Thompson Sampling on Symmetric α-Stable Bandits*

Abhimanyu Dubey and **Alex 'Sandy' Pentland**
Massachusetts Institute of Technology
{dubeya, pentland}@mit.edu

Abstract

Thompson Sampling provides an efficient technique to introduce prior knowledge in the multi-armed bandit problem, along with providing remarkable empirical performance. In this paper, we revisit the Thompson Sampling algorithm under rewards drawn from symmetric α-stable distributions, which are a class of heavy-tailed probability distributions utilized in finance and economics, in problems such as modeling stock prices and human behavior. We present an efficient framework for posterior inference, which leads to two algorithms for Thompson Sampling in this setting. We prove finite-time regret bounds for both algorithms, and demonstrate through a series of experiments the stronger performance of Thompson Sampling in this setting. With our results, we provide an exposition of symmetric α-stable distributions in sequential decision-making, and enable sequential Bayesian inference in applications from diverse fields in finance and complex systems that operate on heavy-tailed features.

1 Introduction

The multi-armed bandit (MAB) problem is a fundamental model in understanding the *exploration-exploitation* dilemma in sequential decision-making. The problem and several of its variants have been studied extensively over the years, and a number of algorithms have been proposed that optimally solve the bandit problem when the reward distributions are well-behaved, i.e. have a finite support, or are sub-exponential.

The most prominently studied class of algorithms are the Upper Confidence Bound (UCB) algorithms, that employ an "optimism in the face of uncertainty" heuristic [Auer *et al.*2002], which have been shown to be optimal (in terms of regret) in several cases [Cappé *et al.*2013, Bubeck *et al.*2013]. Over the past few years, however, there has been a resurgence in interest in the Thompson Sampling (TS) algorithm [Thompson1933], that approaches the problem from a Bayesian perspective.

Rigorous empirical evidence in favor of TS demonstrated by [Chapelle and Li2011] sparked new interest in the theoretical analysis of the algorithm, and the seminal work of [Agrawal and Goyal2012, Agrawal and Goyal2013, Russo and Van Roy2014] demonstrated the optimality of TS when rewards are bounded in $[0, 1]$ or are Gaussian. These results were extended in the work of [Korda *et al.*2013] to more general, exponential family reward distributions. The empirical studies, along with theoretical guarantees, have established TS as a powerful algorithm for the MAB problem.

However, when designing decision-making algorithms for complex systems, we see that interactions in such systems often lead to heavy-tailed and power law distributions, such as modeling stock prices [Bradley and Taqqu2003], preferential attachment in social networks [Mahanti *et al.*2013], and online behavior on websites [Kumar and Tomkins2010].

Specifically, we consider a family of extremely heavy-tailed reward distributions known as α-stable distributions. This family refers to a class of distributions parameterized by the exponent α, that include the Gaussian ($\alpha = 2$), Lévy ($\alpha = 1/2$) and Cauchy ($\alpha = 1$) distributions, all of which are used extensively in economics [Frain2009], finance [Carr and Wu2003] and signal processing [Shao and Nikias1993].

The primary hurdle in creating machine learning algorithms that account for α-stable distributions, however, is their intractable probability density, which cannot be expressed analytically. This prevents even a direct evaluation of the likelihood under this distribution. Their heavy-tailed nature, additionally, often leads to standard algorithms (such as Thompson Sampling assuming Gaussian rewards), concentrating on incorrect arms.

In this paper, we create two algorithms for Thompson Sampling under symmetric α-stable rewards with finite means. Our contributions can be summarized as follows:

1. Using auxiliary variables, we construct a framework for posterior inference in symmetric α-stable bandits that leads to the first efficient algorithm for Thompson Sampling in this setting, which we call α-TS.
2. To the best of our knowledge, we provide the first finite-time polynomial bound on the Bayesian Regret of Thompson Sampling achieved by α-TS in this setting.
3. We improve on the regret by proposing a modified Thompson Sampling algorithm, called Robust α-TS,

[*] Full version (with appendix) available at this link.

that utilizes a truncated mean estimator, and obtain the first $\tilde{O}(N^{\frac{1}{1+\epsilon}})$ Bayesian Regret in the α-stable setting. Our bound matches the optimal bound for $\alpha \in (1, 2)$ (within logarithmic factors).

4. Through a series of experiments, we demonstrate the proficiency of our two Thompson Sampling algorithms for α-stable rewards, which consistently outperform all existing benchmarks.

Our paper is organized as follows: we first give a technical overview of the MAB problem, the Thompson Sampling algorithm and α-stable distributions. Next, we provide the central algorithm α-TS and its analysis, followed by the same for the Robust α-TS algorithm. We then provide experimental results on multiple simulation benchmarks and finally, discuss the related work in this area prior to closing remarks.

2 Preliminaries

2.1 Thompson Sampling

The K-Armed Bandit Problem: In any instance of the K-armed bandit problem, there exists an agent with access to a set of K actions (or "arms"). The learning proceeds in rounds, indexed by $t \in [1, T]$. The total number of rounds, known as the *time horizon* T, is known in advance. The problem is iterative, wherein for each round $t \in [T]$:

1. Agent picks arm $A_t \in [K]$.
2. Agent observes reward $r_{A_t}(t)$ from that arm.

For arm $k \in [K]$, rewards come from a distribution $\mathcal{D}_k$ with mean $\mu_k = \mathbb{E}_{\mathcal{D}_k}[r]$[1]. The largest expected reward is denoted by $\mu^* = \max_{k\in[K]} \mu_k$, and the corresponding arm(s) is denoted as the *optimal* arm(s) k^*. In our analysis, we will focus exclusively on the i.i.d. setting, that is, for each arm, rewards are independently and identically drawn from $\mathcal{D}_k$, every time arm k is pulled.

To measure the performance of any (possibly randomized) algorithm we utilize a measure known as *Regret* $R(T)$, which, at any round T, is the difference of the cumulative mean reward of the algorithm against the expected reward of always playing an optimal arm.

$$R(T) = \mu^* T - \sum_{t=0}^{T} \mu_{A_t} \tag{1}$$

Thompson Sampling (TS): Thompson Sampling [Thompson1933] proceeds by maintaining a posterior distribution over the parameters of the bandit arms. If we assume that for each arm k, the reward distribution $\mathcal{D}_k$ is parameterized by a (possibly vector) parameter θ_k that come from a set $\mathbf{\Theta}$ with a prior probability distribution $p(\theta_k)$ over the parameters, the Thompson Sampling algorithm proceeds by selecting arms based on the posterior probability of the reward under the arms. For each round $t \in [T]$, the agent:

1. Draws parameters $\hat{\theta}_k(t)$ for each arm $k \in [K]$ from the posterior distribution of parameters, given the previous rewards $\mathbf{r}_k(t-1) = \{r_k^{(1)}, r_k^{(2)}, ...\}$ till round $t-1$ (note that the posterior distribution for each arm only depends on the rewards obtained using that arm). When $t = 1$, this is just the prior distribution over the parameters.

$$\hat{\theta}_k(t) \sim p(\theta_k|\mathbf{r}_k(t-1)) \propto p(\mathbf{r}_k(t-1)|\theta_k)p(\theta_k) \tag{2}$$

2. Given the drawn parameters $\hat{\theta}_k(t)$ for each arm, chooses arm a_t with the largest mean reward over the posterior distribution.

$$A_t = \arg\max_{k\in[K]} \mu_k(\hat{\theta}_k(t)) \tag{3}$$

3. Obtains reward r_t after taking action A_t and updates the posterior distribution for arm A_t.

In the Bayesian case, the measure for performance we will utilize in this paper is the Bayes Regret (BR) [Russo and Van Roy2014], which is the expected regret over the priors. For any policy π, this is given by:

$$\text{BayesRegret}(T, \pi) = \mathbb{E}[R(t)]. \tag{4}$$

While the regret provides a stronger analysis, any bound on the Bayes Regret is essentially a bound on the expected regret, since if an algorithm admits a Bayes Regret of $O(g(T))$, then its Regret is also stochastically bounded by $g(\cdot)$ [Russo and Van Roy2014].

2.2 α-Stable Distributions

α-Stable distributions, introduced by Lévy [Lévy1925] are a class of probability distributions defined over $\mathbb{R}$ whose members are closed under linear transformations.

Definition 1 ([Borak *et al.*2005])**.** *Let X_1 and X_2 be two independent instances of the random variable X. X is **stable** if, for $a_1 > 0$ and $a_2 > 0$, $a_1X_1 + a_2X_2$ follows the same distribution as $cX + d$ for some $c > 0$ and $d \in \mathbb{R}$.*

A random variable $X \sim S_\alpha(\beta, \mu, \sigma)$ follows an α-stable distribution described by the parameters $\alpha \in (0, 2]$ (characteristic exponent) and $\beta \in [-1, 1]$ (skewness), which are responsible for the shape and concentration of the distribution, and parameters $\mu \in \mathbb{R}$ (shift) and $\sigma \in \mathbb{R}^+$ (scale) which correspond to the location and scale respectively. While it is not possible to analytically express the density function for generic α-stable distributions, they are known to admit the characteristic function $\phi(x; \alpha, \beta, \sigma, \mu)$:

$$\phi(x; \alpha, \beta, \sigma, \mu) = \exp\left\{ix\mu - |\sigma x|^\alpha \left(1 - i\beta\, \text{sign}(x)\Phi_\alpha(x)\right)\right\},$$

where $\Phi_\alpha(x)$ is given by

$$\Phi_\alpha(x) = \{\ \tan(\tfrac{\pi\alpha}{2}) \text{ when } \alpha \neq 1, \quad -\tfrac{2}{\pi}\log|x|, \text{ when } \alpha = 1$$

For fixed values of α, β, σ and μ we can recover the density function from $\phi(\cdot)$ via the inverse Fourier transform:

$$p(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(x; \alpha, \beta, \sigma, \mu)e^{-izx}dx$$

[1] α-stable distributions with $\alpha \leq 1$ do not admit a finite first moment. To continue with existing measures of regret, we only consider rewards with $\alpha > 1$.

Algorithm 1 Chambers-Mallows-Stuck Generation

```
Input: $V \sim U(-\pi/2, \pi/2), W \sim E(1)$
Output: $X \sim S_\alpha(\beta, \sigma, \mu)$
    Set $B_{\alpha,\beta} = \arctan(\beta \tan(\pi\alpha/2))\alpha^{-1}$
    Set $S_{\alpha,\beta} = \left(1 + \beta^2 \tan^2(\pi\alpha/2)\right)^{1/(2\alpha)}$
    Set $Y = S_{\alpha,\beta} \times \frac{\sin(\alpha(V+B_{\alpha,\beta})}{\cos(V)^{1/\alpha}} \times \left(\frac{\cos(V-\alpha(V+B_{\alpha,\beta})}{W}\right)^{\frac{1-\alpha}{\alpha}}$
    return $X = \sigma Y + \mu$.
```

Most of the attention in the analysis of α-stable distributions has been focused on the stability parameter α, which is responsible for the tail "fatness". It can be shown that asymptotically, the tail behavior ($x \to \pm\infty$) of $X \sim S_\alpha(\beta, \mu, \sigma)$ follows [Borak *et al.*2005]:

$$f(x) \sim \frac{1}{|x|^{1+\alpha}} \left(\sigma^\alpha (1 + \operatorname{sgn}(x)\beta) \sin\left(\frac{\pi\alpha}{2}\right) \frac{\Gamma(\alpha+1)}{\pi} \right)$$

where $\alpha < 2$ and $\Gamma(\cdot)$ denotes the Gamma function. The power-law relationship admitted by the density is responsible for the heaviness of said tails.

Lemma 1 ([Borak *et al.*2005])**.** $X \sim S_\alpha(\beta, \mu, \sigma), \alpha < 2$ *admits a moment of order* λ *only if* $\lambda \in (-\infty, \alpha)$.

From Lemma 1 it follows that α-stable random variables only admit a finite mean for $\alpha > 1$, and also admit a finite variance only when $\alpha = 2$, which corresponds to the family of Gaussian distributions. To continue with existing measures of regret, we restrict our analysis hence to α-stable distributions only with $\alpha > 1$. Note that for all our discussions, $1 < \alpha < 2$, hence, all distributions examined are heavy-tailed, with infinite variance. Additionally, we restrict ourselves to only symmetric ($\beta = 0$) distributions: asymmetric distributions do not allow a scaled mixture representation (which is the basis of our framework, see Section 3.1).

Sampling from α-Stable Densities

For general values of $\alpha, \beta, \sigma, \mu$, it is not possible to analytically express the density of α-stable distributions, and hence we resort to using auxiliary variables for sampling. The Chambers-Mallows-Stuck [Chambers *et al.*1976] algorithm is a straightforward method to generate samples from the density $S_\alpha(\beta, 1, 0)$ (for $\alpha \neq 1$) via a non-linear transformation of a uniform random variable V and an exponential random variable W, which can then be re-scaled to obtain samples from $S_\alpha(\beta, \sigma, \mu)$ (Algorithm 1).

Products of α-Stable Densities

A central property of α-stable densities that will be utilized in the following section is the behavior of products of independent α-stable variables.

Lemma 2 ([Borak *et al.*2005])**.** *Let* Z *and* $Y > 0$ *be independent random variables such that* $Z \sim S_\gamma(0, \sigma_1, \mu_1)$ *and* $Y \sim S_\delta(1, \sigma_2, \mu_2)$. *Then* $ZY^{1/\gamma}$ *is stable with exponent* $\gamma\delta$.

We now begin with the algorithm description and analysis.

3 α-Thompson Sampling

We consider the setting where, for an arm k, the corresponding reward distribution is given by $\mathcal{D}_k = S_\alpha(0, \sigma, \mu_k)$ where $\alpha \in (1, 2), \sigma \in \mathbb{R}^+$ are known in advance, and μ_k is unknown[2]. We set a prior distribution over μ_k. We now derive the form of the prior distribution, and outline an algorithm for Bayesian inference.

3.1 Scale Mixtures of Normals

On setting $\gamma = 2$ (Gaussian) and $\beta = \alpha/2 < 1$ in Lemma 2, the product distribution $X = ZY^{1/2}$ is stable with exponent α. This property is an instance of the general framework of *scale mixtures of normals* (SMiN) [Andrews and Mallows1974], which are described by the following:

$$p_X(x) = \int_0^\infty \mathcal{N}(x|0, \lambda\sigma^2) p_\Lambda(\lambda) d\lambda \tag{5}$$

This framework contains a large class of heavy-tailed distributions which include the exponential power law, Student's-t and symmetric α-stable distributions [Godsill and Kuruoglu1999]. The precise form of the variable X depends on the *mixing distribution* $p_\Lambda(\lambda)$. For instance, when p_Λ is the inverted Gamma distribution (the conjugate prior for a unknown variance Gaussian), the resulting p_X follows a Student's-t distribution.

Bayesian inference directly from $S_\alpha(0, \sigma, \mu)$ is difficult: the non-analytic density prevents a direct evaluation of the likelihood function, and the non-Gaussianity introduces difficulty in its implementation. However, the SMiN representation enables us to draw samples directly from $S_\alpha(0, \sigma, \mu)$ using the auxiliary variable λ:

$$x \sim \mathcal{N}(\mu, \lambda\sigma^2), \lambda \sim S_{\alpha/2}(1, 1, 0) \tag{6}$$

This sampling assists in inference since x is Gaussian conditioned on λ: given samples of λ, we can generate x from the induced Gaussian conditional distribution.

3.2 Posteriors for α-Stable Rewards

Let us examine approximate inference for a particular arm $k \in [K]$. At any time $t \in [T]$, assume this arm has been pulled $n_k(t)$ times previously, and hence we have a vector of reward samples $\mathbf{r}_k(t) = \{r_k^{(1)}, ..., r_k^{(n_k(t))}\}$ observed until time t. Additionally, assume we have $n_k(t)$ samples of an auxiliary variable $\boldsymbol{\lambda}_k(t) = \{\lambda_k^{(1)}, ..., \lambda_i^{(n_k(t))}\}$ where $\lambda_k \sim S_{\alpha/2}(1, 1, 0)$.

Recall that $r_k \sim S_\alpha(0, \sigma, \mu_k)$ for an unknown (but fixed) μ_k. From the SMiN representation, we know that r_k is conditionally Gaussian given the auxiliary variable λ_k, that is $p(r_k|\lambda_k, \mu_k) \sim \mathcal{N}(\mu_k, \lambda_k\sigma^2)$. We can then obtain the conditional likelihood as the following:

$$p(\mathbf{r}_k(t)|\boldsymbol{\lambda}_k(t), \mu_k) \propto \exp\left(-\tfrac{1}{2\sigma^2}\left(\textstyle\sum_{i=1}^{n_k(t)} \frac{(r_k^{(i)}-\mu_k)^2}{\lambda_k^{(i)}}\right)\right). \tag{7}$$

[2] Typically, more general settings have been explored for TS in the Gaussian case, where the variance is also unknown. Our algorithm can also be extended to an unknown scale (σ) using an inverse Gamma prior, similar to [Godsill and Kuruoglu1999].

We can now assume the conjugate prior over μ_k, which is a normal distribution with mean μ_k^0 and variance σ^2. We then obtain the posterior density for μ_k as (full derivation in the Appendix):

$$p(\mu_k|\mathbf{r}_k(t), \boldsymbol{\lambda}_k(t)) \propto \mathcal{N}\left(\hat{\mu}_k(t), \hat{\sigma}_k^2(t)\right) \text{ where,}$$

$$\hat{\sigma}_k^2(t) = \frac{\sigma^2}{\sum_{i=1}^{n_k(t)} \frac{1}{\lambda_k^{(i)}} + 1}, \; \hat{\mu}_k(t) = \frac{\sum_{i=1}^{n_k(t)} \frac{r_k^{(i)}}{\lambda_k^{(i)}} + \mu_k^0}{\sum_{i=1}^{n_k(t)} \frac{1}{\lambda_k^{(i)}} + 1}. \tag{8}$$

We know that $\hat{\sigma}_k^2(t) > 0$ since $\lambda_k^{(i)} > 0 \, \forall i$ as they are samples from a positive stable distribution ($\beta = 1$). Given $\mathbf{r}_k(t)$ and μ_k, we also know that the individual elements of $\boldsymbol{\lambda}_k(t)$ are independent, which provides us with the following decomposition for the conditional density of $\boldsymbol{\lambda}_k(t)$,

$$p(\boldsymbol{\lambda}_k(t)|\mathbf{r}_k(t), \mu_k) = \prod_{i=1}^{n_k(t)} p(\lambda_k^{(i)}|r_k^{(i)}, \mu_k), \text{ where,}$$

$$p(\lambda_k^{(i)}|r_k^{(i)}, \mu_k) \propto \mathcal{N}(r_k^{(i)}|\mu_k, \lambda_k^{(i)}, \sigma^2) f_{\alpha/2,1}(\lambda_k^{(i)}). \tag{9}$$

Here, $f_{\alpha,\beta}(\cdot)$ is the density of a random variable following $S_\alpha(\beta, 1, 0)$. Our stepwise posterior sampling routine is hence as follows. At any time t, after arm k is pulled and we receive reward $r_k^{n_k(t)}$, we set $\mathbf{r}_k(t) = [\mathbf{r}_k(t-1), r_k^{n_k(t)}]$. Then for a fixed Q iterations, we repeat:

1. For $i \in [1, n_k(t)]$, draw $\lambda_k^{(i)} \sim p(\lambda_k^{(i)}|r_k^{(i)}, \mu_k(t))$.
2. Draw $\mu_k(t) \sim p(\mu_k|\mathbf{r}_k(t), \boldsymbol{\lambda}_k(t))$.

Sampling from the conditional posterior of μ_k is straightforward since it is Gaussian. To sample from the complicated posterior of $\lambda_k^{(i)}$, we utilize rejection sampling.

3.3 Rejection Sampling for $\lambda_k^{(i)}$

Sampling directly from the posterior is intractable since it is not analytical. Therefore, to sample $\lambda_k^{(i)}$ we follow the pipeline described in [Godsill and Kuruoglu1999]. We note that the likelihood of the mean-normalized reward $v_k^{(i)} = r_k^{(i)} - \mu_k(t)$ forms a valid rejection function since it is bounded:

$$p\left(v_k^{(i)}|0, \lambda_k^{(i)}\sigma^2\right) \leq \frac{1}{v_k^{(i)}\sqrt{2\pi}} \exp(-1/2) \tag{10}$$

Since $v_k^{(i)} \sim \mathcal{N}(0; \lambda_k^{(i)}\sigma^2)$. Thus, we get the procedure:

1. Draw $\lambda_k^{(i)} \sim S_{\alpha/2}(1, 1, 0)$ (using Algorithm 1).
2. Draw $u \sim \mathcal{U}\left(0, (v_k^{(i)}\sqrt{2\pi})^{-1} \exp(-1/2)\right)$.
3. If $u > p(v_k^{(i)}|0, \lambda_k^{(i)}\sigma^2)$, reject $\lambda_k^{(i)}$ and go to Step 1.

Combining all these steps, we can now outline our algorithm, α-Thompson Sampling (α-TS) as described in Algorithm 2.

It is critical to note that in Algorithm 2, we do not draw from the full vector of $\boldsymbol{\lambda}_k(t)$ at every iteration, but only from the last obtained reward. This is done to accelerate the inference process, and while it leads to a slower convergence of the posterior, we observe that it performs well in practice. Alternatively, one can re-sample $\boldsymbol{\lambda}_k(t)$ over a fixed window of the previous rewards, to prevent the runtime from increasing linearly while enjoying faster convergence.

Algorithm 2 α-Thompson Sampling

```
**Input**: Arms $k \in [K]$, priors $\mathcal{N}(\mu_k^0, \sigma^2)$ for each arm.
Set $D_k = 1, N_k = 0$ for each arm $k$.
**for** For each iteration $t \in [1, T]$ **do**
    Draw $\bar{\mu}_k(t) \sim \mathcal{N}\left(\frac{\mu_k^0+N_k}{D_k}, \frac{\sigma^2}{D_k}\right)$ for each arm $k$.
    Choose arm $A_t = \arg\max_{k\in[K]} \bar{\mu}_k(t)$, and get reward $r_t$.
    **for** $q \in [0, Q)$ **do**
        Calculate $v_{A_t}^{(t)} = r_t - \bar{\mu}_{A_t}$.
        Draw $\lambda_{A_t}^{(t)}$ following Section 3.3.
        Set $D_q = D_{A_t} + 1/\lambda_{A_t}^{(t)}, N_q = N_{A_t} + r_t/\lambda_{A_t}^{(t)}$.
        Draw $\bar{\mu}_{A_t} \sim \mathcal{N}\left(\frac{\mu_{A_t}^0+N_q}{D_q}, \frac{\sigma^2}{D_q}\right)$.
    **end for**
    Set $D_{A_t} = D_{A_t} + 1/\lambda_{A_t}^{(t)}, N_{A_t} = N_{A_t} + r_t/\lambda_{A_t}^{(t)}$.
**end for**
```

3.4 Regret Analysis

In this section, we derive an upper bound on the finite-time Bayesian Regret (BR) incurred by the α-TS algorithm. We continue with the notation used in previous sections, and assume a K armed bandit with T maximum trials. Each arm k follows an α-stable reward $S_\alpha(0, \sigma, \mu_k)$, and without loss of generality, let $\mu^* = \max_{k\in[K]} \mu_i$ denote the arm with maximum mean reward.

Theorem 1 (Regret Bound)**.** *Let $K > 1, \alpha \in (1, 2), \sigma \in \mathbb{R}^+, \mu_{k:k\in[K]} \in [-M, M]$. For a K-armed stochastic bandit with rewards for each arm k drawn from $S_\alpha(0, \sigma, \mu_k)$, we have, asymptotically, for ϵ chosen a priori such that $\epsilon \rightarrow (\alpha - 1)^-$,*

$$\textit{BayesRegret}(T) = O(K^{\frac{1}{1+\epsilon}} T^{\frac{2}{1+\epsilon}})$$

Proof-sketch. We first utilize the characteristic function representation of the probability density to obtain the centered $(1+\epsilon)^{th}$ moment of the reward distribution for each arm. We use this moment to derive a concentration bound on the deviation of the empirical mean of α-stable densities. Next, we proceed with the decomposition of the Bayesian Regret in terms of upper-confidence bounds, as done in [Russo and Van Roy2014]. The complete proof is included in detail in the appendix.

We note the following: First, the only additional assumption we make on the reward distributions is that the true means are bounded, which is a standard assumption [Russo and Van Roy2014] and easily realizable in most application cases. Next, $\epsilon < \alpha - 1$ must be chosen carefully to control the finite-time regret. As $\epsilon \rightarrow \alpha - 1$, we see that while the growth of $T^{\frac{2}{1+\epsilon}}$ decreases, the constants in the finite-time expression grow, and are not finite at $\epsilon = \alpha - 1$. This behavior arises

Algorithm 3 Robust α-Thompson Sampling

```
Input: Arms $k \in [K]$, priors $\mathcal{N}(\mu_k^0, \sigma^2)$ for each arm.
Set $D_k = 1, N_k = 0$ for each arm $k$.
for For each iteration $t \in [1, T]$ do
    Draw $\bar{\mu}_k(t) \sim \mathcal{N}\left(\frac{\mu_k^0 + N_k}{D_k}, \frac{\sigma^2}{D_k}\right)$ for each arm $k$.
    Choose arm $A_t = \arg\max_k \bar{\mu}_k(t)$, and get reward $r_t$.
    If $|r_t| > \left(\frac{H(\epsilon,\alpha,\sigma)\cdot i}{2\log(T)}\right)^{\frac{1}{1+\epsilon}}$, set $r_t = 0$.
    Repeat steps 6-12 of Algorithm 2.
end for
```

from the non-compactness of the set of finite moments for α-stable distributions (see Appendix for detailed analysis).

Compared to the problem-independent regret bound of $O(\sqrt{KT \log T})$ for Thompson Sampling on the multi-armed Gaussian bandit problem demonstrated by [Agrawal and Goyal2013], our bound differs in two aspects: first, we admit a $K^{\frac{1}{1+\epsilon}}$ complexity on the number of arms, which contrasted with the Gaussian bandit is identical when $\epsilon \to 1$. Second, we have a superlinear dependence of order $T^{\frac{2}{1+\epsilon}}$. The term of $T^{\frac{1}{1+\epsilon}}$ approaches the Gaussian case ($\sqrt{T}$) when $\epsilon \to 1$, leaving us with the additional sub-linear term ($T^{\frac{1}{1+\epsilon}}$) compared to the sub-logarithmic term ($\sqrt{\log T}$) from the Gaussian case.

In the next section, we address the issue of polynomial concentration by utilizing the more robust, truncated mean estimator instead of the empirical mean, and obtain a modified, robust version of α-TS.

3.5 Robust α-Thompson Sampling

Assume that for all arms $k \in [K], \mu_k \leq M$. Note that this assumption is equivalent to the boundedness assumption in the analysis of α-TS, and is a reasonable assumption to make in any practical scenario with some domain knowledge of the problem. Let $\delta \in (0, 1), \epsilon \in (0, \alpha - 1)$. Now, consider the estimator $\hat{r}_k^*(t)$ given by:

$$\hat{r}_k^*(t) = \frac{1}{n_k(t)} \sum_{i=1}^{n_k(t)} r_k^{(i)} \mathbf{1}\left\{ |r_k^{(i)}| \leq \left(\frac{H(\epsilon, \alpha, \sigma) \cdot i}{2 \log(T)} \right)^{\frac{1}{1+\epsilon}} \right\}$$

$$\text{where, } H(\epsilon, \alpha, \sigma) = \left(\frac{\epsilon \left(M \cdot \Gamma(-\epsilon/\alpha) + \sigma\alpha\Gamma(1 - \frac{\epsilon+1}{\alpha}) \right)}{\sigma\alpha \sin\left(\frac{\pi\cdot\epsilon}{2}\right) \Gamma(1-\epsilon)} \right) \tag{11}$$

$\hat{r}_k^*(t)$ then describes a truncated mean estimator for an arm k (pulled $n_k(t)$ times), where a reward $r_k^{(i)}$ at any trial i of the arm is discarded if it is larger than the bound. We choose such a form of the truncation since it allows us to obtain an exponential concentration for α-stable densities, which will eventually provide us tighter regret.

As can be seen, the corresponding Robust α-TS algorithm is identical to the basic α-TS algorithm, except for this step of rejecting a reward if it is greater than the bound. The algorithm is outlined in Algorithm 3.

Theorem 2 (Regret Bound). *Let $K > 1, \alpha \in (1, 2), \sigma \in \mathbb{R}^+, \mu_{k:k\in[K]} \in [-M, M]$. For a K-armed bandit with rewards for each arm k drawn from $S_\alpha(0, \sigma, \mu_k)$, we have, for ϵ chosen a priori such that $\epsilon \to (\alpha - 1)^-$ and truncated estimator given in Equation (11),*

$$\textit{BayesRegret}(T) = \tilde{O}\left((KT)^{\frac{1}{1+\epsilon}}\right)$$

Proof-sketch. The truncated mean estimator can be used to obtain a tighter concentration bound on the empirical mean. This can be understood intuitively as by carefully rejecting values that are distant from the mean, our empirical mean is robust to outliers, and would require less samples to concentrate around the true mean. Using the concentration result, we follow an analysis identical to Theorem 1. The full proof is deferred to the Appendix for brevity.

We see that this bound is tight (upto logarithmic factors): it matches the problem-independent lower bound of $O((KT)^{\frac{1}{1+\epsilon}})$ [Bubeck *et al.*2013]. Algorithm 3's improvements increase as α decreases: the likelihood of obtaining confounding outliers increases as $\alpha \to 1$, and can perturb the posterior mean in the naive α-TS algorithm. In the next section, we discuss some experimental results that cement the value of α-TS for α-stable bandits.

4 Experiments

We conduct simulations with the following benchmarks – (i) an ε-greedy agent with linearly decreasing ε, (ii) Regular TS with Gaussian priors and a Gaussian assumption on the data (Gaussian-TS), (iii) Robust-UCB [Bubeck *et al.*2013] for heavy-tailed distributions using a truncated mean estimator, and (iv) α-TS and (v) Robust α-TS, both with Q(iterations for sampling) as 50.

Setting Priors for TS: In all the Thompson Sampling benchmarks, setting the priors are crucial to the algorithm's performance. In the competitive benchmarking, we randomly set the priors for each arm from the same range we use for setting the mean rewards.

Performance against Competitive Benchmarks

We run 100 MAB experiments each for all 5 benchmarks for $\alpha = 1.8$ and $\alpha = 1.3$, and $K = 50$ arms, and for each arm, the mean is drawn from $[0, 2000]$ randomly for each experiment, and $\sigma = 2500$. Each experiment is run for $T = 5000$ iterations, and we report the regret averaged over time, i.e. $R(t)/t$ at any time t.

In Figures 1A and 1B, we see the results of the regret averaged over all 100 experiments. We observe that for $\alpha = 1.3$, there are more substantial variations in the regret (low α implies heavy outliers), yet both algorithms comfortably outperform the other baselines.

In the case of $\alpha = 1.8$, the variations are not that substantial, the performance follows the same trend. It is important to note that regular Thompson Sampling (Gaussian-TS) performs competitively, although in our experiments, we observed that when K is large, the algorithm often concentrates on the incorrect arm, and subsequently earns a larger regret.

Intuitively, we can see that whenever arms have mean rewards close to each other (compared to the variance), ϵ-greedy will often converge to the incorrect arm. Robust-UCB,

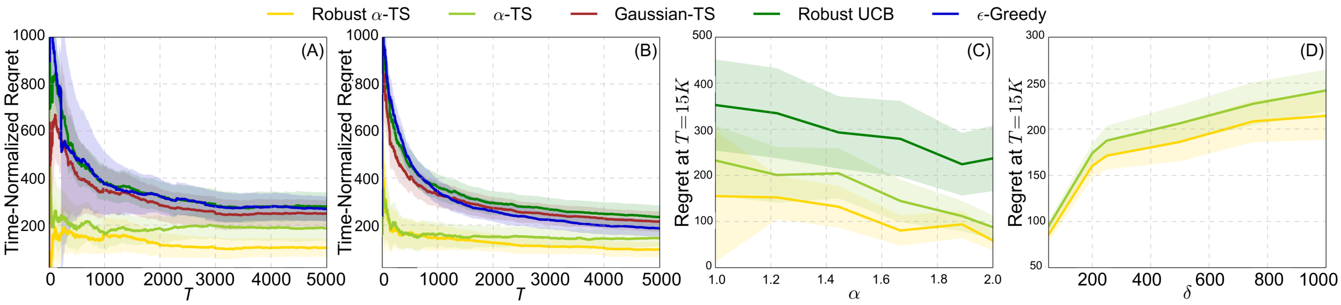

Figure 1: (A) Competitive benchmarking for $\alpha = 1.3$, and (B) $\alpha = 1.8$; (C) Ablation studies for varying α, and (D) varying prior strength. Shaded areas denote variance scaled by 0.25 in (A) and (B), and scaled by 0.5 in (C) and (D). Figures are best viewed in color.

however, makes very weak assumptions on the data distributions, and hence has a much larger exploration phase, leading to larger regret. Compared to α-TS, Robust α-TS is less affected by outlying rewards (as is more evident in $\alpha = 1.3$ vs. $\alpha = 1.8$) and hence converges faster.

Ablation Studies

We additionally run two types of ablation studies - first, we compare the performance of α-TS on the identical set up as before (same K and reward distributions), but with varying values of $\alpha \in (1, 2)$. We report the expected time-normalized regret averaged over 100 trials in Figure 1C, and observe that (i) as α increases, the asymptotic regret decreases faster, and (ii) as expected, for lower values of α there is a substantial amount of variation in the regret.

Secondly, we compare the effect of the sharpness of the priors. In the previous experiments, the priors are drawn randomly from the same range as the means, without any additional information. However, by introducing more information about the means through the priors, we can expect better performance. In this experiment, for each mean μ_k, we randomly draw the prior mean μ_k^0 from $[\mu_k - \delta, \mu_k + \delta]$, and observe the regret after $T = 15K$ trials for δ from 50 to 1000. The results for this experiment are summarized in Figure 1D for $K = 10$ and $\sigma = 25$, and results are averaged over 25 trials each. We see that with uninformative priors, α-TS performs competitively, and gets better as the priors get sharper.

5 Related Work

A version of the UCB algorithm [Auer *et al.*2002] has been proposed in [Bubeck *et al.*2013] coupled with several robust mean estimators to obtain Robust-UCB algorithms with optimal *problem-dependent* (i.e. dependent on individual μ_ks) regret when rewards are heavy-tailed. However, the optimal version of their algorithm has a few shortcomings that α-TS addresses: their algorithm requires the median-of-means estimator, which has an expensive space complexity of $O(\log T)$ and time complexity of $O(\log \log T)$ per update. Secondly, there is no mechanism to incorporate prior information, which can be advantageous, as seen even with weak priors. Additionally, [Vakili *et al.*2013] introduce a deterministic exploration-exploitation algorithm, which achieves same order regret as Robust-UCB for heavy-tailed rewards.

There has been work in analysing Thompson Sampling for specific Pareto and Weibull heavy-tailed distributions in [Korda *et al.*2013], however, the Weibull and Pareto distributions typically have "lighter" tails owing to the existence of more higher order moments, and hence cannot typically be used to model very heavy tailed signals.

In related problems, [Yu *et al.*2018] provide a purely exploratory algorithm for best-arm identification under heavy-tailed rewards, using a finite $(1 + \epsilon)^{th}$ moment assumption. Similarly, [Shao *et al.*2018, Medina and Yang2016] explore heavy-tailed rewards in the linear bandit setting.

6 Conclusion

In this paper, we first designed a framework for efficient posterior inference for the α-stable family, which has largely been ignored in the bandits literature owing to its intractable density function. We formulated the first polynomial problem-independent regret bounds for Thompson Sampling with α-stable densities, and subsequently improved the regret bound to achieve the optimal regret identical to the sub-Gaussian case, providing an efficient framework for decision-making for these distributions.

Additionally, our intermediary concentration results provide a starting point for other machine learning problems that may be investigated in α-stable settings. There is ample evidence to support the existence of α-stability in various modeling problems across economics [Frain2009], finance [Bradley and Taqqu2003] and behavioral studies [Mahanti *et al.*2013]. With tools such as ours, we hope to usher scientific conclusions in problems that cannot make sub-Gaussian assumptions, and can lead to more robust empirical findings. Future work may include viewing more involved decision-making processes, such as MDPs, in the same light, leading to more robust algorithms.

Acknowledgements

We would like to thank Peter Krafft, Yan Leng, Dhaval Adjodah, Ziv Epstein, Michiel Bakker and Ishaan Grover for their comments on this manuscript, the anonymous reviewers for their suggestions, and MIT Quest for Intelligence for their generous support with computational resources.

References

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

[Agrawal and Goyal, 2013] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

[Andrews and Mallows, 1974] David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, 1974.

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[Borak *et al.*, 2005] Szymon Borak, Wolfgang Härdle, and Rafał Weron. Stable distributions. In *Statistical tools for finance and insurance*, pages 21–44. Springer, 2005.

[Bradley and Taqqu, 2003] Brendan O Bradley and Murad S Taqqu. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pages 35–103. Elsevier, 2003.

[Bubeck *et al.*, 2013] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[Cappé *et al.*, 2013] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[Carr and Wu, 2003] Peter Carr and Liuren Wu. The finite moment log stable process and option pricing. *The journal of finance*, 58(2):753–777, 2003.

[Chambers *et al.*, 1976] John M Chambers, Colin L Mallows, and BW Stuck. A method for simulating stable random variables. *Journal of the american statistical association*, 71(354):340–344, 1976.

[Chapelle and Li, 2011] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[Frain, 2009] John C Frain. *Studies on the Application of the Alpha-stable Distribution in Economics*. 2009.

[Godsill and Kuruoglu, 1999] Simon Godsill and Ercan E Kuruoglu. Bayesian inference for time series with heavy-tailed symmetric α-stable noise processes. 1999.

[Korda *et al.*, 2013] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

[Kumar and Tomkins, 2010] Ravi Kumar and Andrew Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web*, pages 561–570. ACM, 2010.

[Lévy, 1925] P Lévy. Calcul des probabilités, vol. 9. *Gauthier-Villars Paris*, 1925.

[Mahanti *et al.*, 2013] Aniket Mahanti, Niklas Carlsson, Anirban Mahanti, Martin Arlitt, and Carey Williamson. A tale of the tails: Power-laws in internet measurements. *IEEE Network*, 27(1):59–64, 2013.

[Medina and Yang, 2016] Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650, 2016.

[Russo and Van Roy, 2014] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[Shao and Nikias, 1993] Min Shao and Chrysostomos L Nikias. Signal processing with fractional lower order moments: stable processes and their applications. *Proceedings of the IEEE*, 81(7):986–1010, 1993.

[Shao *et al.*, 2018] Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, pages 8430–8439, 2018.

[Thompson, 1933] William 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.

[Vakili *et al.*, 2013] Sattar Vakili, Keqin Liu, and Qing 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.

[Yu *et al.*, 2018] Xiaotian Yu, Han Shao, Michael R Lyu, and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Conference on Uncertainty in Artificial Intelligence*, pages 937–946, 2018.