UCB algorithms (e.g., UCB1, UCBV, UCB-Improved, MOSS, and OCUCB). Further, EUCBV incurs a *gap-independent* regret bound of $O\left(\sqrt{KT}\right)$ which is an improvement over that of UCB1, UCBV and UCB-Improved, while being comparable with that of MOSS and OCUCB. Through an extensive numerical study we show that EUCBV significantly outperforms the popular UCB variants (like MOSS, OCUCB, etc.) as well as the Thompson sampling and Bayes-UCB algorithms.

Keywords: Stochastic multi-armed bandits, cumulative regret, UCB-Improved, UCBV.

1 Introduction

In this paper we deal with the stochastic multi-armed bandit (MAB) setting. In its classical form, stochastic MABs represent a sequential learning problem where a learner is exposed to a finite set of actions (or arms) and needs to choose one of the actions at each timestep. After choosing (or pulling) an arm the learner receives a reward, which is conceptualized as an independent random draw from stationary distribution associated with the selected arm. Each of these rewards is random and drawn independently 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 $r_i < r^*, \forall i \in \mathcal{A}$. With this formulation the learner faces the task of balancing exploitation and exploration. In other words, should the learner pull the arm which currently has the best known estimates or explore arms more thoroughly to ensure that a correct decision is being made. 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, and $z_i(T)$ 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, including, but not limited to, mobile channel allocations, active learning and computer simulation games.

1.1 Related Work

Bandit problems has been extensively studied in several earlier work such as [19], [18] and [15]. Lai and Robbins in [15] established an asymptotic lower bound for the cumulative regret. Over the years stochastic MABs has seen several algorithms with strong regret guarantees. For further reference an interested reader can look into [8]. The upper confidence bound algorithms balance the exploration-exploitation dilemma by linking the uncertainty in estimate of an arm with the number of times an arm is pulled, and therefore ensuring sufficient exploration. One of the earliest among these algorithms is UCB1 [5], 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 [2], 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 [2,16]).

The UCB-Improved algorithm, proposed in [7], is a round-based[1] variant of UCB1, that incurs 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 [16], 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 $H_i = \sum_{j=1}^{K} \min\left\{\frac{1}{\Delta_i^2}, \frac{1}{\Delta_j^2}\right\}$, and a gap-independent regret bound of $O\left(\sqrt{KT}\right)$.

[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.

This is the best known gap-dependent and gap-independent regret bounds in the stochastic MAB framework. However, unlike our proposed EUCBV algorithm, OCUCB does not take into account the variance of the arms; as a result, empirically we find that our algorithm outperforms OCUCB in all the environments considered.

In contrast to the above work, the UCBV [4] algorithm utilizes variance estimates to compute the confidence intervals for each arm. UCBV 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, [4] showed that UCBV outperforms UCB1 in several scenarios.

Another notable design principle which has recently gained a lot of popularity is the Thompson Sampling (TS) algorithm ([19], [1]) and Bayes-UCB (BU) algorithm [14]. The TS algorithm maintains a posterior reward distribution for each arm; at each round, the algorithm samples values from these distribution and the arm corresponding to the highest sample value is chosen. Although TS is found to perform extremely well when the reward distributions are Bernoulli, it is established that with Gaussian priors the worst case regret can be as bad as $\Omega\left(\sqrt{KT\log T}\right)$ [16]. The BU algorithm is an extension of the TS algorithm that takes confidence intervals into consideration while choosing arms.

The final design principle we will state is the information theoretic approach of DMED [13] and KL-UCB [12] algorithms. The algorithm KL-UCB uses Kullbeck-Leibler divergence to compute the upper confidence bound for the arms. KL-UCB is stable for a short horizon and is known to reach the [15] lower bound in the special case of Bernoulli distribution. However, [12] showed that KL-UCB, MOSS and UCB1 algorithms are empirically outperformed by UCBV in the exponential distribution as they do not take the variance of the arms into consideration.

1.2 Our Contributions

In this paper we propose the Efficient-UCB-Variance (henceforth referred to as EUCBV) algorithm for the stochastic MAB setting. EUCBV combines the approach of UCB-Improved, CCB [17] 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 [4] 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. Also we carefully design the confidence interval term with the variance estimates along with the pulls allocated to each arm to balance the risk of eliminating the optimal arm against excessive optimism. Theoretically we refine the analysis of [7] and prove that for $T \geq K^{2.4}$ 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 but is poorer than OCUCB. However, EUCBV's 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

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	$O\left(\frac{K\log(T/H_i)}{\Delta}\right)$	$O\left(\sqrt{KT}\right)$

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.

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. Note that except UCBV all the aforementioned algorithms do 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 [7]. In [7] the authors conjectured that an UCB-Improved like arm-elimination algorithm can greatly benefit by taking into consideration the variance of the arms. Also it is the 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 [7] 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 present the EUCBV algorithm. Our main theoretical results are stated in Section 3, while the proofs are established in section 4. Section 5 contains results and discussions from our numerical experiments. Finally, we draw our conclusions in section 6 and discuss about future works.

2 Algorithm: Efficient UCB Variance

2.1 Notations: We denote the set of arms by $\mathcal{A}$, with the individual arms labeled i, where $i = 1, \ldots, K$. 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 ℓ-th time, and $z_i(t)$ is the number of

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

Arm Elimination

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

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

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

Reset Parameters

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

$B_{m+1} := B_m$

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

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

$m := m + 1$

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

end if

end for

times arm i has been pulled until timestep t. We denote the true 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 round-based arm elimination algorithms like Median Elimination [11] and UCB-Improved [7] 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 $\epsilon_m < \frac{\Delta_i}{2}$, where the quantity ϵ_m is initialized to 1 and halved after every round. 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.

EUCBV 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, similar to CCB [17] 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 algorithms (which are based on mean estimation) EUCBV employs mean and variance estimates (as in [4]) for arm elimination. Further, we allow for arm-elimination at every time-step, which is in contrast to the earlier work (e.g., [7]; [11]) where the arm elimination takes place only at the end of the respective exploration rounds.

3 Main Results

The main result of the paper is presented in the following theorem where we establish a regret upper bound for the proposed EUCBV algorithm.

Theorem 1 (*Gap-Dependent Bound*). *For $T \geq K^{2.4}$, $\rho = \frac{1}{2}$ and $\psi = \frac{T}{K^2}$, the regret R_T for EUCBV satisfies*

$$\mathbb{E}[R_T] \leq \sum_{i \in \mathcal{A}: \Delta_i > b} \left\{ 64K + \left(\Delta_i + \frac{64 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right) \right\} + \sum_{i \in \mathcal{A}: 0 < \Delta_i \leq b} 32K + \max_{i \in \mathcal{A}: 0 < \Delta_i \leq b} \Delta_i T$$

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

Proof (Outline). The proof is along the lines of the technique in [7]. It comprises of three modules. In the first module we prove the necessary conditions for arm elimination within a specified number of rounds. However, here we require an additional technical result (see Lemma 1) to bound the length of the confidence intervals. Further, note that our algorithm combines the variance-estimate based approach of [4] with the arm-elimination technique of [7]. Also, while [7] uses Chernoff-Hoeffding bound to derive their regret bound whereas in our work we use Bernstein inequality (as in [4]) to obtain the bound. In the second module we bound the number of pulls required if an arm is eliminated on or before a particular number of rounds. Note that the number of pulls allocated in a round m for each arm is $n_m := \left\lceil \frac{\log(\psi T \epsilon_m^2)}{2\epsilon_m} \right\rceil$ which is much lower than the number of pulls of each arm required by UCB-Improved or Median-Elimination. Finally, the third module deals with case of bounding the regret, given that a sub-optimal arm eliminates the optimal arm. The detailed proof is available in Section 4. ■

Discussion: From the above result we see that the most significant term in the gap-dependent bound is of the order $O\left(\frac{K \log(T\Delta^2/K)}{\Delta}\right)$ which is better than the existing results for UCB1, UCBV, MOSS and UCB-Improved (see Table 1). Audibert et al. in [3] have defined the term $H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$, which is referred to as the hardness of a problem; Bubeck et al. in [8] have conjectured that the gap-dependent regret

upper bound can match $O\left(\frac{K\log(T/H_1)}{\Delta}\right)$. However, in [16] it is 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 $H_i = \sum_{j=1}^{K}\min\left\{\frac{1}{\Delta_i^2},\frac{1}{\Delta_j^2}\right\}$ (OCUCB proposed in [16] achieves this bound). Further, in [16] it is shown that only in the worst case scenario when all the gaps are equal (so that $H_1 = H_i = \sum_{i=1}^{K}\frac{1}{\Delta^2}$) the above two bounds match. In the latter scenario, we see that the gap-dependent bound of our proposed EUCBV simplifies to $O\left(\frac{K\log(T/H_1)}{\Delta}\right)$, thus matching the gap-dependent bound of OCUCB which is order optimal.

Next, we specialize the result of Theorem 1 in Corollary 1 to obtain the gap-independent worst case regret bound.

Corollary 1 (*Gap-Independent Bound*). *When the gaps of all the sub-optimal arms are identical, i.e., $\Delta_i = \Delta = \sqrt{\frac{K\log K}{T}} > \sqrt{\frac{e}{T}}, \forall i \in \mathcal{A}$, the regret of EUCBV is upper bounded by the following gap-independent expression:*

$$\mathbb{E}[R_T] \leq 96K^2 + 64\sqrt{KT}.$$

Discussion: In the non-stochastic scenario, Auer et al. in [6] showed that the bound on the cumulative regret for EXP-3 is $O\left(\sqrt{KT\log K}\right)$. However, in the stochastic case, UCB1 proposed in [5] incurred a regret of order of $O\left(\sqrt{KT\log T}\right)$ which is clearly improvable. From the above result we see that in the gap-independent bound of EUCBV the most significant term is $O\left(\sqrt{KT}\right)$ which matches the upper bound of MOSS and OCUCB, and is better than UCB-Improved, UCB1 and UCBV (see Table 1).

Proof. From [9] we know that the function $x \in [0,1] \mapsto x\exp(-Cx^2)$ is decreasing on $\left[\frac{1}{\sqrt{2C}},1\right]$ for any $C > 0$. Thus, we take $C = \lfloor\frac{T}{e}\rfloor$ and choose $\Delta_i = \Delta = \sqrt{\frac{K\log K}{T}} > \sqrt{\frac{e}{T}}$ for all i.

First, let us recall the result in Theorem 1 below:

$$\mathbb{E}[R_T] \leq \sum_{i\in\mathcal{A}:\Delta_i>b}\left\{64K + \left(\Delta_i + \frac{64\log(\frac{T\Delta_i^2}{K})}{\Delta_i}\right)\right\} + \sum_{i\in\mathcal{A}:0<\Delta_i\leq b} 32K + \max_{i\in\mathcal{A}:0<\Delta_i\leq b}\Delta_i T$$

Now, with $\Delta_i = \Delta = \sqrt{\frac{K\log K}{T}} > \sqrt{\frac{e}{T}}$ we obtain,

$$\sum_{i\in\mathcal{A}:\Delta_i>b}\frac{64\log(\frac{T\Delta_i^2}{K})}{\Delta_i} = \frac{64K\sqrt{T}\log(T\frac{K(\log K)}{TK})}{\sqrt{K\log K}} \leq \frac{64\sqrt{KT}\log(\log K)}{\sqrt{\log K}} \overset{(a)}{\leq} 64\sqrt{KT}$$

where (a) follows from the identity $\frac{\log(\log K)}{\sqrt{\log K}} \leq 1$ for $K \geq 2$. Thus, the total worst case gap-independent bound is given by

$$\mathbb{E}[R_T] \leq 96K^2 + 64\sqrt{KT}.$$

■

4 Proofs

We first present the following technical lemma that is required to prove the result in Theorem 1.

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

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

Proof. The proof is based on contradiction. Suppose

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

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

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

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

Proof of Theorem 1

Proof. For each sub-optimal arm $i \in \mathcal{A}$, let $m_i = \min\left\{m | \sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{4}\right\}$. Also, let $\mathcal{A}^{'} = \{i \in \mathcal{A} : \Delta_i > b\}$ and $\mathcal{A}^{''} = \{i \in \mathcal{A} : \Delta_i > 0\}$. As in [], we bound the regret under the following two cases:

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

– Case (b): an arm $i \in B_{m_i}$ is eliminated in round m_i (or before), or there is no optimal arm $* \in B_{m_i}$

The details of each case are contained in the following sub-sections.

Case (a)**:** For simplicity, let $c_i := \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4n_{m_i}}}$ denote the length of the confidence interval corresponding to arm i in round m_i. Thus, in round m_i (or before) whenever $z_i = n_{m_i} \geq \dfrac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$, we obtain

$$
\begin{aligned}
c_i &\leq \sqrt{\frac{\rho(\hat{v}_i+2)\epsilon_{m_i}\log(\psi T\epsilon_{m_i})}{2\log(\psi T\epsilon_{m_i}^2)}} \overset{(a)}{\leq} \sqrt{\frac{2\rho\epsilon_{m_i}\log(\frac{\psi T\epsilon_{m_i}^2}{\epsilon_{m_i}})}{\log(\psi T\epsilon_{m_i}^2)}} \\
&= \sqrt{\frac{2\rho\epsilon_{m_i}\log(\psi T\epsilon_{m_i}^2) - 2\rho\epsilon_{m_i}\log(\epsilon_{m_i})}{\log(\psi T\epsilon_{m_i}^2)}} \leq \sqrt{2\rho\epsilon_{m_i} - \frac{2\rho\epsilon_{m_i}\log(\frac{1}{2^{m_i}})}{\log(\psi T\frac{1}{2^{2m_i}})}} \\
&\leq \sqrt{2\rho\epsilon_{m_i} + \frac{2\rho\epsilon_{m_i}\log(2^{m_i})}{\log(\psi T) - \log(2^{2m_i})}} \leq \sqrt{2\rho\epsilon_{m_i} + \frac{2\rho\epsilon_{m_i}m_i\log(2)}{\log(\psi T) - 2m_i\log(2)}} \\
&\overset{(b)}{\leq} \sqrt{2\rho\epsilon_{m_i} + 2.\frac{3}{2}\epsilon_{m_i}} < \sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{4}
\end{aligned}
$$

In the above simplification, (a) is due to $\hat{v}_i \in [0,1]$, while (b) is obtained using Lemma 1. Similarly, it can be shown that $c^* < \frac{\Delta_i}{4}$ in round m_i.

Now, the sufficient conditions for arm i to get eliminated by an optimal arm in round m_i is given by

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

Indeed, in round m_i suppose (1) holds, then we have

$$
\begin{aligned}
\hat{r}_i + c_i &\leq r_i + 2c_i = r_i + 4c_i - 2c_i \\
&< r_i + \Delta_i - 2c_i \leq r^* - 2c^* \leq \hat{r}^* - c^*
\end{aligned}
$$

so that a sub-optimal arm $i \in \mathcal{A}^{'}$ gets eliminated. Thus, the probability of the complementary event of these two conditions yields a bound on the probability that arm i is not eliminated in round m_i. A bound on the complementary of the first condition 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

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

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

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

The first term in the LHS of (2) can be bounded using the Bernstein inequality as below:

$$\mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) \leq \exp\left(-\frac{(\bar{c}_i)^2 z_i}{2\sigma_i^2 + \frac{2}{3}\bar{c}_i}\right) \overset{(a)}{\leq} \exp\left(-\rho\left(\frac{3\sigma_i^2 + 3\sqrt{\epsilon_{m_i}} + 6}{6\sigma_i^2 + 2\sqrt{\epsilon_{m_i}}}\right)\log(\psi T \epsilon_{m_i}\right)$$
$$\overset{(b)}{\leq} \exp\left(-\rho\log(\psi T\epsilon_{m_i})\right) \leq \frac{1}{(\psi T\epsilon_{m_i})^\rho} \quad (4)$$

where, (a) is obtained by substituting equation 3 and (b) occurs because for all $\sigma_i^2 \in [0,1]$, $\left(\frac{3\sigma_i^2 + 3\sqrt{\epsilon_{m_i}} + 6}{6\sigma_i^2 + 2\sqrt{\epsilon_{m_i}}}\right) \geq 1$.

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

$$\mathbb{P}\left\{\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right\} \leq \mathbb{P}\left\{\frac{1}{n_i}\sum_{t=1}^{n_i}(X_{i,t} - r_i)^2 - (\hat{r}_i - r_i)^2 \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right\}$$
$$\leq \mathbb{P}\left\{\frac{\sum_{t=1}^{n_i}(X_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right\} \overset{(a)}{\leq} \mathbb{P}\left\{\frac{\sum_{t=1}^{n_i}(X_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \bar{c}_i\right\}$$
$$\overset{(b)}{\leq} \exp\left(-\rho\left(\frac{3\sigma_i^2 + 3\sqrt{\epsilon_{m_i}} + 6}{6\sigma_i^2 + 2\sqrt{\epsilon_{m_i}}}\right)\log(\psi T\epsilon_{m_i})\right) \leq \frac{1}{(\psi T\epsilon_{m_i})^\rho} \quad (5)$$

where inequality (a) is obtained using (3), while (b) follows from the Bernstein inequality.

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, $\mathbb{P}\{\hat{r}^* < r^* - c^*\} \leq \frac{2}{(\psi T\epsilon_{m_i})^\rho}$. Summing the above two contributions, 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

$$\sum_{i\in\mathcal{A}^{'}}\left(\frac{4T\Delta_i}{(\psi T\epsilon_{m_i})^\rho}\right) \leq \sum_{i\in\mathcal{A}^{'}}\left(\frac{4T\Delta_i}{(\psi T\frac{\Delta_i^2}{4.16})^\rho}\right) \leq \sum_{i\in\mathcal{A}^{'}}\left(\frac{2^{2+2\rho}.16^\rho T^{1-\rho}}{\psi^\rho \Delta_i^{2\rho-1}}\right)$$
$$\overset{(a)}{\leq} \sum_{i\in\mathcal{A}^{'}}\left(\frac{2^{2+1}.16^{\frac{1}{2}}T^{1-\frac{1}{2}}}{(\frac{T}{K^2})^{\frac{1}{2}}\Delta_i^{2.\frac{1}{2}-1}}\right) = \sum_{i\in\mathcal{A}^{'}} 32K$$

Here in (a) we substitute the values of ρ and ψ.

Case (b)**:** Here, there are two sub-cases to be considered.
Case $(b1)$ ($* \in B_{m_i}$ ***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,

$$\Delta_i \left\lceil \frac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}} \right\rceil \leq \Delta_i \left\lceil \frac{\log(\psi T (\frac{\Delta_i}{16 \times 256})^4)}{2(\frac{\Delta_i}{4\sqrt{4}})^2} \right\rceil, \text{ since } \sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{4}$$

$$\leq \Delta_i \left(1 + \frac{32 \log(\psi T (\frac{\Delta_i^4}{16384})}{\Delta_i^2}\right) \leq \Delta_i \left(1 + \frac{32 \log(\psi T \Delta_i^4)}{\Delta_i^2}\right)$$

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

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

$$\overset{(a)}{\leq} \sum_{i \in \mathcal{A}^{'}} \Delta_i \left(1 + \frac{64 \log(\frac{T\Delta_i^2}{K})}{\Delta_i^2}\right)$$

We obtain (a) by substituting the value of ψ.

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 $\mathcal{A}^{'}$. 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 $m_b = \min \left\{ m | \sqrt{4\epsilon_m} < \frac{\Delta_b}{4} \right\}$. Hence, the maximal regret after eliminating the arm $*$ is upper bounded by,

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

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

$$\leq \sum_{m_*=0}^{max_{j \in \mathcal{A}^{'}} m_j} \sum_{i \in \mathcal{A}^{''}: m_i > m_*} 32 \left(\frac{T^{1-\rho}}{\psi^\rho \epsilon_{m_*}^{\rho - \frac{1}{2}}}\right)$$

$$\leq \sum_{i \in \mathcal{A}^{''}: m_i > m_*} \sum_{m_*=0}^{\min\{m_i, m_b\}} \left(\frac{32 T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_*}}\right)$$

$$\leq \sum_{i \in \mathcal{A}^{'}} \left(\frac{32 T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_*}}\right) + \sum_{i \in \mathcal{A}^{''} \setminus \mathcal{A}^{'}} \left(\frac{32 T^{1-\rho}}{\psi^\rho 2^{-(\rho - \frac{1}{2}) m_b}}\right)$$

$$
\begin{aligned}
&\leq \sum_{i \in \mathcal{A}'} \left(\frac{32T^{1-\rho} * 2^{\frac{\rho}{2}-\frac{1}{4}}}{\psi^{\rho} \Delta_i^{\rho-\frac{1}{2}}} \right) + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} \left(\frac{32T^{1-\rho_a}}{\psi^{\rho} b^{\rho-\frac{1}{2}}} \right) \\
&\leq \sum_{i \in \mathcal{A}'} \left(\frac{2^{\frac{\rho}{2}+\frac{19}{4}}.T^{1-\rho}}{\psi^{\rho} \Delta_i^{2\rho-1}} \right) + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} \left(\frac{2^{\frac{\rho}{2}+\frac{19}{4}}.T^{1-\rho}}{\psi^{\rho} b^{2\rho_a-1}} \right) \\
&\overset{(a)}{\leq} \sum_{i \in \mathcal{A}'} \left(\frac{2^{\frac{1}{4}+\frac{19}{4}}.T^{1-\frac{1}{2}}}{(\frac{T}{K^2})^{\frac{1}{2}} \Delta_i^{2.\frac{1}{2}-1}} \right) + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} \left(\frac{2^{\frac{1}{4}+\frac{19}{4}}.T^{1-\frac{1}{2}}}{(\frac{T}{K^2})^{\frac{1}{2}} b^{2.\frac{1}{2}-1}} \right) \\
&\leq \sum_{i \in \mathcal{A}'} 32K + \sum_{i \in \mathcal{A}'' \setminus \mathcal{A}'} 32K
\end{aligned}
$$

In the above simplification, (a) is obtained by substituting the values of ψ and ρ.

Finally, summing up the regrets in **Case a** and **Case b**, the total regret is given by

$$
\begin{aligned}
\mathbb{E}[R_T] \leq & \sum_{i \in \mathcal{A}: \Delta_i > b} \left\{ 64K + \left(\Delta_i + \frac{64 \log (\frac{T\Delta_i^2}{K})}{\Delta_i} \right) \right\} \\
& + \sum_{i \in \mathcal{A}: 0 < \Delta_i \leq b} 32K + \max_{i \in \mathcal{A}: 0 < \Delta_i \leq b} \Delta_i T.
\end{aligned}
$$

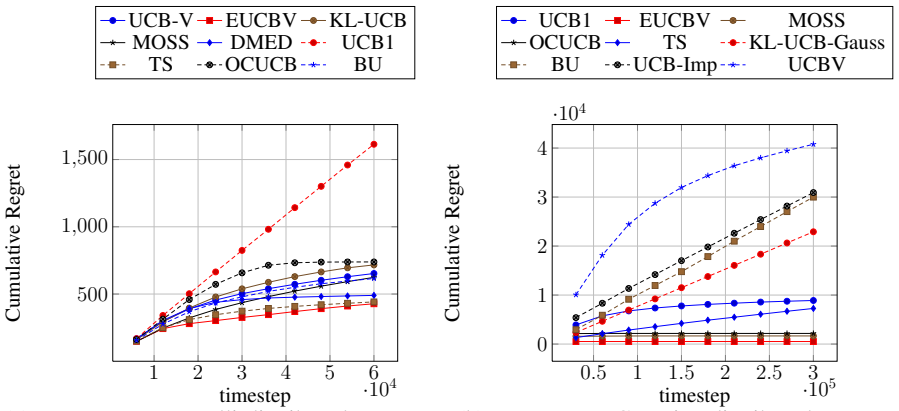
5 Experimental Section

In this section we conduct extensive empirical evaluations of EUCBV against several other popular bandit algorithms. We use cumulative regret as the metric of comparison. We implement the following algorithms: KL-UCB [12], DMED [13], MOSS [2], UCB1 [5], UCB-Improved [7], Median Elimination [11], Thompson Sampling (TS) [1], OCUCB [16], Bayes-UCB (BU) [14] and UCB-V [4][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 1).

Experiment-1 (Bernoulli with uniform gaps): 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 $r_{1:19} = 0.07$ and $r^*_{20} = 0.1$. 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 do 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. EUCBV is only slightly better than TS in this setting but better than other algorithms like Bayes-UCB and KL-UCB.

Experiment-2 (Failure of TS): This experiment is conducted to demonstrate that in certain environments when the horizon is large, gaps are small and the variance of the

[2] The implementation for KL-UCB, Bayes-UCB and DMED were taken from [10]

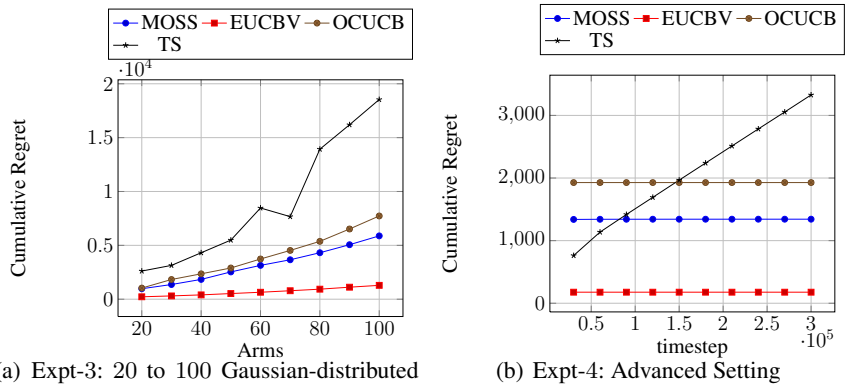


(a) Expt-1: 20 Bernoulli-distributed arms (b) Expt-2: 100 Gaussian-distributed arms

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

optimal arm is high, the Bayesian algorithms (like TS) do not perform well. This setting comprises of a large horizon of $T = 3 \times 10^5$ timesteps and a large set of arms. This testbed comprises of 100 arms involving Gaussian reward distributions with expected rewards of the arms $r_{1:33} = 0.7$, $r_{34:99} = 0.8$ and $r^*_{100} = 0.9$ with variance set as $\sigma^2_{1:33} = 0.7, \sigma^2_{34:99} = 0.1$ and $r^*_{100} = 0.7$. 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 since the gaps are small and the variances of the optimal arm and the arms farthest from the optimal arm are the highest, EUCBV, which allocates pulls proportional to the variances of the arms, outperforms all the non-variance based algorithms MOSS, OCUCB, UCB1, TS, UCB-Improved and Median-Elimination($\epsilon = 0.1, \delta = 0.1$). The performance of Median elimination is extremely weak in comparison with the other algorithms and its plot is not shown in Figure 1(b). We omit its plot in order to more clearly show the difference between EUCBV, MOSS and OCUCB. Also note that the order of magnitude in the y-axis (cumulative regret) is 10^4. The performance of TS is also weak and this is in line with the observation in [16] that the worst case regret of TS when Gaussian prior is used is $\Omega\left(\sqrt{KT\log T}\right)$. Again both Bayes-UCB and KL-UCB-Gauss perform much worse than TS.

Experiment-3 (Gaussian with large horizon and uniform gaps): 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 Gaussian reward distributions, where the expected rewards of the arms are $r_{i_{i\neq *}} = 0.05$ and $r^* = 0.1$ and variances are set as $\sigma^2_{i:i\neq *} = 0.25$ and $\sigma^2_* = 0.7$. 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. We report the performance of MOSS, TS and OCUCB who are the closest competitors of EUCBV 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 TS, OCUCB and MOSS. Because of



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

(b) Expt-4: Advanced Setting

Fig. 2: Further Experiments with EUCBV

the poor performance of Bayes-UCB and KL-UCB-Gauss in the last two experiments we do not implement them in this setup. Also Bayes-UCB and KL-UCB-Gauss are not feasible to be run over such a large horizon as these algorithms are extremely slow in their execution.

Experiment-4 (Advance Setting): This experiment demonstrates that in certain environments, when the variance of the optimal arm is higher than all the other sub-optimal arms, then EUCBV performs exceptionally well. This experiment is conducted on 100 Gaussian distributed arms such that expected rewards of the arms are $r_{1:33} = 0.4$, $r_{34:99} = 0.6$, $r^*_{100} = 0.9$ and the variance is set as $\sigma^2_{1:33} = 0.2$, $\sigma^2_{34:99} = 0.1$, $\sigma^2_{100} = 0.7$ and $T = 3 \times 10^5$. We refer to this setup as Advanced Setting because here the chosen variance values are such that only variance-aware algorithms will perform very well because the variance of the optimal arm is chosen to be higher than all the other arms. The algorithms that are not variance-aware will spent a significant amount of pulls trying to find the optimal arm. The result is shown in Figure 2(b). Predictably EUCBV, which allocates pulls proportional to the variance of the arms, outperforms its closest competitors TS, MOSS and OCUCB. The plot for UCBV, KL-UCB-Gauss and Bayes-UCB are omitted from the figure and their performance is extremely weak in comparison with other algorithms. We omit their plots to clearly show how EUCBV outperforms its nearest competitors TS, MOSS and OCUCB. Note that EUCBV by virtue of its aggressive exploration parameters outperforms UCBV in all the experiments even though UCBV is a variance based algorithm. Also in all the experiments with Gaussian distributions EUCBV significantly outperforms all the Bayesian algorithms.

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 exhibits an order-optimal gap-independent regret bound of $O\left(\sqrt{KT}\right)$. Empirically we show

that EUCBV performs superbly across diverse experimental settings and outperforms most of the bandit algorithms in stochastic MAB setup. Our experiments show that EUCBV is extremely stable for larger horizons and performs consistently well across different types of distributions. One avenue for future work is to remove the constraint of $T \geq K^{2.4}$ required for EUCBV to reach the order optimal regret bound. Another future direction is to come up with an anytime version of EUCBV. An anytime algorithm does not need the horizon T as an input parameter.

Acknowledgement: This work was fully/partially supported by funding from IIT Madras (CSE/14-15/831/RFTP/BRAV).

References

1. Agrawal, S., Goyal, N.: Analysis of thompson sampling for the multi-armed bandit problem. arXiv preprint arXiv:1111.1797 (2011)
2. Audibert, J.Y., Bubeck, S.: Minimax policies for adversarial and stochastic bandits. In: COLT. pp. 217–226 (2009)
3. Audibert, J.Y., Bubeck, S.: Best arm identification in multi-armed bandits. In: COLT-23th COLT-2010. pp. 13–p (2010)
4. Audibert, J.Y., Munos, R., Szepesvári, C.: Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. TCS 410(19), 1876–1902 (2009)
5. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine learning 47(2-3), 235–256 (2002)
6. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multiarmed bandit problem. SICOMP 32(1), 48–77 (2002)
7. Auer, P., Ortner, R.: Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. Periodica Mathematica Hungarica 61(1-2), 55–65 (2010)
8. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721 (2012)
9. Bubeck, S., Munos, R., Stoltz, G.: Pure exploration in finitely-armed and continuous-armed bandits. TCS 412(19), 1832–1852 (2011)
10. Cappe, O., Garivier, A., Kaufmann, E.: pymabandits (2012), `http://mloss.org/software/view/415/`
11. Even-Dar, E., Mannor, S., Mansour, Y.: Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. JMLR 7, 1079–1105 (2006)
12. Garivier, A., Cappé, O.: The kl-ucb algorithm for bounded stochastic bandits and beyond. arXiv preprint arXiv:1102.2490 (2011)
13. Honda, J., Takemura, A.: An asymptotically optimal bandit algorithm for bounded support models. In: COLT. pp. 67–79. Citeseer (2010)
14. Kaufmann, E., Cappé, O., Garivier, A.: On bayesian upper confidence bounds for bandit problems. In: AISTATS. pp. 592–600 (2012)
15. Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Advances in applied mathematics 6(1), 4–22 (1985)
16. Lattimore, T.: Optimally confident ucb: Improved regret for finite-armed bandits. arXiv preprint arXiv:1507.07880 (2015)
17. Liu, Y.C., Tsuruoka, Y.: Modification of improved upper confidence bounds for regulating exploration in monte-carlo tree search. TCS (2016)
18. Robbins, H.: Some aspects of the sequential design of experiments. In: Herbert Robbins Selected Papers, pp. 169–177. Springer (1952)
19. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika pp. 285–294 (1933)