Efficient Learning in Stochastic Bandits

YU, Xiaotian

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Computer Science and Engineering

The Chinese University of Hong Kong
December 2018

Thesis Assessment Committee

Professor CHAN Lai Wan (Chair)

Professor KING Kuo Chin Irwin (Thesis Supervisor)

Professor LYU Rung Tsong Michael (Thesis Co-supervisor)

Professor YIP Yuk Lap (Committee Member)

Professor XU Zeng Lin (External Member)

Abstract of thesis entitled:

Efficient Learning in Stochastic Bandits

Submitted by YU, Xiaotian

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in December 2018

The prevailing decision-making model named as multi-armed bandits (MAB) elegantly characterizes a wide class of practical problems in sequential learning with stochastic feedbacks. A predominant characteristic of MAB is a trade-off between exploration and exploitation in the sequential decision process. The intrinsic trade-off frequently arises in scientific research and various industrial applications, e.g., resource allocation, online advertising and personalized recommendations.

In this thesis, we study efficient learning in stochastic bandits (LSB). The goal of efficient LSB is to develop algorithms with provable performance guarantees, as well as practical implementations. We address three challenges in LSB: mean-variance explorations, decisions with heavy-tailed payoffs and fast learning for nonlinear payoff functions. By attacking these three challenges, we generalize the applicability of bandits to more real-world scenarios.

This thesis makes four main contributions. First, we rigorously prove that the error resulting from the mean-variance estimation is sub-gamma. Then, we develop two efficient algorithms to solve the problem of pure exploration of mean-variance. In addition, based on sub-gamma estimation noises, we derive upper bounds of the probability of error

for the proposed algorithms. By comparing with two algorithms in experiments, we show the superiority and robustness of our algorithms.

Second, we investigate the problem of pure exploration of MAB with heavy-tailed payoffs. Heavy-tailed payoffs in this thesis mean that each feedback after a decision has finite p-th moments, where $p \in (1, +\infty)$. We analyze tail probabilities of empirical average and truncated empirical average for estimating expected payoffs in sequential decisions. In addition, we propose two bandit algorithms to solve the problem of pure exploration of MAB with heavy-tailed payoffs. We derive theoretical guarantees and show the effectiveness of our bandit algorithms.

Third, we focus on the practical problem of regret minimization for linear stochastic bandits with heavy-tailed payoffs. We rigorously analyze the lower bound of the above problem, and develop two novel bandit algorithms such that the regret upper bounds match the lower bound up to polylogarithmic factors. To the best of our knowledge, we are the first to solve the problem optimally in the sense of the polynomial order on the total number of rounds in sequential decisions. The proposed algorithms outperform the state-of-the-art methods.

Finally, we investigate the problem of stochastic bandits with nonlinear payoff functions. We propose a generic algorithm for accelerating the convergence of existing algorithms to learn nonlinear functions. The key of the novel algorithm is to explore a local growth condition of underlying objective functions. The benefits of the proposed acceleration technique are three-fold: 1) it is applicable to both settings of one-point and two-point evaluations; 2) it does not necessarily require strong convexity or smoothness of objective functions; 3) it improves the convergence for a broad family of problems. Empirical studies in various settings show the effectiveness of the proposed algorithm.

Acknowledgement

It is my great honour to take this chance to express my sincere gratitude to the people who have offered me guidance, help and support on the way to my PhD degree.

I would like to sincerely thank my supervisors, Prof. Irwin King and Prof. Michael R. Lyu, for their advices, encouragements, patience and support. I am grateful to my supervisors for all the time and efforts they have devoted to my research in the Chinese University of Hong Kong. Without their kind support and guidance, I would not seize the chance of my PhD study in Hong Kong.

I would like to thank Prof. Tianbao Yang at the Department of Computer Science in University of Iowa, who mentored me when I visited University of Iowa in the summer of 2017. The viewpoints and instructions from Prof. Yang helped me form important research results in this thesis.

I would like to thank my thesis assessment committee members, Prof. Lai-Wan Chan, Prof. Yuk-Lap Yip and Prof. Zenglin Xu for their constructive comments and valuable suggestions to this thesis.

I would like to thank Prof. Minggao Gu at the Department of Statistics in the Chinese University of Hong Kong, who provided me help and suggestions when I had doubts in research and life. I also want to thank Prof. Wan-Sang Wah, who guided me in the first year of my

PhD study and gave me the freedom of transferring my research focus from financial engineering to machine learning.

I thank Han Shao, Haiqin Yang, Xixian Chen and Tong Zhao for their valuable guidance and contribution to the research work in this thesis. Many thanks to my group fellows: Shenglin Zhao, Hongyi Zhang, Yuxin Su, Cuiyun Gao, Hui Xu, Jichuan Zeng, Pinjia He, Jiani Zhang, Hou-Pong Chan, Wang Chen, Yaoman Li, Yue Wang, Shilin He, Pengpeng Liu, Jingjing Li, Yifan Gao, Haoli Bai, Wenxiang Jiao, Weibin Wu, who gave me encouragement and kind help.

I also thank my friends: Zhensong Zhang, Xinying Wang, Weicong Liu, Furui Liu, Fan Yang, Jingxi Xu, Mengyi Zhang, for their help in my PhD study. I am thankful to my roommates, Xin Ye and Chaojie Wang, who brought me a colourful life in the past four years.

I would like to thank my teammates, Gan Tan and Changning Wang, for their sharing, help and support during my PhD journey. I own the thanks to my good friend in Hong Kong, Zebin Luo, for his kindness and great help.

Last but the most important, I would like to sincerely thank my dear family. The deep love and constant support from my father, my mother, my brother and my sisters is the key to the achievements in my PhD study. Without their endless love and support, none of this would be possible.

To my beloved parents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning is a key and fast-developing domain in artificial intelligence which generally means that devices designed by human beings act intelligently (Fetzer, 1990). In many scenarios, the main task of machine learning is to automatically predict or infer an output with respect to certain input data. In the past two decades, there has been a tremendous surge in the study of machine learning, such as support vector machine (Cortes and Vapnik, 1995; Vapnik, 2013), neural networks (Goodfellow et al., 2016; Haykin et al., 2009; LeCun et al., 2015), and reinforcement learning (Mnih et al., 2015; Sutton and Barto, 1998; Szepesvári, 2010).

In the framework of Jordan and Mitchell (2015), machine learning algorithms can be categorized into three paradigms: supervised learning, unsupervised learning and reinforcement learning. Supervised learning problems, which usually turn out to be classification or regression problems, aim at predicting an output for the input of a testing sample based on a model well trained by a finite number of training samples. Each training sample in a supervised learning process consists of an input vector and a label. Basically, unsupervised learning

involves the analysis of unlabeled data, and assumptions about the data's structural properties are necessary. A common example of unsupervised learning is to partition the unlabeled data, i.e., clustering. Clearly, we have the supervised learning paradigm if training samples contain the information of label, and we have the unsupervised learning paradigm if training samples do not contain any information of label. In practice, there do exist scenarios in which training samples provide intermediate information for algorithms, and these scenarios cannot be classified as either supervised learning or unsupervised learning. In other words, machine learning algorithms predict an output in each round of learning, and then receive a feedback with respect to the predicted output. Reinforcement learning characterizes the cases where training samples have intermediate information. Simplified versions of reinforcement learning are known as bandit problems (Jordan and Mitchell, 2015), where training samples are a set of bandits. For bandit problems, it is assumed that a payoff is observed for each play of an arm by machine learning algorithms, and the goal of bandit problems is set as maximizing cumulative payoffs for a certain number of rounds in sequential decisions, or identifying the optimal bandit that is also called the optimal arm in a given decision set.

The problem of learning in stochastic bandits (LSB) belongs to the paradigm of reinforcement learning. The learning process of LSB is to successively play bandits in a decision set, and in each round of playing bandits, machine learning algorithms receive a stochastic payoff with respect to the chosen bandit. Practical applications of LSB include online personalized recommendations (Li et al., 2010) and online resource allocations (Badanidiyuru et al., 2014).

The focus of this thesis is to develop efficient algorithms for LSB

problems. Efficient LSB requires algorithms to have provable performance guarantees, as well as practical implementations. There are three challenges to develop efficient algorithms for LSB problems. The first challenge is LSB with the mean-variance metric, which is less investigated in previous studies. The second challenge is LSB under heavy tails, which generalizes the results in previous studies. The third challenge is LSB with nonlinear payoff functions. By solving these three challenges, we enhance the applicability of LSB in the real world. The rest of this chapter is organized as follows. In Section 1.1, we present the background of this thesis. In Section 1.2, we give five practical examples motivating the study of LSB problems. In Section 1.3, we describe the details of the challenges for LSB problems and summarize the contributions of this thesis. In Section 1.4, we list the common notations used in the ensuing chapters. Finally, we give the thesis structure in Section 1.5.

1.1 Background

The model of multi-armed bandits (MAB), which is the simplest LSB problem, has attracted researchers for almost a century. In the past decade, practitioners have successfully applied MAB to real scenarios with satisfactory performances. The most significant application of MAB is online news personalized recommendations (Li et al., 2010). The original problem of learning in stochastic bandits dates back to 1933 by Thompson (1933), which lies in the domain of probability and statistics. Thompson asked whether or not it was possible to differentiate two probabilities via finite samples, and was motivated by the real problem of identifying a good treatment in clinical trials. The pioneering study by Thompson answers the above question affirmatively.

It is worth mentioning that a formal characterization of bandits was proposed by Robbins (1952), which later is named as MAB. Bandit problems have attracted the attention of researchers and practitioners because the central issue in bandit problems is decision making under uncertainty. Real-world applications of MAB include clinical trials, online recommendations and resource allocations. In the era of big data, samples arrive on the fly. Machine learning algorithms for on-the-fly data can be fundamentally applied into sequential decision making. To lay a basis for our discussion, the model of MAB can be briefly described below.

The model of multi-armed bandits (MAB)

Known parameters: the number of arms K, and the number of rounds $T \geq K$.

Unknown parameters: K probability distributions $\mathcal{P}_1, \cdots, \mathcal{P}_K$ on $[0, 1]$.

For each round $t = 1, \cdots, T$

Select an arm $x_t \in \{1, \cdots, K\}$.

Observe a stochastic payoff of arm x_t, $y_t(x_t) \sim \mathcal{P}_{x_t}$.

From the above learning process of MAB, we observe that the feedback from the chosen arm is noisy and not a true label. This finding reveals that the feedback contains intermediate information, and thus the learning of MAB belongs to the paradigm of reinforcement learning. In Chapter 2 of this thesis, we will further distinguish MAB from the general model of reinforcement learning.

In general, there are two types of goals for MAB problems (Bubeck et al., 2012). One common goal is to maximize the cumulative payoffs over a sequence of rounds for playing bandits. In this case, algorithms

for MAB problems have limited knowledge about the mechanism of generating payoffs, and need to learn the distribution of each arm while playing bandits. Thus, there exists an intrinsic trade-off between exploration and exploitation. The other goal is to identify the best arm among the given decision set, which is called pure exploration of MAB.

In the domain of bandit problems, researchers evaluate an algorithm via theoretical guarantees, as well as empirical verifications. For the goal of maximizing cumulative payoffs, the evaluation metric is regret, which is defined as the summation of differences between a clairvoyance and a bandit algorithm over rounds. For each round of playing bandits, the difference is usually characterized by the gap between the mean of the optimal arm based on the clairvoyance and the mean of the chosen arm by the bandit algorithm. The first theoretical results, including the lower bound of MAB and the upper bound of a bandit algorithm, were formally developed by Lai and Robbins (1985). For the goal of pure exploration, the evaluation metric is probability of error or sample complexity. In the past two decades, it witnessed a lot of interesting results for various MAB problems (Bubeck et al., 2012).

In this thesis, our focus is LSB, which is more general than MAB. The differences between LSB and MAB are briefly discussed as follows. The first difference is the evaluation metric can be generalized into mean-variance in LSB. The second difference comes from noise distributions of arms. In previous studies of MAB, researchers usually assume sub-Gaussian noises in feedbacks. We generalize noise distributions into heavy tails in LSB. The third generalization of LSB is the decision set can be chosen as a continuous convex set. The fourth difference is the underlying payoff functions can be nonlinear in LSB, which is more practical than MAB and linear stochastic bandits.

1.2 Motivation

The model of MAB is a simple abstraction of reality for decision making with uncertainty. In practice, one can usually obtain a (stochastic) feedback after making a decision. The general question is how to make sequential decisions with a good quality, as well as a theoretical guarantee. In this section, we list five motivating examples to support the research topics of MAB and LSB.

1.2.1 News Recommendation

News recommendation is a stylized and significant task in Internet. The problem of news recommendation requires machine learning algorithms to capture human behaviours. Without loss of generality, we assume that a user visits a news website with a database in the backend, and the website equipped with machine learning algorithms recommends a subset of all news for the user. If the user clicks an item among the recommended set, the feedback is recorded by one. Otherwise the feedback is zero. Due to the stochastic behaviour of human beings, the problem of news recommendation for a user can be formalized by an MAB model, where the mean of each arm refers to the true underlying preference of the user to each item in the recommended set.

Since the database of news is dynamic with time evolution, a bandit algorithm needs to conduct online recommendation, which is essentially sequential decision making. More importantly, the current action recommended by the bandit algorithm depends on the previous exploration of human behaviours. An intuitive trade-off between exploration and exploitation occurs in online recommendations of news. The power of MAB in solving the problem of news recommendation has been empirically demonstrated by Li et al. (2010).

The extension of online recommendation of news can be various, e.g., online advertising in Internet shopping and sponsored search. For online advertising, a user visits a website and the website demonstrates one of many advertisements. The problem is how to display an advertisement that will be clicked as much as possible by potential clients. For sponsored search, the problem is to return a subset of links, which are the most useful and significant for the client using the search engine. Recent investigations on advertising via MAB can be found in Schwartz et al. (2017).

In traditional recommendations, the goal is usually to maximize cumulative payoffs, which is related to the identification of the arm with the largest mean among the decision set. However, in practice, the mean information sometimes is not enough, and the variance of each arm should also be considered. A generalization of the traditional bandit problems is to optimize the metric of mean-variance in LSB. Thus, we will investigate the problem of pure exploration of mean-variance in Chapter 3.

1.2.2 Clinical Trials

The problem of clinical trials is to determine a treatment for patients via sequential selections. A patient visits a doctor and the doctor determines one of several potential treatments based on symptoms of the patient. After the patient taking the treatment, the doctor evaluates the treatment effectiveness via the feedback from the patient. The doctor encounters the trade-off between exploring the underlying best treatment and exploiting the empirically satisfying treatment. Intuitively, the best case is that the disease can be cured as fast as possible. The problem of clinical trials can be modelled by an MAB.

Recent studies for clinical trials with MAB can be found in Villar et al. (2015). The problem of clinical trials can be solved by two methodologies: Bayesian approach or frequentist approach. Each methodology enjoys its own characteristics. For example, Bayesian approach needs the assumption of a specific distribution for each arm. We will show more differences of the two approaches in Chapter 2. In fact, these two methodologies can be applied into solving many other variants of MAB.

One notorious issue of clinical trials is that feedbacks from patients are too noisy. The noisy feedbacks might not follow the assumption of sub-Gaussian in the traditional MAB, and even can be heavy tails. It is surprising to find that less effort has been devoted to the problem of MAB with heavy tails. In this thesis, we will solve bandit problems with heavy tails in Chapter 4 and Chapter 5.

1.2.3 Network Routing

In the era of big data, Internet increasingly influences daily life of human beings. A significant problem in Internet is network routing, where algorithms try to direct internet traffic as fast as possible among a large amount of network nodes. Given a package of data, network routing algorithms need to real-time identify the shortest path from the original node to the destination node.

In the problem of network routing, the decision set of bandits is the set of all potential network paths from the original node to the destination node. The stochastic feedback of selecting a path is the realized time consumption for sending a packet via the chosen path. The interesting investigations on network routing with MAB are by Badanidiyuru et al. (2013); Le Ny et al. (2008).

In previous work of Liebeherr et al. (2012), it has been pointed out that network delay affects the performance of routing. Besides, the delay of network routing is heavy-tailed. The problem of network routing also faces the heavy-tailed phenomenon, which reflects the significance of LSB with heavy tails. We notice that the first study of bandits with heavy tails was due to Bubeck et al. (2013a). In this thesis, we will make a progress for bandit problems with heavy tails.

1.2.4 Dynamic Pricing

In marketing, dynamic pricing plays an important role in attracting potential customers. Generally, the purchase intention of customers is greatly affected by the price of a product. A company needs to real-time optimize the price of a product such that the final revenue of the product achieves a predefined target. Some interesting studies of dynamic pricing via MAB have been conducted by Misra et al. (2018); Xu et al. (2017a).

1.2.5 Online Resource Allocations

A common problem in logistics is to allocation resource to an agent, where the consumption of the resource is stochastic. The output of the consumption can be viewed as the payoff. The problem can be modelled by contextual bandits, which is a variant of MAB.

In online resource allocation, one popular application is portfolio management. The portfolio management not only contains noisy feedbacks following heavy tails, but also involves a nonlinear relationship of investment items. Inspired by the nonlinear relationship in payoffs, we investigate the problem of nonlinear stochastic bandits in Chapter 6.

1.3 Challenges and Contributions

With a rapid development in bandits, many variants have been proposed to solve practical problems. Recently, there have been interesting investigations based on the traditional MAB model, such as linear bandits (Auer, 2002; Yu et al., 2017b; Zhao and King, 2016), pure exploration of MAB (Audibert and Bubeck, 2010), risk-averse MAB (Sani et al., 2012; Yu et al., 2017a), cascading bandits (Kveton et al., 2015) and clustering bandits (Korda et al., 2016; Li et al., 2016). We present the challenges in LSB problems and then summarize the contributions of this thesis.

1.3.1 Challenges

Though plenty of achievements have been conducted in the domain of bandits, many open problems exist if we would like to close the gap between the theoretical model of MAB and application scenarios. Currently, there are three significant challenges in LSB shown as follows.

- **Bandit problems with mean-variance**

  Traditionally, the optimal arm in MAB refers to the arm with the highest mean. However, in many practical applications, e.g., portfolio selection, it is not sufficient to only consider the mean information for the optimal decision. Thus, it is a significant challenge to study bandit problems with the metric of mean-variance. Bandit problem with mean-variance leads to three technical issues. The first comes from the analysis of errors due to estimations of mean-variance in bandits. The second is how to design efficient bandit algorithms for solving problems via the analysis of estimation errors. The third is to upper bound the performance

of bandit algorithms. We will tackle the three technical issues in Chapter 3.

- **Bandit problems with heavy tails**

  The model of MAB with sub-Gaussian noises has been well investigated. However, it is surprising to find that less effort has been devoted to the topic of bandits with noises following heavy-tailed distributions. It is an urgent problem to investigate the bandits with heavy tails. In this thesis, we will study pure exploration and regret minimization of bandits with heavy tails, with novel and systematic theoretical guarantees. Specifically, by breaking the assumption of payoffs with sub-Gaussian noises in bandits, we assume that stochastic payoffs from bandits are with finite p-th moments, where $p \in (1, +\infty)$. We show the studies in Chapter 4 and Chapter 5.

- **Bandit problems with nonlinear payoffs**

  We extend the case of stochastic bandits with linear functions to nonlinear functions. It has been a notorious challenge to solve the stochastic bandits with nonlinear functions. We classify two scenarios: one is convex functions and the other is non-convex functions. For convex functions, there exist studies for stochastic bandits with convex functions, which also called stochastic bandits convex optimization. In this case, explicit gradient calculations may be computationally infeasible, expensive, or impossible. Previous algorithms are slower than stochastic optimization with gradient feedbacks due to an unavoidable dependence of their iteration complexities on the dimensionality of the problem. We will accelerate the convergence rate in stochastic bandit convex opti-

mization. Besides, we can extend the results of the convex setting to the non-convex setting. We show the study in Chapter 6.

1.3.2 Contributions

We focus on the problem of LSB in this thesis, and propose efficient bandit algorithms with practical implementation and provable performance guarantee. For a better understanding, we list main contributions of our study on efficient learning in stochastic bandits as follows.

- **Efficient pure exploration of MAB with mean-variance**

  Pure exploration of bandits has the goal of identifying the optimal arm in a given decision-arm set. Traditionally, the optimal arm refers to the arm with the highest mean. However, in many practical applications, e.g., portfolio selection, it is not sufficient to only consider the mean information for the optimal decision. Motivating by exploration of high-order statistics, we study the problem of Pure Exploration of Mean-Variance (PEMV) in bandits. We rigorously prove that the error resulting from the mean-variance estimation is sub-gamma. Then, we develop two efficient algorithms to tackle PEMV. Besides, with sub-gamma estimation noises, we derive upper bounds of the probability of error for the proposed algorithms. Finally, we conduct a series of experiments on synthetic and real-world datasets for the task of pure exploration. By comparing with two state-of-the-art algorithms, we demonstrate the proposed algorithms are superior and robust.

- **Fast pure exploration of MAB with heavy tails**

  Since heavy-tailed distributions are significant in real-world scenarios, we investigate the problem on pure exploration of MAB

with heavy-tailed payoffs by breaking the assumption of payoffs containing sub-Gaussian noises. In particular, we assume that stochastic payoffs from bandits are with finite p-th moments, where $p \in (1, +\infty)$. The main contributions of our study for this problem are three-fold. First, we technically analyze tail probabilities of empirical average and truncated empirical average for estimating expected payoffs in sequential decisions with heavy-tailed noises via martingales. Second, we propose two effective bandit algorithms based on different prior information (i.e., fixed confidence or fixed budget) for pure exploration of MAB generating payoffs with finite p-th moments. Third, we derive theoretical guarantees for the proposed two bandit algorithms, and demonstrate the effectiveness of two algorithms in pure exploration of MAB with heavy-tailed payoffs in synthetic data and real-world financial data.

- **Almost optimal algorithms for linear stochastic bandits with heavy tails**

  It is commonly assumed that payoffs are with sub-Gaussian noises. Under a weaker assumption on noises, we study the problem of Linear stochastic Bandits with hEavy-Tailed payoffs (LinBET), where the distributions have finite moments of order p, where $p \in (1, 2]$. We rigorously analyze the regret lower bound of LinBET as $\Omega(T^{\frac{1}{p}})$, implying that finite moments of order 2 (i.e., finite variances) yield the bound of $\Omega(\sqrt{T})$, with T being the total number of rounds to play bandits. The provided lower bound also indicates that the state-of-the-art algorithms for LinBET are far from optimal. By adopting median of means with a well-designed allocation of decisions and truncation based on historical infor-

mation, we develop two novel bandit algorithms, where the regret upper bounds match the lower bound up to polylogarithmic factors. To the best of our knowledge, we are the first to solve LinBET optimally in the sense of the polynomial order on T. Our proposed algorithms are evaluated based on synthetic datasets, and outperform the state-of-the-art results.

- **Acceleration of stochastic bandit optimization**

  We extend the learning in stochastic bandits with nonlinear payoffs. We investigate two settings: convex functions and nonconvex functions. We propose a generic approach for accelerating the convergence of existing algorithms to solve the problem of stochastic bandit optimization. Standard techniques for accelerating the convergence of stochastic bandit algorithms are by exploring multiple functional evaluations (e.g., two-point evaluation), or by exploiting global conditions of the problem (e.g., smoothness and strong convexity). Nevertheless, these classic acceleration techniques are necessarily restricting the applicability of newly developed algorithms. The key of our proposed generic approach is to explore a local growth condition (or called local error bound condition) of the objective function. The benefits of the proposed acceleration technique are: (i) it is applicable to both settings with one-point evaluation and two-point evaluation; (ii) it does not necessarily require strong convexity or smoothness condition of the objective function; (iii) it yields an improvement on convergence for a broad family of problems. Empirical studies in various settings demonstrate the effectiveness of the proposed acceleration approach.

Table 1.1: Common symbols used in the thesis.

symbol	description
$\triangleq$	definition
$\mathcal{A}$	a bandit algorithm
$\mathbb{N}$, $\mathbb{N}^+$	natural numbers, $\mathbb{N} \triangleq \{0, 1, \cdots\}$ and $\mathbb{N}^+ \triangleq \mathbb{N} \backslash \{0\}$
$\mathbb{R}$, $\mathbb{R}^+$	$\mathbb{R} \triangleq (-\infty, +\infty)$ and $\mathbb{R}^+ \triangleq (0, +\infty)$
$[T]$	$\{1, 2, 3, \cdots, T\}$
$\lvert S \rvert$	the cardinality of a finite set S
$\mathcal{P}$	a probability distribution
$\mathbb{E}[A]$	the expectation of a random variable A
$\mathbb{P}[\mathcal{E}]$	the probability of an event $\mathcal{E}$
$\mathcal{N}(\mu, \sigma^2)$	a normal distribution with mean μ and variance σ^2
$\mathcal{F}_t$	a filtration until time t
$\exp(\cdot)$	the exponential operation
$\text{poly}(\cdot)$	the polynomial operation
$\lVert x \rVert_2$	ℓ_2-norm of a vector x
$\langle x, y \rangle$ or $x^\top y$	inner product of vectors x and y
$\nabla f(x)$	gradient of function $f(x)$
$\partial f(x)$	sub-gradient of function $f(x)$
$\mathbf{0}$ and $\mathbf{1}$	a vector of all elements being zeros and ones

1.4 Notations

In this section, we list the common symbols used in this thesis in Table 1.1. Besides, since each chapter is intended to be self-contained, we will give the detailed descriptions of notations for each chapter.

1.5 Thesis Structure

The rest of this thesis is organized as follows.

- Chapter 2

  In this chapter, we present a survey for the research topic of stochastic bandits, especially for MAB. In Section 2.1, we present the theoretical advancements in stochastic bandits, which include two settings: pure exploration and regret minimization. In Section 2.2, we give a discussion of bandit algorithms based on methodology. The common methodology in bandits is two-fold: one is frequentist approach and the other is Bayesian approach. In Section 2.3, we present a taxonomy of bandits.

- Chapter 3

  In this chapter, we investigate pure exploration of mean-variance of bandits. We give an introduction of pure exploration of bandits in Section 3.1. In Section 3.2, we present the preliminary and related work. In Section 3.3, We list the assumptions for pure exploration of mean-variance of bandits and formally define the problem. In Section 3.4, we propose two bandit algorithms and show their theoretical performance. In Section 3.5, We rigorously prove the theorems of the two algorithms. In Section 3.6, we conduct experiments to verify the performance of the proposed

algorithms. In Section 3.7, we give conclusive remarks of this chapter.

- Chapter 4

  In this chapter, we investigate the problem on pure exploration of MAB with heavy-tailed payoffs by breaking the assumption of payoffs with sub-Gaussian noises in bandits. We assume that stochastic payoffs from bandits are with finite p-th moments, where $p \in (1, +\infty)$. In Section 4.1, we present the background information of the problem. In Section 4.2 and Section 4.3, we give the preliminary and related work. In Section 4.4, we propose two algorithms based on the settings of fixed confidence and fixed budget. In Section 4.5, we show the proofs of the theorems for the proposed algorithms. In Section 4.6, we conduct a series of experiments in synthetic and real-world datasets for demonstrating the efficiency of the algorithms. In Section 4.7, we give conclusive remarks of fast pure exploration of MAB with heavy tails.

- Chapter 5

  In this chapter, we investigate the problem of linear stochastic bandits with heavy-tailed payoffs. The heavy-tailed payoffs refer to the distributions of feedbacks have finite moments of order p with $p \in (1, 2]$. In Section 5.1, we present an introduction and show the significance of the problem. In Section 5.2, we give the preliminary and related work. In Section 5.3, we prove the regret lower bound of the problem. In Section 5.4, we develop two bandit algorithms based on median of means and truncation. In Section 5.5, we give the proofs of the theoretical guarantees. In Section 5.6, we conduct experiments to show the superiority of

the proposed algorithms. In Section 5.7, we conclude the study on the problem of linear stochastic bandits with heavy-tailed payoffs.

- Chapter 6

  In this chapter, we investigate the problem of stochastic bandit optimization, and most of our efforts are on the setting of convex functions. In Section 6.1, we present an introduction of the problem. In Section 6.2 and Section 6.3, we give the related work and then present the notations and preliminary. In Section 6.4, we propose a generic framework to solve stochastic bandit convex optimization. In Section 6.5, we give the theoretical proofs of the algorithm. In Section 6.6, we conduct a series of experiments in real-world datasets to show the efficiency of the proposed algorithm. In Section **??**, we extend the results into the setting of non-convex functions. In Section 6.7, we give the conclusive remarks for the problem of stochastic bandit optimization.

- Chapter 7

  In this chapter, we summarize this thesis and present three potential directions for future work. In particular, we conclude this thesis in Section 7.1. Then, we list three points as future directions in Section 7.2.

□ **End of chapter.**

Chapter 2

Learning in Stochastic Bandits: A Survey

In this chapter, we review the current research progress in the area of MAB. In particular, we give a whole view of theoretical guarantees in bandit problems from two aspects: regret minimization and pure exploration. Then, we discuss the methodology for solving stochastic MAB. Besides, in order to characterize the development of bandits, we give the taxonomy of bandits in the past two decades.

As mentioned in Chapter 1, the model of MAB is to sequentially select an arm among a decision-arm set. The learning of MAB contains bandit feedback because, after each round of selection, algorithms observe a payoff with respect to the chosen arm. Learning in stochastic bandits generally refers to the setting that the feedback from the chosen arm is stochastic. The inherent challenge of learning in stochastic bandits is to balance the trade-off between exploring the true best arm and exploiting the empirically optimal arm. Since the trade-off of exploration and exploitation is ubiquitous, bandit problems have played an essential role in many industrial domains, e.g., online recommenda-

tions.

2.1 Theoretical Advancements

Bandit problems are originated from Thompson (1933), where different treatments are available for a patient and the task is to determine the appropriate treatment for the patient based on historical records. The final goal of MAB can be defined from two perspectives. One is maximization of cumulative payoffs and the other is identification of the best arm among the decision set. The latter one is also called pure exploration of bandits. In the following, we will summarize theoretical advancements based on the two perspectives of the final goal for MAB.

2.1.1 Regret Minimization

To maximize cumulative payoffs over a number of rounds for sequential decisions, we assume the existence of an underlying optimal arm, and analyze the difference between the true optimal arm and the selection arms. To analyze the performance of a bandit algorithm $\mathcal{A}$ over T rounds, we usually compare the realized cumulative rewards via $\mathcal{A}$ with rewards from the true optimal arm. The notion of regret is introduced to investigate the theoretical performance of $\mathcal{A}$, which is defined as

$$\bar{\mathbf{R}}(\mathcal{A}, T) \triangleq \max_{i=1,\cdots,K} \sum_{t=1}^{T} y_t(i) - \sum_{t=1}^{T} y_t(x_t), \tag{2.1}$$

where x_t is the selected arm at round t, $y_t(i)$ is the reward of arm i at round t, and $y_t(x_t)$ is the observed reward of selected arm x_t. Both rewards $y_t(i)$ and the algorithm's choices x_t might be stochastic, which entails the expected regret as

$$\mathbb{E}\left[\bar{\mathbf{R}}(\mathcal{A}, T)\right] = \mathbb{E}\left[\max_{i=1,\cdots,K} \sum_{t=1}^{T} y_t(i) - \sum_{t=1}^{T} y_t(x_t)\right]. \tag{2.2}$$

In practice, a notion of pseudo-regret is adopted, which is

$$\mathbf{R}(\mathcal{A}, T) \triangleq \max_{i=1,\cdots,K} \mathbb{E}\left[\sum_{t=1}^{T} y_t(i) - \sum_{t=1}^{T} y_t(x_t)\right]. \tag{2.3}$$

It is obvious that the pseudo-regret is weaker that the expected regret, i.e., $\mathbf{R}(\mathcal{A}, T) \leq \mathbb{E}[\bar{\mathbf{R}}(\mathcal{A}, T)]$. But the pseudo-regret is more about the statistical optimal strategy. In the following, we always adopt the pseudo-regret rather than the expected regret as a performance measurement for bandit algorithms.

We show the theoretical advancements of stochastic bandits with K arms in Figure 2.1. In particular, for regret minimization, the origin of MAB dates back to 1933 (Robbins, 1952; Thompson, 1933). In Robbins (1952), the MAB problem was formally proposed. In Lai and Robbins (1985), the first distribution-dependent asymptotic analysis of stochastic MAB was developed with the regret lower bound being $\Omega(\log(T))$, and the regret upper bound of the proposed bandit algorithm matched the lower bound up to constant factors. Agrawal (1995) proposed the sample mean index policy and asymptotic analysis of bandits, which naturally leads to a finite-time analysis. Auer et al. (2002a) proposed the upper confidence bound (UCB) algorithm and gave the finite-time analysis for regret of MAB. A study of revisited UCB by Auer and Ortner (2010) developed an improved analysis of upper bounds for bandits, which inspired the investigation in Bubeck et al. (2013b). A new regret lower bound was proved as $\mathbf{R}(\mathcal{A}, T) \geq \Omega\left(\sum_i \log(T\Delta_i^2)/\Delta_i\right)$, and it implies that the worst case of the regret lower bound is $\Omega(\sqrt{T})$ when $\Delta_i = \Theta(1/\sqrt{T})$. It is worth mentioning that bandit problems with Thompson sampling have been investigated by Agrawal and Goyal (2012, 2013a); Kaufmann et al. (2012), where problem-dependent and problem independent bounds were developed.

A natural and important variant of MAB is linear stochastic bandits

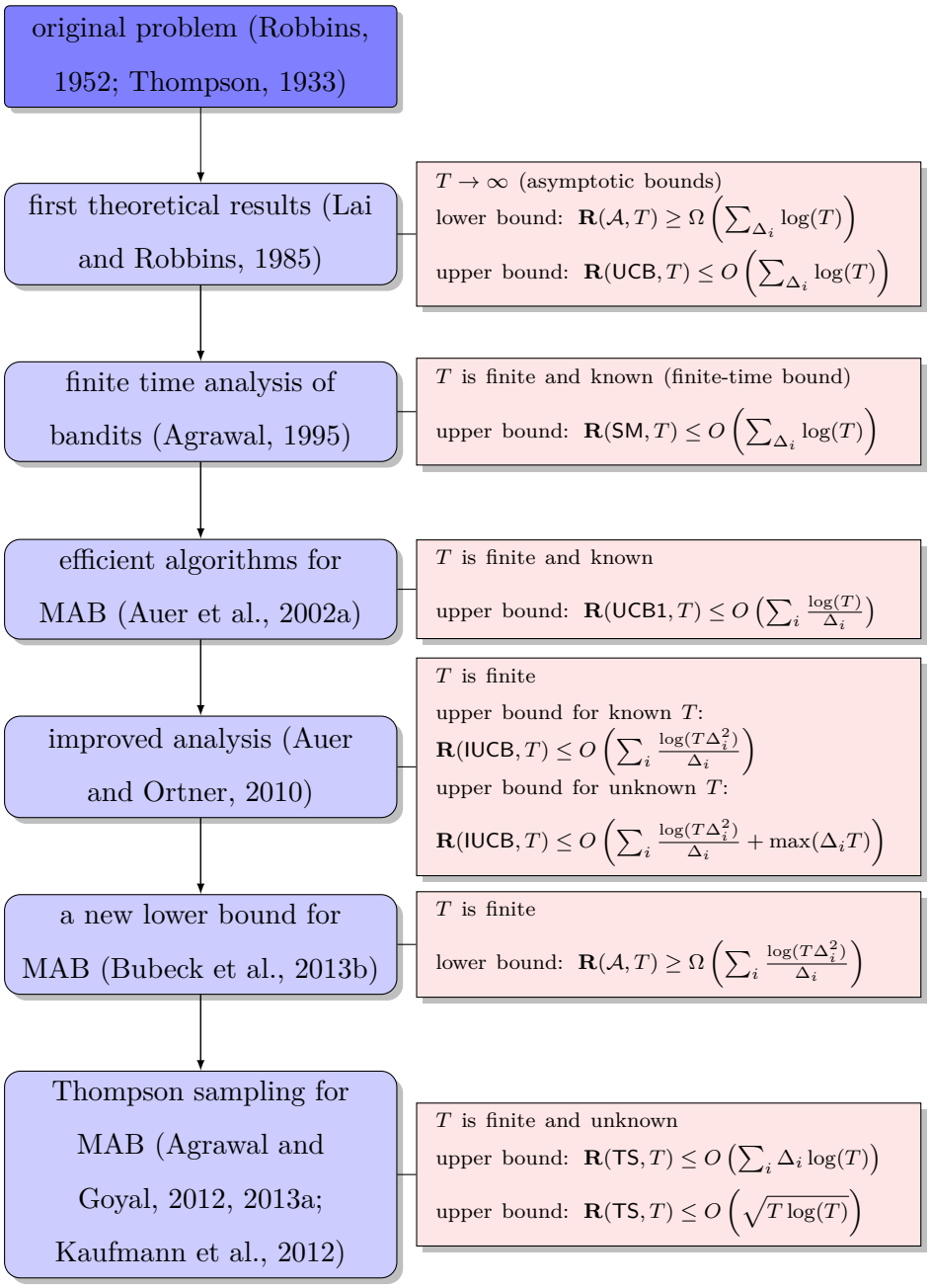


Figure 2.1: Theoretical advancements in MAB, and Δ_i denotes the mean difference between the true optimal arm and the i-th arm with $i \in [K]$.

with the expected payoff of each arm satisfying a linear mapping from the arm information to a real number. The model of linear stochastic

bandits enjoys some good theoretical properties, e.g., there exists a closed-form solution of the linear mapping at each time step in light of ridge regression. The linear mapping of linear stochastic bandits requires a parameter, which is defined as θ_*. Given the decision set as $D \subseteq \mathbb{R}^d$, a bandit algorithm selects an arm $x_t \in D$ at time t, and observes a stochastic payoff $y_t(x_t) \triangleq x_t^\top \theta_* + \epsilon_t$, where ϵ_t is a noise. Then, the pseudo-regret in linear stochastic bandits is defined as

$$\mathbf{R}(\mathcal{A}, T) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} y_t(x_t^*) - \sum_{t=1}^{T} y_t(x_t)\right], \tag{2.4}$$

where x_t^* is the true optimal arm at t.

There are fruitful results in linear stochastic bandits, which are shown in Figure 2.2. The model of linear stochastic bandits, which is also termed as associative reinforcement learning with linear payoff functions by Abe and Long (1999); Auer (2000), is a special case of reinforcement learning (Kaelbling, 1994; Sutton and Barto, 1998). In Auer (2002), a bandit algorithm with linear regression was developed to solve linear stochastic bandits, and the regret upper bound of the algorithm was $O\left(\sqrt{T}\text{poly}(\log(T))\right)$. Online convex optimization under bandit feedbacks was investigated by Flaxman et al. (2005), which covered the case of linear stochastic bandits. Since the method of Flaxman et al. (2005) was based on noisy gradient descent, the proposed algorithm could solve bandit problems when the decision set is continuous and convex. The lower bound of linear stochastic bandits was addressed by Dani et al. (2008a), which was shown as $\Omega(d\sqrt{T})$ with d denoting the dimension of the convex decision set. An improvement on the logarithmic factor for the theoretical analysis in linear stochastic bandits was conducted in Abbasi-Yadkori et al. (2011), and the empirical improvement was also verified. Note that the finite-arm setting was studied by Chu et al. (2011) with an improvement on the dimension fac-

tor. The proposed algorithm LinUCB has been successfully applied into online personalized news recommendations (Li et al., 2010). Recently, the problem of linear stochastic bandits with Thompson sampling was well studied by Agrawal and Goyal (2013b).

A topic close to stochastic bandits is adversarial bandits (Auer et al., 1995). There is no statistical assumption for the generation of payoffs in adversarial bandits. The model of adversarial bandits can be viewed as a game between a player and an adversary, where for each round of the game, the player chooses an arm in the decision set and the adversary determines a payoff for the player. The payoff could be non-stochastic or even arbitrary. Intuitively, the adversary should be oblivious, because if it is malicious, then the regret lower bound of the two-player game is linear. In the line of adversarial bandits, many theoretical results have been developed (Alon et al., 2017; Auer et al., 2002b; Gerchinovitz and Lattimore, 2016; Uchiya et al., 2010). Most of algorithms for adversarial bandits enjoy an upper bound $O(\sqrt{T})$. Since the topic of this thesis is stochastic bandits, we will not list the details in adversarial bandits.

There are researchers trying to develop bandit algorithms for both stochastic and adversarial worlds (Audibert and Bubeck, 2009; Bubeck and Slivkins, 2012). In Audibert and Bubeck (2009), two algorithms were developed for closing the gap of a logarithmic factor between upper bounds of the previous methods and the lower bound in MAB. The first work to develop a unified algorithm for both stochastic and adversarial bandits was by Bubeck and Slivkins (2012). The bandit algorithm named SAO in Bubeck and Slivkins (2012) achieved the regret upper bound $O(\sqrt{T}\text{poly}(\log(T)))$ for adversarial bandits and achieved the regret upper bound $O(\text{poly}(\log(T)))$ for stochastic bandits. Auer

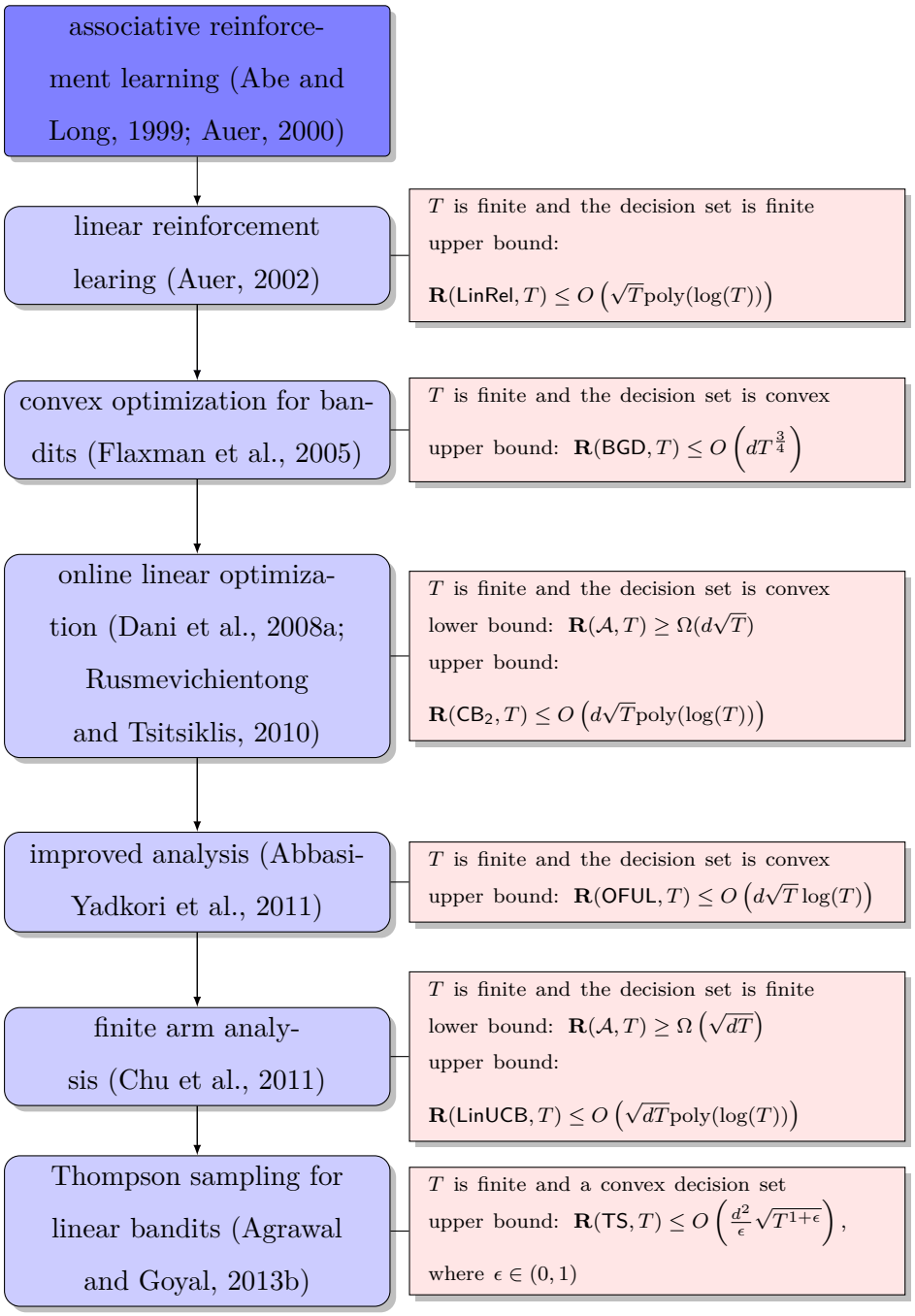


Figure 2.2: Theoretical advancements in linear stochastic bandits.

and Chiang (2016) found that it is impossible to develop algorithms to simultaneously achieve $O(\log(T))$ for stochastic bandits and $O(\sqrt{T})$ for

adversarial bandits. Later, by comparing with SAO, Seldin and Lugosi (2017); Seldin and Slivkins (2014) proposed practical EXP3++ with improved regret upper bounds to solve the problem of both stochastic and adversarial bandits.

2.1.2 Pure Exploration

For pure exploration, its goal is to find the optimal arm after exploration among a given decision-arm set (Audibert and Bubeck, 2010; Bubeck et al., 2009; Chen et al., 2014; Gabillon et al., 2012, 2016; Jamieson and Nowak, 2014). It has been pointed out that pure exploration in MAB has many applications, such as communication networks and online advertising.

In the task of pure exploration, there are two settings: fixed confidence and fixed budget. For the setting of fixed confidence, $\mathcal{A}$ receives the information of the probability of error at the beginning, and $\mathcal{A}$ generates an output when a certain condition related to the probability of error is satisfied. For the setting of fixed budget, $\mathcal{A}$ receives the information of the total number of rounds at the beginning, and $\mathcal{A}$ generates an output at the end of exploration.

For pure exploration of MAB with the sub-Gaussian assumption, theoretical guarantees have been well studied. Specifically, in the setting of fixed confidence, the first distribution-dependent lower bound of sample complexity was developed by Mannor and Tsitsiklis (2004), which was $\sum_{k\in[K]} \Delta_k^{-2}$. Even-Dar et al. (2002) originally proposed a bandit algorithm via successive elimination for bounded payoffs with an upper bound of sample complexity matching the lower bound up to a multiplicative logarithmic factor. Karnin et al. (2013) proposed an improved bandit algorithm, which achieved an upper bound of sample

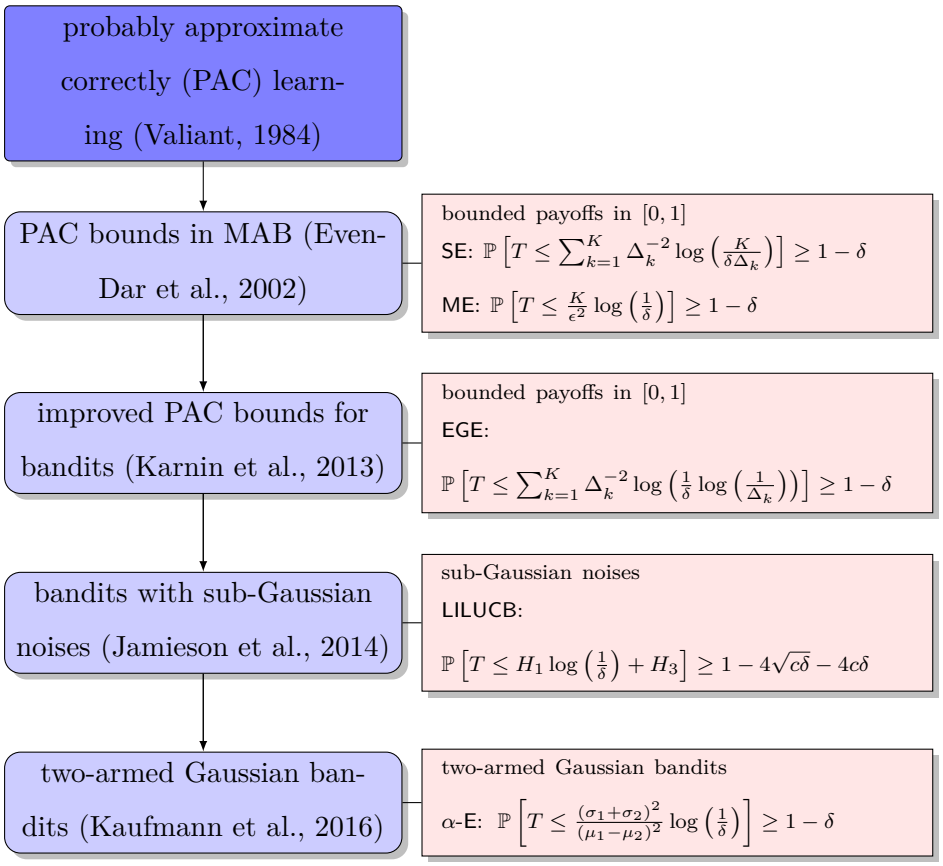


Figure 2.3: Theoretical advancements of pure exploration in MAB with an input parameter of confidence $\delta \in (0,1)$.

complexity matching the lower bound up to a multiplicative doubly-logarithmic factor. Jamieson et al. (2014) proved that it is necessary to have a multiplicative doubly-logarithmic factor in the distribution-dependent lower bound of sample complexity. Jamieson et al. also developed a bandit algorithm via the law of the iterated logarithm algorithm for pure exploration of MAB, which achieved the optimal sample complexity. We show the theoretical advancements of pure exploration in MAB with an input parameter of confidence $\delta \in (0,1)$ in Figure 2.3.

In the setting of fixed budget under the sub-Gaussian assumption,

Audibert and Bubeck (2010) developed a distribution-dependent lower bound of probability of error, and provided two algorithms, which enjoyed the optimality of probability of error up to logarithmic factors. Gabillon et al. (2012) proposed a unified algorithm for fixed budget and fixed confidence, and addressed ϵ-optimal learning in best arm identification of MAB. Karnin et al. (2013) proposed a bandit algorithm via sequential halving to improve probability of error by a multiplicative constant. It is worth mentioning that Kaufmann et al. (2016) investigated best arm identification of MAB under Gaussian or Bernoulli assumption, and provided lower bounds in terms of Kullback-Leibler divergence. We show the theoretical advancements of pure exploration in MAB with an input parameter of budget T in Figure 2.4.

There also exists a generalized variant for the problem of pure exploration of MAB, which is termed combinatorial pure exploration (Chen et al., 2017a,b, 2014; Gabillon et al., 2016).

2.2 Methodology

There are two fundamental methodologies for stochastic MAB, which are UCB and Thompson Sampling. The main idea of UCB is optimism in face of uncertainty. We assume that a bandit algorithm has obtained historical data on the arms and has to decide which arm to play at the next round. An estimate related to the true mean of each arm based on the data is constructed. With high probability, the true values lie in a region around the estimate. The algorithm plays the empirical optimal arm with respect to the supremes of the regions.

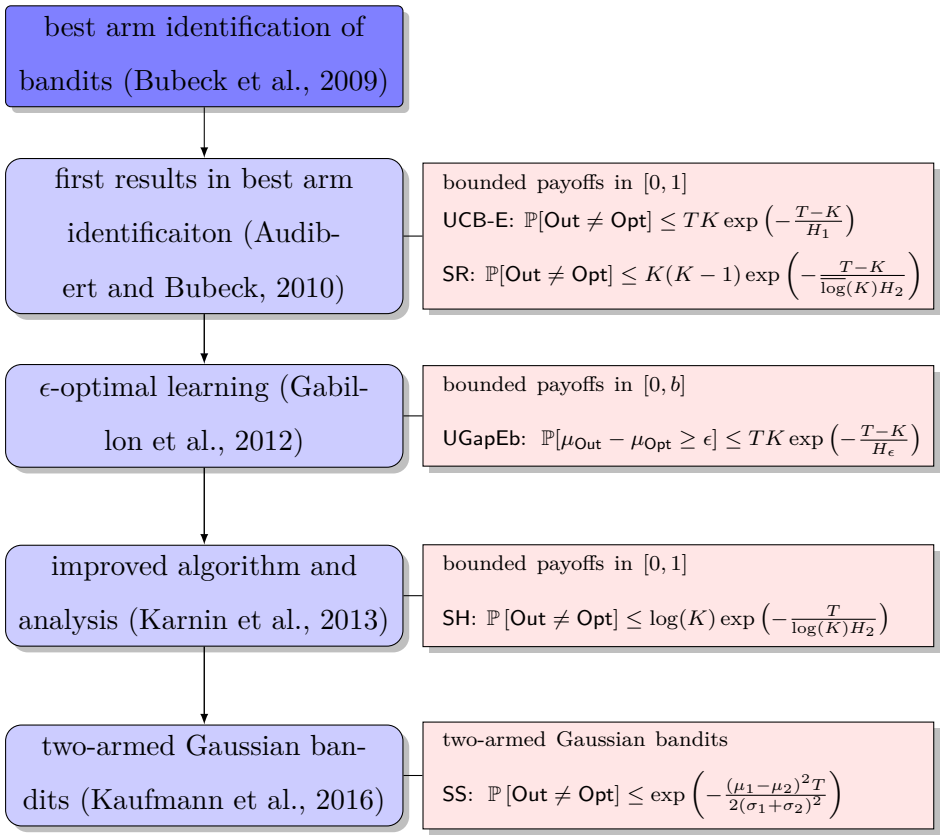


Figure 2.4: Theoretical advancements of pure exploration in MAB with an input parameter of budget T.

2.2.1 Frequentist Approach

The idea of UCB had been investigated in Agrawal (1995); Lai and Robbins (1985). In 2002, Auer et al. (2002a) proposed an algorithm named UCB1 with finite-time analyses. Audibert et al. (2009) took the variances of payoffs into consideration and proposed an algorithm named UCB-V. Audibert and Bubeck (2009) then proposed an algorithm named MOSS based on UCB1 with modified indexes. The algorithm MOSS achieves the distribution-free optimal rate while still having a distribution-dependent rate, which is logarithmic of the number of plays. Garivier and Cappé (2011) proposed KL-UCB, which

satisfied a uniformly better regret bound than UCB1 and other variants and reached the lower bound of Lai and Robbins (1985) in the special case of Bernoulli rewards.

2.2.2 Bayesian Approach

Instead of the method of UCB, there is a bayesian method named Thompson Sampling (TS) (Thompson, 1933), which assumes a simple prior distribution on the parameters of each arm's distribution. But the question of the optimality of TS had been open before Agrawal and Goyal (2012, 2013b); Kaufmann et al. (2012) provided analyses for it. Agrawal and Goyal (2012) derived the first logarithmic upper bound for regret with TS in expectation. Kaufmann et al. (2012) provided an upper bound of TS in terms of finite time for MAB with Bernoulli payoffs, which matched the regret lower bound proposed by Lai and Robbins (1985). Agrawal and Goyal (2013b) provided the first problem-independent regret upper bound of $O\left(\sqrt{T\log(T)}\right)$.

Most algorithms adopting Thompson Sampling are similar (Agrawal and Goyal, 2012, 2013b; Kaufmann et al., 2012). The algorithms construct a posterior distribution for the estimates based on the data. At the next round, the algorithm draws a sample from the posterior distribution of each arm, and selects the empirical optimal arm according to the samples.

The method of Gittins indices is applied to the MAB model where the arms are associated with K Markov processes, each arm with its own state space. The underlying stochastic transition matrices are typically assumed to be known, and thus the method of Gittins indices (**?**) provides a way to efficiently compute the optimal arm.

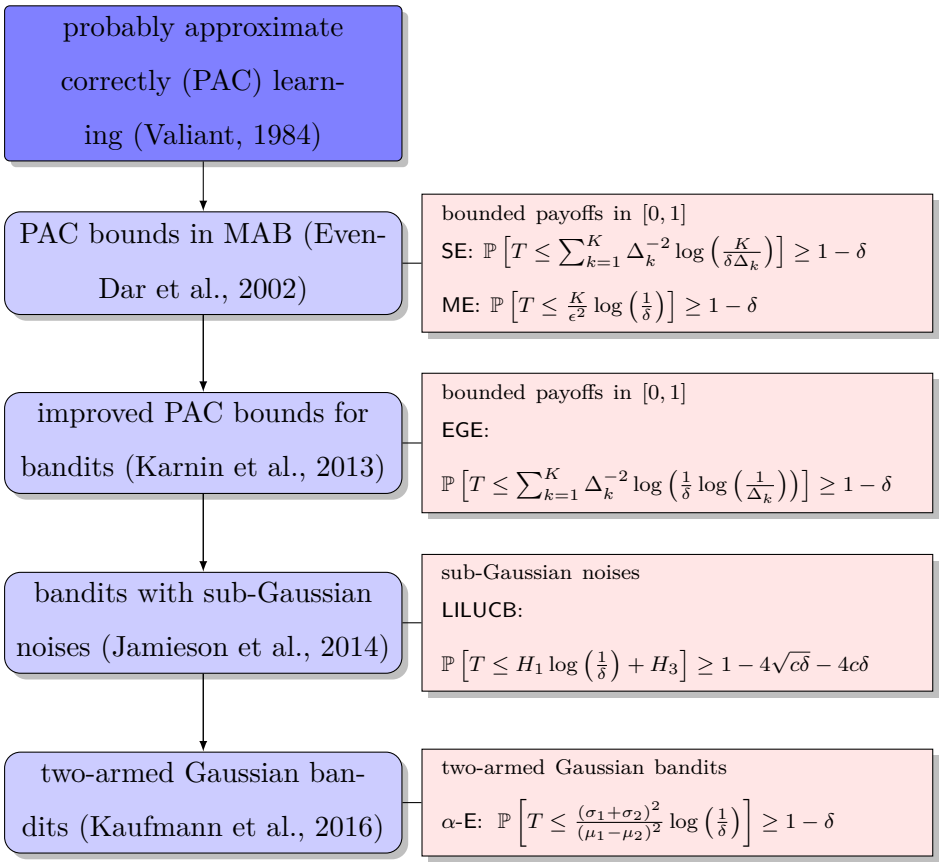


Figure 2.5: A taxonomy of bandits.

2.3 Taxonomy

We construct a taxonomy in Figure 2.3. In the above figure, we know that there are various variants in bandit problems.

□ **End of chapter.**

Chapter 3

Pure Exploration of Mean-Variance

The popular decision-making model of MAB elegantly characterizes a wide class of problems for sequential learning with stochastic feedbacks. A predominant characteristic of MAB is the trade-off between exploration and exploitation for decisions. In previous studies, most of bandit algorithms aim at maximizing cumulative payoffs over a number of rounds. In this chapter, we investigate pure exploration of mean-variance for bandits with K arms.

One non-trivial branch of MAB is pure exploration, of which the goal is to identify the optimal arm in a given decision-arm set. Traditionally, the optimal arm refers to the arm with the highest mean. However, in many practical applications, e.g., portfolio selection, it is not sufficient to only consider the mean information for the optimal decision. With the motivation of exploring high-order statistics, we study the problem of Pure Exploration of Mean-Variance (PEMV) in bandits. The problem of PEMV leads to three technical challenges. The first comes from the analysis of errors due to estimations of mean-variance

in pure exploration of MAB. The second is how to design efficient bandit algorithms for solving PEMV via the analysis of estimation errors. The third is to upper bound the probability of error for selecting a sub-optimal arm by a bandit algorithm. To solve the challenges, we rigorously prove that the error resulting from the mean-variance estimation is sub-gamma. Then, we develop two efficient algorithms to tackle PEMV. Besides, with sub-gamma estimation noises, we derive upper bounds of the probability of error for the proposed algorithms. Finally, we conduct a series of experiments on synthetic and real-world datasets for the task of pure exploration. By comparing with two state-of-the-art algorithms, we demonstrate the proposed algorithms are superior and robust.

3.1 Introduction

The model of MAB well tackles a large number of sequential-decision problems, such as personalized recommendations (Li et al., 2010) and online resource allocations (Lattimore et al., 2014). The inherent characteristic of MAB is the trade-off between exploration and exploitation. The first asymptotic theoretical guarantee of this class of models was developed by Lai and Robbins (1985). Recently, with the growth of research in machine learning and operations research, there emerges a surge of theoretical study on MAB and its variants (Bubeck et al., 2012; Cesa-Bianchi and Lugosi, 2006; Li et al., 2017; Zhou, 2015).

Traditional MAB algorithms aim at maximizing (expected) cumulative payoffs over a sequence of decisions. Given a clairvoyance knowing the optimal decision for each round, we can define a metric of regret, which is the difference of payoffs from the clairvoyance and a bandit algorithm. Intuitively, a small regret indicates good performance of

a bandit algorithm. It has been shown that, for any algorithm, the regret of MAB is at least logarithmic with respect to the total number of sequential decisions (Lai and Robbins, 1985).

To generalize applications of MAB, there have been various interesting investigations on its variants, such as bandits with side observations (Joseph et al., 2016; Langford and Zhang, 2008), pure exploration of MAB (Audibert and Bubeck, 2010; Carpentier and Locatelli, 2016; Chen et al., 2014), risk-averse MAB (Galichet et al., 2013; Sani et al., 2012), and dueling bandits (Yue et al., 2012). With real-world scenarios of medical trials and crowdsourcing, pure exploration is a fundamental variant of MAB for various decision-making problems, which usually sets the goal to find the optimal arm in a given decision-arm set at the end of exploration. It is worth mentioning that, for pure exploration, there is no explicit trade-off between exploration and exploitation for each round decision. It is suitable to view this decision-making model as two phases, i.e., first exploration and then exploitation.

In previous work, the optimal arm for pure exploration of MAB refers to the arm with the highest mean in the given decision set. However, there do exist practical scenarios where we cannot neglect high-order statistics of an arm in pure exploration. As a motivating example, consider two therapies in clinical trials for patients. After a sequential clinical testing, the doctor determines which of the therapies should be suggested for the next patient. The first therapy performs 80 in terms of treatments (with a maximum score of 100 denoting completely curing patients). Besides, the first one faces a risk of 20 to incur failures (or even, death). The second one can heal patients with a bit lower score (e.g., 75) than that of the first one, but facing a risk of 5. Then, a reasonable optimal suggestion for the next patient is to choose

the second therapy. This example reveals that the estimated expected payoff (i.e., the score in the example) should not be the only goal in decisions. It needs to incorporate risks into the optimal decision. The similar scenario does happen in pure exploration of the optimal decision for other real applications.

It is worth pointing out that pure exploration of high-order statistics (e.g., mean-variance) in MAB has been rarely investigated, which might be caused by different reasons. One possible reason is that high-order statistics may bring the divergence of estimation errors in sequential learning. Another reason could be that the probability of error for selecting a sub-optimal arm in pure exploration is too sensitive to high-order statistics, leading to failures in control.

Since pure exploration of MAB has played an important role in various practical applications, it is urgent and meaningful to study exploration of high-order statistics, e.g., mean-variance exploration, in bandits. To the best of our knowledge, whether or not pure exploration of high-order statistics in bandits is feasible remains an open problem.

In this chapter, we focus on the problem of Pure Exploration of Mean-Variance (PEMV) in MAB, and we answer the above question in the affirmative. Specifically, there are three main technical challenges in solving the problem of PEMV. The first comes from the analysis of the error between the estimation of mean-variance and the true mean-variance for pure exploration in MAB. The second is how to design efficient bandit algorithms for solving PEMV with the analysis of estimation errors. The third is how to bound the probability of errors for selecting a sub-optimal arm at the end of exploration. Here the optimal arm refers to the arm with the minimal mean-variance in the given decision set.

To solve the aforementioned challenges, based on the empirical estimation of mean-variance, we rigorously prove that the error resulting from the estimation is sub-gamma. We develop two bandit algorithms to solve the problem of PEMV under the sub-gamma noises. Besides, we derive upper bounds of the probability of error for the proposed algorithms. Based on two baselines in pure exploration, we demonstrate superiority and robustness of the proposed algorithms in synthetic and real-world datasets.

In summary, we make the following contributions in this chapter.

- We rigorously prove that the empirical estimation of mean-variance in sequential decisions incurs sub-gamma noise, where we adopt the technique of martingale.
- We develop two bandit algorithms, which are named respectively as PEMV.CB and PEMV.HALVING, to solve the problem of exploration of mean-variance in bandits. Moreover, we derive upper bounds of the probability of errors for the algorithms in the setting of a fixed budget.
- We evaluate the proposed algorithm via a series of experiments with synthetic and real datasets. By comparing the proposed algorithms with two state-of-the-art baselines, we demonstrate that the two algorithms have superior performance in pure exploration with mean-variance.

3.2 Preliminary and Previous Work

We first present related notations and definitions. Then, we give a literature review on pure exploration of bandits.

3.2.1 Notations and Definitions

The learning process of pure exploration in bandits for the fixed budget setting can be briefly summarized as follows. At the beginning, a learning algorithm $\mathcal{A}$ receives the fixed budget T, which is the total number of rounds to play bandits, and also is given a decision set with K arms. For each round $t \in [T]$ with $[T] \triangleq \{1, 2, \cdots, T\}$, $\mathcal{A}$ decides to play an arm $a_t \in [K]$. At the end of t, the algorithm observes a stochastic payoff $y_t(a_t)$, which is corresponding to the chosen arm. Then, at $t = T$ (sometimes $t < T$ if $\mathcal{A}$ is confident enough), the algorithm is required to output the optimal arm. The challenge is usually to upper bound the probability of error because of selecting a sub-optimal arm. Let $\mathbb{E}[\cdot]$ be the expectation of a random variable, and $\exp(\cdot)$ denote the exponential operation. Given a set $\Phi = \{\Phi_1, \cdots, \Phi_s\}$ with size $s = |\Phi|$, we denote by $\Phi \backslash \Phi_s$ the elimination of Φ_s in Φ. Given $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the greatest integer less than or equal to x, and $\lceil x \rceil$ is the least integer greater than or equal to x.

Definition 3.1. *(see Buldygin and Kozachenko, 1980) A random variable ζ is sub-Gaussian if there exists a constant $\bar{R} > 0$ such that*

$$\mathbb{E}[\exp(\lambda\zeta)] \leq \exp\left(\frac{\lambda^2 \bar{R}^2}{2}\right), \ \forall \lambda \in \mathbb{R}. \tag{3.1}$$

A random variable ζ satisfying Eq. (3.1) is also called $\bar{R}$-sub-Gaussian. Besides, we have $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}[\zeta^2] \leq \bar{R}^2$.

Definition 3.2. *(see Boucheron et al., 2013) A random variable ζ is sub-gamma on the right tail if*

$$\mathbb{E}[\exp(\lambda\zeta)] \leq \exp\left(\frac{\lambda^2 v}{2(1 - c\lambda)}\right), \ \forall \lambda \in \left(0, \frac{1}{c}\right), \tag{3.2}$$

where $v > 0$ is a variance factor, $c > 0$ is a scale parameter.

Definition 3.3. *The measure of mean-variance for a random variable* ζ *is defined as*

$$\omega(\zeta) \triangleq \sigma^2(\zeta) - \kappa\mu(\zeta), \tag{3.3}$$

where $\sigma^2(\zeta)$ *and* $\mu(\zeta)$ *are, respectively, the variance and the mean of* ζ*, and the coefficient* $\kappa \geq 0$ *is the risk tolerance factor. Besides, given* T *samples of random variable* ζ *as* $\{\zeta_t\}_{t=1}^T$*, we directly define the empirical mean and variance, respectively, as* $\hat{\mu}(\zeta_T) \triangleq \sum_{t=1}^T \zeta_t/T$ *and* $\hat{\sigma}^2(\zeta_T) \triangleq \sum_{t=1}^T (\zeta_t - \hat{\mu}(\zeta_T))^2/(T-1)$*. Then, the empirical mean-variance over* T *samples is* $\hat{\omega}(\zeta_T) \triangleq \hat{\sigma}^2(\zeta_T) - \kappa\hat{\mu}(\zeta_T)$*.*

Given K arms, let Opt denote the optimal arm with minimal mean-variance shown as Eq. (3.3). For $a \neq \mathsf{Opt}$, we introduce the sub-optimality metric between arms a and Opt as

$$\Delta_a \triangleq \omega(a) - \omega(\mathsf{Opt}), \tag{3.4}$$

where $a \in [K]$. Based on Eq. (3.4), we further define the minimal sub-optimality as $\Delta_* \triangleq \min_{a \neq \mathsf{Opt}, a \in [K]} \Delta_a$. Clearly, we have $\Delta_a \geq 0$. We introduce the notation $(a) \in [K]$ to denote the a-th best arm (with ties break arbitrarily), thus

$$\Delta_* = \Delta_{(1)} \leq \Delta_{(2)} \leq \Delta_{(3)} \leq \cdots \leq \Delta_{(K)}. \tag{3.5}$$

The sorted sequence of sub-optimality shown in Eq. (3.5) is helpful in analyzing the probability of error in algorithms. Inspired by Audibert and Bubeck (2010), we define the hardness of pure exploration with mean-variance as

$$\mathbf{H}_1 \triangleq \sum_{a=1}^K \frac{1}{\Delta_a^2}, \ \mathbf{H}_2 \triangleq \max_{a \in [K]} a\Delta_{(a)}^{-2}.$$

We generalize the concept of the above hardness into

$$\mathbf{H}_3 \triangleq \sum_{a=1}^K \frac{1}{\Delta_a}, \ \mathbf{H}_4 \triangleq \max_{a \in [K]} a\Delta_{(a)}^{-1}.$$

In the above definitions, we adopt Δ_a and $\Delta_{(a)}$ in denominators. These two generalized concepts are related to the theoretical analyses of the probability of error.

3.2.2 Previous Work

Pure exploration in MAB is an essential branch in decision-making problems, where the goal is to identify the optimal arm after exploration among a given decision set (Audibert and Bubeck, 2010; Bubeck et al., 2009; Chen et al., 2017a, 2014). It can be tailored to many applications, such as resource allocations and online advertising.

As mentioned, in previous work, the optimal arm is generally set as the arm with the highest expected payoff. For the problem PEMV, there has been very rare investigations. But there exist studies of risk control in bandits with strictly assumptions (Yu and Nikolova, 2013), where the payoffs are set as bounded, and the density function of payoffs is assumed to be continuously differentiable. In Yu and Nikolova (2013), a bandit algorithm named CuRisk was proposed to minimize risk for each round of decisions. But it did not discuss on pure exploration of high-order statistics.

It is worth mentioning that there also exists a closely related line of research to pure exploration with risk in MAB, which is called as risk-averse MAB. The investigations on risk-averse MAB aim at maximizing cumulative payoffs during sequential decisions with consideration of variance of payoffs (Even-Dar et al., 2006; Galichet et al., 2013; Sani et al., 2012; Vakili and Zhao, 2016). The first work on risk-averse MAB was by Even-Dar et al. (2006), where two potential definitions on risk of payoffs for decision-making problems were developed. One is Sharpe ratio (Sharpe, 1966) and the other is mean-variance (Markowitz, 1952).

After 2006, the metric of mean-variance has become popular for risk control in MAB. In Vakili and Zhao (2016), risk-averse MAB is directly assumed to have sub-Gaussian noises for high-order statistics, which should be infeasible.

Via the given information in pure exploration of MAB, we can distinguish between two settings, fixed budget and fixed confidence (Gabillon et al., 2012). For the first setting, an algorithm is required to output the optimal arm after playing a given fixed number of rounds. The theoretical guarantee of this setting focuses on the upper bound of the probability of error for selecting a sub-optimal arm (Audibert and Bubeck, 2010). The other setting is to fix a level of confidence to output the optimal arm, and its theoretical guarantee is to minimize the number of rounds for playing arms (Jamieson and Nowak, 2014). Recently, it has been pointed out that these two settings could be equivalent in the sense of sample complexity Chen et al. (2014), and also can be unified into a model Gabillon et al. (2012). Without loss of generality, in this chapter, we focus on the setting of a fixed budget.

3.3 Assumptions and Problem Definition

In this section, we give the assumptions to solve the problem of PEMV. Then, we formally present the problem definition.

3.3.1 Assumptions

We list the assumptions as follows.

Assumption 3.1. *Given a K-arm decision set, if an algorithm $\mathcal{A}$ chooses the arm $a_t \in [K]$, then a stochastic payoff is generated as*

$$r_t(a_t) = \mu(a_t) + \zeta_t, \tag{3.6}$$

where ζ_t is a random noise. Without loss of generality, let a filtration be $\mathcal{F}_t \triangleq \{a_i\}_{i=1}^{t} \cup \{\zeta_i\}_{i=1}^{t-1}$. Then, we assume

$$\mathbb{E}[\exp(\lambda\zeta_t)|\mathcal{F}_t] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \tag{3.7}$$

where $R > 0$. Clearly, we have the result of $\mathbb{E}[\zeta_t|\mathcal{F}_t] = 0$.

Assumption 3.2. *We assume that, for any arm $a \in [K]$, the true variance of arm a is time-invariant and is denoted as $\sigma^2(a) > 0$. Besides, we also assume that the variances among K arms are not all the same. Otherwise, the PEMV problem is trivial because it is equivalent to pure exploration of mean. Based on Assumption 3.1, we have $\mathbb{E}[r_t(a_t)|\mathcal{F}_t] = \mu(a_t)$, and $\sigma^2(a_t) = \mathbb{E}[\zeta_t^2|\mathcal{F}_t] \leq R^2$ for the randomness of the noise ζ_t comes from payoffs of the chosen arm a_t.*

Assumption 3.3. *We assume that the optimal arm in terms of true mean-variance is unique among the given K arms, and thus the optimal arm can be denoted by Opt.*

Remark 3.1. *We would like to briefly address the feasibility of Assumptions 3.1-3.3. Different from previous studies in pure exploration of bandits (Shahrampour et al., 2017), we do not assume independent payoffs for sequential decisions in Assumption 3.1, which is practical. Besides, sub-Gaussian noise ζ_t is general because it encompasses all distributions that are supported on $[0, R]$ as well as many unbounded distributions. Finally, Assumptions 3.2 and 3.3 are reasonable in pure exploration of bandits.*

3.3.2 Problem Definition

We focus on pure exploration with mean-variance in the setting of a fixed budget, which is an essential scenario in MAB (Chen et al., 2014;

Gabillon et al., 2012). Given an algorithm $\mathcal{A}$, the goal of PEMV is to identify the optimal arm Opt with the smallest mean-variance shown in Eq. (3.3). Specifically, with a fixed budget of T, we design bandit algorithms to minimize the probability of error, which is shown as

$$\min \mathbb{P}[a_T \neq \mathsf{Opt}]. \tag{3.8}$$

It is difficult to directly solve the problem of Eq. (3.8). A potential solution is to find its upper bound, which has been a popular alternative in Audibert and Bubeck (2010); Chen et al. (2014). Compared with pure exploration in traditional MAB, the selection sequence of arms with mean-variance will encounter the second-order statistics.

From Eq. (3.3), we notice that there exists a risk tolerance factor in the problem of PEMV. The factor of κ in Eq. (3.3) shows the trade-off between the mean and the variance of stochastic payoffs. It is also worth to mention that, for different κ, the optimal arm could be different in the given decision set. Thus, our focus in this chapter is to find the optimal arm among the decision set with a given value κ, instead of finding optimal κ for mean and variance.

By further investigating the risk tolerance factor of κ, we find that the problem of PEMV is a generalization of the traditional problem of pure exploration for mean. Specifically, we can set different values for specific needs of risk tolerance. On one hand, when we set κ sufficiently large (or even $\kappa \to +\infty$), the dominated term in Eq. (3.3) should be mean because the variances of arms are bounded by R^2 via Assumption 3.1. Then, the problem of PEMV becomes the standard problem of identifying the optimal arm with the highest mean. On the other hand, when we set the parameter as $\kappa = 0$, PEMV becomes minimizing the variance of payoffs.

Algorithm 3.1 PEMV.CB

input: T, K, R, $\mathbf{H}_1$, $\mathbf{H}_3$, κ
$\delta = \min\left(\frac{25(T-2K)}{576(96R^2+\kappa^2)R^2\mathbf{H}_1}, \frac{5(T-2K)}{96R^2\mathbf{H}_3}\right)$
play each arm twice and observe payoffs
for $t = 1, 2, \cdots, T$ **do**
  for $a \in [K]$ **do**
    $\hat{\omega}_t(a) = \hat{\sigma}_t^2(a) - \kappa\hat{\mu}_t(a)$
    $\mathrm{CB}_t(a) = \sqrt{\frac{128R^4(s_t(a)+1)\delta}{(s_t(a)-1)^2} + \frac{4\kappa^2R^2\delta}{s_t(a)}} + \frac{8R^2\delta}{(s_t(a)-1)}$
    $p_t(a) = \hat{\omega}_t(a) - \mathrm{CB}_t(a)$
  end for
  $a_t = \arg\min_{a\in[K]}\ p_t(a)$ ▷ *break ties arbitrarily*
  observe a payoff $y_t(a_t)$ and save information
end for
return: $a_T = \arg\min_{a\in[K]}\ \hat{\omega}_t(a)$

3.4 Two Bandit Algorithms and Analyses

We present two bandit algorithms, which are named as PEMV.CB and PEMV.HALVING. Specifically, we adopt Confidence Bound (CB) technique to develop the algorithm of PEMV.CB, and adopt sequential halving technique to develop PEMV.HALVING.

For any $a \in [K]$, we design mean-variance estimation as

$$\hat{\omega}_t(a) = \hat{\sigma}_t^2(a) - \kappa\hat{\mu}_t(a), \tag{3.9}$$

where $\hat{\mu}_t(a)$ is the estimation of the true expected payoff of $\mu(a)$ at time t. We can calculate $\hat{\mu}_t(a)$ as

$$\hat{\mu}_t(a) = \sum_{i\in\Phi_t(a)} \frac{y_i(a)}{s_t(a)}, \tag{3.10}$$

where $\Phi_t(a)$ is a set to store historical time instants of selecting arm a over $1, \cdots, t-1$, and $s_t(a)$ is the size of the set $\Phi_t(a)$. We can denote

by $\phi_t^i(a)$ the i-th largest element in $\Phi_t(a)$. Note that the values of elements in $\Phi_t(a)$ monotonically increase. Besides, we calculate $\hat{\sigma}_t^2(a)$ as

$$\hat{\sigma}_t^2(a) = \frac{1}{s_t(a)-1} \sum_{i \in \Phi_t(a)} (y_i(a) - \hat{\mu}_t(a))^2. \tag{3.11}$$

3.4.1 Description of PEMV.CB and Results

The confidence bound for the mean-variance estimation is different from that in traditional MAB. With parameters in Eq. (3.31), we set $\mathrm{CB}_t(a) = \sqrt{2v_t(a)\delta} + c_t(a)\delta$, implying

$$\mathrm{CB}_t(a) = \sqrt{\frac{128R^4(s_t(a)+1)\delta}{(s_t(a)-1)^2} + \frac{4\kappa^2 R^2 \delta}{s_t(a)}} + \frac{8R^2\delta}{(s_t(a)-1)},$$

where $\delta > 0$ is a parameter.

We present PEMV.CB in Algorithm 3.1, which is inspired by the CB technique in MAB (Auer and Ortner, 2010; Auer et al., 2002a). The key idea of CB technique is to add or subtract a CB term to empirical estimations. Note that we adopt subtraction because of the minimization of mean-variance. Besides, the input includes the hardness, which is to calculate a parameter of δ. Finally, we are ready to have the time complexity of PEMV.CB as $O(TK)$. The following theorem shows the theoretical guarantee of Algorithm 3.1.

Theorem 3.1. *For pure exploration of mean-variance with K-arm MAB, suppose Assumptions 3.1-3.3 are satisfied. If Algorithm 3.1 is run with a fixed budget of T, we have the upper bound of the probability of error for PEMV.CB as*

$$\mathbb{P}[a_T \neq \mathit{Opt}] \leq 2TK \exp\Big(-\frac{\delta}{5}\Big), \tag{3.12}$$

where $\delta \in \left(0, \min\left(\frac{25(T-2K)}{576(96R^2+\kappa^2)R^2\mathbf{H}_1}, \frac{5(T-2K)}{96R^2\mathbf{H}_3}\right)\right]$.

3.4.2 Description of PEMV.HALVING and Results

Since Algorithm 3.1 requires the information of hardness, we further develop PEMV.HALVING to remove this weakness. The halving technique is popular in machine learning (Kalyanakrishnan et al., 2012; Karnin et al., 2013). We show PEMV.HALVING in Algorithm 3.2, with the key idea of deleting an arm via mean-variance estimation. We can calculate the time complexity of PEMV.HALVING as $O(T + K)$. The following theorem shows the theoretical guarantee of Algorithm 3.2.

Theorem 3.2. *For pure exploration of mean-variance with K-arm MAB, suppose Assumptions 3.1-3.3 are satisfied. If Algorithm 3.2 is run with a fixed budget of T, we have the upper bound of the probability of error for PEMV.HALVING as*

$$\mathbb{P}[a_T \neq \mathit{Opt}] \leq 2K \exp\Big(- \frac{T}{\log_2(K)\mathbf{H}}\Big), \tag{3.13}$$

where $\mathbf{H} = 12(96R^2 + \kappa^2)R^2 \min(\mathbf{H}_4, 3\mathbf{H}_2)$.

3.5 Proofs of Theorems

In this section, we first rigorously prove that the error at t resulting from the mean-variance estimation is sub-gamma. Then, with a fixed budget, we develop upper bounds of probability of error for the proposed bandit algorithms.

Theorem 3.3. *For pure exploration of mean-variance with K-arm MAB, suppose Assumptions 3.1-3.3 are satisfied. We define a random variable as $\rho_t(a) \triangleq \hat{\omega}_t(a) - \omega(a)$ for any $a \in [K]$. Then, we have $\rho_t(a)$ is sub-gamma on the right tail, implying*

$$\mathbb{E}[\exp(\lambda \rho_t(a))] \leq \exp\Big(\frac{\lambda^2 v}{2(1 - c\lambda)}\Big),$$

Algorithm 3.2 PEMV.HALVING

input T, K, κ
construct a decision-arm set $\mathcal{D}_1 = [K]$, $t = 0$
for $k = 1, \cdots, \lceil \log_2(K) \rceil$ **do**
  $T_k = \lfloor \frac{T}{|\mathcal{D}_k| \lceil \log_2(K) \rceil} \rfloor$
  for $a \in \mathcal{D}_k$ **do**
    for $j = 1, \cdots, T_k$ **do**
      $t = t + 1$
      select a and observe $y_j(a)$
    end for
  end for
  if $|\mathcal{D}_k| > 1$ **then**
    for $j = 1, \cdots, \lfloor \frac{|\mathcal{D}_k|}{2} \rfloor$ **do**
      select an arm $a_j = \arg\max_{a \in \mathcal{D}_k} \hat{\omega}_k(a)$
      $\mathcal{D}_k = \mathcal{D}_k \backslash a_j$ ▷ *delete an arm*
    end for
  end if
  $\mathcal{D}_{k+1} = \mathcal{D}_k$
end for
return $a_T = \mathcal{D}_{\lceil \log_2(K) \rceil + 1}$

where $\lambda \in (0, \frac{1}{c})$, $c = 8R^2$, $v = (192R^2 + \kappa^2)R^2$ *for any* $a \in [K]$ *and* $t \in [T]$.

Before we present the proof of Theorem 3.3, we give the following two lemmas. Lemma 3.1 is on the property of square of a sub-Gaussian random variable. Lemma 3.2 is on the moment generating function of the sum of random variables.

Lemma 3.1. *If ζ is $\bar{R}$-sub-Gaussian, we have*

$$\mathbb{E}\Big[\exp(\gamma(\zeta^2 - \mathbb{E}[\zeta^2]))\Big] \leq \exp\Big(\frac{8\gamma^2\bar{R}^4}{1-2\gamma\bar{R}^2}\Big), \tag{3.14}$$

where $\gamma \in (-\frac{1}{2\bar{R}^2}, \frac{1}{2\bar{R}^2})$.

Proof. Since ζ is $\bar{R}$-sub-Gaussian, we have

$$\mathbb{E}[\exp(\lambda\zeta)] \leq \exp\Big(\frac{\lambda^2\bar{R}^2}{2}\Big),\ \forall\lambda \in \mathbb{R}. \tag{3.15}$$

Based on Rivasplata (2012), we have

$$\mathbb{E}[|\zeta|^r] \leq r2^{\frac{r}{2}}\bar{R}^r\Gamma\Big(\frac{r}{2}\Big),\ \forall\ r \geq 0, \tag{3.16}$$

where $\Gamma(r)$ is the Gamma function. Based on the result of Honorio and Jaakkola (2014), we have

$$\mathbb{E}[\exp(\gamma(\zeta - \mathbb{E}[\zeta^2]))] \leq 1 + \frac{8\gamma^2\bar{R}^4}{1-2\gamma\bar{R}^2}, \tag{3.17}$$

where we set $\gamma \in (-\frac{1}{2\bar{R}^2}, \frac{1}{2\bar{R}^2})$. Due to the fact of $1 + x \leq \exp(x)$ for $x \in \mathbb{R}$, we have

$$\mathbb{E}[\exp(\gamma(\zeta - \mathbb{E}[\zeta^2]))] \leq \exp\Big(\frac{8\gamma^2\bar{R}^4}{1-2\gamma\bar{R}^2}\Big), \tag{3.18}$$

which completes the proof. □

Lemma 3.2. *Given two random variables of ζ_1 and ζ_2, for any $\lambda \in \mathbb{R}$ and $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\mathbb{E}[\exp(\lambda(\zeta_1 + \zeta_2))] \leq \Big(\mathbb{E}[\exp(p\lambda\zeta_1)]\Big)^{\frac{1}{p}}\Big(\mathbb{E}[\exp(q\lambda\zeta_2)]\Big)^{\frac{1}{q}}.$$

In particular, by setting $p = q = 2$, we have the result as $\mathbb{E}[\exp(\lambda(\zeta_1 + \zeta_2))] \leq \sqrt{\Big(\mathbb{E}[\exp(2\lambda\zeta_1)]\Big)}\sqrt{\Big(\mathbb{E}[\exp(2\lambda\zeta_2)]\Big)}$.

Proof. The proof can be easily generalized from the result of Hölder's inequality by Finner et al. (1992). □

Proof of Theorem 3.3. The key is the moment generating function of $\rho_t(a) = \hat{\omega}_t(a) - \omega(a)$, which is the estimation error of mean-variance . There are three steps for the proof.

Step 1. moment generating function of mean

For any arm $a \in [K]$, we define

$$\alpha_t(a) \triangleq \hat{\mu}_t(a) - \mu(a), \tag{3.19}$$

where $\alpha_t(a) \in \mathbb{R}$. With the estimation in Eq. (3.10), we have $\alpha_t(a) = \sum_{i\in\Phi_t(a)} \zeta_i / s_t(a)$. Via Assumption 3.1, we obtain

$$\begin{aligned}
\mathbb{E}[\exp(\lambda\alpha_t(a))] &= \mathbb{E}\Big[\exp\Big(\frac{\lambda\sum_{i\in\Phi_t(a)}\zeta_i}{s_t(a)}\Big)\Big] \\
&= \mathbb{E}\Big[\exp\Big(\frac{\lambda\sum_{i\in\Phi_t^1(a)}\zeta_i}{s_t(a)}\Big)\mathbb{E}\Big[\exp\big(\frac{\lambda\zeta_{\phi_t^1(a)}}{s_t(a)}\big)|\mathcal{F}_{\phi_t^1(a)}\Big]\Big] \\
&\leq \exp\Big(\frac{\lambda^2R^2}{2s_t^2(a)}\Big)\mathbb{E}\Big[\exp\Big(\frac{\lambda\sum_{i\in\Phi_t^1(a)}\zeta_i}{s_t(a)}\Big)\Big],
\end{aligned} \tag{3.20}$$

where $\Phi_t^1(a) = \Phi_t(a)\backslash\phi_t^1(a)$ with $\phi_t^1(a)$ the largest element in $\Phi_t^1(a)$. Continuing inductively, we conclude that

$$\mathbb{E}[\exp(\lambda\alpha_t(a))] \leq \exp\Big(\frac{\lambda^2R^2}{2s_t(a)}\Big), \tag{3.21}$$

which implies that $\alpha_t(a)$ is sub-Gaussian. Thus, we have

$$\begin{aligned}
\mathbb{E}[\alpha_t^2(a)] &= \mathbb{E}[\hat{\mu}_t^2(a) - 2\hat{\mu}_t(a)\mu(a) + \mu^2(a)] \\
&= \mathbb{E}\Big[\frac{1}{s_t^2(a)}\Big(\sum_{i\in\Phi_t(a)}(\mu(a)+\zeta_i)\Big)^2\Big] - \mu^2(a) \\
&= \mathbb{E}\Big[\frac{\sum_{i\in\Phi_t(a)}\mathbb{E}[\zeta_i^2|\mathcal{F}_i]}{s_t^2(a)}\Big] = \frac{\sigma^2(a)}{s_t(a)},
\end{aligned} \tag{3.22}$$

where we adopt the fact of $\mathbb{E}[\zeta_i|\mathcal{F}_i] = 0$ for any $i \in \Phi_t(a)$. We know $\mathbb{E}[\alpha_t^2(a)]$ decreases with increasing $s_t(a)$.

Step 2. moment generating function of variance

For any arm $a \in [K]$, we define

$$\beta_t(a) \triangleq \hat{\sigma}_t^2(a) - \sigma^2(a), \tag{3.23}$$

where $\beta_t(a) \in \mathbb{R}$. We also define

$$\bar{\sigma}_t^2(a) \triangleq \frac{1}{s_t(a)-1} \sum_{i \in \Phi_t(a)} \Big(y_i(a) - \mu(a) \Big)^2, \tag{3.24}$$

which implies $\bar{\sigma}_t^2(a) = \frac{\sum_{i \in \Phi_t(a)} \zeta_i^2}{s_t(a)-1}$ via Assumption 3.1. Besides, we are ready to have the result of

$$\begin{aligned}
\bar{\sigma}_t^2(a) &= \frac{\sum_{i \in \Phi_t(a)} \Big(y_i(a) - \hat{\mu}_t(a) + \hat{\mu}_t(a) - \mu(a) \Big)^2}{s_t(a)-1} \\
&= \frac{\sum_{i \in \Phi_t(a)} \Big((y_i(a) - \hat{\mu}_t(a))^2 + (\hat{\mu}_t(a) - \mu(a))^2 \Big)}{s_t(a)-1} \\
&= \frac{\sum_{i \in \Phi_t(a)} \Big((y_i(a) - \hat{\mu}_t(a))^2 \Big)}{s_t(a)-1} + \frac{s_t(a)}{s_t(a)-1} \alpha_t^2(a), \\
&= \hat{\sigma}_t^2(a) + \frac{s_t(a)}{s_t(a)-1} \alpha_t^2(a),
\end{aligned} \tag{3.25}$$

where we adopt the fact $\sum_{i \in \Phi_t(a)} (y_i(a) - \hat{\mu}_t(a)) = 0$ in the second equality, and adopt the definition of $\alpha_t(a)$ in the third equality. Now we have

$$\begin{aligned}
\beta_t(a) &= \bar{\sigma}_t^2(a) - \frac{s_t(a)}{s_t(a)-1} \alpha_t^2(a) - \sigma^2(a) \\
&= \frac{\sum_{i \in \Phi_t(a)} \zeta_i^2}{s_t(a)-1} - \frac{s_t(a)}{s_t(a)-1} \alpha_t^2(a) - \sigma^2(a) \\
&= \frac{\sum_{i \in \Phi_t(a)} (\zeta_i^2 - \sigma^2(a))}{s_t(a)-1} - \frac{s_t(a)}{s_t(a)-1} \Big(\alpha_t^2(a) - \frac{\sigma^2(a)}{s_t(a)} \Big).
\end{aligned}$$

Now we consider separately the two terms on the right hand side of the above equation. Specifically, we have

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\frac{\lambda \sum_{i \in \Phi_t(a)} (\zeta_i^2 - \sigma^2(a))}{s_t(a)-1} \Big) \Big] \\
&= \mathbb{E}\Big[\exp\Big(\lambda \Theta_t^1(a) \Big) \mathbb{E}\Big[\exp\Big(\frac{\lambda(\zeta_{\phi_t^1(a)}^2 - \sigma^2(a))}{s_t(a)-1} \Big) | \mathcal{F}_{\phi_t^1(a)} \Big] \Big] \\
&\leq \exp\Big(\frac{8\gamma_1^2 R^4}{1 - 2\gamma_1 R^2} \Big) \mathbb{E}\Big[\exp\Big(\lambda \Delta_t^1(a) \Big) \Big],
\end{aligned} \tag{3.26}$$

where $\Theta_t^1(a) \triangleq \frac{\sum_{i\in\Phi_t^1(a)}\left(\zeta_i^2-\sigma^2(a)\right)}{s_t(a)-1}$, $\gamma_1 = \lambda/(s_t(a)-1)$ with $\gamma_1 \in (-\frac{1}{2R^2}, \frac{1}{2R^2})$, and the inequality is due to Lemma 3.1 by using the conditional filtration. Similar to the technique in Eq. (3.20), we conclude that

$$\mathbb{E}\Big[\exp\Big(\frac{\lambda\sum_{i\in\Phi_t(a)}(\zeta_i^2-\sigma^2(a))}{s_t(a)-1}\Big)\Big] \le \exp\Big(\frac{8\gamma_1^2R^4s_t(a)}{1-2\gamma_1R^2}\Big).$$

Similarly, based on Eqs. (3.21) and (3.22), we have

$$\mathbb{E}\Big[\exp\Big(\frac{-\lambda s_t(a)}{s_t(a)-1}\Big(\alpha_t^2(a)-\frac{\sigma^2(a)}{s_t(a)}\Big)\Big)\Big] \le \exp\Big(\frac{8\gamma_1^2R^4}{1+2\gamma_1R^2}\Big).$$

Now, with the support of Lemma 3.2, we can calculate

$$\begin{aligned}
\mathbb{E}[\exp(\lambda\beta_t(a))] &= \mathbb{E}\Big[\exp\Big(\bar{\sigma}_t^2(a)-\frac{s_t(a)}{s_t(a)-1}\alpha_t^2(a)-\sigma^2(a)\Big)\Big] \\
&\le \sqrt{\mathbb{E}\Big[\exp\Big(\frac{2\lambda\sum_{i\in\Phi_t(a)}(\zeta_i^2-\sigma^2(a))}{s_t(a)-1}\Big)\Big]}\times \\
&\sqrt{\mathbb{E}\Big[\exp\Big(-\frac{2\lambda s_t(a)}{s_t(a)-1}\Big(\alpha_t^2(a)-\frac{\sigma^2(a)}{s_t(a)}\Big)\Big)\Big]} \\
&\le \exp\Big(\frac{4\gamma_2^2R^4s_t(a)}{1-2\gamma_2R^2}+\frac{4\gamma_2^2R^4)}{1+2\gamma_2R^2}\Big) \\
&= \exp\Big(\frac{4\gamma_2^2R^4\Big(s_t(a)+1-\frac{4}{1/(\gamma_2R^2)+2)}\Big)}{1-2\gamma_2R^2}\Big), \qquad (3.27)
\end{aligned}$$

where $\gamma_2 = 2\lambda/(s_t(a)-1)$ with $\gamma_2 \in (0, \frac{1}{2R^2})$. Clearly, we have $\frac{4}{1/(\gamma_2R^2)+2} \in (0,1)$. Note that we consider $\gamma_2 > 0$ here because $\gamma_2 \in (-\frac{1}{2R^2}, 0)$ will lead to $\frac{4}{1/(\gamma_2R^2)+2} \in (-\infty, 0)$, implying the unbounded moment generating function of $\beta_t(a)$. Thus, we have

$$\mathbb{E}[\exp(\lambda\beta_t(a))] \le \exp\Big(\frac{4\gamma_2^2R^4(s_t(a)+1)}{1-2\gamma_2R^2}\Big). \qquad (3.28)$$

We obtain

$$\mathbb{E}[\exp(\lambda\beta_t(a))] \le \exp\Big(\frac{\lambda^2b_1}{2(1-\lambda b_2)}\Big), \qquad (3.29)$$

where $\lambda \in (0, \frac{1}{b_2})$, $b_1 = 32R^4(s_t(a)+1)/(s_t(a)-1)^2$ and $b_2 = 4R^2/(s_t(a)-1)$. This implies that $\beta_t(a)$ is a sub-gamma random variable on the right tail.

Step 3. moment generating function of $\rho_t(a)$

Since $\rho_t(a) = \hat{\omega}_t(a) - \omega(a) = \beta_t(a) - \kappa\alpha_t(a)$, we have

$$\begin{aligned}\mathbb{E}[\exp(\lambda\rho_t(a))] &\leq \sqrt{\mathbb{E}[\exp(2\lambda\beta_t(a))]}\sqrt{\mathbb{E}[\exp(-2\kappa\lambda\alpha_t(a))]} \\ &\leq \sqrt{\exp\Big(\frac{4\lambda^2 b_1}{2(1-2\lambda b_2)}\Big)}\sqrt{\exp\Big(\frac{4\kappa^2\lambda^2R^2}{2s_t(a)}\Big)} \\ &= \exp\Big(\frac{2\lambda^2 b_1}{2(1-2\lambda b_2)} + \frac{2\kappa^2\lambda^2R^2}{2s_t(a)}\Big) \\ &\leq \exp\Big(\frac{\lambda^2(2b_1 s_t(a) + 2\kappa^2R^2)}{2s_t(a)(1-2\lambda b_2)}\Big), \end{aligned} \tag{3.30}$$

where $\lambda \in (0, \frac{1}{2b_2})$, the first inequality is due to Lemma 3.2, and the last inequality is due to the fact of $1 - 2\lambda b_2 < 1$. Then, we have

$$\mathbb{E}[\exp(\lambda\rho_t(a))] \leq \exp\Big(\frac{\lambda^2 v_t(a)}{2(1-c_t(a)\lambda)}\Big), \tag{3.31}$$

where $\lambda \in \Big(0, \frac{1}{c_t(y)}\Big)$, $c_t(a) = 8R^2/(s_t(a)-1)$ and $v_t(a) = 64R^4(s_t(a)+1)/(s_t(a)-1)^2 + 2\kappa^2R^2/s_t(a)$. Without loss of generality, we can set $s_t(a) \geq 2$. Then we have $c_t(a) \leq 8R^2$ and $v_t(a) \leq (192R^2 + \kappa^2)R^2$ for any $a \in [K]$. Finally, we have

$$\mathbb{E}[\exp(\lambda\rho_t(a))] \leq \exp\Big(\frac{\lambda^2 v}{2(1-c\lambda)}\Big), \tag{3.32}$$

where $\lambda \in (0, \frac{1}{c})$, $c = 8R^2$, $v = (192R^2 + \kappa^2)R^2$ for any $a \in [K]$ and $t \in [T]$. This means that $\rho_t(a)$ is sub-gamma on the right tail, which completes the proof. □

3.5.1 Proof of Theorem 3.1

Proof. From Theorem 3.3, we know $\{\rho_t(a)\}$ are sub-gamma on the right tail for all $a \in [K]$ and $t \in [T]$. Based on Boucheron et al. (2012), we know a sub-gamma random variable $\rho_t(a)$ on the right tail satisfies a Bernstein inequality as

$$\mathbb{P}[\rho_t(a) \geq \sqrt{2v_t(a)\delta} + c_t(a)\delta] \leq \exp(-\delta), \tag{3.33}$$

where $\delta > 0$. Note that here we adopt the parameters in Eq. (3.31). Besides, we have similar results of Eq. (3.33) for $-\rho_t(a)$. Thus, we have the concentration inequality as

$$\mathbb{P}[|\rho_t(a)| \leq \sqrt{2v_t(a)\delta} + c_t(a)\delta] \geq 1 - 2\exp(-\delta). \tag{3.34}$$

Inspired by the above equation, we consider the event as

$$\mathcal{E} = \left\{a \in [K], |\hat{\omega}_t(a) - \omega(a)| \leq \frac{1}{5}\sqrt{2v_t(a)\delta} + \frac{1}{5}c_t(a)\delta\right\},$$

where $t \in [T]$, $c_t(a) = 8R^2/(s_t(a) - 1)$ and $v_t(a) = 64R^4(s_t(a) + 1)/(s_t(a) - 1)^2 + 2\kappa^2R^2/s_t(a)$. The following details are inspired by the reuslt in Audibert and Bubeck (2010). In the proof, we show that the event $\mathcal{E}$ implies that $a_T = \mathsf{Opt}$. Thus, the probability of error in PEMV.CB is equivalent to the upper bound of $1 - \mathbb{P}[\mathcal{E}]$. Since $t \in [T]$ and $a \in [K]$, by adopting a union bound of probability in $\mathcal{E}$, we need to find δ such that

$$\mathbb{P}[\mathcal{E}] \geq 1 - 2TK\exp(-\frac{\delta}{5}), \tag{3.35}$$

where we adopt the result in Eq. (3.34) and the fact of $\frac{1}{5}\sqrt{2v_t(a)\delta} > \frac{1}{25}\sqrt{2v_t(a)\delta}$. It is enough to prove that

$$\frac{1}{5}\sqrt{2v_t(a)\delta} + \frac{1}{5}c_t(a)\delta \leq \frac{\Delta_a}{2}, \tag{3.36}$$

where $a \in [K]$. Then we consider two cases: 1) $\sqrt{2v_t(a)\delta} \geq c_t(a)\delta$, and 2) $\sqrt{2v_t(a)\delta} < c_t(a)\delta$. Note that, in Algorithm 3.1, we design the confidence bound as $\mathrm{CB}_t(a) = \sqrt{2v_t(a)\delta} + c_t(a)\delta$.

Case 1. Since $\sqrt{2v_t(a)\delta} \geq c_t(a)\delta$ and for $\mathcal{E}$, we should prove

$$\frac{2}{5}\sqrt{2v_t(a)\delta} \leq \frac{\Delta_a}{2}. \tag{3.37}$$

Besides, we have

$$\begin{aligned}\sqrt{v_t(a)} &= \sqrt{\frac{64R^4(s_t(a)+1)}{(s_t(a)-1)^2} + \frac{2\kappa^2R^2}{s_t(a)}} \\ &= \sqrt{\frac{64R^4}{s_t(a)-1} + \frac{128R^4}{(s_t(a)-1)^2} + \frac{2\kappa^2R^2}{s_t(a)}} \\ &\leq \sqrt{\frac{64R^4}{s_t(a)-1} + \frac{128R^4}{s_t(a)-1} + \frac{2\kappa^2R^2}{s_t(a)-1}} \\ &= \sqrt{\frac{2(96R^2+\kappa^2)R^2}{s_t(a)-1}}. \end{aligned} \tag{3.38}$$

By setting $t = T$, it is enough to prove

$$\frac{2}{5}\sqrt{2v_T(a)\delta} \leq \frac{4}{5}\sqrt{\frac{(96R^2+\kappa^2)R^2\delta}{s_T(a)-1}} \leq \frac{\Delta_a}{2}. \tag{3.39}$$

Equivalently, we should find δ such that

$$s_T(a) \geq \frac{64(96R^2+\kappa^2)R^2\delta}{25\Delta_a^2} + 1, \ \forall\ a \in [K]. \tag{3.40}$$

Now we need to consider the upper bound of $s_t(a)$ for $a \in [K]$ with $a \neq \mathsf{Opt}$. We prove by induction that

$$s_t(a) \leq \frac{576(96R^2+\kappa^2)R^2\delta}{25\Delta_a^2} + 2, \tag{3.41}$$

where $a \in [K]$ with $a \neq \mathsf{Opt}$, and $t = [T]$.

For $t = 1$, Eq. (3.41) holds obviously. We assume Eq. (3.41) holds at $t-1$, then we prove Eq. (3.41) at t. If $a_t \neq a$ for $\forall\ a \in [K]$ with $a \neq \mathsf{Opt}$, the chosen arm is Opt. Thus, $s_t(a) = s_{t-1}(a)$ holds. If $a_t = a$, we know that

$$\hat{\omega}_{t-1}(a) - 2\sqrt{2v_{t-1}(a)\delta} \leq \hat{\omega}_{t-1}(\mathsf{Opt}) - 2\sqrt{2v_{t-1}(\mathsf{Opt})\delta}. \tag{3.42}$$

Because we consider $\mathcal{E}$, we have $\hat{\omega}_{t-1}(\mathsf{Opt}) - 2\sqrt{2v_{t-1}(\mathsf{Opt})\delta} \leq \omega(\mathsf{Opt})$. Besides, we have $\omega(a) - \frac{12}{5}\sqrt{2v_{t-1}(a)\delta} \leq \hat{\omega}_{t-1}(a) - 2\sqrt{2v_{t-1}(a)\delta}$. Then, we have

$$\omega(a) - \omega(\mathsf{Opt}) \leq \frac{12}{5}\sqrt{2v_{t-1}(a)\delta}, \tag{3.43}$$

implying that $s_t(a) \leq \frac{576(96R^2+\kappa^2)R^2\delta}{25\Delta_a^2} + 2$ with Eq. (3.38), where we adopt the fact of $s_t(a) = s_{t-1}(a) + 1$.

Now we prove by induction that

$$s_t(a) \geq \min\Big(\frac{144(96R^2+\kappa^2)R^2\delta}{25\Delta_a^2}, \frac{9}{16}(s_t(\mathsf{Opt}) - 2)\Big) + 1, \tag{3.44}$$

where $a \in [K]$ and $a \neq \mathsf{Opt}$. For $t = 1$, it is obvious true. We assume Eq. (3.44) holds at $t-1$, then we prove it at t. If $a_t = a$ for $\forall a \in [K]$ with $a \neq \mathsf{Opt}$, we know that it clearly is true since $s_t(a) = s_{t-1}(a) + 1$. If $a_t = \mathsf{Opt}$, then we have

$$\omega(\mathsf{Opt}) - \frac{8}{5}\sqrt{2v_{t-1}(\mathsf{Opt})\delta} \leq \omega(a) - \frac{12}{5}\sqrt{2v_{t-1}(a)\delta}. \tag{3.45}$$

Thus, with Eq. (3.38), we have

$$s_{t-1}(a) - 1 \geq \frac{576}{25}\frac{(96R^2+\kappa^2)R^2\delta}{\Big(\Delta_a + \frac{16}{5}\sqrt{\frac{(96R^2+\kappa^2)R^2\delta}{s_{t-1}(\mathsf{Opt})-1}}\Big)^2}, \tag{3.46}$$

which implies Eq. (3.44). In order to guarantee the result in Eq. (3.40), we only need to prove

$$\frac{9}{16}(s_T(\mathsf{Opt}) - 2) \geq \frac{64(96R^2+\kappa^2)R^2\delta}{25\Delta_*^2}, \tag{3.47}$$

where $\Delta_* = \min_{a\neq\mathsf{Opt}, a\in[K]} \Delta_a$. Based on Eq. (3.41), we have

$$\begin{aligned} s_T(\mathsf{Opt}) - 2 &= T - 2 - \sum_{a\neq\mathsf{Opt}} s_T(a) \\ &\geq T - 2 - 2(K-1) - \frac{576(96R^2+\kappa^2)R^2\delta}{25}\sum_{a\neq\mathsf{Opt}}\frac{1}{\Delta_a^2} \\ &\geq \frac{576(96R^2+\kappa^2)R^2\delta}{25\Delta_*^2} > \frac{1024(96R^2+\kappa^2)R^2\delta}{225\Delta_*^2}, \end{aligned} \tag{3.48}$$

which implies that

$$0 < \delta \leq \frac{25(T-2K)}{576(96R^2+\kappa^2)R^2\mathbf{H}_1}. \tag{3.49}$$

Case 2. Since $\sqrt{2v_t(a)\delta} < c_t(a)\delta$ and for $\mathcal{E}$, we should prove

$$\frac{2}{5}c_t(a)\delta \leq \frac{\Delta_a}{2}. \tag{3.50}$$

Via analyses similar to Case 1, we need to find δ satisfying

$$s_T(a) \geq \frac{32R^2\delta}{5\Delta_a} + 1. \tag{3.51}$$

We can have

$$s_t(a) \leq \frac{96R^2\delta}{5\Delta_a} + 2, \ \forall\ a \in [K] \text{ with } a \neq \mathsf{Opt}, \text{ and } \forall t \in [T]. \tag{3.52}$$

We also have

$$s_t(a) \geq \min\Big(\frac{48R^2\delta}{5\Delta_a}, \frac{3}{2}(s_t(\mathsf{Opt}) - 2)\Big) + 1, \tag{3.53}$$

where $a \in [K]$ with $a \neq \mathsf{Opt}$, and $t \in [T]$. Thus, we need to have

$$\frac{3}{2}(s_T(\mathsf{Opt}) - 2) \geq \frac{32R^2\delta}{5\Delta_*}. \tag{3.54}$$

Then, we have

$$\begin{aligned} s_T(\mathsf{Opt}) - 2 &= T - 2 - \sum_{a \neq \mathsf{Opt}} s_T(a) \\ &\geq T - 2 - 2(K-1) - \frac{96R^2\delta}{5} \sum_{a \neq \mathsf{Opt}} \frac{1}{\Delta_a} \\ &\geq \frac{96R^2\delta}{5\Delta_*} > \frac{64R^2\delta}{15\Delta_*}, \end{aligned} \tag{3.55}$$

which implies that

$$0 < \delta \leq \frac{5(T-2K)}{96R^2\mathbf{H}_3}. \tag{3.56}$$

Because we need to hold $\mathcal{E}$ in both cases for $a \in [K]$, we can set $\delta \in \Big(0, \min\Big(\frac{25(T-2K)}{576(96R^2+\kappa^2)R^2\mathbf{H}_1}, \frac{5(T-2K)}{96R^2\mathbf{H}_3}\Big)\Big)$. Then, we know

$$\mathbb{P}[\mathcal{E}] \geq 1 - 2TK\exp\Big(-\frac{\delta}{5}\Big). \tag{3.57}$$

This implies that

$$\mathbb{P}[a_T \neq \mathsf{Opt}] \leq 2TK\exp\Big(-\frac{\delta}{5}\Big), \tag{3.58}$$

which completes the proof. □

3.5.2 Proof of Theorem 3.2

Proof. Without loss of generality, we assume K is a power of 2. The following analysis can be generalized to any K. From the result in Theorem 3.3, we know $\{\rho_t(a)\}$ are sub-gamma on the right tail for all $a \in [K]$ and $t \in [T]$. Based on Boucheron et al. (2012), we know a sub-gamma random variable $\rho_t(y)$ on the right tail satisfies a Bernstein inequality as

$$\mathbb{P}[\rho_t(a) \geq \sqrt{2v_t(a)\delta} + c_t(a)\delta] \leq \exp(-\delta), \tag{3.59}$$

where $c_t(a) = 8R^2/(s_t(a)-1)$ and $v_t(a) = 64R^4(s_t(a)+1)/(s_t(a)-1)^2 + 2\kappa^2R^2/s_t(a)$. Here we adopt a different technique to combine $v_t(a)$ and $c_t(a)$. We observe that $v_t(a) \geq c_t(a)$ due to $s_t(a) > 0$, then we have $\sqrt{2v_t(a)\delta} + c_t(a)\delta \leq \sqrt{2v_t(a)\delta} + v_t(a)\delta$. By further noticing that $\sqrt{2v_t(a)\delta} + v_t(a)\delta \leq 2\sqrt{v_t(a)\delta} + v_t(a)\delta$, we have

$$\mathbb{P}[\rho_t(a) \geq 2\sqrt{v_t(a)\delta} + v_t(a)\delta] \leq \exp(-\delta). \tag{3.60}$$

Thus, we have

$$\begin{cases} \mathbb{P}[\rho_t(a) \geq 3v_t(a)\delta] \leq \exp(-\delta) & \text{if } v_t(a)\delta \geq 1, \\ \mathbb{P}[\rho_t(a) \geq 3\sqrt{v_t(a)\delta}] \leq \exp(-\delta) & \text{if } 0 < v_t(a)\delta < 1. \end{cases}$$

We can set $\hat{\delta} = 3v_t(a)\delta$ if $v_t(a)\delta \geq 1$, and $\hat{\delta} = 3\sqrt{v_t(a)\delta}$ if $0 < v_t(a)\delta < 1$. Equivalently, we have

$$\begin{cases} \mathbb{P}[\rho_t(a) \geq \hat{\delta}] \leq \exp\left(-\frac{\hat{\delta}}{3v_t(a)}\right) & \text{if } \hat{\delta} \geq 3, \\ \mathbb{P}[\rho_t(a) \geq \hat{\delta}] \leq \exp\left(-\frac{\hat{\delta}^2}{9v_t(a)}\right) & \text{if } 0 < \hat{\delta} < 3. \end{cases} \tag{3.61}$$

Besides, we define

$$\hat{v}_t(a) \triangleq \frac{2(96R^2+\kappa^2)R^2}{s_t(a)-1}. \tag{3.62}$$

It is ready to have $v_t(a) \leq \hat{v}_t(a)$. Thus, we have

$$\begin{cases} \mathbb{P}[\rho_t(a) \geq \hat{\delta}] \leq \exp\Big(- \frac{(s_t(a)-1)\hat{\delta}}{6(96R^2+\kappa^2)R^2} \Big) & \text{if } \hat{\delta} \geq 3, \\ \mathbb{P}[\rho_t(a) \geq \hat{\delta}] \leq \exp\Big(- \frac{(s_t(a)-1)\hat{\delta}^2}{18(96R^2+\kappa^2)R^2} \Big) & \text{if } 0 < \hat{\delta} < 3. \end{cases}$$

In Algorithm 3.2, in each epoch k, we denote the set of deleted arms as $\bar{\mathcal{D}}_k$, i.e., the arm set in Line 14 of PEMV.HALVING. Inspired by Karnin et al. (2013), we can upper bound the probability of error as

$$\begin{aligned} \mathbb{P}[a_T \neq \mathsf{Opt}] &\leq \sum_{k=1}^{\log_2(K)} \sum_{a \in \bar{\mathcal{D}}_k} \mathbb{P}[\hat{\omega}_t(a) \leq \hat{\omega}_t(\mathsf{Opt})] \\ &= \sum_{k=1}^{\log_2(K)} \sum_{a \in \bar{\mathcal{D}}_k} \mathbb{P}[\hat{\omega}_t(\mathsf{Opt}) - \omega(\mathsf{Opt}) - \hat{\omega}_t(a) + \omega(a) \geq \Delta_a]. \end{aligned}$$

We have sub-gamma noise for $\hat{\omega}_t(\mathsf{Opt}) - \omega(\mathsf{Opt}) - \hat{\omega}_t(a) + \omega(a)$ shown as Theorem 3.1 if we consider the noise of $\hat{\omega}_t(\mathsf{Opt}) - \hat{\omega}_t(a) - (\omega(\mathsf{Opt}) - \omega(a))$. This means the tail probability of Eq. (3.61) can be used. For clarity, we can define

$$A \triangleq (96R^2 + \kappa^2)R^2, \tag{3.63}$$

where $A > 0$. Thus,

$$\begin{cases} \mathbb{P}[\rho_t(a) \geq \hat{\delta}] \leq \exp\Big(- \frac{(s_t(a)-1)\hat{\delta}}{6A} \Big) & \text{if } \hat{\delta} \geq 3, \\ \mathbb{P}[\rho_t(a) \geq \hat{\delta}] \leq \exp\Big(- \frac{(s_t(a)-1)\hat{\delta}^2}{18A} \Big) & \text{if } 0 < \hat{\delta} < 3. \end{cases}$$

Without loss of generality, we can set $\Delta_a = \hat{\delta}$. Then, if $\Delta_a \geq 3$, we have

$$\begin{aligned} \mathbb{P}[a_T \neq \mathsf{Opt}] &\leq \sum_{k=1}^{\log_2(K)} \sum_{a \in \bar{\mathcal{D}}_k} \exp\Big(- \frac{(\sum_{i=1}^{k} T_i - 1)\Delta_a}{6A} \Big) \\ &\leq \sum_{k=1}^{\log_2(K)} \sum_{a \in \bar{\mathcal{D}}_k} \exp\Big(- \frac{2^{k-1} T \Delta_a}{6AK \log_2(K)} \Big), \end{aligned}$$

where T_i is the value in Line 4 of Algorithm 3.2 and we adopt the fact of $|\bar{\mathcal{D}}_k| = |\mathcal{D}_k| = \frac{K}{2^{k-1}}$ for $k \geq 2$. Besides, $\mathcal{D}_1 = K$ and $\bar{\mathcal{D}}_1 = 0$. Inspired

by Karnin et al. (2013), we have

$$\begin{aligned}\mathbb{P}[a_T \neq y^*] &\leq \sum_{k=1}^{\log_2(K)} \frac{K}{2^{k-1}} \max_{a\in\bar{D}_1} \exp\Big(-\frac{2^k T\Delta_a}{12AK\log_2(K)}\Big)\\ &\leq \sum_{k=1}^{\log_2(K)} \frac{K}{2^{k-1}} \exp\Big(-\frac{T}{12A\log_2(K)a\Delta_{(a)}^{-1}}\Big)\\ &\leq 2K\exp\Big(-\frac{T}{12A\log_2(K)\mathbf{H}_4}\Big)\\ &= 2K\exp\Big(-\frac{T}{12(96R^2+\kappa^2)R^2\log_2(K)\mathbf{H}_4}\Big).\end{aligned}$$

Similarly, if $0 < \Delta_y < 3$, we obtain

$$\mathbb{P}[a_T \neq \mathsf{Opt}] \leq 2K\exp\Big(-\frac{T}{36(96R^2+\kappa^2)R^2\log_2(K)\mathbf{H}_2}\Big).$$

By taking $\mathbf{H} = 12(96R^2+\kappa^2)R^2\min(\mathbf{H}_4, 3\mathbf{H}_2)$, we have

$$\mathbb{P}[a_T \neq \mathsf{Opt}] \leq 2K\exp\Big(-\frac{T}{\log_2(K)\mathbf{H}}\Big), \tag{3.64}$$

which completes the proof. □

3.6 Experiments

In this section, we conduct a series of experiments via synthetic and real-world datasets to evaluate PEMV.CB and PEMV.HALVING. We compare the proposed algorithms with two state-of-the-art algorithms, i.e., UCBE of Audibert and Bubeck (2010) and CuRisk of Yu and Nikolova (2013). Note that UCBE searches the optimal arm with the highest mean. Since the key idea in CuRisk is to find the optimal arm via empirical estimation, CuRisk here is implemented to find the optimal arm based on minimal mean-variance estimation.

In experiments, we find that the proposed two algorithms have superior performance in pure exploration for synthetic and real-world datasets. Specifically, the algorithms of PEMV.CB and PEMV.HALVING

Table 3.1: Statistics of used synthetic datasets.

dataset	#arm	$\{\mu(y)\}$	$\{\sigma^2(y)\}$
S1	20	$[1.0, 2.9]$ with a uniform gap	$\sigma^2(11) \sim \sigma^2(15) = 0.6$, $\sigma^2(20) = 0.6$, others 0.3
S2	10	random value in $[0.0, 1.0]$	random value in $[1.0, 2.0]$
S3	30	$\mu(1) = 1.0$, $\mu(y) = 1 - \frac{1.0}{2y^2}$	$\sigma^2(1) = 1.0$, $\sigma^2(y) = 2.0 - \frac{1.0}{2y^2}$

Table 3.2: Probability of error with $\kappa = 1.0$ and $T = 1000$.

algorithm	S1	S2	S3
UCBE	0.63 ± 0.12	0.95 ± 0.04	0.95 ± 0.03
CuRisk	0.43 ± 0.06	0.63 ± 0.11	0.38 ± 0.10
PEMV.CB	0.19 ± 0.10	0.55 ± 0.08	**0.17**± 0.06
PEMV.HALVING	**0.05**± 0.01	**0.40**± 0.12	0.23 ± 0.09

always outperform CuRisk and UCBE in terms of probability of error for synthetic datasets. Besides, for real-world datasets, we conduct the experiment of yearly investments via sliding windows with pure exploration on investment alternatives, and find both proposed algorithms have higher cumulative returns than UCBE and CuRisk.

3.6.1 Settings

To evaluate algorithms in synthetic datasets, we calculate the probability of error based on frequency of wrong decision after exploration.

Specifically, we run multiple epochs of experiments, with each epoch containing 20 independent experiments. For each independent experiment, algorithms output an optimal arm at T. We label 1 for an experiment if the output arm is the true optimal arm. Otherwise we label 0. For 20 experiments of an epoch, we evaluate the probability of error in terms of frequency of zero in labels. Clearly, we have an estimated probability of error in an epoch. By running 10 epochs, we obtain an average of probability of error and its standard error.

In real-world financial data, it is difficult to identify the best investment alternative in hindsight. Thus, it is reasonable to evaluate algorithms via future returns of the chosen alternative. Specifically, in yearly investments, we run algorithms over a sliding window to identify the optimal choice for investments at the beginning of a year. We calculate the performance of returns with the chosen alternative at the end of the year. Via yearly sequential investments, we calculate cumulative returns of algorithms. The higher cumulative returns, the better performance of an algorithm.

3.6.2 Synthetic Data and Results

For verifications, we adopt three synthetic datasets (named as S1-S3) in the experiments, of which statistics are shown in Table 4.3. In S1, the variances are set to satisfy Assumption 3.2. In S2, the values are uniformly randomly generated, and S3 is inspired by Karnin et al. (2013).

From experimental results in Tables 3.2 and 3.3, we find superior performance of the proposed algorithms in terms of the probability of error with a fixed budget of T. Note that the data of the table in bold mean the best performance in the dataset among the four algorithms.

Table 3.3: Probability of error with $\kappa = 10.0$ and $T = 1000$.

algorithm	S1	S2	S3
UCBE	0.32 ± 0.04	0.52 ± 0.10	0.47 ± 0.23
CuRisk	0.56 ± 0.12	0.67 ± 0.11	0.52 ± 0.12
PEMV.CB	0.47 ± 0.17	0.62 ± 0.09	$\mathbf{0.24} \pm 0.03$
PEMV.HALVING	$\mathbf{0.08} \pm 0.05$	$\mathbf{0.47} \pm 0.10$	0.31 ± 0.10

We also show one standard deviation of the performance in the tables. Clearly, PEMV.CB and PEMV.HALVING always outperform CuRisk and UCBE.

In Tables 3.2 and 3.3, we show probability of error on synthetic datasets by different bandit algorithms. We show more comparisons of different κ in Tables 3.4 and 3.5. We can find that the proposed two algorithms are robust for different values of κ.

Besides, to verify the performance of two proposed algorithms with different T, we show experimental results in Figure 3.1. From the figure, we know the probability of error almost decreases with the increase of T. The experimental results are consistent with the theoretical analyses in Theorems 3.1 and 3.2. We also conduct experiments for other parameter settings, and find similar observations.

From the above experimental results, we find the robustness of PEMV.CB and PEMV.HALVING. We also conduct experiments with other parameters for the superiority of two algorithms.

Table 3.4: Probability of error with $\kappa = 0.3$ and $T = 1000$.

algorithm	S1	S2	S3
UCBE	0.70 ± 0.07	0.93 ± 0.07	0.97 ± 0.03
CuRisk	0.59 ± 0.12	0.61 ± 0.10	0.30 ± 0.21
PEMV.CB	0.55 ± 0.06	0.59 ± 0.12	0.14 ± 0.05
PEMV.HALVING	$\mathbf{0.31} \pm 0.09$	$\mathbf{0.38} \pm 0.06$	$\mathbf{0.12} \pm 0.04$

Table 3.5: Probability of error with $\kappa = 0.6$ and $T = 1000$.

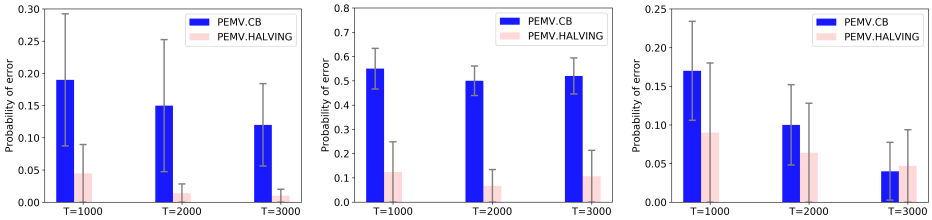
algorithm	S1	S2	S3
UCBE	0.67 ± 0.06	0.96 ± 0.05	0.97 ± 0.03
CuRisk	0.53 ± 0.14	0.62 ± 0.07	0.31 ± 0.08
PEMV.CB	0.36 ± 0.10	0.51 ± 0.10	0.16 ± 0.07
PEMV.HALVING	$\mathbf{0.10} \pm 0.06$	$\mathbf{0.33} \pm 0.12$	$\mathbf{0.15} \pm 0.06$

3.6.3 Financial Data and Results

We conduct yearly sequential investments by adopting the technique of pure exploration. The real data for experiments are historical returns on stocks, bonds and bills of United States from 1928 to 2016[1]. The dataset contains 89 samples of annual returns on SP500, 3-month Treasury Bill and 10-year Treasury Bond, which can be viewed as three arms in bandits for sequential decisions.

We adopt the measure of cumulative returns for performance eval-

[1] http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html



(a) S1 (b) S2 (c) S3

Figure 3.1: Probability of error with different T and $\kappa = 1.0$ in synthetic datasets.

uations, which is defined as

$$C_{ret}(N) = \prod_{i=1}^{N}(1 + r_i), \tag{3.65}$$

where r_i is the realized return for the i-th investment period, and N is the total periods in investments. Clearly, an algorithm performs better if $C_{ret}(N)$ is higher.

For yearly investments of the real dataset, we should output the optimal arm for investments in each year. For example, at the beginning of 2015, we first determine which choice is the best among SP500, 3-month Treasury Bill and 10-year Treasury Bond, and then invest all the available money on that choice. After a year (i.e., at the beginning of 2016), we observe the realized return of the choice in 2015, and sequentially determine the best choice for investments in 2016. This experiment can also be called one-year forward sequential investments.

In Fig. 3.2, we find that PEMV.CB and PEMV.HALVING outperform UCBE and CuRisk in terms of cumulative returns. This reveals that pure exploration of high order statistics in financial scenarios brings better performance in returns.

Now we show more experimental results in Figures 3.3 and 3.4, where $\kappa = 1.0, 1.5, 2.0$ and we adopt different sliding windows. From

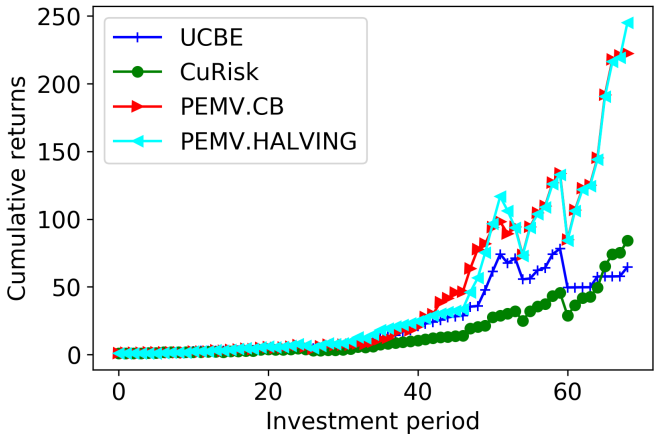


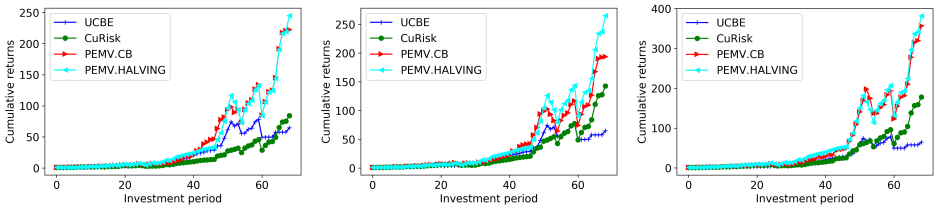
Figure 3.2: Cumulative returns in yearly investments on SP500, 3-month Treasury Bill and 10-year Treasury Bond. The investment is one-year forward from 1947 to 2016.

the figures, we find that the proposed algorithms are robust to the real-world financial data. The algorithms of PEMV.CB and PEMV.HALVING always outperform UCBE and CuRisk in terms of cumulative returns.

Overall, by comparing with state-of-the-art algorithms in pure exploration of MAB, we demonstrate the superiority of the proposed PEMV.CB and PEMV.HALVING in synthetic and real-world datasets. From comprehensive comparisons, we also find that the proposed algorithms are robust in pure exploration with high order statistics.

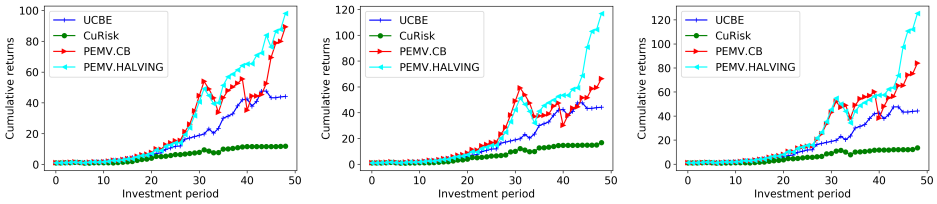
3.7 Conclusion

In this chapter, motivating by optimization of high order statistics in bandits, we investigated the problem of Pure Exploration of Mean-Variance (PEMV). The problem contains three technical challenges, where the core challenge is the analysis of estimation errors due to mean-variance. We have solved the challenges by rigorously proving



(a) $\kappa = 1.0$ (b) $\kappa = 1.5$ (c) $\kappa = 2.0$

Figure 3.3: Cumulative returns in yearly investments on SP500, 3-month Treasury Bill and 10-year Treasury Bond with sliding window $W = 20$. The investment is one-year forward from 1947 to 2016.



(a) $\kappa = 1.0$ (b) $\kappa = 1.5$ (c) $\kappa = 2.0$

Figure 3.4: Cumulative returns in yearly investments on SP500, 3-month Treasury Bill and 10-year Treasury Bond with sliding window $W = 40$. The investment is one-year forward from 1967 to 2016.

that the error resulting from the mean-variance estimation is sub-gamma. Besides, we developed two efficient algorithms to tackle PEMV. With the sub-gamma noises, we derived upper bounds of the probability of error for the proposed algorithms. By conducting a series of experiments on synthetic and real-world datasets, we demonstrated the two algorithms are superior and robust.

Chapter 4

Pure Exploration with Heavy Tails

The model of MAB with sub-Gaussian noises has been well investigated. However, it is surprising to find that less effort has been devoted to the topic of bandits with noises following heavy-tailed distributions. Inspired by heavy-tailed distributions in practical scenarios, we investigate the problem on pure exploration of MAB with heavy-tailed payoffs by breaking the assumption of payoffs with sub-Gaussian noises in bandits, and assuming that stochastic payoffs from bandits are with finite p-th moments, where $p \in (1, +\infty)$.

The main contributions in this chapter are three-fold. First, we technically analyze tail probabilities of empirical average and truncated empirical average (TEA) for estimating expected payoffs in sequential decisions with heavy-tailed noises via martingales. Second, we propose two effective bandit algorithms based on different prior information (i.e., fixed confidence or fixed budget) for pure exploration of MAB generating payoffs with finite p-th moments. Third, we derive theoretical guarantees for the proposed two bandit algorithms, and demonstrate

the effectiveness of two algorithms in pure exploration of MAB with heavy-tailed payoffs in synthetic data and real-world financial data of crytocurrency.

4.1 Introduction

The prevailing decision-making model MAB elegantly characterizes a wide class of practical problems on sequential learning with partial feedbacks, which was first formally proposed and investigated in Robbins (1952). Most algorithms in MAB are primarily developed to maximize cumulative payoffs during a number of rounds for sequential decisions. Recently, there have been interesting investigations on various variants of the traditional MAB model, such as linear bandits (Auer, 2002; Yu et al., 2017b; Zhao and King, 2016), pure exploration of MAB (Audibert and Bubeck, 2010), risk-averse MAB (Sani et al., 2012; Yu et al., 2017a), cascading bandits (Kveton et al., 2015) and clustering bandits (Korda et al., 2016; Li et al., 2016).

One non-trivial branch of MAB is pure exploration, where the goal is to find the optimal arm in a given decision-arm set at the end of exploration. In this case, there is no explicit trade-off between exploration and exploitation for sequential decisions, which means that the exploration phase and the exploitation phase are separated. The problem of pure exploration is motivated by real scenarios which prefer to identify an optimal arm instead of maximizing cumulative payoffs. Recent advances in pure exploration of MAB have found potential applications in many practical domains including communication networks and commercialized products (Audibert and Bubeck, 2010; Chen et al., 2014).

In previous studies on pure exploration of MAB, a common as-

sumption is that noises in observed payoffs are sub-Gaussian. The sub-Gaussian assumption encompasses cases of all bounded payoffs and many unbounded payoffs in MAB, e.g., payoffs of an arm following a Gaussian distribution. However, there exist non-sub-Gaussian noises in observed payoffs for bandits, e.g., high-probability extreme payoffs in sequential decisions which are called heavy-tailed payoffs. A practical motivation example for MAB with heavy-tailed payoffs is the distribution of delays in end-to-end network routing (Liebeherr et al., 2012). Pure exploration of MAB with heavy-tailed payoffs is important, especially for identifications of the potential optimal investment target for practical financial applications. It is worth mentioning that the case of maximizing cumulative payoffs of MAB with heavy tails has been extensively investigated in Bubeck et al. (2013a); Carpentier and Valko (2014); Lattimore (2017); Medina and Yang (2016); Vakili et al. (2013). In Bubeck et al. (2013a), the setting of sequential payoffs with bounded p-th moments was investigated for regret minimization in MAB, where $p \in (1, 2]$. Vakili et al. (2013) introduced bounded p-th moments with the support over $(1, +\infty)$, and provided a complete regret guarantee in MAB. In Medina and Yang (2016), regret guarantee in linear bandits with heavy-tailed payoffs was investigated, which is still scaled by parameters of bounded moments. Recently, payoffs in bandits with bounded kurtosis were discussed in Lattimore (2017).

In this chapter, we investigate the problem on pure exploration of MAB with heavy-tailed payoffs characterized by the bound of p-th moments. It is surprising to find that less effort has been devoted to pure exploration of MAB with heavy-tailed payoffs. Compared with previous work on pure exploration of MAB, the problem of best arm identifcation with heavy-tailed payoffs has three challenges. The first

challenge is the estimate of expected payoffs of an arm in MAB. It might not be sufficient to adopt an empirical average (EA) of observed payoffs with heavy-tailed noises for estimating a true mean. The second challenge is the probability of error for the estimate of expected payoffs, which affects performance of bandit algorithms in pure exploration of MAB. The third challenge is to develop effective bandit algorithms with theoretical guarantees for pure exploration of MAB with heavy-tailed stochastic payoffs.

To solve the above three challenges, we need to introduce a general assumption that stochastic payoffs in MAB are with finite p-th moments, where $p \in (1, +\infty)$. Note that the case of $p \in (1, 2]$ is weaker than the classic assumption of payoffs with sub-Gaussian noises in MAB. Then, under the assumption of finite p-th moments, we present theoretical behaviours of empirical average, and analyze the estimate of truncated empirical average (TEA). Based on different prior information, i.e., fixed confidence or fixed budget, we propose two bandit algorithms in pure exploration of bandits with heavy-tailed payoffs. Finally, based on synthetic data with noises from standard *Student's t-distribution* and real-world financial data, we demonstrate the effectiveness of the proposed bandit algorithms. To the best of our knowledge, this is the first systematic investigation on pure exploration of MAB with heavy-tailed payoffs. For reading convenience, we list contributions of this chapter below.

- We technically analyze tail probabilities of EA and TEA to estimate true mean of arms in MAB with the general assumption of conditionally independent payoffs.

- We propose two bandit algorithms for pure exploration of MAB with heavy-tailed stochastic payoffs characterized by finite p-th

moments, where $p \in (1, +\infty)$.

- We derive theoretical results of the proposed bandit algorithms, as well as demonstrating effectiveness of two algorithms via synthetic data and real-world financial data.

4.2 Preliminaries

In this section, we first present related notations and definitions in this chapter. Then, we present assumptions and the problem definition for pure exploration of MAB with heavy-tailed payoffs.

4.2.1 Notations

Let $\mathcal{A}$ be a bandit algorithm for pure exploration of MAB, which contains K arms at the beginning of exploration. For pure exploration, let Opt be the true optimal arm among K arms, where $\mathsf{Opt} \in [K]$ with $[K] \triangleq \{1, 2, \cdots, K\}$. The total number of sequential rounds for $\mathcal{A}$ to play bandits is T, which is also called as sample complexity. The confidence parameter is denoted by $\delta \in (0, 1)$, which means that, with probability at least $1 - \delta$, $\mathcal{A}$ generates an output optimal arm Out equivalent to Opt, where $\mathsf{Out} \in [K]$. In other words, it happens with a small probability δ that $\mathsf{Opt} \neq \mathsf{Out}$, and δ can be also called the probability of error.

There are two settings based on different prior information given at the beginning of exploration, i.e., fixed confidence or fixed budget. For the setting of fixed confidence, $\mathcal{A}$ receives the information of δ at the beginning, and $\mathcal{A}$ generates Out when a certain condition related to δ is satisfied. For the setting of fixed budget, $\mathcal{A}$ receives the information of T at the beginning, and $\mathcal{A}$ generates Out at the end of T.

We present the learning process on pure exploration of MAB as follows. For $t = 1, 2, \cdots, T$, $\mathcal{A}$ decides to play an arm $a_t \in [K]$ with historical information of $\{a_1, \pi_1(a_1), \cdots, a_{t-1}, \pi_{t-1}(a_{t-1})\}$. Then, $\mathcal{A}$ observes a stochastic payoff $\pi_t(a_t) \in \mathbb{R}$ with respect to a_t, of which the expectation conditional on $\mathcal{F}_{t-1}$ is $\mu(a_t)$ with $\mathcal{F}_{t-1} \triangleq \{a_1, \pi_1(a_1), \cdots, a_{t-1}, \pi_{t-1}(a_{t-1}), a_t\}$ and $\mathcal{F}_0$ being an empty set. Based on $\pi_t(a_t)$, $\mathcal{A}$ updates parameters to proceed with the exploration at $t+1$. We store time index t of playing arm a_t in $\Phi(a_t)$, which is a set with increasing integers.

Given an event $\mathcal{E}$ and a random variable ξ, let $\mathbb{P}[\mathcal{E}]$ be the probability of $\mathcal{E}$ and $\mathbb{E}[\xi]$ be the expectation of ξ. For $x \in \mathbb{R}$, we denote by $|x|$ the absolute value of x, and for a set S, we denote by $|S|$ the cardinality of S. For an event $\mathcal{E}$, let $\mathbb{1}_{[\mathcal{E}]}$ be the indicator function of $\mathcal{E}$.

Definition 4.1. *(Heavy-tailed payoffs in MAB) Given MAB with K arms, let $\pi(k)$ be a stochastic payoff drawn from any arm $k \in [K]$. For $t = 1, \cdots, T$, conditional on $\mathcal{F}_{t-1}$, MAB has heavy-tailed payoffs with the p-th raw moment bounded by B, or the p-th central moment bounded by C, where $p \in (1, +\infty)$, $B, C \in (0, +\infty)$ and $k \in [K]$.*

4.2.2 Problem Definition

It is general to assume that payoffs during sequential decisions contain noises in many practical scenarios. We list the assumptions in this chapter for pure exploration of MAB with heavy-tailed payoffs as follows.

Assumption 4.1. *Assume that $\mathsf{Opt} \triangleq \arg\max_{k \in [K]} \mu(k)$ is unique for pure exploration of MAB with K arms.*

Assumption 4.2. *Assume that MAB has heavy-tailed payoffs with the p-th raw or central moment conditional on $\mathcal{F}_{t-1}$ bounded by B or C, for $t = 1, \cdots, T$.*

Assumption 4.3. *Assume that the sequence of stochastic payoffs from arm $k \in [K]$ has noises with zero mean conditional on $\mathcal{F}_{t-1}$ in pure exploration of MAB. For any time instant $t \in [T]$ and the selected arm a_t, we define random noise of a true payoff as $\xi_t(a_t) \triangleq \pi_t(a_t) - \mu(a_t)$, and assume $\mathbb{E}[\xi_t(a_t)|\mathcal{F}_{t-1}] = 0$.*

Now we present a problem definition for pure exploration of MAB as follows. Given K arms satisfying Assumptions 1–3, the problem in this chapter is to develop a bandit algorithm $\mathcal{A}$ generating an arm $\mathsf{Out}_T \in [K]$ after T pullings of bandits such that $\mathbb{P}[\mathsf{Out}_T \neq \mathsf{Opt}] \leq \delta$, where $\delta \in (0, 1)$.

We discuss theoretical guarantees in two settings for best arm identification of bandits. One is to derive the theoretical guarantee of T by fixing the value of δ, which is called fixed confidence. The other is to derive the theoretical guarantee of δ by fixing the value of T, which is called fixed budget.

For simplicity of notations, we enumerate the arms according to their expected payoffs as a sequence of $\mu(1) > \mu(2) \geq \cdots \geq \mu(K)$. In the ranked sequence, we know that $\mathsf{Opt} = 1$. Note that the ranking operation does not affect our theoretical guarantees. For any arm $k \neq \mathsf{Opt}$ and $k \in [K]$, we define the sub-optimality as $\Delta_k \triangleq \mu(\mathsf{Opt}) - \mu(k)$, which leads to a sequence of sub-optimality as $\{\Delta_k\}_{k=2}^K$. To obtain K terms in sub-optimality, which helps theoretical analyses, we further define $\Delta_1 \triangleq \Delta_2$. Inspired by Audibert and Bubeck (2010), we define the hardness for pure exploration of MAB with heavy-tailed payoffs by quantities as

$$H_2^p \triangleq \max_{k \in [K]} k^{p-1} \Delta_k^{-p}. \tag{4.1}$$

4.3 Related Work

Pure exploration in MAB, aiming at finding the optimal arm after exploration among a given decision-arm set, has become an attracting branch in the decision-making domain (Audibert and Bubeck, 2010; Bubeck et al., 2009; Chen et al., 2014; Gabillon et al., 2012, 2016; Jamieson and Nowak, 2014). It has been pointed out that pure exploration in MAB has many applications, such as communication networks and online advertising.

For pure exploration of MAB with payoffs under sub-Gaussian noises, theoretical guarantees have been well studied. Specifically, in the setting of fixed confidence, the first distribution-dependent lower bound of sample complexity was developed by Mannor and Tsitsiklis (2004), which is $\sum_{k\in[K]} \Delta_k^{-2}$. Even-Dar et al. (2002) originally proposed a bandit algorithm via successive elimination for bounded payoffs with an upper bound of sample complexity matching the lower bound up to a multiplicative logarithmic factor. Karnin et al. (2013) proposed an improved bandit algorithm, which achieves an upper bound of sample complexity matching the lower bound up to a multiplicative doubly-logarithmic factor. Jamieson et al. (2014) proved that it is necessary to have a multiplicative doubly-logarithmic factor in the distribution-dependent lower bound of sample complexity. Jamieson et al. also developed a bandit algorithm via the law of iterated logarithm algorithm for pure exploration of MAB, which achieved the optimal sample complexity of the problem.

In the setting of fixed budget with payoffs under sub-Gaussian noises, Audibert and Bubeck (2010) developed a distribution-dependent lower bound of probability of error, and provided two algorithms, which achieve optimal probability of error up to logarithmic factors. Gabillon

et al. (2012) proposed a unified algorithm for fixed budget and fixed confidence, which discusses ϵ-optimal learning in best arm identification of MAB. Karnin et al. (2013) proposed a bandit algorithm via sequential halving to improve probability of error by a multiplicative constant. It is worth mentioning that Kaufmann et al. (2016) investigated best arm identification of MAB under Gaussian or Bernoulli assumption, and provided lower bounds in terms of Kullback-Leibler divergence. We also notice that there are extensions of best arm identification of MAB, which is multiple-arm identification (Bubeck et al., 2013c; Chen et al., 2014).

To the best of our knowledge, there is no investigation on pure exploration of MAB without the strict assumption of payoffs under sub-Gaussian noises. There are some potential reasons for this fact. One main reason can be that, without sub-Gaussian noises, the tail probabilities of estimates for expected payoffs can be heavy because Chernoff-Hoeffding inequalities of estimates do not hold in general. The failure of Chernoff-Hoeffding inequalities of estimates is a big challenge in pure exploration of MAB. In this chapter, we investigate theoretical performance of pure exploration of MAB with heavy-tailed stochastic payoffs characterized by finite p-th moments, where $p \in (1, +\infty)$. We will put more efforts on $p \in (1, 2]$ because the case of $p \in (2, +\infty)$ enjoys a similar format of $p = 2$. To compare our work with prior studies, we list the distributional assumptions and theoretical guarantees in pure exploration of MAB in Table 4.2. Finally, it is worth mentioning that the case of maximizing expected cumulative payoffs of MAB with heavy tails has been extensively investigated by Bubeck et al. (2013a); Carpentier and Valko (2014); Medina and Yang (2016); Vakili et al. (2013).

Table 4.1: Comparisons on distributional assumptions and theoretical guarantees in pure exploration of MAB for the setting of fixed confidence. Note we omit constant factors in the following inequalities, and H_1, H_2 and H_3 can refer to the corresponding work.

setting	work	assumption on payoffs	algorithm	theoretical guarantee
fixed δ	Even-Dar et al. (2002)	bounded payoffs in $[0,1]$	SE	$\mathbb{P}\left[T \leq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{K}{\delta \Delta_k}\right)\right] \geq 1-\delta$
			ME	$\mathbb{P}\left[T \leq \frac{K}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right] \geq 1-\delta$
	Karnin et al. (2013)	bounded payoffs in $[0,1]$	EGE	$\mathbb{P}\left[T \leq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{1}{\delta} \log\left(\frac{1}{\Delta_k}\right)\right)\right] \geq 1-\delta$
	Jamieson et al. (2014)	sub-Gaussian noise	LILUCB	$\mathbb{P}\left[T \leq H_1 \log\left(\frac{1}{\delta}\right) + H_3\right] \geq 1 - 4\sqrt{c\delta} - 4c\delta$
	Kaufmann et al. (2016)	two-armed Gaussian bandits	α-E	$\mathbb{P}\left[T \leq \frac{(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2} \log\left(\frac{1}{\delta}\right)\right] \geq 1-\delta$
	our work	finite p-th moments	SE-δ(EA)	$\mathbb{P}\left[T \leq \sum_{k=1}^{K} \left(\frac{2^{2p+1} KC}{\Delta_k^p \delta}\right)^{\frac{1}{p-1}}\right] \geq 1-\delta$
		with $p \in (1,2]$	SE-δ(TEA)	$\mathbb{P}\left[T \leq \sum_{k=1}^{K} \left(\frac{20 B^{\frac{1}{p}}}{\Delta_k}\right)^{\frac{p}{p-1}} \log\left(\frac{2K}{\delta}\right)\right] \geq 1-\delta$
		finite p-th central/raw moment	SE-δ(EA)	$\mathbb{P}\left[T \geq \sum_{k=1}^{K} \left(\frac{2^{2p} C_p KC}{\Delta_k^p \delta}\right)^{\frac{2}{p}}\right] \leq 1-\delta$
		with $p \in (2,+\infty)$	SE-δ(TEA)	$\mathbb{P}\left[T \geq \sum_{k=1}^{K} \left(\frac{20 B^{\frac{1}{p}}}{\Delta_k}\right)^{2} \left(\log\left(\frac{2K}{\delta}\right)\right)^{\frac{2p-2}{p}}\right] \leq 1-\delta$

Table 4.2: Comparisons on distributional assumptions and theoretical guarantees in pure exploration of MAB for the setting of fixed budget. Note we omit constant factors in the following inequalities, and H_1, H_2 and H_3 can refer to the corresponding work.

setting	work	assumption on payoffs	algorithm	theoretical guarantee
fixed T	Audibert and Bubeck (2010)	bounded payoffs in $[0,1]$	UCB-E	$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq TK \exp\left(-\frac{T-K}{H_1}\right)$
			SR	$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq K(K-1) \exp\left(-\frac{T-K}{\overline{\log}(K) H_2}\right)$
	Gabillon et al. (2012)	bounded payoffs in $[0,b]$	UGapEb	$\mathbb{P}[\mu_{\mathsf{Out}} - \mu_{\mathsf{Opt}} \geq \epsilon] \leq TK \exp\left(-\frac{T-K}{H_\epsilon}\right)$
	Karnin et al. (2013)	bounded payoffs in $[0,1]$	SH	$\mathbb{P}\left[\mathsf{Out} \neq \mathsf{Opt}\right] \leq \log(K) \exp\left(-\frac{T}{\log(K) H_2}\right)$
	Kaufmann et al. (2016)	two-armed Gaussian bandits	SS	$\mathbb{P}\left[\mathsf{Out} \neq \mathsf{Opt}\right] \leq \exp\left(-\frac{(\mu_1-\mu_2)^2 T}{2(\sigma_1+\sigma_2)^2}\right)$
	our work	finite p-th moments with $p \in (1,2]$	SE-T(EA)	$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq 2^{p+1} CK(K-1) H_2^p \left(\frac{\bar{K}}{T-K}\right)^{p-1}$
			SE-T(TEA)	$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq 2K(K-1) \exp\left(-\frac{(T-K)\bar{B}_1}{\bar{K} K \underline{\Delta}^{p/(1-p)}}\right)$
		finite p-th central/raw moment with $p \in (2,+\infty)$	SE-T(EA)	$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq 2^{p-1} C_p CK(K-1) \bar{H}_2^p \left(\frac{\bar{K}}{T-K}\right)^{\frac{1}{2}}$
			SE-T(TEA)	$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq K(K-1) \exp\left(-\frac{(T-K)\bar{B}_2}{\bar{K} K \underline{\Delta}^{-2}}\right)$

4.4 Algorithms and Analyses

In this section, we first investigate two estimates, i.e., EA and TEA, for expected payoffs of bandits, and derive tail probabilities for EA and TEA under sequential payoffs. Then, we develop two bandit algorithms for best arm identification of MAB in the spirit of successive elimination (SE) and successive rejects (SR). In particular, SE is for the setting of fixed confidence and SR is for the setting of fixed budget. Finally, we derive theoretical guarantees for each bandit algorithm, where we take advantage of EA or TEA.

4.4.1 Empirical Estimates

In SE and SR, it is common for $\mathcal{A}$ to maintain a subset of arms $S_t \subseteq [K]$ at time $t = 1, 2, \cdots$ and $\mathcal{A}$ will output an arm when a certain condition is satisfied, e.g., $|S_t| = 1$ in the setting of fixed confidence. Similar to the most frequently used estimates for expected payoffs in MAB, we consider the following EA to estimate expected payoffs for any arm $k \in S_t$:

$$\hat{\mu}_t(k) \triangleq \frac{1}{s_{t,k}} \sum_{i \in \Phi(k)} \pi_i(k), \tag{4.2}$$

where $s_{t,k} \triangleq |\Phi(k)|$ at time t. Note that the number of elements in $\Phi(k)$ will increase or hold with time evolution, and the elements in $\Phi(k)$ may not successively increase. We also investigate the following estimator TEA for any arm $k \in S_t$:

$$\hat{\mu}_t^{\dagger}(k) \triangleq \frac{1}{s_{t,k}} \sum_{i \in \Phi(k)} \pi_i(k) \mathbb{1}_{[|\pi_i(k)| \leq b_i]}, \tag{4.3}$$

where $b_i > 0$ is a truncating parameter, and b_i will be completely discussed in the ensuing theoretical analyses.

We do not discuss the estimator called median of means (MoM) discussed by Bubeck et al. (2013a), because theoretical guarantees of MoM enjoy similar formats to those of TEA. Before we prove concentration inequalities for estimates via martingales, we have results as below.

Proposition 4.1. *(Dharmadhikari et al., 1968; von Bahr et al., 1965) Let* $\{\nu_i\}_{i=1}^t$ *be random variables satisfying* $\mathbb{E}[|\nu_i|^p] \leq C$ *and* $\mathbb{E}[\nu_i|\mathcal{F}_{i-1}] = 0$*. If* $p \in (1, 2]$*, then we have* $\mathbb{E}\left[\left|\sum_{i=1}^t \nu_i\right|^p\right] \leq 2tC$*. If* $p \in (2, +\infty)$*, then we have* $\mathbb{E}\left[\left|\sum_{i=1}^t \nu_i\right|^p\right] \leq C_p C t^{p/2}$*, where* $C_p \triangleq (8(p-1)\max(1, 2^{p-3}))^p$*.*

Proposition 4.2. *(Seldin et al., 2012) Let* $\{\nu_i\}_{i=1}^t$ *be random variables satisfying* $|\nu_i| \leq b_i$ *with* $\{b_i\}_{i=1}^t$ *being a non-decreasing sequence,* $\mathbb{E}[\nu_i|\mathcal{F}_{i-1}] = 0$ *and* $\mathbb{E}[\nu_i^2|\mathcal{F}_{i-1}]$ *is bounded. Then, with probability* $1-\delta$*, we have* $\left|\sum_{i=1}^t \nu_i\right| \leq b_t \log(2/\delta) + V_t/b_t$*, and* $V_t = \sum_{i=1}^t \mathbb{E}[\nu_i^2|\mathcal{F}_{i-1}]$*.*

Lemma 4.1. *In pure exploration of MAB with* K *arms, for any* $t \in [T]$ *and any arm* $k \in S_t$*, with probability* $1-\delta$

- *for EA, we have*

$$\begin{cases} |\hat{\mu}_t(k) - \mu(k)| \leq \left(\frac{2C}{s_{t,k}^{p-1}\delta}\right)^{\frac{1}{p}}, \ 1 < p \leq 2, \\ |\hat{\mu}_t(k) - \mu(k)| \leq \left(\frac{C_p C}{s_{t,k}^{p/2}\delta}\right)^{\frac{1}{p}}, \ p > 2; \end{cases}$$

- *for TEA, we have*

$$\begin{cases} |\hat{\mu}_t^\dagger(k) - \mu(k)| \leq 5B^{\frac{1}{p}} \left(\frac{\log(2/\delta)}{s_{t,k}}\right)^{\frac{p-1}{p}}, \ 1 < p \leq 2, \\ |\hat{\mu}_t^\dagger(k) - \mu(k)| \leq 5B^{\frac{1}{p}} \left(\frac{\log(2/\delta)}{s_{t,k}}\right)^{\frac{1}{2}}, \ p > 2. \end{cases}$$

Proof. We first prove the results with the estimator $\hat{\mu}_t(k)$ with $k \in S_t$. By Chebyshev's inequality, we have

$$\mathbb{P}[|\hat{\mu}_t(k) - \mu(k)| \geq \delta] \leq \frac{\mathbb{E}[|\hat{\mu}_t(k) - \mu(k)|^p]}{\delta^p} = \frac{\mathbb{E}[|\sum_{i\in\Phi(k)} \pi_i(k) - \mu(k)|^p]}{s_{t,k}^p \delta^p}, \tag{4.4}$$

where $\delta \in (0,1)$ and $s_{t,k}$ is fixed at time t.

Based on Assumption 4.2, we have $\mathbb{E}[|\xi_i(k)|^p] \leq C$ and $\mathbb{E}[\xi_i(k)|\mathcal{F}_{i-1}] = 0$ for any $i \in \Phi(k)$ at t. For $p \in (1,2]$,

$$\mathbb{P}[|\hat{\mu}_t(k) - \mu(k)| \geq \delta] \leq \frac{\mathbb{E}\left[\left|\sum_{i\in\Phi(k)} \xi_i\right|^p\right]}{s_{t,k}^p \delta^p} \leq \frac{2C}{s_{t,k}^{p-1}\delta^p},$$

where we adopt Proposition 4.1. Thus, for any arm $k \in S_t$, with probability at least $1-\delta$

$$|\hat{\mu}_t(k) - \mu(k)| \leq \left(\frac{2C}{s_{t,k}^{p-1}\delta}\right)^{\frac{1}{p}}. \tag{4.5}$$

For $p \in (2, +\infty)$, we have

$$\mathbb{P}[|\hat{\mu}_t(k) - \mu(k)| \geq \delta] \leq \frac{C_p C}{s_{t,k}^{p/2}\delta^p}, \tag{4.6}$$

where we adopt Proposition 4.1. With probability $1-\delta$

$$|\hat{\mu}_t(k) - \mu(k)| \leq \left(\frac{C_p C}{s_{t,k}^{p/2}\delta}\right)^{\frac{1}{p}}. \tag{4.7}$$

Now we prove the results with the estimator $\hat{\mu}_t^\dagger(k)$, where $k \in S_t$. Considering b_i in Eq. (4.3), we define $\mu_i^\dagger(k) \triangleq \mathbb{E}\left[\pi_i(k)\mathbb{1}_{[|\pi_i(k)|\leq b_i]}|\mathcal{F}_{i-1}\right]$, and $\zeta_i(k) \triangleq \mu_i^\dagger(k) - \pi_i(k)\mathbb{1}_{[|\pi_i(k)|\leq b_i]}$, for any $i \in \Phi(k)$. We have $|\zeta_i(k)| \leq 2b_i$, $\mathbb{E}[\zeta_i(k)|\mathcal{F}_{i-1}] = 0$ and $\mathbb{E}\left[\pi_i(k)\mathbb{1}_{[|\pi_i(k)|>b_i]}|\mathcal{F}_{i-1}\right] \leq B/b_i^{p-1}$. Besides, we also have

$$\begin{aligned}\mu(k) - \hat{\mu}_t^\dagger(k) &= \frac{1}{s_{t,k}}\sum_{i\in\Phi(k)}\left[\mu(k) - \mu_i^\dagger(k)\right] + \frac{1}{s_{t,k}}\sum_{i\in\Phi(k)}\left[\mu_i^\dagger(k) - \pi_i(k)\mathbb{1}_{[|\pi_i(k)|\leq b_i]}\right] \\ &= \frac{1}{s_{t,k}}\sum_{i\in\Phi(k)}\left(\mathbb{E}\left[\pi_i(k)\mathbb{1}_{[|\pi_i(k)|>b_i]}|\mathcal{F}_{i-1}\right] + \zeta_i(k)\right),\end{aligned}$$

which implies the inequality of $\mu(k) - \hat{\mu}_t^\dagger(k) \leq \frac{1}{s_{t,k}}\sum_{i\in\Phi(k)}\left(\frac{B}{b_i^{p-1}} + \zeta_i(k)\right)$. For $p \in (1,2]$, we have $\mathbb{E}[\zeta_i^2(k)|\mathcal{F}_{i-1}] \leq \mathbb{E}\left[\pi_i^2(k)\mathbb{1}_{[|\pi_i(k)|\leq b_i]}|\mathcal{F}_{i-1}\right] \leq \frac{B}{b_i^{p-2}}$.

Based on Proposition 4.2, with probability at least $1-\delta$

$$\begin{aligned}\left|\sum_{i\in\Phi(k)}\zeta_i(k)\right| &\leq 2b_t\log(2/\delta)+\frac{1}{2b_t}\sum_{i\in\Phi(k)}\mathbb{E}[\zeta_i^2(k)|\mathcal{F}_{i-1}]\\ &\leq 2b_t\log(2/\delta)+s_{t,k}\frac{B}{2b_t^{p-1}},\end{aligned} \tag{4.8}$$

where we adopt the design of $\{b_i\}_{i\in\Phi(k)}$ as a non-decreasing sequence, i.e., $b_1\leq b_2\leq\cdots\leq b_t$. Thus, by setting $b_t=\left(\frac{Bs_{t,k}}{\log(2/\delta)}\right)^{\frac{1}{p}}$, with probability at least $1-\delta$, we have

$$|\hat{\mu}_t^{\dagger}(k)-\mu(k)|\leq 5B^{\frac{1}{p}}\left(\frac{\log(2/\delta)}{s_{t,k}}\right)^{\frac{p-1}{p}}, \tag{4.9}$$

where we adopt the fact of

$$\frac{1}{s_{t,k}}\sum_{i\in\Phi(k)}\frac{B}{b_i^{p-1}}\leq 2B^{\frac{1}{p}}\left(\frac{\log(2/\delta)}{s_{t,k}}\right)^{\frac{p-1}{p}}. \tag{4.10}$$

For $p\in(2,+\infty)$, by Jensen's inequality, we have

$$\mathbb{E}[\zeta_i^2(k)|\mathcal{F}_{i-1}]\leq B^{\frac{2}{p}}. \tag{4.11}$$

By converting the condition in $p\in(2,+\infty)$ to the condition in $p=2$ with Jensen's inequality and using Eq. (4.9), with probability at least $1-\delta$, we have

$$|\hat{\mu}_t^{\dagger}(k)-\mu(k)|\leq 5B^{\frac{1}{p}}\left(\frac{\log(2/\delta)}{s_{t,k}}\right)^{\frac{1}{2}}, \tag{4.12}$$

which completes the proof. □

Remark 4.1. *In Bubeck et al. (2013a); Vakili et al. (2013), the Bernstein inequality without martingales is adopted with an implicit assumption of sampling payoffs of an arm being independent of sequential decisions, which is informal. By contrast, in Lemma 4.1, conditional on $\mathcal{F}_{t-1}$, the subset S_t is fixed, and we adopt Bernstein inequality with martingales. Thus, we break the assumption of independent payoffs in*

previous work, and prove formal theoretical results of tail probabilities of estimators EA and TEA. Note that the superiority of martingales in sequential decisions has been fully discussed by Zhao et al. (2016).

Remark 4.2. *The concentration results with martingales in Lemma 4.1 for $p \in (1, +\infty)$ can also be applied into regret minimization of heavy-tailed payoffs and other applications in sequential decisions. In particular, we observe that the concentration inequality of $p = 2$ recovers that of payoffs under sub-Gaussian noises. Besides, when $p > 2$, the concentration results indicate constant variations with respect to B. Note that, in Lemma 4.1, we analyze concentration results when $p > 2$, which has not been analyzed by Bubeck et al. (2013a). Compared to Vakili et al. (2013), the concentration result in our work for TEA when $p > 2$ enjoys a constant improvement. Since the case of $p \in (2, +\infty)$ can be resolved by $p = 2$, we will focus on $p \in (1, 2]$ in bandit algorithms for pure exploration of MAB with heavy-tailed payoffs.*

4.4.2 Fixed Confidence

In this subsection, we present a bandit algorithm for pure exploration of MAB with heavy-tailed payoffs under a fixed confidence. Then, we derive upper bounds of sample complexity of the bandit algorithms.

Description of SE-δ

In fixed confidence, we design our bandit algorithm for pure exploration of MAB with heavy-tailed payoffs based on the idea of SE, which is inspired by Even-Dar et al. (2002). For SE-δ(EA), the algorithmic procedures are almost the same as that in Even-Dar et al. (2002), which are omitted here. For SE-δ(TEA), $\mathcal{A}$ will output an arm Out when $|S_t| = 1$ with computation details shown in Algorithm 4.1, where

Algorithm 4.1 Successive Elimination-δ (SE-δ(TEA))

input: δ, K, p, B
initialization: $\hat{\mu}_1^\dagger(k) \leftarrow 0$ for any arm $k \in [K]$, $S_1 \leftarrow [K]$, and $b_1 \leftarrow 0$
$t \leftarrow 1$ ▷ *begin to explore arms in* $[K]$
while $|S_t| > 1$ **do**
  $c_t \leftarrow 5B^{\frac{1}{p}} \left(\frac{\log(2K/\delta)}{t}\right)^{\frac{p-1}{p}}$ ▷ *update confidence bound*
  $b_t \leftarrow \left(\frac{Bt}{\log(2K/\delta)}\right)^{\frac{1}{p}}$ ▷ *update truncating parameter*
  for $k \in S_t$ **do**
    play arm k and observe a payoff $\pi_t(k)$
    $\hat{\mu}_t^\dagger(k) \leftarrow \frac{1}{t}\sum_i^t \pi_i(k)\mathbb{1}_{[|\pi_i(k)| \le b_i]}$ ▷ *calculate TEA*
  end for
  $a_t \leftarrow \arg\max_{k\in[K]} \hat{\mu}_t^\dagger(k)$ ▷ *choose the best arm at* t
  $S_{t+1} \leftarrow \emptyset$ ▷ *create a new arm set for* $t+1$
  for $k \in S_t$ **do**
    if $\hat{\mu}_t^\dagger(a_t) - \hat{\mu}_t^\dagger(k) \le 2c_t$ **then**
      $S_{t+1} \leftarrow S_{t+1} + \{k\}$ ▷ *add arm* k *to* S_{t+1}
    end if
  end for
  $t \leftarrow t+1$ ▷ *update time index*
end while
Out $\leftarrow S_t[0]$ ▷ *assign the first entry of* S_t *to* Out
return: Out

δ is a given parameter. The idea is to eliminate the arm which has the farthest deviation compared with the empirical best arm in S_t.

Theoretical guarantee of SE-δ

We derive upper bounds of sample complexity of SE-δ with estimators of EA and TEA. Note that T is the time complexity of SE-δ.

Theorem 4.1. *For pure exploration in MAB with K arms, with probability at least $1-\delta$, Algorithm SE-δ identifies the optimal arm Opt*

with sample complexity as

- *for* SE-δ*(EA)*

$$T \leq \sum_{k=1}^{K} \left(\frac{2^{2p+1} KC}{\Delta_k^p \delta} \right)^{\frac{1}{p-1}};$$

- *for* SE-δ*(TEA)*

$$T \leq \sum_{k=1}^{K} \left(\frac{20 B^{\frac{1}{p}}}{\Delta_k} \right)^{\frac{p}{p-1}} \log\left(\frac{2K}{\delta} \right),$$

where $p \in (1, 2]$.

4.4.3 Fixed Budget

In this subsection, we present a bandit algorithm for pure exploration of MAB with heavy-tailed payoffs under a fixed budget. Then, we derive upper bounds of probability of error for the bandit algorithms.

Description of SR-T

For SR-T(EA), we omit the algorithm because it is almost the same as that of Audibert and Bubeck (2010). For SR-T(TEA), we design a bandit algorithm for pure exploration of MAB with heavy-tailed payoffs based on the idea of SR, with computation details shown in Algorithm 4.2, where T is a given parameter. The high-level idea is to conduct non-uniform pulling of arms by $K-1$ phases, and SR-T rejects a worst empirical arm for each phase. The reject operation is based on EA or TEA, and we distinguish the two cases by SR-T(EA) and SR-T(TEA).

For simplicity, we show SR-T(TEA) in Algorithm 4.2, where $\underline{\Delta} > 0$ is a design parameter for the estimator of TEA. The design parameter $\underline{\Delta}$ helps to calculate the truncating parameter b in SR-T(TEA). Usually, we set $\underline{\Delta} \leq \Delta_k$ for any $k \in [K]$.

Algorithm 4.2 Successive Rejects-T (SR-T(TEA))

input T, K, p, B, $\underline{\Delta} > 0$
initialization: $\hat{\mu}^{\dagger}(k) \leftarrow 0$ for any arm $k \in [K]$, $S_1 \leftarrow [K]$, $n_0 \leftarrow 0$, $b \leftarrow 0$ and $\bar{K} \leftarrow \sum_{i=1}^{K} \frac{1}{i}$
$b \leftarrow \left(\frac{3Bp}{\underline{\Delta}}\right)^{\frac{1}{p-1}}$ ▷ *calculate truncating parameter*
for $k \in S_1$ **do**
  $\Phi(k) \leftarrow \emptyset$ ▷ *construct sets to store time index*
end for
for $k \in [K-1]$ **do**
  $n_k \leftarrow \lceil \frac{T-K}{\bar{K}(K+1-k)} \rceil$ ▷ *calculate n_k at stage k*
  $n \leftarrow n_k - n_{k-1}$ ▷ *calculate the number of times to pull arms*
  for $y \in S_k$ **do**
    for $i \in [n]$ **do**
      $t \leftarrow t + 1$
      play arm y, and observe a payoff $\pi_t(y)$
      $\Phi(y) \leftarrow \Phi(y) + \{t\}$ ▷ *store time index for arm y*
    end for
    $\hat{\mu}_k^{\dagger}(y) \leftarrow \frac{1}{|\Phi(y)|} \sum_{i \in \Phi(y)} \pi_i(y) \mathbb{1}_{[|\pi_i(y)| \le b]}$
  end for
  $a_k \leftarrow \arg\min_{y \in S_k} \hat{\mu}_t^{\dagger}(y)$ ▷ *choose the worst arm at k*
  $S_{k+1} \leftarrow S_k - \{a_k\}$ ▷ *successively reject arm a_k*
end for
Out $\leftarrow S_K[0]$ ▷ *assign the first entry of S_K to* Out
return: Out

Theoretical guarantee of SR-T

We derive upper bounds of probability of error for SR-T with estimators of EA and TEA. We have the following theorem for SR-T.

Theorem 4.2. *For pure exploration in MAB with K arms, if Algorithm SR-T is run with a fixed budget T, we have probability of error for $p \in (1, 2]$ as*

- *for SR-T(EA)*

$$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq 2^{p+1} CK(K-1) H_2^p \left(\frac{\bar{K}}{T-K}\right)^{p-1};$$

- *for SR-T(TEA)*

$$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq 2K(K-1) \exp\left(-\frac{(T-K)\bar{B}_1}{\bar{K} K \underline{\Delta}^{p/(1-p)}}\right),$$

where $\bar{B}_1 = \frac{p-1}{4(2^p 3Bp^p)^{\frac{1}{p-1}}}$.

4.5 Proofs of Theorems

In this section, we present the proof of theorems.

4.5.1 Proof of Theorem 4.1

Proof. We first consider EA in Eq. (4.2) for estimating the expected payoffs in MAB. For $p \in (1, 2]$, for any arm $k \in S_t$, we have

$$\mathbb{P}[|\hat{\mu}_t(k) - \mu(k)| \geq \delta] \leq \frac{2C}{t^{p-1}\delta^p}, \tag{4.13}$$

where we adopt $s_{t,k} = t$ in SE-δ(EA). We notice the inherent characteristic of SE that, for any arm $k \in S_t$, we have $\Phi(k) = \{1, 2, \cdots, t\}$.

Based on Lemma 4.1, for $t = 1, 2, \cdots$, with probability at least $1 - \delta/K$, the following event holds

$$\mathcal{E}_t \triangleq \{k \in S_t, |\hat{\mu}_t(k) - \mu(k)| \leq c_{t_k}\},$$

where $c_{t_k} = \left(2KC/(t_k^{p-1}\delta)\right)^{\frac{1}{p}}$ is a confidence interval. To eliminate a sub-optimal arm k, we need to play any arm $k \in [K] \backslash \mathsf{Opt}$ with t_k times such that

$$\hat{\Delta}_k \triangleq \hat{\mu}_{t_k}(\mathsf{Opt}) - \hat{\mu}_{t_k}(k) \geq 2c_{t_k}. \tag{4.14}$$

Based on Lemma 4.1, with a high probability, we have

$$\hat{\Delta}_k \geq \mu(\mathsf{Opt}) - c_{t_k} - (\mu(k) + c_{t_k}) = \Delta_k - 2c_{t_k},$$

where c_{t_k} is a confidence interval. To satisfy Eq. (4.14), we are ready to set

$$\Delta_k - 2c_{t_k} \geq 2c_{t_k}. \tag{4.15}$$

To solve the above inequality, we are ready to have that $t_k = \left(\frac{2^{2p+1}KC}{\Delta_k^p \delta}\right)^{\frac{1}{p-1}}$ is sufficient. The total sample complexity is $T = t_2 + \sum_{k=2}^{K} t_k$, because the number of pulling the optimal arm $t_1 = t_2$. This implies, with probability at least $1 - \delta$, we have

$$T \leq \sum_{k=1}^{K} \left(\frac{2^{2p+1}KC}{\Delta_k^p \delta}\right)^{\frac{1}{p-1}}. \tag{4.16}$$

Now we consider TEA in Eq. (4.3) for estimating the expected payoffs in MAB. Similarly, for $p \in (1, 2]$, with probability at least $1 - \delta$, we have

$$T \leq \sum_{k=1}^{K} \left(\frac{20B^{\frac{1}{p}}}{\Delta_k}\right)^{\frac{p}{p-1}} \log\left(\frac{2K}{\delta}\right), \tag{4.17}$$

which completes the proof. □

4.5.2 Proof of Theorem 4.2

Proof. We first consider EA in Eq. (4.2) for estimating the expected payoffs in MAB. For $p \in (1, 2]$, we have

$$\begin{aligned}
\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] &\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}\left[\hat{\mu}_k(\mathsf{Opt}) \leq \hat{\mu}_k(i)\right] \\
&\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}\left[\hat{\mu}_k(i) - \mu(i) + \mu(\mathsf{Opt}) - \hat{\mu}_k(\mathsf{Opt}) \geq \Delta_i\right] \\
&\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \frac{4C}{n_i^{p-1}\left(\frac{\Delta_i}{2}\right)^p} \qquad (4.18) \\
&\leq \sum_{k=1}^{K-1} \frac{2^{p+2}Ck}{n_k^{p-1}\Delta_{K+1-k}^p}, \qquad (4.19)
\end{aligned}$$

where the inequality of Eq. (4.18) is due to the results in Lemma 1 by setting $s_{t,k} = n_k$. Besides, we notice that

$$n_k^{p-1}\Delta_{K+1-k}^p \geq \frac{1}{H_2^p}\left(\frac{T-K}{\bar{K}}\right)^{p-1},$$

which implies that

$$\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq 2^{p+1}CK(K-1)H_2^p\left(\frac{\bar{K}}{T-K}\right)^{p-1}.$$

Now we consider TEA in Eq. (4.3) for estimating the expected payoffs in MAB. By considering the design of b in SR-T(TEA), we have a similar result of Lemma 4.1. Then, for $p \in (1, 2]$, we have probability of error as

$$\begin{aligned}
\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] &\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}\left[\hat{\mu}_k^\dagger(\mathsf{Opt}) \leq \hat{\mu}_k^\dagger(i)\right] \\
&\leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}\left[\hat{\mu}_k^\dagger(i) - \mu(i) + \mu(\mathsf{Opt}) - \hat{\mu}_k^\dagger(\mathsf{Opt}) \geq \underline{\Delta}\right] \\
&\leq 2K(K-1)\exp\left(-\frac{(T-K)\bar{B}_1}{\bar{K}K\underline{\Delta}^{p/(1-p)}}\right), \qquad (4.20)
\end{aligned}$$

which completes the proof. □

Table 4.3: Statistics of used synthetic data.

dataset	#arms	$\{\mu(k)\}$	heavy-tailed $\{p, B, C\}$
S1	10	one arm is 2.0 and nine arms are over [0.7, 1.5] with a uniform gap	{2, 7, 3}
S2	10	one arm is 2.0 and nine arms are over [1.0, 1.8] with a uniform gap	{2, 7, 3}

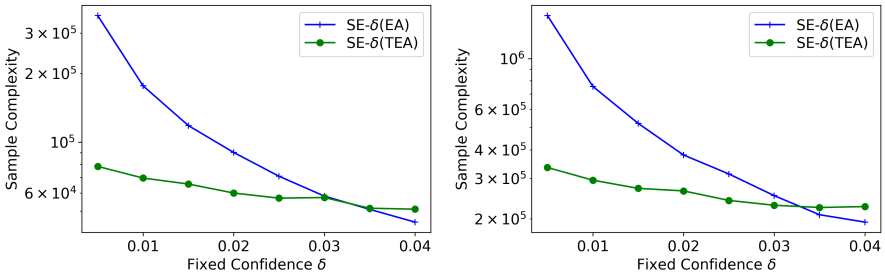
4.6 Experiments

In this section, we conduct experiments via synthetic and real-world data to evaluate the performance of the proposed bandit algorithms. We run experiments in a personal computer with Intel CPU@3.70GHz and 16GB memory. For the setting of fixed confidence, we compare the sample complexities of SE-δ(EA) and SE-δ(TEA). For the setting of fixed budget, we compare the error probabilities of SR-T(EA) and SR-T(TEA).

4.6.1 Synthetic Data and Results

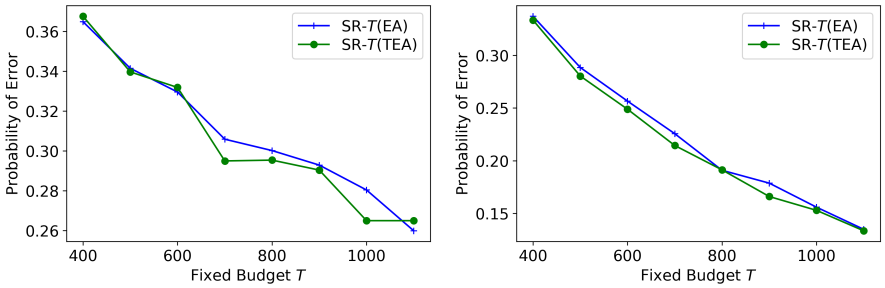
For verifications, we adopt two synthetic data (named as S1-S2) in the experiments, of which statistics are shown in Table 4.3. The data are generated from *Student's t-distribution* with 3 degrees of freedom. In experiments, we run multiple epochs for each dataset, with each epoch containing ten independent experiments for best arm identification of MAB. Besides, we set the value of fixed confidence from 0.005 to 0.040 with a uniform gap of 0.005. We set the value of fixed budget from 400 to 1100 with a uniform gap of 100.

We show experimental results in Figures 4.1 and 4.2, where both



(a) S1 (b) S2

Figure 4.1: Sample complexity for SE-δ in pure exploration of MAB with heavy-tailed payoffs.



(a) S1 (b) S2

Figure 4.2: Probability of error for SR-T in pure exploration of MAB with heavy-tailed payoffs.

proposed algorithms are effective for pure exploration of MAB with heavy-tailed payoffs. In particular, in fixed-confidence setting, sample complexity decreases with increasing value of δ. In fixed-budget setting, probability of error converges to zero with increasing value of T. Besides, for fixed-confidence setting, SE-δ(TEA) beats SE-δ(EA) in both datasets with small δ due to a better control of confidence interval. The experimental results also reflect that the concentration properties of EA are much weaker than those of TEA. For fixed-budget setting, SR-T(TEA) is comparable to SR-T(EA) due to the selection

of truncating parameter.

4.6.2 Financial Data and Results

It has been pointed out that financial data show the inherent characteristic of heavy tails (Panahi, 2016), because the probability of events with a large deviation is high in financial markets. Due to the availability of financial data on the Internet, we choose a financial application of identifying the most profitable cryptocurrency over a period of time in a given pool of digital currencies. The identification for the most profitable cryptocurrency among the top ten cryptocurrency in terms of market value is motivated by the practical scenario that an investor would like to invest a fixed budget of money in a cryptocurrency and get return as much as possible.

We get hourly price data of cryptocurrency via the Internet[1], which include the open price, the closed price, the highest price and the lowest price of each hour. From historical financial data in digital currency, we observe that high fluctuations of price of cryptocurrency reflect a significant characteristic of heavy tails, which is pretty ideal for the task of pure exploration of MAB with heavy-tailed payoffs. For experiments, we choose top ten cryptocurrencies in terms of market value, with basic information shown in Table 4.4.

We show the statistics of real data in Table 4.5. In the table, we conduct a statistical analysis in hindsight with hourly returns of cryptocurrency from February 3rd, 2018 to April 27th, 2018. From the table, we find that the optimal option in hindsight is EOS in terms of the maximal empirical mean of hourly payoffs. Besides, we conduct Kolmogrov-Smirnov (KS) test to fit real data of a cryptocurrency to a

[1] `https://www.cryptocompare.com/`

Table 4.4: Ten selected cryptocurrencies in experiments.

full name	symbol	market value in April 2018 (unit: billion US dollar)
Bitcoin	BTC	155
Ethereum Classic	ETC	66
Ripple	XRP	32
Bitcoin Cash	BCH	23
EOS	EOS	15
Litecoin	LTC	8
Cardano	ADA	8
Stellar	XLM	7
IOTA	IOT	5
NEO	NEO	5

distribution. In particular, via KS test, we know that the null hypothesis of real data following a Gaussian distribution is rejected, because $\bar{p}$-value is smaller than a significant level of 0.05. We observe that real data of cryptocurrency are likely to follow a *Student's t-distribution* via KS test in Table 4.5.

With the above statistical analyses, we can fit real data of cryptocurrency to a *Student's t-distribution*, and obtain distribution pa-

Table 4.5: Statistical property of ten selected cryptocurrencies with hourly returns from Feb. 3rd, 2018 to Apr. 27th, 2018. KS-test1 denotes Kolmogrov-Smirnov (KS) test with a null hypothesis that real data follow a Gaussian distribution. KS-test2 denotes KS test with a null hypothesis that real data follow a *Student's t-distribution*.

symbol	empirical statistics (mean$\times 10^3$, variance$\times 10^3$)	KS-test1 (statistic, $\bar{p}$-value)	KS-test2 (statistic, $\bar{p}$-value)
BTC	$(0.36, 0.54)$	$(0.08, 0.005)$	$(0.05, 0.20)$
ETC	$(0.29, 1.03)$	$(0.07, 0.02)$	$(0.03, 0.89)$
XRP	$(0.33, 0.94)$	$(0.09, 0.0004)$	$(0.03, 0.61)$
BCH	$(0.78, 0.92)$	$(0.08, 0.001)$	$(0.03, 0.64)$
EOS	$(\mathbf{1.56}, 1.18)$	$(0.09, 0.0002)$	$(0.03, 0.88)$
LTC	$(0.68, 0.86)$	$(0.10, 0.0002)$	$(0.04, 0.49)$
ADA	$(0.02, 1.22)$	$(0.07, 0.03)$	$(0.02, 0.99)$
XLM	$(0.62, 0.12)$	$(0.07, 0.02)$	$(0.03, 0.80)$
IOT	$(0.68, 0.11)$	$(0.07, 0.02)$	$(0.04, 0.57)$
NEO	$(-0.31, 1.26)$	$(0.10, 0.0002)$	$(0.04, 0.53)$

rameters shown in Table 4.6. Based on the property of *Student's t-distribution*, we can set $p = 2$, and estimate B and C as shown in the table. The estimated parameter of p reflects that financial data contain a finite variance, which is reasonable.

Via a similar setting to that of synthetic data, we show the results on pure exploration of top ten cryptocurrencies in Figure 4.3. Note that, due to limitation of data points in the setting of fixed confidence, we generate payoffs from *Student's t-distributions* fitting to real data.

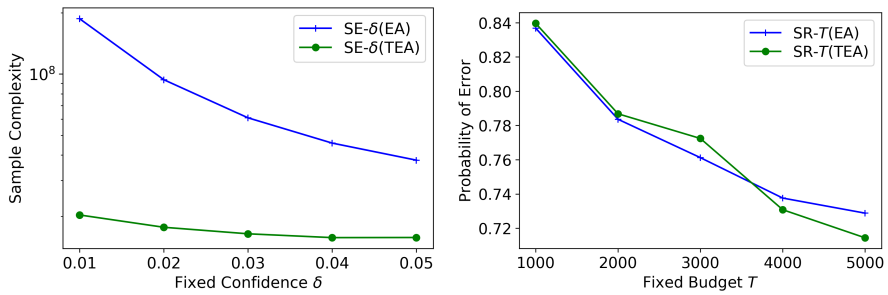
Table 4.6: Estimated parameters for ten cryptocurrencies.

symbol	degree of freedom	(p, B, C) in experiments
BTC	3.50	
ETC	3.81	
XRP	2.53	
BCH	3.00	
EOS	2.90	
LTC	2.75	$(2, 1.577\times10^{-3}, 1.575\times10^{-3})$
ADA	3.55	
XLM	3.81	
IOT	4.66	
NEO	3.13	

But in the setting of fixed budget, we adopt exactly real financial data. We have similar observations as those in synthetic data. It is worth mentioning that, TEA algorithm outperforms EA algorithm in fixed-confidence setting when the value of δ is small. Besides, TEA is comparable to EA in fixed-budget setting because the truncating parameter in Algorithm 4.2 only has budget information and does not increase with the number of samples. Overall, with synthetic and real-world data, we have verified the effectiveness of our two algorithms.

4.7 Conclusion

In this chapter, we broke the assumption of payoffs under sub-Gaussian noises in pure exploration of MAB, and investigated best arm identification of MAB with a general assumption that the p-th moments of



(a) fixed confidence (b) fixed budget

Figure 4.3: Pure exploration of cryptocurrency.

stochastic payoffs are bounded, where $p \in (1, +\infty)$. We have technically analyzed tail probabilities of empirical average and truncated empirical average for estimating expected payoffs in sequential decisions. Besides, we proposed two bandit algorithms for pure exploration of MAB with heavy-tailed payoffs based on SE and SR. Finally, we derived theoretical guarantees of the proposed bandit algorithms, and demonstrated the effectiveness of bandit algorithms in pure exploration of MAB with heavy-tailed payoffs.

Chapter 5

Linear Stochastic Bandits with Heavy Tails

For linear stochastic bandits, where the expected payoff of an arm is a linear function, theoretical guarantees under sub-Gaussian assumptions, e.g., the observed stochastic payoffs are bounded, have been well investigated. The model of linear stochastic bandits is an important variant of MAB. However, many practical applications for sequential decisions contain non-sub-Gaussian noises, especially heavy-tailed noises with finite moments for the p-th order, where $p \in (1, 2]$.

In previous studies, it is commonly assumed that payoffs are with sub-Gaussian noises for linear stochastic bandits. In this chapter, under a weaker assumption on noises, we study the problem of Linear stochastic Bandits with hEavy-Tailed payoffs (LinBET), where the distributions have finite moments of order p, for some $p \in (1, 2]$. We rigorously analyze the regret lower bound of LinBET as $\Omega(T^{\frac{1}{p}})$, implying that finite moments of order 2 (i.e., finite variances) yield the bound of $\Omega(\sqrt{T})$, with T being the total number of rounds to play bandits. The provided lower bound also indicates that the state-of-the-art algorithms

for LinBET are far from optimal. By adopting median of means with a well-designed allocation of decisions and truncation based on historical information, we develop two novel bandit algorithms, where the regret upper bounds match the lower bound up to polylogarithmic factors. To the best of our knowledge, we are the first to solve LinBET optimally in the sense of the polynomial order on T. Our proposed algorithms are evaluated based on synthetic datasets, and outperform the state-of-the-art results.

5.1 Introduction

Bandit algorithms usually aims at maximizing cumulative payoffs over a sequence of rounds. A natural and important variant of MAB is linear stochastic bandits with the expected payoff of each arm satisfying a linear mapping from the arm information to a real number. The model of linear stochastic bandits enjoys some good theoretical properties, e.g., there exists a closed-form solution of the linear mapping at each time step in light of ridge regression. Many practical applications take advantage of MAB and its variants to control decision performance, e.g., online personalized recommendations (Li et al., 2010) and resource allocations (Lattimore et al., 2014).

In most previous studies of MAB and linear stochastic bandits, a common assumption is that noises in observed payoffs are supposed to be sub-Gaussian conditional on historical information (Abbasi-Yadkori et al., 2011; Bubeck et al., 2012), which encompasses cases of all bounded payoffs and many unbounded payoffs, e.g., payoffs of an arm following a Gaussian distribution. However, there do exist practical scenarios of non-sub-Gaussian noises in observed payoffs for sequential decisions, such as high-probability extreme returns in investments for financial

markets (Cont and Bouchaud, 2000) and fluctuations of neural oscillations (Roberts et al., 2015), which are called heavy-tailed noises. Thus, it is significant to completely study theoretical behaviours of sequential decisions in the case of heavy-tailed noises.

Many practical distributions, e.g., Pareto distributions and Weibull distributions, are heavy-tailed, which perform high tail probabilities compared with exponential family distributions. We consider a general characterization of heavy-tailed payoffs in bandits, where the distributions have finite moments of order p, where $p \in (1, 2]$. When $p = 2$, stochastic payoffs are generated from distributions with finite variances. When $1 < p < 2$, stochastic payoffs are generated from distributions with infinite variances (Shao and Nikias, 1993). Note that, different from sub-Gaussian noises in the traditional bandit setting, noises from heavy-tailed distributions do not enjoy exponentially decaying tails, and thus make it more difficult to learn a parameter of an arm.

The regret of MAB with heavy-tailed payoffs has been well addressed by Bubeck et al. (2013a), where stochastic payoffs have bounds on raw or central moments of order p. For MAB with finite variances (i.e., $p = 2$), the regret of truncation algorithms or median of means recovers the optimal regret for MAB under the sub-Gaussian assumption. Recently, Medina and Yang (2016) investigated theoretical guarantees for the problem of Linear stochastic Bandits with hEavy-Tailed payoffs (LinBET). It is surprising to find that, for $p = 2$, the regret of bandit algorithms by Medina and Yang (2016) to solve LinBET is $\widetilde{O}(T^{\frac{3}{4}})$ [1], which is far away from the regret of the state-of-the-art algorithms (i.e., $\widetilde{O}(\sqrt{T})$) in linear stochastic bandits under the sub-Gaussian assumption (Abbasi-Yadkori et al., 2011; Dani et al., 2008a). Thus, the most

[1] We omit polylogarithmic factors of T for $\widetilde{O}(\cdot)$.

interesting and non-trivial question is

Is it possible to recover the regret of $\widetilde{O}(\sqrt{T})$ *when* $p = 2$ *for LinBET?*

In this chapter, we answer this question affirmatively. Specifically, we investigate the problem of LinBET characterized by finite p-th moments, where $p \in (1, 2]$. The problem of LinBET raises several interesting challenges. The first challenge is the lower bound of the problem, which remains unknown. The technical issues come from the construction of an elegant setting for LinBET, and the derivation of a lower bound with respect to p. The second challenge is how to develop a robust estimator for the parameter in LinBET, because heavy-tailed noises greatly affect errors of the conventional least-squares estimator. It is worth mentioning that Medina and Yang (2016) has tried to tackle this challenge, but their estimators do not make full use of the contextual information of chosen arms to eliminate the effect from heavy-tailed noises, which eventually leads to large regrets. The third challenge is how to successfully adopt median of means and truncation to solve LinBET with regret upper bounds matching the lower bound as closely as possible.

To solve the above challenges, first of all, we rigorously analyze the lower bound on the problem of LinBET, which enjoys a polynomial order on T as $\Omega(T^{\frac{1}{p}})$. The lower bound provides two essential hints: one is that finite variances in LinBET yield a bound of $\Omega(\sqrt{T})$, and the other is that algorithms by Medina and Yang (2016) are sub-optimal. Then, we develop two novel bandit algorithms to solve LinBET based on the basic techniques of median of means and truncation. Both the algorithms adopt the optimism in the face of uncertainty principle, which is common in bandit problems (Abbasi-Yadkori et al., 2011; Munos et al., 2014). The regret upper bounds of the proposed two algorithms, which

are $\widetilde{O}(T^{\frac{1}{p}})$, match the lower bound up to polylogarithmic factors. To the best of our knowledge, we are the first to solve LinBET almost optimally. We conduct experiments based on synthetic datasets, which are generated by Student's t-distribution and Pareto distribution, to demonstrate the effectiveness of our algorithms. Experimental results show that our algorithms outperform the state-of-the-art results. The contributions of this chapter are summarized as follows:

- We provide the lower bound for the problem of LinBET characterized by finite p-th moments, where $p \in (1, 2]$. In the analysis, we construct an elegant setting of LinBET, which results in a regret bound of $\Omega(T^{\frac{1}{p}})$ in expectation for any bandit algorithm.
- We develop two novel bandit algorithms, which are named as MENU and TOFU (with details shown in Section 5.4). The MENU algorithm adopts median of means with a well-designed allocation of decisions and the TOFU algorithm adopts truncation via historical information. Both algorithms achieve the regret $\widetilde{O}(T^{\frac{1}{p}})$ with high probability.
- We conduct experiments based on synthetic datasets to demonstrate the effectiveness of our proposed algorithms. By comparing our algorithms with the state-of-the-art results, we show improvements on cumulative payoffs for MENU and TOFU, which are strictly consistent with theoretical guarantees in this chapter.

5.2 Preliminaries and Related Work

In this section, we first present preliminaries, i.e., notations and learning setting of LinBET. Then, we give a detailed discussion on the line of research for bandits with heavy-tailed payoffs.

5.2.1 Notations

For a positive integer K, $[K] \triangleq \{1, 2, \cdots, K\}$. Let the ℓ-norm of a vector $x \in \mathbb{R}^d$ be $\|x\|_\ell \triangleq (x_1^\ell + \cdots + x_d^\ell)^{\frac{1}{\ell}}$, where $\ell \geq 1$ and x_i is the i-th element of x with $i \in [d]$. For $r \in \mathbb{R}$, its absolute value is $|r|$, its ceiling integer is $\lceil r \rceil$, and its floor integer is $\lfloor r \rfloor$. The inner product of two vectors x, y is denoted by $x^\top y = \langle x, y \rangle$. Given a positive definite matrix $A \in \mathbb{R}^{d \times d}$, the weighted Euclidean norm of a vector $x \in \mathbb{R}^d$ is $\|x\|_A = \sqrt{x^\top A x}$. $\mathbb{B}(x, r)$ denotes a Euclidean ball centered at x with radius $r \in \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of positive numbers. Let e be Euler's number, and $I_d \in \mathbb{R}^{d \times d}$ an identity matrix. Let $\mathbb{1}_{\{\cdot\}}$ be an indicator function, and $\mathbb{E}[X]$ the expectation of X.

5.2.2 Learning Setting

For a bandit algorithm $\mathcal{A}$, we consider sequential decisions with the goal to maximize cumulative payoffs, where the total number of rounds for playing bandits is T. For each round $t = 1, \cdots, T$, the bandit algorithm $\mathcal{A}$ is given a decision set $D_t \subseteq \mathbb{R}^d$ such that $\|x\|_2 \leq D$ for any $x \in D_t$. $\mathcal{A}$ has to choose an arm $x_t \in D_t$ and then observes a stochastic payoff $y_t(x_t)$. For notation simplicity, we also write $y_t = y_t(x_t)$. The expectation of the observed payoff for the chosen arm satisfies a linear mapping from the arm to a real number as $y_t(x_t) \triangleq \langle x_t, \theta_* \rangle + \eta_t$, where θ_* is an underlying parameter with $\|\theta_*\|_2 \leq S$ and η_t is a random noise. Without loss of generality, we assume $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1} \triangleq \{x_1, \cdots, x_t\} \cup \{\eta_1, \cdots, \eta_{t-1}\}$ is a σ-filtration and $\mathcal{F}_0 = \emptyset$. Clearly, we have $\mathbb{E}[y_t(x_t) | \mathcal{F}_{t-1}] = \langle x_t, \theta_* \rangle$. For an algorithm $\mathcal{A}$, to maximize cumulative payoffs is equivalent to minimizing the regret as

$$\mathbf{R}(\mathcal{A}, T) \triangleq \left(\sum_{t=1}^{T} \langle x_t^*, \theta_* \rangle \right) - \left(\sum_{t=1}^{T} \langle x_t, \theta_* \rangle \right) = \sum_{t=1}^{T} \langle x_t^* - x_t, \theta_* \rangle, \qquad (5.1)$$

where x_t^* denotes the optimal decision at time t for θ_*, i.e., $x_t^* \in \arg\max_{x \in D_t} \langle x, \theta_* \rangle$. In this chapter, we will provide high-probability upper bound of $\mathbf{R}(\mathcal{A}, T)$ with respect to $\mathcal{A}$, and provide the lower bound for LinBET in expectation for any algorithm. The problem of LinBET is defined as below.

Definition 5.1 (LinBET)**.** *Given a decision set D_t for time step $t = 1, \cdots, T$, an algorithm $\mathcal{A}$, of which the goal is to maximize cumulative payoffs over T rounds, chooses an arm $x_t \in D_t$. With $\mathcal{F}_{t-1}$, the observed stochastic payoff $y_t(x_t)$ is conditionally heavy-tailed, i.e., $\mathbb{E}\left[|y_t|^p | \mathcal{F}_{t-1}\right] \leq b$ or $\mathbb{E}\left[|y_t - \langle x_t, \theta_* \rangle|^p | \mathcal{F}_{t-1}\right] \leq c$, where $p \in (1, 2]$, and $b, c \in (0, +\infty)$.*

5.2.3 Related Work

The model of MAB dates back to 1952 with the original work by Robbins (1952), and its inherent characteristic is the trade-off between exploration and exploitation. The asymptotic lower bound of MAB was developed by Lai and Robbins (1985), which is logarithmic with respect to the total number of rounds. An important technique called upper confidence bound was developed to achieve the lower bound (Agrawal, 1995; Lai and Robbins, 1985). Other related techniques to solve the problem of sequential decisions include Thompson sampling (Agrawal and Goyal, 2012; Chapelle and Li, 2011; Thompson, 1933) and Gittins index (Gittins et al., 2011).

The problem of MAB with heavy-tailed payoffs characterized by finite p-th moments has been well investigated (Bubeck et al., 2013a; Vakili et al., 2013; Yu et al., 2018). Bubeck et al. (2013a) pointed out that finite variances in MAB are sufficient to achieve regret bounds of the same order as the optimal regret for MAB under the sub-Gaussian

assumption, and the order of T in regret bounds increases when ϵ decreases. The lower bound of MAB with heavy-tailed payoffs has been analyzed (Bubeck et al., 2013a), and robust algorithms by Bubeck et al. (2013a) are optimal. Theoretical guarantees by Bubeck et al. (2013a); Vakili et al. (2013) are for the setting of finite arms. In Vakili et al. (2013), primary theoretical results were presented for the case of $p > 2$. We notice that the case of $p > 2$ is not interesting, because it reduces to the case of finite variances in MAB.

For the problem of linear stochastic bandits, which is also named linear reinforcement learning by Auer (2002), the lower bound is $\Omega(d\sqrt{T})$ when contextual information of arms is from a d-dimensional space (Dani et al., 2008b). Bandit algorithms matching the lower bound up to polylogarithmic factors have been well developed (Abbasi-Yadkori et al., 2011; Auer, 2002; Chu et al., 2011; Dani et al., 2008a). Notice that all these studies assume that stochastic payoffs contain sub-Gaussian noises. More variants of MAB can be discussed by Bubeck et al. (2012).

It is surprising to find that the lower bound of LinBET remains unknown. In Medina and Yang (2016), bandit algorithms based on truncation and median of means were presented. When $p \in (1, 2]$ for LinBET, the algorithms by Medina and Yang (2016) cannot recover the bound of $\widetilde{O}(\sqrt{T})$ which is the regret of the state-of-the-art algorithms in linear stochastic bandits under the sub-Gaussian assumption. Medina and Yang (2016) conjectured that it is possible to recover $\widetilde{O}(\sqrt{T})$ with p being a finite number for LinBET. Thus, it is urgent to conduct a thorough analysis of the conjecture in consideration of the importance of heavy-tailed noises in real scenarios. Solving the conjecture generalizes the practical applications of bandit models. Practical motivating examples for bandits with heavy-tailed payoffs include delays

in end-to-end network routing (Liebeherr et al., 2012) and sequential investments in financial markets (Cont and Bouchaud, 2000).

Recently, the assumption in stochastic payoffs of MAB was relaxed from sub-Gaussian noises to bounded kurtosis (Lattimore, 2017), which can be viewed as an extension of Bubeck et al. (2013a). The interesting point of Lattimore (2017) is the scale free algorithm, which might be practical in applications. Besides, Carpentier and Valko (2014) investigated extreme bandits, where stochastic payoffs of MAB follow Fréchet distributions. The setting of extreme bandits fits for the real scenario of anomaly detection without contextual information. The order of regret in extreme bandits is characterized by distributional parameters, which is similar to the results by Bubeck et al. (2013a).

It is worth mentioning that, for linear regression with heavy-tailed noises, several interesting studies have been conducted. Hsu and Sabato (2016) proposed a generalized method in light of median of means for loss minimization with heavy-tailed noises. Heavy-tailed noises in Hsu and Sabato (2016) might come from contextual information, which is more complicated than the setting of stochastic payoffs in this chapter. Therefore, linear regression with heavy-tailed noises usually requires a finite fourth moment. In Audibert et al. (2011), the basic technique of truncation was adopted to solve robust linear regression in the absence of exponential moment condition. The related studies in this line of research are not directly applicable for the problem of LinBET.

5.3 Lower Bound

In this section, we provide the lower bound for LinBET. We consider heavy-tailed payoffs with finite p-th raw moments in the analysis. In particular, we construct the following setting. Assume $d \geq 2$ is even

(when d is odd, similar results can be easily derived by considering the first $d-1$ dimensions). For $D_t \subseteq \mathbb{R}^d$ with $t \in [T]$, we fix the decision set as $D_1 = \cdots = D_T = D_{(d)}$. Then, the fixed decision set is constructed as $D_{(d)} \triangleq \{(x_1, \cdots, x_d) \in \mathbb{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$, which is a subset of intersection of the cube $[0,1]^d$ and the hyperplane $x_1 + \cdots + x_d = d/2$. We define a set $S_d \triangleq \{(\theta_1, \cdots, \theta_d) : \forall i \in [d/2], (\theta_{2i-1}, \theta_{2i}) \in \{(2\Delta, \Delta), (\Delta, 2\Delta)\}\}$ with $\Delta \in (0, 1/d]$. The payoff functions take values in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ such that, for every $x \in D_{(d)}$, the expected payoff is $\theta_*^\top x$. To be more specific, we have the payoff function of x as

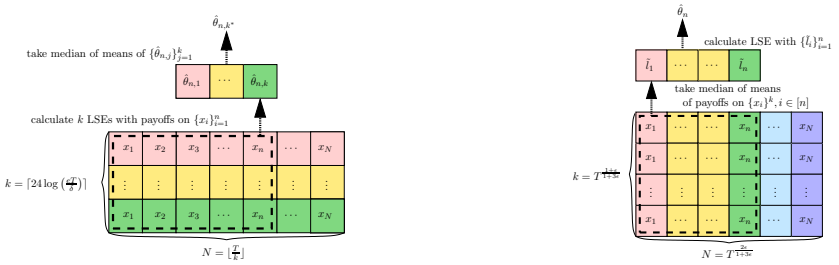
$$y(x) = \begin{cases} \left(\frac{1}{\Delta}\right)^{\frac{1}{p-1}} & \text{with a probability of } \Delta^{\frac{1}{p-1}} \theta_*^\top x, \\ 0 & \text{with a probability of } 1 - \Delta^{\frac{1}{p-1}} \theta_*^\top x. \end{cases} \tag{5.2}$$

We have the theorem for the lower bound of LinBET as below.

Theorem 5.1. *If θ_* is chosen uniformly at random from S_d, and the payoff for each $x \in D_{(d)}$ is in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ with mean $\theta_*^\top x$, then for any algorithm $\mathcal{A}$ and every $T \geq (d/12)^{\frac{p-1}{p}}$, we have*

$$\mathbb{E}\left[R(\mathcal{A}, T)\right] \geq \frac{d}{192} T^{\frac{1}{p}}. \tag{5.3}$$

In the proof of Theorem 5.1, we first prove the lower bound when $d = 2$, and then generalize the argument to any $d > 2$. We notice that the parameter in the original d-dimensional space is rearranged to $d/2$ tuples, each of which is a 2-dimensional vector as $(\theta_{2i-1}, \theta_{2i}) \in \{(2\Delta, \Delta), (\Delta, 2\Delta)\}$ with $i \in [d/2]$. If the i-th tuple of the parameter is selected as $(2\Delta, \Delta)$, then the i-th tuple of the optimal arm is $(x_{*,2i-1}, x_{*,2i}) = (1, 0)$. In this case, if we define the i-th tuple of the chosen arm as $(x_{t,2i-1}, x_{t,2i})$, the instantaneous regret is $\Delta(1 - x_{t,2i-1})$. Then, the regret can be represented as an integration of $\Delta(1 - x_{t,2i-1})$



(a) Framework of MENU (b) Framework of MoM

Figure 5.1: Framework comparison between our MENU and MoM by Medina and Yang (2016).

over $D_{(d)}$. Finally, with common inequalities in information theory, we obtain the regret lower bound by setting $\Delta = T^{-\frac{p-1}{p}}/12$.

We notice that martingale differences to prove the lower bound for linear stochastic bandits in (Dani et al., 2008a) are not directly feasible for the proof of lower bound in LinBET, because under our construction of heavy-tailed payoffs (i.e., Eq. (5.4)), the information of p is excluded. Besides, our proof is partially inspired by Bubeck (2010). We show the detailed proof of Theorem 5.1 in Section 5.5.

Remark 5.1. *The above lower bound provides two essential hints for bandit algorithms: one is that finite variances in LinBET yield a bound of $\Omega(\sqrt{T})$, and the other is that algorithms proposed by Medina and Yang (2016) are far from optimal. The result in Theorem 5.1 strongly indicates that it is possible to design bandit algorithms recovering $\widetilde{O}(\sqrt{T})$ with finite variances.*

5.4 Algorithms and Upper Bounds

In this section, we develop two novel bandit algorithms to solve LinBET, which turns out to be almost optimal. We rigorously prove regret

Algorithm 5.1 MENU

input d, c, p, δ, λ, S, T, $\{D_n\}_{n=1}^N$
initialization: $k = \lceil 24 \log\left(\frac{eT}{\delta}\right) \rceil$, $N = \lfloor \frac{T}{k} \rfloor$, $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(\mathbf{0}, S)$
for $n = 1, 2, \cdots, N$ **do**
  $(x_n, \tilde{\theta}_n) = \arg\max_{(x,\theta) \in D_n \times C_{n-1}} \langle x, \theta \rangle$ ▷ *to select an arm*
  Play x_n with k times and observe payoffs $y_{n,1}, y_{n,2}, \cdots, y_{n,k}$
  $V_n = V_{n-1} + x_n x_n^\top$
  For $j \in [k]$, $\hat{\theta}_{n,j} = V_n^{-1} \sum_{i=1}^n y_{i,j} x_i$ ▷ *to calculate LSE for the j-th group*
  For $j \in [k]$, let r_j be the median of $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} : s \in [k] \backslash j\}$
  $k^* = \arg\min_{j \in [k]} r_j$ ▷ *to take median of means of estimates*
  $\beta_n = 3\left((9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right)$
  $C_n = \{\theta : \|\theta - \hat{\theta}_{n,k^*}\|_{V_n} \leq \beta_n\}$ ▷ *to update the confidence region*
end for

upper bounds for the proposed algorithms. In particular, our core idea is based on the Optimism in the Face of Uncertainty principle (OFU). The first algorithm is MEdian of meaNs under optimism in the face of Uncertainty (MENU) shown in Algorithm 5.1, and the second algorithm is Truncation under Optimism in the Face of Uncertainty (TOFU) shown in Algorithm 5.2. For comparisons, we directly name the bandit algorithm based on median of means in Medina and Yang (2016) as MoM, and name the bandit algorithm based on confidence region with truncation in Medina and Yang (2016) as CRT.

Both algorithms in this chapter adopt the tool of ridge regression. At time step t, let $\hat{\theta}_t$ be the ℓ^2-regularized least-squares estimate (LSE) of θ_* as $\hat{\theta}_t = V_t^{-1} X_t^\top Y_t$, where $X_t \in \mathbb{R}^{t \times d}$ is a matrix of which rows are $x_1^\top, \cdots, x_t^\top$, $V_t \triangleq X_t^\top X_t + \lambda I_d$, $Y_t \triangleq (y_1, \cdots, y_t)$ is a vector of the historical observed payoffs until time t and $\lambda > 0$ is a regularization parameter.

5.4.1 MENU and Regret

Description of MENU. To conduct median of means in LinBET, it is common to allocate T pulls of bandits among $N \leq T$ epochs, and for each epoch the same arm is played multiple times to obtain an estimate of θ_*. We find that there exist different ways to contruct the epochs. We design the framework of MENU in Figure 5.1(a), and show the framework of MoM designed by Medina and Yang (2016) in Figure 5.1(b). For MENU and MoM, we have the following three differences. First, for each epoch $n = 1, \cdots, N$, MENU plays the same arm x_n by $O(\log(T))$ times, while MoM plays the same arm by $O(T^{\frac{p}{3p-2}})$ times. Second, at epoch n with historical payoffs, MENU conducts LSEs by $O(\log(T))$ times, each of which is based on $\{x_i\}_{i=1}^n$, while MoM conducts LSE by one time based on intermediate payoffs calculated via median of means of observed payoffs. Third, MENU adopts median of means of LSEs, while MoM adopts median of means of the observed payoffs. Intuitively, the execution of multiple LSEs will lead to the improved regret of MENU. With a better trade-off between k and N in Figure 5.1(a), we derive an improved upper bound of regret in Theorem 5.2.

In light of Figure 5.1(a), we develop algorithmic procedures in Algorithm 5.1 for MENU. We notice that, in order to guarantee the median of means of LSEs not far away from the true underlying parameter with high probability, we construct the confidence interval in Line 10 of Algorithm 5.1. Now we have the following theorem for the regret upper bound of MENU.

Theorem 5.2. *Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \leq D$, $\|\theta_*\|_2 \leq S$, $|x_t^\top \theta_*| \leq L$ and $\mathbb{E}[|\eta_t|^p | \mathcal{F}_{t-1}] \leq c$. Then, with probability at least $1-\delta$, for every $T \geq 256 + 24\log(e/\delta)$, the regret of the MENU*

algorithm satisfies

$$\mathbf{R}(\mathit{MENU}, T)$$
$$\leq 6\left((9dc)^{\frac{1}{p}} + \lambda^{\frac{1}{2}}S + L\right)T^{\frac{1}{p}}\left(24\log\left(\frac{eT}{\delta}\right)+1\right)^{\frac{p-1}{p}}\sqrt{2d\log\left(1+\frac{TD^2}{\lambda d}\right)}.$$

The technical challenges in MENU (i.e., Algorithm 5.1) and its proofs are discussed as follows. Based on the common techniques in linear stochastic bandits (Abbasi-Yadkori et al., 2011), to guarantee the instantaneous regret in LinBET, we need to guarantee $\|\theta_* - \hat{\theta}_{n,k^*}\|_{V_n} \leq \beta_n$ with high probability. We attack this issue by guaranteeing $\|\theta_* - \hat{\theta}_{n,j}\|_{V_n} \leq \beta_n/3$ with a probability of 3/4, which could reduce to a problem of bounding a weighted sum of historical noises. Interestingly, by conducting singular value decomposition on X_n (of which rows are $x_1^\top, \cdots, x_n^\top$), we find that 2-norm of the weights is no greater than 1. Then the weighted sum can be bounded by a term as $O\left(n^{\frac{2-p}{2p}}\right)$. With a standard analysis in linear stochastic bandits from the instantaneous regret to the regret, we achieve the above results for MENU. We show the detailed proof of Theorem 5.2 in Section 5.5.

Remark 5.2. *For MENU, we adopt the assumption of heavy-tailed payoffs on central moments, which is required in the basic technique of median of means (Bubeck et al., 2013a). Besides, there exists an implicit mild assumption in Algorithm 5.1 that, at each epoch n, the decision set must contain the selected arm x_n at least k times, which is practical in applications, e.g., online personalized recommendations (Li et al., 2010). The condition of $T \geq 256 + 24\log(e/\delta)$ is required for $T \geq k$. The regret upper bound of MENU is $\widetilde{O}(T^{\frac{1}{p}})$, which implies that finite variances in LinBET are sufficient to achieve $\widetilde{O}(\sqrt{T})$.*

Algorithm 5.2 TOFU

input d, b, p, δ, λ, T, $\{D_t\}_{t=1}^T$
initialization: $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(\mathbf{0}, S)$
for $t = 1, 2, \cdots, T$ **do**
  $b_t = \left(\frac{b}{\log\left(\frac{2T}{\delta}\right)}\right)^{\frac{1}{p-1}} t^{\frac{2-p}{2p}}$
  $(x_t, \tilde{\theta}_t) = \arg\max_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$ ▷ *to select an arm*
  Play x_t and observe a payoff y_t
  $V_t = V_{t-1} + x_t x_t^\top$ and $X_t^\top = [x_1, \cdots, x_t]$
  $[u_1, \cdots, u_d]^\top = V_t^{-1/2} X_t^\top$
  for $i = 1, \cdots, d$ **do**
    $Y_i^\dagger = (y_1 \mathbb{1}_{u_{i,1} y_1 \le b_t}, \cdots, y_t \mathbb{1}_{u_{i,t} y_t \le b_t})$ ▷ *to truncate the payoffs*
  end for
  $\theta_t^\dagger = V_t^{-1/2}(u_1^\top Y_1^\dagger, \cdots, u_d^\top Y_d^\dagger)$
  $\beta_t = 4\sqrt{d} b^{\frac{1}{p}} \left(\log\left(\frac{2dT}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S$
  Update $C_t = \{\theta : \|\theta - \theta_t^\dagger\|_{V_t} \le \beta_t\}$ ▷ *to update the confidence region*
end for

5.4.2 TOFU and Regret

Description of TOFU. We demonstrate the algorithmic procedures of TOFU in Algorithm 5.2. We point out two subtle differences between our TOFU and the algorithm of CRT as follows. In TOFU, to obtain the accurate estimate of θ_*, we need to trim all historical payoffs for each dimension individually. Besides, the truncating operations depend on the historical information of arms. By contrast, in CRT, the historical payoffs are trimmed once, which is controlled only by the number of rounds for playing bandits. Compared to CRT, our TOFU achieves a tighter confidence interval, which can be found from the setting of β_t. Now we have the following theorem for the regret upper bound of the TOFU algorithm.

Theorem 5.3. *Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \le D$,*

$\|\theta_*\|_2 \leq S$, $|x_t^\top \theta_*| \leq L$ *and* $\mathbb{E}[|y_t|^p|\mathcal{F}_{t-1}] \leq b$. *Then, with probability at least* $1-\delta$, *for every* $T \geq 1$, *the regret of the* TOFU *algorithm satisfies*

$$\begin{aligned}&\mathbf{R}(\textit{TOFU}, T)\\ &\leq 2T^{\frac{1}{p}}\left(4\sqrt{d}b^{\frac{1}{p}}\left(\log\left(\frac{2dT}{\delta}\right)\right)^{\frac{p-1}{p}} + \lambda^{\frac{1}{2}}S + L\right)\sqrt{2d\log\left(1+\frac{TD^2}{\lambda d}\right)}.\end{aligned}$$

Similarly to the proof in Theorem 5.2, we can achieve the above results for TOFU. Due to space limitation, we show the detailed proof of Theorem 5.3 in Appendix 5.5.3.

Remark 5.3. *For* TOFU, *we adopt the assumption of heavy-tailed payoffs on raw moments. It is worth pointing out that, when* $\epsilon = 1$, *we have regret upper bound for* TOFU *as* $\widetilde{O}(d\sqrt{T})$, *which implies that we recover the same order of* d *as that under sub-Gaussian assumption (Abbasi-Yadkori et al., 2011). A weakness in* TOFU *is high time complexity, because for each round* TOFU *needs to truncate all historical payoffs. The time complexity might be reasonably reduced by dividing* T *into multiple epochs, each of which contains only one truncation.*

5.5 Proofs of Theorems

In this section, we show the proofs of theorems.

5.5.1 Proof of Theorem 5.1

We prove the lower bound for $d \geq 2$. Assume d is even (when d is odd, similar results can be easily derived by considering the first $d-1$ dimensions). For $D_t \subseteq \mathbb{R}^d$ with $t \in [T]$, we fix the decision set as $D_1 = \cdots = D_T = D_{(d)}$. Then, the fixed decision set is constructed as $D_{(d)} \triangleq \{(x_1, \cdots, x_d) \in \mathbb{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$,

which is a subset of intersection of the cube $[0,1]^d$ and the hyperplane $x_1 + \cdots + x_d = d/2$. We define a set $S_d \triangleq \{(\theta_1, \cdots, \theta_d) : \forall i \in [d/2], (\theta_{2i-1}, \theta_{2i}) \in \{(2\Delta, \Delta), (\Delta, 2\Delta)\}\}$ with $\Delta \in (0, 1/d]$. The payoff functions take values in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ with $\epsilon \in (0,1]$, for every $x \in D_{(d)}$, the expected payoff is $\theta_*^\top x$, where θ_* is the underlying parameter drawn from S_d. To be more specific, we have the payoff function of x as

$$y(x) = \begin{cases} \left(\frac{1}{\Delta}\right)^{\frac{1}{p-1}} & \text{with a probability of } \Delta^{\frac{1}{p-1}}\theta_*^\top x, \\ 0 & \text{with a probability of } 1 - \Delta^{\frac{1}{p-1}}\theta_*^\top x. \end{cases} \tag{5.4}$$

In this setting, the p-th raw moments of payoffs are bounded by d and $|\theta_*^\top x| \leq 1$. We start the proof with the 2-dimensional case in Subsection 5.5.1. Its extension to the general case (i.e., $d > 2$) is provided in Subsection 5.5.1. Though we set a fixed decision set in the proofs, we can easily extend the lower bound here to the setting of time-varying decision sets, as discussed by Dani et al. (2008a).

$d = 2$ Case

Let $\mu_0 = (\Delta, \Delta)$, $\mu_1 = (2\Delta, \Delta)$ and $\mu_2 = (\Delta, 2\Delta)$. The 2-dimensional decision set is $D_{(2)} = \{(x_1, x_2) \in \mathbb{R}_+^2 : x_1 + x_2 = 1\}$. Our payoff functions take values in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$, and for every $x \in D_{(2)}$, the expected payoff is $\theta_*^\top x$, where θ_* is chosen uniformly at random from $\{\mu_1, \mu_2\}$. It is easy to find $\mu_j^\top x = \Delta(1 + x_j)$ which is maximized at $x_j = 1$ for $j \in \{1, 2\}$, and $\mu_0^\top x = \Delta$ for any $x \in D_{(2)}$.

Lemma 5.1. *If θ_* is chosen uniformly at random from $\{\mu_1, \mu_2\}$, and the payoff for each $x \in D_{(2)}$ is in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ with mean $\theta_*^\top x$, then for every algorithm $\mathcal{A}$ and every $T \geq 1$, the regret satisfies*

$$\mathbb{E}[R(\mathcal{A}, T)] \geq \frac{1}{96} T^{\frac{1}{p}}. \tag{5.5}$$

Proof. We consider a deterministic algorithm $\mathcal{A}$ first. Let $q_{x,T} = T(x)/T$, where $T(x)$ denotes the number of pulls of arm x. $\mathcal{Q}_T$ is the empirical distribution of arms with respect to $q_{x,T}$ and X is drawn from $\mathcal{Q}_T$. We let $\mathcal{P}_j$ and $\mathbb{E}_j$ denote, respectively, the probability distribution of X conditional on $\theta_* = \mu_j$ and the expectation conditional on $\theta_* = \mu_j$, where $j \in \{0, 1, 2\}$. Thus, we have $\mathcal{P}_j(X \in \mathcal{E}) = \mathbb{E}_j[\sum_{x\in\mathcal{E}} T(x)]/T$ for any $\mathcal{E} \subseteq D_{(2)}$. At each time step t, $x_t = (x_{t,1}, x_{t,2})$ is selected. We let $y_t^* = \langle x_t^*, \theta_* \rangle$. Hence, for $j \in \{1, 2\}$, we have

$$\begin{aligned}
&\mathbb{E}_j\left[\sum_{t=1}^{T}(y_t^* - y_t(x_t))\right] = \sum_{t=1}^{T}\mathbb{E}_j\left[\Delta(1 - x_{t,j})\right] = T\int_{D_{(2)}}\Delta(1 - x_j)\mathrm{d}\mathcal{P}_j(x) \\
&= T\Delta\left(1 - \int_{D_{(2)}} x_j \mathrm{d}\mathcal{P}_j(x)\right) \\
&= T\Delta\left(1 - \left(\int_{0\le x_j\le\frac{1}{2}} x_j\mathrm{d}\mathcal{P}_j(x) + \int_{\frac{1}{2}<x_j\le 1} x_j\mathrm{d}\mathcal{P}_j(x)\right)\right) \\
&\ge T\Delta\left(1 - \left(\frac{1}{2}\mathcal{P}_j\left(0 \le X_j \le \frac{1}{2}\right) + \mathcal{P}_j\left(\frac{1}{2} < X_j \le 1\right)\right)\right), \qquad (5.6)
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbb{E}[\mathbf{R}(\mathcal{A}, T)] = \mathbb{E}_{\theta_*}\left[\mathbb{E}_j\left[\sum_{t=1}^{T}(y_t^* - y_t(x_t))\right]\right] \\
&\ge T\Delta\left(1 - \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{2}\mathcal{P}_j\left(0 \le X_j \le \frac{1}{2}\right) + \mathcal{P}_j\left(\frac{1}{2} < X_j \le 1\right)\right)\right). \qquad (5.7)
\end{aligned}$$

According to Pinsker's inequality, for any $\mathcal{E} \subseteq D_{(2)}$, we have

$$\mathcal{P}_j(X \in \mathcal{E}) \le \mathcal{P}_0(X \in \mathcal{E}) + \sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j)}, \qquad (5.8)$$

where $\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j)$ denotes the Kullback-Leibler divergence (simply KL divergence). Hence,

$$\begin{aligned}
&\mathbb{E}[\mathbf{R}(\mathcal{A}, T)] \\
&\ge T\Delta\left(1 - \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{2}\mathcal{P}_0\left(0 \le X_j \le \frac{1}{2}\right) + \mathcal{P}_0\left(\frac{1}{2} < X_j \le 1\right) + \frac{3}{2}\sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j)}\right)\right) \\
&= T\Delta\left(\frac{1}{4} - \frac{3}{4}\sum_{j=1}^{2}\sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j)}\right). \qquad (5.9)
\end{aligned}$$

Since $\mathcal{A}$ is deterministic, the sequence of received rewards denoted by $W_T \triangleq (y_1, y_2, \cdots, y_T) \in \{0, (1/\Delta)^{\frac{1}{p-1}}\}^T$ uniquely determines the empirical distribution $\mathcal{Q}_T$ and thus, $\mathcal{Q}_T$ conditional on W_T is the same for any θ_*. We let $\mathcal{P}_j^t$ be the probability distribution of $W_t = (y_1, y_2, \cdots, y_t)$ conditional on $\theta_* = \mu_j$. Based on the chain rule for KL divergence, we have

$$\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j) \leq \mathrm{KL}(\mathcal{P}_0^T, \mathcal{P}_j^T). \tag{5.10}$$

Further, iteratively using the chain rule for KL divergence, we have

$$\begin{aligned}
&\mathrm{KL}(\mathcal{P}_0^T, \mathcal{P}_j^T) \\
&= \mathrm{KL}(\mathcal{P}_0^1, \mathcal{P}_j^1) + \sum_{t=2}^{T} \int_{W_{t-1}} \mathrm{KL}\left(\mathcal{P}_0^t(\cdot|w_{t-1}), \mathcal{P}_j^t(\cdot|w_{t-1})\right) \mathrm{d}\mathcal{P}_0^{t-1}(W_{t-1}) \\
&= \mathrm{KL}(\mathcal{P}_0^1, \mathcal{P}_j^1) + \sum_{t=2}^{T} \int_{x_t \in D_{(2)}} \int_{W_{t-1}|x_{t,j}=x_j} && (5.11) \\
&\mathrm{KL}\left(\Delta^{\frac{p}{p-1}}, \Delta^{\frac{p}{p-1}}(1+x_j)\right) \mathrm{d}\mathcal{P}_0^{t-1}(W_{t-1}|x_{t,j}=x_j) \mathrm{d}\mathcal{P}_0^{t-1}(x_{t,j}=x_j) && (5.12) \\
&\leq 2\Delta^{\frac{p}{p-1}} + && (5.13) \\
&\sum_{t=2}^{T} \int_{x_t \in D_{(2)}} \int_{W_{t-1}|x_{t,j}=x_j} 2\Delta^{\frac{p}{p-1}} \mathrm{d}\mathcal{P}_0^{t-1}(W_{t-1}|x_{t,j}=x_j) \mathrm{d}\mathcal{P}_0^{t-1}(x_{t,j}=x_j) && (5.14) \\
&= 2T\Delta^{\frac{p}{p-1}}, && (5.15)
\end{aligned}$$

where Eq. (5.14) could be derived by setting $\Delta \leq (1/2)^{\frac{p-1}{p}}$. Note that for any $p, q \in (0,1)$, let $\mathcal{P}$ and $\mathcal{Q}$ denote the Bernoulli distribution with parameters a and b respectively. We denote $\mathrm{KL}(\mathcal{P}, \mathcal{Q})$ as $\mathrm{KL}(a, b)$ in Eq. (5.12). Therefore, we have

$$\mathbb{E}[\mathbf{R}(\mathcal{A}, T)] \geq T\Delta \left(\frac{1}{4} - \frac{3}{2}\sqrt{T\Delta^{\frac{p}{p-1}}}\right) \geq \frac{1}{96} T^{\frac{1}{p}}, \tag{5.16}$$

where we set $\Delta = T^{-\frac{p-1}{p}}/12$.

So far we have discussed the case where $\mathcal{A}$ is a deterministic algorithm. When $\mathcal{A}$ is a randomized algorithm, the result is the same. In particular, let $\mathbb{E}_{\mathcal{A}}$ denote the expectation with respect to the randomness of $\mathcal{A}$. Then, we have

$$\mathbb{E}[\mathbf{R}(\mathcal{A}, T)] = \mathbb{E}_{\mathcal{A}}\left[\mathbb{E}_{\theta_*}\left[\mathbb{E}_j\left[\sum_{t=1}^{T}(y_t^* - y_t(x_t))\right]\right]\right]. \tag{5.17}$$

If we fix the realization of the algorithm's randomization, the results of the previous steps for a deterministic algorithm apply and we know that $\mathbb{E}_{\theta_*}\left[\mathbb{E}_i\left[\sum_{t=1}^{T}(y_t^* - y_t(x_t))\right]\right]$ could be lower bounded as before. Hence, $\mathbb{E}[\mathbf{R}(\mathcal{A}, T)]$ is lower bounded as Eq. (5.16). □

General Case ($d > 2$)

Now we suppose $d > 2$ is even. If d is odd, we just take the first $d-1$ dimensions into consideration. Then we consider the contribution to the total expected regret from the choice of (x_{2i-1}, x_{2i}), for all $i \in [d/2]$. We call (x_{2i-1}, x_{2i}) the i-th component of x.

Analogously to the $d = 2$ case, we set $(\theta_{*,2i-1}, \theta_{*,2i}) \in \{\mu_1, \mu_2\}$. The decision region is $D_{(d)} = \{(x_1, \cdots, x_d) \in \mathbb{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$. Then, by following the proof for $d = 2$ case, we could derive the regret due to the i-th component of x as

$$\mathbb{E}\left[\mathbf{R}^{(i)}(\mathcal{A}, T)\right] \geq \frac{1}{96}T^{\frac{1}{p}}, \tag{5.18}$$

where $i \in [d/2]$. Summing over the $d/2$ components of Eq. (5.18) completes the proof for Theorem 5.1.

5.5.2 Proof of Theorem 5.2

To prove Theorem 5.2, we start with proving the following two lemmas. Recall that the algorithm in the chapter is based on least-squares estimate (LSE).

Lemma 5.2 (Confidence Ellipsoid of LSE)**.** *Let $\hat{\theta}_n$ denote the LSE of θ_* with the sequence of decisions $x_1, \cdots, x_n$ and observed payoffs $y_1, \cdots, y_n$. Assume that for all $\tau \in [n]$ and all $x_\tau \in D_\tau \subseteq \mathbb{R}^d$, $\mathbb{E}[|\eta_\tau|^p | \mathcal{F}_{\tau-1}] \leq c$ and $\|\theta_*\|_2 \leq S$. Then $\hat{\theta}_n$ satisfies*

$$\mathbb{P}\left[\|\hat{\theta}_n - \theta_*\|_{V_n} \leq (9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right] \geq \frac{3}{4}, \tag{5.19}$$

where $\lambda > 0$ is a regularization parameter and $V_n = \lambda I_d + \sum_{\tau=1}^n x_\tau x_\tau^\top$.

Proof. The singular value decomposition of X_n is $U\Sigma_n V^\top$, where U is an $n \times d$ matrix with orthonormal columns, V is a $d \times d$ unitary matrix and Σ_n is an $n \times n$ diagonal matrix with non-negative entries. We calculate $V_n = V(\Sigma_n^2 + \lambda I_d)V^\top$ and

$$V_n^{-\frac{1}{2}} X_n^\top = V\left(\Sigma_n^2 + \lambda I_d\right)^{-\frac{1}{2}} \Sigma_n U^\top. \tag{5.20}$$

Let $u_i^\top$ denote the i-th row of $V\left(\Sigma_n^2 + \lambda I_d\right)^{-\frac{1}{2}} \Sigma_n U^\top$, which leads to $\|u_i\|_2 \leq 1$. More importantly, by optimization, we have $\|u_i\|_p \leq n^{\frac{2-p}{2p}}$. By letting $Y_n = (y_1, \cdots, y_n)$, we have

$$\begin{aligned}
\|\hat{\theta}_n - \theta_*\|_{V_n} &= \|V_n^{-1} X_n^\top (Y_n - X_n\theta_*) - \lambda V_n^{-1}\theta_*\|_{V_n} \\
&\leq \|V_n^{-\frac{1}{2}} X_n^\top (Y_n - X_n\theta_*)\|_2 + \lambda \|\theta_*\|_{V_n^{-1}} \\
&\leq \sqrt{\sum_{i=1}^d \left(u_i^\top (Y_n - X_n\theta_*)\right)^2} + \lambda^{\frac{1}{2}} S.
\end{aligned} \tag{5.21}$$

Inspired by Bubeck et al. (2013a); Medina and Yang (2016), we bound the desired probability by using a union bound as

$$\begin{aligned}
&\mathbb{P}\left[\sum_{i=1}^d \left(\sum_{\tau=1}^n u_{i,\tau}\eta_\tau\right)^2 > \gamma^2\right] \\
&\leq \mathbb{P}\left[\exists i, \tau, |u_{i,\tau}\eta_\tau| > \gamma\right] + \mathbb{P}\left[\sum_{i=1}^d \left(\sum_{\tau=1}^n u_{i,\tau}\eta_\tau \mathbb{1}_{|u_{i,\tau}\eta_\tau| \leq \gamma}\right)^2 > \gamma^2\right],
\end{aligned} \tag{5.22}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function. By using a union bound and Markov's inequality, the first term could be bounded as

$$\mathbb{P}\left(\exists i,\tau,|u_{i,\tau}\eta_\tau|>\gamma\right) \leq \sum_{i=1}^{d}\sum_{\tau=1}^{n}\mathbb{P}(|u_{i,\tau}\eta_\tau|>\gamma) \tag{5.23}$$

$$\leq \frac{\sum_{i=1}^{d}\sum_{\tau=1}^{n}\mathbb{E}[|u_{i,\tau}\eta_\tau|^p]}{\gamma^p} \tag{5.24}$$

$$\leq \frac{\sum_{i=1}^{d}\sum_{\tau=1}^{n}|u_{i,\tau}|^{1+\epsilon}c}{\gamma^p} \leq \frac{dcn^{\frac{1-\epsilon}{2}}}{\gamma^p}. \tag{5.25}$$

Based on Markov's inequality, we bound the second term as

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{d}\left(\sum_{\tau=1}^{n}u_{i,\tau}\eta_\tau\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}\right)^2>\gamma^2\right)\\
&\leq \frac{\mathbb{E}\left[\sum_{i=1}^{d}(\sum_{\tau=1}^{n}u_{i,\tau}\eta_\tau\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma})^2\right]}{\gamma^2}\\
&=\sum_{i=1}^{d}\frac{\mathbb{E}\left[\sum_{\tau=1}^{n}(u_{i,\tau}\eta_\tau)^2\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}\right]}{\gamma^2}+\\
&\sum_{i=1}^{d}2\frac{\mathbb{E}\left[\sum_{\tau'>\tau}(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}(u_{i,\tau'}\eta_{\tau'})\mathbb{1}_{|u_{i,\tau'}\eta_{\tau'}|\leq\gamma}\right]}{\gamma^2}\\
&\leq\sum_{i=1}^{d}\frac{\mathbb{E}\left[\sum_{\tau=1}^{n}(u_{i,\tau}\eta_\tau)^2\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}\right]}{\gamma^2}+\\
&\sum_{i=1}^{d}2\frac{\sum_{\tau'>\tau}\mathbb{E}[(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}]\mathbb{E}[(u_{i,\tau'}\eta_{\tau'})\mathbb{1}_{|u_{i,\tau'}\eta_{\tau'}|\leq\gamma}|\mu_{i,\tau}\eta_\tau]}{\gamma^2}\\
&\leq\sum_{i=1}^{d}\left(\frac{\sum_{\tau=1}^{n}|u_{i,\tau}|^pc}{\gamma^p}+\left(\frac{\sum_{\tau=1}^{n}|u_{i,\tau}|^pc}{\gamma^p}\right)^2\right) \qquad (5.26)\\
&\leq\frac{dcn^{\frac{2-p}{2}}}{\gamma^p}+d\left(\frac{n^{\frac{2-p}{2}}c}{\gamma^p}\right)^2. \qquad (5.27)
\end{aligned}$$

Note that Eq. (5.26) uses the fact as follows.

$$\mathbb{E}[(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}|\mathcal{F}_{\tau-1}] = -\mathbb{E}[(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|>\gamma}|\mathcal{F}_{\tau-1}]. \tag{5.28}$$

Finally, setting $\gamma = (9dc)^{\frac{1}{p}}n^{\frac{2-p}{2p}}$ completes the proof. □

Lemma 5.3. *Recall $\hat{\theta}_{n,j}$, $\hat{\theta}_{n,k^*}$ and V_n in* MENU *(i.e., Algorithm 5.1). If there exists a $\gamma > 0$ such that $\mathbb{P}\left[\|\hat{\theta}_{n,j} - \theta_*\|_{V_n} \leq \gamma\right] \geq \frac{3}{4}$ holds for all $j \in [k]$ with $k \geq 1$, then with probability at least $1 - e^{-\frac{k}{24}}$, $\|\hat{\theta}_{n,k^*} - \theta_*\|_{V_n} \leq 3\gamma$.*

Proof. The proof is inspired by Hsu and Sabato (2014). We define $b_j \triangleq \mathbb{1}_{\|\hat{\theta}_{n,j} - \theta_*\|_{V_n} > \gamma}$, $p_j \triangleq \mathbb{P}(b_j = 1)$ and $\mathbb{B}_{V_n}(\theta_*, \gamma) \triangleq \{\theta : \|\theta - \theta_*\|_{V_n} \leq \gamma\}$. We know that $p_j < 1/4$. By Azuma-Hoeffding's inquality, we have

$$\mathbb{P}\left[\sum_{j=1}^{k} b_j \geq \frac{k}{3}\right] < \mathbb{P}\left[\sum_{j=1}^{k} b_j - p_j \geq \frac{k}{12}\right] \leq e^{-\frac{k}{24}}, \tag{5.29}$$

which means that more than $2/3$ of $\{\hat{\theta}_{n,1}, \cdots, \hat{\theta}_{n,k}\}$ are contained in $\mathbb{B}_{V_n}(\theta_*, \gamma)$ (denoting by this event $\mathcal{E}$) with probability at least $1 - e^{-\frac{k}{24}}$. Note that the value $k/3$ in Eq. (5.29) could also be set as other values in $(k/4, k/2)$. Conditional on the event $\mathcal{E}$, by letting r_j be the median of $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} : s \in [k] \backslash j\}$, we have

- If $\hat{\theta}_{n,j} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} \leq 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j \leq 2\gamma$.
- If $\hat{\theta}_{n,j} \notin \mathbb{B}_{V_n}(\theta_*, 3\gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} > 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j > 2\gamma$.

Combining the above two cases completes proof. □

Based on Lemmas 5.2 and 5.3, by setting $k = \lceil 24 \log(eT/\delta) \rceil$, we are ready to have $\|\hat{\theta}_{n,k^*} - \theta_*\|_{V_n} \leq 3\left((9dc)^{\frac{1}{1+\epsilon}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right)$ with probability at least $1 - \delta/T$. The following part of proof is standard (Abbasi-Yadkori et al., 2011; Dani et al., 2008a). We include it for the sake of completeness. By letting $\beta_n = 3\left((9dc)^{\frac{1}{1+\epsilon}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right)$, we can

decompose the instantaneous regret as follows:

$$\begin{aligned} r_n &= \theta_*^\top x_* - \theta_*^\top x_n \\ &\le \tilde{\theta}_n^\top x_n - \theta_*^\top x_n \\ &\le \left(\|\tilde{\theta}_n - \hat{\theta}_{n-1,k^*}\|_{V_{n-1}} + \|\hat{\theta}_{n-1,k^*} - \theta_*\|_{V_{n-1}} \right) \|x_n\|_{V_{n-1}^{-1}} \\ &\le 2\beta_{n-1} \|x_n\|_{V_{n-1}^{-1}}, \end{aligned} \tag{5.30}$$

where we recall that $(x_n, \tilde{\theta}_n)$ is optimistic in MENU. Note that, for $n = 1$, the above inequality also holds with $V_0 = \lambda I_d$. On the other hand, by considering $|x_t^\top \theta_*| \le L$, we always have

$$r_n \le 2L. \tag{5.31}$$

We can get that

$$r_n \le 2 \min\{\beta_{n-1} \|x_n\|_{V_{n-1}^{-1}}, L\} \le 2(\beta_{n-1} + L) \min\{\|x_n\|_{V_{n-1}^{-1}}, 1\}. \tag{5.32}$$

Following Lemma 11 of Abbasi-Yadkori et al. (2011), we know that

$$\begin{aligned} \sum_{n=1}^{N} \min\{\|x_n\|^2_{V_{n-1}^{-1}}, 1\} &\le 2 \sum_{n=1}^{N} \log(1 + \|x_n\|^2_{V_{n-1}^{-1}}) \\ &= 2 \log \left(\frac{\det(V_N)}{\det(V_0)} \right) \\ &\le 2d \log \left(1 + \frac{ND^2}{\lambda d} \right), \end{aligned} \tag{5.33}$$

where N is the number of epochs in MENU. Therefore, the total regret can be upper bounded by

$$\begin{aligned} &\mathbf{R}(\mathsf{MENU}, T) \\ &\le k \sum_{n=1}^{N} r_n \le k \sqrt{N \sum_{n=1}^{N} r_n^2} \\ &\le 2kN^{\frac{1}{2}} (\beta_N + L) \sqrt{\sum_{n=1}^{N} \min\{\|x_n\|^2_{V_{n-1}^{-1}}, 1\}} \\ &\le 6 \left((12dc)^{\frac{1}{p}} + \lambda^{\frac{1}{2}} S + L \right) T^{\frac{1}{p}} \left(24 \log \left(\frac{eT}{\delta} \right) + 1 \right)^{\frac{p-1}{p}} \sqrt{2d \log \left(1 + \frac{TD^2}{\lambda d} \right)}. \end{aligned} \tag{5.34}$$

The condition of $T \geq 256 + 24\log(e/\delta)$ is required for $T \geq k$, which completes the proof.

5.5.3 Proof of Theorem 5.3

Lemma 5.4. *With the sequence of decisions $x_1, \cdots, x_t$, the truncated payoffs $\{Y_i^\dagger\}_{i=1}^d$ and the parameter estimate $\theta_t^\dagger$ are defined in* TOFU *(i.e., Algorithm 5.2). Assume that for all $\tau \in [t]$ and all $x_\tau \in D_\tau \subseteq \mathbb{R}^d$, $\mathbb{E}[|y_\tau|^p|\mathcal{F}_{\tau-1}] \leq b$ and $\|\theta_*\|_2 \leq S$. With probability at least $1-\delta$, we have*

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leq 4\sqrt{d}b^{\frac{1}{p}}\left(\log\left(\frac{2d}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}}S, \tag{5.35}$$

where $\lambda > 0$ is a regularization parameter and $V_t = \lambda I_d + \sum_{\tau=1}^t x_\tau x_\tau^\top$.

Proof. Similarly to Eq. (5.21), we have

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leq \sqrt{\sum_{i=1}^d \left(u_i^\top (Y_i^\dagger - X_t\theta_*)\right)^2} + \lambda^{\frac{1}{2}}S. \tag{5.36}$$

We let $y_\tau^\dagger$ denote $Y_{i,\tau}^\dagger$ for notation simplicity as the following proof holds

for all $i \in [d]$. Then with probability at least $1 - \delta/d$, we have

$$
\begin{aligned}
& u_i^\top \left(Y_i^\dagger - X_t \theta_*\right) && (5.37)\\
&= \sum_{\tau=1}^{t} u_{i,\tau} \left(y_\tau^\dagger - \mathbb{E}[y_\tau | \mathcal{F}_{\tau-1}]\right) && (5.38)\\
&= \sum_{\tau=1}^{t} u_{i,\tau} \left(y_\tau^\dagger - \mathbb{E}\left[y_\tau^\dagger | \mathcal{F}_{\tau-1}\right] - \mathbb{E}\left[y_\tau \mathbb{1}_{|u_{i,\tau} y_\tau| > b_t} | \mathcal{F}_{\tau-1}\right]\right) \\
&\leq \left| \sum_{\tau=1}^{t} u_{i,\tau} (y_\tau^\dagger - \mathbb{E}[y_\tau^\dagger | \mathcal{F}_{\tau-1}]) \right| + \left| \sum_{\tau=1}^{t} u_{i,\tau} \mathbb{E}[y_\tau \mathbb{1}_{|u_{i,\tau} y_\tau| > b_t} | \mathcal{F}_{\tau-1}] \right| \\
&\leq \left| 2b_t \log\left(\frac{2d}{\delta}\right) + \frac{1}{2b_t} \sum_{\tau=1}^{t} \mathbb{E}\left[u_{i,\tau}^2 \left(y_\tau^\dagger - \mathbb{E}\left[y_\tau^\dagger | \mathcal{F}_{\tau-1}\right]\right)^2 | \mathcal{F}_{\tau-1}\right] \right| \\
&+ \left| \sum_{\tau=1}^{t} \mathbb{E}[u_{i,\tau} y_\tau \mathbb{1}_{|u_{i,\tau} y_\tau| > b_t} | \mathcal{F}_{\tau-1}] \right| && (5.39)\\
&\leq 2b_t \log\left(\frac{2d}{\delta}\right) + \frac{\sum_{\tau=1}^{t} |u_{i,\tau}|^p b}{2b_t^{p-1}} + \frac{\sum_{\tau=1}^{t} |u_{i,\tau}|^p b}{b_t^{p-1}} \\
&\leq 4b^{\frac{1}{p}} \left(\log\left(\frac{2d}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}}, && (5.40)
\end{aligned}
$$

where Eq. (5.39) is obtained by applying Bernstein's inequality for martingales (Seldin et al., 2012) and Eq. (5.40) is obtained by the fact that $\|u_i\|_p \leq t^{\frac{2-p}{2p}}$ and by setting $b_t = (b/\log(2d/\delta))^{\frac{1}{p}} t^{\frac{2-p}{2p}}$. Combining Eq. (5.36) and Eq. (5.40) completes the proof. □

With similar procedures to the proof of Theorem 5.2, we have the regret of TOFU as follows.

$$
\begin{aligned}
& \mathbf{R}(\mathsf{TOFU}, T) \\
& \leq 2T^{\frac{1}{p}} \left(4\sqrt{d} b^{\frac{1}{p}} \left(\log\left(\frac{2dT}{\delta}\right)\right)^{\frac{p-1}{p}} + \lambda^{\frac{1}{2}} S + L\right) \sqrt{2d \log\left(1 + \frac{TD^2}{\lambda d}\right)}, && (5.41)
\end{aligned}
$$

which completes the proof.

Table 5.1: Statistics of synthetic datasets in experiments. For Student's t-distribution, ν denotes the degree of freedom, l_p denotes the location, s_p denotes the scale. For Pareto distribution, α denotes the shape and s_m denotes the scale. NA denotes not available.

dataset	D_t {#arms, #dimensions}	distribution {parameters}	$\{p, b, c\}$	mean of the optimal arm
S1	{20,10}	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{2.00, NA, 3.00}	4.00
S2	{100,20}	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{2.00, NA, 3.00}	7.40
S3	{20,10}	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{1.50, 7.72, NA}	3.10
S4	{100,20}	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{1.50, 54.37, NA}	11.39

5.6 Experiments

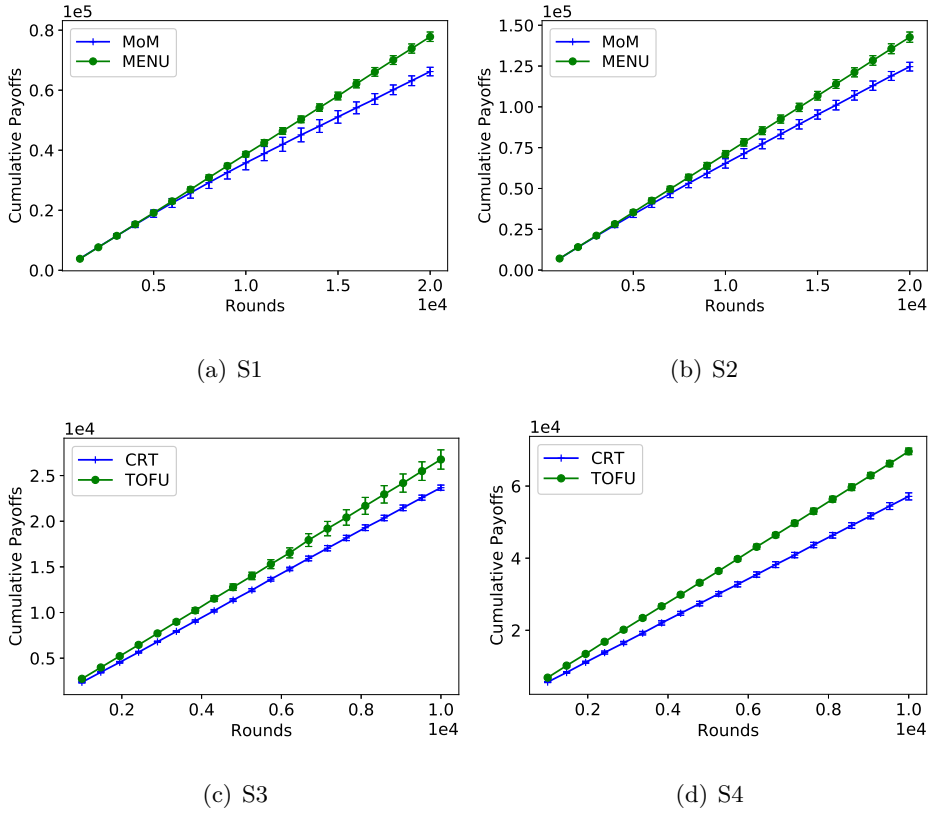
In this section, we conduct experiments based on synthetic datasets to evaluate the performance of our proposed bandit algorithms: MENU and TOFU. For comparisons, we adopt two baselines: MoM and CRT proposed by Medina and Yang (2016). We run multiple independent repetitions for each dataset in a personal computer under Windows 7 with Intel CPU@3.70GHz and 16GB memory.

5.6.1 Datasets and Setting

To show effectiveness of bandit algorithms, we will demonstrate cumulative payoffs with respect to number of rounds for playing bandits over a fixed finite-arm decision set. For verifications, we adopt four synthetic datasets (named as S1–S4) in the experiments, of which statistics are shown in Table 5.1. The experiments on heavy tails require p, b or p, c to be known, which corresponds to the assumptions of Theorem 5.2 or Theorem 5.3. According to the required information, we can apply MENU or TOFU into practical applications. We adopt Student's t and Pareto distributions because they are common in practice. For Student's t-distributions, we easily estimate c, while for Pareto distributions, we easily estimate b. Besides, we can choose different parameters (e.g., larger values) in the distributions, and recalculate the parameters of b and c.

For S1 and S2, which contain different numbers of arms and different dimensions for the contextual information, we adopt standard Student's t-distribution to generate heavy-tailed noises. For the chosen arm $x_t \in D_t$, the expected payoff is $x_t^\top \theta_*$, and the observed payoff is added a noise generated from a standard Student's t-distribution. We generate each dimension of contextual information for an arm, as well as the underlying parameter, from a uniform distribution over $[0, 1]$. The standard Student's t-distribution implies that the bound for the second central moment of S1 and S2 is 3.

For S3 and S4, we adopt Pareto distribution, where the shape parameter is set as $\alpha = 2$. We know $x_t^\top \theta_* = \alpha s_m/(\alpha - 1)$ implying $s_m = x_t^\top \theta_*/2$. Then, we set $p = 1.5$ leading to the bound of raw moment as $\mathbb{E}\left[|y_t|^{1.5}\right] = \alpha s_m^{1.5}/(\alpha - 1.5) = 4 s_m^{1.5}$. We take the maximum of $4 s_m^{1.5}$ among all arms as the bound of the 1.5-th raw moment. We



(a) S1 (b) S2

(c) S3 (d) S4

Figure 5.2: Comparison of cumulative payoffs for synthetic datasets S1-S4 with four algorithms.

generate arms and the parameter similar to S1 and S2.

In figures, we show the average of cumulative payoffs with time evolution over ten independent repetitions for each dataset, and show error bars of a standard variance for comparing the robustness of algorithms. For S1 and S2, we run MENU and MoM and set $T = 2 \times 10^4$. For S3 and S4, we run TOFU and CRT and set $T = 1 \times 10^4$. For all algorithms, we set $\lambda = 1.0$, and $\delta = 0.1$.

5.6.2 Results and Discussions

We show experimental results in Figure 5.2. From the figure, we clearly find that our proposed two algorithms outperform MoM and CRT, which is consistent with the theoretical results in Theorems 5.2 and 5.3. We also evaluate our algorithms with other synthetic datasets, as well as different λ and δ, and observe similar superiority of MENU and TOFU. Finally, for further comparison on regret, complexity and storage of four algorithms, we list the results shown in Table 5.2.

Table 5.2: Comparison on regret, complexity and storage of four algorithms.

algorithm	MoM	MENU	CRT	TOFU
regret	$\widetilde{O}(T^{\frac{2p-1}{3p-2}})$	$\widetilde{O}(T^{\frac{1}{p}})$	$\widetilde{O}(T^{\frac{1}{2}+\frac{1}{2p}})$	$\widetilde{O}(T^{\frac{1}{p}})$
complexity	$O(T)$	$O(T\log T)$	$O(T)$	$O(T^2)$
storage	$O(1)$	$O(\log T)$	$O(1)$	$O(T)$

5.7 Conclusion

We have studied the problem of LinBET, where stochastic payoffs are characterized by finite p-th moments with $p \in (1, 2]$. We broke the traditional assumption of sub-Gaussian noises in payoffs of bandits, and derived theoretical guarantees based on the prior information of bounds on finite moments. We rigorously analyzed the lower bound of LinBET, and developed two novel bandit algorithms with regret upper bounds matching the lower bound up to polylogarithmic factors. Two novel algorithms were developed based on median of means and truncation. In the sense of polynomial dependence on T, we provided optimal algo-

rithms for the problem of LinBET, and thus solved an open problem, which has been pointed out by Medina and Yang (2016). Finally, our proposed algorithms have been evaluated based on synthetic datasets, and outperformed the state-of-the-art results. Since both algorithms in this chapter require a priori knowledge of p, future directions in this line of research include automatic learning of LinBET without information of distributional moments, and evaluation of our proposed algorithms in real-world scenarios.

Chapter 6

Nonlinear Stochastic Bandits

The decision set in MAB can be generalized into a convex set, and the payoff function can be generalized into a non-linear function. In this chapter, we investigate the problem of learning on non-linear stochastic bandits. Given a convex decision set, we consider the stochastic bandits under convex and non-convex payoff functions. Due to the elegant structure of convex functions, we mainly focus on pure exploration of stochastic bandits with convex functions, and extend the results into the scenario of non-convex functions.

We name the problem of pure exploration of stochastic bandits with convex functions as Stochastic Bandit Convex Optimization (SBCO). The problem of SBCO, which is also known as stochastic zeroth-order optimization, has been extensively studied in the literature. It is worth mentioning that stochastic bandit optimization is an LSB problem with a closed compact domain. In stochastic bandit optimization, we will consider two settings: convex objective functions and non-convex objective functions. In particular, we first focus on the study on the convex case and then extend the results to the non-convex case. The goal of this chapter is to develop algorithms with faster convergence rates for

LSB problems in a closed compact domain, where the convergence rate means the number of iterations to train a model.

In particular, we propose a generic approach for accelerating the convergence of existing algorithms to solve the problem of stochastic zeroth-order convex optimization (SBCO). Standard techniques for accelerating the convergence of stochastic zeroth-order algorithms are by exploring multiple functional evaluations (e.g., two-point evaluation), or by exploiting global conditions of the problem (e.g., smoothness and strong convexity). Nevertheless, these classic acceleration techniques are necessarily restricting the applicability of newly developed algorithms. The key of our proposed generic approach is to explore a local growth condition (or called local error bound condition) of the objective function in SBCO. The benefits of the proposed acceleration technique are: (i) it is applicable to both settings with one-point evaluation and two-point evaluation; (ii) it does not necessarily require strong convexity or smoothness condition of the objective function; (iii) it yields an improvement on convergence for a broad family of problems. Empirical studies in various settings demonstrate the effectiveness of the proposed acceleration approach.

6.1 Introduction

We focus on the case of Stochastic Bandit Convex Optimization (SBCO). Then, we extend the results of SBCO to non-convex functions learning. The problem of SBCO has been extensively studied in the literature due to its attractiveness in applications where explicit gradient calculations may be computationally infeasible, expensive, or impossible. However, stochastic zeroth-order algorithms are notoriously slower than stochastic first-order algorithms due to an unavoidable dependence of their

iteration complexities on the dimensionality of the problem.

We consider the following problem of SBCO:

$$\min_{x\in\Omega} f(x) \triangleq \mathbb{E}_{\xi}[f(x;\xi)], \tag{6.1}$$

where $\Omega \subseteq \mathbb{R}^d$ is a closed compact convex set, $f(x;\xi)$ is a stochastic convex function depending on random noise ξ. This problem has broad applications in computer science and engineering. For example, many practical problems in machine learning can be cast into a stochastic convex optimization, where ξ denotes a random data point and x denotes the parameter of a prediction model. A standard approach for solving the problem of Eq. (6.1) is to adopt the stochastic (sub)gradient of $f(x;\xi)$ (Nemirovski et al., 2009). However, there exist situations where the first-order gradient information is computationally infeasible, expensive, or impossible, while the zeroth-order functional information can be easily obtained. For example, in online auctions and advertisement selections, only function values are revealed as feedbacks for algorithms (Wibisono et al., 2012). In stochastic structured predictions, explicit differentiations may be difficult to perform while the functional evaluations of predicted structures are easily obtained (Sokolov et al., 2016). The optimization problem of Eq. (6.1) in such situations is referred to SBCO.

A key concern in the development of iterative stochastic zeroth-oder algorithms for solving Eq. (6.1) is the order of the necessary number of functional evaluations in the form of $f(x;\xi)$, which is termed as sample complexity or iteration complexity. Flaxman et al. (2005) should be the first work related to SBCO. They studied a closely related setting namely online bandit convex optimization where only One-Point Evaluation (OPE) is available for the cost function at each iteration. Applied to the stochastic setting, their algorithm suffers from an iteration

complexity of $O(d^2/\epsilon^4)$ for finding an ϵ-optimal solution $\widehat{x}$ such that $\mathbb{E}[f(\widehat{x}) - \min_{x\in\Omega} f(x)] \leq \epsilon$. Since then, there have been a number of studies (Agarwal et al., 2010; Duchi et al., 2015; Nesterov and Spokoiny, 2017; Shamir, 2013, 2017) trying to improve the iteration complexity of Flaxman et al. (2005) in online bandit setting or in stochastic optimization setting. A useful technique to accelerate the convergence of SBCO is by leveraging Two-Point Evaluation (TPE) at each iteration. Another technique is to explore the strong convexity or the smoothness condition of the random function $f(x;\xi)$. Clearly, both techniques impose strong restrictions of their developed algorithms, and thus the applicability of the resultant algorithms is limited.

The goal of this chapter is to design a generic approach for accelerating existing SBCO algorithms which is applicable to both settings with OPE and TPE, and to cases even without smoothness and strong convexity assumptions of the objective function. A novel contribution is to explore a generic local growth condition (or called local error bound condition) of the objective function, which specifies how fast the objective function grows in a local neighborhood of optimal solutions. In particular, we propose a generic algorithmic framework for accelerating existing SBCO algorithms in various settings by leveraging the local error bound condition. This is accomplished by a novel synthesis of existing SBCO algorithms and a multi-stage adaptive technique, which consists of three components: using a multi-stage scheme with each stage running existing algorithms, warm starting each stage using the solution from previous stage, and adaptively changing the algorithm's parameters after each stage (e.g., step size, the smoothing parameter). Depending on the Local Error Bound (LEB) condition, the improvement over existing results is up to a factor of $1/\epsilon^2$. Em-

pirical studies in various settings demonstrate the effectiveness of the proposed acceleration approach.

6.2 Related Work and Our Results

A quick comparison between our obtained upper bounds of iteration complexities under different settings and previous upper bounds is shown in Table 6.1. Zeroth-order convex optimization has been studied in two closely related paradigms, namely online bandit optimization and stochastic optimization. Using the standard online-to-batch conversion (Shalev-Shwartz et al., 2012), the regret bounds for online bandit optimization can easily be translated into convergence results for stochastic optimization. Hence, we focus on the discussion of results for stochastic zeroth-order convex optimization. However, it is worth mentioning that related chapters studied for online optimization may contain more results for adversarial setting, which is beyond the scope of this thesis.

In Flaxman et al. (2005), the authors developed the first method for online bandit convex optimization, which updates the solutions based on the functional evaluation at a single point. Central to the algorithm and the analysis is a noisy gradient estimator, which is proved to be an unbiased gradient estimator of a smoothed function. By using the analysis of online gradient method, as well as the properties of the noisy gradient estimator and the smoothed function, they established a regret bound of $O(d^{1/2}T^{3/4})$ for Lipschitz continuous cost functions.

Table 6.1: A comparison between our results and existing works for SBCO in the setting of OPE. LC: Lipschitz Continuous, SC: Strong Convexity, SM: SMoothness, and LEB: Local Error Bound.

algorithm	assumption	iteration complexity	high probability or expectation
(Flaxman et al., 2005)	LC	$O\left(\frac{d^2}{\epsilon^4}\right)$	expectation
(Agarwal et al., 2010)	LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon^3}\right)$	expectation
	LC + SC + SM	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	expectation
our work	LC + LEB	$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(2-\theta)}}\right)$, $\theta \in (0, \frac{1}{2}]$	expectation
		$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(2-\theta)}}\right)$, $\theta \in (0, 1]$	high probability
our work	LC + LEB + SM	$\widetilde{O}\left(\frac{d^2}{\epsilon^{3-2\theta}}\right)$, $\theta \in (0, \frac{1}{2}]$	expectation
		$\widetilde{O}\left(\frac{d^2}{\epsilon^{3-2\theta}}\right)$, $\theta \in (0, 1]$	high probability

Table 6.2: A comparison between our results and existing works for SBCO in the setting of TPE. LC: Lipschitz Continuous, SC: Strong Convexity, SM: SMoothness, and LEB: Local Error Bound.

algorithm	assumption	iteration complexity	high probability or expectation
(Agarwal et al., 2010)	LC	$O\left(\frac{d^2}{\epsilon^2}\right)$	high probability
	LC + SC	$\widetilde{O}\left(\frac{d^2}{\epsilon}\right)$	high probability
(Nesterov and Spokoiny, 2017)	LC	$\widetilde{O}\left(\frac{d^2}{\epsilon^2}\right)$	expectation
	LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	expectation
(Duchi et al., 2015)	LC	$\widetilde{O}\left(\frac{d\log d}{\epsilon^2}\right)$	expectation
	LC + SM	$O\left(\frac{d}{\epsilon^2}\right)$	expectation
(Shamir, 2017)	LC	$O\left(\frac{d}{\epsilon^2}\right)$	expectation
our work	LC + LEB	$\widetilde{O}\left(\frac{d^2}{\epsilon^{2(1-\theta)}}\right)$, $\theta \in (0, 1]$	high probability
our work	LC + LEB	$\widetilde{O}\left(\frac{d}{\epsilon^{2(1-\theta)}}\right)$, $\theta \in (0, \frac{1}{2}]$	expectation

Lower bounds for SBCO have been also studied in several works under various settings (Dani et al., 2008b; Duchi et al., 2015; Shamir, 2013). We will show that our proposed algorithm's performance in certain settings matches the existing lower bounds. For example, for stochastic zeroth-order *linear* optimization with OPE, our obtained upper bound of iteration complexity is $\widetilde{O}(d^2/\epsilon^2)$[1], which matches the lower bound by Dani et al. (2008b). In addition, the best upper bound in this chapter for SBCO in the setting with OPE without smoothness assumption is $\widetilde{O}(d^2/\epsilon^2)$, which matches the lower bound by Shamir (2013) up to a logarithmic factor. It is also notable that the best upper bound achieved in this chapter can be as good as $\min(O(d^2\log(1/\epsilon)), \widetilde{O}(d/\epsilon))$. However, we note that our result does not contradict to the lower bound by Duchi et al. (2015) because either their considered random functions do not necessarily have bounded gradients as assumed in this chapter or their considered problem does not satisfy the LEB condition that yields our best result. Finally, we note that the LEB condition has been explored in (stochastic) convex optimization for improving the convergence of first-order methods (Xu et al., 2017b; Yang and Lin, 2015). To the best of our knowledge, this is the first work that leverages the LEB condition for improving the convergence of SBCO.

6.3 Notations and Preliminaries

In this section, we present some notations and preliminaries for SBCO. Let the ℓ-norm of a vector x be $\|x\|_\ell$ (where $\ell \geq 1$). The inner product of two vectors x, y is denoted by $x^\top y = \langle x, y\rangle$. The notation of $\mathbb{B}(x, r)$ denotes a Euclidean ball centered at x with radius $r > 0$. The ceiling integer of a real number r is $\lceil r \rceil$.

[1] We omit a poly-logarithmic factor for $\widetilde{O}(\cdot)$.

Let $\partial f(x)$ and $\nabla f(x)$ denote, respectively, the subgradient of a non-smooth function and the gradient of a smooth function. $f(x)$ is G-Lipschitz continuous if $\exists G > 0$ such that $|f(x)-f(y)| \leq G\|x-y\|_2$, $\forall x, y \in \Omega$, i.e., $\|\partial f(x)\|_2 \leq G, \forall x \in \Omega$. $f(x)$ is L-smooth if it is differentiable and has L-Lipschitz-continuous gradient, i.e., $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2$, $\forall x, y \in \Omega$. $f(x)$ is convex if $f(x) \geq f(y) + \langle \partial f(y), x-y\rangle$, $\forall x, y \in \Omega$. $f(x)$ is σ-strongly convex if $f(x) \geq f(y) + \langle \partial f(y), x-y\rangle + \sigma\|x-y\|_2^2/2$, $\forall x, y \in \Omega$ and $\sigma \geq 0$.

Let $u \sim \mathbb{B}(\mathbf{0}, 1)$ denote a noise vector uniformly sampled from a unit sphere, and $u \sim \mathcal{N}(0, 1)$ denote a noise vector sampled from a standard Gaussian distribution. Given a noise vector u, let $\hat{f}(x;\xi) \triangleq \mathbb{E}_u[f(x+\delta u;\xi)]$ denote a smoothed function with smoothing parameter $\delta > 0$ and $\hat{f}(x) \triangleq \mathbb{E}_{u;\xi}[f(x+\delta u;\xi)]$. Let Ω_* denote the optimal solution set for Eq. (6.1), and $f_* \triangleq \min_{x\in\Omega} f(x)$. In the sequel, we will make the following assumption.

Assumption 6.1. *Assume that there exist $x_0 \in \Omega$ and $\epsilon_0 > 0$ such that $f(x_0) - \min_{x\in\Omega} f(x) \leq \epsilon_0$. For any $\delta \in (0, +\infty)$, there exists $B > 0$ such that $|f(x+\delta u;\xi)| \leq B$ for any $x \in \Omega$ and ξ, where $u \sim \mathbb{B}(\mathbf{0}, 1)$.*

6.3.1 Noisy Gradient Estimators

The noisy gradient estimator in the setting with OPE proposed by Flaxman et al. (2005) is given as:

$$g_t^{\mathrm{f}} = \frac{d}{\delta} f(x_t + \delta u_t; \xi_t) u_t, \tag{6.2}$$

where $u_t \sim \mathbb{B}(\mathbf{0}, 1)$ and $\delta > 0$. The properties of g_t^{f} and $\hat{f}(x;\xi)$ are stated below.

Lemma 6.1 ((Flaxman et al., 2005)). *Given $u \sim \mathbb{B}(\mathbf{0}, 1)$, we have $\mathbb{E}_u[g_t^f] = \nabla \hat{f}(x_t;\xi_t)$, and $\|g_t^f\|_2 \leq dB/\delta$. If $f(x;\xi)$ is G-Lipschitz con-*

tinuous, we have $|f(x;\xi) - \hat{f}(x;\xi)| \leq G\delta$. *If* $f(x;\xi)$ *is L-smooth, we have* $|f(x;\xi) - \hat{f}(x;\xi)| \leq L\delta^2/2$.

For the setting with TPE, there are different gradient estimators used in previous studies. For example, Agarwal et al. (2010); Shamir (2017) used the following noisy gradient estimator with $u_t \sim \mathbb{B}(\mathbf{0}, 1)$:

$$g_t^{\mathrm{a}} = \frac{d}{2\delta}\Big(f(x_t + \delta u_t; \xi_t) - f(x_t - \delta u_t; \xi_t)\Big)u_t. \tag{6.3}$$

Duchi et al. (2015); Nesterov and Spokoiny (2017) considered the following noisy gradient estimator for TPE with $u_t \sim \mathcal{N}(0, 1)$:

$$g_t^{\mathrm{n}} = \frac{1}{\delta}(f(x_t + \delta u_t; \xi_t) - f(x_t; \xi_t))u_t. \tag{6.4}$$

The properties of estimators in Eqs. (6.3) and (6.4) are summarized as below.

Lemma 6.2. *(Agarwal et al., 2010; Shamir, 2017) Given* $u \sim \mathbb{B}(\mathbf{0}, 1)$, *we have* $\mathbb{E}_u[g_t^a] = \nabla\hat{f}(x_t; \xi_t)$. *If* $f(x;\xi)$ *is G-Lipschitz continuous, we have* $\|g_t^a\|_2 \leq Gd$, $\mathbb{E}_u[\|g_t^a\|_2^2] \leq db^2G^2C$, *and* $|f(x;\xi) - \hat{f}(x;\xi)| \leq G\delta$, *where* C *is a universal constant and* b *is a constant such that* $(\mathbb{E}[\|u\|_2^4])^{1/4} \leq b$. *If* $f(x;\xi)$ *is L-smooth, we have* $|f(x;\xi) - \hat{f}(x;\xi)| \leq L\delta^2/2$.

Lemma 6.3. *(Nesterov and Spokoiny, 2017) Considering* $u \sim \mathcal{N}(0, 1)$, *we have* $\mathbb{E}_u[g_t^n] = \nabla\hat{f}(x_t; \xi_t)$. *If* $f(x;\xi)$ *is G-Lipschitz continuous, we have* $\mathbb{E}_u[\|g_t^n\|_2^2] \leq G^2(d+4)^2$, *and* $|f(x;\xi_t) - \hat{f}(x;\xi_t)| \leq \delta Gd^{1/2}$. *If* $f(x;\xi)$ *is G-Lipschitz continuous and L-smooth, we have* $\mathbb{E}_u[\|g_t^n\|_2^2] \leq \delta^2(d+6)^3L^2/2 + 2(d+4)G^2$, *and* $|f(x;\xi) - \hat{f}(x;\xi)| \leq \delta^2Ld/2$.

Remark 6.1. *The absolute upper bound of the noisy gradient estimators is needed for high probability analysis and the variance bound of the noisy gradient estimators is useful for expectational convergence analysis.*

The iterative update in the previous studies takes the following form:

$$x_{t+1} = \Pi_\Omega[x_t - \eta g_t], \tag{6.5}$$

where $\eta > 0$ is a step size, g_t is a gradient estimator and Π_Ω denotes the Euclidean projection onto the set Ω. We synthesize the convergence analysis of stochastic optimization in the following proposition, which, combined with properties of different gradient estimators, yields corresponding convergence results in previous studies.

Proposition 6.1. *Considering the update in Eq. (6.5) with an initial point of $x_1 \in \Omega$, for any $x \in \Omega$, we have*

$$\begin{aligned}
&\sum_{t=1}^{T} f(x_t;\xi_t) - f(x;\xi_t) \\
&\leq 2\sum_{t=1}^{T} \sup_{x\in\Omega} |f(x;\xi_t) - \hat{f}(x;\xi_t)| + \sum_{t=1}^{T} g_t^\top (x_t - x) + (\nabla \hat{f}(x_t;\xi_t) - g_t)^\top (x_t - x),
\end{aligned}$$

and $\sum_{t=1}^{T} g_t^\top (x_t - x) \leq \frac{\|x_1 - x\|_2^2}{2\eta} + \sum_{t=1}^{T} \frac{\eta \|g_t\|_2^2}{2}$.

6.3.2 Local Error Bound Condition

Definition 6.1. *A problem of Eq. (6.1) satisfies the LEB condition on a compact set Ω if there exist $\theta \in (0, 1]$ and $c > 0$ such that for any $x \in \Omega$*

$$dist(x, \Omega_*) \leq c(f(x) - \min_{x\in\Omega} f(x))^\theta, \tag{6.6}$$

where $dist(x, \Omega_*) \triangleq \min_{\mathbf{v}\in\Omega_*} \|\mathbf{v} - x\|_2$.

Note that the LEB condition has been studied thoroughly by Bolte et al. (2015); Xu et al. (2017b); Yang and Lin (2015). It is satisfied for a broad family of problems. For example, when $f(x)$ is continuous and semi-algebraic (or sub-analytic), the LEB condition holds on any compact set (Bolte et al., 2015). Below, we consider several instances

of problems that satisfy the LEB condition. More interesting examples in machine learning can be found in Xu et al. (2017b); Yang and Lin (2015).

Example 1: When $f(x;\xi) = x^\top \xi$ is a linear function and Ω is a polyhedral set (e.g., hypercube), then the problem of Eq. (6.1) satisfies the LEB with $\theta = 1$ (Yang and Lin, 2015). These functions are considered in online bandit linear optimization (Dani et al., 2008b). More generally, if $f(x)$ is a polyhedral function and Ω is a polyhedral set, then LEB with $\theta = 1$ holds (Yang and Lin, 2015). For instance, $f(x) = \sum_{i=1}^n |\mathbf{a}_i^\top x - b_i|/n$ and $\Omega = \{\|x\|_1 \leq s\}$.

Example 2: When $f(x)$ is strongly convex, then the LEB condition holds with $\theta = 1/2$ (Xu et al., 2017b).

Example 3: Even when $f(x)$ is not strongly convex, the LEB condition with $\theta = 1/2$ may still hold, such as $f(x) = \sum_{i=1}^n (\mathbf{a}_i^\top x - b_i)^2/n$ and Ω is a polyhedral set.

6.4 Our Approach and Results

In this section, we propose a generic algorithm for accelerating the convergence of SBCO and its main results in various settings. In order to achieve improved high probability convergence results, we need to use the following update to control the last term in Proposition 6.1:

$$x_{t+1} = \Pi_{\mathbb{D}}[x_t - \eta g_t], \tag{6.7}$$

where $\mathbb{D} = \Omega \cap \mathbb{B}(x^1, D)$ with x^1 being a reference point and D being the radius of the ball. The proposed acceleration framework is presented in Algorithm 6.1, which is a multi-stage adaptive scheme consisting of three key components: (i) a multi-stage scheme with each stage running existing updates, (ii) warm starting each stage using the solution from

Algorithm 6.1 A generic approach for accelerating SBCO

initialization x_0, K, η_1, δ_1, D_1
for $k = 1, \cdots, K$ **do**
  $x_k^1 = x_{k-1}$, $\mathbb{D}_k = \Omega \cap \mathbb{B}(x_k^1, D_k)$
  for $\tau = 1, \cdots, t$ **do**
    compute a gradient estimator in light of Eq. (6.2) or Eq. (6.3) or Eq. (6.4)
    compute x_k^τ according to Eq. (6.5) or Eq. (6.7) with a step size η_k, a parameter δ_k, and a domain $\mathbb{D}_k$
  end for
  let $x_k = \sum_{\tau=1}^t x_k^\tau / t$
  update δ_{k+1}, D_{k+1} and η_{k+1}
end for
return x_K

previous stage, and (iii) adaptively changing the parameters after each stage. Next, we present the iteration complexities of Algorithm 6.1 in various settings. Let $\epsilon_k = \epsilon_0 / 2^k$ be a sequence.

Theorem 6.1 (Results for OPE)**.** *Let Algorithm 6.1 run with Eq. (6.2) as the noisy gradient estimator and $K = \lceil \log_2(\epsilon_0/\epsilon) \rceil$ stages. We have the following results.*

- *R-I: if $f(x;\xi)$ is G-Lipschitz continuous, by employing Eq. (6.5) and setting $t = O(d^2/\epsilon^{2(2-\theta)})$, $\delta_k = \epsilon_k/(6G)$, $\eta_k = \epsilon_k^3/(54G^2d^2B^2)$, then Algorithm 6.1 enjoys an iteration complexity of $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$ in expectation for problems satisfy the LEB condition with $\theta \in (0, 1/2]$;*
- *R-II: if $f(x;\xi)$ is G-Lipschitz continuous and L-smooth, by employing Eq. (6.5) and setting $t = O(d^2/\epsilon^{3-2\theta})$, $\delta_k = \sqrt{\epsilon_k}/(\sqrt{3L})$,*

$\eta_k = 2\epsilon_k^2/(9Ld^2B^2)$, *then Algorithm 6.1 enjoys an iteration complexity of* $\widetilde{O}\Big(d^2/\epsilon^{3-2\theta}\Big)$ *in expectation for problems satisfy the LEB condition with* $\theta \in (0, 1/2]$;

- *R-III: if* $f(x;\xi)$ *is G-Lipschitz continuous, by employing Eq. (6.7) and setting* δ_k, η_k *similarly as in R-I and* $t = \widetilde{O}(d^2/\epsilon^{2(2-\theta)}), D_k = O(\epsilon_{k-1}^{\theta})$, *then Algorithm 6.1 enjoys an iteration complexity of* $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$ *with high probability* $1-p$, *where we set* $p \in (0,1)$, *for problems satisfy the LEB condition with* $\theta \in (0, 1]$;

- *R-IV: if* $f(x;\xi)$ *is G-Lipschitz continuous and L-smooth, by employing Eq. (6.7) and setting* δ_k, η_k *similarly as in R-II and* $t = \widetilde{O}(d^2/\epsilon^{3-2\theta}), D_k = O(\epsilon_{k-1}^{\theta})$, *then Algorithm 6.1 enjoys an iteration complexity of* $\widetilde{O}(d^2/\epsilon^{3-2\theta})$ *with high probability* $1-p$, *where we set* $p \in (0,1)$, *for problems satisfy the LEB condition with* $\theta \in (0, 1]$.

Remark 6.2. *For the statement of high probability results, we omit a poly-logarithmic factor of* $\log(K/p)$ *in* t. *In the above results, we assume the fact that* $|f(x_t + \delta u)| \leq B$, *which is mild considering that* $\delta \in O(1/T^{\alpha})$. *Another way is to assume* $|f(x)| \leq B$ *for any* $x \in \Omega$ *and* $\mathbb{B}_r \subseteq \Omega$, *where* $\mathbb{B}_r$ *is a ball centered at the origin with radius* r. *For every iteration, we project the solution into* $(1-\varepsilon)\Omega$. *By assuming that* $\varepsilon = \delta/r$, *we have* $x + \delta u \in \Omega$ *for any* $x \in (1-\varepsilon)\Omega$. *The improvements on iteration complexity still hold. Our iteration complexities by leveraging the LEB condition are better than those in Agarwal et al. (2010); Flaxman et al. (2005). For LEB with* $\theta = 1/2$ *that is weaker than the strong convexity assumption, our iteration complexities match that in Agarwal et al. (2010) for strongly convex functions. For problems with* $f(x;\xi)$ *being a linear function and* Ω *being a polyhedral set, the LEB with* $\theta = 1$ *holds and we achieve an iteration complexity of*

$\widetilde{O}\Big(d^2/\epsilon^2\Big)$ *with high probability, which matches the lower bound in Dani et al. (2008b). Besides, one may get expectational results for* $\theta > 1/2$ *from high probability results R-III and R-IV following the Corollary 3 in (Xu et al., 2016).*

Theorem 6.2 (Results for TPE)**.** *Let Algorithm 6.1 run with* $K = \lceil \log_2(\epsilon_0/\epsilon) \rceil$ *stages. We have the following results.*

- *R-I: if* $f(x;\xi)$ *is G-Lipschitz continuous, by employing the noisy gradient estimator of Eq. (6.3) and the update of Eq. (6.5) and setting* $t = O(d/\epsilon^{2(1-\theta)})$, $\delta_k = \epsilon_k/(6G)$, $\eta_k = 2\epsilon_k/(3db^2G^2C)$ *where* b *and* C *are parameters discussed in Lemma 6.2, then Algorithm 6.1 enjoys an iteration complexity of* $\widetilde{O}(d/\epsilon^{2(1-\theta)})$ *in expectation for problems satisfy the LEB condition with* $\theta \in (0, 1/2]$*;*

- *R-II: if* $f(x;\xi)$ *is G-Lipschitz continuous and L-smooth, by employing the noisy gradient estimator of Eq. (6.4) and the update of Eq. (6.5) and setting* $t = O(d/\epsilon^{2(1-\theta)})$, $\delta_k = \sqrt{\epsilon_k}/(2\sqrt{dL})$, $\eta_k = \min\{\epsilon_k/(4(d+4)G^2), 2d/((d+6)^3L)\}$*, then Algorithm 6.1 enjoys an iteration complexity of* $\widetilde{O}(d/\epsilon^{2(1-\theta)})$ *in expectation for problems satisfy the LEB condition with* $\theta \in (0, 1/2]$*;*

- *R-III: if* $f(x;\xi)$ *is G-Lipschitz continuous, by employing the noisy gradient estimator of Eq. (6.3) and the update of Eq. (6.7) and setting* $\delta_k = \epsilon_k/(8G)$, $\eta_k = \epsilon_k/(2d^2G^2)$, $t = \widetilde{O}(d^2/\epsilon^{2(1-\theta)})$, *and* $D_k = O(\epsilon_{k-1}^{\theta})$*, then Algorithm 6.1 enjoys an iteration complexity of* $\widetilde{O}(d^2/\epsilon^{2(1-\theta)})$ *with high probability* $1-p$*, where we set* $p \in (0, 1)$*, for problems satisfy the LEB condition with* $\theta \in (0, 1]$*.*

Remark 6.3. *It is notable that in the setting with TPE, the smoothness of the random function does not improve the convergence (by comparing R-I and R-II). The reason is that, for R-I, we utilize the refined analysis*

in (Shamir, 2017) to bound the variance of the noisy gradient estimator $\mathbb{E}[\|g_t^a\|_2^2] \leq O(d)$ *(see Lemma 6.2), which is in the same order to that of the noisy gradient estimator* g_t^n *with small enough* δ *as established in (Nesterov and Spokoiny, 2017) (see Lemma 6.3). The expectational results R-I and R-II have better dependence on* d *compared to the high probability result R-III. The reason is that, for high probability analysis, we have to use the absolute bound of* g_t^a*. However, the expectational results R-I and R-II cannot enjoy better dependence on* ϵ *for* $\theta > 1/2$ *as in the high probability result R-III. We notice that one can obtain similar expectational results for* $\theta > 1/2$ *in light of R-III with the same technique in Remark 2.*

Finally, we would like to point out that although the above results require knowing the value of θ in the LEB condition, we can use another level of restarting on top of Algorithm 6.1 and an increasing sequence of t for the outer loop similar to that by Xu et al. (2017b); Yang and Lin (2015), which still enjoy improved iteration complexities compared with previous results. Due to limitation of space, this result and the related proofs are both omitted here.

6.5 Proofs of Theorems

In this section, we present the proofs of Proposition 6.1, Theorem 6.1 and Theorem 6.2.

6.5.1 Proof of Proposition 6.1

Proof. We adopt the update in Eq. (6.5) to find an ϵ-optimal solution for SBCO. After T iterations, we have

$$\begin{aligned}\sum_{t=1}^{T}\hat{f}(x_t;\xi_t)-\hat{f}(x;\xi_t) &\leq \sum_{t=1}^{T}\nabla\hat{f}(x_t;\xi_t)^\top(x_t-x)\\ &\leq \sum_{t=1}^{T}(\nabla\hat{f}(x_t;\xi_t)-g_t)^\top(x_t-x)+\sum_{t=1}^{T}g_t^\top(x_t-x)\\ &\leq \sum_{t=1}^{T}(\nabla\hat{f}(x_t;\xi_t)-g_t)^\top(x_t-x)+\\ &\sum_{t=1}^{T}\frac{\|x-x_t\|_2^2-\|x-x_{t+1}\|_2^2}{2\eta}+\sum_{t=1}^{T}\frac{\eta\|g_t\|_2^2}{2},\end{aligned}$$

where $x\in\Omega$, η is the learning rate, and the last inequality is due to Zinkevich (2003). By taking the upper bound between $f(x;\xi_t)$ and $\hat{f}(x;\xi_t)$, we have

$$\begin{aligned}&\sum_{t=1}^{T}f(x_t;\xi_t)-f(x;\xi_t)\\ &\leq \sum_{t=1}^{T}\hat{f}(x_t;\xi_t)-\hat{f}(x;\xi_t)+2\sum_{t=1}^{T}\sup_{x\in\Omega}|f(x;\xi_t)-\hat{f}(x;\xi_t)|.\end{aligned}$$

Thus, we are ready to have

$$\begin{aligned}&\sum_{t=1}^{T}f(x_t;\xi_t)-f(x;\xi_t)\\ &\leq 2\sum_{t=1}^{T}\sup_{x\in\Omega}|f(x;\xi_t)-\hat{f}(x;\xi_t)|+\frac{\|x_1-x\|_2^2}{2\eta}+\sum_{t=1}^{T}\frac{\eta\|g_t\|_2^2}{2}+\\ &(\nabla\hat{f}(x_t;\xi_t)-g_t)^\top(x_t-x),\end{aligned}$$

where $x\in\Omega$. □

6.5.2 Proof of Theorem 6.1

Proof. We present the proof of the theorem based on different conditions as follows. For expectational results, we adopt the update

in Eq. (6.5). For high probability results, we adopt the update in Eq. (6.7).

i) Proof of R-I: Expectational results when $f(x;\xi)$ is G-Lipschitz continuous.

If $f(x;\xi)$ is G-Lipschitz continuous, based on Proposition 6.1 and Lemma 6.1, we have

$$\sum_{t=1}^{T} f(x_t;\xi_t) - f(x;\xi_t) \leq 2TG\delta + \frac{\eta T d^2 B^2}{2\delta^2} + \frac{\|x_1 - x\|_2^2}{2\eta} + \sum_{t=1}^{T} (\nabla \hat{f}(x_t;\xi_t) - g_t^{\mathrm{f}})^\top (x_t - x).$$

By setting $\hat{x}_T = \sum_{t=1}^{T} x_t/T$ and taking the expectation over randomness in u and ξ, we have

$$\mathbb{E}[f(\hat{x}_T) - f(x)] \leq \frac{\mathbb{E}[\|x_1 - x\|_2^2]}{2\eta T} + \frac{\eta d^2 B^2}{2\delta^2} + 2G\delta.$$

By adopting the generic framework in Algorithm 6.1, for the k-th stage, we have

$$\mathbb{E}[f(x_k) - f(x)] \leq \frac{\mathbb{E}[\|x_{k-1} - x\|_2^2]}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k,$$

where we use t iterations in inner loops of Algorithm 6.1.

For $\theta \in (0, 1/2]$, based on Jensen's inequality, we have $\mathbb{E}[\|x - x_*\|_2^2] \leq c^2 \mathbb{E}[(f(x) - f_*)^{2\theta}] \leq c^2 (\mathbb{E}[f(x) - f_*])^{2\theta}$, with $x_* \in \Omega_*$ and $x \in \Omega$. Note that here we adopt the LEB condition over Ω. Then, we show that $\mathbb{E}[f(x_k) - f_*] \leq \epsilon_k$ holds by induction, where $\epsilon_k = \epsilon_0/2^k$.

If $k = 0$, we clearly have $\mathbb{E}[f(x_0) - f_*] \leq \epsilon_0$. Conditioned on the inequality of $\mathbb{E}[f(x_{k-1}) - f_*] \leq \epsilon_{k-1}$, we will show that $\mathbb{E}[f(x_k) - f_*] \leq \epsilon_k$.

Let $x_{k-1,*} = \arg\min_{\mathbf{v}\in\Omega_*} \|\mathbf{v} - x_{k-1}\|_2$. We have

$$\begin{aligned}\mathbb{E}[f(x_k) - f(x_{k-1,*})] &\leq \frac{\mathbb{E}[\|x_{k-1} - x_{k-1,*}\|_2^2]}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k \\ &\leq \frac{c(\mathbb{E}[f(x_{k-1}) - f(x_{k-1,*})])^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k \\ &\leq \frac{c\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + 2G\delta_k.\end{aligned}$$

To establish $\mathbb{E}[f(x_k) - f_*] \leq \epsilon_k$, we set

$$\begin{aligned}\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{6} &\Rightarrow t \geq \frac{1296 d^2 B^2 G^2 c^2}{\epsilon_{k-1}^{2(2-\theta)}}, \\ \frac{\eta_k d^2 B^2}{2\delta_k^2} \leq \frac{\epsilon_k}{3} &\Rightarrow \eta_k \leq \frac{\epsilon_k^3}{54 G^2 d^2 B^2}, \\ 2G\delta_k \leq \frac{\epsilon_k}{3} &\Rightarrow \delta_k \leq \frac{\epsilon_k}{6G}.\end{aligned}$$

By setting $\epsilon_K = \epsilon_0/2^K = \epsilon$, we have $K = \lceil \log(\epsilon_0/\epsilon) \rceil$. Thus, we have $\mathbb{E}[f(x_K) - f_*] \leq \epsilon_K = \epsilon$. As a result, the total iteration complexity is $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$.

ii) Proof of R-II: Expectational results when $f(x;\xi)$ is G-Lipschitz continuous and L-smooth.

If $f(x;\xi)$ is L-smooth, with Lemma 6.1, we have

$$\mathbb{E}[f(x_k) - f(x_{k-1,*})] \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + L\delta_k^2.$$

To establish $\mathbb{E}[f(x_k) - f_*] \leq \epsilon_k$, we set

$$\begin{aligned}\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{6} &\Rightarrow t \geq \frac{54 d^2 B^2 L c^2}{\epsilon_{k-1}^{3-2\theta}}, \\ \frac{\eta_k d^2 B^2}{2\delta_k^2} \leq \frac{\epsilon_k}{3} &\Rightarrow \eta_k \leq \frac{2\epsilon_k^2}{9 L d^2 B^2}, \\ L\delta_k^2 \leq \frac{\epsilon_k}{3} &\Rightarrow \delta_k \leq \frac{\sqrt{\epsilon_k}}{\sqrt{3L}}.\end{aligned}$$

Thus, with $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, the total iteration complexity is $\widetilde{O}(d^2/\epsilon^{3-2\theta})$.

iii) Proof of R-III: High probability results when $f(x;\xi)$ is G-Lipschitz continuous.

We prove high probability convergence for $\theta \in (0, 1]$. Similar to Proposition 6.1, we can derive

$$\sum_{t=1}^{T} \Big(\hat{f}(x_t) - \hat{f}(x) \Big) \leq \sum_{t=1}^{T} \langle g_t^{\mathrm{f}}, (x_t - x) \rangle + \sum_{t=1}^{T} (\nabla \hat{f}(x_t) - g_t^{\mathrm{f}})^\top (x_t - x).$$

Since $\mathbb{E}_{u,\xi}[g_t] = \nabla \hat{f}(x_t)$ and $\|g_t^{\mathrm{f}}\|_2 \leq dB/\delta$, based on Lemma 14 in Hazan and Kale (2014), we have the following result. Given $x_1 \in \Omega$, we apply T iterations of Eq. (6.2) and the update in Eq. (6.7). For any fixed $x \in \Omega \cap \mathbb{B}(x_1, D)$ and $\tilde{p} \in (0, 1)$, with a probability at least $1 - \tilde{p}$, the following inequality holds

$$\hat{f}(\hat{x}_T) - \hat{f}(x) \leq \frac{\|x_1 - x\|_2^2}{2\eta T} + \frac{\eta d^2 B^2}{2\delta^2} + \frac{4dBD\sqrt{3\log(\frac{1}{p})}}{\sqrt{T}\delta},$$

where $\hat{x}_T = \sum_{t=1}^{T} x_t/T$. Then, we are ready to have

$$f(x_k) - f(x_{k-1,*}) \leq \frac{c^2 \epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + \frac{4dBc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{p})}}{\sqrt{t}\delta_k} + 2G\delta_k,$$

where we use $D_k = c\epsilon_{k-1}^{\theta}$. We can easily establish $f(x_k) - f(x_*) \leq \epsilon_k$ with high probability by induction. By setting $\delta_k = O(\epsilon_k)$, $\eta_k = O(\epsilon_k^3/d^2)$, $t = O(d^2 \log(K/\tilde{p})/\epsilon^{2(2-\theta)})$ and $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, we have the iteration complexity as $\widetilde{O}(d^2/\epsilon^{2(2-\theta)})$ with high probability of $1 - p$, where $\tilde{p} = p/K$.

iv) Proof of R-IV: High probability results when $f(x; \xi)$ is G-Lipschitz continuous and L-smooth.

For smooth objective functions, we have

$$f(x_k) - f(x_{k-1,*}) \leq \frac{c^2 \epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 B^2}{2\delta_k^2} + \frac{4dBc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{p})}}{\sqrt{t}\delta_k} + L\delta_k^2.$$

We establish $f(x_k) - f(x_*) \leq \epsilon_k$ with high probability by induction. By setting $D_k = c\epsilon_{k-1}^{\theta}$, $\delta_k = O(\sqrt{\epsilon_k})$, $\eta_k = O(\epsilon_k^2/d^2)$, $t = O(d^2 \log(K/\tilde{p})/\epsilon^{3-2\theta})$ and $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, we have the iteration complexity as $\widetilde{O}(d^2/\epsilon^{3-2\theta})$ with high probability of $1 - p$, where $\tilde{p} = p/K$. □

6.5.3 Proof of Theorem 6.2

Proof. In the setting with TPE, we present the proofs as follows. Again, for expectational results, we adopt the update in Eq. (6.5). For high probability results, we adopt the update in Eq. (6.7).

i) Proof of R-I: Expectational results when $f(x;\xi)$ is G-Lipschitz continuous.

We consider the noisy gradient estimator of Eq. (6.3). Based on the LEB condition, for $\theta \in (0, 1/2]$, we have $\mathbb{E}[\|x - x_*\|_2^2] \leq c^2\mathbb{E}[(f(x) - f_*)^{2\theta}] \leq c^2(\mathbb{E}[f(x)-f_*])^{2\theta}$. With results in Proposition 6.1 and Lemma 6.2, and the analysis of Theorem 6.1, we have

$$\mathbb{E}[f(x_k) - f(x_{k-1,*})] \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d b^2 G^2 C}{2} + 2G\delta_k,$$

where b and C are parameters discussed in Lemma 6.2.

Similarly, we can easily establish the relationship of $\mathbb{E}[f(x_k) - f_*] \leq \epsilon_k$ by induction. We can set

$$\begin{aligned}
\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{6} &\Rightarrow t \geq \frac{9dG^2b^2Cc^2}{\epsilon_{k-1}^{2(1-\theta)}},\\
\frac{\eta_k d b^2 G^2 C}{2} \leq \frac{\epsilon_k}{3} &\Rightarrow \eta_k \leq \frac{2\epsilon_k}{3db^2G^2C},\\
2G\delta_k \leq \frac{\epsilon_k}{3} &\Rightarrow \delta_k \leq \frac{\epsilon_k}{6G}.
\end{aligned}$$

Then, with $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, the total iteration complexity is $\widetilde{O}(d/\epsilon^{2(1-\theta)})$.

ii) Proof of R-II: Expectational results when $f(x;\xi)$ is G-Lipschitz continuous and L-smooth.

If $f(x;\xi)$ is L-smooth, we adopt the noisy gradient estimator of Eq. (6.4) and the update of Eq. (6.5) to solve SBCO. For $\theta \in (0, 1/2]$, with the results in Lemma 6.3, we are ready to have

$$\mathbb{E}[f(x_k) - f(x_{k-1,*})] \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k}{2}\left(\frac{\delta_k^2(d+6)^3L^2}{2} + 2(d+4)G^2\right) + L\delta_k^2 d,$$

To establish the induction $\mathbb{E}[f(x_k) - f_*] \leq \epsilon_k$, we set

$$\frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_{k-1}}{8} \Rightarrow t \geq \frac{32(d+4)G^2c^2}{\epsilon_{k-1}^{2(1-\theta)}},$$
$$\frac{\eta_k}{2}\left(\frac{\delta_k^2(d+6)^3L^2}{2}\right) \leq \frac{\epsilon_k}{4} \Rightarrow \eta_k \leq \frac{4d}{(d+6)^3L},$$
$$\frac{\eta_k}{2}(2(d+4)G^2) \leq \frac{\epsilon_k}{4} \Rightarrow \eta_k \leq \frac{\epsilon_k}{4(d+4)G^2},$$
$$L\delta_k^2 d \leq \frac{\epsilon_k}{4} \Rightarrow \delta_k \leq \frac{\sqrt{\epsilon_k}}{2\sqrt{dL}}.$$

Here we can set $\eta_k = \min\{\frac{\epsilon_k}{4(d+4)G^2}, \frac{2d}{(d+6)^3L}\}$. Since ϵ_k goes to ϵ, the term $\frac{\epsilon_k}{4(d+4)G^2}$ is dominant in iteration complexity and t is calculated via $\eta_k \leq \frac{\epsilon_k}{4(d+4)G^2}$. Thus, with $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, the total iteration complexity is $\widetilde{O}(d/\epsilon^{2(1-\theta)})$.

iii) Proof of R-III: High probability results when $f(x;\xi)$ is G-Lipschitz continuous.

For high probability analysis, we adopt the noisy gradient estimator of Eq. (6.3) and the update of Eq. (6.7) to solve SBCO. Similar to the analysis of R-III in Theorem 6.1, with high probability at least $1 - \tilde{p}$,

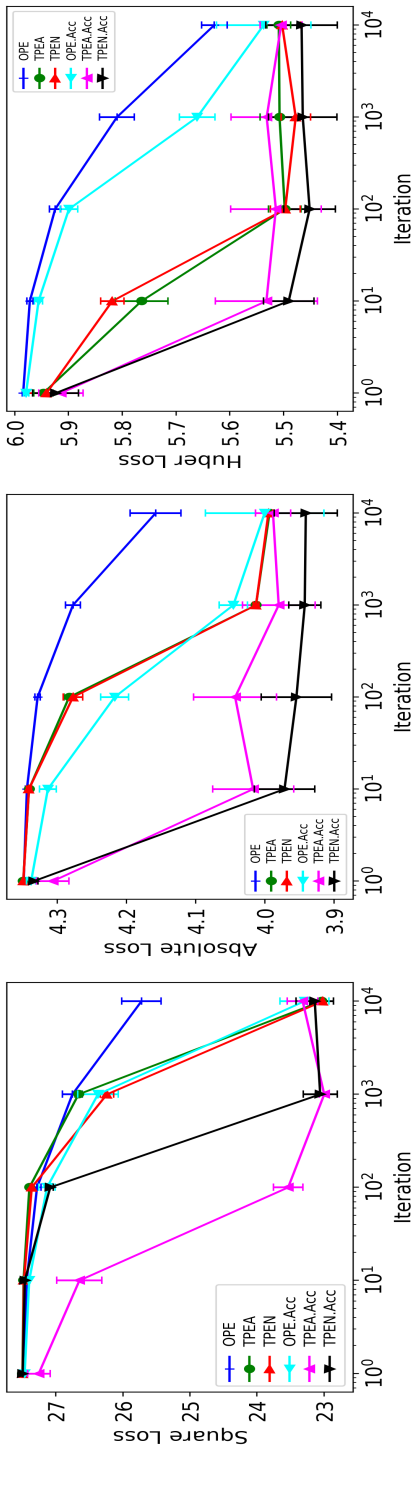
$$f(x_k) - f(x_{k-1,*}) \leq \frac{c^2\epsilon_{k-1}^{2\theta}}{2\eta_k t} + \frac{\eta_k d^2 G^2}{2} + \frac{4dGc\epsilon_{k-1}^{\theta}\sqrt{3\log(\frac{1}{p})}}{\sqrt{t}} + 2G\delta_k,$$

where we set $D_k = c\epsilon_{k-1}^{\theta}$. To establish the induction $f(x_k) - f_* \leq \epsilon_k$, we set $\delta_k = \epsilon_k/(8G)$, $\eta_k = \epsilon_k/(2d^2G^2)$, $t = O(d^2\log(K/\tilde{p})/\epsilon^{2(1-\theta)})$ and $K = \lceil \log(\epsilon_0/\epsilon) \rceil$, where $\tilde{p} = p/K$. As a result, the total iteration complexity is $\widetilde{O}(d^2/\epsilon^{2(1-\theta)})$ with probability of $1 - p$. □

6.6 Experiments

In this section, we conduct experiments on real-world datasets in various settings to demonstrate the superior performance of the proposed acceleration approach in Algorithm 6.1. We run experiments in a personal computer with Intel CPU@3.70GHz and 16GB memory.

To compare the efficiency of the acceleration framework with prior methods, we will show the evolution of function values with respect to the number of iterations. We adopt three baselines: the first is OPE from (Flaxman et al., 2005); the second is TPEA from (Agarwal et al., 2010); and the third is TPEN from (Nesterov and Spokoiny, 2017). We add a term "Acc" to denote our acceleration version for each baseline in experiments. To show experimental results, we run experiments ten times with the same initialization point, and show the average of function values. For the first experiment on real-world datasets, we also show error bars of a standard variance.



(a) $f(x) = \frac{\sum_{i=1}^{N}(w_i^\top x - r_i)^2}{N}$, $\theta = 0.5$,angle=0 (b) $f(x) = \frac{\sum_{i=1}^{N}|w_i^\top x - r_i|}{N}$, $\theta = 1$ (c) $f(x)$ is averaged Huber loss, $\theta = 1$

Figure 6.1: Comparisons of convergence with three objective functions for music recommendation competition data and $T = 10^4$.

6.6.1 Music Recommendation Competition Data

We consider the ensemble learning setting of recommendations as a black-box optimization problem discussed in Lian et al. (2016). In particular, we blend the existing models in Chen et al. (2011) for music recommendation competition in KDD-Cup 2011, which turns out to be a linear regression problem. Since true ratings for the test set are unknown in competition, the feedback is the evaluation of the linear regression prediction of the blended model. Thus, this ensemble learning case is SBCO.

We get predicted ratings of individual models in Chen et al. (2011) for the test set in KDD-Cup 2011, with 237 models and 6,005,940 predictions for each model. In addition to a square loss (Lian et al., 2016), we also consider an absolute loss and a huber loss (Zadorozhnyi et al., 2016) as objective functions. For better demonstrations of convergence rate, we sample 10 models from 237 models with predicted ratings denoted by $w \in \mathbb{R}^{10}$ in ensemble learning, and set the number of training points as $N = 10^5$. The ground truth of sample i is denoted by r_i.

We show the superior convergence of our proposed acceleration approach with different objective functions in Figure 6.1. From the standard variance error bar, we clearly find that our approach stably accelerates the existing SBCO algorithms with order improvements.

6.6.2 Industrial Data on Ceramic Thin Films

We consider industrial data on crystallization of ceramic thin films in Nakamura et al. (2017). The goal for the industrial application on crystallization of ceramic thin films is to determine an optimal setting for the volume of tetraethylene glycol (TEG), temperature (T), and the time of heat to a temperature in time (HTI), which is in fact a SBCO

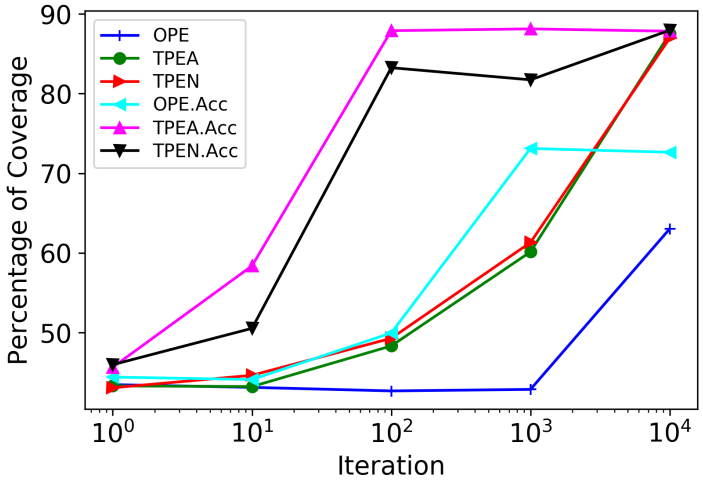


Figure 6.2: Growth of ceramic thin films with $T = 10^4$.

problem. The objective of the experiment is a quadratic function. For more details, please refer to Nakamura et al. (2017); Wang et al. (2017).

By updating the values of TEG, T and HTI, we show the growth of ceramic thin films with the number of iterations in Figure 6.2. The superior performance of the acceleration via Algorithm 6.1 is clear. We also test different intensity of noises, and find that the acceleration is robust. Note that, in ceramic thin films, we solve a concave function and thus the function value increases.

6.7 Conclusion

In this chapter, we have developed a generic acceleration approach to solve the problem of SBCO. We tackled the SBCO problem with the core idea of exploring an LEB condition of objective functions, which is frequently encountered in real applications. The benefits of the proposed acceleration technique are three-fold: wide applicability, weak assumption and improvements on iteration complexity. With LEB con-

dition, the best upper bound here can be $\min(O(d^2 \log(1/\epsilon)), \widetilde{O}(d/\epsilon))$, and the improvement over existing results is up to a factor of $1/\epsilon^2$. Experimental results have shown superior and robust performance of the proposed acceleration approach.

□ **End of chapter.**

Chapter 7

Conclusion and Discussions

In this chapter, we conclude the study of efficient learning in stochastic bandits. First, we give a summary of contributions in this thesis. Then, we list three interesting future directions in bandits.

7.1 Main Contributions

The topic of efficient learning in stochastic bandits lies in the domain of machine learning. We developed efficient algorithms for problems of learning in stochastic bandits. We have demonstrated the effectiveness and efficiency of the algorithms by theoretical analyses and experiments. In particular, we have the following summarizations.

In Chapter 3, we investigated the problem of pure exploration of mean-variance. We have solved three technical challenges by rigorously proving that the error resulting from the mean-variance estimation is sub-gamma. Besides, we developed two efficient algorithms to tackle the problem. With the sub-gamma noises, we derived upper bounds of the probability of error for the proposed algorithms. By conducting a series of experiments on synthetic and real-world datasets, we demonstrated the two algorithms are superior and robust.

In Chapter 4, we broke the assumption of payoffs under sub-Gaussian noises in pure exploration of MAB, and investigated best arm identification of MAB with a general assumption that the p-th moments of stochastic payoffs are bounded, where $p \in (1, +\infty)$. We have technically analyzed tail probabilities of empirical average and truncated empirical average for estimating expected payoffs in sequential decisions. Besides, we proposed two bandit algorithms for pure exploration of MAB with heavy-tailed payoffs. Finally, we derived theoretical guarantees of the proposed bandit algorithms, and demonstrated the effectiveness of bandit algorithms in pure exploration of MAB with heavy-tailed payoffs.

In Chapter 5, we studied the problem of linear stochastic bandits with heavy tails. We rigorously analyzed the lower bound of the problem, and developed two novel bandit algorithms with regret upper bounds matching the lower bound up to polylogarithmic factors. In the sense of polynomial dependence on T, we provided optimal algorithms for the problem. Finally, our proposed algorithms have been evaluated based on synthetic datasets, and outperformed the state-of-the-art results.

In Chapter 6, we developed a generic acceleration approach to solve the problem of stochastic bandit optimization. We tackled the problem with the core idea of exploring an error bound condition of objective functions. With the error bound condition, the best upper bound here can be $\min(O(d^2 \log(1/\epsilon)), \widetilde{O}(d/\epsilon))$, and the improvement over existing results is up to a factor of $1/\epsilon^2$. Experimental results have shown superior and robust performance of the proposed acceleration approach.

7.2 Future Directions

In machine learning, automatically learning is an important branch, especially for sequential decisions in practical scenarios. From the perspective of automatically learning of bandits, where payoffs can be sub-Gaussian or heavy tails, our high-level goal is to develop efficient automatic bandit algorithms with good theoretical guarantees for real applications, e.g., adaptive online recommendations and automatic portfolio selection.

In this thesis, we have fully investigated efficient learning in stochastic bandits. There are many other interesting open problems in bandits. Specifically, we list the following three sub-problems for future directions:

1. A direct extension of this thesis is to investigate the problem of bandits with outliers, where some of the payoffs could change arbitrarily without the constraints p.

2. The state-of-the-art algorithms for heavy-tailed payoffs assuming the boundedness of the moments of order p, i.e., $\mathbb{E}\left[|y_t|^p|\mathcal{F}_{t-1}\right] \leq b$ or $\mathbb{E}\left[|\eta_t|^p|\mathcal{F}_{t-1}\right] \leq c$, require the prior knowledge on p and b or c. In real data, we usually do not have access to this knowledge. Thus, a potential problem we would like to investigate is automatically learning p and b or c.

3. The third potential direction is that bandits can have time-varying tails, e.g., change from sub-Gaussian tails to heavy tails. To derive theoretical results for automatically learning this problem is a big challenge.

Appendix A

List of Publications

[1] **Xiaotian Yu**, Haiqin Yang, Irwin King and Michael R. Lyu. Online non-negative dictionary learning via moment information for sparse poisson coding. In *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 5094–5101, 2016.

[2] **Xiaotian Yu**, Michael R. Lyu and Irwin King. CBRAP: Contextual Bandits with RAndom Projection. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 2859–2866, 2017.

[3] **Xiaotian Yu**, Irwin King and Michael R. Lyu. Risk control of best arm identification in multi-armed bandits via successive rejects. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, pages 1147–1152, 2017.

[4] **Xiaotian Yu**, Irwin King, Michael R. Lyu and Tianbao Yang. A generic approach for accelerating stochastic zeroth-order convex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3040–3046, 2018.

[5] **Xiaotian Yu**, Han Shao, Michael R. Lyu and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Pro-*

ceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI), pages 937–946, 2018.

[6] Han Shao*, **Xiaotian Yu***, Irwin King and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages (to appear), 2018. **Spotlight presentation**.

* denotes equal contributions.

□ **End of chapter.**

Bibliography

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Annual Conference on Neural Information Processing Systems*, pages 2312–2320, 2011.

N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11, 1999.

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40, 2010.

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

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013a.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013b.

N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour,

and O. Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pages 13–p, 2010.

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

J.-Y. Audibert, O. Catoni, et al. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

P. Auer. Using upper confidence bounds for online learning. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 270–279. IEEE, 2000.

P. Auer. Using confidence bounds for exploitation-exploration tradeoffs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

P. Auer and C.-K. Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120, 2016.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the

multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 207–216. IEEE, 2013.

A. Badanidiyuru, J. Langford, and A. Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134, 2014.

J. Bolte, T. P. Nguyen, J. Peypouquet, and B. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *CoRR*, abs/1510.08234, 2015.

S. Boucheron, M. Thomas, et al. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17(51):1–12, 2012.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

S. Bubeck. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2010.

S. Bubeck and A. Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic

and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013a.

S. Bubeck, V. Perchet, and P. Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134, 2013b.

S. Bubeck, T. Wang, and N. Viswanathan. Multiple identifications in multi-armed bandits. In *Proceedings of ICML*, pages 258–265, 2013c.

V. V. Buldygin and Y. V. Kozachenko. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32(6):483–489, 1980.

A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604, 2016.

A. Carpentier and M. Valko. Extreme bandits. In *Proceedings of NIPS*, pages 1089–1097, 2014.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

L. Chen, A. Gupta, J. Li, M. Qiao, and R. Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pages 482–534, 2017a.

L. Chen, J. Li, and M. Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110, 2017b.

P.-L. Chen, C.-T. Tsai, Y.-N. Chen, K.-C. Chou, C.-L. Li, C.-H. Tsai, K.-W. Wu, Y.-C. Chou, C.-Y. Li, W.-S. Lin, et al. A linear ensemble of individual and blended models for music rating prediction. In *KDD-Cup*, pages 21–60, 2011.

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

R. Cont and J.-P. Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pages 355–366, 2008a.

V. Dani, S. M. Kakade, and T. P. Hayes. The price of bandit information for online optimization. In *NIPS*, pages 345–352, 2008b.

S. Dharmadhikari, V. Fabian, and K. Jogdeo. Bounds on the moments of martingales. *The Annals of Mathematical Statistics*, pages 1719–1723, 1968.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-

armed bandit and Markov decision processes. In *Proceedings of COLT*, pages 255–270. Springer, 2002.

E. Even-Dar, M. Kearns, and J. Wortman. Risk-sensitive online learning. In *Algorithmic Learning Theory*, pages 199–213, 2006.

J. H. Fetzer. What is artificial intelligence? In *Artificial Intelligence: Its Scope and Limits*, pages 3–27. Springer, 1990.

H. Finner et al. A generalization of holder's inequality and some probability inequalities. *The Annals of Probability*, 20(4):1893–1901, 1992.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Annual Conference on Neural Information Processing Systems*, pages 3212–3220, 2012.

V. Gabillon, A. Lazaric, M. Ghavamzadeh, R. Ortner, and P. Bartlett. Improved learning complexity in combinatorial pure exploration bandits. In *Artificial Intelligence and Statistics*, pages 1004–1012, 2016.

N. Galichet, M. Sebag, and O. Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems*,

pages 1198–1206, 2016.

J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin. *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, 2009.

E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

J. Honorio and T. Jaakkola. Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pages 384–392, 2014.

D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pages 37–45, 2014.

D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1): 543–582, 2016.

K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Annual Conference on Information Science and Systems*, pages 1–6, 2014.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of COLT*, pages 423–439, 2014.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Annual Conference on Neural Information Processing Systems*, pages 325–333, 2016.

L. P. Kaelbling. Associative reinforcement learning: A generate and test algorithm. *Machine Learning*, 15(3):299–319, 1994.

S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 655–662, 2012.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of ICML*, pages 1238–1246, 2013.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

N. Korda, B. Szörényi, and L. Shuai. Distributed clustering of linear bandits in peer to peer networks. In *Proceedings of ICML*, pages 1301–1309, 2016.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, 2015.

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

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed

bandits with side information. In *Annual Conference on Neural Information Processing Systems*, pages 817–824, 2008.

T. Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. In *Proceedings of NIPS*, pages 1583–1592, 2017.

T. Lattimore, K. Crammer, and C. Szepesvári. Optimal resource allocation with semi-bandit feedback. *arXiv preprint arXiv:1406.3840*, 2014.

J. Le Ny, M. Dahleh, and E. Feron. Multi-uav dynamic routing with partial observations using restless bandit allocation indices. In *American Control Conference, 2008*, pages 4220–4225. IEEE, 2008.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553): 436, 2015.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

L. Li, Y. Lu, and D. Zhou. Provable optimal algorithms for generalized linear contextual bandits. *arXiv preprint arXiv:1703.00048*, 2017.

S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *Proceedings of ACM SIGIR*, pages 539–548, 2016.

X. Lian, H. Zhang, C.-J. Hsieh, Y. Huang, and J. Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In *NIPS*, pages 3054–3062, 2016.

J. Liebeherr, A. Burchard, and F. Ciucu. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2):1010–1024, 2012.

S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration

in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

A. M. Medina and S. Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of ICML*, pages 1642–1650, 2016.

K. Misra, E. M. Schwartz, and J. Abernethy. Dynamic online pricing with incomplete information using multi-armed bandit experiments. 2018.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

R. Munos et al. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

N. Nakamura, J. Seepaul, J. B. Kadane, and B. Reeja-Jayan. Design for low-temperature microwave-assisted crystallization of ceramic thin films. *Applied Stochastic Models in Business and Industry*, 2017.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17 (2):527–566, 2017.

H. Panahi. Model selection test for the heavy-tailed distributions under censored samples with application in financial data. *International Journal of Financial Studies*, 4(4):24, 2016.

O. Rivasplata. Subgaussian random variables: An expository note. *Internet publication, PDF*, 2012.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

J. A. Roberts, T. W. Boonstra, and M. Breakspear. The heavy tail of the human brain. *Current Opinion in Neurobiology*, 31:164–172, 2015.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Annual Conference on Neural Information Processing Systems*, pages 3275–3283, 2012.

E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Y. Seldin and G. Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759, 2017.

Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *ICML*, pages 1287–1295, 2014.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

S. Shahrampour, M. Noshad, and V. Tarokh. On sequential elimination algorithms for best-arm identification in multi-armed bandits. *IEEE Transactions on Signal Processing*, 2017.

S. Shalev-Shwartz et al. Online learning and online convex optimiza-

tion. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *COLT*, pages 3–24, 2013.

O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

M. Shao and C. L. Nikias. Signal processing with fractional lower order moments: stable processes and their applications. *Proceedings of the IEEE*, 81(7):986–1010, 1993.

W. F. Sharpe. Mutual fund performance. *The Journal of Business*, 39 (1):119–138, 1966.

A. Sokolov, J. Kreutzer, S. Riezler, and C. Lo. Stochastic structured prediction under bandit feedback. In *NIPS*, pages 1489–1497, 2016.

R. Sutton and A. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054, 1998.

C. Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

T. Uchiya, A. Nakamura, and M. Kudo. Algorithms for adversarial bandit problems with multiple plays. In *International Conference on Algorithmic Learning Theory*, pages 375–389. Springer, 2010.

S. Vakili and Q. Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

S. Vakili, K. Liu, and Q. Zhao. Deterministic sequencing of exploration

and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):759–767, 2013.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

B. von Bahr, C.-G. Esseen, et al. Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.

Y. Wang, S. Du, S. Balakrishnan, and A. Singh. Stochastic zeroth-order optimization in high dimensions. *arXiv preprint arXiv:1710.10551*, 2017.

A. Wibisono, M. J. Wainwright, M. I. Jordan, and J. C. Duchi. Finite sample convergence rates of zero-order stochastic optimization methods. In *NIPS*, pages 1439–1447, 2012.

L. Xu, C. Jiang, Y. Qian, Y. Zhao, J. Li, and Y. Ren. Dynamic privacy pricing: A multi-armed bandit approach with time-variant rewards. *IEEE Transactions on Information Forensics and Security*, 12(2): 271–285, 2017a.

Y. Xu, Q. Lin, and T. Yang. Accelerated stochastic subgradient methods under local error bound condition. *arXiv preprint arXiv:1607.01027*, 2016.

Y. Xu, Q. Lin, and T. Yang. Stochastic convex optimization: Faster

local growth implies faster global convergence. In *ICML*, pages 3821–3830, 2017b.

T. Yang and Q. Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *arXiv preprint arXiv:1512.03107*, 2015.

J. Y. Yu and E. Nikolova. Sample complexity of risk-averse bandit-arm selection. In *International Joint Conference on Artificial Intelligence*, 2013.

X. Yu, I. King, and M. R. Lyu. Risk control of best arm identification in multi-armed bandits via successive rejects. In *Proceedings of IEEE ICDM*, pages 1147–1152, 2017a.

X. Yu, M. R. Lyu, and I. King. CBRAP: Contextual bandits with random projection. In *Proceedings of AAAI*, pages 2859–2866, 2017b.

X. Yu, H. Shao, M. R. Lyu, and I. King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 937–946. AUAI Press, 2018.

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5): 1538–1556, 2012.

O. Zadorozhnyi, G. Benecke, S. Mandt, T. Scheffer, and M. Kloft. Huber-norm regularization for linear prediction models. In *ECML*, pages 714–730, 2016.

S. Zhao, E. Zhou, A. Sabharwal, and S. Ermon. Adaptive concentration inequalities for sequential decision problems. In *Proceedings of NIPS*, pages 1343–1351, 2016.

T. Zhao and I. King. Locality-sensitive linear bandit model for online social recommendation. In *Proceedings of ICONIP*, pages 80–90,

2016.

L. Zhou. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*, 2015.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.