Efficient-UCBV: An Almost Optimal Algorithm Using Variance Estimates

Subhojyoti Mukherjee, 1 K. P. Naveen, 2 Nandan Sudarsanam, 3,4 Balaraman Ravindran 1,4
1Department of Computer Science & Engineering, Indian Institute of Technology Madras
2Department of Electrical Engineering, Indian Institute of Technology Tirupati
3Department of Management Studies, Indian Institute of Technology Madras
4 Robert Bosch Centre for Data Science and AI (RBC-DSAI), Indian Institute of Technology Madras
1 subho@cse.iitm.ac.in, 2 naveenkp@iittp.ac.in, 3 pandan@iitm.ac.in, 4ravi@cse.iitm.ac.in

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 (Auer and Ortner 2010), while taking into account the variance estimates to compute the arms’ confidence bounds, similar to UCBV (Audibert, Munos, and Szepesvári 2009). Through a theoretical analysis we establish that EUCBV incurs a gap-dependent regret bound of $O \left( \frac{K \sigma_{\max} \log(T \Delta^2 / K)}{\Delta} \right)$ after $T$ trials, where $\Delta$ 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 (such as UCB1, UCB-Improved, UCBV, MOSS). 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 Thompson sampling and Bayes-UCB algorithms.

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. 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 $\bar{r}$ is denoted by $r^*$ such that $r_i < r^*, \forall i \in A$, where $A$ is the set of arms such that $|A| = K$. With this formulation the learner faces the task of balancing exploration and exploitation. 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:

$$R_T = r^*T - \sum_{i=1}^{K} r_i z_i(T),$$

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,

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

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 industries, including, but not limited to, mobile channel allocations, online advertising and computer simulation games.

1.1 Related Work
Bandit problems have been extensively studied in several earlier works such as Thompson (1933), Robbins (1952) and Lai and Robbins (1985). Lai and Robbins (1985) established an asymptotic lower bound for the cumulative regret. Over the years stochastic MABs have seen several algorithms with strong regret guarantees. For further reference an interested reader can look into Bubeck and Cesa-Bianchi (2012). 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 (Auer, Cesa-Bianchi, and Fischer 2002), 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...
KLUCB, 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 approaches of UCB-Improved, CCB (Liu and Tsuruoka 2016) and UCBV algorithms. EUCBV, by virtue of taking into account the empirical variance of the arms, exploration parameters and non-uniform arm selection (as opposed to UCB-Improved), performs significantly better than the existing algorithms in the stochastic MAB setting. EUCBV outperforms UCBV (Audibert, Munos, and Szepesvári 2009) which also takes into account empirical variance but is less powerful than EUCBV because of the usage of exploration regulatory factor 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 Auer and Ortner (2010) and prove that for $T \geq K^{2.4}$ our algorithm is order optimal and achieves a worst case gap-independent regret bound of $O \left( \sqrt{KT} \right)$ which is same as that of MOSS and OCUCB but better than that of UCBV, UCB1 and UCB-Improved. Also, the gap-dependent regret bound of EU-CBV 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 exploration parameters for the best performance of EUCBV. Our proof technique is highly generic and can be easily extended to other MAB settings. In Table 1 we show the regret bounds of different algorithms.

<table>
<thead>
<tr>
<th>Algo</th>
<th>Gap-Dependent</th>
<th>Gap-Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUCBV</td>
<td>$O \left( K \sigma_{\text{max}} \log \left( \frac{T \Delta^2}{K} \right) \right)$</td>
<td>$O (\sqrt{KT})$</td>
</tr>
<tr>
<td>UCB1</td>
<td>$O \left( \frac{K \log T}{\Delta} \right)$</td>
<td>$O (\sqrt{KT \log T})$</td>
</tr>
<tr>
<td>UCBV</td>
<td>$O \left( \frac{K \sigma_{\text{max}} \log T}{\Delta} \right)$</td>
<td>$O (\sqrt{KT \log T})$</td>
</tr>
<tr>
<td>UCB- Imp</td>
<td>$O \left( \frac{K \log(T \Delta^2)}{\Delta} \right)$</td>
<td>$O (\sqrt{KT \log K})$</td>
</tr>
<tr>
<td>MOSS</td>
<td>$O \left( \frac{K^2 \log(T \Delta^2 / K)}{\Delta} \right)$</td>
<td>$O (\sqrt{KT})$</td>
</tr>
<tr>
<td>OCUCB</td>
<td>$O \left( \frac{K \log(T / \Delta)}{\Delta} \right)$</td>
<td>$O (\sqrt{KT})$</td>
</tr>
</tbody>
</table>

Empirically, we show that EUCBV, owing to its estimating the variance of the arms, exploration parameters and non-uniform arm pull, performs significantly better than MOSS, OCUCB, UCB-Improved, UCB1, UCBV, TS, BU.

---

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.
DMED, KLUCB and Median Elimination algorithms. Note that except UCBV, TS, KLUCB and BU (the last three with Gaussian priors) all the aforementioned algorithms do not take into account the empirical variance estimates of the arms. Also, for the optimal performance of TS, KLUCB and BU one has to have the prior knowledge of the type of distribution, but UCBV requires no such prior knowledge. EU-CBV 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), where the authors conjectured that an UCB-Improved like arm-elimination algorithm can greatly benefit by taking into consideration the variance of the arms. A similar variance based arm-elimination algorithm has been proposed before for minimizing the expected loss in pure-exploration thresholding bandit setup in Mukherjee et al. (2017). Also, EUCBV 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 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. We draw our conclusions in section 6 and section 7 is Appendix (supplementary material).

2 Algorithm: Efficient UCB Variance

2.1 Notations: We denote the set of arms by $A$, 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 $\ast$. We denote the sample mean of the rewards for an arm $i$ at time instant $t$ by $\hat{r}_i(t) = \frac{1}{z_i(t)} \sum_{t' = 1}^{z_i(t)} X_{i,t'}$, where $X_{i,t'}$ is the reward sample received when arm $i$ is pulled for the $t'$-th time, and $z_i(t)$ is the number of times arm $i$ has been pulled until timestep $t$. We denote the true variance of an arm by $\sigma_i^2$, while $\hat{\sigma}_i(t)$ is the estimated variance, i.e., $\hat{\sigma}_i(t) = \frac{1}{z_i(t)} \sum_{t' = 1}^{z_i(t)} (X_{i,t'} - \hat{r}_i(t))^2$. Whenever there is no ambiguity about the underlying time index $t$, for simplicity we neglect $t$ from the notations and simply use $\hat{r}_i$, $\hat{\sigma}_i$, 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 (Even-Dar, Mannor, and Mansour 2006) and UCB-Improved 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 $\epsilon_m < \frac{\Delta}{2\sqrt{n}}$, where the quantity $\epsilon_m$ is initialized to 1 and halved after every round. In the worst case scenario when $K$ is large, and the gaps are uniform ($r_{1} = r_{2} = \cdots = r_{K-1} < r_{\ast}$) and small this results in very high regret.

The EUCBV algorithm, which is mainly based on the arm elimination technique of the UCB-Improved algorithm, remedies these by employing exploration regulatory factor $\psi$ and arm elimination parameter $\rho$ for aggressive elimination of sub-optimal arms. Along with these, similar to 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 algorithms (which are based on mean estimation) EUCBV employs mean and variance estimates (as in Audibert, Munos, and Szepesvári (2009)) for arm elimination. Further, we allow for arm-elimination at every exploration rounds.

Algorithm 1 EUCBV

**Input:** Time horizon $T$, exploration parameters $\rho$ and $\psi$.

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

**Pull each arm once** for $t = K + 1, \ldots, T$

**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,

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

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

**Reset Parameters**

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

$$B_{m+1} := B_m$$

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

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

$m := m + 1$

end if

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

end for

The main result of the paper is presented in the following theorem, where we establish a regret upper bound for the
Theorem 1 (Gap-Dependent Bound) For $T \geq K^{2.4}$, $\psi = \frac{T}{K^2}$, and $\rho = \frac{1}{2}$, the regret $R_T$ for EUCBV satisfies

$$E[R_T] \leq \sum_{i \in A, \Delta_i > b} \left\{ C_0 K^4 \frac{T^4}{\Delta_i^3} + \left( \Delta_i + \frac{32\sigma_i^2 \log \left( \frac{T \Delta_i^2}{K} \right)}{\Delta_i} \right) \right\}$$

$$+ \sum_{i \in A, 0 < \Delta_i \leq b} \left\{ C_2 K^5 \frac{T^2}{\Delta_i} + \max_{i \in A, 0 < \Delta_i \leq b} \Delta_i T \right\}$$

for all $b \geq \sqrt{\frac{T}{2}}$ and $C_0, C_2$ are integer constants.

Proof 1 (Outline) The proof is along the lines of the technique in Auer and Ortner (2010). 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 some additional technical results (see Lemma 1 and Lemma 2) to bound the length of the confidence intervals. Further, note that our algorithm combines the variance-estimate based approach of Audibert, Munos, and Szepesvári (2009) with the arm-elimination technique of Auer and Ortner (2010) (see Lemma 3). Also, while Auer and Ortner (2010) uses Chernoff-Hoeffding bound to derive their regret bound whereas in our work we use Bernstein inequality (as in Audibert, Munos, and Szepesvári (2009)) to obtain the bound. To bound the probability of the non-uniform arm selection before it gets eliminated we use Lemma 4 and Lemma 5. 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 $n_m := \left\lfloor \frac{\log(\psi T^2)}{2\epsilon_m} \right\rfloor$ which is much lower than the number of pulls of each arm required by UCB-Improved or Median-Elimination. We introduce the variance term in the most significant term in the bound by Lemma 6. Finally, the third module deals with case of bounding the regret, given that a sub-optimal arm eliminates the optimal arm.

Discussion: From the above result we see that the most significant term in the gap-dependent bound is of the order of $O \left( \frac{K^{2.4} \log(T \Delta_i^2 / K)}{\Delta_i} \right)$ which is better than the existing results for UCB1, UCBV, MOSS and UCB-Improved (see Table 1). Also, like UCBV, this term scales with the variance. Audibert and Bubeck (2010) 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 and Cesa-Bianchi (2012) have conjectured that the gap-dependent regret upper bound can match $O \left( \frac{K \log(T)}{\Delta} \right)$. However, in Lattimore (2015) it is proved that the gap-dependent regret bound cannot be lower than $O \left( \sum_{i=2}^{K} \frac{\log(T \Delta_i^2 / H_1)}{\Delta_i^2} \right)$, where $H_1 = \sum_{i=1}^{K} \min \left\{ \frac{1}{\Delta_i^2}, \frac{1}{\Delta_i} \right\}$ (OCUCB proposed in Lattimore (2015) achieves this bound). Further, in Lattimore (2015) it is shown that only in the worst case scenario when all the gaps are equal (so that $H_1 = H_2 = \sum_{i=1}^{K} \frac{1}{\Delta_i}$) the above two bounds match. In the latter scenario, considering $\sigma_{\Delta}^2 \leq \frac{1}{4}$ as all rewards are bounded in $[0, 1]$, we see that the gap-dependent bound of 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., $\Delta_i = \Delta = \sqrt{\frac{K \log K}{T}}, \forall i \in A$ and $C_3$ being an integer constant, the regret of EUCBV is upper bounded by the following gap-independent expression:

$$E[R_T] \leq C_5 K^5 \frac{T^2}{\epsilon_m} + 80\sqrt{KT}.$$ 

The proof is given in Appendix.

Discussion: In the non-stochastic scenario, Auer et al. (2002) showed that the bound on the cumulative regret for EXP-4 is $O \left( \sqrt{KT \log K} \right)$. However, in the stochastic case, UCB1 proposed in Auer, Cesa-Bianchi, and Fischer (2002) 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).

4 Proofs

We first present a few technical lemmas that are 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 $m \leq \frac{1}{2} \log_{2} \left( \frac{T}{\epsilon} \right)$, then,

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

Lemma 2 If $T \geq K^{2.4}$, $\psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$, $m_i = \min\{ m \sqrt{4 \epsilon_m} < \frac{\Delta_i}{4} \}$ and $c_i = \sqrt{\frac{\rho (\psi_i + 2) \log(\psi T \epsilon_m)}{4 \epsilon_i}}$, then,

$$c_i < \frac{\Delta_i}{4}.$$ 

Lemma 3 If $m_i = \min\{ m \sqrt{4 \epsilon_m} < \frac{\Delta_i}{4} \}$, $c_i = \sqrt{\frac{\rho (\psi_i + 2) \log(\psi T \epsilon_m)}{4 \epsilon_i}}$ and $n_m = \frac{\log(\psi T \epsilon_m)}{2 \epsilon_m}$, then we can show that in the $m_i$-th round,

$$P(\hat{r}_i > r_i + c_i) \leq \frac{2}{(\psi T \epsilon_m) \frac{\Delta_i}{4}}.$$

Lemma 4 If $m_i = \min\{ m \sqrt{4 \epsilon_m} < \frac{\Delta_i}{4} \}, \psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$, $c_i = \sqrt{\frac{\rho (\psi_i + 2) \log(\psi T \epsilon_m)}{4 \epsilon_i}}$, and $n_m = \frac{\log(\psi T \epsilon_m)}{2 \epsilon_m}$, then in the $m_i$-th round,

$$P(e_i^* > c_i) \leq \frac{182\sqrt{K^4 \epsilon_m}}{T^2 \sqrt{\epsilon_m}}.$$ 

Lemma 5 If $m_i = \min\{ m \sqrt{4 \epsilon_m} < \frac{\Delta_i}{4} \}, \psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$, $c_i = \sqrt{\frac{\rho (\psi_i + 2) \log(\psi T \epsilon_m)}{4 \epsilon_i}}$, and $n_m = \frac{\log(\psi T \epsilon_m)}{2 \epsilon_m}$, then in the $m_i$-th round,

$$P(z_i < n_m) \leq \frac{182\sqrt{K^4 \epsilon_m}}{T^2 \sqrt{\epsilon_m}}.$$ 

6420
Lemma 6  For two integer constants $c_1$ and $c_2$, if $20c_1 \leq c_2$ then,
\[
c_1 \left( \frac{4\sigma_i^2 + 4 \log \left( \frac{T \Delta_i^2}{K} \right)}{\Delta_i} \right) \leq c_2 \left( \frac{\sigma_i^2 \log \left( \frac{T \Delta_i^2}{K} \right)}{\Delta_i} \right).
\]

The proofs of lemmas 1 - 6 can be found in Appendix.

Proof of Theorem 1

Proof 2  I For each sub-optimal arm $i \in \mathcal{A}$, let $m_i = \min \{m_i \mid \sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{2} \}$. Also, let $\mathcal{A}' = \{i \in \mathcal{A} : \Delta_i > 0\}$ and $\mathcal{A}'' = \{i \in \mathcal{A} : \Delta_i = 0\}$. Note that as all rewards are bounded in $[0, 1]$, it implies that $0 \leq \sigma_i^2 \leq \frac{1}{4}, \forall i \in \mathcal{A}$. Now, as in Auer and Ortner (2010), 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 $\star \in B_{m_i}$
- Case (b): an arm $i \in B_{m_i}$ is eliminated in round $m_i$ (or before), or there is no optimal arm $\star \in B_{m_i}$

The details of each case are contained in the following subsections.

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

Now, the sufficient conditions for arm $i$ to get eliminated by an optimal arm in round $m_i$ is given by
\[
\hat{r}_i \leq r_i + c_i \leq r_i + \Delta_i - 2\epsilon_i \geq r^* - c^* \mbox{ and } z_i \geq n_{m_i}.
\]

Indeed, in round $m_i$ suppose (1) holds, then we have
\[
\hat{r}_i + c_i \leq r_i + 2\epsilon_i = r_i + 4\epsilon_i - 2\epsilon_i < r_i + \Delta_i - 2\epsilon_i \leq r^* - 2\epsilon_i \leq r^* - c^*
\]

so that a sub-optimal arm $i \in \mathcal{A}'$ gets eliminated. Thus, the probability of the complementary event of these four conditions in (1) yields a bound on the probability that arm $i$ is not eliminated in round $m_i$. Following the proof of Lemma 1 of Audibert, Munos, and Szepesvári (2009) we can show that a bound on the complementary of the first condition is given by,
\[
P(\hat{r}_i > r_i + c_i) \leq P(\hat{r}_i > r_i + \bar{c}_i) + P(\bar{c}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}) \leq \frac{1}{2} \frac{\sqrt{\rho i T^2 \log (\psi T \epsilon_{m_i})}}{4 \epsilon_{m_i}}.
\]

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

From Lemma 3 we can show that $P(\hat{r}_i > r_i + c_i) \leq P(\hat{r}_i > r_i + \bar{c}_i) + P(\bar{c}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}) \leq \frac{2}{(\psi T \epsilon_{m_i})^2}$. Similarly, $P(\hat{r}_i > r_i + \bar{c}_i) \leq \frac{2}{(\psi T \epsilon_{m_i})^2}$. Summing the above two contributions, the probability that a sub-optimal arm $i$ is not eliminated on or before $m_i$-th round by the first two conditions in (1) is,
\[
\leq \frac{4}{(\psi T \epsilon_{m_i})^2}.
\]

Again, from Lemma 4 and Lemma 5 we can bound the probability of the complementary of the event $c_i \geq c^*$ and $z_i \geq n_{m_i}$, by
\[
\frac{182K^4}{T^4 \sqrt{\epsilon_{m_i}}} + \frac{182K^4}{T^4 \sqrt{\epsilon_{m_i}}} \leq \frac{364K^4}{T^4 \sqrt{\epsilon_{m_i}}}.
\]

Also, for eq. (3) we can show that for any $\epsilon_{m_i} \in [\sqrt{\frac{2}{7}}, 1]$,
\[
\left( \frac{4}{(\psi T \epsilon_{m_i})^2} \right)^{(a)} \leq \left( \frac{4}{(\psi T \epsilon_{m_i})^2} \right)^{(b)} \leq \left( \frac{4K^4}{(T^4 \sqrt{\epsilon_{m_i}})} \right)^{(b)} \leq \frac{4K^4}{(T^4 \sqrt{\epsilon_{m_i}})}
\]

Here, in (a) we substitute the values of $\psi$ and $\rho$ and (b) follows from the identity $c_i^2 \geq \frac{\Delta_i}{2 T \epsilon_{m_i}}$. Set $\epsilon_{m_i} \geq \sqrt{\frac{2}{7}}$.

Summing up over all arms in $\mathcal{A}'$ and bounding the regret for all the four arm elimination conditions in (1) by (4) + (5) for each arm $i \in \mathcal{A}'$ trivially by $T \Delta_i$, we obtain
\[
\sum_{i \in \mathcal{A}'} \left( \frac{4K^4 T \Delta_i}{T^4 \sqrt{\epsilon_{m_i}}} \right) + \sum_{i \in \mathcal{A}''} \left( \frac{364K^4 T \Delta_i}{T^4 \sqrt{\epsilon_{m_i}}} \right) \leq \sum_{i \in \mathcal{A}'} \left( c_i K^4 \right).
\]

Here, (a) happens because $\sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{4}$, and in (b), $C_3$ denotes a constant integer value.

Case (b): Here, there are two sub-cases to be considered.

Case (b1) $\epsilon \in B_{m_i}$ 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,
\[
z_i \leq \left[ \log \left( \frac{\psi T \epsilon_{m_i}^2}{2 \epsilon_{m_i}} \right) \right].
\]

So, the total contribution of i until round $m_i$ is given by,
\[
\Delta_i \left[ \frac{\log (\psi T \epsilon_{m_i}^2)}{2 \epsilon_{m_i}} \right] \leq \Delta_i \left[ \frac{\log (\psi T \epsilon_{m_i}^2)}{2 \epsilon_{m_i}} \right] \leq \Delta_i \left[ \frac{\log \left( \frac{\Delta_i}{16 \times 256} \right)^4}{2 \left( \frac{\Delta_i}{4 \epsilon_{m_i}^2} \right)^2} \right]
\]
\[
\leq \Delta_i \left( 1 + \frac{32 \log \left( \frac{\psi T \epsilon_{m_i}^2}{2 \epsilon_{m_i}} \right)}{\Delta_i} \right)
\]

\[
\leq \Delta_i \left( 1 + \frac{32 \log \left( \frac{\psi T \epsilon_{m_i}^2}{2 \epsilon_{m_i}} \right)}{\Delta_i} \right).
\]

6421
Here, (a) happens because \( \sqrt{4 \epsilon_m} < \frac{\Delta_i}{4} \). Summing over all arms in \( \mathcal{A} \), the total regret is given by,

\[
\sum_{i \in \mathcal{A}'} \Delta_i \left( 1 + \frac{32 \log \left( \psi T \Delta_i^4 \right)}{\Delta_i^2} \right)
= \sum_{i \in \mathcal{A}'} \left( \Delta_i + \frac{32 \log \left( \psi T \Delta_i^4 \right)}{\Delta_i} \right)
\leq \sum_{i \in \mathcal{A}'} \left( \Delta_i + \frac{64 \log \left( \frac{T \Delta_i^2}{K} \right)}{\Delta_i} \right)
\leq \sum_{i \in \mathcal{A}'} \left( \Delta_i + \frac{16(4\sigma_i^2 + 4) \log \left( \frac{T \Delta_i^2}{K} \right)}{\Delta_i} \right)
\leq \sum_{i \in \mathcal{A}'} \left( \Delta_i + \frac{320\sigma_i^2 \log \left( \frac{T \Delta_i^2}{K} \right)}{\Delta_i} \right).
\]

We obtain (a) by substituting the value of \( \psi \), (b) from 0 \( \leq \sigma_i^2 \leq \frac{\Delta_i}{4} \), \( \forall i \in \mathcal{A} \) and (c) from Lemma 6.

**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 \( m_* \) 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 \( a_0 \) 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 \( \Delta_j \) among the remaining arms \( j \) with \( m_j \geq m_* \). Let \( m_b = \min \left\{ m \, | \, \sqrt{4 \epsilon_m} < \frac{\Delta_j}{4} \right\} \). Hence, the maximal regret after eliminating the arm * is upper bounded by,

\[
\max_{m_* = 0}^{m_j} \sum_{m = 0}^{m_j} \left( \frac{368K^4}{T^4 \epsilon_m} \right)^{T \Delta_j} \max_{j \in \mathcal{A}' \setminus m_j} \Delta_j
\leq \sum_{m = 0}^{m_j} \left( \frac{368K^4}{T^4 \epsilon_m} \right)^{T \Delta_j} \mathcal{A}' \epsilon_m
\leq \sum_{m = 0}^{m_j} \left( \frac{C_2K^4}{T^4 \epsilon_m^2} \right)^{T \Delta_j}
\leq \sum_{m = 0}^{m_j} \left( \frac{C_2K^4}{T^4} \right)^{T \Delta_j} + \sum_{m = 0}^{m_j} \left( \frac{C_2K^4}{T^4} \right)^{T \Delta_j}.
\]

Here at (a), \( C_2 \) denotes an integer constant.

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

\[
\mathbb{E}[R_T] \leq \sum_{i \in \mathcal{A}, \Delta_i > b} \left\{ \frac{C_0K^4}{T^4} + \left( \Delta_i + \frac{320\sigma_i^2 \log \left( \frac{T \Delta_i^2}{K} \right)}{\Delta_i} \right) \right\}
+ \sum_{i \in \mathcal{A}, 0 < \Delta_i \leq b} \left( \frac{C_2K^4}{T^4} + \max_{i \in \mathcal{A}, 0 < \Delta_i \leq b} \Delta_i T \right)
\]

where \( C_0, C_1, C_2 \) are integer constants s.t. \( C_0 = C_1 + C_2 \).

### 5 Experiments

In this section, we conduct extensive empirical evaluations of EUCBV against several other popular MAB algorithms. We use expected cumulative regret as the metric of comparison. The comparison is conducted against the following algorithms: KLUCB+ (Garivier and Cappe 2011), DMED (Honda and Takemura 2010), MOSS (Audibert and Bubeck 2009), UCB1 (Auer, Cesa-Bianchi, and Fischer 2002), UCB-Improved (Auer and Ortner 2010), Median Elimination (Even-Dar, Mannor, and Mansour 2006), Thompson Sampling (TS) (Agrawal and Goyal 2011), OCUCB (Lattimore 2015), Bayes-UCB (BU) (Kaufmann, Cappe, and Garivier 2012) and UCB-V (Audibert, Munos, and Szepesvári 2009). 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). Note that KLUCB+ empirically outperforms KLUCB (see Garivier and Cappe (2011)).

![Figure 1: A comparison of the cumulative regret incurred by the various bandit algorithms.](image)

**Experiment-1 (Bernoulli with uniform gaps):** This experiment is conducted to observe the performance of EU-CBV over a short horizon. The horizon \( T \) is set to 60000. The testbed comprises of 20 Bernoulli distributed arms with expected rewards of the arms as \( r_{1:19} = 0.07 \) and \( r_{20} = 0.1 \) and these type of cases are frequently encountered in web-advertising domain (see Garivier and Cappe (2011)). The regret is averaged over 100 independent runs and is shown in Figure 1(a). EUCBV, MOSS, OCUCB, UCB1, UCB-V, KLUCB+, TS, BU and DMED are run in this experimental setup. Not only do we observe that EUCBV performs better

\[\text{The implementation for KLUCB, Bayes-UCB and DMED were taken from Cappe, Garivier, and Kaufmann (2012)}\]
than all the non-variance based algorithms such as 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 compare with UCB-Improved and Median Elimination.

**Experiment-2 (Gaussian 3 Group Mean Setting):** This experiment is conducted to observe the performance of EU

CBV over a large horizon in Gaussian distribution tested. 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 in 3 groups, \( r_{1:66} = 0.07 \), \( r_{67:99} = 0.01 \) and \( r_{100} = 0.09 \) with variance set as \( \sigma_{1:66}^2 = 0.01 \), \( \sigma_{67:99}^2 = 0.25 \) and \( \sigma_{100}^2 = 0.25 \). 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, 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 EU

CBV, MOSS and OCUCB. Also note that the order of magnitude in the y-axis (cumulative regret) of Figure 1(b) is \( 10^3 \). KLUCB-Gauss+ (denoted by KLUCB-G+), TS-G and BU-G are initialized with Gaussian priors. Both KLUCB-G+ and UCBV which is a variance-aware algorithm perform much worse than TS-G and EUCBV. The performance of DMED is similar to KLUCB-G+ in this setup and its plot is omitted.

![Figure 2: Further Experiments with EUCBV](image)

**Experiment-3 (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 optimal arm is high, the Bayesian algorithms (like TS) do not perform well but EUCBV performs exceptionally well. This experiment is conducted on 100 Gaussian distributed arms such that expected rewards of the arms \( r_{1:10} = 0.045 \), \( r_{11:99} = 0.04 \), \( r_{100} = 0.05 \) and the variance is set as \( \sigma_{1:10}^2 = 0.01 \), \( \sigma_{11:99}^2 = 0.25 \) and \( T = 4 \times 10^5 \). The variance of the arms \( i = 11 : 99 \) are chosen uniform randomly between \([0.2, 0.24]\). TS and BU with Gaussian priors fail because here the chosen variance values are such that only variance-aware algorithms with appropriate exploration factors will perform well or otherwise it will get bogged down in costly exploration. The algorithms that are not variance-aware will spend a significant amount of pulls trying to find the optimal arm. The result is shown in Figure 2(a). Predictably EUCBV, which allocates pulls proportional to the variance of the arms, outperforms its closest competitors TS-G, BU-G, UCBV, MOSS and OCUCB. The plots for KLUCB-G+, DMED, UCB1, UCB-Improved and Median Elimination are omitted from the figure as their performance is extremely weak in comparison with other algorithms. We omit their plots to clearly show how EUCBV outperforms its nearest competitors. Note that EUCBV by virtue of its aggressive exploration parameters outperforms UCBV in all the experiments even though UCBV is a variance-based algorithm. The performance of TS-G is also weak and this is in line with the observation in Lattimore (2015) that the worst case regret of TS when Gaussian prior is used is \( \Omega(\sqrt{KT \log T}) \).

**Experiment-4 (Gaussian 3 Group Variance setting):** This experiment is conducted to show that when the gaps are uniform and variance of the arms is the only discriminative factor then the EUCBV performs extremely well over a very large horizon and over a large number of arms. This testbed comprises of 100 arms with Gaussian reward distributions, where the expected rewards of the arms are \( r_{1:99} = 0.09 \) and \( r_{100} = 0.1 \). The variances of the arms are divided into 3 groups. The group 1 consist of arms \( i = 1 : 49 \) where the variances are chosen uniform randomly between \([0.0, 0.05]\), group 2 consist of arms \( i = 50 : 99 \) where the variances are chosen uniform randomly between \([0.19, 0.24]\) and for the optimal arm \( i = 100 \) (group 3) the variance is set as \( \sigma_{i}^2 = 0.25 \). We report the cumulative regret averaged over 100 independent runs. The horizon is set at \( T = 4 \times 10^5 \) timesteps. We report the performance of MOSS, BU-G, UCBV, TS-G and OCUCB who are the closest competitors of EUCBV over this uniform gap setup. From the results in Figure 2(b), it is evident that the growth of regret for EUCBV is much lower than that of TS-G, MOSS, BU-G, OCUCB and UCBV. Because of the poor performance of KLUCB-G+ in the last two experiments we do not implement it in this setup. Also, note that for optimal performance BU-G, TS-G and KLUCB-G+ require the knowledge of the type of distribution to set their priors. Also, in all the experiments with Gaussian distributions EU

CBV significantly outperforms all the Bayesian algorithms initialized with Gaussian priors.

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 in conjunction with non-uniform arm selection (as opposed to UCB-Improved) to eliminate sub-optimal arms. Our theoretical analysis conclusively established that EUCBV exhibits an order-optimal gap-independent regret bound of \( O(\sqrt{KT}) \). Empirically, we show that EUCBV performs superbly across diverse experimental settings and outperforms most of the bandit algorithms in a stochastic MAB setup. Our experiments show that EUCBV is extremely stable for large 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 which does not require horizon \( T \) as input parameter.

**Acknowledgements:** This work is supported by a funding from Robert Bosch Centre for Data Science and Artificial Intelligence (RBC-DSAI) at IIT Madras. The work of the second author is supported by an INSPIRE Faculty Award of the Department of Science and Technology, Govt. of India.

**References**


