Linear Stochastic Bandits with Heavy-Tailed Payoffs

SHAO, Han

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

The Chinese University of Hong Kong
March 2019

Thesis Assessment Committee

Professor ZHANG Shengyu (Chair)

Professor KING Kuo Chin Irwin (Thesis Supervisor)

Professor LYU Rung Tsong Michael (Thesis Co-supervisor)

Professor CHAN Siu On (Committee Member)

Professor YEUNG Dit-Yan (External Member)

Abstract of thesis entitled:

Linear Stochastic Bandits with Heavy-Tailed Payoffs

Submitted by SHAO, Han

for the degree of Master of Philosophy

at The Chinese University of Hong Kong in March 2019

In this thesis, we center around bandit models, especially linear stochastic bandits. Bandit models can tackle numerous problems of sequential learning with feedback of instantaneous payoffs. With such an ability, bandits have been applied into many applications, such as clinical trials, online recommendations and portfolio managements.

The main problem solved in this thesis is linear stochastic bandits with heavy-tailed payoffs. This problem is motivated by the wide application of linear stochastic bandits and the phenomenon of heavy-tailed distributions in various scenarios, e.g., network routing and financial markets. However, the problem has not been studied well in the previous work.

In linear stochastic bandits, it is commonly assumed that payoffs are with sub-Gaussian noises. In this thesis, we study the problem of linear stochastic bandits with non-sub-Gaussian payoffs. We assume that the distributions of the payoffs have finite moments of order p, with $p \in (1, 2]$. First, we analyze the regret lower bound of $\Omega(T^{\frac{1}{p}})$, where T denotes the number of rounds. This provides us with two hints: one is that the prior algorithms are far from optimal and the other is that finite variances (which refer to finite moments of order 2) lead to the regret of $\Omega(\sqrt{T})$. Then we propose two algorithms based on the techniques of median of means and truncation for two slightly different assumptions, which are bounded central moments and bounded raw moments. Both algorithms achieve the optimal regret upper bounds on the polynomial order of T. As far as we know, we are the first to derive the lower bound for this problem and develop almost optimal algorithms. Finally, we conduct experiments on synthetic datasets to demonstrate the superior performance of our algorithms.

Acknowledgement

I would like to express my deepest appreciation to my supervisors, Prof. Irwin King and Prof. Michael R. Lyu, for their advices, understanding, encouragements and support. I enjoy the days working with them at the Chinese University of Hong Kong. Without their help, I cannot find my interests in research or complete my MPhil study.

I would like to thank Dr. Emilie Kaufmann at INRIA Lille, who mentored me when I was an intern at INRIA. Her recognition of my research encourages me and gives me confidence. Her instructions guide me to solve problems. I am lucky enough to work with her.

I would like to thank my thesis assessment committee members, Prof. Shengyu Zhang, Prof. Siu On Chan and Prof. Dit-Yan Yeung, for their instructive comments and suggestions to this thesis.

I thank Xiaotian Yu, who guides me and collaborates with me on the research work in this thesis. I thank Xixian Chen and Tong Zhao for their guidance and help in my research. I thank all my group fellows: Shenglin Zhao, Hongyi Zhang, Hui Xu, Yuxin Su, Cuiyun Gao, Jichuan Zeng, Pinjia He, Jiani Zhang, Hou-Pong Chan, Jian Li, Wang Chen, Yaoman Li, Yue Wang, Shilin He, Pengpeng Liu, Jingjing Li, Yifan Gao, Haoli Bai, Wenxiang Jiao, Weibin Wu, Tianyi Yang, Xinyu Fu, Ziqiao Meng, for their kind help, especially during the days when I got my knees injured.

I also thank my friends Huan Shi, Honghui Zhang, Junyin Ru and Yixuan Ding for bringing me fun and color in my leisure time.

Last but not the least, I would like to thank my parents sincerely for their understanding and support. They always support every decision I make. Their endless love makes me brave and strong.

To my beloved parents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The bandit problem is a fundamental problem in the area of reinforcement learning. The origin of bandits dates back to its application in clinical trials by Thompson (1933). In this problem, a doctor has access to a set of treatments (corresponding to arms in bandits), of which performance follows unknown probability distributions. At each time step, after a patient's arrival, the doctor has to select one of the treatments and observes whether the patient is cured or not before the next time step. The goal is to maximize cumulative number of cured patients, which naturally addresses a trade-off between exploration and exploitation.

1.1 Background

Bandits, also called online learning with bandit feedback or optimization with bandit feedback, can be used to solve various sequential decision making problems. The basic model of bandits is K-armed multi-armed bandits (MAB). In the K-armed MAB model, there are K arms corresponding to K distributions $\{v_1, \cdots, v_K\}$ with means $\{u_1, \cdots, u_K\}$. Usually, we assume that the distributions are lighttailed. An agent has to choose arms sequentially for T rounds. At each round t, the agent chooses one arm $I_t \in \{1, \cdots, K\}$ without knowing the means and then observes feedback of arm I_t. The feedback is usually a stochastic reward of the chosen arm $y_t(I_t)$, which is drawn from v_{I_t}. Note that the rewards of one arm are identically and independently distributed. The learning process of K-armed MAB is sketched as follows.

Learning process of K-armed MAB

```
Input: the arm set $\{1, \cdots, K\}$, the number of rounds $T \geq K$.
For time $t = 1, \cdots, T$,
  Select an arm $I_t \in \{1, \cdots, K\}$.
  Observe a stochastic reward $y_t(I_t) \sim v_{I_t}$ of the chosen arm $I_t$.
```

One general goal of bandits is to maximize cumulative rewards, which naturally addresses a trade-off between exploration and exploitation. To be specific, exploration means selecting the arms the agent has not pulled enough to gain more information. Exploitation means selecting the empirical optimal arm to obtain the instantaneous reward as high as possible at the current round. We denote by $u_* \triangleq \max_{i \in \{1, \cdots, K\}} u_i$ the largest value among the expected rewards of arms. A performance metric for an algorithm $\mathcal{A}$ named regret is defined as

$$\bar{\mathbf{R}}(\mathcal{A}, T) \triangleq \max_{i=1,\cdots,K} \sum_{t=1}^{T} y_t(i) - \sum_{t=1}^{T} y_t(I_t), \tag{1.1}$$

which represents the difference between the cumulative rewards of always choosing the optimal arm at every around and the rewards of the algorithm $\mathcal{A}$. In the basic K-armed MAB setting aforementioned, the arm set is discrete and unchanged. The rewards of arms are unrelated. To break these constraints, an extension of MAB called structured bandits is proposed, where we usually assume that the rewards of arms follow a structure, e.g., linearity (Dani et al., 2008a), unimodality (Combes and Proutiere, 2014) and Lipschitz structure (Magureanu et al., 2014). Linear stochastic bandits (LSB) are a common class of structured bandits with rewards being linear mappings from arms to real numbers with an underlying linear parameter θ. In LSB, each arm is represented by a d-dimensional vector x and the expected reward of arm x is $x^\top \theta$. The observed reward is the summation of $x^\top \theta$ and a stochastic noise, which is usually assumed to be sub-Gaussian conditional on historical information. In LSB, the number of arms can be infinite and the arm set can change over time. For a better understanding, the learning process of LSB is sketched as follows.

```
Learning process of LSB
Input: the number of rounds $T$.
For time $t = 1, \cdots, T$,
  Given the arm set $\mathcal{D}_t \subseteq \mathbb{R}^d$, select an arm $x_t \in \mathcal{D}_t$.
  Observe a stochastic reward $y_t(x_t) = x_t^\top \theta + \eta_t$, where $\eta_t$ is a stochastic noise.
```

1.2 Motivation

Bandits can be applied to solve many real-world sequential decision making problems with feedback, such as clinical trials, online personalized recommendations and portfolio managements. In these problems, decision makers are provided with a set of choices. At each time, they make a decision and then receive a noisy feedback. The question is how to design a strategy to optimize a certain criterion.

1.2.1 Clinical Trials

As mentioned at the beginning of this chapter, clinical trials are the primal application of MAB (Thompson, 1933). A doctor has K treatments for a disease without knowing which treatment is optimal. Assume that the performance of each treatment is stochastic. At each time, a patient comes to the doctor. The doctor selects a treatment and treats the patient with it. The patient may be cured or die. The doctor collects the feedback of the patient and then treats the next patient. The goal is to save as many lives as possible.

1.2.2 Online Personalized Recommendations

Personalized recommendation, including news recommendation, advertisements in online shopping and sponsored search, learns users' preferences in a sequential process and recommends appealing items to the users. Taking news recommendation as an example, the recommender has a set of news and selects one piece from the set to a user at each time. If the user clicks the news, the recommender regards it as a positive reward; otherwise regards it negative. We assume that the feedback is drawn from Bernoulli distributions with expectations being the user's preferences. Therefore, the problem can be formulated into a bandit problem.

Furthermore, the recommender usually has feature information of users and news. It is common to assume that a user's preference for a piece of news is a linear function of the features of the user and the news. LSB can be used to solve this problem by taking the features into consideration. One advantage of LSB in news recommendation is to make recommendations for new users and with the latest news. For a new user without any history data, it is hard to recommend news feed-

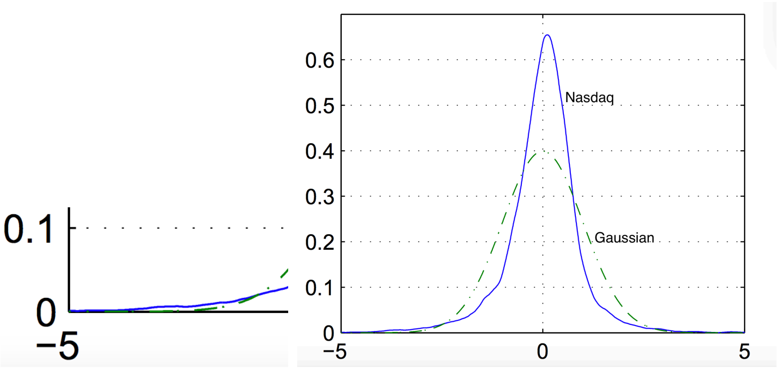


Figure 1.1: Simulation result of Nasdaq returns in the last 20 years.

ing the user's appetite without using the user's features. Besides, the pool of news evolves with time and thus how to recommend the latest news which have never been recommended before is difficult without considering the news' features. The effectiveness of LSB in personalized recommendation has been empirically validated by Li et al. (2010); Schwartz et al. (2017).

1.2.3 Portfolio Managements

Bandits can also find its application in finance, such as portfolio managements. An investor has fixed budget to invest in some financial products. At each time, the investor selects the weights of budget to invest, and then receives the returns, which are stochastic due to the randomness of financial markets. Hence, the portfolio managements can be formulated into a bandit problem (Badanidiyuru et al., 2013). With the features of the products, LSB can be adopted.

In practice, the noises of returns usually do not follow sub-Gaussian distributions. In Figure 1.1, it shows that the simulation result of Nasdaq returns in the last 20 years follows a heavy-tailed distribution. Hence, study of bandits with heavy-tailed payoffs is important.

MAB with heavy-tailed payoffs have been investigated by Bubeck et al. (2013). However, it is surprising to find that LSB with heavy-tailed payoffs have not been solved well. In this thesis, we study the problem of LSB with heavy-tailed payoffs.

1.3 Notations

In this section, we list all common symbols used in this thesis in Table 1.1.

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 $\mathbb{R}^+ \triangleq (0, +\infty)$
$\mathbb{E}[A]$	the expectation of a random variable A
$\mathbb{P}[\mathcal{E}]$	the probability of an event $\mathcal{E}$
$\mathrm{KL}(v_1, v_2)$	the Kullback-Leibler divergence between v_1 and v_2
Opt	the optimal arm
Out	the output
Δ_i	gap between arm i and the optimal arm, $\Delta_i = u_* - u_i$
$\langle x, y\rangle$ or $x^\top y$	the inner product of vectors x and y

1.4 Thesis Structure

The rest of this thesis is organized as follows.

- Chapter 2

  In this chapter, we present a survey of bandits. In Section 2.1, we discuss the theoretical developments of K-armed MAB with two general goals, which are regret minimization and pure exploration. In Section 2.2, we discuss the problems of structured bandits, including LSB and other important classes of structured bandits. Especially, we present the theoretical developments of LSB. In Section 2.3, we present some important variants of bandits. In Section 2.4, we construct a taxonomy of bandits.

- Chapter 3

  In this chapter, we show the results of our investigation for the problem of regret minimization in LSB with heavy-tailed payoffs. In Section 3.1, we introduce the background of the problem and our contributions in this problem. In Section 3.2, we give the preliminary of the problem, including the background, formal definition of the problem and basic algorithm framework to solve LSB. In Section 3.3, we develop the regret lower bound of the problem in this setting. In Section 3.4, we propose two almost optimal algorithms to solve the problem under two slightly different assumptions: one is bounded central moments of payoffs and the other is bounded raw moments. In Section 3.5, we give the proofs for the worst-case regret lower bound and the regret upper bounds of the two algorithms. In Section 3.6, we demonstrate the empirical study of our algorithms on synthetic datasets. In Section 3.7, we conclude our study in this problem.

- Chapter 4

  In this chapter, we summarize this thesis and present three potential directions for future work.

□ **End of chapter.**

Chapter 2

A Survey of Bandits

In the chapter, we review the present research progress in bandit problems. First, we give the literature review of theoretical developments in K-armed MAB and structured bandits. Then, we discuss some important variants of bandits. Finally, we construct a taxonomy of bandits.

2.1 K-armed Multi-Armed Bandits

Stochastic bandits have two common goals: one is regret minimization (equivalent to rewards maximization) as mentioned in Section 1.1 and the other is pure exploration, which usually refers to best arm identification. For regret minimization, we have to balance the decisions all over the time to maximize the cumulative rewards, which addresses the trade-off between exploration and exploitation. For pure exploration, we only take into accounts the final output at the end of the learning process, which addresses exploration only. We present the theoretical developments of K-armed MAB with these two goals in this section. In this thesis, we mainly focus on regret minimization.

2.1.1 Regret Minimization

The regret is defined in Eq. (1.1). The expected regret of an algorithm $\mathcal{A}$ is

$$\mathbb{E}\left[\bar{\mathbf{R}}(\mathcal{A},T)\right] = \mathbb{E}\left[\max_{i=1,\cdots,K}\sum_{t=1}^{T} y_t(i) - \sum_{t=1}^{T} y_t(I_t)\right]. \tag{2.1}$$

A frequently used metric named pseudo-regret is defined as

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

Actually, the pseudo-regret is more statistically meaningful than the expected regret. Therefore, the pseudo-regret is more common to use than the expected regret. In the following of this thesis, we adopt the pseudo-regret as the performance metric. When we mention regret, we refer to pseudo-regret.

We show the theoretical developments of regret minimization in K-armed MAB in Table 2.1. The origin of K-armed MAB dates back to 1933 (Thompson, 1933). The MAB problem was proposed formally by Robbins (1952) for the first time. Then by Lai and Robbins (1985), the problem-dependent asymptotic lower bound and an algorithm based on <u>u</u>pper <u>c</u>onfidence <u>b</u>ound (UCB) index with asymptotically optimal upper bound were proposed. But this index is hard to compute. After that, Agrawal (1995) proposed a simpler and more general sample mean index policy, which also achieves asymptotically optimal upper bound. Auer et al. (2002a) proposed an efficient index policy named UCB1 that can achieve logarithmic regret uniformly over time. All these index policies mentioned adopt the idea of UCB. The main idea of UCB is optimism in face of uncertainty. The UCB-based policies estimate the means of arms and construct the confidence intervals of the true means based on historical data. With high probability,

Table 2.1: Theoretical developments of regret minimization in MAB. $\Delta_i = u_* - u_i$.

work	results
(Thompson, 1933) (Robbins, 1952)	original formalization
(Lai and Robbins, 1985)	the first theoretical analysis $\lim_{T\to\infty}\frac{\mathbf{R}(\mathcal{A},T)}{\log(T)} \geq \sum_{\Delta_i>0}\frac{\Delta_i}{\mathrm{KL}(u_i,u_*)}$ $\lim_{T\to\infty}\frac{\mathbf{R}(\mathsf{UCB},T)}{\log(T)} \leq \sum_{\Delta_i>0}\frac{\Delta_i}{\mathrm{KL}(u_i,u_*)}$
(Agrawal, 1995)	a simpler algorithm $\lim_{T\to\infty}\frac{\mathbf{R}(\mathsf{SM},T)}{\log(T)} \leq \sum_{\Delta_i>0}\frac{\Delta_i}{\mathrm{KL}(u_i,u_*)}$
(Auer et al., 2002a)	finite-time analysis $\mathbf{R}(\mathsf{UCB1},T) = O\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right)$ $\mathbf{R}(\mathsf{UCB1},T) = O\left(\sqrt{T}\right)$
(Agrawal and Goyal, 2012)	Bernoulli payoffs $\mathbf{R}(\mathsf{TS},T) = O\left(\left(\sum_{\Delta_i>0}\frac{1}{\Delta_i}^2\right)^2\log(T)\right)$
(Kaufmann et al., 2012)	Bernoulli payoffs $\lim_{T\to\infty}\frac{\mathbf{R}(\mathsf{TS},T)}{\log(T)} \leq \sum_{\Delta_i>0}\frac{\Delta_i}{\mathrm{KL}(u_i,u_*)}$
(Garivier et al., 2018)	finite-time lower bound small T: lower bound $\mathbf{R}(\mathcal{A},T) \geq \sum_{\Delta_i>0}\frac{\Delta_i T}{2K}$ large T: lower bound $\mathbf{R}(\mathcal{A},T) = \Omega\left(\sum_{\Delta_i>0}\frac{\Delta_i\log(T)}{\mathrm{KL}(u_i,u_*)}\right)$

the true means lie in the confidence intervals. Then the UCB-based algorithms choose the arm with the largest value among supremes of the confidence intervals.

Another line of methodology is Thompson sampling (TS), which has been investigated by Agrawal and Goyal (2012); Kaufmann et al. (2012). The idea of TS is to construct posterior distributions for the means of arms based on history. At each round, the algorithm draws a sample from the posterior distribution of each arm, and then selects the arm with the largest value of samples.

More recently, the non-asymptotic regret lower bound was proposed by Garivier et al. (2018), which proved that the regret lower bound is linear when the total number of rounds is small and that the regret lower bound is logarithmic when the total number of rounds is large.

2.1.2 Pure Exploration

Pure exploration is to output a solution to a question after exploration among the arm set. The most common question is identifying the best arm. There are two general settings of pure exploration: fixed confidence and fixed budget.

In the fixed confidence setting, given a probability threshold $\delta \in (0, 1)$, the agent has to output an optimal arm at the end of learning process with probability of error no greater than δ. The performance metric is sample complexity, which refers to the total number of rounds. In the fixed budget setting, given the total number of rounds T, the agent has to output an optimal arm after T rounds of learning. The performance metric is probability of error. We show the theoretical developments of pure exploration in K-armed MAB in Table 2.2.

In the fixed confidence setting, Even-Dar et al. (2002) proposed an algorithm based on successive elimination with upper bound of sample complexity matching the lower bound, which was proposed by Mannor and Tsitsiklis (2004), up to a logarithmic factor. The logarithmic factor was improved to doubly-logarithmic factor by Karnin et al. (2013). These two studies require the constraints of bounded payoffs. Jamieson et al. (2014) proposed a UCB-based algorithm to solve the problem of payoffs with sub-Gaussian noises. Kaufmann et al. (2016) provided a lower bound of sample complexity that involves information-theoretic divergences and proposed an algorithm with upper bound of sample complexity matching the lower bound in two-armed Gaussian bandits.

In the setting of fixed budget setting, Audibert and Bubeck (2010) proposed a lower bound of probability of error and two algorithms matching the lower bound up to a logarithmic factor. One of the algorithms called UCB-E is based on UCB and the other is based on the idea of successive rejects. The results were improved by Karnin et al. (2013) via sequential halving. Kaufmann et al. (2016) proposed an improved lower bound of probability of error in Gaussian cases.

Beyond identifying the best arm, bandits can do pure exploration to output solutions to other questions. For example, Yu et al. (2017) explored to output the arm with largest mean-variance, Garivier et al. (2016) explored to output the best actions in game trees and Kaufmann et al. (2018) explored to output whether the lowest value among the expected rewards of arms is smaller than a given threshold.

2.2 Structured Bandits

In the previous section, we assume that all the expected rewards of arms are unrelated. However, if we have infinite number of arms, it is necessary to have some structure over the expected rewards of arms, such as linearity, unimodality and Lipschitz structure. Linear stochastic bandits are a very important class of structured bandits. In this section, we present the theoretical developments of regret minimization in LSB and introduce other important classes of structured bandits.

2.2.1 Linear Stochastic Bandits

The theoretical developments of LSB are shown in Table 2.3. The problem of LSB was first studied by Abe and Long (1999); Auer (2000). Auer (2002) proposed an algorithm named LinRel based on UCB index for K-armed LSB. Dani et al. (2008a) investigated the problem of stochastic linear optimization with bandit feedback, which is LSB with

Table 2.2: Theoretical developments of pure exploration in MAB. H_1 and H_2 are hardness parameters of the problem.

work	results
(Even-Dar et al., 2002)	bounded payoffs $\mathbb{P}\left[T \geq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{K}{\delta\Delta_k}\right)\right] \leq \delta$
(Audibert and Bubeck, 2010)	bounded payoffs $\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq TK \exp\left(-\frac{T-K}{H_1}\right)$
(Karnin et al., 2013)	bounded payoffs $\mathbb{P}\left[T \geq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{1}{\delta} \log\left(\frac{1}{\Delta_k}\right)\right)\right] \leq \delta$ $\mathbb{P}\left[\mathsf{Out} \neq \mathsf{Opt}\right] \leq \log(K) \exp\left(-\frac{T}{\log(K) H_2}\right)$
(Jamieson et al., 2014)	sub-Gaussian noises $\mathbb{P}\left[T \geq H_1 \log\left(\frac{1}{\delta}\right) + H_3\right] \leq 4\sqrt{c\delta} + 4c\delta$
(Kaufmann et al., 2016)	two-armed Gaussian bandits $\lim_{\delta\to 0} \frac{\mathbb{E}[T]}{\log\left(\frac{1}{\delta}\right)} \geq \frac{2(\sigma_1+\sigma_2)^2}{(u_1-u_2)^2}$ $\lim_{\delta\to 0} \frac{\mathbb{E}[T]}{\log\left(\frac{1}{\delta}\right)} \leq \frac{2(\sigma_1+\sigma_2)^2}{(u_1-u_2)^2}$ $\lim_{T\to\infty} \sup -\frac{\log(\mathbb{P}[\mathsf{Out}\neq\mathsf{Opt}])}{T} \leq \frac{(u_1-u_2)^2}{2(\sigma_1+\sigma_2)^2}$

arms belonging to a compact and convex set. They derived the worst-case regret lower bound and proposed an algorithm with worst-case regret upper bound matching the lower bound up to a polylogarithmic factor. The algorithm also achieved a polylogarithmic problem-dependent regret upper bound in the K-armed case. The upper bounds were improved by Abbasi-Yadkori et al. (2011) in the polylogarithmic factor. For K-armed LSB, Chu et al. (2011) derived the worst-case regret lower bound and proposed an algorithm named LinUCB similar to LinRel. Agrawal and Goyal (2013) studied TS for LSB. Lattimore and Szepesvari (2017) proposed a problem-dependent regret lower bound and an asymptotically optimal algorithm.

Table 2.3: Theoretical developments in LSB. $\Delta = \min_i u_* - u_i$.

work	results
(Abe and Long, 1999) (Auer, 2000)	original formalization
(Auer, 2002)	first theoretical analysis; K arms $\mathbf{R}(\mathsf{LinRel}, T) = O\left(\sqrt{Td}\log^{\frac{3}{2}}(KT\log(T))\right)$
(Dani et al., 2008a)	compact and convex arm set; bounded payoffs $\mathbf{R}(\mathcal{A}, T) = \Omega\left(d\sqrt{T}\right)$ $\mathbf{R}(\mathsf{CB}_2, T) = O\left(d\sqrt{T}\log^{\frac{3}{2}}(T)\right)$ $\mathbf{R}(\mathsf{CB}_2, T) = O\left(\frac{d^2}{\Delta}\log^3(T)\right)$, if $\Delta > 0$
(Abbasi-Yadkori et al., 2011)	compact and convex arm set; sub-Gaussian noises $\mathbf{R}(\mathsf{OFUL}, T) = O\left(d\sqrt{T}\log(T)\right)$
(Chu et al., 2011)	K arms; bounded payoffs $\mathbf{R}(\mathcal{A}, T) = \Omega\left(\sqrt{dT}\right)$ $\mathbf{R}(\mathsf{LinUCB}, T) = O\left(\sqrt{dT}\log^{\frac{3}{2}}(KT\log(T))\right)$
(Agrawal and Goyal, 2013)	K arms; sub-Gaussian noises $\mathbf{R}(\mathsf{TS}, T) = O\left(d^2\sqrt{T}\log(dT)\right)$
(Lattimore and Szepesvari, 2017)	K arms; Gaussian payoffs $\lim_{T\to\infty} \frac{\mathbf{R}(\mathcal{A},T)}{\log(T)} \geq c(\mathcal{A}, \theta)$ $\lim_{T\to\infty} \frac{\mathbf{R}(\mathsf{OA},T)}{\log(T)} \leq c(\mathcal{A}, \theta)$

2.2.2 Other Classes of Structured Bandits

Numerous types of structures have been investigated in bandits. Here we list some important classes. Lipschitz bandits, where the expected rewards are a Lipschitz function of arms, were studied by Bubeck et al. (2011); Magureanu et al. (2014). Stochastic convex optimization with bandit feedback was investigated by Agarwal et al. (2011). In this setting, the arm set is a compact and convex set and the expected rewards are a 1-Lipschitz convex function of arms. Unimodal bandits, where the expected rewards of arms follow a unimodal structure, were studied by Combes and Proutiere (2014). Rank-1 bandits, where the expected rewards follow a special unimodal structure, were investigated by Katariya et al. (2017). Dueling bandits were studied by Komiyama et al. (2015); Yue et al. (2012). In the problem of dueling bandits, at each time two arms are pulled and only the relative feedback is revealed. Combes et al. (2017) proposed an asymptotically optimal algorithm for general structured bandits.

2.3 Variants of Bandits

In this section, we present several important variants of bandit problems. All bandits mentioned above have stochastic payoffs. Besides stochastic payoffs, another type is adversarial payoffs, which are generated by an adversary arbitrarily (Auer et al., 1995, 2002b). Bandits with adversarial payoffs are called adversarial bandits. At each round, the agent chooses an arm and the adversary generates a payoff for the chosen arm at the same time. An intermediate type of payoffs between stochastic payoffs and adversarial payoffs is contaminated stochastic payoffs (Seldin and Lugosi, 2017; Seldin and Slivkins, 2014). In this setting, most of the payoffs are stochastic and some of the payoffs are contaminated to be adversarial under some specific assumptions.

Bandit problems can also be called online learning with bandit feedback, because in this setting, an agent can only obtain the feedback of the chosen arm. If the agent gets feedback of all arms, the problem is called online learning with full information (Hazan et al., 2016). If the agent gets feedback of partial arms (i.e., more than one arm), the problem is called online learning with semi-bandit feedback (Lattimore et al., 2014). Semi-bandit feedback is popular in the problems of combinatorial bandits, where the agent chooses more than one arm at each time and the reward is a function of all chosen arms. It is common to assume that the agent observes the feedback of all chosen arms (Chen et al., 2014, 2013; Kveton et al., 2015).

Usually we assume that there is only one agent who selects arms. However, the number of agents can be more than one in practical applications, which leads to the problem of multi-player bandits (Besson and Kaufmann, 2018; Kalathil et al., 2014). In multi-player bandits, the expected rewards of arms for different agents are different. At each round, each arm can only be selected by one agent and thus, avoiding collision is an essential issue in this problem.

2.4 Taxonomy

The taxonomy of bandit problems is constructed in Figure 2.1. We can classify a bandit problem according to the categories of goal, arm set, feedback format and number of agents basically. According to different goals of bandits, we can classify bandit problems into pure exploration and regret minimization. According to the arm set, we have bandit problems with discrete arms or continuous arms. If expected rewards of arms follow a structure, we have structured bandits, such as linear bandits, unimodal bandits and etc. According to feedback generating methods, we have adversarial bandits and stochastic bandits. With further assumptions on the distribution of payoffs in stochastic bandits, we have bandits with sub-Gaussian payoffs, heavy-tailed payoffs and etc. According to the amount of feedback, we have online learning problems with full information feedback, semi-bandit feedback and bandit feedback. According to the number of agents, we have multi-player bandits and classical bandits with single agent.

A specific bandit problem can be classified by these categories, e.g., regret minimization for linear stochastic bandits with sub-Gaussian

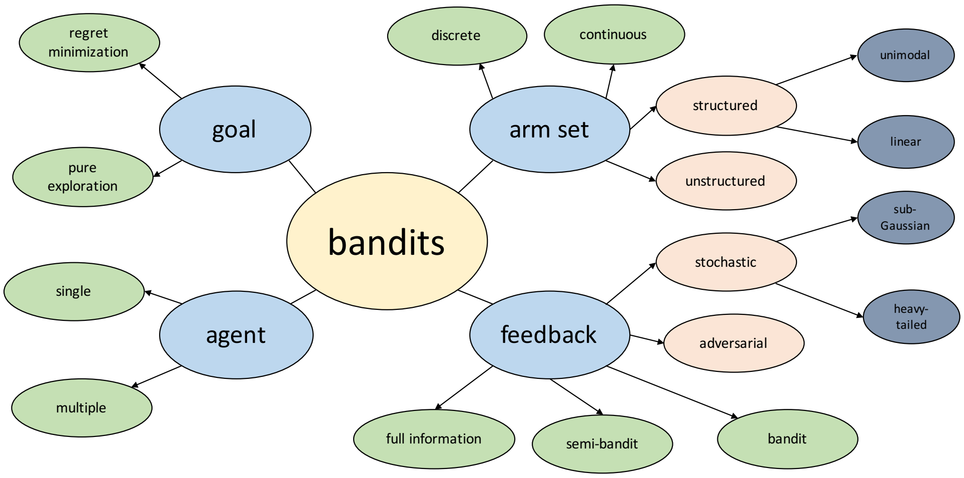


Figure 2.1: Taxonomy of bandits.

payoffs. Also, a specific problem can have further assumptions in different settings.

□ **End of chapter.**

Chapter 3

Linear Stochastic Bandits with Heavy Tails

Linear stochastic bandits are an important class of structured bandits. Previously, studies on LSB usually assumed that the noises follow sub-Gaussian distributions. In practice, many real applications with sequential decision making encounter noises, which follow non-sub-Gaussian distributions.

In this chapter, we study the problem of LSB under a different noise assumption. Specifically, we assume the distributions of payoffs have finite moments of order p, for some $p \in (1, 2]$, which is a weaker assumption compared with the sub-Gaussian noise assumption. The problem is called $\underline{\text{lin}}$ear stochastic $\underline{\text{b}}$andits with h$\underline{\text{e}}$avy-$\underline{\text{t}}$ailed payoffs (LinBET).

We analyze the regret lower bound as $\Omega(T^{\frac{1}{p}})$, which provides us with two hints: one is that the prior algorithms are far from optimal and the other is that finite variances (which refer to finite moments of order 2) lead to the regret of $\Omega(\sqrt{T})$. Then we propose two algorithms based on the techniques of median of means and truncation. Both algorithms achieve the optimal regret upper bounds on the polynomial order of T. As far as we know, we are the first to derive the lower bound for this problem and develop the almost optimal algorithms. Finally, we conduct experiments on synthetic datasets to demonstrate the superior performance of our algorithms.

In this chapter, we first present the background of linear stochastic bandits with heavy-tailed payoffs, the challenges and our contributions. Then we provide the preliminary for the problem and the technical results of this chapter. Finally, we give the conclusion for linear stochastic bandits with heavy-tailed payoffs.

3.1 Introduction

In the domain of bandits, algorithms are usually designed for maximizing cumulative payoffs in a sequence of decisions. Linear stochastic bandits have the assumption the the rewards are a linear mapping from arm space to real number space. There are good theoretical properties with the linear assumption, such as a closed-form solution for the linear parameter estimation. In prior work, linear stochastic bandits have been applied into many practical scenarios, e.g., online personalized recommendations (Li et al., 2010) and resource allocations (Lattimore et al., 2014).

In the traditional investigation of bandits, researchers usually assume that payoffs in decisions have noises following sub-Gaussian distributions (Abbasi-Yadkori et al., 2011; Bubeck et al., 2012). Note that sub-Gaussian noises are common, which include all bounded payoffs and many unbounded payoffs, such as Gaussian distributions.

However, in practice, we will encounter many cases with non-sub-Gaussian noises. An intuitive example is high-probability extreme returns in sequential investments in financial markets (Cont and Bouchaud, 2000). These events have higher probability to generate extreme values and we call that the events have heavy-tailed noises.

For a better understanding, we show two examples of heavy-tailed distributions. One is Pareto distributions, and the other is Weibull distributions. Both cases are with higher tails compared with sub-Gaussian distributions.

In this chapter, we consider a general definition of heavy-tailed noises. In particular, we investigate heavy-tailed payoffs in bandits with finite moments of order p, where $p \in (1, 2]$. When $p = 2$, stochastic payoffs in bandits are generated from distributions with finite variances. We notice that sub-Gaussian noises also have finite variances. Thus the question of the connection between the results of $p = 2$ and those of sub-Gaussian noises arises. When $1 < p < 2$, stochastic payoffs are generated from distributions with infinite variances (Shao and Nikias, 1993). In this case, noises from heavy-tailed distributions do not enjoy exponentially decaying tails. Clearly, it is difficult to learn a parameter of an arm when payoffs have heavy-tailed noises.

The regret minimization in K-arm MAB with heavy-tailed payoffs has been studied by Bubeck et al. (2013). For linear stochastic bandits, we adopt the same definition of the noises for heavy-tailed MAB proposed by Bubeck et al. (2013). Bubeck et al. (2013) proposed two algorithms with the regrets of $\widetilde{O}(\sqrt{T})$ [1] for MAB with finite moments of order 2. This result gives us a hint of the relation between the results of payoffs with finite variances and sub-Gaussian noises. Taking into consideration the importance of heavy tails in real applications, such as network routing with delays (Liebeherr et al., 2012) and sequential

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

investments in financial markets (Cont and Bouchaud, 2000), it is urgent and necessary to conduct a rigorous analysis of LinBET. Solving the problem of LinBET generalizes the applications of linear stochastic bandits.

Recently, Medina and Yang (2016) studied the problem of LinBET. The proposed algorithms only achieved the regret of $\widetilde{O}(T^{\frac{3}{4}})$ for the case of finite variances. Clearly, this result is very 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, we have an interesting and essential question as

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. In particular, we study the problem of LinBET characterized by finite p-th moments, where $p \in (1, 2]$. The problem of LinBET intrinsically has several interesting challenges. The first challenge is the lower bound of the problem. As far as we know, the regret lower bound of the problem remains unknown. The technical issues of the lower bound come from the construction of an elegant setting, including the arm set, payoffs and the linear parameter, for LinBET, which conserves the information of p, and the derivation of a lower bound with respect to p. The second challenge is to develop a estimator for the linear parameter that will be stable under the heavy-tailed noises as heavy-tailed noises increase the errors of least-squares estimator largely. We consider about adopting the techniques of median of means and truncation to construct such a robust estimator. The third challenge is how to adopt median of means and truncation to solve the problem with regret upper bounds matching the lower bound as closely as possible.

It is worth mentioning that the prior work by Medina and Yang (2016) has tried to solve this problem with median of means and truncation, but their estimators did not make full use of the contextual information of chosen arms to eliminate the effect from heavy-tailed noises, which eventually caused large regrets.

To solve the aforementioned challenges, we have the following three contributions. First, we rigorously analyze the worst-case regret lower bound on the problem of LinBET, and prove it as $\Omega(T^{\frac{1}{p}})$. The lower bound indicates that finite variances are possible to result in a regret bound of $\Omega(\sqrt{T})$ and that the prior results are sub-optimal according to the order of T. Second, based on the common techniques of median of means and truncation, we develop two novel bandit algorithms to solve LinBET. Both of the algorithms take advantage of the optimism in the face of uncertainty principle and build on the framework of OFUL proposed by Abbasi-Yadkori et al. (2011). The regret upper bounds of the proposed two algorithms match the lower bound up to a polylogarithmic factor. As far as we know, we are the first to solve LinBET almost optimally. Finally, we conduct experiments on synthetic datasets. The noises in the data are generated from Student's t-distribution and Pareto distribution, and then we demonstrate the effectiveness of our algorithms. Experimental results clearly demonstrate that our algorithms outperform the state-of-the-art results. For a better understanding, the main contributions of this chapter are summarized as follows.

- We provide the worst-case lower bound for the problem of LinBET characterized by finite moments of order p, where $p \in (1, 2]$. In the analysis, we construct an elegant setting of arms, linear parameters and payoffs in LinBET. Then we prove that for any bandit algorithm, the expectation of regret is at least $\Omega(T^{\frac{1}{p}})$.

- We develop two new bandit algorithms for LinBET, which are named as MENU and TOFU (with technical details shown in Section 3.4). The MENU algorithm adopts median of means with a well-designed grouping of payoffs and the TOFU algorithm adopts truncation by setting truncation threshold based on historical information. Both algorithms achieve the regret $\widetilde{O}(T^{\frac{1}{p}})$ with high probability.
- We run our algorithms on synthetic datasets to show the effectiveness of our proposed algorithms. The comparisons between our algorithms and two baselines MoM and CRT demonstrate the improvements on cumulative rewards for MENU and TOFU. It shows that our algorithms outperform the two baselines empirically.

3.2 Preliminary and Related Work

In this section, we first present the preliminary of the study, 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.

3.2.1 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 $i \in [d]$. 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 $\mathbb{1}_{\{.\}}$ be an indicator function, and $\mathbb{E}[X]$ the expectation of X. For $r \in \mathbb{R}$, its absolute value is $|r|$, its ceiling integer is $\lceil r \rceil$, and its floor integer is $\lfloor r \rfloor$.

3.2.2 Learning Setting

For a bandit algorithm $\mathcal{A}$, we consider sequential decisions in a given decision set. In a sequence of T decisions, the ultimate goal is to maximize cumulative rewards. In particular, 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 $x \in D_t$ and some $D \geq 0$. The algorithm $\mathcal{A}$ has to choose an arm $x_t \in D_t$ and then observes a stochastic payoff $y_t(x_t)$ of the chosen arm. For notation simplicity, we also write $y_t = y_t(x_t)$. In the linear setting, the expectation of the observed payoff for the chosen arm is a linear function of the arm 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 $\mathbb{E}\left[\eta_t | \mathcal{F}_{t-1}\right] = 0$, 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 $\mathbb{E}[y_t(x_t)|\mathcal{F}_{t-1}] = \langle x_t, \theta_* \rangle$. As mentioned in previous chapters, for an algorithm $\mathcal{A}$, to maximize cumulative payoffs is equivalent to minimizing the regret as

$$\mathbf{R}(\mathcal{A}, T) \triangleq \left(\sum_{t=1}^{T} \langle x_t^*, \theta_* \rangle \right) - \left(\sum_{t=1}^{T} \langle x_t, \theta_* \rangle \right) = \sum_{t=1}^{T} \langle x_t^* - x_t, \theta_* \rangle, \qquad (3.1)$$

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 worst-case upper bounds of $\mathbf{R}(\mathcal{A}, T)$ with respect to T, and provide the worst-case lower bound for LinBET in expectation for any algorithm. The formal definition for the problem of LinBET is shown as follows.

Definition 3.1 (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 rewards 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 $b, c \in (0, +\infty)$.*

3.2.3 Related Work

The origin model of MAB dates back to 1933 with study by Robbins (1952); Thompson (1933). One of the most important characteristics of MAB is addressing the trade-off between exploration and exploitation. The problem-dependent asymptotic lower bound of bandits was proposed by Lai and Robbins (1985). In bandits, there are two important methodologies. One is upper confidence bound, and the other is Thompson sampling. The methodology of UCB was developed in Agrawal (1995); Lai and Robbins (1985) to match the lower bound as close as possible. Other related techniques can refer to (Agrawal and Goyal, 2012; Chapelle and Li, 2011; Gittins et al., 2011; Thompson, 1933).

The problem of MAB with heavy-tailed payoffs characterized by finite moments of order p has been well studied by Bubeck et al. (2013); Vakili et al. (2013); Yu et al. (2018). Bubeck et al. (2013) stated that finite moments of order 2 in MAB were able to achieve regret bounds of $\widetilde{O}(\sqrt{T})$. This is the same polynomial order of T as the sub-Gaussian case. Besides, Bubeck et al. (2013) also pointed out that the polynomial order of T in regret bounds decreased with p. Bubeck et al. (2013) constructed a specific setting to prove the lower bound of MAB with heavy-tailed payoffs and then proved that algorithms with the proposed robust estimators were optimal. Note that Bubeck et al. (2013); Vakili et al. (2013) derived a robust estimator for the expected payoff of each arm individually, which required using the technique of median of means and truncation for scalars. In our problem, we can have infinite number of arms and all arms are represented by d-dimensional vectors. Vakili et al. (2013) also presented the theoretical results for the case of $p > 2$. But we note that the case of $p > 2$ is not much interesting, because it can reduce to the case of $p = 2$ via Jensen's inequality. Recently, Yu et al. (2018) investigated pure exploration of MAB with heavy-tailed payoffs.

For the problem of linear stochastic bandits, which is also named linear reinforcement learning by Auer (2002), the worst-case lower bound is $\Omega(d\sqrt{T})$ when arms are represented by d-dimensional vectors (Dani et al., 2008b). Almost optimal bandit algorithms in linear stochastic bandits, which match the lower bound up to a polylogarithmic factor, have been well developed by Abbasi-Yadkori et al. (2011); Auer (2002); Chu et al. (2011); Dani et al. (2008a) in the sub-Gaussian setting.

Although we are not the first to study the problem of LinBET, the lower bound of this problem has never been studied before. Medina and Yang (2016) studied LinBET with $p \in (1, +\infty)$ and proposed two bandit algorithms based on median of means and confidence region with truncation respectively, which we just name as MoM and CRT. The algorithm of MoM achieved the regret of $\widetilde{O}(T^{\frac{2p-1}{3p-2}})$ and the algorithm of CRT achieved the regret of $\widetilde{O}(T^{\frac{1}{2}+\frac{1}{2p}})$. Both of the algorithms cannot recover the regret of $\widetilde{O}(\sqrt{T})$ when $p = 2$. The algorithm of CRT only achieved the regret of $\widetilde{O}(\sqrt{T})$ when $p \to +\infty$. Medina and Yang (2016) conjectured that it was possible to recover the regret upper bound of $\widetilde{O}(\sqrt{T})$ with p being a finite number.

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. (2013). The interesting point of Lattimore (2017) is the scale free algorithm, which might be practical in applications. Besides, Carpentier and Valko (2014) investigated the problem of extreme bandits, where stochastic payoffs of MAB follow Fréchet distributions. The setting of extreme bandits well fits for the real scenario of anomaly detection without contextual information. The order of regrets in extreme bandits was characterized by distributional parameters, and the regrets of algorithms for extreme bandits were similar to the results by Bubeck et al. (2013).

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. Hence, 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.

3.2.4 The Basic Algorithm OFUL

Our two algorithms are developed based on the framework of OFUL proposed by Abbasi-Yadkori et al. (2011). The idea of OFUL is using ridge regression to estimate the linear parameter θ and constructing a self-normalized confidence ellipsoid of the estimate. In this subsection,

Algorithm 3.1 OFUL

```
input $R$, $d$, $\delta$, $\lambda$, $S$, $T$, $\{D_t\}_{t=1}^T$
initialization: $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(\mathbf{0}, S)$
for $t = 1, 2, \cdots, T$ do
    $(x_t, \tilde{\theta}_t) = \arg\max_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$    ▷ to select an arm
    $V_t = V_{t-1} + x_t x_t^\top$
    $\hat{\theta}_t = V_t^{-1} X_t^\top Y_t$
    $\beta_t = R\sqrt{d \log\left(\frac{1+tD^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}} S$
    $C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{V_t} \le \beta_t\}$    ▷ to update the confidence region
end for
```

we present the prior algorithm of OFUL and two lemmas about the theoretical results of OFUL, which are the basics for the analysis of our algorithms in Section 3.4. Lemma 3.1 shows a high probability self-normalized confidence region of the least square estimate.

Lemma 3.1 (Theorem 2 in (Abbasi-Yadkori et al., 2011))**.** *Let $\hat{\theta}_t$ denote the least square estimate of θ_* with the sequence of decisions $x_1, \cdots, x_t$ and observed payoffs $y_1, \cdots, y_t$. We assume that $\|\theta_*\|_2 \le S$. Assume that for all $t \in [T]$ and all $x_t \in D_t \subseteq \mathbb{R}^d$, η_t is $\mathcal{F}_{t-1}$ measurable and η_t is conditionally R-sub-Gaussian for some $R \ge 0$. Then $\hat{\theta}_t$ satisfies*

$$\mathbb{P}\left[\|\hat{\theta}_t - \theta_*\|_{V_t} \le R\sqrt{d \log\left(\frac{1+tD^2/\lambda}{\delta}\right)} + \lambda^{\frac{1}{2}} S\right] \ge 1 - \delta, \qquad (3.2)$$

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

The upper bound of the instantaneous regret $\langle x_t^* - x_t, \theta_* \rangle$ at time t can be decomposed as $\|x_t\|_{V_t^{-1}} \|\hat{\theta}_t - \theta_*\|_{V_t}$. With Lemma 3.1, it is direct to derive the following result for the regret of OFUL.

Lemma 3.2 (Theorem 3 in (Abbasi-Yadkori et al., 2011))**.** *Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \le D$, $\|\theta_*\|_2 \le S$, $|x_t^\top \theta_*| \le 1$*

and η_t is conditionally R-sub-Gaussian for some $R \geq 0$. Then, with probability at least $1-\delta$, for every $T \geq 0$, the regret of the OFUL *algorithm satisfies*

$$\begin{aligned}&\mathbf{R}(\mathsf{OFUL}, T)\\ &\leq 4\sqrt{Td\log\left(\lambda + \frac{TD}{d}\right)}\left(R\sqrt{d\log\left(1+\frac{nD}{\lambda d}\right) + 2\log\left(\frac{1}{\delta}\right)} + \lambda^{\frac{1}{2}}S\right).\end{aligned}$$

In Section 3.4, we develop two algorithms based on OFUL and adopt the similar proof techniques.

3.3 Lower Bound

In this section, we provide the lower bound for the problem LinBET. We construct a setting where the payoffs are heavy-tailed distributions with finite p-th raw moments. Assume $d \geq 2$ is even (when d is odd, the results of $d-1$ dimensions can be directly applied). For $D_t \subseteq \mathbb{R}^d$ with $t \in [T]$, we set the decision set fixed as $D_1 = \cdots = D_T = D_{(d)} \triangleq \{(x_1, \cdots, x_d) \in \mathbb{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$. This 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 $\theta_*^\top x$. In particular, we have the payoff function of x as

$$y(x) = \begin{cases} \left(\frac{1}{\Delta}\right)^{\frac{1}{p-1}} & \text{with a probability of } \Delta^{\frac{1}{p-1}}\theta_*^\top x, \\ 0 & \text{with a probability of } 1 - \Delta^{\frac{1}{p-1}}\theta_*^\top x. \end{cases} \tag{3.3}$$

In this setting, we can derive the worst-case lower bound of LinBET as follows.

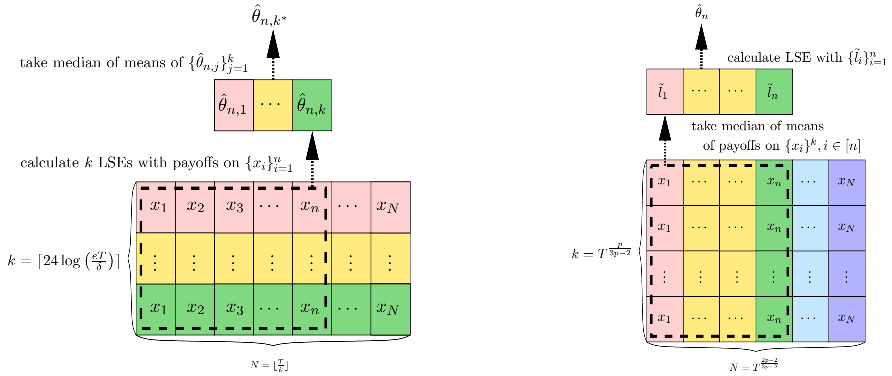
Theorem 3.1. *If θ_* is chosen uniformly at random from S_d, and the payoff for each $x \in D_{(d)}$ is in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ with mean $\theta_*^\top x$, then for any algorithm $\mathcal{A}$ and every $T \geq (d/12)^{\frac{p-1}{p}}$, we have*

$$\mathbb{E}\left[R(\mathcal{A}, T)\right] \geq \frac{d}{192} T^{\frac{1}{p}}. \tag{3.4}$$

In the proof of Theorem 3.1, we first prove the lower bound for the case of $d = 2$, 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 $(x_{*,2i-1}, x_{*,2i}) = (1, 0)$. Thus, 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 expectation of regret can be represented as an integration of $\Delta(1 - x_{t,2i-1})$ over $D_{(d)}$. Finally, with fundamental inequalities in information theory, it is easy to obtain the worst-case regret lower bound by taking $\Delta = T^{-\frac{p-1}{p}}/12$.

Dani et al. (2008a) proposed a method of taking martingale differences to prove the lower bound for linear stochastic bandits. But it is not directly feasible for the proof of lower bound in LinBET, because under our construction of heavy-tailed payoffs (i.e., Eq. (3.5)), the information of p will be excluded in the computation of martingale differences. In addition, our proof is partially inspired by Bubeck (2010). The detailed proof of Theorem 3.1 is shown in Section 3.5.

Remark 3.1. *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 3.1*


$\hat{\theta}_{n,k^*}$
take median of means of $\{\hat{\theta}_{n,j}\}_{j=1}^k$
$\hat{\theta}_{n,1}$ $\cdots$ $\hat{\theta}_{n,k}$
calculate k LSEs with payoffs on $\{x_i\}_{i=1}^n$
x_1 x_2 x_3 $\cdots$ x_n $\cdots$ x_N
$k = \lceil 24 \log\left(\frac{eT}{\delta}\right) \rceil$
$N = \lfloor \frac{T}{k} \rfloor$
$\hat{\theta}_n$
calculate LSE with $\{\tilde{l}_i\}_{i=1}^n$
$\tilde{l}_1$ $\cdots$ $\cdots$ $\tilde{l}_n$
take median of means
of payoffs on $\{x_i\}^k, i \in [n]$
x_1 $\cdots$ $\cdots$ x_n $\cdots$ x_N
$k = T^{\frac{p}{3p-2}}$
$N = T^{\frac{2p-2}{3p-2}}$

(a) Framework of MENU

(b) Framework of MoM

Figure 3.1: Framework comparison between our algorithm MENU and prior algorithm MoM by Medina and Yang (2016).

strongly indicates that it is possible to design bandit algorithms recovering $\widetilde{O}(\sqrt{T})$ with finite variances.

3.4 Algorithms and Upper Bounds

In this section, we develop two novel bandit algorithms to solve LinBET. Both are proved to be almost optimal with respect to the order of T. We prove regret upper bounds for the two algorithms. In particular, our core idea is based on the optimism in the face of uncertainty principle (OFU). Our algorithms are built based on the algorithm OFUL. The first algorithm is median of means under optimism in the face of uncertainty (MENU), which is shown in Algorithm 3.2, and the second algorithm is truncation under optimism in the face of uncertainty (TOFU), which is shown in Algorithm 3.3.

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

Algorithm 3.2 MENU

```
input $d$, $c$, $p$, $\delta$, $\lambda$, $S$, $T$, $\{D_n\}_{n=1}^{N}$
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
    $\beta_n = 3\left((9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right)$
    $C_n = \{\theta : \|\theta - \hat{\theta}_{n,k^*}\|_{V_n} \le \beta_n\}$    ▷ to update the confidence region
end for
```

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.

3.4.1 MENU and Regret

In this section, we first give the details of the design of MENU. Then, we develop the regret upper bound of the algorithm MENU.

Description of Algorithm MENU

We adopt median of means for heavy-tailed payoffs to obtain a robust estimate. To conduct median of means in LinBET, it is common to divide T pulls of bandits into $N \le T$ epochs. In each epoch, once an arm is chosen to be pulled, it will be played for multiple times to obtain an estimate of θ_*. We find that there exist different ways to construct the epochs. We design the framework of MENU in Figure 3.1(a), and show the framework of MoM designed by Medina and Yang (2016) in Figure 3.1(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 3.1(a), we derive an improved upper bound of regret in Theorem 3.2.

In light of Figure 3.1(a), we develop detailed algorithmic procedures in Algorithm 3.2 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 3.2. Now we have the following theorem for the regret upper bound of MENU.

Theorem 3.2. *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*

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

The technical challenges in MENU (i.e., Algorithm 3.2) and its proofs are discussed as follows. Based on the common techniques in linear stochastic bandits (Abbasi-Yadkori et al., 2011), in order to guarantee the instantaneous regret in LinBET, we need to guarantee

$\|\theta_* - \hat{\theta}_{n,k^*}\|_{V_n} \le \beta_n$ with high probability. We attack this issue by guaranteeing $\|\theta_* - \hat{\theta}_{n,j}\|_{V_n} \le \beta_n/3$ 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 3.2 in Section 3.5.

Remark 3.2. *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., 2013). In addition, there exists an implicit mild assumption in Algorithm 3.2 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 $T \ge 256 + 24\log(e/\delta)$ is required for $T \ge k$. The regret upper bound of MENU is $\widetilde{O}(T^{\frac{1}{p}})$, which implies that finite variances in LinBET are sufficient to achieve $\widetilde{O}(\sqrt{T})$.*

3.4.2 TOFU and Regret

In this section, we first describe the design of TOFU. Then, we present the regret upper bound of the algorithm TOFU.

Description of Algorithm TOFU

We demonstrate the algorithmic procedures of TOFU in Algorithm 3.3. For a better understanding, 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

Algorithm 3.3 TOFU

```
input $d$, $b$, $p$, $\delta$, $\lambda$, $T$, $\{D_t\}_{t=1}^T$
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
```

dimension individually. Besides, the truncating operations depend on the historical information of arms. By contrast, in the prior algorithm 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 3.3. *Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \le D$, $\|\theta_*\|_2 \le S$, $|x_t^\top \theta_*| \le L$ and $\mathbb{E}[|y_t|^p | \mathcal{F}_{t-1}] \le b$. Then, with probability at least $1 - \delta$, for every $T \ge 1$, the regret of the TOFU algorithm satisfies*

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

Similarly to the proof in Theorem 3.2, we can achieve the above results for the algorithm TOFU. We show the detailed proof of Theorem 3.3 in Section 3.5.

Remark 3.3. *For TOFU, we adopt the assumption of heavy-tailed payoffs on raw moments. It is worth pointing out that, when $p = 2$, we have regret upper bound for TOFU as $\widetilde{O}(d\sqrt{T})$. This implies that we can recover the same order of d as that under sub-Gaussian assumption (Abbasi-Yadkori et al., 2011). We notice that 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.*

3.5 Proofs of Theorems

In this section, we show the proofs of theorems.

3.5.1 Proof of Theorem 3.1

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 $p \in (1, 2]$, 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

$$y(x) = \begin{cases} \left(\frac{1}{\Delta}\right)^{\frac{1}{p-1}} & \text{with a probability of } \Delta^{\frac{1}{p-1}}\theta_*^\top x, \\ 0 & \text{with a probability of } 1 - \Delta^{\frac{1}{p-1}}\theta_*^\top x. \end{cases} \tag{3.5}$$

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 3.5.1. Its extension to the general case (i.e., $d > 2$) is provided in Subsection 3.5.1. 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 3.3. *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*

$$\mathbb{E}[R(\mathcal{A}, T)] \geq \frac{1}{96} T^{\frac{1}{p}}. \tag{3.6}$$

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

$$\begin{aligned}
&\mathbb{E}_j\left[\sum_{t=1}^{T}(y_t^* - y_t(x_t))\right] = \sum_{t=1}^{T} \mathbb{E}_j\left[\Delta(1 - x_{t,j})\right] = T\int_{D_{(2)}} \Delta(1 - x_j)\mathrm{d}\mathcal{P}_j(x) \\
&= T\Delta\left(1 - \int_{D_{(2)}} x_j \mathrm{d}\mathcal{P}_j(x)\right) \\
&= T\Delta\left(1 - \left(\int_{0 \le x_j \le \frac{1}{2}} x_j \mathrm{d}\mathcal{P}_j(x) + \int_{\frac{1}{2} < x_j \le 1} x_j \mathrm{d}\mathcal{P}_j(x)\right)\right) \\
&\ge T\Delta\left(1 - \left(\frac{1}{2}\mathcal{P}_j\left(0 \le X_j \le \frac{1}{2}\right) + \mathcal{P}_j\left(\frac{1}{2} < X_j \le 1\right)\right)\right), \qquad (3.7)
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbb{E}[\mathbf{R}(\mathcal{A}, T)] = \mathbb{E}_{\theta_*}\left[\mathbb{E}_j\left[\sum_{t=1}^{T}(y_t^* - y_t(x_t))\right]\right] \\
&\ge T\Delta\left(1 - \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{2}\mathcal{P}_j\left(0 \le X_j \le \frac{1}{2}\right) + \mathcal{P}_j\left(\frac{1}{2} < X_j \le 1\right)\right)\right). \qquad (3.8)
\end{aligned}$$

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

$$\mathcal{P}_j(X \in \mathcal{E}) \le \mathcal{P}_0(X \in \mathcal{E}) + \sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j)}, \qquad (3.9)$$

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

$$\begin{aligned}
&\mathbb{E}[\mathbf{R}(\mathcal{A}, T)] \\
&\ge T\Delta\left(1 - \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{2}\mathcal{P}_0\left(0 \le X_j \le \frac{1}{2}\right) + \mathcal{P}_0\left(\frac{1}{2} < X_j \le 1\right) + \frac{3}{2}\sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j)}\right)\right) \\
&= T\Delta\left(\frac{1}{4} - \frac{3}{4}\sum_{j=1}^{2}\sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j)}\right). \qquad (3.10)
\end{aligned}$$

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

$$\mathrm{KL}(\mathcal{P}_0, \mathcal{P}_j) \leq \mathrm{KL}(\mathcal{P}_0^T, \mathcal{P}_j^T). \tag{3.11}$$

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

$$\begin{aligned}
&\mathrm{KL}(\mathcal{P}_0^T, \mathcal{P}_j^T) \\
&= \mathrm{KL}(\mathcal{P}_0^1, \mathcal{P}_j^1) + \sum_{t=2}^{T} \int_{W_{t-1}} \mathrm{KL}\left(\mathcal{P}_0^t(\cdot|w_{t-1}), \mathcal{P}_j^t(\cdot|w_{t-1})\right) \mathrm{d}\mathcal{P}_0^{t-1}(W_{t-1}) \\
&= \mathrm{KL}(\mathcal{P}_0^1, \mathcal{P}_j^1) + \sum_{t=2}^{T} \int_{x_t \in D_{(2)}} \int_{W_{t-1}|x_{t,j}=x_j} \qquad (3.12) \\
&\mathrm{KL}\left(\Delta^{\frac{p}{p-1}}, \Delta^{\frac{p}{p-1}}(1+x_j)\right) \mathrm{d}\mathcal{P}_0^{t-1}(W_{t-1}|x_{t,j}=x_j)\mathrm{d}\mathcal{P}_0^{t-1}(x_{t,j}=x_j) \\
&\qquad (3.13) \\
&\leq 2\Delta^{\frac{p}{p-1}} + \qquad (3.14) \\
&\sum_{t=2}^{T} \int_{x_t \in D_{(2)}} \int_{W_{t-1}|x_{t,j}=x_j} 2\Delta^{\frac{p}{p-1}} \mathrm{d}\mathcal{P}_0^{t-1}(W_{t-1}|x_{t,j}=x_j)\mathrm{d}\mathcal{P}_0^{t-1}(x_{t,j}=x_j) \\
&\qquad (3.15) \\
&= 2T\Delta^{\frac{p}{p-1}}, \qquad (3.16)
\end{aligned}$$

where Eq. (3.15) 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. (3.13). Therefore, we have

$$\mathbb{E}[\mathbf{R}(\mathcal{A}, T)] \geq T\Delta\left(\frac{1}{4} - \frac{3}{2}\sqrt{T\Delta^{\frac{p}{p-1}}}\right) \geq \frac{1}{96}T^{\frac{1}{p}}, \tag{3.17}$$

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

$$\mathbb{E}[\mathbf{R}(\mathcal{A},T)] = \mathbb{E}_{\mathcal{A}}\left[\mathbb{E}_{\theta_*}\left[\mathbb{E}_j\left[\sum_{t=1}^{T}(y_t^* - y_t(x_t))\right]\right]\right]. \tag{3.18}$$

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. (3.17). □

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

$$\mathbb{E}\left[\mathbf{R}^{(i)}(\mathcal{A},T)\right] \geq \frac{1}{96}T^{\frac{1}{p}}, \tag{3.19}$$

where $i \in [d/2]$. Summing over the $d/2$ components of Eq. (3.19) completes the proof for Theorem 3.1.

3.5.2 Proof of Theorem 3.2

To prove Theorem 3.2, we start with proving the following two lemmas. Recall that the algorithm in the chapter is based on least-squares estimate (LSE).

Lemma 3.4 (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*

$$\mathbb{P}\left[\|\hat{\theta}_n - \theta_*\|_{V_n} \leq (9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right] \geq \frac{3}{4}, \tag{3.20}$$

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

$$V_n^{-\frac{1}{2}} X_n^\top = V\left(\Sigma_n^2 + \lambda I_d\right)^{-\frac{1}{2}} \Sigma_n U^\top. \tag{3.21}$$

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

$$\begin{aligned}\|\hat{\theta}_n - \theta_*\|_{V_n} &= \|V_n^{-1} X_n^\top (Y_n - X_n\theta_*) - \lambda V_n^{-1}\theta_*\|_{V_n} \\ &\leq \|V_n^{-\frac{1}{2}} X_n^\top (Y_n - X_n\theta_*)\|_2 + \lambda\|\theta_*\|_{V_n^{-1}} \\ &\leq \sqrt{\sum_{i=1}^d \left(u_i^\top (Y_n - X_n\theta_*)\right)^2} + \lambda^{\frac{1}{2}} S. \end{aligned} \tag{3.22}$$

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

$$\begin{aligned}&\mathbb{P}\left[\sum_{i=1}^d \left(\sum_{\tau=1}^n u_{i,\tau}\eta_\tau\right)^2 > \gamma^2\right] \\ &\leq \mathbb{P}\left[\exists i, \tau, |u_{i,\tau}\eta_\tau| > \gamma\right] + \mathbb{P}\left[\sum_{i=1}^d \left(\sum_{\tau=1}^n u_{i,\tau}\eta_\tau \mathbb{1}_{|u_{i,\tau}\eta_\tau| \leq \gamma}\right)^2 > \gamma^2\right],\end{aligned} \tag{3.23}$$

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

$$\begin{aligned} \mathbb{P}\left[\exists i, \tau, |u_{i,\tau}\eta_\tau| > \gamma\right] &\leq \sum_{i=1}^{d}\sum_{\tau=1}^{n} \mathbb{P}[|u_{i,\tau}\eta_\tau| > \gamma] && (3.24)\\ &\leq \frac{\sum_{i=1}^{d}\sum_{\tau=1}^{n} \mathbb{E}[|u_{i,\tau}\eta_\tau|^p]}{\gamma^p} && (3.25)\\ &\leq \frac{\sum_{i=1}^{d}\sum_{\tau=1}^{n} |u_{i,\tau}|^p c}{\gamma^p} \leq \frac{dcn^{\frac{2-p}{2}}}{\gamma^p}. && (3.26)\end{aligned}$$

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

$$\begin{aligned} &\mathbb{P}\left[\sum_{i=1}^{d}\left(\sum_{\tau=1}^{n} u_{i,\tau}\eta_\tau \mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}\right)^2 > \gamma^2\right]\\ &\leq \frac{\mathbb{E}\left[\sum_{i=1}^{d}(\sum_{\tau=1}^{n} u_{i,\tau}\eta_\tau \mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma})^2\right]}{\gamma^2}\\ &= \sum_{i=1}^{d}\frac{\mathbb{E}\left[\sum_{\tau=1}^{n}(u_{i,\tau}\eta_\tau)^2\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}\right]}{\gamma^2} +\\ &\sum_{i=1}^{d} 2\frac{\mathbb{E}\left[\sum_{\tau'>\tau}(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}(u_{i,\tau'}\eta_{\tau'})\mathbb{1}_{|u_{i,\tau'}\eta_{\tau'}|\leq\gamma}\right]}{\gamma^2}\\ &\leq \sum_{i=1}^{d}\frac{\mathbb{E}\left[\sum_{\tau=1}^{n}(u_{i,\tau}\eta_\tau)^2\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}\right]}{\gamma^2} +\\ &\sum_{i=1}^{d} 2\frac{\sum_{\tau'>\tau}\mathbb{E}[(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}]\mathbb{E}[(u_{i,\tau'}\eta_{\tau'})\mathbb{1}_{|u_{i,\tau'}\eta_{\tau'}|\leq\gamma}|\mu_{i,\tau}\eta_\tau]}{\gamma^2}\\ &\leq \sum_{i=1}^{d}\left(\frac{\sum_{\tau=1}^{n}|u_{i,\tau}|^p c}{\gamma^p} + \left(\frac{\sum_{\tau=1}^{n}|u_{i,\tau}|^p c}{\gamma^p}\right)^2\right) && (3.27)\\ &\leq \frac{dcn^{\frac{2-p}{2}}}{\gamma^p} + d\left(\frac{n^{\frac{2-p}{2}}c}{\gamma^p}\right)^2. && (3.28)\end{aligned}$$

Note that Eq. (3.27) uses the fact as follows.

$$\mathbb{E}[(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|\leq\gamma}|\mathcal{F}_{\tau-1}] = -\mathbb{E}[(u_{i,\tau}\eta_\tau)\mathbb{1}_{|u_{i,\tau}\eta_\tau|>\gamma}|\mathcal{F}_{\tau-1}]. \qquad (3.29)$$

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

Lemma 3.5. *Recall $\hat{\theta}_{n,j}$, $\hat{\theta}_{n,k^*}$ and V_n in* MENU *(i.e., Algorithm 3.2). 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 inequality, we have

$$\mathbb{P}\left[\sum_{j=1}^{k} b_j \geq \frac{k}{3}\right] < \mathbb{P}\left[\sum_{j=1}^{k} b_j - p_j \geq \frac{k}{12}\right] \leq e^{-\frac{k}{24}}, \tag{3.30}$$

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. (3.30) 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 3.4 and 3.5, 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:

$$\begin{aligned} r_n &= \theta_*^\top x_* - \theta_*^\top x_n \\ &\le \tilde{\theta}_n^\top x_n - \theta_*^\top x_n \\ &\le \left(\|\tilde{\theta}_n - \hat{\theta}_{n-1,k^*}\|_{V_{n-1}} + \|\hat{\theta}_{n-1,k^*} - \theta_*\|_{V_{n-1}} \right) \|x_n\|_{V_{n-1}^{-1}} \\ &\le 2\beta_{n-1} \|x_n\|_{V_{n-1}^{-1}}, \end{aligned} \tag{3.31}$$

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

$$r_n \le 2L. \tag{3.32}$$

We can get that

$$r_n \le 2 \min\{\beta_{n-1}\|x_n\|_{V_{n-1}^{-1}}, L\} \le 2(\beta_{n-1} + L) \min\{\|x_n\|_{V_{n-1}^{-1}}, 1\}. \tag{3.33}$$

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

$$\begin{aligned} \sum_{n=1}^{N} \min\{\|x_n\|^2_{V_{n-1}^{-1}}, 1\} &\le 2 \sum_{n=1}^{N} \log(1 + \|x_n\|^2_{V_{n-1}^{-1}}) \\ &= 2 \log \left(\frac{\det(V_N)}{\det(V_0)} \right) \\ &\le 2d \log \left(1 + \frac{ND^2}{\lambda d} \right), \end{aligned} \tag{3.34}$$

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

$$\begin{aligned} &\mathbf{R}(\mathsf{MENU}, T) \\ &\le k \sum_{n=1}^{N} r_n \le k \sqrt{N \sum_{n=1}^{N} r_n^2} \\ &\le 2kN^{\frac{1}{2}} (\beta_N + L) \sqrt{\sum_{n=1}^{N} \min\{\|x_n\|^2_{V_{n-1}^{-1}}, 1\}} \\ &\le 6 \left((12dc)^{\frac{1}{p}} + \lambda^{\frac{1}{2}} S + L \right) T^{\frac{1}{p}} \left(24 \log \left(\frac{eT}{\delta} \right) + 1 \right)^{\frac{p-1}{p}} \sqrt{2d \log \left(1 + \frac{TD^2}{\lambda d} \right)}. \end{aligned} \tag{3.35}$$

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

3.5.3 Proof of Theorem 3.3

Lemma 3.6. *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 3.3). 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*

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leq 4\sqrt{d}b^{\frac{1}{p}}\left(\log\left(\frac{2d}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}}S, \tag{3.36}$$

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. (3.22), we have

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leq \sqrt{\sum_{i=1}^d \left(u_i^\top (Y_i^\dagger - X_t\theta_*)\right)^2} + \lambda^{\frac{1}{2}}S. \tag{3.37}$$

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

$$
\begin{aligned}
& u_i^\top \left(Y_i^\dagger - X_t\theta_*\right) && (3.38)\\
&= \sum_{\tau=1}^{t} u_{i,\tau}\left(y_\tau^\dagger - \mathbb{E}[y_\tau|\mathcal{F}_{\tau-1}]\right) && (3.39)\\
&= \sum_{\tau=1}^{t} u_{i,\tau}\left(y_\tau^\dagger - \mathbb{E}\left[y_\tau^\dagger|\mathcal{F}_{\tau-1}\right] - \mathbb{E}\left[y_\tau \mathbb{1}_{|u_{i,\tau}y_\tau|>b_t}|\mathcal{F}_{\tau-1}\right]\right)\\
&\le \left|\sum_{\tau=1}^{t} u_{i,\tau}(y_\tau^\dagger - \mathbb{E}[y_\tau^\dagger|\mathcal{F}_{\tau-1}])\right| + \left|\sum_{\tau=1}^{t} u_{i,\tau}\mathbb{E}[y_\tau \mathbb{1}_{|u_{i,\tau}y_\tau|>b_t}|\mathcal{F}_{\tau-1}]\right|\\
&\le \left|2b_t \log\left(\frac{2d}{\delta}\right) + \frac{1}{2b_t}\sum_{\tau=1}^{t}\mathbb{E}\left[u_{i,\tau}^2\left(y_\tau^\dagger - \mathbb{E}\left[y_\tau^\dagger|\mathcal{F}_{\tau-1}\right]\right)^2|\mathcal{F}_{\tau-1}\right]\right|\\
&+ \left|\sum_{\tau=1}^{t}\mathbb{E}[u_{i,\tau}y_\tau \mathbb{1}_{|u_{i,\tau}y_\tau|>b_t}|\mathcal{F}_{\tau-1}]\right| && (3.40)\\
&\le 2b_t \log\left(\frac{2d}{\delta}\right) + \frac{\sum_{\tau=1}^{t}|u_{i,\tau}|^p b}{2b_t^{p-1}} + \frac{\sum_{\tau=1}^{t}|u_{i,\tau}|^p b}{b_t^{p-1}}\\
&\le 4b^{\frac{1}{p}}\left(\log\left(\frac{2d}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}}, && (3.41)
\end{aligned}
$$

where Eq. (3.40) is obtained by applying Bernstein's inequality for martingales (Seldin et al., 2012) and Eq. (3.41) is obtained by the fact that $\|u_i\|_p \le t^{\frac{2-p}{2p}}$ and by setting $b_t = (b/\log(2d/\delta))^{\frac{1}{p}}\, t^{\frac{2-p}{2p}}$. Combining Eq. (3.37) and Eq. (3.41) completes the proof. □

With similar procedures to the proof of Theorem 3.2, we have the regret of TOFU as follows.

$$
\begin{aligned}
&\mathbf{R}(\mathsf{TOFU}, T)\\
&\le 2T^{\frac{1}{p}}\left(4\sqrt{d}b^{\frac{1}{p}}\left(\log\left(\frac{2dT}{\delta}\right)\right)^{\frac{p-1}{p}} + \lambda^{\frac{1}{2}}S + L\right)\sqrt{2d\log\left(1+\frac{TD^2}{\lambda d}\right)},
\end{aligned}
\tag{3.42}
$$

which completes the proof.

Table 3.1: 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

3.6 Experiments

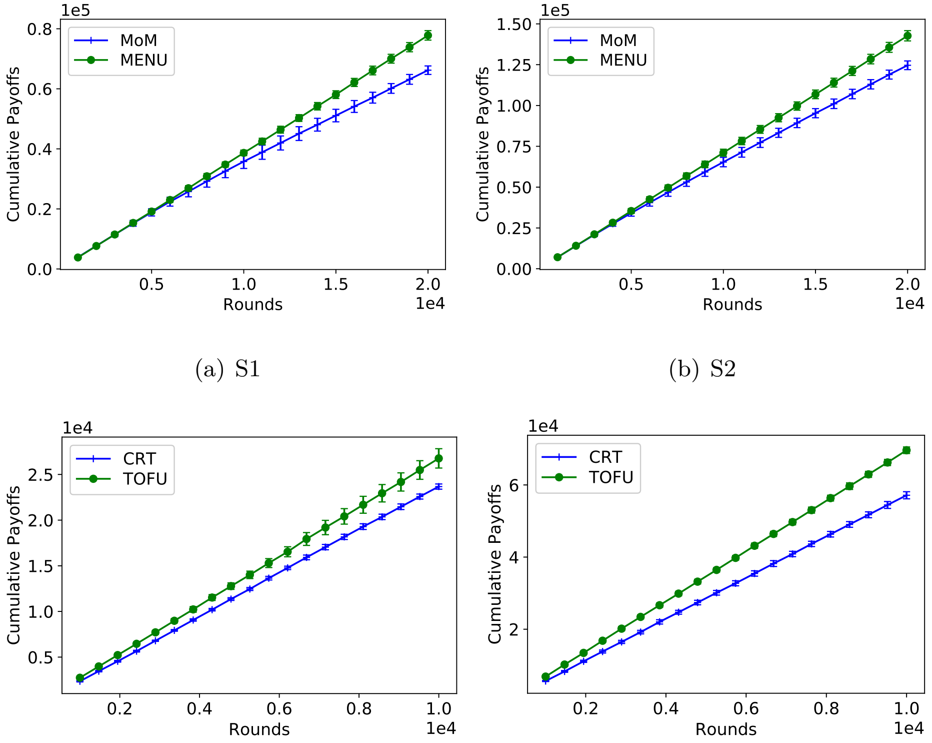
In this section, we conduct a series of experiments in light of synthetic datasets to evaluate the performance of our proposed bandit algorithms: MENU and TOFU. We compare our algorithms with MoM and CRT proposed by Medina and Yang (2016). We run multiple independent repetitions for each dataset in a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

3.6.1 Datasets and Setting

To show effectiveness of bandit algorithms, we show cumulative payoffs with respect to number of rounds for playing bandits over a fixed finite-arm decision set. We run ten independent experiments for calculating the average cumulative payoffs with a standard deviation. For verifications, we adopt four synthetic datasets (named as S1–S4) in the experiments, of which statistics are shown in Table 3.1. In the experiments, we need to know p, b or p, c, which correspond to the assumptions of Theorem 3.2 or Theorem 3.3. 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 can estimate c, and for Pareto distributions, we can estimate b. In addition, we can choose different parameters (e.g., larger values) in the distributions, and recalculate the parameters of b and c.

For datasets of S1 and S2, they contain different numbers of arms and different dimensions for the contextual information. Then, we adopt standard Student's t-distribution to generate heavy-tailed noises. For the chosen arm $x_t \in D_t$ at time t, the expected payoff is $x_t^\top \theta_*$. For the realized payoffs, we add a noise generated from a standard Student's t-distribution to the expected payoffs. Without loss of generality, based on a uniform distribution over $[0, 1]$, we generate each dimension of contextual information for an arm, as well as the underlying parameter. 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 mo-



(a) S1

(b) S2

(c) S3

(d) S4

Figure 3.2: Comparison of cumulative payoffs for synthetic datasets S1-S4 with four algorithms.

ment as $\mathbb{E}\left[|y_t|^{1.5}\right] = \alpha s_m^{1.5}/(\alpha - 1.5) = 4s_m^{1.5}$. We take the maximum of $4s_m^{1.5}$ among all arms as the bound of the 1.5-th raw moment. We generate arms and the parameter for S3 and S4 similar to S1 and S2.

In Figure 3.2, 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 independently run MENU and MoM and set $T = 2 \times 10^4$. For S3 and S4, we independently run TOFU and CRT and set $T = 1 \times 10^4$. For all algorithms, we set $\lambda = 1.0$, and $\delta = 0.1$. We also test other parameters in the experiments.

3.6.2 Results and Discussions

For comparisons, we show experimental results in Figure 3.2. From the figure, it is easy to find that our proposed two algorithms outperform the previous algorithms MoM and CRT. The observations are consistent with the theoretical results in Theorems 3.2 and 3.3. We further evaluate our algorithms with other synthetic datasets, as well as different λ and δ, and also observe similar superiority of MENU and TOFU. Finally, for comparison on regret, complexity and storage of four algorithms, we list all the comparison results as shown in Table 3.2.

Table 3.2: 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	$O(1)$	$O(\log T)$	$O(1)$	$O(T)$

3.7 Conclusion

In this chapter, we have studied the problem of LinBET. In particular, stochastic payoffs in this problem are characterized by finite p-th moments with $p \in (1, 2]$. Different from prior work, we broke the traditional assumption of sub-Gaussian noises in payoffs of bandits, and derived theoretical guarantees in light of the prior information of bounds on finite moments. We rigorously analyzed the lower bound of LinBET, which filled the space in this domain, and developed two novel bandit algorithms with regret upper bounds matching the lower bound up to a polylogarithmic factor. Two novel algorithms were proposed by taking advantage of median of means and truncation. In terms of polynomial dependence on T, we provided optimal algorithms for the problem of LinBET, and thus solved an open problem. This open problem has been pointed out by Medina and Yang (2016). Finally, our proposed algorithms have been verified in light of synthetic datasets, and beaten the state-of-the-art results.

Chapter 4

Conclusions and Future Directions

In this chapter, we conclude this thesis. First, we present the main contributions of this thesis. Then we introduce three potential directions for future work.

4.1 Main Contributions

We investigated linear stochastic bandits with heavy-tailed payoffs in this thesis. We developed two algorithms based on the techniques of median of means and truncation. Theoretically, our two algorithms achieved the optimal worst-case regret uppers matching the lower bound up to a logarithmic factor with respect to the number of rounds T. Empirically, we conducted experiments on synthetic datasets to demonstrate the effectiveness of our two algorithms.

4.2 Future Directions

For future work, we list three potential directions.

1. For LinBET, there are still some open problems unsolved yet. First, in this thesis, we only derived the worst-case regret bounds. A trivial extension of current theoretical analysis cannot lead to polylogarithmic problem-dependent regret upper bounds. How to derive the problem-dependent results is still unclear. Second, our results are almost optimal for the number of rounds T but not for dimension d. The impact of d in LinBET is unclear.

2. Another problem worth investigation is adaptive learning on parameters. In LinBET, we assume that we have prior knowledge of the moment bound parameters b and c. However, in real cases, without any prior knowledge, we have to estimate these parameters. This problem not only exists in LinBET, but also in a lot of other bandit problems. Therefore, how to learn these parameters adaptively in the learning process is very meaningful and challenging.

3. Finally, it has been pointed out that all index policies and Thompson sampling cannot achieve the optimal problem-dependent regret in linear bandits by (Lattimore and Szepesvari, 2017). In this work, they provided an asymptotically optimal algorithm with forced exploration, which forces to pull the arms in each dimension for $\sqrt{\log(T)}$ rounds. However, whether such forced exploration can be removed or not is unclear and this algorithm is extremely inefficient and even infeasible sometimes. Designing an efficient optimal algorithm for linear bandits is important and worth efforts. Not only linear bandits, but also other structured bandits face the same efficiency problem.

□ **End of chapter.**

Appendix A

List of Publications

[1] **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 8430–8439, 2018. **Spotlight presentation**.

[2] Xiaotian Yu, **Han Shao**, Michael R. Lyu and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 937–946, 2018.

□ **End of chapter.**

Bibliography

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

N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of International Conference on Machine Learning*, pages 3–11, 1999.

A. Agarwal, D. P. Foster, D. J. Hsu, S. M. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1035–1043, 2011.

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 *Proceedings of Conference on Learning Theory*, pages 39–1, 2012.

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

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *Proceedings of Conference on Learning Theory*, pages

13–29, 2010.

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 *Proceedings of Annual Symposium on Foundations of Computer Science*, pages 270–279, 2000.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of Annual Symposium on Foundations of Computer Science*, pages 322–331, 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 *Proceedings of Annual Symposium on Foundations of Computer Science*, pages 207–216, 2013.

L. Besson and E. Kaufmann. Multi-player bandits revisited. In *Proceedings of International Conference on Algorithmic Learning Theory*, 2018.

S. Bubeck. *Bandits games and clustering foundations.* PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2010.

S. Bubeck, G. Stoltz, and J. Y. Yu. Lipschitz bandits without the lipschitz constant. In *Proceedings of International Conference on Algorithmic Learning Theory*, pages 144–158, 2011.

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, 2013.

A. Carpentier and M. Valko. Extreme bandits. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1089–1097, 2014.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *Proceedings of Advances in Neural Information Processing Systems*, pages 379–387, 2014.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of International Conference on Machine Learning*, pages 151–159, 2013.

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

R. Combes and A. Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of International Conference on Machine Learning*, pages 521–529, 2014.

R. Combes, S. Magureanu, and A. Proutiere. Minimal exploration in structured stochastic bandits. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1763–1771, 2017.

R. Cont and J.-P. Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(2):170–196, 2000.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of 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 *Proceedings of Advances in Neural Information Processing Systems*, pages 345–352, 2008b.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *Proceedings of International Conference on Computational Learning Theory*, pages 255–270, 2002.

A. Garivier, E. Kaufmann, and W. M. Koolen. Maximin action identification: A new bandit framework for games. In *Proceedings of Conference on Learning Theory*, pages 1028–1050, 2016.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018.

J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

E. Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In *Proceedings of 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, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of Conference on Learning Theory*, pages 423–439, 2014.

D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of International Conference on Machine Learning*, pages 1238–1246, 2013.

S. Katariya, B. Kveton, C. Szepesvari, C. Vernade, and Z. Wen. Stochastic rank-1 bandits. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 392–401, 2017.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of International Conference on Algorithmic Learning Theory*, pages 199–213, 2012.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

E. Kaufmann, W. M. Koolen, and A. Garivier. Sequential test for the lowest mean: From Thompson to Murphy sampling. In *Proceedings of Advances in Neural Information Processing Systems*, pages 6335–6345, 2018.

J. Komiyama, J. Honda, H. Kashima, and H. Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Proceedings of Conference on Learning Theory*, pages 1141–1154, 2015.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of Inter-*

national Conference on Artificial Intelligence and Statistics, pages 535–543, 2015.

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

T. Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1583–1592, 2017.

T. Lattimore and C. Szepesvari. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 728–737, 2017.

T. Lattimore, K. Crammer, and C. Szepesvári. Optimal resource allocation with semi-bandit feedback. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 477–486, 2014.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the Nineteenth International Conference on World Wide Web*, pages 661–670, 2010.

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. Magureanu, R. Combes, and A. Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of Conference on Learning Theory*, pages 975–999, 2014.

S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

A. M. Medina and S. Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of International Conference on Machine Learning*, pages 1642–1650, 2016.

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.

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 *Proceedings of Conference on Learning Theory*, pages 1743–1759, 2017.

Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of International Conference on Machine Learning*, 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.

H. Shao, X. Yu, I. King, and M. R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Proceedings of Advances in Neural Information Processing Systems*, pages 8430–8439, 2018.

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. 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.

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.

X. Yu, I. King, and M. R. Lyu. Risk control of best arm identification in multi-armed bandits via successive rejects. In *Proceedings of International Conference on Data Mining*, pages 1147–1152, 2017.

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, 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.