# Pure Exploration of Multi-Armed Bandits with Heavy-Tailed Payoffs

**Xiaotian Yu, Han Shao, Michael R. Lyu, Irwin King**
[1]Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications,
Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China

## Abstract

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

## 1 Introduction

The prevailing decision-making model named Multi-Armed Bandits (MAB) elegantly characterizes a wide class of practical problems on sequential learning with partial feedbacks, which was first formally proposed and investigated in (Robbins, 1952). In general, a predominant characteristic of MAB is a trade-off between exploration and exploitation for sequential decisions, which has been frequently encountered in scientific research and various industrial applications, e.g., resource allocation, online advertising and personalized recommendations (Auer et al., 2002; Bubeck et al., 2012; Chu et al., 2011; Lattimore et al., 2015; Wu et al., 2016).

Most algorithms in MAB are primarily developed to maximize cumulative payoffs during a number of rounds for sequential decisions. Recently, there have been interesting investigations on various variants of the traditional MAB model, such as linear bandits (Auer, 2002), pure exploration in MAB (Audibert and Bubeck, 2010), risk-averse MAB (Sani et al., 2012), cascading bandits (Kveton et al., 2015) and clustering bandits (Korda et al., 2016; Li et al., 2016).

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

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

In this paper, we investigate the problem on pure exploration of MAB with heavy-tailed payoffs characterized by the bound of $p$-th moments. It is surprising to find that less effort has been devoted to pure exploration of MAB with heavy-tailed payoffs. In particular, it is still unknown about theoretical behaviours of pure exploration of MAB, which generates stochastic payoffs with bounded $p$-th moments. Compared with previous work on pure exploration of MAB, the problem of best arm identification with heavy-tailed payoffs has three challenges. The first challenge is the estimate of expected payoffs of an arm in MAB. It might not be sufficient to adopt an empirical average (EA) of observed payoffs with heavy-tailed noises for estimating a true mean. The second challenge is the probability of error for the estimate of expected payoffs, which affects performance of bandit algorithms in pure exploration of MAB. The third challenge is to develop effective bandit algorithms with theoretical guarantees for best arm identification of MAB with heavy-tailed stochastic payoffs.

To solve the above three challenges, we need to introduce a general assumption that stochastic payoffs in MAB are with finite $p$-th moments, where $p \in (1, +\infty)$. Note that the case of $p \in (1, 2]$ is weaker than the classic assumption of payoffs with sub-Gaussian noises in MAB. Then, we analyze theoretical behaviours of empirical average, which needs the assumption of finite $p$-th central moments and is an estimate of expected payoffs in sequential decisions. Besides, we also analyze the estimate of truncated empirical average (TEA), which needs the assumption of finite $p$-th raw moments. Based on different prior information, i.e., fixed confidence or fixed budget, we propose two bandit algorithms in pure exploration with heavy-tailed payoffs, where we fully take advantage of EA and TEA. Finally, based on synthetic data with noises from standard *Student's t-Distributions* and real-world financial data, we demonstrate the effectiveness of the proposed bandit algorithms for pure exploration of bandits generating payoffs with finite $p$-th moments. To the best of our knowledge, this is the first systematic investigation on pure exploration of MAB with heavy-tailed payoffs. For reading convenience, we summarize three contributions of this paper as follows.

- We technically analyze tail probabilities of EA and TEA to estimate true mean of arms in MAB with the general assumption of conditionally independent payoffs.
- We propose two bandit algorithms for pure exploration of MAB with heavy-tailed stochastic payoffs characterized by finite $p$-th moments, where $p \in (1, +\infty)$.
- We derive theoretical results of the proposed bandit algorithms, as well as demonstrating effectiveness of two algorithms via synthetic data and real-world financial data.

## 2 Preliminary and Related Work

In this section, we first present related notations and definitions used in this paper. Then, we present assumptions and a problem definition for pure exploration of MAB with heavy-tailed payoffs. Finally, we give a brief literature review on pure exploration of MAB and regret minimization in bandits, where stochastic payoffs have finite $p$-th moments.

### 2.1 Notations

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

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

We present the process of pure exploration of MAB as follows. For $t = 1, 2, \cdots, T$, $\mathcal{A}$ decides to play an arm $a_t \in [K]$ among a decision-arm set with historical information of $\{a_1, \pi_1(a_1), \cdots, a_{t-1}, \pi_{t-1}(a_{t-1})\}$. For MAB with $K$ arms, let $\mu(k)$ be the expected payoff for any arm $k \in [K]$. Then, $\mathcal{A}$ observes a stochastic payoff $\pi_t(a_t) \in \mathbb{R}$ with respect to $a_t$, of which the expectation conditional on $\mathcal{F}_{t-1}$ is $\mu(a_t)$ with $\mathcal{F}_{t-1} \triangleq$

$\{a_1, \pi_1(a_1), \cdots, a_{t-1}, \pi_{t-1}(a_{t-1}), a_t\}$ and $\mathcal{F}_0$ being an empty set. Based on $\pi_t(a_t)$, $\mathcal{A}$ updates parameters to proceed with the exploration at $t+1$. We store time index $t$ of playing arm $a_t$ in $\Phi(a_t)$, which is a set with increasing integers.

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

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

### 2.2 Assumptions and Problem Definition

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

1. Assume that $\mathsf{Opt} \triangleq \arg\max_{k\in[K]} \mu(k)$ is unique for pure exploration of MAB with $K$ arms.
2. Assume that MAB has heavy-tailed payoffs with the $p$-th raw or central moment conditional on $\mathcal{F}_{t-1}$ bounded by $B$ or $C$, for $t = 1, \cdots, T$.
3. Assume that the sequence of stochastic payoffs from arm $k \in [K]$ has noises with zero mean conditional on $\mathcal{F}_{t-1}$ in pure exploration of MAB. For any time instant $t \in [T]$ and the selected arm $a_t$, we define random noise from a true payoff as $\xi_t(a_t) \triangleq \pi_t(a_t) - \mu(a_t)$, and assume $\mathbb{E}[\xi_t(a_t)|\mathcal{F}_{t-1}] = 0$.

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

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

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

$$H_2^p \triangleq \max_{k\in[K]} k^{p-1}\Delta_k^{-p}, \bar{H}_2^p \triangleq \max_{k\in[K]} \sqrt{k}\Delta_k^{-p}. \quad (1)$$

### 2.3 Related Work

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

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

In the setting of fixed budget with payoffs under sub-Gaussian noises, (Audibert and Bubeck, 2010) developed a distribution-dependent lower bound of probability of error, and provided two algorithms, which achieve optimal probability of error up to logarithmic factors. Gabillon et al. (2012) proposed a unified algorithm for fixed budget and fixed confidence, which discusses $\epsilon$-optimal learning in best arm identification of MAB. Karnin et al. (2013) proposed a bandit algorithm via sequential halving to improve probability of error by a multiplicative constant. It is worth mentioning that (Kaufmann et al., 2016) investigated best arm identification of MAB under Gaussian or Bernoulli assumption, and provided lower bounds in terms of Kullback-Leibler diver-

Table 1: Comparisons on distributional assumptions and theoretical guarantees in pure exploration of MAB. Note we omit constant factors in the following inequalities, and $H_1$, $H_2$ and $H_3$ can refer to the corresponding work.

| setting | work | assumption on payoffs | algorithm | theoretical guarantee |
|---|---|---|---|---|
| fixed $\delta$ | Even-Dar et al. (2002) | bounded payoffs in $[0, 1]$ | SE | $\mathbb{P}\left[T \leq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{K}{\delta\Delta_k}\right)\right] \geq 1-\delta$ |
| | | | ME | $\mathbb{P}\left[T \leq \frac{K}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right] \geq 1-\delta$ |
| | Karnin et al. (2013) | bounded payoffs in $[0, 1]$ | EGE | $\mathbb{P}\left[T \leq \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{1}{\delta} \log\left(\frac{1}{\Delta_k}\right)\right)\right] \geq 1-\delta$ |
| | Jamieson et al. (2014) | sub-Gaussian noise | LILUCB | $\mathbb{P}\left[T \leq H_1 \log\left(\frac{1}{\delta}\right) + H_3\right] \geq 1 - 4\sqrt{c\delta} - 4c\delta$ |
| | Kaufmann et al. (2016) | two-armed Gaussian bandits | $\alpha$-E | $\mathbb{P}\left[T \leq \frac{(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2} \log\left(\frac{1}{\delta}\right)\right] \geq 1-\delta$ |
| | our work | finite $p$-th moments | SE-$\delta$(EA) | $\mathbb{P}\left[T \leq \sum_{k=1}^{K} \left(\frac{2^{2p+1}KC}{\Delta_k^p \delta}\right)^{\frac{1}{p-1}}\right] \geq 1-\delta$ |
| | | with $p \in (1, 2]$ | SE-$\delta$(TEA) | $\mathbb{P}\left[T \leq \sum_{k=1}^{K} \left(\frac{20B^{\frac{1}{p}}}{\Delta_k}\right)^{\frac{p}{p-1}} \log\left(\frac{2K}{\delta}\right)\right] \geq 1-\delta$ |
| fixed $T$ | Audibert and Bubeck (2010) | bounded payoffs in $[0, 1]$ | UCB-E | $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq TK \exp\left(-\frac{T-K}{H_1}\right)$ |
| | | | SR | $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq K(K-1) \exp\left(-\frac{T-K}{\overline{\log}(K)H_2}\right)$ |
| | Gabillon et al. (2012) | bounded payoffs in $[0, b]$ | UGapEb | $\mathbb{P}[\mu_{\text{Out}} - \mu_{\text{Opt}} \geq \epsilon] \leq TK \exp\left(-\frac{T-K}{H_\epsilon}\right)$ |
| | Karnin et al. (2013) | bounded payoffs in $[0, 1]$ | SH | $\mathbb{P}\left[\text{Out} \neq \text{Opt}\right] \leq \log(K) \exp\left(-\frac{T}{\log(K)H_2}\right)$ |
| | Kaufmann et al. (2016) | two-armed Gaussian bandits | SS | $\mathbb{P}\left[\text{Out} \neq \text{Opt}\right] \leq \exp\left(-\frac{(\mu_1-\mu_2)^2 T}{2(\sigma_1+\sigma_2)^2}\right)$ |
| | our work | finite $p$-th moments | SE-$T$(EA) | $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq 2^p CK(K-1)H_2^p \left(\frac{\bar{K}}{T-K}\right)^{p-1}$ |
| | | with $p \in (1, 2]$ | SE-$T$(TEA) | $\mathbb{P}[\text{Out} \neq \text{Opt}] \leq K(K-1) \exp\left(-\frac{(T-K)\bar{B}_1}{\bar{K}K\underline{\Delta}^{p/(1-p)}}\right)$ |

gence. We also notice that there are extensions of best arm identification of MAB, which is multiple-arm identification (Bubeck et al., 2013b; Chen et al., 2014).

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

# 3 Algorithms and Analyses

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

## 3.1 Empirical Estimations of Expected Payoffs

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

$$\hat{\mu}_t(k) \triangleq \frac{1}{s_{t,k}} \sum_{i \in \Phi(k)} \pi_i(k), \tag{2}$$

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

$$\hat{\mu}_t^{\dagger}(k) \triangleq \frac{1}{s_{t,k}} \sum_{i\in\Phi(k)} \pi_i(k)\mathbb{1}_{[|\pi_i(k)|\le b_i]}, \tag{3}$$

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

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

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

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

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

- *for EA, we have*

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

- *for TEA, we have*

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

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

$$\begin{aligned} \mathbb{P}[|\hat{\mu}_t(k) - \mu(k)| \ge \delta] &\le \frac{\mathbb{E}[|\hat{\mu}_t(k) - \mu(k)|^p]}{\delta^p} \\ &= \frac{\mathbb{E}[|\sum_{i\in\Phi(k)} \pi_i(k) - \mu(k)|^p]}{s_{t,k}^p\delta^p}, \end{aligned} \tag{4}$$

where $\delta \in (0,1)$.

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

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

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

$$|\hat{\mu}_t(k) - \mu(k)| \le \left(\frac{2C}{s_{t,k}^{p-1}\delta}\right)^{\frac{1}{p}}. \tag{5}$$

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

$$\mathbb{P}[|\hat{\mu}_t(k) - \mu(k)| \ge \delta] \le \frac{C_pC}{s_{t,k}^{p/2}\delta^p}, \tag{6}$$

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

$$|\hat{\mu}_t(k) - \mu(k)| \le \left(\frac{C_pC}{s_{t,k}^{p/2}\delta}\right)^{\frac{1}{p}}. \tag{7}$$

Now we prove the results with the estimator $\hat{\mu}_t^{\dagger}(k)$, where $k \in S_t$. Considering $b_i$ in Eq. (3), we define $\mu^{\dagger}(k) \triangleq \mathbb{E}\left[\pi_i(k)\mathbb{1}_{[|\pi_i(k)|\le b_i]}|\mathcal{F}_{i-1}\right]$, and $\zeta_i(k) \triangleq \mu^{\dagger}(k) - \pi_i(k)\mathbb{1}_{[|\pi_i(k)|\le b_i]}$, for any $i \in \Phi(k)$. We have $|\zeta_i(k)| \le 2b_i$, $\mathbb{E}[\zeta_i(k)|\mathcal{F}_{i-1}] = 0$ and $\mathbb{E}\left[\pi_i(k)\mathbb{1}_{[|\pi_i(k)|>b_i]}\right] \le B/b_i^{p-1}$. Besides, we also have

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

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

Based on Proposition 2, with probability at least $1-\delta$

$$\begin{aligned} &\left|\sum_{i\in\Phi(k)} \zeta_i(k)\right| \le 2b_t\log(2/\delta) + \frac{1}{2b_t}\sum_{i\in\Phi(k)} \mathbb{E}[\zeta_i^2(k)|\mathcal{F}_{i-1}] \\ &\le 2b_t\log(2/\delta) + s_{t,k}\frac{B}{2b_t^{p-1}}, \end{aligned} \tag{8}$$

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

$$|\hat{\mu}_t^{\dagger}(k) - \mu(k)| \le 5B^{\frac{1}{p}}\left(\frac{\log(2/\delta)}{s_{t,k}}\right)^{\frac{p-1}{p}}, \tag{9}$$

where we adopt the fact of

$$\frac{1}{s_{t,k}} \sum_{i \in \Phi(k)} \frac{B}{b_i^{p-1}} \leq 2B^{\frac{1}{p}} \left( \frac{\log(2/\delta)}{s_{t,k}} \right)^{\frac{p-1}{p}}. \tag{10}$$

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

$$\mathbb{E}[\zeta_i^2(k) | \mathcal{F}_{i-1}] \leq B^{\frac{2}{p}}. \tag{11}$$

By setting $p = 2$ and using Eq. (9), with probability at least $1 - \delta$, we have

$$|\hat{\mu}_t^{\dagger}(k) - \mu(k)| \leq 5B^{\frac{1}{p}} \left( \frac{\log(2/\delta)}{s_{t,k}} \right)^{\frac{1}{2}}, \tag{12}$$

which completes the proof. □

**Remark 1.** In (Bubeck et al., 2013a; Vakili et al., 2013), the Bernstein inequality without martingales is adopted with an implicit assumption of sampling payoffs of an arm being independent of sequential decisions, which is informal. By contrast, in Lemma 1, conditional on $\mathcal{F}_{t-1}$, the subset $S_t$ is fixed, and we adopt Bernstein inequality with martingales. Thus, we break the assumption of independent payoffs in previous work, and prove formal theoretical results of tail probabilities of estimators EA and TEA. Note that the superiority of martingales in sequential decisions has been fully discussed in (Zhao et al., 2016).

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

### 3.2 Pure Exploration with Fixed Confidence

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

#### 3.2.1 Description of SE-$\delta$

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

**Algorithm 1** Successive Elimination-$\delta$ (SE-$\delta$(TEA))

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

#### 3.2.2 Theoretical Guarantee of SE-$\delta$

We derive upper bounds of sample complexity of SE-$\delta$ with estimators of EA and TEA. We denote by $T$ the largest $t$ in SE-$\delta$.

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

- *for SE-$\delta$(EA)*

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

- *for SE-$\delta$(TEA)*

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

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

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

$k \in S_t$, we have

$$\mathbb{P}[|\hat{\mu}_t(k) - \mu(k)| \geq \delta] \leq \frac{2C}{t^{p-1}\delta^p}, \tag{13}$$

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

For any $t \in [T]$, with probability at least $1 - \delta/K$, the following event holds

$$\mathcal{E}_t \triangleq \left\{ k \in S_t, |\hat{\mu}_t(k) - \mu(k)| \leq \left(\frac{2KC}{t^{p-1}\delta}\right)^{\frac{1}{p}} \right\}.$$

To eliminate a sub-optimal arm $k$, we need to play any arm $k \in [K] \backslash \mathsf{Opt}$ with $t_k$ times such that

$$\hat{\Delta}_k \triangleq \hat{\mu}_{t_k}(\mathsf{Opt}) - \hat{\mu}_{t_k}(k) \geq 2\left(\frac{2KC}{t_k^{p-1}\delta}\right)^{\frac{1}{p}}. \tag{14}$$

Based on Lemma 1, with a high probability, we have

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

where $c_{t_k}$ is a confidence interval. To satisfy Eq. (14), we are ready to set

$$\Delta_k - 2c_{t_k} \geq 2\left(\frac{2KC}{t_k^{p-1}\delta}\right)^{\frac{1}{p}}. \tag{15}$$

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

$$T \leq \sum_{k=1}^K \left(\frac{2^{2p+1}KC}{\Delta_k^p\delta}\right)^{\frac{1}{p-1}}. \tag{16}$$

Now we consider TEA in Eq. (3) for estimating the expected payoffs in MAB. Similarly, for $p \in (1, 2]$, with probability at least $1 - \delta$, we have

$$T \leq \sum_{k=1}^K \left(\frac{20B^{1/p}}{\Delta_k}\right)^{\frac{p}{p-1}} \log\left(\frac{2K}{\delta}\right), \tag{17}$$

which completes the proof. □

### 3.3 Pure Exploration with Fixed Budget

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

**Algorithm 2** Successive Rejects-$T$ ($\mathsf{SR}$-$T$(TEA))

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

#### 3.3.1 Description of SR-$T$

For SR-$T$(EA), we omit the algorithm because it is almost the same as that in (Audibert and Bubeck, 2010). For SR-$T$(TEA), we design a bandit algorithm for pure exploration of MAB with heavy-tailed payoffs based on the idea of SR, with computation details shown in Algorithm 2, where $T$ is a given parameter. The high-level idea is to conduct non-uniform pulling of arms by $K - 1$ phases, and SR-$T$ rejects a worst empirical arm for each phase. The reject operation is based on EA or TEA, and we distinguish the two cases by SR-$T$(EA) and SR-$T$(TEA).

For simplicity, we show SR-$T$(TEA) in Algorithm 2, where $\underline{\Delta} \in (0, 1)$ is a design parameter for the estimator of TEA. The design parameter $\underline{\Delta}$ helps to calculate the truncating parameter $b$ in SR-$T$(TEA). Usually, we set $\underline{\Delta} \leq \Delta_k$ for any $k \in [K]$.

#### 3.3.2 Theoretical Guarantee of SR-$T$

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

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

- *for SR-T(EA)*

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

- *for SR-T(TEA)*

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

*where* $\bar{B}_1 = \frac{1}{4}\left[\left(\frac{\underline{\Delta}^p}{3Bp}\right)^{\frac{1}{p-1}} - \left(\frac{\underline{\Delta}^p}{3Bp^p}\right)^{\frac{1}{p-1}}\right]$.

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

$$\begin{aligned}
&\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq \sum_{k=1}^{K-1}\sum_{i=K+1-k}^{K} \mathbb{P}[\hat{\mu}_k(\mathsf{Opt}) \leq \hat{\mu}_k(i)] \\
&\leq \sum_{k=1}^{K-1}\sum_{i=K+1-k}^{K} \mathbb{P}[\hat{\mu}_k(i) - \mu(i) + \mu(\mathsf{Opt}) - \hat{\mu}_k(\mathsf{Opt}) \geq \Delta_i] \\
&\leq \sum_{k=1}^{K-1}\sum_{i=K+1-k}^{K} \frac{2C}{n_i^{p-1}\left(\frac{\Delta_i}{2}\right)^p} \qquad (18) \\
&\leq \sum_{k=1}^{K-1} \frac{2^{p+1}Ck}{n_k^{p-1}\Delta_{K+1-k}^p}, \qquad (19)
\end{aligned}$$

where the inequality of Eq. (18) is due to the results in Lemma 1 by setting $s_{t,k} = n_k$. Besides, we notice that

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

which implies that

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

Now we consider TEA in Eq. (3) for estimating the expected payoffs in MAB. For $p \in (1, 2]$, we have probability of error as

$$\begin{aligned}
&\mathbb{P}[\mathsf{Out} \neq \mathsf{Opt}] \leq \sum_{k=1}^{K-1}\sum_{i=K+1-k}^{K} \mathbb{P}[\hat{\mu}_k^{\dagger}(\mathsf{Opt}) \leq \hat{\mu}_k^{\dagger}(i)] \\
&\leq \sum_{k=1}^{K-1}\sum_{i=K+1-k}^{K} \mathbb{P}[\hat{\mu}_k^{\dagger}(i) - \mu(i) + \mu(\mathsf{Opt}) - \hat{\mu}_k^{\dagger}(\mathsf{Opt}) \geq \underline{\Delta}] \\
&\leq K(K-1)\exp\left(-\frac{(T-K)\bar{B}_1}{\bar{K}K\underline{\Delta}^{p/(1-p)}}\right), \qquad (20)
\end{aligned}$$

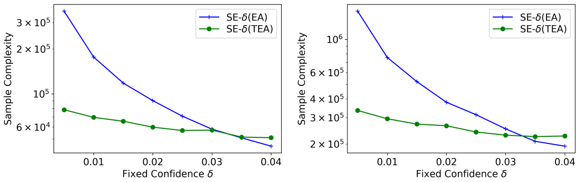
which completes proofs. □

# 4 Experiments

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

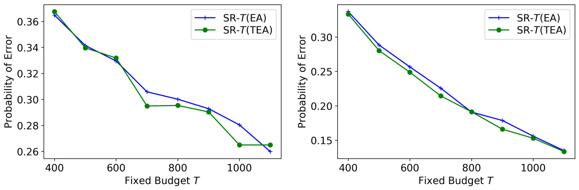
Table 2: Statistics of used synthetic data.

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



(a) S1 (b) S2

Figure 1: Sample complexity for SE-$\delta$ in pure exploration of MAB with heavy-tailed payoffs.



(a) S1 (b) S2

Figure 2: Probability of error for SR-$T$ in pure exploration of MAB with heavy-tailed payoffs.

## 4.1 Synthetic Data and Setting

For verifications, we adopt two synthetic data (named as S1-S2) in the experiments, of which statistics are shown in Table 2. The data are generated from *Student's t-Distribution* with 3 degrees of freedom. In experiment, we run multiple epochs of experiments, with each epoch containing ten independent experiments for best arm identification of MAB. Besides, we set the value of fixed confidence from 0.005 to 0.040 with a uniform gap of 0.005. We set the value of fixed budget from 400 to 1100 with a uniform gap of 100.

We show experimental results in Figures 1 and 2, where both proposed algorithms are effective for pure exploration of MAB with heavy-tailed payoffs. In particular, in fixed-confidence setting, sample complexity decreases with increasing value of $\delta$. In fixed-budget setting, probability of error converges to zero with increasing value of $T$. Besides, for fixed-confidence setting, SE-$\delta$(TEA) beats SE-$\delta$(EA) in both datasets with small $\delta$ due to a better control of confidence interval. The experimental results also reflect that the concentration properties of EA are much weaker than those of TEA. For fixed-budget setting, SR-$T$(TEA) is comparable to SR-$T$(EA) due to

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

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

Table 4: Estimated parameters for ten cryptocurrencies.

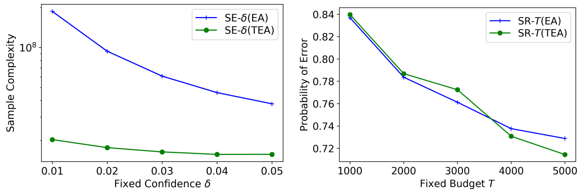
| symbol | degree of freedom | $(p, B, C)$ in experiments |
|---|---|---|
| BTC | 3.50 | (2,1.577×$10^{-3}$,1.575×$10^{-3}$) |
| ETC | 3.81 | |
| XRP | 2.53 | |
| BCH | 3.00 | |
| EOS | 2.90 | |
| LTC | 2.75 | |
| ADA | 3.55 | |
| XLM | 3.81 | |
| IOT | 4.66 | |
| NEO | 3.13 | |

the selection of truncating parameter.

### 4.2 Financial Data and Setting

It has been pointed out that financial data show the inherent characteristic of heavy tails (Panahi, 2016). We choose a financial application of identifying the most profitable cryptocurrency in a given pool of digital currencies. The identification for the most profitable cryptocurrency among the top ten cryptocurrency in terms of market value is motivated by the practical scenario that an investor would like to invest a fixed budget of money in a cryptocurrency and get return as much as possible.

For experiments, we get hourly price data of the ten selected cryptocurrencies[1], and show the statistics of real data in Table 3. In the table, we conduct a statistical analysis in hindsight with hourly returns of cryptocurrency from February 3rd, 2018 to April 27th, 2018. From the table, we find that the optimal option in hindsight is EOS in terms of the maximal empirical mean of hourly payoffs. Besides, we conduct Kolmogrov-Smirnov (KS) test to fit real data of a cryptocurrency to a distribution. In particular, via KS test, we know that the null hypothesis of real data following a Gaussian distribution is rejected, because $\bar{p}$-value is smaller than a significant level of 0.05. We observe that real data of cryptocurrency are likely to follow a *Student's t distribution* via KS test in Table 3.



(a) fixed confidence (b) fixed budget

Figure 3: Pure exploration of cryptocurrency.

With the above statistical analyses, we can fit real data of cryptocurrency to a *Student's t distribution*, and obtain distribution parameters shown in Table 4. Based on the property of *Student's t distribution*, we can choose $p = 2$ and estimate $B$ and $C$ shown in the table.

By setting a similar experimental setting of synthetic data, we show the results on pure exploration of top ten cryptocurrencies in Figure 3. Note that, due to limitation of data points in the setting of fixed confidence, we generate payoffs from *Student's t distributions* fitting to real data. But in the setting of fixed budget, we adopt exactly real financial data. We have similar observations as those in synthetic data. It is worth mentioning that, TEA algorithm outperforms EA algorithm in fixed-confidence setting when the value of $\delta$ is small. Besides, TEA is comparable to EA in fixed-budget setting because the truncated parameter in Algorithm 2 only has budget information and does not increase with the number of samples. Overall, with synthetic and real-world data, we have verified the effectiveness of our two algorithms.

## 5 Conclusion

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

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

## Acknowledgments

The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14208815 and No. CUHK 14234416 of the General Research Fund), and Microsoft Research Asia (2018 Microsoft Research Asia Collaborative Research Award).

## References

J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *COLT*, pages 13–p, 2010.

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

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

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *ALT*, pages 23–37. Springer, 2009.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59 (11):7711–7717, 2013a.

S. Bubeck, T. Wang, and N. Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265, 2013b.

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

S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. In *NIPS*, pages 379–387, 2014.

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, pages 208–214, 2011.

S. Dharmadhikari, V. Fabian, and K. Jogdeo. Bounds on the moments of martingales. *The Annals of Mathematical Statistics*, pages 1719–1723, 1968.

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

V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS*, pages 3212–3220, 2012.

V. Gabillon, A. Lazaric, M. Ghavamzadeh, R. Ortner, and P. Bartlett. Improved learning complexity in combinatorial pure exploration bandits. In *Artificial Intelligence and Statistics*, pages 1004–1012, 2016.

K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *CISS*, pages 1–6. IEEE, 2014.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

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

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.

N. Korda, B. Szörényi, and L. Shuai. Distributed clustering of linear bandits in peer to peer networks. In *ICML*, volume 48, pages 1301–1309. International Machine Learning Societ, 2016.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *ICML*, pages 767–776, 2015.

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

T. Lattimore, K. Crammer, and C. Szepesvári. Linear multi-resource allocation with semi-bandit feedback. In *NIPS*, pages 964–972, 2015.

S. Li, A. Karatzoglou, and C. Gentile. Collaborative filtering bandits. In *SIGIR*, pages 539–548. ACM, 2016.

J. Liebeherr, A. Burchard, and F. Ciucu. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2):1010–1024, 2012.

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

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

H. Panahi. Model selection test for the heavy-tailed distributions under censored samples with application in financial data. *International Journal of Financial Studies*, 4(4):24, 2016.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58 (5):527–535, 1952.

A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *NIPS*, pages 3275–3283, 2012.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

S. 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.

B. von Bahr, C.-G. Esseen, et al. Inequalities for the $r$-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.

Q. Wu, H. Wang, Q. Gu, and H. Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538. ACM, 2016.

S. Zhao, E. Zhou, A. Sabharwal, and S. Ermon. Adaptive concentration inequalities for sequential decision problems. In *NIPS*, pages 1343–1351, 2016.