LaTeX CLASS FOR DISSERTATIONS SUBMITTED TO IITM

A THESIS

submitted by

NAME

for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHYSICS
INDIAN INSTITUTE OF TECHNOLOGY MADRAS

MONTH 2009

THESIS CERTIFICATE

This is to certify that the thesis titled **LaTeX CLASS FOR DISSERTATIONS SUBMITTED TO IIT-M**, submitted by **Author**, to the Indian Institute of Technology, Madras, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. 1
Research Guide
Professor
Dept. of Physics
IIT-Madras, 600 036

Place: Chennai

Date: 19th January 2009

ACKNOWLEDGEMENTS

Thanks to all those who made TEX and LATEX what it is today.

ABSTRACT

KEYWORDS: LaTeX; Thesis; Style files; Format.

A LaTeX class along with a simple template thesis are provided here. These can be used to easily write a thesis suitable for submission at IIT-Madras. The class provides options to format PhD, MS, M.Tech. and B.Tech. thesis. It also allows one to write a synopsis using the same class file. Also provided is a BIBTeX style file that formats all bibliography entries as per the IITM format.

The formatting is as (as far as the author is aware) per the current institute guidelines.

Contents

List of Tables

List of Figures

ABBREVIATIONS

IITM	Indian Institute of Technology, Madras
RTFM	Read the Fine Manual

NOTATION

r	Radius, m
α	Angle of thesis in degrees
β	Flight path in degrees

Chapter 1

Introduction

This document provides a simple template of how the provided `iitmdiss.cls` LaTeX class is to be used. Also provided are several useful tips to do various things that might be of use when you write your thesis.

Before reading any further please note that you are strongly advised against changing any of the formatting options used in the class provided in this directory, unless you are absolutely sure that it does not violate the IITM formatting guidelines. *Please do not change the margins or the spacing.* If you do change the formatting you are on your own (don't blame me if you need to reprint your entire thesis). In the case that you do change the formatting despite these warnings, the least I ask is that you do not redistribute your style files to your friends (or enemies).

It is also a good idea to take a quick look at the formatting guidelines. Your office or advisor should have a copy. If they don't, pester them, they really should have the formatting guidelines readily available somewhere.

To compile your sources run the following from the command line:

```
% latex thesis.tex
% bibtex thesis
% latex thesis.tex
% latex thesis.tex
```

Modify this suitably for your sources.

To generate PDF's with the links from the `hyperref` package use the following command:

```
% dvipdfm -o thesis.pdf thesis.dvi
```

1.1 Package Options

Use this thesis as a basic template to format your thesis. The `iitmdiss` class can be used by simply using something like this:

```
\documentclass[PhD]{iitmdiss}
```

To change the title page for different degrees just change the option from `PhD` to one of `MS`, `MTech` or `BTech`. The dual degree pages are not supported yet but should be quite easy to add. The title page formatting really depends on how large or small your thesis title is. Consequently it might require some hand tuning. Edit your version of `iitmdiss.cls` suitably to do this. I recommend that this be done once your title is final.

To write a synopsis simply use the `synopsis.tex` file as a simple template. The synopsis option turns this on and can be used as shown below.

```
\documentclass[PhD,synopsis]{iitmdiss}
```

Once again the title page may require some small amount of fine tuning. This is again easily done by editing the class file.

This sample file uses the `hyperref` package that makes all labels and references clickable in both the generated DVI and PDF files. These are very useful when reading the document online and do not affect the output when the files are printed.

1.2 Example Figures and tables

Fig. 1.1 shows a simple figure for illustration along with a long caption. The formatting of the caption text is automatically single spaced and indented. Table 1.1 shows a sample table with the caption placed correctly. The caption for this should always be placed before the table as shown in the example.

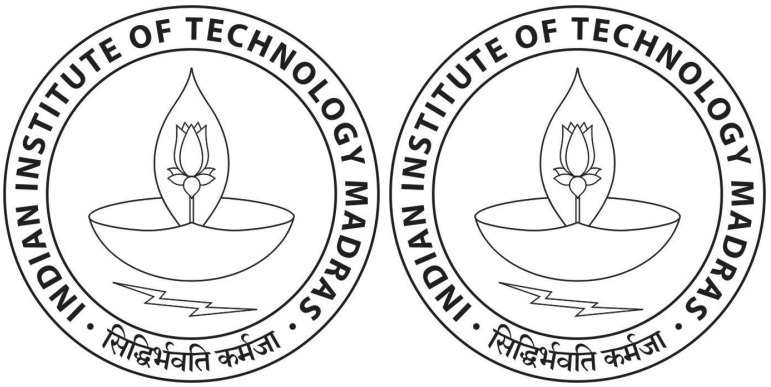


Figure 1.1: Two IITM logos in a row. This is also an illustration of a very long figure caption that wraps around two two lines. Notice that the caption is single-spaced.

Table 1.1: A sample table with a table caption placed appropriately. This caption is also very long and is single-spaced. Also notice how the text is aligned.

x	x^2
1	1
2	4
3	9
4	16
5	25
6	36
7	49
8	64

1.3 Bibliography with BIBTEX

I strongly recommend that you use BIBTEX to automatically generate your bibliography. It makes managing your references much easier. It is an excellent way to organize your references and reuse them. You can use one set of entries for your references and cite them in your thesis, papers and reports. If you haven't used it anytime before please invest some time learning how to use it.

I've included a simple example BIBTEX file along in this directory called `refs.bib`. The `iitmdiss.cls` class package which is used in this thesis and for the synopsis uses the `natbib` package to format the references along with a customized bibliography style provided as the `iitm.bst` file in the directory containing `thesis.tex`. Documentation for the `natbib` package should be available in your distribution of LATEX. Basically, to cite the author along with the author name and year use `\cite{key}` where `key` is the citation key for your bibliography entry. You can also use `\citet{key}` to get the same effect. To make the citation without the author name in the main text but inside the parenthesis use `\citep{key}`. The following paragraph shows how citations can be used in text effectively.

More information on BIBTEX is available in the book by **?**. There are many references (**??**) that explain how to use BIBTEX. Read the `natbib` package documentation for more details on how to cite things differently.

Here are other references for example. **?** presents a Python based visualization system called MayaVi in a conference paper. **?** illustrates a journal article with multiple authors. Python (**?**) is a programming language and is cited here to show how to cite something that is best identified with a URL.

1.4 Other useful LATEX packages

The following packages might be useful when writing your thesis.

- It is very useful to include line numbers in your document. That way, it is very easy for people to suggest corrections to your text. I recommend the use of the `lineno` package for this purpose. This is not a standard package but can be

obtained on the internet. The directory containing this file should contain a lineno directory that includes the package along with documentation for it.

- The `listings` package should be available with your distribution of LaTeX. This package is very useful when one needs to list source code or pseudo-code.
- For special figure captions the `ccaption` package may be useful. This is specially useful if one has a figure that spans more than two pages and you need to use the same figure number.
- The notation page can be entered manually or automatically generated using the `nomencl` package.

More details on how to use these specific packages are available along with the documentation of the respective packages.

Chapter 2

Thresholding Bandits

2.1 Introduction

Stochastic multi-armed bandit (MAB) problems are instances of the classic sequential decision-making scenario; specifically an MAB problem comprises of a learner and a collection of actions (or arms), denoted $\mathcal{A}$. In each trial the learner plays (or pulls) an arm $i \in \mathcal{A}$ which yields independent and identically distributed (i.i.d.) reward samples from a distribution (corresponding to arm i), whose expectation is denoted by r_i. The learner's objective is to identify an arm corresponding to the maximum expected reward, denoted r^*. Thus, at each time-step the learner is faced with the *exploration vs. exploitation dilemma*, where it can pull an arm which has yielded the highest mean reward (denoted $\hat{r}_i$) thus far (*exploitation*) or continue to explore other arms with the prospect of finding a better arm whose performance has not been observed sufficiently (*exploration*).

Pure-exploration MAB problems are unlike their traditional (exploration vs. exploitation) counterparts where the objective is to minimize the cumulative regret (which is the total loss incurred by the learner for not playing the optimal arm throughout the time horizon T). Instead, in pure-exploration problems a learning algorithm, until time T, can invest entirely on exploring the arms without being concerned about the loss incurred while exploring; the objective is to minimize the probability that the arm recommended at time T is not the best arm. In this paper, we further consider a combinatorial version of the pure-exploration MAB, called the thresholding bandit problem (TBP). Here, the learning algorithm is provided with a threshold τ, and the objective, after exploring for T rounds, is to output all arms i whose r_i is above τ. It is important to emphasize that the *thresholding* bandit problem is different from the *threshold* bandit setup studied in Abernethy *et al.* (2016), where the learner receives an unit reward whenever the value of an observation is above a threshold.

Formally, the problem we consider is the following. First, we define the set $S_\tau = \{i \in \mathcal{A} : r_i \geq \tau\}$. Note that, S_τ is the set of all arms whose reward mean is greater than τ. Let S_τ^c denote the complement of S_τ, i.e., $S_\tau^c = \{i \in \mathcal{A} : r_i < \tau\}$. Next, let $\hat{S}_\tau = \hat{S}_\tau(T) \subseteq \mathcal{A}$ denote the recommendation of a learning algorithm (under consideration) after T time units of exploration, while $\hat{S}_\tau^c$ denotes its complement. The performance of the learning agent is measured by the accuracy with which it can classify the arms into S_τ and S_τ^c after time horizon T. Equivalently, using $\mathbb{I}(E)$ to denote the indicator of an event E, the *loss* $\mathcal{L}(T)$ is defined as

$$\mathcal{L}(T) = \mathbb{I}\big(\{S_\tau \cap \hat{S}_\tau^c \neq \emptyset\} \cup \{\hat{S}_\tau \cap S_\tau^c \neq \emptyset\}\big).$$

Finally, the goal of the learning agent is to minimize the expected loss:

$$\mathbb{E}[\mathcal{L}(T)] = \mathbb{P}\big(\{S_\tau \cap \hat{S}_\tau^c \neq \emptyset\} \cup \{\hat{S}_\tau \cap S_\tau^c \neq \emptyset\}\big).$$

Note that the expected loss is simply the *probability of mis-classification* (i.e., error), that occurs either if a good arm is rejected or a bad arm is accepted as a good one.

The above TBP formulation has several applications, for instance, from areas ranging from anomaly detection and classification (see Locatelli *et al.* (2016)) to industrial application. Particularly in industrial applications a learners objective is to choose (i.e., keep in operation) all machines whose productivity is above a threshold. The TBP also finds applications in mobile communications (see Audibert and Bubeck (2010)) where the users are to be allocated only those channels whose quality is above an acceptable threshold.

2.1.1 Related Work

Significant amount of literature is available on the stochastic MAB setting with respect to minimizing the cumulative regret. While the seminal work of Robbins (1952), Thompson (1933), and Lai and Robbins (1985) prove asymptotic lower bounds on the cumulative regret, the more recent work of Auer *et al.* (2002*a*) propose the UCB1 algorithm that provides finite time-horizon guarantees. Also, subsequent work such as

Audibert and Bubeck (2009) and Auer and Ortner (2010) have improved the upper bounds on the cumulative regret. The authors in Auer and Ortner (2010) have proposed a *round-based*[1] version of the UCB algorithm, referred to as UCB-Improved. Of special mention is the work of Audibert *et al.* (2009) where the authors have introduced a *variance-aware* UCB algorithm, referred to as UCB-V; it is shown that the algorithms that take into account variance estimation along with mean estimation tends to perform better than the algorithms that solely focuses on mean estimation, for instance, such as UCB1. For a more detail survey of literature on UCB algorithms, we refer the reader to Bubeck and Cesa-Bianchi (2012).

In this work we are particularly interested in *pure-exploration MABs*, where the focus in primarily on simple regret rather than the cumulative regret. The relationship between cumulative regret and simple regret is proved in Bubeck *et al.* (2011) where the authors prove that minimizing the simple regret necessarily results in maximizing the cumulative regret. The pure exploration problem has been explored mainly under the following two settings:

1. Fixed Budget setting: Here the learning algorithm has to suggest the best arm(s) within a fixed time-horizon T, that is usually given as an input. The objective is to maximize the probability of returning the best arm(s). This is the scenario we consider in our paper. In Audibert and Bubeck (2010) the authors propose the UCBE and the Successive Reject (SR) algorithm, and prove simple-regret guarantees for the problem of identifying the single best arm. In the combinatorial fixed budget setup Gabillon *et al.* (2011) propose the GapE and GapE-V algorithms that suggest, with high probability, the best m arms at the end of the time budget. Similarly, Bubeck *et al.* (2013) introduce the Successive Accept Reject (SAR) algorithm, which is an extension of the SR algorithm; SAR is a round based algorithm whereby at the end of each round an arm is either accepted or rejected (based on certain confidence conditions) until the top m arms are suggested at the end of the budget with high probability. A similar combinatorial setup was explored in Chen *et al.* (2014) where the authors propose the Combinatorial Successive Accept Reject (CSAR) algorithm, which is similar in concept to SAR but with a more general setup.

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

2. Fixed Confidence setting: In this setting the learning algorithm has to suggest the best arm(s) with a fixed confidence (given as input) with as fewer number of attempts as possible. The single best arm identification has been studied in Even-Dar *et al.* (2006), while for the combinatorial setup Kalyanakrishnan *et al.* (2012) have proposed the LUCB algorithm which, on termination, returns m arms which are at least ϵ close to the true top-m arms with probability at least $1 - \delta$. For a detail survey of this setup we refer the reader to Jamieson and Nowak (2014).

Apart from these two settings some unified approaches has also been suggested in Gabillon *et al.* (2012) which proposes the algorithms UGapEb and UGapEc which can work in both the above two settings. The thresholding bandit problem is a specific instance of the pure-exploration setup of Chen *et al.* (2014). In the latest work of Locatelli *et al.* (2016) Anytime Parameter-Free Thresholding (APT) algorithm comes up with an improved anytime guarantee than CSAR for the thresholding bandit problem.

2.1.2 Our Contribution

In this paper we propose the Augmented UCB (AugUCB) algorithm for the fixed-budget setting of a specific combinatorial, pure-exploration, stochastic MAB called the thresholding bandit problem. AugUCB essentially combines the approach of UCB-Improved, CCB (Liu and Tsuruoka, 2016) and APT algorithms. Our algorithm takes into account the empirical variances of the arms along with mean estimates; to the best of our knowledge this is the first variance-based algorithm for the considered TBP. Thus, we also address an open problem discussed in Auer and Ortner (2010) of designing an algorithm that can eliminate arms based on variance estimates. In this regard, note that both CSAR and APT are not variance-based algorithms.

Our theoretical contribution comprises proving an upper bound on the expected loss incurred by AugUCB (Theorem 1). In Table 2.1 we compare the upper bound on the losses incurred by the various algorithms, including AugUCB. The terms H_1, H_2, $H_{CSAR,2}, H_{\sigma,1}$ and $H_{\sigma,2}$ represent various problem complexities, and are as defined in

Section 2.3. From Section 2.3 we note that, for all $K \geq 8$, we have

$$\log\left(K \log K\right) H_{\sigma,2} > \log(2K) H_{\sigma,2} \geq H_{\sigma,1}.$$

Thus, it follows that the upper bound for UCBEV is better than that for AugUCB. However, implementation of UCBEV algorithm requires $H_{\sigma,1}$ as input, whose computation is not realistic in practice. In contrast, our AugUCB algorithm requires no such complexity factor as input.

Proceeding with the comparisons, we emphasize that the upper bound for AugUCB is, in fact, not comparable with that of APT and CSAR; this is because the complexity term $H_{\sigma,2}$ is not explicitly comparable with either H_1 or $H_{CSAR,2}$. However, through extensive simulation experiments we find that AugUCB significantly outperforms both APT, CSAR and other non variance-based algorithms. AugUCB also outperforms UCBEV under explorations where non-optimal values of $H_{\sigma,1}$ are used. In particular, we consider experimental scenarios comprising large number of arms, with the variances of arms in S_τ being large. AugUCB, being variance based, exhibits superior performance under these settings.

The remainder of the paper is organized as follows. In section 2.2 we present our AugUCB algorithm. Section 2.3 contains our main theorem on expected loss, while section 2.4 contains simulation experiments. We finally draw our conclusions in section 2.5.

Table 2.1: AugUCB vs. State of the art

Algorithm	Upper Bound on Expected Loss
AugUCB	$\exp\left(-\frac{T}{4096\log(K\log K)H_{\sigma,2}} + \log\left(2KT\right)\right)$
UCBEV	$\exp\left(-\frac{1}{512}\frac{T-2K}{H_{\sigma,1}} + \log\left(6KT\right)\right)$
APT	$\exp\left(-\frac{T}{64H_1} + 2\log((\log(T)+1)K)\right)$
CSAR	$\exp\left(-\frac{T-K}{72\log(K)H_{CSAR,2}} + 2\log(K)\right)$

2.2 Augmented-UCB Algorithm

Notation and assumptions: $\mathcal{A}$ denotes the set of arms, and $|\mathcal{A}| = K$ is the number of arms in $\mathcal{A}$. For arm $i \in \mathcal{A}$, we use r_i to denote the true mean of the distribution from which the rewards are sampled, while $\hat{r}_i(t)$ denotes the estimated mean at time t. Formally, using $n_i(t)$ to denote the number of times arm i has been pulled until time t, we have $\hat{r}_i(t) = \frac{1}{n_i(t)} \sum_{z=1}^{n_i(t)} X_{i,z}$, where $X_{i,z}$ is the reward sample received when arm i is pulled for the z-th time. Similarly, we use σ_i^2 to denote the true variance of the reward distribution corresponding to arm i, while $\hat{v}_i(t)$ is the estimated variance, i.e., $\hat{v}_i(t) = \frac{1}{n_i(t)} \sum_{z=1}^{n_i(t)} (X_{i,z} - \hat{r}_i)^2$. Whenever there is no ambiguity about the underlaying time index t, for simplicity we neglect t from the notations and simply use $\hat{r}_i, \hat{v}_i$, and n_i, to denote the respective quantities. Let $\Delta_i = |\tau - r_i|$ denote the distance of the true mean from the threshold τ. Also, the rewards are assumed to take values in $[0, 1]$.

The Algorithm: The Augmented-UCB (AugUCB) algorithm is presented in Algorithm 1. AugUCB is essentially based on the arm elimination method of the UCB-Improved Auer and Ortner (2010), but adapted to the thresholding bandit setting proposed in Locatelli *et al.* (2016). However, unlike the UCB improved (which is based on mean estimation) our algorithm employs *variance estimates* (as in Audibert *et al.* (2009)) for arm elimination; to the best of our knowledge this is the first variance-aware algorithm for the thresholding bandit problem. Further, we allow for arm-elimination at each time-step, which is in contrast to the earlier work (e.g., Auer and Ortner (2010); Chen *et al.* (2014)) where the arm elimination task is deferred to the end of the respective exploration rounds. The details are presented below.

The active set B_0 is initialized with all the arms from $\mathcal{A}$. We divide the entire budget T into rounds/phases like in UCB-Improved, CCB, SAR and CSAR. At every time-step AugUCB checks for arm elimination conditions, while updating parameters at the end of each round. As suggested by Liu and Tsuruoka (2016) to make AugUCB to overcome too much early exploration, we no longer pull all the arms equal number of times in each round. Instead, we choose an arm in the active set B_m that minimizes $(|\hat{r}_i - \tau| - 2s_i)$ where

$$s_i = \sqrt{\frac{\rho \psi_m (\hat{v}_i + 1) \log(T \epsilon_m)}{4 n_i}}$$

Algorithm 1 AugUCB

Input: Time budget T; parameter ρ; threshold τ
Initialization: $B_0 = \mathcal{A}$; $m = 0$; $\epsilon_0 = 1$;

$$M = \left\lfloor \frac{1}{2} \log_2 \frac{T}{e} \right\rfloor; \;\; \psi_0 = \frac{T\epsilon_0}{128\Big(\log(\frac{3}{16}K \log K)\Big)^2};$$

$$\ell_0 = \left\lceil \frac{2\psi_0 \log(T\epsilon_0)}{\epsilon_0} \right\rceil; \;\; N_0 = K\ell_0$$

Pull each arm once

for $t = K+1, .., T$ **do**
  Pull arm $j \in \arg\min_{i \in B_m} \Big\{ |\hat{r}_i - \tau| - 2s_i \Big\}$
  $t \leftarrow t + 1$
  for $i \in B_m$ **do**
    if $(\hat{r}_i + s_i < \tau - s_i)$ or $(\hat{r}_i - s_i > \tau + s_i)$ **then**
      $B_m \leftarrow B_m \backslash \{i\}$ (Arm deletion)
    end if
  end for

  if $t \geq N_m$ and $m \leq M$ **then**
    Reset Parameters
    $\epsilon_{m+1} \leftarrow \frac{\epsilon_m}{2}$
    $B_{m+1} \leftarrow B_m$
    $\psi_{m+1} \leftarrow \frac{T\epsilon_{m+1}}{128(\log(\frac{3}{16}K \log K))^2}$
    $\ell_{m+1} \leftarrow \left\lceil \frac{2\psi_{m+1} \log(T\epsilon_{m+1})}{\epsilon_{m+1}} \right\rceil$
    $N_{m+1} \leftarrow t + |B_{m+1}|\ell_{m+1}$
    $m \leftarrow m + 1$
  end if
end for
Output: $\hat{S}_\tau = \{i : \hat{r}_i \geq \tau\}$.

with ρ being the arm elimination parameter and ψ_m being the exploration regulatory factor. The above condition ensures that an arm closer to the threshold τ is pulled; parameter ρ can be used to fine tune the elimination interval. The choice of exploration factor, ψ_m, comes directly from Audibert and Bubeck (2010) and Bubeck *et al.* (2011) where it is stated that in pure exploration setup, the exploring factor must be linear in T (so that an exponentially small probability of error is achieved) rather than being logarithmic in T (which is more suited for minimizing cumulative regret).

2.3 Theoretical Results

Let us begin by recalling the following definitions of the *problem complexity* as introduced in Locatelli *et al.* (2016):

$$H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2} \text{ and } H_{CSAR,2} = \min_{i \in \mathcal{A}} \frac{i}{\Delta_{(i)}^2}$$

where $(\Delta_{(i)} : i \in \mathcal{A})$ is obtained by arranging $(\Delta_i : i \in \mathcal{A})$ in an increasing order. Also, from Chen *et al.* (2014) we have

$$H_{CSAR,2} = \max_{i \in \mathcal{A}} \frac{i}{\Delta_{(i)}^2}.$$

$H_{CSAR,2}$ is the complexity term appearing in the bound for the CSAR algorithm. The relation between the above complexity terms are as follows (see Locatelli *et al.* (2016)):

$$H_1 \leq \log(2K) H_2 \text{ and } H_1 \leq \log(K) H_{CSAR,2}.$$

As ours is a variance-aware algorithm, we require H_1^σ (as defined in Gabillon *et al.* (2011)) that incorporates reward variances into its expression as given below:

$$H_{\sigma,1} = \sum_{i=1}^{K} \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}.$$

Finally, analogous to $H_{CSAR,2}$, in this paper we introduce the complexity term $H_{\sigma,2}$, which is given by

$$H_{\sigma,2} = \max_{i \in \mathcal{A}} \frac{i}{\tilde{\Delta}_{(i)}^2}$$

where $\tilde{\Delta}_i^2 = \frac{\Delta_i^2}{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}$, and $(\tilde{\Delta}_{(i)})$ is an increasing ordering of $(\tilde{\Delta}_i)$. Following the results in Audibert and Bubeck (2010), we can show that

$$H_{\sigma,2} \leq H_{\sigma,1} \leq \overline{\log}(K) H_{\sigma,2} \leq \log(2K) H_{\sigma,2}.$$

Our main result is summarized in the following theorem where we prove an upper

bound on the expected loss.

Theorem 1 *For $K \geq 4$ and $\rho = 1/3$, the expected loss of the AugUCB algorithm is given by,*

$$\mathbb{E}[\mathcal{L}(T)] \leq 2KT \exp\bigg(- \frac{T}{4096 \log(K \log K) H_{\sigma,2}} \bigg).$$

Proof 1 *The proof comprises of two modules. In the first module we investigate the necessary conditions for arm elimination within a specified number of rounds, which is motivated by the technique in Auer and Ortner (2010). Bounds on the arm-elimination probability is then obtained; however, since we use variance estimates, we invoke the Bernstein inequality (as in Audibert* et al. *(2009)) rather that the Chernoff-Hoeffding bounds (which is appropriate for the UCB-Improved (Auer and Ortner, 2010)). In the second module, as in Locatelli* et al. *(2016), we first define a favourable event that will yield an upper bound on the expected loss. Using union bound, we then incorporate the result from module-1 (on the arm elimination probability), and finally derive the result through a series of simplifications. The details are as follows.*

Arm Elimination: *Recall the notations used in the algorithm, Also, for each arm $i \in \mathcal{A}$, define $m_i = \min\big\{m | \sqrt{\rho \epsilon_m} < \frac{\Delta_i}{2}\big\}$. In the m_i-th round, whenever $n_i = \ell_{m_i} \geq \frac{2\psi_{m_i} \log(T\epsilon_{m_i})}{\epsilon_{m_i}}$, we obtain (as $\hat{v}_i \in [0,1]$)*

$$s_i \leq \sqrt{\frac{\rho(\hat{v}_i + 1)\epsilon_{m_i}}{8}} \leq \frac{\sqrt{\rho \epsilon_{m_i}}}{2} < \frac{\Delta_i}{4}. \tag{2.1}$$

First, let us consider a bad arm $i \in \mathcal{A}$ (i.e., $r_i < \tau$). We note that, in the m_i-th round whenever $\hat{r}_i \leq r_i + 2s_i$, then arm i is eliminated as a bad arm. This is easy to verify as follows: using (2.1) we obtain,

$$\hat{r}_i \leq r_i + 2s_i < r_i + \Delta_i - 2s_i = \tau - 2s_i$$

which is precisely one of the elimination conditions in Algorithm 1. Thus, the probability that a bad arm is not eliminated correctly in the m_i-th round (or before) is given by

$$\mathbb{P}(\hat{r}_i > r_i + 2s_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + 2\bar{s}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\rho\epsilon_{m_i}}\right) \tag{2.2}$$

where

$$\bar{s}_i = \sqrt{\frac{\rho\psi_{m_i}(\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}} + 1)\log(T\epsilon_{m_i})}{4n_i}}$$

Note that, substituting $n_i = \ell_{m_i} \geq \frac{2\psi_{m_i}\log(T\epsilon_{m_i})}{\epsilon_{m_i}}$, $\bar{s}_i$ *can be simplified to obtain,*

$$2\bar{s}_i \leq \frac{\sqrt{\rho\epsilon_{m_i}(\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}} + 1)}}{2} \leq \sqrt{\rho\epsilon_{m_i}}. \tag{2.3}$$

The first term in the LHS of (3.2) can be bounded using the Bernstein inequality as below:

$$\begin{aligned}
&\mathbb{P}\left(\hat{r}_i > r_i + 2\bar{s}_i\right) \\
&\leq \exp\left(-\frac{(2\bar{s}_i)^2 n_i}{2\sigma_i^2 + \frac{4}{3}\bar{s}_i}\right) \\
&\leq \exp\left(-\frac{\rho\psi_{m_i}(\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}} + 1)\log(T\epsilon_{m_i})}{2\sigma_i^2 + \frac{2}{3}\sqrt{\rho\epsilon_{m_i}}}\right) \\
&\overset{(a)}{\leq} \exp\left(-\frac{3\rho T\epsilon_{m_i}}{256a^2}\left(\frac{\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}} + 1}{3\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}}}\right)\log(T\epsilon_{m_i})\right) \\
&:= \exp(-Z_i)
\end{aligned} \tag{2.4}$$

where, for simplicity, we have used α_i *to denoted the exponent in the inequality* (a)*. Also, note that* (a) *is obtained by using* $\psi_{m_i} = \frac{T\epsilon_{m_i}}{128a^2}$*, where* $a = (\log(\frac{3}{16}K\log K))$*.*

The second term in the LHS of (3.2) can be simplified as follows:

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

$$\overset{(b)}{\leq} \exp\left(-\frac{3\rho\psi_{m_i}}{2}\left(\frac{\sigma_i^2+\sqrt{\rho\epsilon_{m_i}}+1}{3\sigma_i^2+\sqrt{\rho\epsilon_{m_i}}}\right)\log(T\epsilon_{m_i})\right)$$
$$= \exp(-Z_i) \tag{2.5}$$

where inequality (a) is obtained using (A.2), while (b) follows from the Bernstein inequality.

Thus, using (A.3) and (A.4) in (3.2) we obtain $\mathbb{P}(\hat{r}_i > r_i + 2s_i) \leq 2\exp(-Z_i)$. Proceeding similarly, for a good arm $i \in \mathcal{A}$, the probability that it is not correctly eliminated in the m_i-th round (or before) is also bounded by $\mathbb{P}(\hat{r}_i < r_i - 2s_i) \leq 2\exp(-Z_i)$. In general, for any $i \in \mathcal{A}$ we have

$$\mathbb{P}(|\hat{r}_i - r_i| > 2s_i) \leq 4\exp(-Z_i). \tag{2.6}$$

Favourable Event: *Following the notation in Locatelli* et al. *(2016) we define the event*

$$\xi = \left\{\forall i \in \mathcal{A}, \forall t = 1, 2, .., T : |\hat{r}_i - r_i| \leq 2s_i\right\}.$$

Note that, on ξ each arm $i \in \mathcal{A}$ is eliminated correctly in the m_i-th round (or before). Thus, it follows that $\mathbb{E}[\mathcal{L}(T)] \leq P(\xi^c)$. Since ξ^c can be expressed as an union of the events $(|\hat{r}_i - r_i| > 2s_i)$ for all $i \in \mathcal{A}$ and all $t = 1, 2, \cdots, T$, using union bound we can write

$$\mathbb{E}[\mathcal{L}(T)] \leq \sum_{i\in\mathcal{A}}\sum_{t=1}^{T}\mathbb{P}(|\hat{r}_i - r_i| > 2s_i)$$
$$\leq \sum_{i\in\mathcal{A}}\sum_{t=1}^{T} 4\exp(-Z_i)$$
$$\leq 4T\sum_{i\in\mathcal{A}}\exp\left(-\frac{3\rho T\epsilon_{m_i}}{256a^2}\left(\frac{\sigma_i^2+\sqrt{\rho\epsilon_{m_i}}+1}{3\sigma_i^2+\sqrt{\rho\epsilon_{m_i}}}\right)\log(T\epsilon_{m_i})\right)$$
$$\overset{(a)}{\leq} 4T\sum_{i\in\mathcal{A}}\exp\left(-\frac{3T\Delta_i^2}{4096a^2}\left(\frac{4\sigma_i^2+\Delta_i+4}{12\sigma_i^2+\Delta_i}\right)\log(\frac{3}{16}T\Delta_i^2)\right)$$
$$\overset{(b)}{\leq} 4T\sum_{i\in\mathcal{A}}\exp\left(-\frac{12T\Delta_i^2}{(12\sigma_i+12\Delta_i)}\frac{\log(\frac{3}{16}K\log K)}{4096a^2}\right)$$

$$
\begin{aligned}
&\overset{(c)}{\leq} 4T \sum_{i \in \mathcal{A}} \exp\left(- \frac{T\Delta_i^2 \log(\frac{3}{16} K \log K)}{4096(\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i})a^2} \right) \\
&\overset{(d)}{\leq} 4T \sum_{i \in \mathcal{A}} \exp\left(- \frac{T \log(\frac{3}{16} K \log K)}{4096 \tilde{\Delta}_i^{-2} a^2} \right) \\
&\overset{(e)}{\leq} 4T \sum_{i \in \mathcal{A}} \exp\left(- \frac{T \log(\frac{3}{16} K \log K)}{4096 \max_j (j \tilde{\Delta}_{(j)}^{-2})(\log(\frac{3}{16} K \log K))^2} \right) \\
&\overset{(f)}{\leq} 4KT \exp\left(- \frac{T}{4096 \log(K \log K) H_{\sigma,2}} \right).
\end{aligned}
$$

The justification for the above simplifications are as follows:

• (a) *is obtained by noting that in round* m_i *we have* $\frac{\Delta_i}{4} \leq \sqrt{\rho \epsilon_{m_i}} < \frac{\Delta_i}{2}$.

• *For* (b)*, we note that the function* $x \mapsto x \exp(-Cx^2)$*, where* $x \in [0, 1]$*, is decreasing on* $[1/\sqrt{2C}, 1]$ *for any* $C > 0$ *(see Bubeck* et al. *(2011); Auer and Ortner (2010)). Thus, using* $C = \lfloor T/e \rfloor$ *and* $\min_{j \in \mathcal{A}} \Delta_j = \Delta = \sqrt{\frac{K \log K}{T}} > \sqrt{\frac{e}{T}}$*, we obtain (b).*

• *To obtain* (c) *we have used the inequality* $\Delta_i \leq \sqrt{\sigma_i^2 + (16/3)\Delta_i}$ *(which holds because* $\Delta_i \in [0, 1]$*).*

• (d) *is obtained simply by substituting* $\tilde{\Delta}_i = \frac{\Delta_i^2}{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}$ *and* $a = \log(\frac{3}{16} K \log K)$.

• *Finally, to obtain* (e) *and* (f)*, note that* $\tilde{\Delta}_i^{-2} \leq i \tilde{\Delta}_i^{-2} \leq \max_{j \in \mathcal{A}} j \Delta_{(j)}^{-2} = H_{\sigma,2}$.

2.4 Numerical Experiments

In this section, we empirically compare the performance of AugUCB against APT, UCBE, UCBEV, CSAR and the uniform-allocation (UA) algorithms. A brief note about these algorithms are as follows:

• APT: This algorithm is from Locatelli *et al.* (2016); we set $\epsilon = 0.05$, which is the margin-of-error within which APT suggests the set of good arms.

• AugUCB: This is the Augmented-UCB algorithm proposed in this paper; as in Theorem 2 we set $\rho = \frac{1}{3}$.

• UCBE: This is a modification of the algorithm in Audibert *et al.* (2009) (as it was originally proposed for the best arm identification problem); here, we set $a = \frac{T-K}{H_1}$, and

at each time-step an arm $i \in \arg\min \left\{ |\hat{r}_i - \tau| - \sqrt{\frac{a}{n_i}} \right\}$ is pulled.

• UCBEV: This is a modification of the algorithm in Gabillon *et al.* (2011) (proposed for the TopM problem); its implementation is identical to UCBE, but with $a = \frac{T-2K}{H_{\sigma,1}}$. As mentioned earlier, note that UCBEV's implementation would not be possible in real scenarios, as it requires computing the problem complexity $H_{\sigma,1}$. However, for theoretical reasons we show the best performance achievable by UCBEV. In experiment 6 we perform further explorations of UCBEV with alternate settings of a.

• CSAR: Modification of the successive-reject algorithm in Chen *et al.* (2014); here, we reject the arm farthest from τ after each round.

• UA: The naive strategy where at each time-step an arm is uniformly sampled from $\mathcal{A}$ (the set of all arms); however, UA is known to be optimal if all arms are equally difficult to classify.

Motivated by the settings considered in Locatelli *et al.* (2016), we design six different experimental scenarios that are obtained by varying the arm means and variances. Across all experiments consists of $K = 100$ arms (indexed $i = 1, 2, \cdots, 100$) of which $S_\tau = \{6, 7, \cdots, 10\}$, where we have fixed $\tau = 0.5$. In all the experiments, each algorithm is run independently for 10000 time-steps. At every time-step, the output set, $\hat{S}_\tau$, suggested by each algorithm is recorded; the output is counted as an error if $\hat{S}_\tau \neq S_\tau$. In Figure 1, for each experiment, we have reported the percentage of error incurred by the different algorithms as a function of time; Error percentage is obtained by repeating each experiment independently for 500 iterations, and then respectively computing the fraction of errors. The details of the considered experiments are as follows.

Experiment-1: The reward distributions are Gaussian with means $r_{1:4} = 0.2 + (0:3) \cdot 0.05$, $r_5 = 0.45$, $r_6 = 0.55$, $r_{7:10} = 0.65 + (0:3) \cdot 0.05$ and $r_{11:100} = 0.4$. Thus, the means of the first 10 arms follow an arithmetic progression. The remaining arms have identical means; this setting is chosen because now a significant budget is required in exploring these arms, thus increasing the problem complexity.

The corresponding variances are $\sigma^2_{1:5} = 0.5$ and $\sigma^2_{6:10} = 0.6$, while $\sigma^2_{11:100}$ is chosen independently and uniform in the interval $[0.38, 0.42]$; note that, the variances of the arms in S_τ are higher than those of the other arms. The corresponding results are

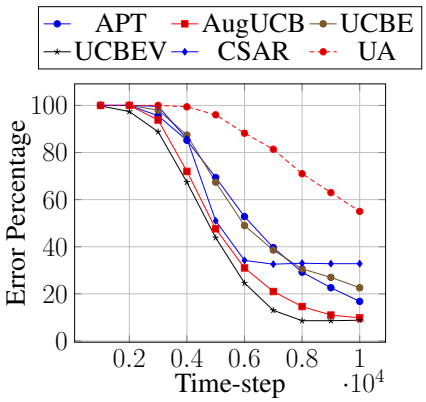
shown in Figure 2.1(a), from where we see that UCBEV, which has access to the problem complexity while being variance-aware, outperforms all other algorithm (including UCBE which also has access to the problem complexity but does not take into account the variances of the arms). Interestingly, the performance of our AugUCB (without requiring any complexity input) is comparable with UCBEV, while it outperforms UCBE, APT and the other non variance-aware algorithms that we have considered.

Experiment-2: We again consider Gaussian reward distributions. However, here the means of the first 10 arms constitute a geometric progression. Formally, the reward means are $r_{1:4} = 0.4 - (0.2)^{1:4}$, $r_5 = 0.45$, $r_6 = 0.55$, $r_{7:10} = 0.6 + (0.2)^{5-(1:4)}$ and $r_{11:100} = 0.4$; the arm variances are as in experiment-1. The corresponding results are shown in Figure 2.1(b). We again observe AugUCB outperforming the other algorithms, except UCBEV.

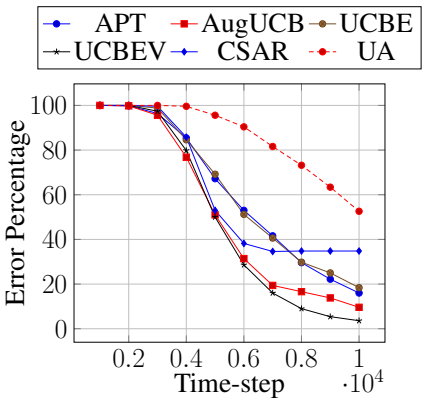
Experiment-3: Here, the first 10 arms are partitioned into three groups, with all arms in a group being assigned the same mean; the reward distributions are again Gaussian. Specifically, the reward means are $r_{1:3} = 0.1$, $r_{4:7} = \{0.35, 0.45, 0.55, 0.65\}$ and $r_{8:10} = 0.9$; as before, $r_{11:100} = 0.4$ and all the variances are as in Experiment-1. The results for this scenario are presented in Figure 2.1(c). The observations are inline with those made in the previous experiments.

Experiment-4: The setting is similar to that considered in Experiment-3, but with the first 10 arms partitioned into two groups; the respective means are $r_{1:5} = 0.45$, $r_{6:10} = 0.55$. The corresponding results are shown in Figure 2.1(d), from where the good performance of AugUCB is again validated.

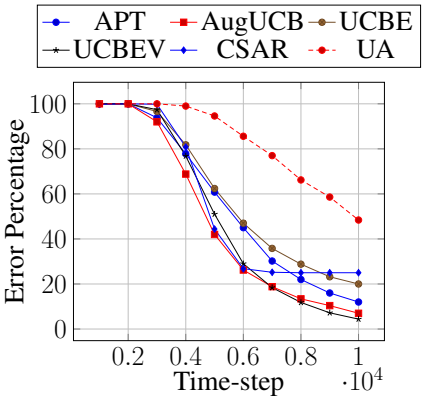
Experiment-5: This is again the two group setting involving Gaussian reward distributions. The reward means are as in Experiment-4, while the variances are $\sigma^2_{1:5} = 0.3$ and $\sigma^2_{6:10} = 0.8$; $\sigma^2_{11:100}$ are independently and uniformly chosen in the interval $[0.2, 0.3]$. The corresponding results are shown in Figure 2.1(e). We refer to this setup as *Advanced* because here the chosen variance values are such that only variance-aware algorithms will perform well.Hence, we see that UCBEV performs very well in comparison with the other algorithms. However, it is interesting to note that the performance of AugUCB catches-up with UCBEV as the time-step increases, while significantly outperforming the other non-variance aware algorithms.



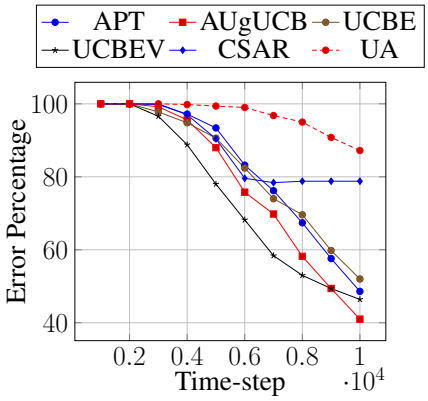
(a) Expt-1: Arithmetic Progression (Gaussian)



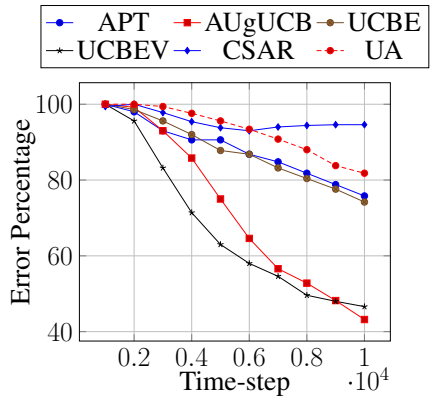
(b) Expt-2: Geometric Progression (Gaussian)



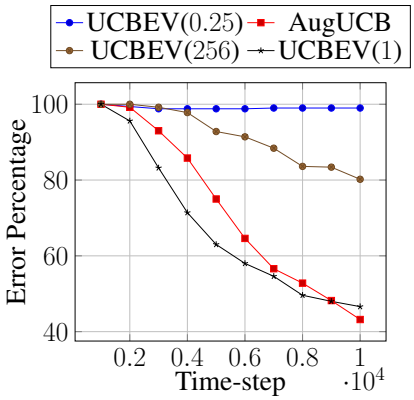
(c) Expt-3: Three Group Setting (Gaussian)



(d) Expt-4: Two Group Setting (Gaussian)



(e) Expt-5: Two Group Setting (Advance)



(f) Expt-6: Two Group Setting (Advance)

Figure 2.1: Performances of the various TBP algorithms in terms of error percentage vs. time-step, for six different experimental scenarios.

Experiment-6: We use the same setting as in Experiment-5, but conduct more exploration of UCBEV with different values of the exploration parameter a. The corresponding results are shown in Figure 2.1(f). As studied in Locatelli *et al.* (2016), we implement UCBEV with $a_i = 4^i \frac{T-2K}{H_{\sigma,1}}$ for $i = -1, 0, 4$. Here, a_0 corresponds to UCBEV(1) (in Figure 2.1(f)) which is UCBEV run with the optimal choice of $H_{\sigma,1}$.

For other choices of a_i we see that UCBEV(a_i) is significantly outperformed by AugUCB.

Finally, note that in all the above experiments, the CSAR algorithm, although performs well initially, quickly exhausts its budget and saturates at a higher error percentage. This is because it pulls all arms equally in each round, with the round lengths being non-adaptive.

2.5 Conclusion

We proposed the AugUCB algorithm for a fixed-budget, pure-exploration TBP. Our algorithm employs both mean and variance estimates for arm elimination. This, to our knowledge is the first variance-based algorithm for the specific TBP that we have considered. We first prove an upper bound on the expected loss incurred by AugUCB. We then conduct simulation experiments to validate the performance of AugUCB. In comparison with APT, CSAR and other non variance-based algorithms, we find that the performance of AugUCB is significantly better. Further, the performance of AugUCB is comparable with UCBEV (which is also variance-based), although the latter exhibits a slightly better performance. However, UCBEV is not implementable in practice as it requires computing problem complexity, $H_{\sigma,1}$, while AugUCB (requiring no such inputs) can be easily deployed in real-life scenarios. It would be an interesting future work to design an anytime version of the AugUCB algorithm.

2.6 Summary

In this chapter we looked at the Augmented-UCB (AugUCB) algorithm for a fixed-budget version of the thresholding bandit problem (TBP), where the objective is to identify a set of arms whose expected mean is above a threshold. A key feature of AugUCB is that it uses both mean and variance estimates to eliminate arms that have been sufficiently explored; to the best of our knowledge this is the first algorithm to employ such an approach for the considered TBP. Theoretically, we obtain an upper bound on the loss (probability of mis-classification) incurred by AugUCB. Although UCBEV in

literature provides a better guarantee, it is important to emphasize that UCBEV has access to problem complexity (whose computation requires arms' mean and variances), and hence is not realistic in practice; this is in contrast to AugUCB whose implementation does not require any such complexity inputs. We conduct extensive simulation experiments to validate the performance of AugUCB. Through our simulation work, we establish that AugUCB, owing to its utilization of variance estimates, performs significantly better than the state-of-the-art APT, CSAR and other non variance-based algorithms.

Chapter 3

Efficient UCB Variance

3.1 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 $*$ is denoted by r^* such that $r_i < r^*, \forall i \in \mathcal{A}$, where $\mathcal{A}$ is the set of arms such that $|\mathcal{A}| = K$. 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 industries, in-

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

3.1.1 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 in 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 *et al.*, 2002*a*), 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[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 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 $H_i = \sum_{j=1}^{K}\min\left\{\frac{1}{\Delta_i^2},\frac{1}{\Delta_j^2}\right\}$, and a gap-independent regret bound of $O\left(\sqrt{KT}\right)$. This is the best known gap-dependent and gap-independent regret bounds in the stochastic MAB framework. However, unlike our

[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.

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 (Audibert *et al.*, 2009) 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, Audibert *et al.* (2009) 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 ((Thompson, 1933), (Agrawal and Goyal, 2011)) and Bayes-UCB (BU) algorithm (Kaufmann *et al.*, 2012). 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 $\Omega\left(\sqrt{KT\log T}\right)$ (Lattimore, 2015). The BU algorithm is an extension of the TS algorithm that takes quartile deviations into consideration while choosing arms.

The final design principle we state is the information theoretic approach of DMED (Honda and Takemura, 2010) and KLUCB (Garivier and Cappé, 2011) algorithms. The algorithm KLUCB uses Kullbeck-Leibler divergence to compute the upper confidence bound for the arms. KLUCB is stable for a short horizon and is known to reach the Lai and Robbins (1985) lower bound in the special case of Bernoulli distribution. However, Garivier and Cappé (2011) showed that 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.

3.1.2 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

Table 3.1: Regret upper bound of different algorithms

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

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 *et al.*, 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 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 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 3.1.

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, 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 EUCBV requires no such prior knowledge. 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), where 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 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 3.2 we present the EUCBV algorithm. Our main theoretical results are stated in section 3.3, while the proofs are established in section 3.4. Section 3.5 contains results and discussions from our numerical experiments. We draw our conclusions in section 3.6 and section A.1 is Appendix (supplementary material).

3.2 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 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

Algorithm 2 EUCBV

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

Initialization: Set $m := 0$, $B_0 := \mathcal{A}$, $\epsilon_0 := 1$, $M = \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$, $n_0 = \lceil \frac{\log(\psi T \epsilon_0^2)}{2\epsilon_0} \rceil$ 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 by Mean Estimation

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

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

$B_{m+1} := B_m$

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

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

$m := m + 1$

end if

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

end for

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 (Even-Dar *et al.*, 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 $\epsilon_m < \frac{\Delta_i}{2}$, 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.

The 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 (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 *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.

3.3 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 2 (*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*

$$\mathbb{E}[R_T] \leq \sum_{i \in \mathcal{A}:\Delta_i > b} \left\{ \frac{C_0 K^4}{T^{\frac{1}{4}}} + \left(\Delta_i + \frac{320\sigma_i^2 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right) \right\} + \sum_{i \in \mathcal{A}:0<\Delta_i \leq b} \frac{C_2 K^4}{T^{\frac{1}{4}}} + \max_{i \in \mathcal{A}:0<\Delta_i \leq b} \Delta_i T.$$

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

Proof 2 (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 et al. *(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* et al. *(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 $n_m := \left\lceil \frac{\log(\psi T \epsilon_m^2)}{2\epsilon_m} \right\rceil$ which is much lower than the number of pulls of each arm required by UCB-Improved or Median-Elimination. 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 $O\left(\frac{K\sigma^2_{\max}\log(T\Delta^2/K)}{\Delta}\right)$ which is better than the existing results for UCB1, UCBV, MOSS and UCB-Improved (see Table 3.1). Also as 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 $O\left(\frac{K\log(T/H_1)}{\Delta}\right)$. However, in Lattimore (2015) it is proved that the gap-dependent regret bound cannot be lower than $O\left(\sum_{i=2}^{K}\frac{\log(T/H_i)}{\Delta_i}\right)$, where $H_i = \sum_{j=1}^{K}\min\left\{\frac{1}{\Delta_i^2}, \frac{1}{\Delta_j^2}\right\}$ (OCUCB proposed in 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 $H_1 = H_i = \sum_{i=1}^{K}\frac{1}{\Delta^2}$) the above two bounds match. In the latter scenario, considering $\sigma^2_{\max} \leq \frac{1}{4}$ 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 2 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., $\Delta_i = \Delta = \sqrt{\frac{K\log K}{T}} > \sqrt{\frac{e}{T}}, \forall i \in \mathcal{A}$ and C_3 being an integer constant,*

the regret of EUCBV is upper bounded by the following gap-independent expression:

$$\mathbb{E}[R_T] \leq \frac{C_3 K^5}{T^{\frac{1}{4}}} + 320\sqrt{KT}.$$

The proof is given in Appendix A.1.7.

Discussion: In the non-stochastic scenario, Auer *et al.* (2002*b*) 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 *et al.* (2002*a*) 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 3.1).

3.4 Proofs

We first present a few technical lemmas that is required to prove the result in Theorem 2.

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,*

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

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* $c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$, *then,*

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

Lemma 3 *If* $m_i = min\{m|\sqrt{4\epsilon_m} < \frac{\Delta_i}{4}\}$, $c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$ *and* $n_{m_i} = \frac{\log(\psi T\epsilon_{m_i})}{2\epsilon_{m_i}}$ *then we can show that,*

$$\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \frac{2}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}}.$$

Lemma 4 *If* $m_i = min\{m|\sqrt{4\epsilon_m} < \frac{\Delta_i}{4}\}$, $\psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$, $c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$ *and* $n_{m_i} = \frac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ *then in the* m_i*-th round,*

$$\mathbb{P}\{c^* > c_i\} \leq \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}.$$

Lemma 5 *If* $m_i = min\{m|\sqrt{4\epsilon_m} < \frac{\Delta_i}{4}\}$,$\psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$, $c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$ *and* $n_{m_i} = \frac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ *then in the* m_i*-th round,*

$$\mathbb{P}\{z_i < n_{m_i}\} \leq \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}.$$

Lemma 6 *For two integer constants* c_1 *and* c_2*, if* $20c_1 \leq c_2$ *then,*

$$c_1\frac{4\sigma_i^2+4}{\Delta_i}\log\left(\frac{T\Delta_i^2}{K}\right) \leq c_2\frac{\sigma_i^2}{\Delta_i}\log\left(\frac{T\Delta_i^2}{K}\right).$$

The proofs of lemmas 1 - 6 can be found in Appendix A.1.1, A.1.2, A.1.3, A.1.4, A.1.5 and A.1.6 respectively.

Proof of Theorem 1

Proof 1 *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* $\mathcal{A}^{''} = \{i \in \mathcal{A} : \Delta_i > 0\}$*. 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* $* \in B_{m_i}$
- *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* $c_i := \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$ *denote the length of the confidence interval corresponding to arm* i *in round* m_i*. Thus, in round* m_i *(or before) whenever* $z_i \geq n_{m_i} \geq \frac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$*, 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

$$\hat{r}_i \leq r_i + c_i,\ \hat{r}^* \geq r^* - c^*,\ c_i \geq c^* \text{ and } z_i \geq n_{m_i}. \tag{3.1}$$

Indeed, in round m_i suppose (3.1) 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 four conditions in (3.1) yields a bound on the probability that arm i is not eliminated in round m_i. Following the proof of Lemma 1 of Audibert et al. *(2009) we can show that a bound on the complementary of the first condition is given by,*

$$\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right) \tag{3.2}$$

where

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

From Lemma 3 we can show that $\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right) \leq \frac{2}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}}$. *Similarly,* $\mathbb{P}\{\hat{r}^* < r^* - c^*\} \leq \frac{2}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}}$. *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 (3.1) is,*

$$\left(\frac{4}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}}\right). \tag{3.3}$$

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,

$$\frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}} + \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}} \leq \frac{364K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}. \tag{3.4}$$

Also, for eq. (3.3) *we can show that for any* $\epsilon_{m_i} \in [\sqrt{\frac{e}{T}}, 1]$

$$\begin{aligned}\left(\frac{4}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}}\right) &\overset{(a)}{\leq} \left(\frac{4}{(\frac{T^2}{K^2}\epsilon_{m_i})^{\frac{3}{4}}}\right) \leq \left(\frac{4K^{\frac{3}{2}}}{(T^{\frac{3}{2}}\epsilon_{m_i}^{\frac{1}{4}}\sqrt{\epsilon_{m_i}})}\right) \\ &\overset{(b)}{\leq} \left(\frac{4K^{\frac{3}{2}}}{(T^{\frac{3}{2}-\frac{1}{8}}\sqrt{\epsilon_{m_i}})}\right) \leq \frac{4K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}.\end{aligned} \tag{3.5}$$

Here, in (a) *we substitute the values of* ψ *and* ρ *and* (b) *follows from the identity* $\epsilon_{m_i}^{\frac{1}{4}} \geq (\frac{e}{T})^{\frac{1}{8}}$ *as* $\epsilon_{m_i} \geq \sqrt{\frac{e}{T}}$.

Summing up over all arms in $\mathcal{A}'$ *and bounding the regret for all the four arm elimination conditions in (3.1) by* $(3.4) + (3.5)$ *for each arm* $i \in \mathcal{A}'$ *trivially by* $T\Delta_i$*, we obtain*

$$\begin{aligned}&\sum_{i\in\mathcal{A}'}\left(\frac{4K^4T\Delta_i}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}\right) + \sum_{i\in\mathcal{A}'}\left(\frac{364K^4T\Delta_i}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}\right) \\ &\overset{(a)}{\leq} \sum_{i\in\mathcal{A}'}\left(\frac{368K^4T\Delta_i}{T^{\frac{5}{4}}\left(\frac{\Delta_i^2}{4.16}\right)^{\frac{1}{2}}}\right) \overset{(b)}{\leq} \sum_{i\in\mathcal{A}'}\left(\frac{C_1K^4}{(T)^{\frac{1}{4}}}\right).\end{aligned}$$

Here, (a) *happens because* $\sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{4}$*, and in* (b)*,* C_1 *denotes a constant integer value.*

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,*

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

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

$$\Delta_i \left\lceil \frac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}} \right\rceil \overset{(a)}{\leq} \Delta_i \left\lceil \frac{\log(\psi T (\frac{\Delta_i}{16 \times 256})^4)}{2(\frac{\Delta_i}{4\sqrt{4}})^2} \right\rceil$$
$$\leq \Delta_i \left(1 + \frac{32 \log(\psi T (\frac{\Delta_i^4}{16384})}{\Delta_i^2}\right) \leq \Delta_i \left(1 + \frac{32 \log(\psi T \Delta_i^4)}{\Delta_i^2}\right).$$

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

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

We obtain (a) by substituting the value of ψ, (b) from $0 \leq \sigma_i^2 \leq \frac{1}{4}, \forall i \in \mathcal{A}$ 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_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,*

$$
\begin{aligned}
&\sum_{m_*=0}^{max_{j\in\mathcal{A}'}m_j} \sum_{i\in\mathcal{A}'':m_i>m_*} \left(\frac{368K^4}{(T^{\frac{5}{4}}\sqrt{\epsilon_{m_*}})}\right).T \max_{j\in\mathcal{A}'':m_j\geq m_*} \Delta_j \\
&\leq \sum_{m_*=0}^{max_{j\in\mathcal{A}'}m_j} \sum_{i\in\mathcal{A}'':m_i>m_*} \left(\frac{368K^4\sqrt{4}}{(T^{\frac{5}{4}}\sqrt{\epsilon_{m_*}})}\right).T.4\sqrt{\epsilon_{m_*}} \\
&\overset{(a)}{\leq} \sum_{m_*=0}^{max_{j\in\mathcal{A}'}m_j} \sum_{i\in\mathcal{A}'':m_i>m_*} \left(\frac{C_2K^4}{T^{\frac{1}{4}}\epsilon_{m_*}^{\frac{1}{2}-\frac{1}{2}}}\right) \\
&\leq \sum_{i\in\mathcal{A}'':m_i>m_*} \sum_{m_*=0}^{\min\{m_i,m_b\}} \left(\frac{C_2K^4}{T^{\frac{1}{4}}}\right) \\
&\leq \sum_{i\in\mathcal{A}'} \left(\frac{C_2K^4}{T^{\frac{1}{4}}}\right) + \sum_{i\in\mathcal{A}''\setminus\mathcal{A}'} \left(\frac{C_2K^4}{T^{\frac{1}{4}}}\right).
\end{aligned}
$$

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*

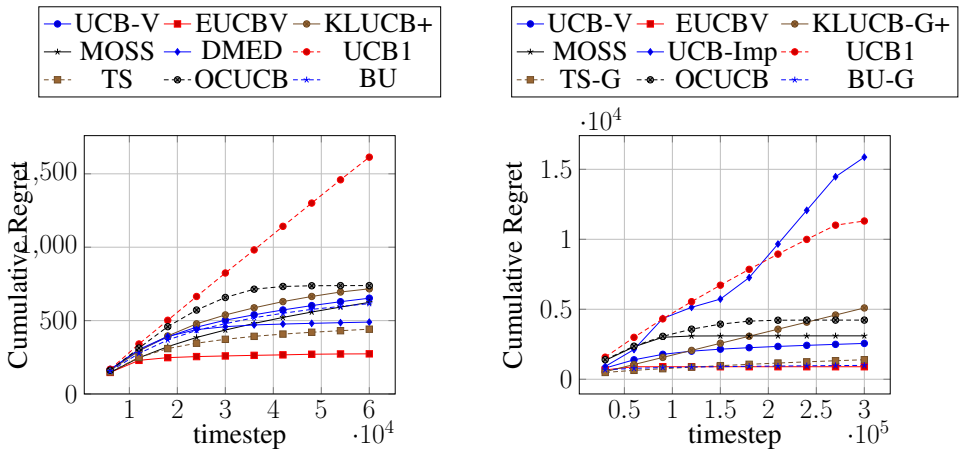
$$
\begin{aligned}
\mathbb{E}[R_T] \leq &\sum_{i\in\mathcal{A}:\Delta_i>b} \left\{\frac{C_0K^4}{T^{\frac{1}{4}}} + \left(\Delta_i + \frac{320\sigma_i^2\log(\frac{T\Delta_i^2}{K})}{\Delta_i}\right)\right\} \\
&+ \sum_{i\in\mathcal{A}:0<\Delta_i\leq b} \frac{C_2K^4}{T^{\frac{1}{4}}} + \max_{i\in\mathcal{A}:0<\Delta_i\leq b} \Delta_i T
\end{aligned}
$$

where C_0, C_1, C_2 *are integer constants s.t.* $C_0 = C_1 + C_2$*.*

3.5 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 Cappé, 2011), DMED (Honda and Takemura, 2010), MOSS (Audibert and Bubeck, 2009), UCB1 (Auer *et al.*, 2002*a*), UCB-Improved (Auer and

Ortner, 2010), Median Elimination (Even-Dar *et al.*, 2006), Thompson Sampling (TS) (Agrawal and Goyal, 2011), OCUCB (Lattimore, 2015), Bayes-UCB (BU) (Kaufmann *et al.*, 2012) and UCB-V (Audibert *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 1). Note that KLUCB+ empirically outperforms KLUCB (as shown in Garivier and Cappé (2011)).



(a) Expt-1: 20 Bernoulli-distributed arms (b) Expt-2: 3 Group Mean Setting

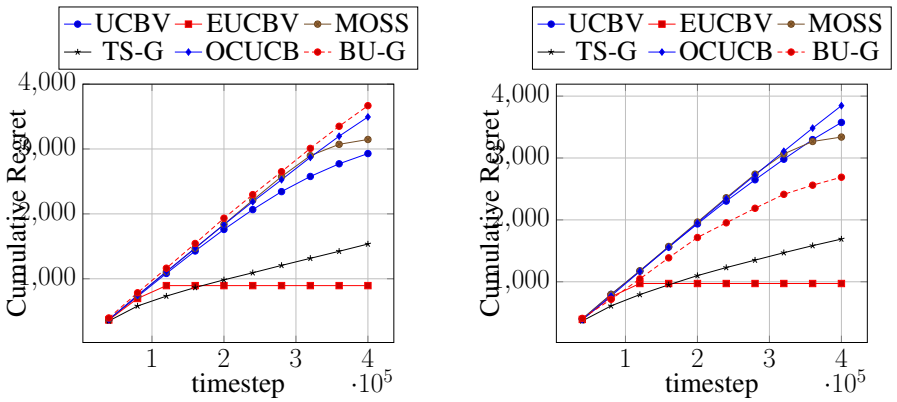
Figure 3.1: A comparison of the cumulative regret incurred by the various bandit algorithms.

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. 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 Cappé (2011)). The regret is averaged over 100 independent runs and is shown in Figure 3.1(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 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 for this test-case.

Experiment-2 (Gaussian 3 Group Mean Setting): This experiment is conducted to observe the performance of EUCBV over a large horizon in Gaussian distribution

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

testbed. 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 $\sigma^2_{1:66} = 0.01, \sigma^2_{67:99} = 0.25$ and $\sigma^2_{100} = 0.25$. The regret is averaged over 100 independent runs and is shown in Figure 3.1(b). From the results in Figure 3.1(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 3.1(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) of Figure 3.1(b) is 10^4. 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.



(a) Expt-3: Failure of TS (b) Expt-4: 3 Group Variance Setting

Figure 3.2: Further Experiments with EUCBV

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 $\sigma^2_{1:10} = 0.01$, $\sigma^2_{100} = 0.25$ 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 3.2(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 $\Omega\left(\sqrt{KT\log T}\right)$.

Experiment-4 (Gaussian 3 Group Variance setting): This experiment is conducted to show that when the gaps are uniform and variance of the arms are 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 $\sigma^2_* = 0.25$. 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 3.2(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 EUCBV significantly outperforms all the Bayesian algorithms initialized with Gaussian priors.

3.6 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 $O\left(\sqrt{KT}\right)$. 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 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.

3.7 Summary

In this chapter we looked at 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 *et al.*, 2009). Through a theoretical analysis we establish that EUCBV incurs a *gap-dependent* regret bound of $o\left(\frac{K\sigma^2_{\max}\log(T\Delta^2/K)}{\Delta}\right)$ after T trials, 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 (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.

Appendix A

APPENDIX

A.1 Appendix for EUCBV

A.1.1 Proof of Lemma 1

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

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

Proof 3 *The proof is based on contradiction. Suppose*

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

Then, with $\psi = \dfrac{T}{K^2}$ *and* $\rho = \dfrac{1}{2}$, *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* $m \leq \dfrac{1}{2}\log_2\left(\dfrac{T}{e}\right)$, *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.* ■

A.1.2 Proof of Lemma 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* $c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$, *then,*

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

Proof 4 *In the* m_i*-th round since* $z_i \geq n_{m_i}$, *by substituting* z_i *with* n_{m_i} *we can show that,*

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

In the above simplification, (a) *is due to* $\hat{v}_i \in [0,1]$, *while* (b) *is obtained using Lemma 1.*

■

A.1.3 Proof of Lemma 3

Lemma 3 *If* $m_i = min\{m|\sqrt{4\epsilon_m} < \frac{\Delta_i}{4}\}$, $c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$ *and* $n_{m_i} = \frac{\log(\psi T\epsilon_{m_i})}{2\epsilon_{m_i}}$ *then we can show that,*

$$\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \frac{2}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}}.$$

Proof 5 *We start by recalling from equation (3.2) that,*

$$\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right) \tag{A.1}$$

where

$$c_i = \sqrt{\frac{\rho(\hat{v}_i + 2)\log(\psi T \epsilon_{m_i})}{4z_i}} \text{ and}$$
$$\bar{c}_i = \sqrt{\frac{\rho(\sigma_i^2 + \sqrt{\epsilon_{m_i}} + 2)\log(\psi T \epsilon_{m_i})}{4z_i}}.$$

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

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

The first term in the LHS of (A.1) can be bounded using the Bernstein inequality as below:

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

where, (a) *is obtained by substituting equation A.2 and* (b) *occurs because for all* $\sigma_i^2 \in [0, \frac{1}{4}]$, $\left(\frac{3\sigma_i^2 + 3\sqrt{\epsilon_{m_i}} + 6}{6\sigma_i^2 + 2\sqrt{\epsilon_{m_i}}}\right) \geq \frac{3}{2}$.

The second term in the LHS of (A.1) can be simplified as follows:

$$\begin{aligned}
&\mathbb{P}\left\{\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right\} \\
&\leq \mathbb{P}\left\{\frac{1}{n_i}\sum_{t=1}^{n_i}(X_{i,t} - r_i)^2 - (\hat{r}_i - r_i)^2 \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right\} \\
&\leq \mathbb{P}\left\{\frac{\sum_{t=1}^{n_i}(X_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right\}
\end{aligned}$$

$$\overset{(a)}{\leq} \mathbb{P}\left\{ \frac{\sum_{t=1}^{n_i}(X_{i,t} - r_i)^2}{n_i} \geq \sigma_i^2 + \bar{c}_i \right\}$$
$$\overset{(b)}{\leq} \exp\left(-\rho \left(\frac{3\sigma_i^2 + 3\sqrt{\epsilon_{m_i}} + 6}{6\sigma_i^2 + 2\sqrt{\epsilon_{m_i}}} \right) \log(\psi T \epsilon_{m_i}) \right) \leq \frac{1}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}} \tag{A.4}$$

where inequality (a) is obtained using (A.2), while (b) follows from the Bernstein inequality.

Thus, using (A.3) and (A.4) in (A.1) we obtain $\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \frac{2}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}}$. ■

A.1.4 Proof of Lemma 4

Lemma 4 *If* $m_i = min\{m | \sqrt{4\epsilon_m} < \frac{\Delta_i}{4}\}$, $\psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$, $c_i = \sqrt{\frac{\rho(\hat{v}_i + 2)\log(\psi T \epsilon_{m_i})}{4 z_i}}$ *and* $n_{m_i} = \frac{\log(\psi T \epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ *then in the m_i-th round,*

$$\mathbb{P}\{c^* > c_i\} \leq \frac{182 K^4}{T^{\frac{5}{4}} \sqrt{\epsilon_{m_i}}}.$$

Proof 6 *From the definition of c_i we know that $c_i \propto \frac{1}{z_i}$ as ψ and T are constants. Therefore in the m_i-th round,*

$$\mathbb{P}\{c^* > c_i\} \leq \mathbb{P}\{z^* < z_i\}$$
$$\leq \sum_{m=0}^{m_i} \sum_{z^*=1}^{n_m} \sum_{z_i=1}^{n_m} \left(\mathbb{P}\{\hat{r}^* < r^* - c^*\} + \mathbb{P}\{\hat{r}_i > r_i + c_i\} \right)$$

Now, applying Bernstein inequality and following the same way as in Lemma 3 we can show that,

$$\mathbb{P}\{\hat{r}^* < r^* - c^*\} \leq \exp(-\frac{(c^*)^2}{2\sigma_*^2 + \frac{2c^*}{3}} z^*) \leq \frac{4}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}}$$
$$\mathbb{P}\{\hat{r}_i > r_i + c_i\} \leq \exp(-\frac{(c_i)^2}{2\sigma_i^2 + \frac{2c_i}{3}} z_i) \leq \frac{4}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}}$$

Hence, summing everything up,

$$\mathbb{P}\{c^* > c_i\}$$

$$
\begin{aligned}
&\leq \sum_{m=0}^{m_i}\sum_{z^*=1}^{n_m}\sum_{z_i=1}^{n_m}\Bigg(\mathbb{P}\{\hat{r}^* < r^* - c^*\} + \mathbb{P}\{\hat{r}_i > r_i + c_i\}\Bigg) \\
&\overset{(a)}{\leq} \sum_{m=0}^{m_i}|B_m| n_m \Bigg(\mathbb{P}\{\hat{r}^* < r^* - c^*\} + \mathbb{P}\{\hat{r}_i > r_i + c_i\}\Bigg) \\
&\overset{(b)}{\leq} \sum_{m=0}^{m_i}\frac{4K}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}}\frac{\log(\psi T\epsilon_m^2)}{2\epsilon_m}\times \\
&\Bigg(\mathbb{P}\{\hat{r}^* < r^* - c^*\} + \mathbb{P}\{\hat{r}_i > r_i + c_i\}\Bigg) \\
&\overset{(c)}{\leq} \sum_{m=0}^{m_i}\frac{4K}{(\psi T\epsilon_m)^{\frac{3\rho}{2}}}\frac{\log(T)}{\epsilon_m}\left[\frac{4}{(\psi T\epsilon_m)^{\frac{3\rho}{2}}} + \frac{4}{(\psi T\epsilon_m)^{\frac{3\rho}{2}}}\right] \\
&\leq \sum_{m=0}^{m_i}\frac{32K\log T}{(\psi T\epsilon_m)^{3\rho}\epsilon_m} \leq \frac{32K\log T}{(\psi T)^{3\rho}}\sum_{m=0}^{m_i}\frac{1}{\epsilon_m^{3\rho+1}} \\
&\overset{(d)}{\leq} \sum_{m=0}^{m_i}\frac{32K\log T}{(\psi T)^{3\rho}}\left(\sum_{m=0}^{m_i}\frac{1}{\epsilon_m}\right)^{3\rho+1} \\
&\overset{(e)}{\leq} \frac{32K\log T}{(\frac{T^2}{K^2})^{\frac{3}{2}}}\left[\left(1 + \frac{2(2^{\frac{1}{2}\log_2\frac{T}{e}} - 1)}{2-1}\right)^{\frac{5}{2}}\right] \\
&\leq \frac{182K^4T^{\frac{5}{4}}\log T}{T^3} \overset{(f)}{\leq} \frac{182K^4}{T^{\frac{5}{4}}} \overset{(g)}{\leq} \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}
\end{aligned}
$$

where, (a) comes from the total pulls allocated for all $i \in B_m$ till the m-th round, in (b) the arm count $|B_m|$ can be bounded by using equation (3.3) and then we substitute the value of n_m, (c) happens by substituting the value of ψ and considering $\epsilon_m \in [\sqrt{\frac{e}{T}}, 1]$, (d) follows as $\frac{1}{\epsilon_m} \geq 1, \forall m$, in (e) we use the standard geometric progression formula and then we substitute the values of ρ and ψ, (f) follows from the inequality $\log T \leq \sqrt{T}$ and (g) is valid for any $\epsilon_{m_i} \in [\sqrt{\frac{e}{T}}, 1]$.

■

A.1.5 Proof of Lemma 5

Lemma 5 *If $m_i = min\{m|\sqrt{4\epsilon_m} < \frac{\Delta_i}{4}\}$, $\psi = \frac{T}{K^2}$, $\rho = \frac{1}{2}$, $c_i = \sqrt{\frac{\rho(\hat{v}_i+2)\log(\psi T\epsilon_{m_i})}{4z_i}}$ and $n_{m_i} = \frac{\log(\psi T\epsilon_{m_i}^2)}{2\epsilon_{m_i}}$ then in the m_i-th round,*

$$\mathbb{P}\{z_i < n_{m_i}\} \leq \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}.$$

Proof 7 *Following a similar argument as in Lemma 4, we can show that in the m_i-th round,*

$$\begin{aligned}
&\mathbb{P}\{z_i < n_{m_i}\} \\
&\leq \sum_{m=0}^{m_i} \sum_{z_i=1}^{n_m} \sum_{z^*=1}^{n_m} \left(\mathbb{P}\{\hat{r}^* > r^* - c^*\} + \mathbb{P}\{\hat{r}_i < r_i + c_i\} \right) \\
&\leq \frac{32K \log T}{(\psi T)^{3\rho}} \sum_{m=0}^{m_i} \frac{1}{\epsilon_m^{3\rho+1}} \leq \frac{182K^4}{T^{\frac{5}{4}} \sqrt{\epsilon_{m_i}}}.
\end{aligned}$$

■

A.1.6 Proof of Lemma 6

Lemma 6 *For two integer constants c_1 and c_2, if $20c_1 \leq c_2$ then,*

$$c_1 \frac{4\sigma_i^2 + 4}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right) \leq c_2 \frac{\sigma_i^2}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right).$$

Proof 8 *We again prove this by contradiction. Suppose,*

$$c_1 \frac{4\sigma_i^2 + 4}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right) > c_2 \frac{\sigma_i^2}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right).$$

Further reducing the above two terms we can show that,

$$\begin{aligned}
&4c_1\sigma_i^2 + 4c_1 > c_2\sigma_i^2 \\
&\Rightarrow 4c_1.\frac{1}{4} + 4c_1 \overset{(a)}{>} \frac{c_2}{4} \\
&\Rightarrow 20c_1 > c_2.
\end{aligned}$$

Here, (a) occurs because $0 \leq \sigma_i^2 \leq \frac{1}{4}, \forall i \in \mathcal{A}$. But, we already know that $20c_1 \leq c_2$. Hence,

$$c_1 \frac{4\sigma_i^2 + 4}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right) \leq c_2 \frac{\sigma_i^2}{\Delta_i} \log\left(\frac{T\Delta_i^2}{K}\right).$$

■

A.1.7 Proof of Corollary 1

Corollary 1 ***(Gap-Independent Bound)*** *When the gaps of all the sub-optimal arms are identical, i.e.,* $\Delta_i = \Delta = \sqrt{\frac{K \log K}{T}} > \sqrt{\frac{e}{T}}, \forall i \in \mathcal{A}$ *and* C_3 *being an integer constant, the regret of EUCBV is upper bounded by the following gap-independent expression:*

$$\mathbb{E}[R_T] \leq \frac{C_3 K^5}{T^{\frac{1}{4}}} + 320\sqrt{KT}.$$

Proof 9 *From Bubeck* et al. *(2011) 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* $C = \left\lfloor \frac{T}{e} \right\rfloor$ *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 2 below:

$$\mathbb{E}[R_T] \leq \sum_{i \in \mathcal{A}: \Delta_i > b} \left\{ \frac{C_0 K^4}{T^{\frac{1}{4}}} + \left(\Delta_i + \frac{320\sigma_i^2 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \right) \right\} + \sum_{i \in \mathcal{A}: 0 < \Delta_i \leq b} \frac{C_2 K^4}{T^{\frac{1}{4}}} + \max_{i \in \mathcal{A}: 0 < \Delta_i \leq b} \Delta_i T.$$

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

$$\sum_{i \in \mathcal{A}: \Delta_i > b} \frac{320\sigma_i^2 \log\left(\frac{T\Delta_i^2}{K}\right)}{\Delta_i} \leq \frac{320\sigma_{\max}^2 K\sqrt{T} \log\left(T\frac{K(\log K)}{TK}\right)}{\sqrt{K \log K}}$$
$$\leq \frac{320\sigma_{\max}^2 \sqrt{KT} \log(\log K)}{\sqrt{\log K}} \overset{(a)}{\leq} 320\sigma_{\max}^2 \sqrt{KT}$$

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

$$\mathbb{E}[R_T] \overset{(a)}{\leq} \frac{C_3 K^5}{T^{\frac{1}{4}}} + 320\sigma_{\max}^2 \sqrt{KT}$$
$$\overset{(b)}{\leq} \frac{C_3 K^5}{T^{\frac{1}{4}}} + 320\sqrt{KT}$$

where, in(a)*,* C_3 *is an integer constant such that* $C_3 = C_0 + C_2$ *and* (b) *occurs because* $\sigma_i^2 \in [0, \frac{1}{4}], \forall i \in \mathcal{A}$*.*

■

Bibliography

1. **Abernethy, J. D.**, **K. Amin**, and **R. Zhu**, Threshold bandits, with and without censored feedback. *In Advances In Neural Information Processing Systems*. 2016.

2. **Agrawal, S.** and **N. Goyal** (2011). Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*.

3. **Audibert, J.-Y.** and **S. Bubeck**, Minimax policies for adversarial and stochastic bandits. *In COLT*. 2009.

4. **Audibert, J.-Y.** and **S. Bubeck**, Best arm identification in multi-armed bandits. *In COLT-23th Conference on Learning Theory-2010*. 2010.

5. **Audibert, J.-Y.**, **R. Munos**, and **C. Szepesvári** (2009). Exploration–exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, **410**(19), 1876–1902.

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

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

8. **Auer, P.** and **R. Ortner** (2010). Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, **61**(1-2), 55–65.

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

10. **Bubeck, S.**, **R. Munos**, and **G. Stoltz** (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, **412**(19), 1832–1852.

11. **Bubeck, S.**, **T. Wang**, and **N. Viswanathan**, Multiple identifications in multi-armed bandits. *In ICML (1)*. 2013.

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

13. **Chen, S.**, **T. Lin**, **I. King**, **M. R. Lyu**, and **W. Chen**, Combinatorial pure exploration of multi-armed bandits. *In Advances in Neural Information Processing Systems*. 2014.

14. **Even-Dar, E.**, **S. Mannor**, and **Y. Mansour** (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, **7**, 1079–1105.

15. **Gabillon, V.**, **M. Ghavamzadeh**, and **A. Lazaric**, Best arm identification: A unified approach to fixed budget and fixed confidence. *In Advances in Neural Information Processing Systems*. 2012.

16. **Gabillon, V.**, **M. Ghavamzadeh**, **A. Lazaric**, and **S. Bubeck**, Multi-bandit best arm identification. *In Advances in Neural Information Processing Systems*. 2011.

17. **Garivier, A.** and **O. Cappé** (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. *arXiv preprint arXiv:1102.2490*.

18. **Honda, J.** and **A. Takemura**, An asymptotically optimal bandit algorithm for bounded support models. *In COLT*. Citeseer, 2010.

19. **Jamieson, K.** and **R. Nowak**, Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. *In Information Sciences and Systems (CISS), 2014 48th Annual Conference on*. IEEE, 2014.

20. **Kalyanakrishnan, S.**, **A. Tewari**, **P. Auer**, and **P. Stone**, Pac subset selection in stochastic multi-armed bandits. *In Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 2012.

21. **Kaufmann, E.**, **O. Cappé**, and **A. Garivier**, On bayesian upper confidence bounds for bandit problems. *In AISTATS*. 2012.

22. **Lai, T. L.** and **H. Robbins** (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, **6**(1), 4–22.

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

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

25. **Locatelli, A.**, **M. Gutzeit**, and **A. Carpentier** (2016). An optimal algorithm for the thresholding bandit problem. *arXiv preprint arXiv:1605.08671*.

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

27. **Thompson, W. R.** (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 285–294.

LIST OF PAPERS BASED ON THESIS

1. Authors.... Title... *Journal*, Volume, Page, (year).