Finite-time Analysis of Frequentist Strategies for Multi-armed Bandits

A THESIS

submitted by

SUBHOJYOTI MUKHERJEE

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

January 2018

THESIS CERTIFICATE

This is to certify that the thesis titled **Finite-time Analysis of Frequentist Strategies for Multi-armed Bandits**, submitted by **Subhojyoti Mukherjee**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Science (Research)**, 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.

Dr. Balaraman Ravindran
Research Guide
Associate Professor
Dept. of Computer Science
IIT-Madras, 600 036

Dr. Nandan Sudarsanam
Research Co-Guide
Assistant Professor
Dept. of Management Studies
IIT-Madras, 600 036

Place: Chennai

Date: January 12, 2018

ACKNOWLEDGEMENTS

I am grateful to my primary thesis advisor Dr. Balaraman Ravindran from Computer Science and Engineering Department at IIT Madras. Without his constant guidance, motivation and perseverance with me, none of these works would have been completed. His courses of Introduction to Machine Learning and Introduction to Reinforcement Learning were my first step into the research world. Not only these courses carried a wealth of information on the current state of literature in the respective fields but they used to be hugely competitive which used to constantly motivate me to strive for excellence. Apart from our formal meetings every week, I used to catch him in his office or in the corridors or anywhere in the department whenever I had any doubts and we used to converse on them. One of his greatest influences on me is to inculcate the constant urge to collaborate with as many interested people as possible, both inside and outside of my immediate field of research. This became a huge contributing factor in the research that I conducted for this thesis.

I am also grateful to my co-advisor Dr. Nandan Sudarsanam from Department of Management Studies at IIT Madras. It is from him that I first learned the most important factor of publication, how to properly write research papers in a concise and crisp manner. I fondly remember that how intensely he used to scrutinize my drafts for errors and incoherent ideas. His dedication and trust in my work are evident from this one incident during the ICML 2017 submission, when at 2 am in the night he called me up to correct portions of my draft. Also, his courses in Department of Management Studies are very enjoyable to attend and are very informative. Nobody can make you understand a complex idea in a simple way and yet retain all its subtleties as Nandan sir can and his courses used to reflect this idea.

Another person who had a profound influence on me is Dr. K.P. Naveen from Electrical Engineering Department at IIT Tirupati who advised me on a significant portion of this thesis. I first attended one of his courses on Introduction to Probability Theory

here at IIT Madras (when he was INSPIRE Faculty member here) and instantly liked the way he taught the complex idea of Probability theory in simple terms. Even though the area of Multi-armed bandits is not his core research area, yet we collaborated on this topic which led to multiple publications. I fondly remember his rigor and patience with me while correcting the theoretical portions of my draft, the most important section of any bandit research paper.

I am also grateful to Dr. L.A. Prashanth from Computer Science and Engineering Department at IIT Madras in my initial days of research for advising me on a significant portion of stochastic multi-armed bandits. He is up-to-date with all the research work in this area and my interactions with him always led to new ideas and a broader way to look at a problem. All collaborations do not always lead to publications and I hope that our future interactions will lead to a more fruitful outcome.

While writing this thesis I must also acknowledge the contributions of Dr. Odalric-Ambrym Maillard from INRIA, SequeL Lab at Lille, France where I went for winter internship from September to December 2017. He is one of the most inspiring researchers that I met in my foray into the research world. He was always available at his office, always working on interesting ideas and always interested to listen to any idea that I came up with. The best thing about him that I can fondly relate to is that he never rejected any idea even though they were initially outrageous, but used to guide me slowly towards the reason on why they would fail in the long run. Similarly, I must also thank Dr. Branislav Kveton from Adobe Research, San Jose, whom I met on the sidelines of IJCAI 2017 who gave me a lot of feedback on how to improve my work on Thresholding Bandits.

I am thankful to the other members of my committee for their constant support and guidance. I am especially thankful to Dr. Sutanu Chakraborti for his trust on me. I fondly remember that whenever we used to meet we used to converse in Bengali and he used to constantly motivate me to work more. His course on Natural Language Processing is a very enjoyable course and the project on humor generation that I did under his supervision is one of most interesting projects I did in my IIT Madras coursework. I am also thankful to my Dr. Nirav P Bhatt and my committee chair Dr. P Sreenivasa Kumar for being always available for me.

I am thankful towards a host of my colleagues in the RISE Lab at Computer Science and Engineering Department where I spent a significant time while conducting research at IIT Madras. Thanks Patanjali for creating such an environment in the lab, conducive to research. I also thank my first roommate Priyesh for making my first semester so livable on the campus. In the course of time, I made more friends who contributed in making my life at IIT Madras memorable; Shashank, Siddharth, Deepak, Tarun, Madhuri, Ditty, Ahana, without you people life at IIT Madras would have been boring. There are several members of IPCV lab to whom I am grateful for making my stay at IIT Madras very memorable especially I must thank Nimisha, Abhijith, and Karthik for being such wonderful friends. I am also grateful towards my father, mother, brother, and sister-in-law Sarah for providing me constant motivation in my research and personal life.

Last but not the least, I must thank Dr. A.P. Vijay Rengarajan from Electrical Engineering Department at IIT Madras. He was also conducting research for his doctoral studies in the IPCV lab and it his from him that I learned the true meaning of dedication towards research. This dedication of him, of course, showed up in all the publications he got but most of all he was always there, listening to all my concerns on a host of issues that I faced in my research and personal life. Vijay without you life at IIT Madras would not have been same.

Finally, I must acknowledge the influence of Dr. Sven Koenig whom I met on the sidelines of IJCAI 2017 in the "Lunch with A Fellow" event. That four-hour interaction with him is a memory I cherish because we talked about every possible aspect of a researcher's life and how to conduct good research. In the end, he said to me the three mantras of conducting good research; shed ego, collaborate and learn something new every day.

ABSTRACT

KEYWORDS: Reinforcement Learning, Stochastic Bandits, UCB, UCBV, EUCBV, Thresholding Bandits, APT, AugUCB

This thesis studies the following topics in the area of Reinforcement Learning: Multi-armed bandits in stationary distribution with the goal of cumulative regret minimization and Thresholding bandits in pure exploration setting. The common underlying theme is the study of bandit theory and its application in various types of environments. In the first part of the thesis, we study the classic multi-armed bandit problem with a stationary distribution, one of the first settings studied by the bandit community and which successively gave rise to several new directions in bandit theory. We propose a novel algorithm in this setting and compare both theoretically and empirically its performance against the available algorithms. Our proposed algorithm termed as Efficient-UCB-Variance (EUCBV) is the first arm-elimination algorithm which uses variance estimation to eliminate arms as well as achieve an order optimal regret bound. Empirically, we show that EUCBV outperforms most of the state-of-the-art algorithms in the considered environments. In the next part, we study a specific type of stochastic multi-armed bandit setup called the thresholding bandit problem and discuss its usage, available state-of-the-art algorithms on this setting and our solution to this problem. We propose the Augmented-UCB (AugUCB) algorithm which again uses variance and mean estimation along with arm elimination technique to conduct exploration. We give theoretical guarantees on the expected loss of our algorithm and also analyze its performance against state-of-the-art algorithms in numerical simulations in multiple synthetic environments.

Contents

List of Tables

List of Figures

ABBREVIATIONS

APT	Anytime Parameter-free Thresholding
AugUCB	Augmented Upper Confidence Bound
BU	Bayes-UCB
CCB	Combined Confidence Bound
CSAR	Combinatorial Successive Accept Reject
DMED	Deterministic Minimum Empirical Divergence
DTS	Discounted Thompson Sampling
DUCB	Discounted UCB
EUCBV	Efficient Upper Confidence Bound Variance
EXP3	Exponential Exploration Exploitation
EXP3IX	EXP3 Implicit Exploration
GCTS	Global Change-ponit Thompson Sampling
KL	Kullback–Leibler
KLUCB	Kullback–Leibler Upper Confidence Bound
LUCB	Lower Upper Confidence Bound
MAB	Multi-armed Bandit
MDP	Markov Decision Process
MOSS	Minimax Optimal Strategy in the Stochastic case
OCUCB	Optimally Confident Upper Confidence Bound
RL	Reinforcement Learning
RExp3	Restarting EXP3
SAR	Successive Accept Reject
SMAB	Stochastic Multi-armed Bandit
SR	Successive Reject
SW-UCB	Switching Window UCB
TBP	Thresholding Bandit Problem
TS	Thompson Sampling

UA	Uniform Allocation
UCB	Upper Confidence Bound
UCBE	Upper Confidence Bound Exploration
UCBEV	Upper Confidence Bound Exploration Variance
UCBV	Upper Confidence Bound Variance

NOTATION

T	Time horizon
$\mathcal{A}$	Set of arms
K	Total number of arms
π	Policy of an agent
D_i	Reward distribution of i-th arm
r_i	Expected mean of D_i for the i-th arm
$\hat{r}_i$	Sample mean of the i-th arm
$X_{i,t}$	Reward obtained for the i-th arm at the t-th timestep
z_i	Number of times arm i is pulled
σ_i	Standard Deviation of the i-th arm
v_i	Variance of the i-th arm
Δ_i	Gap of the i-th arm
Δ	Minimal gap amongst all arms
τ	Threshold
H_1	Hardness with Mean-1
H_2	Hardness with Mean-2
$H_{\sigma,1}$	Hardness with Mean and Standard Deviation-1
$H_{\sigma,2}$	Hardness with Mean and Standard Deviation-2
ρ	Parameter for tuning length of of confidence interval
ψ	Exploration parameter
m	Round number
R_T	Cumulative regret till horizon T
SR_t	Simple regret at t-th timestep

Chapter 1

Introduction to Bandits

1.1 Reinforcement Learning

In today's world, artificial intelligence has proved to be a game-changer in designing agents that interact with an evolving environment and make decisions on the fly. The main goal of artificial intelligence is to design artificial agents that make dynamic decisions in an evolving environment. In pursuit of these, the agent can be thought of as making a series of sequential decisions by interacting with the dynamic environment which provides it with some sort of feedback after every decision which the agent incorporates into its decision-making strategy to formulate the next decision to be made. A large number of problems in science and engineering, robotics and game playing, resource management, financial portfolio management, medical treatment design, ad placement, website optimization and packet routing can be modeled as sequential decision-making under uncertainty. Many of these real-world interesting sequential decision-making problems can be formulated as reinforcement learning (RL) problems ((Bertsekas and Tsitsiklis, 1996), (Sutton and Barto, 1998)). In an RL problem, an agent interacts with a dynamic, stochastic, and unknown environment, with the goal of finding an action-selection strategy or policy that optimizes some long-term performance measure. Every time when the agent interacts with the environment it receives a signal/reward from the environment based on which it modifies its policy. The agent learns to optimize the choice of actions over several time steps which is learned from the sequences of data that it receives from the environment. This is the crux of online sequential learning.

This is in contrast to supervised learning methods that deal with labeled data which are independently and identically distributed (i.i.d.) samples from the considered domain and train some classifier on the entire training dataset to learn the pattern of this distribution to predict the labels of future samples (test dataset) with the assumption

that it is sampled from the same domain. In contrast to this, an RL agent learns from the samples that are collected from the trajectories generated by its sequential interaction with the system. For an RL agent, the trajectory consists of a series of sequential interactions whereby it transitions from one state to another following some dynamics intrinsic to the environment while collecting the reward till some stopping condition is reached. This is known as an episode. Here, for an action i_t taken by the agent at the t-th timestep, the agent transitions from its current state denoted by $S_{i,t}$ to state $S_{i,t+1}$ and observes the reward $X_{i,t}$. An illustrative image depicting the reinforcement learning scenario is shown in Figure 1.1.

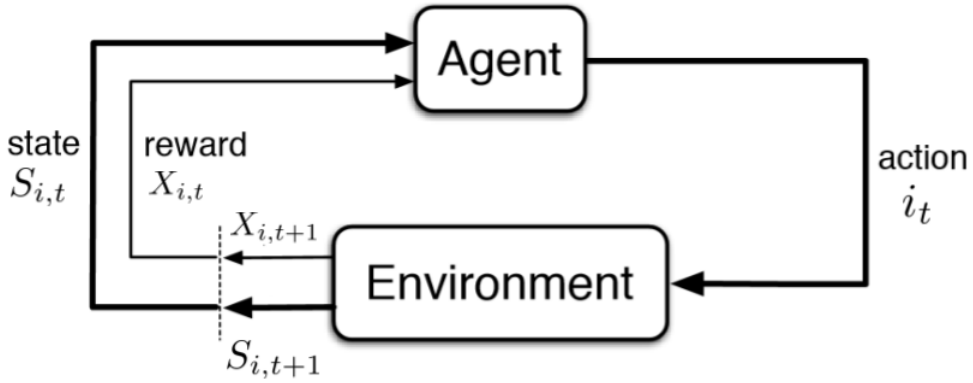


Figure 1.1: Reinforcement Learning

1.2 Connection between Reinforcement Learning and Bandits

As defined in the previous section, an episode consists of a series of sequential interaction whereas agent transitions from one state to another based on some intrinsic dynamics of the environment while collecting the rewards and choosing actions based on some action-selecting strategy. The MAB model can be considered as a special case where there is a single looping state and the agent after taking an action and observing the reward transitions back to the same state. That single looping state consists of several finite number of actions which are called as arms.

The name bandit originated from the concept of casino slot machine where there are levers which are called as arms and the learner can pull one lever and observe the

reward associated with that arm which is sampled from a distribution associated with the specific arm. This game is repeated T times and the goal of the learner is to maximize its profit.

1.3 Why study Bandits?

There are multiple reasons to study the interesting area of bandits. First of all, bandits are the cornerstone of understanding the general reinforcement learning area. In fact, bandits help us to understand the idea of *exploration-exploitation* dilemma which is the basis to build full, multi-state, general reinforcement learning ideas. Secondly, as stated in Maillard (2011), even 50 years after Robbins (1952) introduced the first idea of bandits, there are many interesting and fruitful areas where bandit concept can be extended in both practical and theoretical terms. Finally, there are several real-life industrial applications ranging from recommendation systems, game theory to anomaly detection where bandit applications have been found to perform exceptionally well. All of these forces us to delve deep into a systemic research of bandits.

1.4 Motivation

The MAB model fits very well in various real-world scenarios that can be modeled as decision-making under uncertainty problems. Some of which are as follows:-

1. *Online Shop Domain:* In the online shop domain (Ghavamzadeh *et al.*, 2015), a retailer aims to maximize profit by sequentially suggesting products to online shopping customers. In this scenario, at every timestep, the retailer displays an item to a customer from a pool of items which has the highest probability of being selected by the customer. The one-step interaction ends when the customer selects or does not select a product (which will be considered as a loss to the retailer). This feedback is incorporated by the learner as a feedback from the environment and it modifies its policy for the next suggestion. This process is repeated till a pre-specified number of times with the retailer gathering valuable information regarding the customer from this behaviour and modifying its policy to display other items to different customers. In its simplest form, this can be modeled as a stochastic MAB problem which is studied in the first part of the thesis.
2. *Medical Treatment Design:* Another interesting domain that MAB model was first studied was for the medical treatment design (Thompson, 1933),(Thompson,

1935). Here at every timestep, the learner chooses to administer one out of several treatments sequentially on a stream of patients who are suffering from the same ailment (say). Let's also assume that there is a single treatment which will be able to alleviate the patients from their disease. Here, the one-step interaction ends when the patient responds well or does not respond well to the treatment whereby the learner modifies its policy for the suggestion to the next patient. The goal of the learner is to quickly converge on the best treatment so that whenever a new patient comes with the same ailment, the learner can suggest the best treatment which can relieve the patient of its ailment with a high probability. This problem can also be modeled as a stochastic MAB problem.

3. *Financial Portfolio Management:* In financial portfolio management MAB models can also be used. Here, the learner is faced with the choice of selecting the most profitable stock option out of several stock options. The simplest strategy where we can employ a bandit model is this; at the start of every trading session, the learner suggests a stock to purchase worth Re 1, while at the closing of the trading session it sells off the stock to witness its value after a day's trading. The profit recorded is treated as the reward revealed by the environment and the learner modifies its policy for the next day. Let's assume that no new stock options are being introduced over the considered time horizon and there is a single best stock option which if selected in perpetuity will always give the best returns. Then, the goal of the learner is reduced to identifying the best stock option as quickly as possible. This is another interesting variation which can be modeled as a stochastic MAB problem.

4. *Product Selection:* A company wants to introduce a new product in the market and there is a clear separation of the test phase from the commercialization phase. In this case, the company tries to minimize the loss it might incur in the commercialization phase by testing as much as possible in the test phase. So from the several variants of the product that is in the test phase, the learning learner must suggest the product variant(s) whose qualities are above a particular threshold τ at the end of the test phase that has the highest probability of minimizing the loss in the commercialization phase. A similar problem has been discussed for single best product variant identification without threshold in Bubeck *et al.* (2011). This problem can be modeled as a stochastic thresholding MAB problem which is studied in the second part of the thesis.

5. *Mobile Phone Channel Allocation:* Another similar problem as above concerns channel allocation for mobile phone communications (Audibert *et al.*, 2009). Here there is a clear separation between the allocation phase and communication phase whereby in the allocation phase a learner has to explore as many channels as possible to suggest the best possible set of channel(s) whose qualities are above a particular threshold τ. The threshold may depend on the subscription level of the customer such that with the higher subscription the customer is allowed better channel(s) with the τ set high. Each evaluation of a channel is noisy and the learning algorithm must come up with the best possible set of suggestions within a very small number of attempts. This setting can also be modeled as a stochastic thresholding MAB problem.

6. *Anomaly Detection and Classification:* MABs can also be used for anomaly detection where the goal is to seek out extreme values in the data. Anomalies may not always be naturally concentrated which was shown in Steinwart *et al.* (2005). To implement a MAB model the best possible way is to define a cut-off level τ and classify the samples above this level τ as anomalous along with a tolerance factor which gives it a degree of flexibility. Such an approach has already been mentioned in Streeter and Smith (2006) and further studied in Locatelli *et al.* (2016). Finally, this is also an interesting variation which can be modeled as a stochastic thresholding MAB problem.

1.5 Types of Information Feedback

In an online sequential setting, the feedback that the learner receives from the environment can be characterized into three broad categories, full information feedback, partial information feedback and bandit feedback.

To illustrate the different types of feedback we will take help of the following example. Let a learner be given a set of finite actions $i \in \mathcal{A}$ such that $|\mathcal{A}| = K$. Let, the environment be such that each action has a probability distribution D_i attached to it which is fixed throughout the time horizon T. The learning proceeds as follows, at every timestep the learner selects $m \in \mathcal{A}$ actions and observes some form of feedback vector R_t^{obs} (which will be characterized later). Before the learner selects the set of arms the environment draws the feedback vector $F_t^{env} \in [0,1]^K$ of K i.i.d random rewards for all actions $i \in \mathcal{A}$ from $D_i, \forall i \in \mathcal{A}$ which it decides to reveal in particular format depending on the form of feedback chosen for the game, that is full information, partial information or bandit feedback. This game is shown in algorithm 1.

Algorithm 1 An online sequential game

Input: Time horizon T, K number of arms with unknown parameters of reward distribution

for each timestep $t = 1, 2, \ldots, T$ **do**
The environment chooses a reward vector $F_t^{env} = \left[r_{i,t} \sim^{i.i.d} D_i, \forall i \in \mathcal{A}\right]$.
The learner chooses m actions such that $m < K$ following some policy π, where $\mathcal{A}$ is the set of arms and $|\mathcal{A}| = K$.
The learner observes the reward vector $R_t^{obs} \subseteq F_t^{env}$.

end for

1.5.1 Full information feedback

In full information feedback, when a learner selects m actions then the environment reveals the rewards of all the actions $i \in \mathcal{A}$. Hence, in this form of feedback the learner observes $R_t^{obs} = F_t^{env} = [r_{i,t}, \forall i \in \mathcal{A}]$. This has been studied in many forms previously in Takimoto and Warmuth (2003), Kalai and Vempala (2005) or in online prediction with full information feedback in Cesa-Bianchi and Lugosi (2006).

1.5.2 Partial information feedback

In partial information feedback, when a learner selects m actions then the environment reveals the rewards of only those m actions for $m \in \mathcal{A}$. Hence, in this form of feedback the learner observes $R_t^{obs} = [r_{m,t}, \forall m \in \mathcal{A}]$. This is also sometimes called the semi-bandit feedback and has been studied in Awerbuch and Kleinberg (2004), McMahan and Blum (2004) and György *et al.* (2007).

1.5.3 Bandit feedback

In bandit feedback, when a learner selects m actions then the environment reveals a cumulative reward of those m actions for $m \in \mathcal{A}$. Hence, in this form of feedback the learner observes $R_t^{obs} = \sum_{q=1}^{m} r_{q,t}$. Note, that when $m = 1$, then the learner observes the reward of only that action that it has chosen out of K actions. Bandit feedback for single action has been extensively studied in literature with many open problems and we focus on its various interpretations in this thesis.

1.6 Different types of Bandits

In this section, we discuss the various types of bandits that are available in the literature.

1.6.1 Types of Bandits based on Environment

Stochastic Bandits

In stochastic bandits, the distribution associated with each of the arms remains fixed throughout the time horizon T. Some of the notable works associated with this type of setup are Robbins (1952), Lai and Robbins (1985), Agrawal (1995), Auer *et al.* (2002*a*), UCB-Improved Auer and Ortner (2010), Audibert and Bubeck (2009), Lattimore (2015), etc which gives us a broad class of algorithms suited for this setting. Chapter 2 and Chapter 3 is based on this setup where we discuss extensively on the latest state-of-the-art algorithms and discuss their empirical and theoretical performances.

Non-stochastic Bandits

In non-stochastic setting, the distribution associated with each arm varies over the duration of the play. Two notable examples of this are:-

1. **Adversarial bandits:** One of the first settings that has greatly motivated the studies in bandit literature is the *adversarial setting*. In this setting, at every timestep, an adversary chooses the reward for each arm and then the learner selects an arm without the knowledge of the adversary's choice. The adversary may or may not be oblivious to the learner's strategy and this forces the learner to employ a randomized algorithm to confuse the adversary. Previous works on this have focused on constructing different types of exponential weighting algorithms that are based on the Hedge algorithm that has been proposed before in Littlestone and Warmuth (1994),Freund and Schapire (1995) and analyzed in Auer *et al.* (2000). Further variants of this strategy called EXP3 (Auer *et al.*, 2002*b*), (Auer, 2002) and EXP3IX (Kocák *et al.*, 2014) have also been proposed which incorporates different strategies for exploration to minimize the loss of the learner.

2. **Piece-wise stationary:** Striding between the two contrasting settings of stochastic and adversarial bandits is the *piece-wise stochastic multi-armed bandit setting* where there are a finite number of changepoints when the distribution associated with each arm changes abruptly. Hence, this setting is neither as pessimistic as adversarial setting nor as optimistic as the stochastic setting. Therefore, the two broad class of algorithms mentioned before fail to perform optimally in this setting. Several interesting solutions have been proposed before for this setting which can be broadly divided into two categories, passively adaptive and actively adaptive strategies. The passively adaptive strategies like Discounted UCB (DUCB) (Kocsis and Szepesvári, 2006), Switching Window UCB (SW-UCB) (Garivier and Moulines, 2011) and Discounted Thompson Sampling (DTS) first

proposed in Raj and Kalyani (2017) do not actively try to locate the changepoints but rather try to minimize their losses by concentrating on past few observations. Similarly, algorithms like Restarting Exp3 (RExp3) (Besbes *et al.*, 2014) behave pessimistically as like Exp3 but restart after pre-determined phases. Hence, RExp3 can also be termed as a passively adaptive algorithm. On the other hand, actively adaptive strategies like Adapt-EVE (Hartland *et al.*, 2007), Windowed-Mean Shift (Yu and Mannor, 2009), EXP3.R (Allesiardo *et al.*, 2017), CUSUM-UCB (Liu *et al.*, 2017) try to locate the changepoints and restart the chosen bandit algorithms. Also, there are Bayesian strategies like Global Change-Point Thompson Sampling (GCTS)(Mellor and Shapiro, 2013) which uses Bayesian changepoint detection to locate the changepoints.

1.6.2 Types of Bandits based on goal

In bandit literature, based on the goal we can divide bandits into several categories. To illustrate this we put forward a simple scenario let us consider a stochastic bandit scenario where there are K arms labeled $i = 1, 2, \ldots, K$ with their expected means of reward distributions (D_i) be denoted by r_i. Also let there be single optimal arm $*$ such that $r^* = \max_{i \in \mathcal{A}} r_i$.

Cumulative regret minimization

In cumulative regret minimization the goal of the bandit is to minimize the cumulative regret which is the total loss suffered by the learner throughout the time horizon T for not choosing the optimal arm. Formally, we can define the cumulative regret as,

$$R_T = \sum_{t=1}^{T} r^* - \sum_{i \neq *} r_i n_{i,T} \tag{1.1}$$

where, $n_{i,T}$ is the number of times the learner has chosen arm i over the entire horizon T. We can further reduce equation 1.1 to obtain,

$$R_T = \sum_{t=1}^{T} r^* - \sum_{i \neq *} r_i n_{i,T} = \sum_{i=1}^{K} \Delta_i n_{i,T}$$

where $\Delta_i = r^* - r_i$ is called the gap between the optimal and the sub-optimal arm.

Simple regret minimization

In simple regret minimization the goal of the bandit is to minimize the instantaneous regret that is suffered at any timestep by the learner. Formally, the simple regret at t-th timestep where $J_n \in \mathcal{A}$ is the recommendation by the learner at timestep t is defined,

$$SR_t = r^* - r_{J_n} = \Delta_{J_n}$$

where Δ_{J_n} is the instantaneous gap between the expected mean of the optimal arm and the recommended arm by the learner. In the pure exploration setting the learner tries to minimize the simple regret and we study a very similar setting in chapter 4 and chapter 5.

Weak Regret minimization

In the non-stochastic scenario, when the distribution associated with each arm changes, the notion of regret is defined differently than cumulative regret. In this scenario, considering that there is a single best arm, the learner is more interested in minimizing the worst-case regret. Formally, for any sequence of actions $(j_1, \ldots, j_T)$ chosen by the learner over the time horizon T, the weak regret for single best action is defined as the difference between,

$$G_{\max}(j_1, \ldots, j_T) - G_\pi(T)$$

where, $G_{\max}(j_1, \ldots, j_T) = \max_{i \in \mathcal{A}} \sum_{t=1}^{T} X_{i_t}$ is the return of the globally best action over the entire horizon T, X_{i_t} is the reward observed for the i-th arm at the t-th timestep and $G_\pi(T)$ is the return following the policy π over the horizon T instead of choosing $j_1, \ldots, j_T$.

1.6.3 Contextual Bandits

Another interesting variation of the MAB model is the contextual bandit setup, where there are contexts or features associated with each arm. We can envision this with an example of online news article recommendation where there are users and articles and the goal of the learner is to map the correct article to a user so as to generate the user's interest. Following a similar work in Langford and Zhang (2007) this problem can be formulated as a contextual MAB problem such that at every timestep $t = 1, \ldots, T$

1. The learner observes an user u_t and the set of arms (articles) $i \in \mathcal{A}$ along with their feature vectors $v_{i,t}, \forall i \in \mathcal{A}_t$. This vector contains information about both the users and the arms and is referred as the context.

2. On the basis of previous trials the learner pulls an arm $i_t \in \mathcal{A}$ at the t-th timestep and observes the reward $X_{i,t}$ for only the arm $i_t \in \mathcal{A}$.

3. The algorithm then improves its prediction for the next trial with the new observation, $(v_{i,t}, i_t, X_{i,t})$.

This type of settings have been extensively studied in Li *et al.* (2010) and Beygelzimer *et al.* (2011).

1.6.4 Collaborative Bandits

Distributed bandits is a special setting where a network of bandits collaborates with each other to identify the best set of arms. The contextual MAB model discussed before naturally extends into this setting where a network of bandits try to map articles to a large number of users by collaborating between themselves (see Awerbuch and Kleinberg (2008); Liu and Zhao (2010); Szörényi *et al.* (2013); Hillel *et al.* (2013)). In this setting, bandits at the end of specific phases share information synchronously or asynchronously amongst each other to identify the best set of arms. Further, to learn more complicated structures and interaction between the user and article feature vectors, clustering can be used to cluster the articles and users based on their features and this has been studied in Bui *et al.* (2012), Cesa-Bianchi *et al.* (2013), Gentile *et al.* (2014).

1.6.5 Bandits with Corrupt Feedback

Another interesting area in the bandit setting is a variant of the stochastic MAB problem in which the rewards are corrupted. In certain recommender systems, it is sometimes vital to preserve the privacy of the users. Motivated by these, bandits with corrupt feedback assumes that the rewards it is receiving is corrupted by a stochastic corruption process of known parameters and the goal of the learner is again to maximize the reward by suggesting the best items to the users in this framework. This setting has been analyzed in Gajane *et al.* (2017).

1.6.6 Conservative Bandits

This setting is motivated by the scenario where there is one default safe arm which will always provide the learner with a good reward, but there are several unexplored arms which might provide the learner with better rewards if explored more. But the learner cannot do unconstrained exploration as its budget is limited and every time it pulls an arm it has to pay a cost. Hence, it must balance between pulling the safe arm and constrained exploration. This type of exploration under constraint has been termed as conservative bandits and is studied in Wu *et al.* (2016).

1.7 Objectives of Thesis

The main objectives of the thesis are as follows:-

1. The first objective of this thesis is to study the area of stochastic multi-armed bandit (SMAB) and how to minimize cumulative regret in this setup. We intend to give strong gap-dependent and gap-independent regret guarantees in the SMAB setting. We also intend to provide the algorithm in the SMAB setting that outperforms the current state-of-the-art algorithms in this setting.

2. The second objective of this thesis is to study the area of thresholding bandit problem (TBP) setting where the goal is to minimize the expected loss at the end of a fixed budget provided as input. We intend to provide strong guarantees with respect to expected loss and also propose the algorithm that does not require any problem complexity as an input. We also intend to provide strong empirical evaluations of the algorithm proposed for the TBP setting.

1.8 Contributions of Thesis

The main contributions of the thesis are as follows:-

1. We proposed a novel algorithm for the stochastic multi-armed bandit (MAB) problem. Our proposed Efficient UCB Variance method, referred to as EUCBV is an arm elimination algorithm based on UCB-Improved and UCBV strategy which takes into account the empirical variance of the arms and along with aggressive exploration factors eliminate sub-optimal arms. Through a theoretical analysis, we establish that EUCBV achieves a better gap-dependent regret upper bound than UCB-Improved, MOSS, UCB1, and UCBV algorithms. EUCBV enjoys an order optimal gap-independent regret bound same as that of OCUCB and MOSS, and better than UCB-Improved, UCB1 and UCBV. Empirically, in several considered environments EUCBV outperforms most of the state-of-the-art algorithms.

2. We proposed 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 quality 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. This is the first algorithm to employ such an approach for the considered TBP. Furthermore, in numerical evaluations, we establish in several considered environments that AugUCB outperforms all the algorithms that do not take into consideration the variance of the arms in their action selection strategy.

1.9 Outline of the Thesis

In this chapter, we gave an overview of the various types of bandits available in the literature and also discussed about the main objectives of the thesis and our contributions. In this section, we give a general outline of the thesis that is to follow. In chapter 2 we give a detailed overview of the stochastic multi-armed bandit model and the latest available algorithms in this setting. In the next chapter 3 we introduce our algorithm Efficient UCB Variance (EUCBV) for the stochastic multi-armed bandit model. We give theoretical guarantees on the performance of EUCBV and also show in numerical simulations that it indeed performs very well as compared to the state-of-the-art algorithms. In the subsequent chapter 4 we introduce a new variant of pure exploration multi-armed stochastic bandit called the thresholding bandit problem. We analyze the connections between thresholding bandit problem and pure exploration problem and also discuss

several existing algorithms in both the settings that are relevant to carefully analyze the thresholding bandit problem. Then in chapter 5 we introduce our solution for the thresholding bandit problem, called the Augmented UCB (AugUCB) algorithm. We analyze our algorithm AugUCB and derive theoretical guarantees for it as well as show in numerical experiments that it indeed outperforms several state-of-the-art algorithms in the thresholding bandit setting. Finally, in chapter 6 we conclude by briefly summarizing all the problems covered in the thesis and discussing some future directions in which the stated problems can be further extended.

Chapter 2

Stochastic Multi-armed Bandits

2.1 Introduction to SMAB

In this chapter, we deal with the stochastic multi-armed bandit (SMAB) 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. Also, note that in SMAB, the distribution associated with each arm is fixed throughout the entire duration of the horizon denoted by T. This SMAB formulation is shown in algorithm 2.

Algorithm 2 SMAB formulation

Input: Time horizon T, K number of arms with unknown parameters of reward distribution

for each timestep $t = 1, 2, \ldots, T$ **do**
The learner chooses an arm $i \in \mathcal{A}$, where $\mathcal{A}$ is the set of arms and $|\mathcal{A}| = K$.
The learner observes the reward $X_{i,t} \sim^{i.i.d} D_i$ where, D_i is the distribution associated with the arm i.

end for

The rest of the chapter is organized as follows. We specify all the notations and assumptions in section 2.2. Then we define the problem statement for the SMAB setting in section 2.3. In the next section 2.4 we discuss the motivations behind the SMAB setting. In section 2.5 we discuss extensively on the various state-of-the-art algorithms available for the SMAB setting. Finally, we summarize in section 2.6.

2.2 Notations and assumptions

Assumption 1 *In the considered SMAB setting we assume the optimal arm to be unique and it is denoted by* $*$.

Assumption 2 *We assume the rewards of all arms are bounded in* $[0, 1]$.

Notations: The mean of the reward distribution D_i 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 $|\mathcal{A}| = K$. We denote the individual arms labeled i, where $i = 1, \ldots, K$. 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 underlying time index t, for simplicity we neglect t from the notations and simply use $\hat{r}_i, \hat{v}_i$, and z_i to denote the respective quantities. Also, Δ denotes the minimum gap such that $\Delta = \min_{i \in \mathcal{A}} \{\Delta_i\}$.

2.3 Problem Definition

With the formulation of SMAB stated in algorithm 2, the learner seeks to identify the optimal arm as quickly as possible to maximize its rewards. In the pursuit of this, 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 (exploit) or explores arms more thoroughly to ensure that a correct decision is being made. This is termed as the *exploration-exploitation dilemma*, one of the fundamental challenges of reinforcement learning as discussed in chapter 1.

The objective of the learner in the SMAB setting is to maximize his rewards or in other words, to minimize the cumulative regret, which is defined as follows:

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

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,

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

where $\Delta_i = r^* - r_i$ is the gap between the means of the optimal arm and the i-th arm. In the theoretical analysis of each algorithm, we try to obtain bounds on this cumulative regret. These bounds can be both asymptotic or for a finite horizon. Again, these regret bounds can be either gap-dependent or gap-independent bounds.

1. **Asymptotic regret bounds:** These type of regret bounds are valid for a large horizon T tending to infinity. In other words, if the guarantees of these bounds to be held true then an infinite number of samples needs to be collected.

2. **Finite horizon regret bounds:** These type of regret bounds are valid for a finite horizon when a limited number of samples are allowed to be collected. Note, that the knowledge of horizon may or may not be known to the learner.

3. **Gap-Dependent regret bounds:** In gap-dependent or problem dependent regret bounds the regret is obtained as a measure of the gap $\Delta_i = r^* - r_i$ for an arm $i \in \mathcal{A}$ along with the time horizon and number of arms. It is so called because the regret bound depends explicitly on the means of the arms considered for that environment along with the stated assumptions on the distribution.

4. **Gap-Independent regret bounds:** In gap-independent regret bound the regret does not contain the gaps and is stated explicitly in terms of the number of arms and the horizon. This is because the regret depends only on the distributional assumption, but not on the means of the arms considered. In fact, gap-independent regret bounds point to something more general and informative. These type of bounds actually give us the maximum possible regret such that no matter what is the policy, there will be an environment on which the policy achieves almost the same regret as the gap-independent regret upper bound. This leads to the notion of minimax regret.

5. **Minimax regret bounds:** For a finite horizon T, K number of arms, for all set of possible policies $\pi_{T,K}$ over T and K and all possible environment class $\mathcal{E}$ the minimax regret is given by,

   $$R_T(\mathcal{E}) = \inf_{\pi \in \pi_{T,K}} \sup_{E \in \mathcal{E}} R_T(\pi, E).$$

   Hence, this value is independent of any specific choice of a policy π but only depends on T, K and $\mathcal{E}$ where the dependence on K is hidden in $\mathcal{E}$.

2.4 Motivation

There has been a significant amount of research in the area of stochastic MABs. One of the earliest work can be traced to Thompson (1933), which deals with the problem of choosing between two treatments to administer to patients who come in sequentially. In Thompson (1935) this work was extended to include more general cases of finitely many treatments. In recent years the SMAB setting has garnered extensive popularity because of its simple learning model and its practical applications in a wide-range of industries, including, but not limited to, mobile channel allocations, online advertising and computer simulation games. Some of these problems have been already discussed in chapter 1, section 1.4 and an interested reader can refer to it.

2.5 Related Work in SMAB

2.5.1 Lower bound in SMAB

SMAB problems have been extensively studied in several earlier works such as Thompson (1933), Thompson (1935), Robbins (1952) and Lai and Robbins (1985). Lai and Robbins in Lai and Robbins (1985) established an asymptotic lower bound for the cumulative regret. It showed that for any consistent allocation strategy, we can have

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T]}{\log T} \geq \sum_{\{i:r_i<r^*\}} \frac{(r^* - r_i)}{KL(Q_i||Q^*)}$$

where $KL(Q_i||Q^*)$ is the Kullback-Leibler divergence between the reward densities Q_i and Q^*, corresponding to arms with mean r_i and r^*, respectively.

2.5.2 The Upper Confidence Bound approach

Over the years SMABs have seen several algorithms with strong regret guarantees. For further reference, an interested reader can look into Bubeck and Cesa-Bianchi (2012). In the next few subsections, we will explicitly focus on the upper confidence bound algorithms which is a type of non-Bayesian algorithm widely used in SMAB setting.

The upper confidence bound or UCB algorithms balance the exploration-exploitation dilemma by linking the uncertainty in the estimate of an arm with the number of times an arm is pulled and therefore ensuring sufficient exploration.

UCB1 Algorithm

One of the earliest among these algorithms is UCB1 algorithm proposed first in Agrawal (1995) and subsequently analyzed in Auer *et al.* (2002*a*). The UCB1 algorithm (as stated in Auer *et al.* (2002*a*)) is mentioned in algorithm 3.

Algorithm 3 UCB1

1: **Input:** K number of arms with unknown parameters of reward distribution
2: Pull each arm once
3: **for** $t = K + 1, ..., T$ **do**
4: Pull the arm such that $\arg\max_{i \in A} \left\{ \hat{r}_i + \sqrt{\frac{2\log(t)}{n_i}} \right\}$
5: $t := t + 1$
6: **end for**

The intuition behind this algorithm is simple and it follows from the ideas of concentration inequalities in probability measure theory. The term $\sqrt{\frac{2\log(t)}{n_i}}$ is called the confidence interval of the arm i and it signifies a measure of uncertainty over the arm i based on the history of observed rewards for that arm. Therefore, lesser the confidence interval, higher is our confidence that the estimated mean $\hat{r}_i$ is lying close to the expected mean r_i of the arm i. Also, note that the confidence interval decreases at the rate of $O\left(\frac{1}{\sqrt{n_i}}\right)$ which signifies the rate of convergence of $\hat{r}_i$ to r_i and depends on the number of time the arm has been pulled.

UCB1 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 $O\left(\sqrt{KT\log T}\right)$.

UCB-Improved Algorithm

The UCB-Improved stated in algorithm 4, proposed in Auer and Ortner (2010), is a round-based variant of UCB1. 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

Algorithm 4 UCB-Improved

1: **Input:** Time horizon T, K number of arms with unknown parameters of reward distribution
2: **Initialization:** Set $B_0 := \mathcal{A}$ and $\epsilon_0 := 1$.
3: **for** $m = 0, 1, ..\lfloor \frac{1}{2}\log_2 \frac{T}{e} \rfloor$ **do**
4: Pull each arm in B_m, $n_m = \left\lceil \frac{2\log(T\epsilon_m^2)}{\epsilon_m} \right\rceil$ number of times.
5: ***Arm Elimination by Mean Estimation***
6: For each $i \in B_m$, delete arm i from B_m if,

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

7: Set $\epsilon_{m+1} := \frac{\epsilon_m}{2}$, Set $B_{m+1} := B_m$
8: Stop if $|B_m| = 1$ and pull $i \in B_m$ till n is reached.
9: **end for**

sub-optimal. Note, that in this algorithm the confidence interval term is $\sqrt{\frac{\log(T\epsilon_m^2)}{2n_m}}$ which is constant in the m-th round as n_m is fixed for that round and all arms are being pulled an equal number of times in each round. This is unlike UCB1 algorithm where the confidence interval term depends on n_i which is a random variable. Also, note that in UCB-Improved the knowledge of horizon is required before-hand to calculate the confidence intervals whereas no such input is required for UCB1. An illustrative flowchart depicting the main steps is given in Figure 2.1.

UCB-Improved 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 $O\left(\sqrt{KT\log K}\right)$.Empirically, UCB-Improved is outperformed by UCB1 in almost all environments. This stems from the fact that UCB-Improved is pulling all arms equal number of times in each round and hence spends a significant number of pulls in initial exploration as opposed to UCB1 thereby incurring higher regret.

MOSS Algorithm

In the later work of Audibert and Bubeck (2009), the authors propose the MOSS algorithm (see algorithm 5) and showed that the worst case gap-independent regret bound of

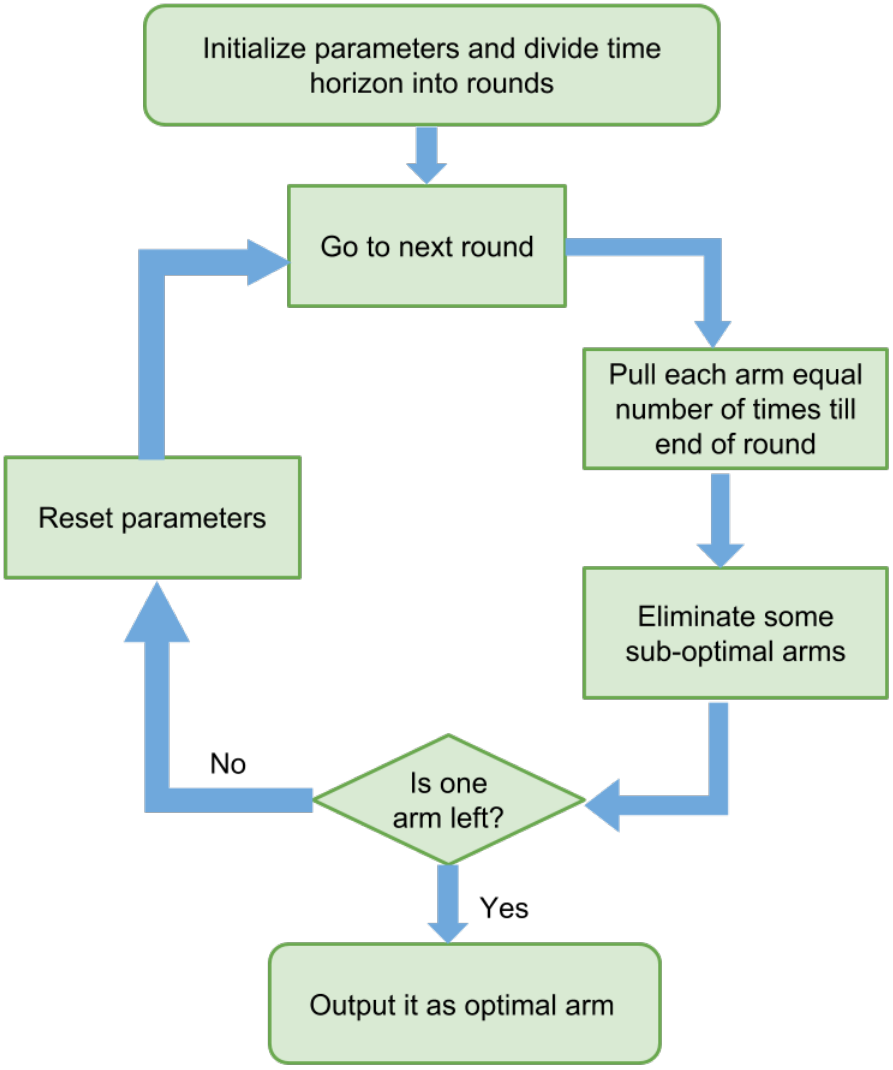


Figure 2.1: Flowchart of UCB-Improved

Algorithm 5 MOSS

1: **Input:** Time horizon T, K number of arms with unknown parameters of reward distribution
2: Pull each arm once
3: **for** $t = K + 1, ..., T$ **do**
4: Pull the arm such that $\arg\max_{i \in \mathcal{A}} \left\{ \hat{r}_i + \sqrt{\frac{\max\{0, \log(\frac{T}{Kn_i})\}}{n_i}} \right\}$
5: $t := t + 1$
6: **end for**

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)).

OCUCB Algorithm

Algorithm 6 OCUCB

Input: Time horizon T, K number of arms with unknown parameters of reward distribution, exploration parameter α and ψ
Pull each arm once
for $t = K+1, ..., T$ **do**
    Pull the arm such that $\arg\max_{i \in \mathcal{A}} \left\{ \hat{r}_i + \sqrt{\frac{\alpha \log(\psi \frac{T}{t})}{n_i}} \right\}$
    $t := t + 1$
end for

Recently in Lattimore (2015), the authors showed that the algorithm Optimally Confident UCB (OCUCB) (see algorithm 6) 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 proposed EU-CBV algorithm (in chapter 3), 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.

UCB-Variance algorithm

Algorithm 7 UCBV

Input: K number of arms with unknown parameters of reward distribution
Pull each arm once
for $t = K+1, ..., T$ **do**
    Pull the arm such that $\max_{i \in A} \left\{ \hat{r}_i + \sqrt{\frac{2\hat{v}_i \log(t)}{s_i}} + \frac{3\log(t)}{2} \right\}$
    $t := t + 1$
end for

In contrast to the above work, the UCB-Variance (UCBV) algorithm in Audibert *et al.* (2009) utilizes variance estimates to compute the confidence intervals for each arm. In UCBV (see algorithm 7) the confidence interval term is given by $\sqrt{\frac{2\hat{v}_i \log(t)}{s_i}} + \frac{3\log(t)}{3}$ where $\hat{v}_i$ denotes the empirical variance of the arm i. Hence, the confidence interval makes sure that the arms whose variances are high are pulled more often to get a better estimates of their $\hat{r}_i$.

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.

2.5.3 Bayesian Approach

Algorithm 8 Bernoulli Thompson Sampling

Input: K number of arms with unknown parameters of reward distribution
Initialization: For each arm $i := 1$ to K set $S_i = 0$ and $F_i = 0$

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

  for $i = 1, .., K$ **do**
    Sample $\theta_i(t)$ from the $Beta(S_i + 1, F_i + 1)$ distribution.
  end for
  Play the arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward $X_{i,t}$.
  if $X_{i,t} = 1$ **then** $S_i(t) = S_i(t) + 1$
  else$F_i(t) = F_i(t) + 1$
  end if
end for

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). This TS is stated in algorithm 8. The TS algorithm maintains a posterior reward distribution for each arm; at each round, the algorithm samples values from these distributions 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.

2.5.4 Information Theoretic approach

The final design principle we state is the information theoretic approach of DMED (Honda and Takemura, 2010) and KLUCB (Garivier and Cappé, 2011),(Cappé *et al.*,

2013) 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.

2.5.5 Discussion on the various confidence intervals

A comparative analysis of the confidence interval of the UCB algorithms is discussed in table 2.1.

Table 2.1: Confidence interval of different algorithms

Algorithm	Confidence interval	Horizon as input	Remarks
UCB1	$\sqrt{\frac{2\log(t)}{n_i}}$	No	Loose confidence interval leading to high regret upper bounds.
UCBV	$\sqrt{\frac{2\hat{v}_i\log(t)}{s_i}} + \frac{3\log(t)}{2}$	No	Confidence interval uses variance estimation.
UCB-Imp	$\sqrt{\frac{\log(T\epsilon_m^2)}{2n_m}}$	Yes	Same confidence interval for all arms in a particular round.
MOSS	$\sqrt{\frac{\max\{0, \log(\frac{T}{Kn_i})\}}{n_i}}$	Yes	Confidence interval is based on dividing the horizon uniformly for K arms.
OCUCB	$\sqrt{\frac{2\log(\frac{2T}{t})}{n_i}}$	Yes	Tightest confidence interval with exploration parameter $\alpha = 2$, $\psi = 2$ leading to order-optimal regret bounds.

2.6 Summary

In this chapter, we looked at the stochastic multi-armed bandit (SMAB) setting and discussed how it is important in the general reinforcement learning setup. We also looked at the various state-of-the-art algorithms in the literature for the SMAB setting and discussed the advantages and disadvantages of them. The regret bounds that have been proven for the said algorithms have also been discussed at length and their confidence intervals have also been compared against each other. In the next chapter, we provide our solution to this SMAB setting which achieves an almost order-optimal regret bound.

Chapter 3

Efficient UCB Variance: An Almost Optimal Algorithm in SMAB Setting

3.1 Introduction

In this chapter, we look 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 (SMAB) 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.

The rest of the chapter is organized as follows. We elaborate our contributions in section 3.2 and in section 3.3 we present the EUCBV algorithm. Our main theoretical results are stated in section 3.4, while the proofs are established in section 3.5. Section 3.6 contains results and discussions from our numerical experiments and finally we summarize in section 3.7.

3.2 Our Contributions

We propose the Efficient-UCB-Variance (henceforth referred to as EUCBV) algorithm for the stochastic MAB setting. EUCBV combines the approach of UCB-Improved, CCB (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),

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

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

3.3 Algorithm: Efficient UCB Variance

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.

Algorithm 9 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

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. An illustrative flowchart depicting the main steps is shown in Figure 3.1.

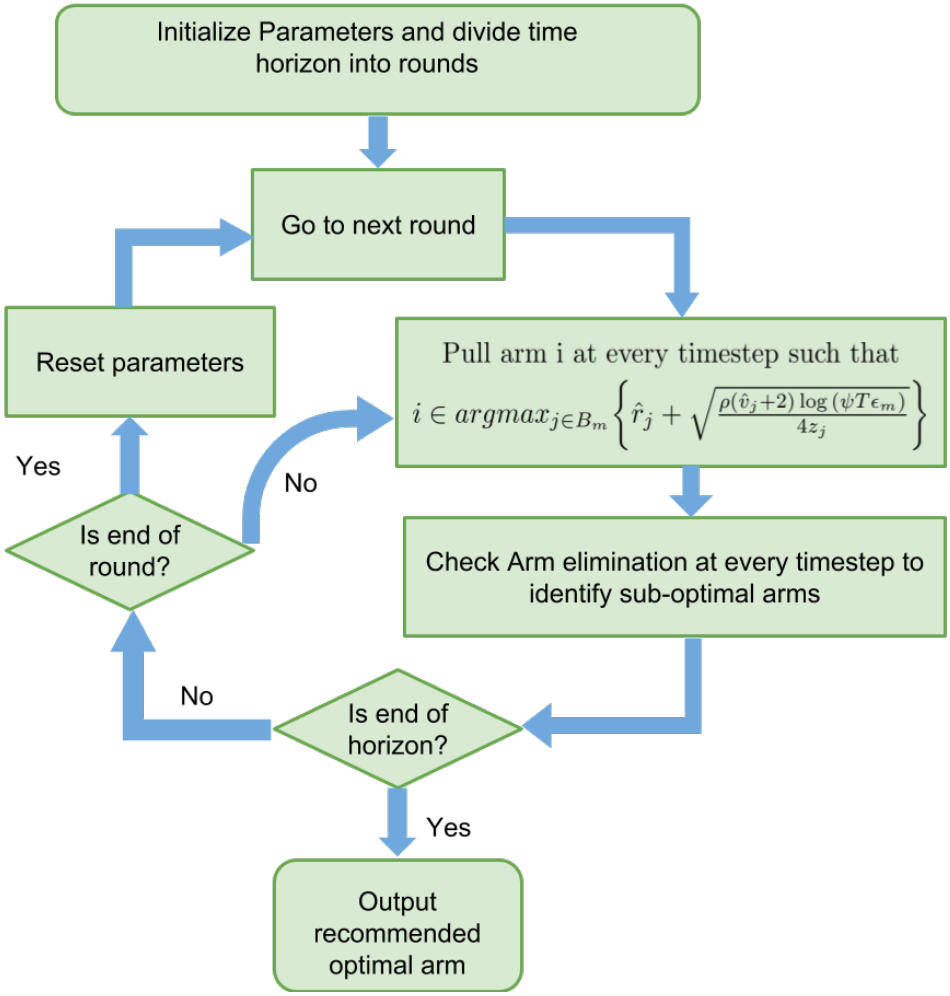


Figure 3.1: Flowchart of EUCBV algorithm

3.4 Main Results

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

Gap-Dependent bound of EUCBV

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

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

$$+ \sum_{i \in \mathcal{A}: 0 < \Delta_i \leq b} \frac{C_2 K^4}{T^{\frac{1}{4}}} + \max_{i \in \mathcal{A}: 0 < \Delta_i \leq b} \Delta_i T.$$

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

Proof (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 (see A.2.2) to derive their regret bound whereas in our work we use Bernstein inequality (as in Audibert* et al. *(2009), see A.2.3) 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 1 *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* et al. *(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.

Gap-Independent bound of EUCBV

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

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

$$\mathbb{E}[R_T] \leq \frac{C_3 K^5}{T^{\frac{1}{4}}} + 80\sqrt{KT}.$$

Proof *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 1 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,

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

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*

$$\mathbb{E}[R_T] \overset{(a)}{\leq} \frac{C_3 K^5}{T^{\frac{1}{4}}} + 320\sigma^2_{\max}\sqrt{KT} \overset{(b)}{\leq} \frac{C_3 K^5}{T^{\frac{1}{4}}} + 80\sqrt{KT}$$

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

■

Discussion 2 *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).*

3.5 Proofs

We first present a few technical lemmas that are required to prove the result in Theorem 1.

In Lemma 1 we use the constraint on the horizon, that is $T \geq K^{2.4}$ to derive an inequality which we re-use in Lemma 2 to bound the length of the confidence interval c_i of the i-th sub-optimal arm till the m_i-th round.

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

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

Proof *The proof is based on contradiction. Suppose*

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

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

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

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

In Lemma 2 we bound the length of the confidence interval c_i for the i-th arm till the m_i-th round.

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 *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,*

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

$$\overset{(b)}{\leq} \sqrt{2\rho\epsilon_{m_i} + 2.\frac{3}{2}\epsilon_{m_i}} < \sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{4}.$$

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

■

In Lemma 3 we bound the probability of the deviation of the sample mean $\hat{r}_i$ from its expectation r_i till the m_i-th round.

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 in the* m_i*-th round,*

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

Proof *We start by recalling from equation (3.6) that,*

$$\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right) \tag{3.1}$$

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,*

$$\bar{c}_i \leq \sqrt{\frac{\rho\epsilon_{m_i}(\sigma_i^2+\sqrt{\epsilon_{m_i}}+2)}{2}} \leq \sqrt{\epsilon_{m_i}}. \tag{3.2}$$

The first term in the LHS of (3.1) can be bounded using the Bernstein inequality as below:

$$\mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) \leq \exp\left(-\frac{(\bar{c}_i)^2 z_i}{2\sigma_i^2 + \frac{2}{3}\bar{c}_i}\right)$$

$$\overset{(a)}{\leq} \exp\left(-\rho\left(\frac{3\sigma_i^2+3\sqrt{\epsilon_{m_i}}+6}{6\sigma_i^2+2\sqrt{\epsilon_{m_i}}}\right)\log(\psi T\epsilon_{m_i}\right)$$
$$\overset{(b)}{\leq} \exp\left(-\rho\log(\psi T\epsilon_{m_i})\right) \leq \frac{1}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}} \tag{3.3}$$

where, (a) *is obtained by substituting equation 3.2 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 (3.1) can be simplified as follows:

$$\begin{aligned}\mathbb{P}\left\{\hat{v}_i \geq \sigma_i^2+\sqrt{\epsilon_{m_i}}\right\} &\leq \mathbb{P}\left\{\frac{1}{n_i}\sum_{t=1}^{n_i}(X_{i,t}-r_i)^2-(\hat{r}_i-r_i)^2 \geq \sigma_i^2+\sqrt{\epsilon_{m_i}}\right\}\\ &\leq \mathbb{P}\left\{\frac{\sum_{t=1}^{n_i}(X_{i,t}-r_i)^2}{n_i} \geq \sigma_i^2+\sqrt{\epsilon_{m_i}}\right\}\\ &\overset{(a)}{\leq} \mathbb{P}\left\{\frac{\sum_{t=1}^{n_i}(X_{i,t}-r_i)^2}{n_i} \geq \sigma_i^2+\bar{c}_i\right\}\\ &\overset{(b)}{\leq} \exp\left(-\rho\left(\frac{3\sigma_i^2+3\sqrt{\epsilon_{m_i}}+6}{6\sigma_i^2+2\sqrt{\epsilon_{m_i}}}\right)\log(\psi T\epsilon_{m_i})\right) \leq \frac{1}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}}\end{aligned} \tag{3.4}$$

where inequality (a) *is obtained using (3.2), while* (b) *follows from the Bernstein inequality.*

Thus, using (3.3) and (3.4) in (3.1) we obtain $\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \frac{2}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}}$. ■

In Lemma 4 we bound the probability of the confidence interval of the optimal arm c^* being greater than the confidence interval c_i of the i-th sub-optimal arm till the m_i-th round.

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,*

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

Proof *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,*

$$\mathbb{P}\{c^* > c_i\} \leq \mathbb{P}\{z^* < z_i\}$$

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

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,

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

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* (3.7) *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 $\epsilon_{m_i} \in [\sqrt{\frac{e}{T}}, 1]$.

■

In Lemma 5 we bound the probability of the number of pulls of the i-th sub-optimal arm, z_i, of not being greater then the allocated n_{m_i} number of pulls till the m_i-th round.

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

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

■

In Lemma 6 we prove the inequality required to introduce the variance term in the number of pulls of a sub-optimal arm i till the m_i-th round.

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 *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,*

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

■

Proof of Theorem 1

Proof *For each sub-optimal arm* $i \in \mathcal{A}$*, let* $m_i = \min\left\{m | \sqrt{4\epsilon_{m_i}} < \frac{\Delta_i}{4}\right\}$*. Also, let* $\mathcal{A}^{'} = \{i \in \mathcal{A} : \Delta_i > b\}$ *and* $\mathcal{A}^{''} = \{i \in \mathcal{A} : \Delta_i > 0\}$*. 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*

$$
\hat{r}_i \leq r_i + c_i, \hat{r}^* \geq r^* - c^*, c_i \geq c^* \text{ and } z_i \geq n_{m_i}. \tag{3.5}
$$

Indeed, in round m_i suppose (3.5) 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 (3.5) 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,*

$$\mathbb{P}(\hat{r}_i > r_i + c_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + \bar{c}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\epsilon_{m_i}}\right) \tag{3.6}$$

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 (3.5) is,*

$$\left(\frac{4}{(\psi T \epsilon_{m_i})^{\frac{3\rho}{2}}}\right). \tag{3.7}$$

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,

$$\frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}} + \frac{182K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}} \leq \frac{364K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}. \tag{3.8}$$

Also, for eq. (3.7) *we can show that for any* $\epsilon_{m_i} \in \left[\sqrt{\frac{e}{T}}, 1\right]$

$$\begin{aligned}\left(\frac{4}{(\psi T\epsilon_{m_i})^{\frac{3\rho}{2}}}\right) &\overset{(a)}{\leq} \left(\frac{4}{(\frac{T^2}{K^2}\epsilon_{m_i})^{\frac{3}{4}}}\right) \leq \left(\frac{4K^{\frac{3}{2}}}{(T^{\frac{3}{2}}\epsilon_{m_i}^{\frac{1}{4}}\sqrt{\epsilon_{m_i}})}\right) \\ &\overset{(b)}{\leq} \left(\frac{4K^{\frac{3}{2}}}{(T^{\frac{3}{2}-\frac{1}{8}}\sqrt{\epsilon_{m_i}})}\right) \leq \frac{4K^4}{T^{\frac{5}{4}}\sqrt{\epsilon_{m_i}}}.\end{aligned} \tag{3.9}$$

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 (3.5) by* $(3.8) + (3.9)$ *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,

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

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

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

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,

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

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.*

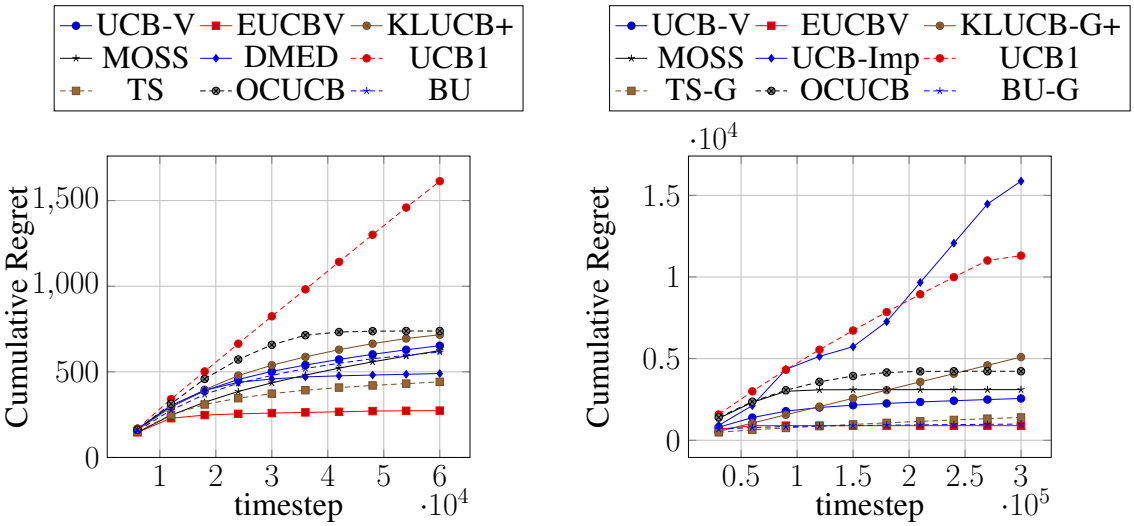
Finally, summing up the regrets in ***Case a*** *and* ***Case b****, the total regret is given by*

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

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

3.6 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)[1]. 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 (see Garivier and Cappé (2011)).



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

Figure 3.2: A comparison of the cumulative regret incurred by the various bandit algorithms.

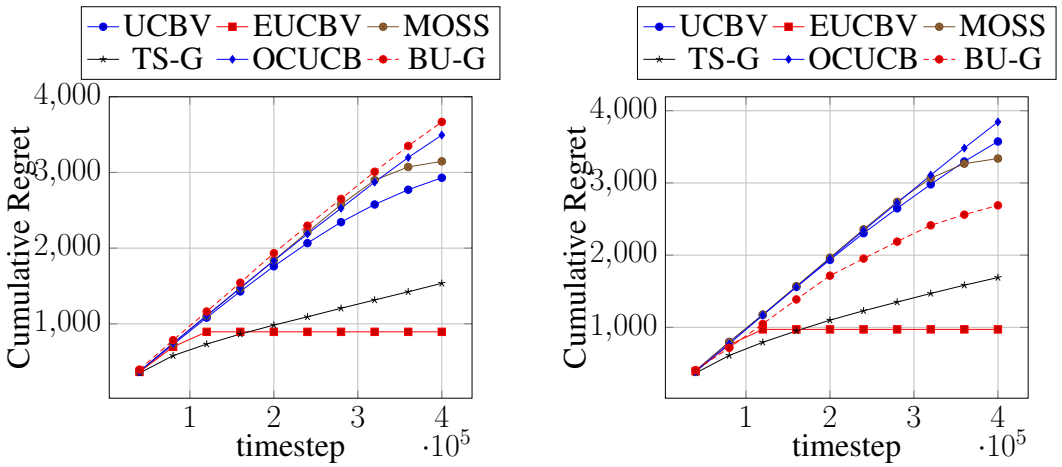
Experiment-1 (Bernoulli with uniform gaps): This experiment is conducted to observe the performance of EUCBV over a short horizon. The horizon T is set to 60000. The testbed comprises of 20 Bernoulli distributed arms with expected rewards of the arms as $r_{1:19} = 0.07$ and $r^*_{20} = 0.1$ and these 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 3.2(a). EUCBV, MOSS, OCUCB, UCB1, UCB-V, KLUCB+, TS, BU and DMED are run in this experimental setup. Not only

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

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 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 3.2(b). From the results in Figure 3.2(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 3.2(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 3.2(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.

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 $[0.2, 0.24]$.



(a) Expt-3: Failure of TS

(b) Expt-4: 3 Group Variance Setting

Figure 3.3: Further Experiments with EUCBV

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 3.3(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 3.3(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.

3.7 Summary

In this chapter, 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 an SMAB setup. Our experiments showed that EUCBV is extremely stable for large horizons and performs consistently well across different types of distributions.

Chapter 4

Thresholding Bandits

4.1 Introduction to Thresholding Bandits

In the previous chapters 2 and 3 we studied the stochastic multi-armed bandit (SMAB) setting with the goal of minimizing cumulative regret. In this chapter, we will study another setting called Pure-exploration multi-armed bandits. An interested reader can read through the previous chapters or can continue from here. Though we re-use the ideas from SMABs, the goal of pure exploration setup is distinctly different from that of cumulative regret minimization of SMABs and the required algorithms to understand this setup are mentioned in this chapter itself. Pure-exploration MAB problems are unlike their traditional (exploration vs. exploitation) counterparts, the SMABs, where the objective is to minimize the cumulative regret. The cumulative regret is the total loss incurred by the learner for not playing the optimal arm throughout the time horizon T. 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 chapter, 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 a unit reward whenever the value of an observation is above a threshold.

The rest of the chapter is organized as follows. We specify all the notations and assumptions in section 4.2. Then we define the problem statement for the TBP setting in section 4.3. In the next section 4.4 we discuss the motivations behind the TBP setting. In section 4.5 we discuss extensively on the various state-of-the-art algorithms available

for the pure exploration setting, in section 4.6 we draw the connection between pure exploration and thresholding bandit setting and then in section 4.7 we discuss the latest works done in the TBP setting. Finally, we summarize in section 4.8.

4.2 Notations and Assumptions

To benefit the reader, we again recall the notations we stated in chapter 2 and also a few additional notations. $\mathcal{A}$ 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 $z_i(t)$ to denote the number of times arm i has been pulled until time t, we have $\hat{r}_i(t) = \frac{1}{z_i(t)} \sum_{b=1}^{z_i(t)} X_{i,b}$, where $X_{i,b}$ is the reward sample received when arm i is pulled for the b-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., $\hat{v}_i(t) = \frac{1}{z_i(t)} \sum_{b=1}^{z_i(t)} (X_{i,b} - \hat{r}_i)^2$. Whenever there is no ambiguity about the underlaying time index t, for simplicity we neglect t from the notations and simply use $\hat{r}_i, \hat{v}_i$, and z_i, to denote the respective quantities. 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 $[0, 1]$.

4.3 Problem Definition

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.

4.4 Motivation

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 *et al.* (2010)) where the users are to be allocated only those channels whose quality is above an acceptable threshold. Again, some of these problems have been already discussed in chapter 1, section 1.4 and an interested reader can refer to it. In some cases the TBP problem is more relevant than the variants of p-best problem (identifying the best p arms from K given arms).

4.5 Related Work in Pure Exploration Problem

A significant amount of literature is available on the stochastic MAB setting with respect to minimizing the cumulative regret. Chapter 2 and 3 deals with that. In this work, we are particularly interested in *pure-exploration MABs*, where the focus is 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 in the following two settings:

4.5.1 Fixed Budget setting

Here the learning algorithm has to suggest the best arm(s) within a fixed budget or time-horizon T, that is given as an input. The objective is to maximize the probability of returning the best arm(s). This is the scenario we consider in this chapter. Some of the important algorithms used in pure exploration setting are discussed in the next part.

UCB-Exploration Algorithm

One of the first algorithms proposed for the fixed budget setting is the UCB-Exploration (UCBE) algorithm in Audibert *et al.* (2010) used for identifying a single best arm. This is shown in algorithm 10.

Algorithm 10 UCBE

1: **Input:** The budget T, exploration parameter a
2: Pull each arm once
3: **for** $t = K + 1, ..., T$ **do**
4: Pull the arm such that $\arg\max_{i \in \mathcal{A}} \left\{ \hat{r}_i + \sqrt{\frac{a}{z_i}} \right\}$, where $a = \frac{25(T-K)}{36 H_1}$ and $H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$.
5: $t := t + 1$
6: **end for**

This algorithm is quite similar to the UCB1 algorithm discussed in Auer *et al.* (2002*a*) (see algorithm 3). The major difference between the two algorithms is the confidence interval such that for UCB1 it is designed for minimizing the cumulative regret but for UCBE it is designed for minimizing simple regret. An illustrative table comparing the two is provided in table 4.1.

Successive Reject Algorithm

The Successive Reject (SR) algorithm has also been proposed in Audibert *et al.* (2010) and is used for identifying a single best arm. This algorithm is quite different than upper confidence bound based algorithms because it does not rely on any explicit confidence interval to select arm at every timestep. It is shown in algorithm 11.

Table 4.1: Confidence interval and exploration parameters of different algorithms

Algorithm	Confidence interval	Exploration Parameter(a)	Remarks
UCB1	$\sqrt{\frac{a}{z_i}}$	$a = 2\log(t)$	a is logarithmic in t to minimize cumulative regret. This achieves a balance between exploration and exploitation. Hence, the cumulative regret grows logarithmically with t.
UCBE	$\sqrt{\frac{a}{z_i}}$	$a = \frac{25(T-K)}{36H_1}$, where $H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$	a is linear in T to minimize simple regret. Here, the main concern is to minimize the probability of error at the end of budget T and conduct as much exploration as possible. Hence, a large a helps to reach exponentially low probability of error.

Algorithm 11 Successive Reject(SR)

1: **Input:** The budget T
2: **Initialization:** $n_0 = 0$
3: **Definition:** $\overline{\log K} = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $n_m = \frac{1}{\overline{\log K}} \frac{T-K}{K+1-m}$
4: **for** For each phase $m = 1, ..., K-1$ **do**
5: For each $i \in B_m$, select arm i for $n_m - n_{m-1}$ timesteps.
6: Let $B_{m+1} = B_m \setminus \arg\min_{i \in B_m} \hat{r}_i$ (remove one element from B_m , if there is a tie, select randomly the arm to dismiss among the worst arms).
7: $m := m + 1$
8: **end for**
9: Output the single remaining $i \in B_m$.

From algorithm 11 we see that SR is a round based algorithm quite similar to UCB-Improved (see algorithm 4). Similar to UCB-Improved, SR pulls all arms equal number of times in each round and then discards some arm that it deems to be sub-optimal until it is left with a single best arm. However, SR does not have any explicit confidence interval as UCBE, rather the idea of the confidence interval is hidden in the number of pulls allocated to each arm in every round. The number of times each arm is pulled in every round, that is $n_m - n_{m-1}$ timesteps makes sure that the optimal arm is not eliminated in the m-th round with high probability.

4.5.2 Successive Accept Reject Algorithm for best-p arms setting

The goal in the best p-arms setting is to identify the top p arms out of K given arms where p is supplied as an input. Several algorithms have been proposed for this setup starting with Gabillon *et al.* (2011) where the authors proposed the GapE and GapE-V algorithms that suggest, with high probability, the best-p arms at the end of the time budget. These algorithms are similar to the UCBE type algorithm discussed in the previous section. In this section, we will discuss the Successive Accept Reject strategy shown in algorithm 12.

Algorithm 12 Successive Accept Reject(SAR)

1: **Input:** The budget T, p
2: **Initialization:** $n_0 = 0$
3: **Definition:** $\overline{\log K} = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $n_m = \frac{1}{\overline{\log} K} \frac{T-K}{K+1-m}$
4: **for** For each phase $m = 1, ..., K-1$ **do**
5: For each $i \in B_m$, select arm i for $n_m - n_{m-1}$ timesteps.
6: Let B'_m be the set that contains arms in decreasing order of their sample means $\hat{r}_i, \forall i \in B_m$ such that $\hat{r}_{B'_m(1),n_m} \geq \hat{r}_{B'_m(2),n_m} \geq \ldots \geq \hat{r}_{B'_m(K+1-m),n_m}$.
7: Define the new empirical gaps $\forall i \in B'_m$ such that for $1 \leq r \leq K+1-m$,

$$\Delta_{B'_m(r),n_m} = \begin{cases} \hat{r}_{B'_m(r),n_m} - \hat{r}_{B'_m(p(m)+1),n_m}, & \text{if } r \leq p(m+1) \\ \hat{r}_{B'_m(p(m)),n_m} - \hat{r}_{B'_m(r),n_m}, & \text{if } r > p(m)+1 \end{cases}$$

8: Let $i_m \in \arg\max_{i \in B'_m} \hat{\Delta}_{i,n_m}$, then $B_{m+1} = B_m \setminus i_m$ (deactivate i_m with ties broken arbitrarily).
9: If $\hat{r}_{i_m,n_m} > \hat{r}_{B'_m(p(m)+1)}$, then accept i_m and set $p(m+1) = p(m) - 1$, $J_{p-p(m+1)} = i_m$.
10: $m := m + 1$
11: **end for**
12: Output the p-accepted arms $J_1, J_2, \ldots, J_p$.

Bubeck *et al.* (2013) introduced 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 p arms are suggested at the end of the budget with high probability. Like SR algorithm, the SAR algorithm divides the budget into rounds where in each round it pulls all the arms equal number of times that is for $n_m - n_{m-1}$ timesteps. At the end of the round it orders the arms into a decreasing sequence of their empirical means in B'_m and computes for the $p(m)$ empirical best arms in B'_m the distance (in

terms of their empirical means) to the $(p(m)+1)$-th empirical best arm in B'_m. Again for the arms not in the $p(m)$ empirical best arms, SAR computes the distance to the $p(m)$-th empirical best arm. Finally, SAR deactivates the arm i_m that has the maximum empirical distance. If i_m is the empirical best arm in the m-th round, then SAR accepts i_m and sets $p(m+1) = p(m) - 1$ and $J_p - p(m+1) = i_m$, or otherwise SAR rejects i_m.

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.

4.5.3 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 detailed survey of this setup, we refer the reader to Jamieson and Nowak (2014).

4.5.4 Unified Setting

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 fixed budget setting and fixed confidence setting.

4.6 TBP connection to Pure Exploration Problem

The thresholding bandit problem is a specific instance of the pure-exploration setup of Chen *et al.* (2014). To put it in perspective, the considered TBP setup lies at the intersection of the larger pure exploration setting and the stochastic multi-armed bandit

(SMAB) setting discussed in chapter 2 and chapter 3. This is shown in Figure 4.1.

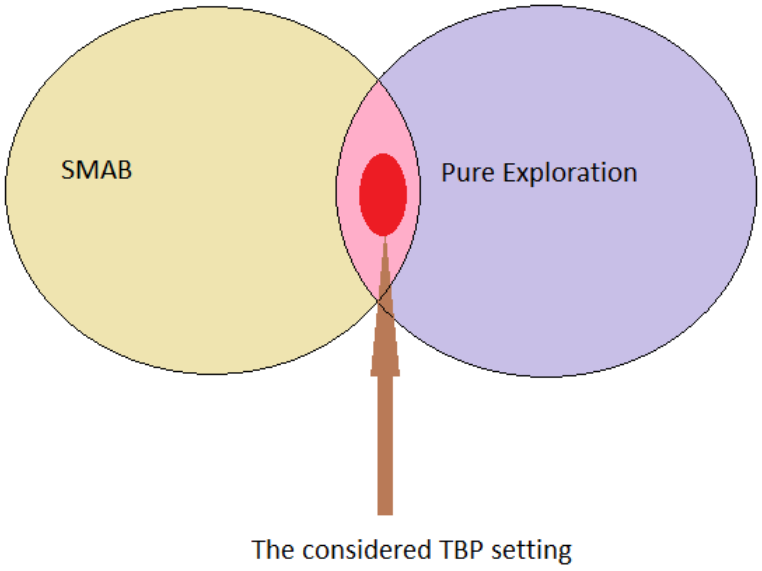


Figure 4.1: Connection between TBP, Pure Exploration and SMAB

Challenges in the TBP seeting

Further, if we look closely into the TBP setting we will see that there are several similarities between the challenges in SMAB setting and the TBP setting. These challenges are as follows:-

1. Closer an arm's expected reward mean (r_i) to $\tau \Rightarrow$ Harder is the problem. This stems from the fact that it becomes increasingly difficult to discriminate between the arms lying above and below τ.
2. Lesser the budget $T \Rightarrow$ Harder is the problem. This is because there are lesser number of pulls available and so the number of samples collected tends to be low.
3. Higher the variance of an arm's reward distribution $D_i \Rightarrow$ Harder is the problem. This is similar to the first case as it becomes harder to discriminate between the arms lying close to the threshold.

In the theoretical section in chapter 5 we will try to characterize this hardness and give guarantees that are almost optimal.

4.7 Related Work in Thresholding Bandits

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. APT is stated in algorithm 13.

Algorithm 13 APT

Input: Time horizon T, threshold τ, tolerance factor $\epsilon \geq 0$
Pull each arm once

for $t = K + 1, .., T$ **do**
Pull arm $j \in \arg\min_{i \in A} \left\{ \left(|\hat{r}_i - \tau| + \epsilon \right) \sqrt{z_i} \right\}$ and observe the reward for arm j.
end for
Output: $\hat{S}_\tau = \{i : \hat{r}_i \geq \tau\}$.

The APT algorithm is very simple to implement and the logic behind the arm pull directly follows from the challenges in the TBP setting discussed before. Note, that the most difficult arms to discriminate are the arms whose expected means are lying close to the threshold τ, hence APT pulls those arms whose sample means $\hat{r}_i$ are lying close to the threshold and the arms which have not been pulled often. The second condition is satisfied by the $\sqrt{z_i}$ term which acts very similar to the confidence interval term discussed for UCBE (see algorithm 10). The tolerance level $\epsilon \geq 0$ gives the algorithm a degree of flexibility in pulling the arms close to the threshold.

4.8 Summary

In this chapter, we looked at the pure exploration MAB and thresholding bandit (TBP) setting which is a special case of combinatorial pure exploration MAB. We then looked at the various state-of-the-art algorithms in the literature for the pure-exploration setting and discussed the advantages and disadvantages of them. Then we looked at the latest algorithm for the TBP setting. The expected loss that has been proven for the said algorithms have also been discussed at length and their exploration parameters have also been compared against each other. In the next chapter, we provide our solution to this TBP setting which uses variance estimation to find the set of arms above the threshold.

Chapter 5

Augmented UCB for Thresholding Bandit Problem

5.1 Introduction

In this chapter we look 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.

The rest of the chapter is organized as follows. We elaborate our contributions in section 5.2 and we present the AugUCB algorithm in section 5.3. Our main theoretical result on expected loss is stated in section 5.4. Section 5.5 contains numerical simulations on various testbeds to show the performance of AugUCB against state-of-the-art algorithms and finally, we summarize in section 5.6.

5.2 Our Contribution

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 2). In Table 5.1 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 5.4. From Section 5.4 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

Table 5.1: 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	$\exp\left(-\frac{T-K}{72\log(K)H_{CSAR,2}} + 2\log(K)\right)$

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.

5.3 Augmented-UCB Algorithm

The Algorithm: The Augmented-UCB (AugUCB) algorithm is presented in Algorithm 14. 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

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

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

Algorithm 14 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: $\hat{S}_\tau = \{i : \hat{r}_i \geq \tau\}$.

factor, ψ_m, comes directly from Audibert *et al.* (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).

A simplified illustrative flowchart highlighting the main steps of AugUCB is provided in Figure 5.1. Also note the similarity between UCB-Improved (Figure 2.1) and AugUCB in this flowchart.

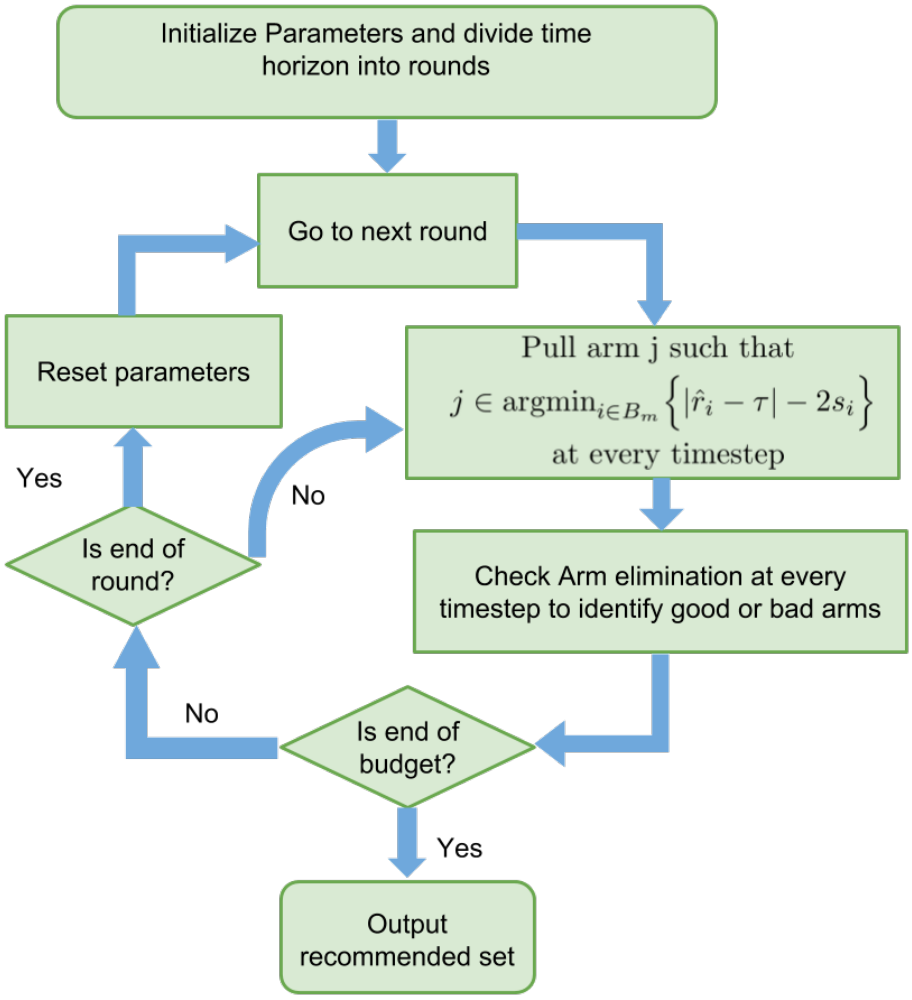


Figure 5.1: Flowchart for AugUCB

5.4 Theoretical Results

Problem Complexity

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

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

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

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

$H_{CSAR,2}$ 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

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

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 *et al.* (2010), we can show that

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

Proof of expected loss of AugUCB

Our main result is summarized in the following theorem where we prove an upper bound on the expected loss.

Theorem 2 *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 ***(Proof Outline)*** *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), see A.2.3) rather that the Chernoff-Hoeffding bounds (which is appropriate for the UCB-Improved (Auer and Ortner, 2010), see A.2.2). 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 of the proof as stated in the proof outline 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* $z_i = \ell_{m_i} \geq \frac{2\psi_{m_i} \log(T\epsilon_{m_i})}{\epsilon_{m_i}}$*, we obtain (as* $\hat{v}_i \in [0,1]$*)*

$$s_i \leq \sqrt{\frac{\rho(\hat{v}_i + 1)\epsilon_{m_i}}{8}} \leq \frac{\sqrt{\rho \epsilon_{m_i}}}{2} < \frac{\Delta_i}{4}. \tag{5.1}$$

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

$$\mathbb{P}(\hat{r}_i > r_i + 2s_i) \leq \mathbb{P}\left(\hat{r}_i > r_i + 2\bar{s}_i\right) + \mathbb{P}\left(\hat{v}_i \geq \sigma_i^2 + \sqrt{\rho\epsilon_{m_i}}\right) \tag{5.2}$$

where

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

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

$$2\bar{s}_i \leq \frac{\sqrt{\rho\epsilon_{m_i}(\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}} + 1)}}{2} \leq \sqrt{\rho\epsilon_{m_i}}. \tag{5.3}$$

The first term in the LHS of (3.6) can be bounded using the Bernstein inequality as below:

$$\begin{aligned}
&\mathbb{P}\left(\hat{r}_i > r_i + 2\bar{s}_i\right) \\
&\leq \exp\left(-\frac{(2\bar{s}_i)^2 z_i}{2\sigma_i^2 + \frac{4}{3}\bar{s}_i}\right) \\
&\leq \exp\left(-\frac{\rho\psi_{m_i}(\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}} + 1)\log(T\epsilon_{m_i})}{2\sigma_i^2 + \frac{2}{3}\sqrt{\rho\epsilon_{m_i}}}\right) \\
&\overset{(a)}{\leq} \exp\left(-\frac{3\rho T\epsilon_{m_i}}{256a^2}\left(\frac{\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}} + 1}{3\sigma_i^2 + \sqrt{\rho\epsilon_{m_i}}}\right)\log(T\epsilon_{m_i})\right) \\
&:= \exp(-Z_i)
\end{aligned} \tag{5.4}$$

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 (5.2) can be simplified as follows:

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

$$\overset{(b)}{\leq} \exp\left(-\frac{3\rho\psi_{m_i}}{2}\left(\frac{\sigma_i^2+\sqrt{\rho\epsilon_{m_i}}+1}{3\sigma_i^2+\sqrt{\rho\epsilon_{m_i}}}\right)\log(T\epsilon_{m_i})\right)$$
$$= \exp(-Z_i) \tag{5.5}$$

where inequality (a) is obtained using (5.3), while (b) follows from the Bernstein inequality.

Thus, using (5.4) and (5.5) in (5.2) 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

$$\mathbb{P}(|\hat{r}_i - r_i| > 2s_i) \leq 4\exp(-Z_i). \tag{5.6}$$

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

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

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

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* $\tilde{\Delta}_i^{-2} \leq i \tilde{\Delta}_i^{-2} \leq \max_{j \in \mathcal{A}} j \Delta_{(j)}^{-2} = H_{\sigma,2}$.

5.5 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.

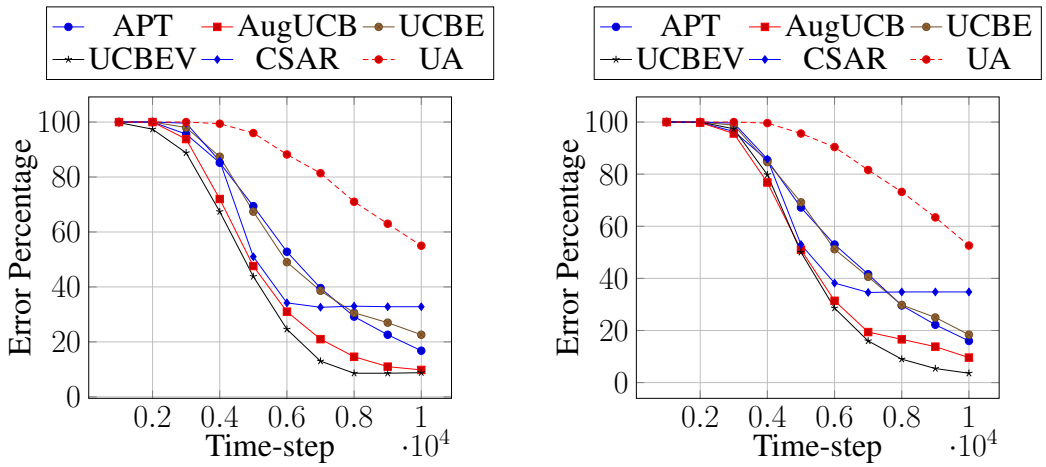
• UA: The naive strategy where at each time-step an arm is uniformly sampled from $\mathcal{A}$ (the set of all arms); however, UA is known to be optimal if all arms are equally difficult to classify.

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.

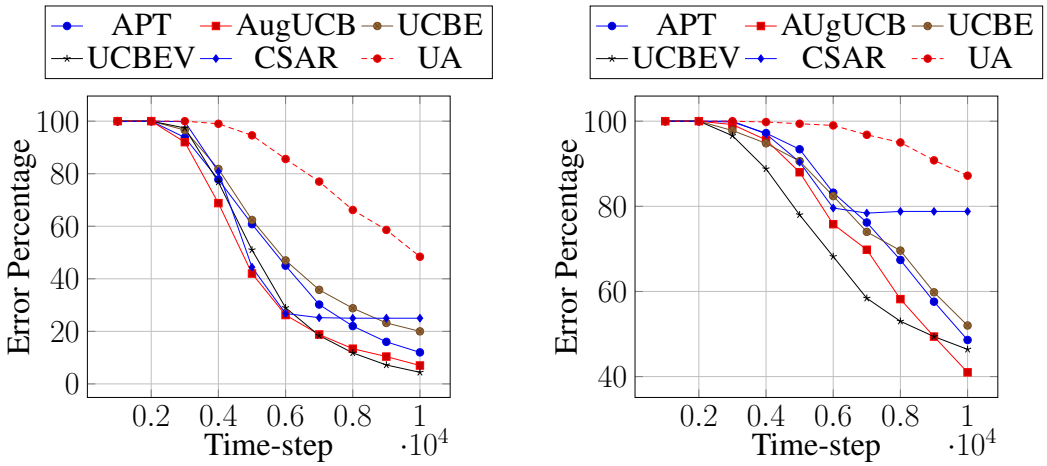
The corresponding variances are $\sigma^2_{1:5} = 0.5$ and $\sigma^2_{6:10} = 0.6$, while $\sigma^2_{11:100}$ is chosen independently and uniform in the interval $[0.38, 0.42]$; note that, the variances of the arms in S_τ are higher than those of the other arms. The corresponding results are

shown in Figure 5.2(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.



(a) Expt-1: Arithmetic Progression (Gaussian) (b) Expt-2: Geometric Progression (Gaussian)

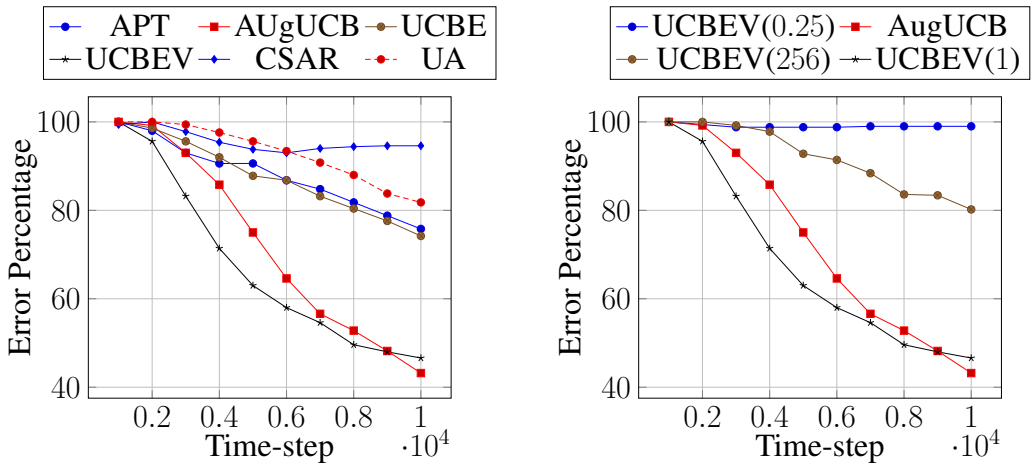
Figure 5.2: Performances of the various TBP algorithms in terms of error percentage vs. time-step in Arithmetic and Geometric Progression Environments.



(a) Expt-3: Three Group Setting (Gaussian) (b) Expt-4: Two Group Setting (Gaussian)

Figure 5.3: Performances of the various TBP algorithms in terms of error percentage vs. time-step in three group and two group Gaussian environments.

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 $r_{1:4} = 0.4 - (0.2)^{1:4}$, $r_5 = 0.45$, $r_6 = 0.55$, $r_{7:10} = 0.6 + (0.2)^{5-(1:4)}$



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

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

Figure 5.4: Performances of the various TBP algorithms in terms of error percentage vs. time-step in Advance setting Gaussian Environment.

and $r_{11:100} = 0.4$; the arm variances are as in experiment-1. The corresponding results are shown in Figure 5.2(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 5.3(a). 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 5.3(b), 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 5.4(a). 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.

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 5.4(b). 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 5.4(b)) 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.

5.6 Summary

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.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

In this thesis, we studied two complex bandit problems, the stochastic multi-armed bandit (SMAB) with the goal of cumulative regret minimization and pure exploration stochastic thresholding bandit problem (TBP) with the goal of expected loss minimization. For the first problem, we devised a novel algorithm called Efficient UCB Variance (EUCBV) which enjoys an order optimal regret bound and performs superbly in diverse stochastic environments. In the second part, the thresholding bandit problem, we came up with the novel algorithm called Augmented UCB (AugUCB) which is the first algorithm to use variance estimation for the considered TBP setting and also empirically outperforms most of the other algorithms.

6.2 Future Directions

There are several directions in which the work done in this thesis can be extended. Starting with the SMAB setting, there are many fundamental questions that need to be answered. Though EUCBV reached an order optimal regret bound of $80\sqrt{KT}$, still the constant associated with the bound is quite large and can be reduced by finer analysis. 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. Also, EUCBV does not have any asymptotic guarantee and we do not know whether it can reach the Lai and Robbins (1985) asymptotic lower bound discussed in chapter 2. Recently another algorithm called KL-UCB++ (Ménard and Garivier, 2017) has been proved to be both minimax optimal and asymptotically optimal. Further, EUCBV requires the knowledge of horizon as input and it will be interesting to find an anytime version of EUCBV. Similar anytime version

of MOSS (Degenne and Perchet, 2016) and OCUCB (Lattimore, 2016) has also been proposed in literature.

The thresholding bandit problem is also being intensely studied in the bandit community and there are several directions where this work can be extended. One way is to modify the APT algorithm itself and come up with a variance adaptive version of APT. This has bee recently studied in Kano *et al.* (2017). Also, currently there are no lower bounds for the TBP setting considering only variance estimation and it will be interesting to derive a lower bound for this setting. Again, whereas APT is an anytime algorithm AugUCB is not anytime and it needs several modifications to obtain an anytime version of AugUCB. Further, we can also derive a gap-independent and gap-dependent bounds for AugUCB as like APT. In Lattimore (2015) the authors showed that APT like UCB1 enjoys a gap-dependent cumulative regret bound of $O\left(\frac{K\log T}{\Delta^2}\right)$ and gap-independent regret bound of $O\left(\sqrt{KT\log T}\right)$.

Finally, to summarize everything, the bandit community is actively researching several of these open problems discussed here and we hope to answer some of these problems in near future. Further, several interesting variations of the problems discussed here are also being studied such as the Contextual Thresholding Bandit problem, Combinatorial Bandit problems (Cesa-Bianchi and Lugosi, 2012) and more powerful algorithms for changepoint detection.

Appendix A

Appendix on Concentration Inequalities

A.1 Sub-Gaussian Distribution

Let a random variable $X \in \mathbb{R}$ with variance as σ^2. Then X is said to be σ-sub-gaussian for $\sigma \geq 0$ such that $\mathbb{E}[X] = 0$ and its moment generating function satisfies for all $\lambda \in \mathbb{R}$ the following condition,

$$\mathbb{E}[\exp \lambda X] \leq \exp\left(-\frac{\lambda^2\sigma^2}{2}\right)$$

Also, note that sub-gaussian distribution is a class of distribution rather than a distribution itself.

Remark 1 *A random variable $X \in [0,1]$ is said to be $\frac{1}{2} - sub - gaussian$ with its moment generating function satisfying the condition,*

$$\mathbb{E}[\exp \lambda X] \leq \exp\left(-\frac{\lambda^2}{8}\right), \forall \lambda \in \mathbb{R}.$$

A.2 Concentration Inequalities

In this section we state some of the concentration inequalities used in the proofs in several chapters of the thesis. Concentration inequality deals with the control of the deviation of the average of independent random variables from their expected mean.

Let, $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables defined on a probability space $(\omega, \mathcal{F}, \mathbb{P})$, is bounded in $[a_i, b_i], \forall i = 1, 2, \ldots, n$. Let S_n denote

the sum of the random variables such that $S_n = X_1 + X_2 + \ldots + X_n$, $\hat{r} = \frac{S_n}{n}$ and $E[S_n] = r$. Let $\mathcal{F}_n$ be an increasing sequence of σ-fields of $\mathcal{F}$ such that for each n, $\sigma(X_1, \ldots, X_n) \subset \mathcal{F}_t$ and for $q > t$, X_q is independent of $\mathcal{F}_n$.

A.2.1 Markov's Inequality

Markov's inequality states that, for any $\epsilon > 0$,

$$\mathbb{P}[S_n > \epsilon] \leq \frac{\mathbb{E}[S_n]}{\epsilon}.$$

The Markov;s inequality gives us a very loose bound which is further tightened by the Chernoff-Hoeffding Bound and Bernstein Inequality.

A.2.2 Chernoff-Hoeffding Bound

Chernoff-Hoeffding gives us the following inequality regarding the sums of independent random variables S_n and their deviation from their expectation $\mathbb{E}[S_n] = r$, for any $\epsilon > 0$,

$$\mathbb{P}\{S_n - n\mathbb{E}[S_n] \geq \epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{n\sum_{i=1}^{n}(a_i - b_i)}\right),$$
$$\mathbb{P}\{S_n - n\mathbb{E}[S_n] \leq -\epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{n\sum_{i=1}^{n}(a_i - b_i)}\right).$$

Considering all the random variables bounded in $[0, 1]$, then both the right and left tail inequality can be reduced to,

$$\mathbb{P}\left\{\left|\frac{S_n}{n} - \mathbb{E}[S_n]\right| \geq \epsilon\right\} \leq 2\exp\left(-2\epsilon^2 n\right).$$

Hence, we obtain that,

$$\mathbb{P}\{|\hat{r} - r| \geq \epsilon\} \leq 2\exp\left(-2\epsilon^2 n\right).$$

A.2.3 Empirical Bernstein Inequality

Similar to Chernoff-Hoeffding bound, empirical Bernstein inequality gives us the following inequality regarding the sums of independent random variables S_n and their deviation from their expectation $\mathbb{E}[S_n] = r$, for any $\epsilon > 0$,

$$\mathbb{P}\{S_n - n\mathbb{E}[S_n] \geq \epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{\left(2\sigma^2 + \frac{2b_{\max}\epsilon}{3}\right) n \sum_{i=1}^{n}(a_i - b_i)}\right),$$
$$\mathbb{P}\{S_n - n\mathbb{E}[S_n] \leq -\epsilon\} \leq \exp\left(-\frac{2\epsilon^2}{\left(2\sigma^2 + \frac{2b_{\max}\epsilon}{3}\right) n \sum_{i=1}^{n}(a_i - b_i)}\right).$$

Considering all the random variables bounded in $[0, 1]$, then both the right and left tail inequality can be reduced to,

$$\mathbb{P}\left\{\left|\frac{S_n}{n} - \mathbb{E}[S_n]\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2\epsilon^2 n}{\left(2\sigma^2 + \frac{2\epsilon}{3}\right)}\right).$$

Hence, we obtain that,

$$\mathbb{P}\{|\hat{r} - r| \geq \epsilon\} \leq 2\exp\left(-\frac{2\epsilon^2 n}{\left(2\sigma^2 + \frac{2\epsilon}{3}\right)}\right).$$

Bibliography

1. **Abernethy, J. D.**, **K. Amin**, and **R. Zhu** (2016). Threshold bandits, with and without censored feedback. *29th Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 4889–4897. URL http://papers.nips.cc/paper/6149-threshold-bandits-with-and-without-censored-feedback.

2. **Agrawal, R.** (1995). Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, **27**(4), 1054–1078.

3. **Agrawal, S.** and **N. Goyal** (2011). Analysis of thompson sampling for the multi-armed bandit problem. *CoRR*, **abs/1111.1797**. URL http://arxiv.org/abs/1111.1797.

4. **Allesiardo, R.**, **R. Féraud**, and **O. Maillard** (2017). The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, **3**(4), 267–283. URL https://doi.org/10.1007/s41060-017-0050-5.

5. **Audibert, J.** and **S. Bubeck** (2009). Minimax policies for adversarial and stochastic bandits. *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 217–226. URL http://www.cs.mcgill.ca/~colt2009/papers/022.pdf#page=1.

6. **Audibert, J.**, **S. Bubeck**, and **R. Munos** (2010). Best arm identification in multi-armed bandits. *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, 41–53. URL http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=49.

7. **Audibert, J.-Y.**, **R. Munos**, and **C. Szepesvári** (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, **410**(19), 1876–1902. URL https://doi.org/10.1016/j.tcs.2009.01.016.

8. **Auer, P.** (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, **3**(Nov), 397–422. URL http://www.jmlr.org/papers/v3/auer02a.html.

9. **Auer, P.**, **N. Cesa-Bianchi**, and **P. Fischer** (2002*a*). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, **47**(2-3), 235–256. URL https://doi.org/10.1023/A:1013689704352.

10. **Auer, P.**, **N. Cesa-Bianchi**, **Y. Freund**, and **R. E. Schapire** (2000). Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Electronic Colloquium on Computational Complexity (ECCC)*, **7**(68). URL http://eccc.hpi-web.de/eccc-reports/2000/TR00-068/index.html.

11. **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. URL https://doi.org/10.1137/S0097539701398375.

12. **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. URL https://doi.org/10.1007/s10998-010-3055-6.

13. **Awerbuch, B.** and **R. Kleinberg** (2008). Competitive collaborative learning. *J. Comput. Syst. Sci.*, **74**(8), 1271–1288. URL https://doi.org/10.1016/j.jcss.2007.08.004.

14. **Awerbuch, B.** and **R. D. Kleinberg** (2004). Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, 45–53. URL http://doi.acm.org/10.1145/1007352.1007367.

15. **Bertsekas, D. P.** and **J. N. Tsitsiklis**, *Neuro-dynamic programming*, volume 3 of *Optimization and neural computation series*. Athena Scientific, 1996. ISBN 1886529108. URL http://www.worldcat.org/oclc/35983505.

16. **Besbes, O.**, **Y. Gur**, and **A. J. Zeevi** (2014). Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *CoRR*, **abs/1405.3316**. URL http://arxiv.org/abs/1405.3316.

17. **Beygelzimer, A.**, **J. Langford**, **L. Li**, **L. Reyzin**, and **R. E. Schapire** (2011). Contextual bandit algorithms with supervised learning guarantees. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, 19–26. URL http://www.jmlr.org/proceedings/papers/v15/beygelzimer11a/beygelzimer11a.pdf.

18. **Bubeck, S.** and **N. Cesa-Bianchi** (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, **5**(1), 1–122. URL https://doi.org/10.1561/2200000024.

19. **Bubeck, S.**, **R. Munos**, and **G. Stoltz** (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, **412**(19), 1832–1852. URL https://doi.org/10.1016/j.tcs.2010.12.059.

20. **Bubeck, S.**, **T. Wang**, and **N. Viswanathan** (2013). Multiple identifications in multi-armed bandits. *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, **28**, 258–265. URL http://jmlr.org/proceedings/papers/v28/bubeck13.html.

21. **Bui, L.**, **R. Johari**, and **S. Mannor** (2012). Clustered bandits. *CoRR*, **abs/1206.4169**. URL http://arxiv.org/abs/1206.4169.

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

23. **Cappé, O.**, **A. Garivier**, **O.-A. Maillard**, **R. Munos**, **G. Stoltz**, *et al.* (2013). Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, **41**(3), 1516–1541.

24. **Cesa-Bianchi, N.**, **C. Gentile**, and **G. Zappella** (2013). A gang of bandits. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, 737–745. URL http://papers.nips.cc/paper/5006-a-gang-of-bandits.

25. **Cesa-Bianchi, N.** and **G. Lugosi**, *Prediction, learning, and games*. Cambridge University Press, 2006. ISBN 978-0-521-84108-5.

26. **Cesa-Bianchi, N.** and **G. Lugosi** (2012). Combinatorial bandits. *J. Comput. Syst. Sci.*, **78**(5), 1404–1422. URL https://doi.org/10.1016/j.jcss.2012.01.001.

27. **Chen, S.**, **T. Lin**, **I. King**, **M. R. Lyu**, and **W. Chen** (2014). Combinatorial pure exploration of multi-armed bandits. *27th Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 379–387. URL http://papers.nips.cc/paper/5433-combinatorial-pure-exploration-of-multi-armed-bandits.

28. **Degenne, R.** and **V. Perchet** (2016). Anytime optimal algorithms in stochastic multi-armed bandits. *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 1587–1595. URL http://jmlr.org/proceedings/papers/v48/degenne16.html.

29. **Even-Dar, E.**, **S. Mannor**, and **Y. Mansour** (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, **7**, 1079–1105. URL http://www.jmlr.org/papers/v7/evendar06a.html.

30. **Freund, Y.** and **R. E. Schapire** (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. *European conference on computational learning theory*, 23–37.

31. **Gabillon, V.**, **M. Ghavamzadeh**, and **A. Lazaric** (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 3221–3229. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.420.6716&rep=rep1&type=pdf.

32. **Gabillon, V.**, **M. Ghavamzadeh**, **A. Lazaric**, and **S. Bubeck** (2011). Multi-bandit best arm identification. *25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, 2222–2230. URL http://papers.nips.cc/paper/4478-multi-bandit-best-arm-identification.

33. **Gajane, P.**, **T. Urvoy**, and **E. Kaufmann** (2017). Corrupt bandits for privacy preserving input. *CoRR*, **abs/1708.05033**. URL http://arxiv.org/abs/1708.05033.

34. **Garivier, A.** and **O. Cappé** (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. *COLT 2011 - The 24th Annual Conference on Learning Theory,*

June 9-11, 2011, Budapest, Hungary, **19**, 359–376. URL http://www.jmlr.org/proceedings/papers/v19/garivier11a/garivier11a.pdf.

35. **Garivier, A.** and **E. Moulines** (2011). On upper-confidence bound policies for switching bandit problems. *Algorithmic Learning Theory - 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, **6925**, 174–188. URL https://doi.org/10.1007/978-3-642-24412-4_16.

36. **Gentile, C.**, **S. Li**, and **G. Zappella** (2014). Online clustering of bandits. *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, **32**, 757–765. URL http://jmlr.org/proceedings/papers/v32/gentile14.html.

37. **Ghavamzadeh, M.**, **S. Mannor**, **J. Pineau**, and **A. Tamar** (2015). Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, **8**(5-6), 359–483. URL https://doi.org/10.1561/2200000049.

38. **György, A.**, **T. Linder**, **G. Lugosi**, and **G. Ottucsák** (2007). The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, **8**(Oct), 2369–2403. URL http://dl.acm.org/citation.cfm?id=1314575.

39. **Hartland, C.**, **N. Baskiotis**, **S. Gelly**, **M. Sebag**, and **O. Teytaud** (2007). Change point detection and meta-bandits for online learning in dynamic environments. *CAp*, 237–250.

40. **Hillel, E.**, **Z. S. Karnin**, **T. Koren**, **R. Lempel**, and **O. Somekh** (2013). Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, 854–862. URL http://papers.nips.cc/paper/4919-distributed-exploration-in-multi-armed-bandits.

41. **Honda, J.** and **A. Takemura** (2010). An asymptotically optimal bandit algorithm for bounded support models. *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, 67–79. URL http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=75.

42. **Jamieson, K. G.** and **R. D. Nowak** (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. *48th Annual Conference on Information Sciences and Systems, CISS 2014, Princeton, NJ, USA, March 19-21, 2014*, 1–6. URL https://doi.org/10.1109/CISS.2014.6814096.

43. **Kalai, A.** and **S. Vempala** (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, **71**(3), 291–307. URL https://doi.org/10.1016/j.jcss.2004.10.016.

44. **Kalyanakrishnan, S.**, **A. Tewari**, **P. Auer**, and **P. Stone** (2012). PAC subset selection in stochastic multi-armed bandits. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. URL http://icml.cc/2012/papers/359.pdf.

45. **Kano, H.**, **J. Honda**, **K. Sakamaki**, **K. Matsuura**, **A. Nakamura**, and **M. Sugiyama** (2017). Good arm identification via bandit feedback. *arXiv preprint arXiv:1710.06360*.

46. **Kaufmann, E.**, **O. Cappé**, and **A. Garivier** (2012). On bayesian upper confidence bounds for bandit problems. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, **22**, 592–600. URL http://jmlr.csail.mit.edu/proceedings/papers/v22/kaufmann12.html.

47. **Kocák, T.**, **G. Neu**, **M. Valko**, and **R. Munos** (2014). Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 613–621. URL http://cs.bme.hu/~gergo/files/KNVM14.pdf.

48. **Kocsis, L.** and **C. Szepesvári** (2006). Discounted ucb. *2nd PASCAL Challenges Workshop*, 784–791.

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

50. **Langford, J.** and **T. Zhang** (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 817–824. URL http://hunch.net/~jl/projects/interactive/sidebandits/bandit.pdf.

51. **Lattimore, T.** (2015). Optimally confident UCB : Improved regret for finite-armed bandits. *CoRR*, **abs/1507.07880**. URL http://arxiv.org/abs/1507.07880.

52. **Lattimore, T.** (2016). Regret analysis of the anytime optimally confident UCB algorithm. *CoRR*, **abs/1603.08661**. URL http://arxiv.org/abs/1603.08661.

53. **Li, L.**, **W. Chu**, **J. Langford**, and **R. E. Schapire** (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, 661–670. URL http://doi.acm.org/10.1145/1772690.1772758.

54. **Littlestone, N.** and **M. K. Warmuth** (1994). The weighted majority algorithm. *Inf. Comput.*, **108**(2), 212–261. URL https://doi.org/10.1006/inco.1994.1009.

55. **Liu, F.**, **J. Lee**, and **N. B. Shroff** (2017). A change-detection based framework for piecewise-stationary multi-armed bandit problem. *CoRR*, **abs/1711.03539**. URL http://arxiv.org/abs/1711.03539.

56. **Liu, K.** and **Q. Zhao** (2010). Distributed learning in multi-armed bandit with multiple players. *IEEE Trans. Signal Processing*, **58**(11), 5667–5681. URL https://doi.org/10.1109/TSP.2010.2062509.

57. **Liu, Y.** and **Y. Tsuruoka** (2016). Modification of improved upper confidence bounds for regulating exploration in monte-carlo tree search. *Theoretical Computer Science*, **644**, 92–105. URL https://doi.org/10.1016/j.tcs.2016.06.034.

58. **Locatelli, A.**, **M. Gutzeit**, and **A. Carpentier** (2016). An optimal algorithm for the thresholding bandit problem. *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, NY, USA, June 19-24, 2016*, **48**, 1690–1698. URL http://jmlr.org/proceedings/papers/v48/locatelli16.html.

59. **Maillard, O.-A.** (2011). *LEARNING S 'E QUENTIAL: Bandits, Statistics and Reinforcement.*. Ph.D. thesis, University of Science and Technology of Lille-Lille I.

60. **McMahan, H. B.** and **A. Blum** (2004). Online geometric optimization in the bandit setting against an adaptive adversary. *Learning Theory, 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004, Proceedings*, 109–123. URL https://doi.org/10.1007/978-3-540-27819-1_8.

61. **Mellor, J. C.** and **J. Shapiro** (2013). Thompson sampling in switching environments with bayesian online change point detection. *CoRR*, **abs/1302.3721**. URL http://arxiv.org/abs/1302.3721.

62. **Ménard, P.** and **A. Garivier** (2017). A minimax and asymptotically optimal algorithm for stochastic bandits. *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, **76**, 223–237. URL http://proceedings.mlr.press/v76/m%C3%A9nard17a.html.

63. **Raj, V.** and **S. Kalyani** (2017). Taming non-stationary bandits: A bayesian approach. *CoRR*, **abs/1707.09727**. URL http://arxiv.org/abs/1707.09727.

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

65. **Steinwart, I.**, **D. R. Hush**, and **C. Scovel** (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, **6**, 211–232. URL http://www.jmlr.org/papers/v6/steinwart05a.html.

66. **Streeter, M. J.** and **S. F. Smith** (2006). Selecting among heuristics by solving thresholded k-armed bandit problems. *ICAPS 2006*, 123–127.

67. **Sutton, R. S.** and **A. G. Barto**, *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 0262193981. URL http://www.worldcat.org/oclc/37293240.

68. **Szörényi, B.**, **R. Busa-Fekete**, **I. Hegedüs**, **R. Ormándi**, **M. Jelasity**, and **B. Kégl** (2013). Gossip-based distributed stochastic bandit algorithms. *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 19–27. URL http://jmlr.org/proceedings/papers/v28/szorenyi13.html.

69. **Takimoto, E.** and **M. K. Warmuth** (2003). Path kernels and multiplicative updates. *Journal of Machine Learning Research*, **4**, 773–818. URL http://www.jmlr.org/papers/v4/takimoto03a.html.

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

71. **Thompson, W. R.** (1935). On the theory of apportionment. *American Journal of Mathematics*, **57**(2), 450–456.

72. **Wu, Y.**, **R. Shariff**, **T. Lattimore**, and **C. Szepesvári**, Conservative bandits. *In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. 2016. URL `http://jmlr.org/proceedings/papers/v48/wu16.html`.

73. **Yu, J. Y.** and **S. Mannor** (2009). Piecewise-stationary bandit problems with side observations. *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, 1177–1184. URL `http://doi.acm.org/10.1145/1553374.1553524`.

LIST OF PAPERS BASED ON THESIS

1. Subhojyoti Mukherjee, K.P. Naveen, Nandan Sudarsanam, and Balaraman Ravindran, "*Thresholding Bandit with Augmented UCB*", *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017,2515-2521*.

2. Subhojyoti Mukherjee, K.P. Naveen, Nandan Sudarsanam, and Balaraman Ravindran, "*Efficient UCBV: An Almost Optimal Algorithm using Variance Estimates*", *To appear in Proceedings of the Thirty-Second Association for the Advancement of Artificial Intelligence, AAAI 2018, New Orleans, Louisiana, USA, February 2-7*.