Thompson Sampling on Asymmetric α-stable Bandits

Zhendong Shi, Ercan E Kuruoglu and Xiaoli Wei
Data Science and Information Technology Center, Tsinghua-Berkeley Shenzhen Institute, Tsinghua University SIGS, China

Keywords: Thompson Sampling, Multi-Armed Bandit Problem, Asymmetric Reward, Reinforcement Learning, α-stable Distribution.

Abstract: In algorithm optimization in reinforcement learning, how to deal with the exploration-exploitation dilemma is particularly important. Multi-armed bandit problem can be designed to realize the dynamic balance between exploration and exploitation by changing the reward distribution. Thompson Sampling has been proposed in the literature for the solution of the multi-armed bandit problem by sampling rewards from posterior distributions. Recently, it was used to process non-Gaussian data with heavy tailed distributions. It is a common observation that various real-life data such as social network data and financial data demonstrate not only impulsive but also asymmetric characteristics. In this paper, we consider the Thompson Sampling approach for multi-armed bandit problem, in which rewards conform to an asymmetric α-stable distribution with unknown parameters and explore their applications in modelling financial and recommendation system data.

1 INTRODUCTION

Sequential decision-making plays a key role in many fields, such as quantitative finance and robotics. In order to make real-time decisions under unknown environments, decision makers must carefully design algorithms to balance the trade-off between exploration and exploitation. Many decision algorithms have been designed and widely used, such as financial decision-making (Shailesh, 2015) and personalized news recommendation (Li et al., 2010).

The multi-armed bandits (MAB) have an important potential in solving the dilemma of exploration and exploitation in the sequential decision-making problem in which a fixed limited set of resources must be allocated between competing (alternative) choices in a way that maximizes their expected gain. Different data may require different reward distribution. Over the years, various reward distribution functions ranging from Bernoulli distribution and Gaussian distribution to sub-exponential family, have been proposed and corresponding fast processing algorithms such as UCB-Rad (Jia et al., 2021) (specifically for MAB with sub-exponential rewards) have been developed.

However, when we design decision-making algorithms for complex systems, the reward distribution function (such as Bernoulli distribution and Gaussian distribution) is inconsistent with the probability distribution which each arm obeys. According to the research on these complex system data, one can observe that interactions often lead to heavy-tailed or power law distributions (Lehmkuhl and Promies, 2020). When dealing with practical problems, we find that many data (e.g. financial data (Embrechts et al., 2003) and social mobile traffic data (Qi et al., 2016)) have characteristics such as heavy tail and negative skewness, which cannot be perfectly described by the Gaussian distribution. These deviations from Gaussian distribution to more complex and practical reward distributions allow for the opportunity to develop significantly more efficient algorithms than were possible in the original setting as long as we capture the correct reward distribution in various real world applications.

Existing machine learning algorithms find it difficult to deal with the problem of multi armed bandits with complicated reward distributions. This is because the probability density of such reward distributions can not be obtained analytically. When real data has characteristics such as heavy tails or asymmetry, the standard algorithms which make conservative statistical assumptions lead to the choice of wrong arms.

In the past few years, there have been a number of studies on the MAB problem with heavy tailed distributions (Liu and Zhao, 2011),(Bubeck et al., 2013). Compared with algorithms that are optimized through repeated trial and error tuning parameters (such as the *epsilon*-greedy algorithm and UCB algorithm), the consideration of heavy-tail distributions and the devel-

Shi, Z., Kuruoglu, E. and Wei, X.
Thompson Sampling on Asymmetric a-stable Bandits.
DOI: 10.5220/0011684200003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 434-441
ISBN: 978-989-758-623-1; ISSN: 2184-433X

opment of algorithms for MAB problems involving heavy tailed data has provided a much more realistic framework leading to higher performance. These work (e.g. (Lee et al., 2020)) investigated the best arm identification of MAB with a general assumption that p-th moments of stochastic rewards, analyzed tail probabilities of average and proposed different bandit algorithms, such as deterministic sequencing of exploration and exploitation (Liu and Zhao, 2011) and truncated empirical average method (Yu et al., 2018).

Instead of tail probabilities, (Dubey and Pentland, 2019) proposed an algorithm based on the symmetric α-stable distribution and demonstrated the success with accurate assumptions and normalized iterative process. α-stable distributions is a family of distributions with power law heavy tails, which can provide with a better reward distribution (Dubey and Pentland, 2019) and can be applied to the exploration of complex systems. This family of distributions stand out among rival non-Gaussian models (Chen et al., 2016) since they satisfy the generalised central limit theorem. α-stable distributions have become state of the art models for various real data such as financial data (Embrechts et al., 2003), sensor noise (Nguyen et al., 2019), radiation from Poisson field of point sources(Win et al., 2009), astronomy data (Herranz et al., 2004), and electric disturbances on power lines (Karakuş et al., 2020).

Motivated by the presence of asymmetric characteristics in various real life data (Kuruoglu, 2003) and the success in reinforcement learning and other directions due to the introduction of asymmetry (Baisero and Amato, 2021), in this work, we propose a statistic model, for which the reward distribution is both heavy-tailed and asymmetric, named asymmetric alpha-Thompson sampling algorithm.

2 BACKGROUND INFORMATION

The multi-armed bandit (Auer et al., 2002) is a theoretical model that has been widely used in machine learning and optimization, and various algorithms have been proposed for optimal solution when the reward distributions are Gaussian-distribution or exponential-distribution (Korda et al., 2013). However, these reward distributions do not hold for those complex systems with impulsive data. For example, when we model stock prices or deal with behaviour in social networks, the interactive data often lead to heavy tail and negative skewness (Oja, 1981).

2.1 Multi-Armed Bandit Problem

Suppose that there are several slot machines available for an agent, who can select, for each round one to pull and record the rewards. Assuming that each slot machine is not exactly the same, after multiple rounds of operation, we can mine some statistical information of each slot machine, and then select the slot machine that gives the expected highest reward.

Learning is carried out in rounds and indexed by $t \in [T]$. The total number of rounds called time range T is known in advance. This problem is iterative, the agent picks arm $a_t \in [N]$ and then observes reward $r_{a_t}(t)$ from that arm in each round of $t \in [T]$. For each arm $n \in [N]$, rewards independently come from a distribution D_n with mean $\mu_n = E_{D_n}[r]$. The largest expected reward is denoted by $\mu^\star = \max_{n\in[N]} \mu_n$, and the corresponding arm(s) is(are) denoted as the optimal arm(s) n^*.

To quantify the performance, the regret $R(T)$ is used, which refers to the difference between the ideal total reward the agent can achieve and the total reward the agent actually gets.

$$R(T) = \mu^\star T - \sum_{t=0}^{T} \mu_{a_t}. \tag{1}$$

2.2 Thompson Sampling Algorithm for Multi-Armed Bandit Problem

There are various exploration algorithms, including the ε-greedy algorithm, UCB algorithm, and Thompson sampling. ε-greedy algorithm(Korte and Lovász, 1984) uses both exploitations to take advantage of prior knowledge and exploration to look for new options, while the UCB algorithm(Cappé et al., 2013) simply pulls the arm that has the highest empirical reward estimate up to that point plus some term which is inversely proportional to the number of times the arm has been played.

Assuming that for each arm $n \in [N]$, the reward distribution is D_n parameterized by $\theta_n \in \Theta$ (μ_n may not be an appropriate parameter) and that the parameter has a prior probability distribution $p(\theta_n)$, Thompson sampling algorithm updates the prior distribution of θ_n based on the observed reward for the arm n, and then selects the arm based on the derived posterior probability of the reward under the arm n.

According to the Bayes rule, $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta')p(\theta')d\theta'}$, where θ is the model parameter and x is the observation. $p(\theta|x)$ is the posterior distribution, $p(x|\theta)$ is likelihood function, $p(\theta)$ is the prior distribution and $p(x)$ is the evidence.

For each round $t \in [T]$, the agent draws the parameter $\hat{\theta}_n(t)$ for each arm $n \in [N]$ from the posterior distribution of parameters given the previous rewards up to time $t-1$, $\boldsymbol{r}_n(t-1) = \{r_n^{(1)}, r_n^{(2)}, \cdots, r_n^{(k_n(t-1))}\}$, where $k_n(t)$ is the number of the arm n that has been pulled up at time t:

$$\hat{\theta}_n(t) \sim p(\theta_n|\boldsymbol{r}_n(t-1)) \propto p(\boldsymbol{r}_n(t-1)|\theta_n)p(\theta_n). \quad (2)$$

Given the drawn parameters $\hat{\theta}_n$(t) of each arm, the agent selects the arm a_t with the largest average return on the posterior distribution, receives the return r_{a_t}, and then updates the posterior distribution of the arm action a_t.

$$a_t = \arg\max_{n\in[N]} \mu_n(\theta_n(t)) \quad (3)$$

We will use the Bayesian Regret (Russo and Van, 2014) for the measurement of the performance in order to compare with the symmetric case. Bayesian Regret (BR) is the estimated regret over the priors. Denoting the parameters over all arms as $\overline{\theta} = \{\theta_1, ..., \theta_N\}$ and their corresponding product distribution as $\overline{D} = \prod_i D_i$, the Bayesian Regret is expressed as:

$$BR(T, \pi) = E_{\overline{\theta}\sim\overline{D}}[R(T)] \quad (4)$$

2.3 α-stable Distribution

An important generalization of the Gaussian distribution is the α-stable distribution, which is often used to model both impulsive and skewed data. It has a non-analytic density and therefore, usually is described with the characteristic function. We say a random variable X is $S_\alpha(\sigma, \beta, \delta)$ if X has characteristic function

$$\begin{aligned}\mathbb{E}[e^{iuX}] = &\exp(-\sigma^\alpha |u|^\alpha (1 + i\beta \text{sign}(u) \\ &(|\sigma u|^{1-\alpha} - 1)) + iu\delta)\end{aligned} \quad (5)$$

where α is the characteristic exponent defining the impulsiveness of the distribution, the parameter β corresponds to the skewness, γ is the scale parameter and δ is the location parameter. (See (Embrechts et al., 2003)).

2.4 Symmetric α-thompson Sampling

In MAB with symmetric α-stable reward distributions, the corresponding reward distribution for each arm n is given by $D_n = S_\alpha(\sigma, \beta = 0, \delta_n)$, where $\alpha \in (1,2)$, $\sigma \in \mathbb{R}^+$ are known in advance, and δ_n is unknown ((Dubey and Pentland, 2019)). In this case, $\mathbb{E}[r_n] = \delta_n$. They set a prior Gaussian distribution $p(\delta_n)$ over the parameter δ_n. Since the only unknown parameter for the reward distributions is δ_n, D_n is parameterized by $\theta_n = \delta_n$.

(Dubey and Pentland, 2019) propose two algorithms for Thompson Sampling. One is called Symmetric α-Thompson Sampling, which is based on the scale mixtures of normals (SMiN) representation. The other is called robust Symmetric α–Thompson Sampling. It is similar to the basic α-Thompson sampling, except for rejecting a reward when the received reward exceeds the threshold. This strategy yields a tighter regret bound than the basic Symmetric α-Thompson Sampling. These algorithms, however, do not apply for asymmetric α-stable MABs since SMiN representation does not hold.

3 ASYMMETRIC α-THOMPSON SAMPLING

Both our algorithm and symmetric α-Thompson sampling algorithm are constructed under the framework of Thompson algorithm. The biggest difference is the assumed reward distribution. Our corresponding reward distribution for each arm n is given by $D_n = S_\alpha(\sigma, \beta, \delta_n)$, where $\alpha \in (1,2)$.

Suppose x is observed data, δ_n is the unknown parameter, we can obtain the posterior density for δ_n from prior distribution through the equation (2). However, as x is assumed to conform to α-stable distribution, density function $f(x|\delta_n)$ is unavailable. (Dubey and Pentland, 2019) take advantage of the symmetry of the distribution and achieve the iterative process through scale mixtures of normals representation. This method is not applicable when β is not equal to 0.

We solve the sampling from the posterior density problem through Gibbs sampling, which requires obtaining conditional distribution. However, it is challenging to obtain the conditional distribution of α-stable distribution which was circumvented by introducing an auxiliary variable leading to a decomposition proposed by (Buckle, 1995).

Theorem 1. *Let* $f : (-\infty, 0) \times (-\frac{1}{2}, l_{\alpha,\beta}) \cup (0, \infty) \times (l_{\alpha,\beta}, \frac{1}{2}) \to (0, \infty)$ *be the bivariate probability density function of* $\hat{X}$ *and* $\hat{Y}$*, conditional on* α*,* β*,* σ *and* δ*.*

$$\begin{aligned}f(x, y|\alpha, \beta, \sigma, \delta) = &\frac{\alpha}{|\alpha - 1|} \exp\{-|\frac{z}{t_{\alpha,\beta}(y)}|^{\frac{\alpha}{\alpha-1}}\} \\ &\times |\frac{z}{t_{\alpha,\beta}(y)}|^{\frac{\alpha}{\alpha-1}} \frac{1}{|z|}\end{aligned} \quad (6)$$

where $z = \frac{x-\delta}{\sigma}$, $\eta_{\alpha,\beta} = \frac{\beta(2-\alpha)\pi}{2}$, $l_{\alpha,\beta} = -\frac{\eta_{\alpha,\beta}}{\pi\alpha}$.

$$\begin{aligned}t_{\alpha,\beta}(y) = &\Big(\frac{\sin[\pi\alpha y + \eta_{\alpha,\beta}]}{\cos[\pi y]}\Big) \times \\ &\Big(\frac{\cos \pi y}{\cos[\pi(\alpha - 1)y + \eta_{\alpha,\beta}]}\Big)^{\frac{\alpha}{\alpha-1}}.\end{aligned} \quad (7)$$

Then f is a proper bivariate probability density for distribution of (X,Y), and marginal distribution of X is $S_\alpha(\sigma,\beta,\delta)$.

Now we are ready to study Bayesian inference for the arm $n \in [N]$. Suppose that at time t, the arm n has been pulled for $k_n(t)$ times and hence we have $k_n(t)$ vectors of rewards $\boldsymbol{r}_n(t) = \{r_n^{(1)}, \cdots, r_n^{k_n(t)}\}$. According to Theorem 1 and the Bayesian rule, we derive the posterior density of δ_n conditional on $\boldsymbol{r}_n(t)$ by the following equation(Buckle, 1995):

$$p(\alpha,\beta,\sigma,\delta_n|\boldsymbol{r}_n(t)) \propto \int (\frac{\alpha}{|\alpha-1|\sigma})^{k_n(t)} \exp\Big\{-\sum_{i=1}^{k_n(t)} |\frac{z_i}{t_{\alpha,\beta}(y_i)}|^{\frac{\alpha}{\alpha-1}}\Big\} \times \prod_{i=1}^{k_n(t)} \Big|\frac{z_i}{t_{\alpha,\beta}(y_i)}\Big|^{\frac{\alpha}{\alpha-1}} \frac{1}{|z_i|} \times p(\alpha,\beta,\sigma,\delta_n) dy \tag{8}$$

where $z_i = \frac{r_n - \delta_n}{\sigma}$ and $p(\delta_n)$ is the prior distribution for δ_n. We can simplify the formula further as α,β,σ are known.

$$p(\delta_n|\alpha,\beta,\sigma,\boldsymbol{r}_n,\boldsymbol{y}_n) \propto \exp\{-\sum_{i=1}^{k_n(t)} |\frac{z_i}{t_{\alpha,\beta}(y_i)}|^{\alpha/(\alpha-1)}\} \times \prod_{i=1}^{k_n(t)} |\frac{z_i}{t_{\alpha,\beta}(y_i)}|^{\alpha/(\alpha-1)} \times \frac{1}{|z_i|} p(\delta_n) \tag{9}$$

Through this formula, we have completed the method to obtain the posterior distribution under the assumption of asymmetric α-stable distribution.

In our algorithm we first estimate parameters α,β,σ and choose normal distribution as the prior distribution of δ. Suppose that we have a model driven by the parameter vector $(\alpha,\beta,\delta,\sigma)$, and that we have observed $x = (x_1, x_2, ..., x_n)$. By taking a set of starting values we can generate μ^1 from $\pi(\delta|\alpha^0,\beta^0,\sigma^0,x)$, α^1 from $\pi(\alpha|\delta^1,\beta^0,\sigma^0,x)$, and so on continuing to other parameters thereby performing one iteration producing the sample $(\alpha^1,\beta^1,\delta^1,\sigma^1)$. The prior distribution of σ is also taken to be a Gaussian distribution, while the prior distributions of α,β are chosen to follow beta distribution.

The conditional distributions of α-stable parameters are obtained as follows:

$$p(\alpha_n|\delta,\beta,\sigma,\boldsymbol{r}_n) \propto (\frac{\alpha}{|\alpha-1|})^n \exp(-\sum_{i=1}^{k_n(t)} |\frac{z_i}{v_i}|^{\frac{\alpha}{\alpha-1}}) \times \prod_{i=1}^{k_n(t)} |\frac{z_i}{v_i}|^{\frac{\alpha}{\alpha-1}} |\frac{dt_{\alpha,\beta}}{dy}|^{-1}_{t_{\alpha,\beta}(y_i)=\Phi_i(r_n^{(i)}-\delta_n)} p(\alpha_n) \tag{10}$$

$$p(\beta_n|\alpha,\delta,\sigma,\boldsymbol{r}_n) \propto |\frac{dt_{\alpha,\beta}}{dy}|^{-1}_{t_{\alpha,\beta}(y_i)=\Phi_i(r_n^{(i)}-\delta_n)} p(\beta_n) \tag{11}$$

$$p(\sigma_n|\alpha,\delta,\beta,\boldsymbol{r}_n) \propto |\frac{dt_{\alpha,\beta}}{dy}|^{-1}_{t_{\alpha,\beta}(y_i)=\Phi_i(r_n^{(i)}-\delta_n)} p(\sigma_n) \tag{12}$$

Algorithm 1: Asymmetric α-Thompson Sampling.

Input: Arms n ∈ [N], priors α,β,σ for each arm, auxiliary variable y

estimate α,β,σ by empirical characteristic function method and deduce prior distribution $p(\delta)$
for each arms $n \in$ [N] **do**
  for each iteration $t \in [k_n(t)]$ **do**
    draw $\delta_n(t)$ from prior distribution
    Generate u from a Uniform(0,1)
    If $u < p(\hat{\delta}_n(t)|\alpha,\beta,\sigma,\boldsymbol{r}_n(t)) \times p(\delta_n(t)|\hat{\delta}_n(t)) / (p(\delta_n(t)|\alpha,\beta,\sigma,\boldsymbol{r}_n(t)) p(\hat{\delta}_n(t)|\delta_n(t)))$ then $\delta_n(t+1) = \hat{\delta}_n(t)$; otherwise, $\delta_n(t+1) = \delta_n(t)$
    choose the arm that maximizes the reward $r_n^{(t)}$
    Update distribution $p(\delta_n(t+1))$ by (9)
    Update distribution $p(\alpha_n(t+1))$ by (10)
    Update distribution $p(\beta_n(t+1))$ by (11)
    Update distribution $p(\sigma_n(t+1))$ by (12)

3.1 Regret Analysis

Bayesian Regret. In this section, we provide a formula for the Bayesian Regret (BR) incurred by the asymmetric α-Thompson Samplings algorithm.

In order to simplify the calculation formula, we introduce the upper bound confidence and lower bound confidence to show the Bayesian Regret. We generalize the upper and lower confidence bounds on an arm's expected rewards at a certain time t (given history H_t): respectively, $U(a,H_t)$ and $L(a,H_t)$. There are two properties we want these functions to have, for some $\gamma > 0$ to be specified later:

$$\forall a,t \quad \mathbb{E}[[U(a,H_t) - \mu(a,t)]^-] \le \gamma \tag{13}$$

$$\forall a,t \quad \mathbb{E}[[\mu(a,t) - L(a,H_t)]^-] \le \gamma \tag{14}$$

Assuming we have lower and upper bound functions that satisfy those two properties, the Bayesian Regret of Thompson sampling can be bounded as follows:

$$BR(T) \le 2\gamma \times T \times N + \sum_{t=1}^{T} \mathbb{E}[[U(a,H_t) - L(a,H_t)]] \tag{15}$$

Theorem 2. *Let $N > 1$, $\alpha \in (1,2)$,$\sigma \in \mathbb{R}^+$. Assume that $\delta_{n\in[N]}$ is uniformly bounded by $M > 0$, i.e. $\delta_{n\in[N]} \in [-M,M]$. Then for a N-armed bandit with rewards for each arm n independently drawn from $S_\alpha(\beta,\sigma,\delta_n)$, for ε chosen a priori such that $\varepsilon \to (\alpha-1)^-$,*

$$BR(T,\pi^{TS}) = O(N^{\frac{1}{1+\varepsilon}} T^{\frac{1+\varepsilon}{1+2\varepsilon}}) \tag{16}$$

Proof. For all heavy-tailed distributions such as α-stable distributions, variance does not exist, so we need to build our own controllable σ. The key difference between our BayesRegret and the results from (2) lies in the moments of α-stable distribution. We have for $X \sim S_\alpha(\sigma,0,0)$, $\varepsilon \in (0,\alpha-1)$

$$E[|X|^{(1+\varepsilon)}] = C(1+\varepsilon,\alpha)|\sigma|^{(1+\varepsilon)/\alpha} \tag{17}$$

where $C((1+\varepsilon),\alpha) = \frac{2^\varepsilon \Gamma(\frac{\varepsilon}{2})\Gamma(-\frac{(1+\varepsilon)}{\alpha})}{\alpha\sqrt{\pi}\Gamma(-\frac{(1+\varepsilon)}{2})}$. while $\beta \neq 0$, by Proposition 1 in (Kuruoglu, 2001), we have for $X \sim S_\alpha(\sigma,\beta,0)$, $p \in (0,\alpha)$.

$$E[|X|^{(1+\varepsilon)}] = C((1+\varepsilon),\alpha,\beta)|\sigma|^{(1+\varepsilon)/\alpha} \tag{18}$$

where

$C((1+\varepsilon),\alpha,\beta) = \frac{\Gamma(1-\frac{(1+\varepsilon)}{\alpha})}{\Gamma((-\varepsilon))}\left|\frac{1}{\cos\theta}\right|^{(1+\varepsilon)/\alpha}$
$\times \frac{\cos(\frac{(1+\varepsilon)\theta}{\alpha})}{\cos(\frac{(1+\varepsilon)\pi}{2})}$ and $\theta = \arctan(\beta\tan\frac{\alpha\pi}{2})$.

Let $x_1, x_2, ..., x_n$ be a real i.i.d. sequence with finite mean μ. Assume for some $\varepsilon \in (0,1]$, $v \geq 0$ and $u \geq 0$, one has $\mathbb{E}[|X-\mu|^{1+\varepsilon}] \leq v$ and $\mathbb{E}[|X|^{1+\varepsilon}] \leq u$.

Let $\hat{\mu}$ be empirical mean, then for any $\delta \in (0,1)$, with probability at least $1-\delta$. One has $\hat{\mu} \leq \mu + (\frac{3*v}{\delta*n^\varepsilon})^{\frac{1}{1+\varepsilon}}$ (Bubeck et al., 2013).

Thus, through the definition of upper bound confidence, lower bound confidence and γ, we obtained $\gamma \leq 2*NM*\delta*T$ where $|U(a,H_t)| \leq M$.

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[[U(a,H_t)-L(a,H_t)]] &\leq 2*\mathbb{E}[\sum_{n=1}^{N}\sum_{t=1}^{T}\mathbb{I}[A_t=k] \\ &(\frac{3*v}{\delta*n^\varepsilon})^{\frac{1}{1+\varepsilon}}] \\ &\leq 2*(\frac{3*C(1+\varepsilon,\alpha,\beta)}{\delta*n^\varepsilon})^{\frac{1}{1+\varepsilon}} \\ &\mathbb{E}[\sum_{n=1}^{N}\int_{s=0}^{k_n(T)}(\frac{1}{s^\varepsilon})^{1+\varepsilon}ds] \\ &= 2(1+\varepsilon)(\frac{3C(1+\varepsilon,\alpha,\beta)}{\delta*n^\varepsilon})^{\frac{1}{1+\varepsilon}} \\ &*(NT)^{\frac{1}{1+\varepsilon}}\end{aligned} \tag{19}$$

$$\begin{aligned}BR(T,\pi^{TS}) \leq 4\Big(\frac{3*C(1+\varepsilon,\alpha,\beta)}{\delta}\Big)^{\frac{1}{1+\varepsilon}} \\ (NT)^{\frac{1}{1+\varepsilon}} + 4NMT^2\delta,\end{aligned} \tag{20}$$

where $\delta \in (0,1)$. By choosing suitable δ, we obtain the desired equation (16).

□

In particular, when $\beta = 0$ or $\beta = \pm 1$.

$$\begin{aligned}BR(T,\beta=0) \leq 4\Big(\frac{3*\frac{\Gamma(1-\frac{p}{\alpha})}{\Gamma(1-p)}\frac{1}{\cos(\frac{p\pi}{2})}}{\delta}\Big)^{\frac{1}{1+\varepsilon}}(NT)^{\frac{1}{1+\varepsilon}} \\ +4NMT^2\delta\end{aligned} \tag{21}$$

$$\begin{aligned}BR(T,\beta=\pm1) \leq 4\Big(\frac{3*\frac{\Gamma(1-\frac{p}{\alpha})}{\Gamma(1-p)}(\frac{1}{\cos(\frac{p\pi}{2})})^{p/\alpha}}{\delta}\Big)^{\frac{1}{1+\varepsilon}} \\ (NT)^{\frac{1}{1+\varepsilon}} + 4NMT^2\delta\end{aligned} \tag{22}$$

where (21) is consistent with the results obtained in the symmetric case.

Although the skewness parameter β has an impact on the regret bound, it does not change the upper confidence bound of the regret bound, which is shown in (16).

4 EXPERIMENTAL STUDIES

In order to show the efficiency and stability of the asymmetric α-TS algorithm in a specific data field, we will use the ε-greedy algorithm, bootstrapped UCB algorithm, symmetry α-TS algorithm and asymmetric algorithm in different data sets for comparative experiments.

To test the efficiency of asymmetric α-TS algorithm, we use synthetic α-stable data, stock prices data and recommendation data as detailed next.

4.1 Synthetic Asymmetric α-stable Data

We generated a simulated data set with 100 arms. We generated $x_{t,a} \in R$ from alpha-stable distributions for all arms a. The true parameters were firstly simulated from an alpha-stable distribution with mean 0. The resulting reward associated with the optimal arm was 0.994 and the mean reward was 0.195. We averaged the experiments over 100 runs.

This asymmetric data set is generated using the Chambers-Mallows-Stuck algorithm (Weron, 1996). We conducted multi-armed bandit experiments with following benchmarks – (i) an ε-greedy agent, (ii) Boostrapped-UCB agent, (iii) Symmetric α-TS and (iv) Asymmetric α-TS. The average value of each arm is randomly selected, where $\alpha = 1.3$ and $\sigma = 500$ respectively of each experiment.

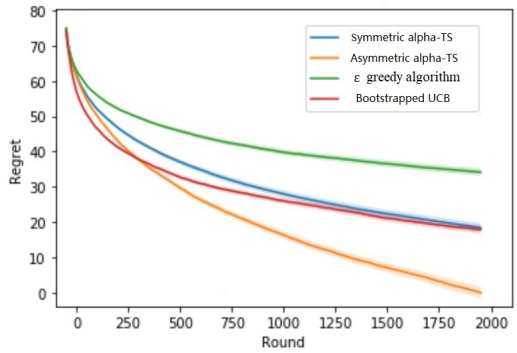

Figure 1: Regret for asymmetric data, the green line is greedy strategy, blue one is common alpha-TS method, red one is Bootstrapped UCB while orange line shows our asymmetric α-TS method.

The test results of the asymmetric data which are shown in Figure 1 meet our expectations, and the symmetric algorithm is worse than our method in time and space efficiency. Under the assumption that the return distribution conforms to the asymmetric α-stable distribution, we obtain reward each iteration independently come from the reward distribution.

4.2 Stock Selection

In this experiment, 100 shares listed in Shenzhen Stock Exchange through Tushare using Python had been chosen as risk assets, and the stock codes are from 000010.SZ to 300813.SZ. We choose the closed stock price from 2016/07/01 to 2020/07/18.

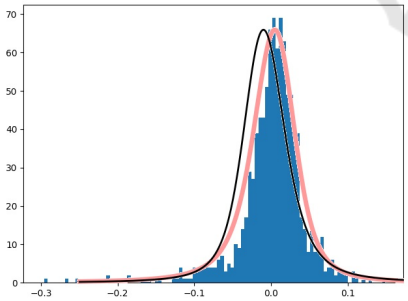

Figure 2: Stock price data, the blue histogram represents the dataset, the black curve represents the fitted symmetric α stable distribution and the red one represents the fitted asymmetric α stable distribution.

In the financial field, reward distribution can be regarded as the distribution of return on each stock. In each iteration, we get the parameters that are more consistent with the actual distribution under the assumption conditions and the maximum arm obtained by sampling. The regret tells the difference between the ideal total reward we can achieve and the total reward we actually gets.

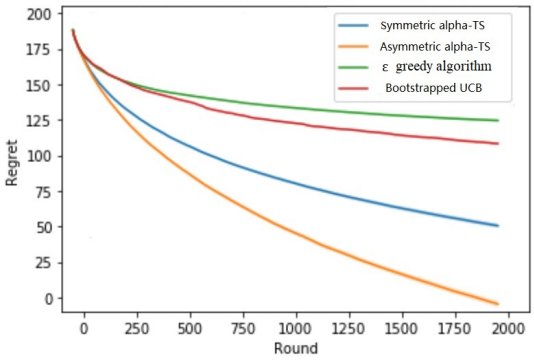

Figure 3: regret for stock price data, the green line is greedy strategy, blue one is common alpha-TS method, red one is Bootstrapped UCB while orange line shows our asymmetric α-TS method.

Table 1: Comparison of Different Strategies.

Strategy	AR	SR	MaxD
$\varepsilon - greedy$	3.36%	9.2%	3.48%
$Boostrapped - UCB$	6.47%	8.3%	6.35%
$Symmetric - TS$	7.76%	17.8%	**3.39%**
$Asymmetric - TS$	**9.68%**	**23.5%**	3.5%

The performance of our trading strategies are compared with $\varepsilon - greedy$, Boostrapped-UCB and Symmetric-TS through Annual Return (AR), Sharpe Ratio (SR), and Maximum Drawdown (MaxD, namely the maximum portfolio value loss from the peak to the bottom). The performances of AR, SR, and MaxD. are shown in Table 1.

We can see the excellent performance of the asymmetric-TS algorithm from Figure 3 in the field of stock selection as the log return of stock is consistent with asymmetric α-stable distribution. The regret is reduced to close to 0, which means that the asymmetric α-stable distribution can almost perfectly fit the distribution of log return of stock prices. Our algorithm also gets the optimal Annual Return (AR) and the maximum Sharpe Ratio (SR), which means that it has good profitability and stability.

In order to illustrate the versatility of Thompson sampling for bandit settings more complicated than just one original data each time, one may consider stochastic Stock portfolio selection problems that relate to the correlation between actions.

4.3 Recommendation System

Recommendation systems are also common applications of Multi-armed Bandits. The items to be recommended are modeled as the arms to be pulled. The

recommendation system gets a score according to its own scoring system, and we regard the distribution of score as our reward distribution. Thus, the main goal is also to maximize the expected reward achieved after T times.

In this section, we have utilized two benchmark datasets (MovieLens 100K) of the real-world in recommender systems to implement the model practically. MovieLens 100K contains 100,000 ratings $R \in \{1,2,3,4,5\}$, 1682 movies (items) rated by 943 users.

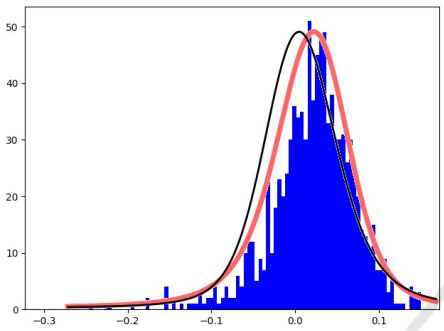

Figure 4: Recommendation data, the blue histogram represents the dataset, the black curve represents the fitted symmetric α stable distribution and the red one represents the fitted asymmetric α stable distribution.

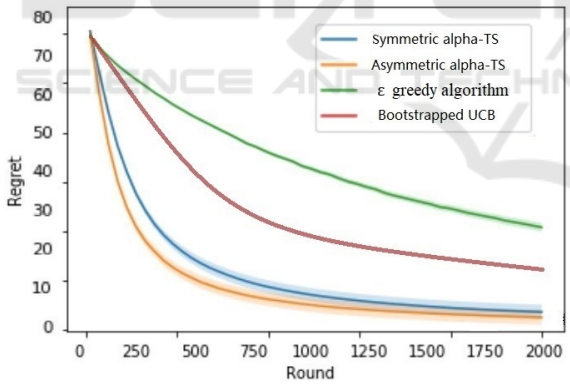

Figure 5: Regret for recommendation data, the green line is Bootstrapped UCB, blue one is common alpha-TS method while orange line shows our asymmetric α-TS method.

Thompson Sampling algorithms learn the rating distributions of films in few rounds, while $\varepsilon - greedy$ and Boostrapped-UCB fall into local optima. Figure 5 shows that Thompson Sampling strategy is more appropriate than ε-greedy and UCB strategy in a noise-free environment. The difference between symmetric and asymmetric algorithms is not significant, which may be due to the fact that the movie dataset conforms to the symmetric situation, or it may be due to the constrains of dataset rating $R \in \{1,2,3,4,5\}$.

5 CONCLUSIONS

In view of the complexity of action/observation space in many problems, we designed an asymmetric α Thompson sampling algorithm using Bayesian inference for stable distribution and verified the conjecture through the asymmetric data ,real stock price data and recommendation data.

Asymmetric α-stable algorithm can also be used to process symmetric data because it has no restrictions on β, but because it uses complicated Bayesian inference formula (in the symmetric α Thompson algorithm, the iteration from prior distribution to posterior distribution can be greatly simplified through the characteristics of symmetry and alternative variables), the iteration speed can not be compared with symmetric α Thompson algorithm which can iterate from prior distribution to posterior distribution immediately under symmetric conjecture and auxiliary variables.

We develop a regret bound for asymmetric α one in the parameter, action and observation spaces. Our algorithms only require the existence of bounded $1+\varepsilon$ moment of payoffs, and achieve $O(N^{\frac{1}{1+\varepsilon}} T^{\frac{1+\varepsilon}{1+2\varepsilon}})$ regret bound which can be used to determine the rationality of the data.

Applying the algorithm to stock price returns and recommendation systems, we demonstrate that asymmetric stable distribution is a better data model, which can explain the existence of skewness.

REFERENCES

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

Baisero, A. and Amato, C. (2021). Unbiased Asymmetric Actor-Critic for Partially Observable Reinforcement Learning. *The Computing Research Repository*.

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

Buckle, D. (1995). Bayesian inference for stable distributions. *Journal of the American Statistical Association*, 90(430):605–613.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, page 1516–1541.

Chen, Y., So, H. C., and Kuruoglu, E. E. (2016). Variance analysis of unbiased least lp-norm estimator in non-gaussian noise. *Signal Processing*, 122:190 – 203.

Dubey, A. and Pentland, A. (2019). Thompson Sampling on Symmetric alpha-Stable Bandits. *International Joint Conference on Artificial Intelligence*.

Embrechts, P., Lindskog, F., McNeil, A., and Rachev, S. (2003). Handbook of heavy tailed distributions in finance. *Modelling Dependence with Copulas and Applications to Risk Management. Handbooks in Finance: Book*, 1:329–385.

Herranz, D., Kuruoğlu, E. E., and Toffolatti, L. (2004). An alpha-stable approach to the study of the P(D) distribution of unresolved point sources in CMB sky maps. *Astronomy & Astrophysics*, 424(3):1081–1096.

Jia, H., Shi, C., and Shen, S. (2021). Multi-armed Bandit with Sub-exponential Rewards. *Operations Research Letters*, 49(5):728–733.

Karakuş, O., Kuruoğlu, E. E., and Altınkaya, M. A. (2020). Modelling impulsive noise in indoor powerline communication systems. *Signal, Image and Video Processing*, 14(8):1655 – 1661.

Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson sampling for 1-dimensional exponential family bandits. *Proceedings of NIPS*.

Korte, B. and Lovász, L. (1984). Greedoids-a structural framework for the greedy algorithm. *Progress in combinatorial optimization*, pages 221–243.

Kuruoglu, E. E. (2001). Density parameter estimation of skewed/spl alpha/-stable distributions. *IEEE Transactions on signal processing*, 49(10):2192–2201.

Kuruoglu, E. E. (2003). Analytical representation for positive /spl alpha/-stable densities. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 6, pages VI–729.

Lee, K., Yang, H., and Lim, S. (2020). Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462.

Lehmkuhl, M. and Promies, N. (2020). Frequency distribution of journalistic attention for scientific studies and scientific sources: An input–output analysis. *PloS one*, 15(11):e0241376.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *Proc.of Intl Conf.on World Wide Web*, pages 661–670.

Liu, K. and Zhao, Q. (2011). Multi-armed bandit problems with heavy-tailed reward distributions. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 485–492. IEEE.

Nguyen, N. H., Doğançay, K., and Kuruoğlu, E. E. (2019). An Iteratively Reweighted Instrumental-Variable Estimator for Robust 3-D AOA Localization in Impulsive Noise. *IEEE Transactions on Signal Processing*, 67(18):4795–4808.

Oja, H. (1981). On Location, Scale, Skewness and Kurtosis of Univariate Distributions. *Scandinavian Journal of Statistics*, 8(3):154–68.

Qi, C., Zhao, Z., Li, R., and Zhang, H. (2016). Characterizing and modeling social mobile data traffic in cellular networks. *IEEE Vehicular Technology Conference*, pages 1–5.

Russo, D. J. and Van, R. B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Shailesh, M. (2015). Decision Making-Investment, Financial and Risk Analysis in Mining Projects. *National Institute of Technology, Rourkela*.

Weron, R. (1996). On the Chambers-Mallows-Stuck method for simulating skewed stable random variables. *Statistics and Probability Letters*, 28(2):165–171.

Win, M. Z., Pinto, P. C., and Shepp, L. A. (2009). A mathematical theory of network interference and its applications. *Proc. IEEE*, 97(2):205–230.

Yu, X., Shao, H., Lyu, M. R., and King, I. (2018). Pure exploration of multi-armed bandits with heavy-tailed payoffs. *Uncertainty in Artificial Intelligence*.