Thompson Sampling Algorithms for Mean-Variance Bandits

Qiuyu Zhu [1] **Vincent Y. F. Tan** [1 2 3]

Abstract

The multi-armed bandit (MAB) problem is a classical learning task that exemplifies the exploration-exploitation tradeoff. However, standard formulations do not take into account *risk*. In online decision making systems, risk is a primary concern. In this regard, the mean-variance risk measure is one of the most common objective functions. Existing algorithms for mean-variance optimization in the context of MAB problems have unrealistic assumptions on the reward distributions. We develop Thompson Sampling-style algorithms for mean-variance MAB and provide comprehensive regret analyses for Gaussian and Bernoulli bandits with fewer assumptions. Our algorithms achieve the best known regret bounds for mean-variance MABs and also attain the information-theoretic bounds in some parameter regimes. Empirical simulations show that our algorithms significantly outperform existing LCB-based algorithms for all risk tolerances.

1. Introduction

The MAB problem studies the problem of online learning with partial feedback. This problem has a large number of real-world applications, such as online advertising, clinical trials, and financial portfolio design. The most widely-used MAB model is the stochastic MAB model. A player chooses among several arms, each defined by an independent reward distribution. In each period, the player plays one arm and obtains a random reward observation from that arm. The player faces a dilemma between exploiting the current information by playing the arm with highest estimated reward and exploring all arms to collecting reward information. The primary concern of this body of literature is to find a learning algorithm which can maximize the expected cumulative reward. However, scant attention has been paid to *risk*. In many practical problems, such as clinical trials, an algorithm that yields a lower expected payout but is less risky may be preferable.

To date, there has been little agreement on the definition of risk. For example, under the MAB setting, a solution with guarantees over multiple runs of an algorithm may not satisfy the desire for a solution with low variability over a single implementation of an algorithm. Indeed, there are various risk modeling paradigms, such as, the *expected utility theory* and the *mean-variance paradigm*. In this paper, we focus on the mean-variance paradigm, which was introduced by Markowitz (1952). We seek to understand the role of risk in the mean-variance MAB problem.

1.1. Related Work

Although problems involving bandits have a long history, dating back to Thompson (1933), risk-aware bandits problem have been studied only recently. Even-Dar et al. (2006) incorporated risk into online learning problem. They initiated the investigation of explicit risk considerations in the standard models of worst-case online learning by considering the *Sharpe ratio* and the *mean-variance* criterion. Audibert et al. (2009) analyzed the expected regret and its distribution, showed that an anytime UCB and UCB-V might have large regret with non-negligible probability. Salomon & Audibert (2011) further extended this result by showing that no anytime algorithm can achieve a regret with both a small expected regret and exponential tails. These result are important steps towards the analysis of risk in MAB problems, but they are limited to the case that an algorithm's objective is to find the arm with the highest expectation.

Sani et al. (2012) considered risk-aversion in MAB problems. In particular, they studied the mean-variance risk criteria and presented the MV-LCB algorithm. Vakili & Zhao (2015) and Vakili & Zhao (2016) considered mean-variance minimization under a regret minimization framework and completed the regret analysis of Sani et al. (2012). In the best arm identification setting, David & Shimkin (2016) and David et al. (2018) studied VaR-based risk criteria. Moreover, some coherent risk measures—for example CVaR—

[1] Institute of Operations Research and Analytics, National University of Singapore, Singapore [2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore [3] Department of Mathematics, National University of Singapore, Singapore. Correspondence to: Qiuyu Zhu <qiuyu_zhu@u.nus.edu>, Vincent Y. F. Tan <vtan@nus.edu.sg>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

have been studied by Kolla et al. (2019), Xu et al. (2018). Galichet et al. (2013) presented the MaRaB algorithm which uses CVaR_α in its implementation. However, they analyzed the regret under the assumptions that $\alpha = 0$ and that the CVaR_α and average optimal arms coincide. Maillard (2013) proposed and analyzed RA-UCB which is based on the measure of entropic risk with a parameter λ. Cassel et al. (2018) proposed a general approach for MAB under risk criteria, they used *Empirical Distribution Performance Measures* as the performance metric of the algorithm. They presented and analyzed the U-UCB algorithm. All algorithms above are based on UCB or LCB ideas. To the best of our knowledge, there is no work on using Thompson Sampling in MAB whilst incorporating risk.

Another line of research concerns variations of the assumptions on the reward distributions. Bubeck et al. (2013) showed that finite moments of order 2 (i.e. finite variance) are sufficient to obtain regret bounds of the same order as if one assumes sub-Gaussian rewards. Liu & Zhao (2011) proposed the DSEE approach to complement existing work on MAB by providing results under a set of relaxed conditions on the reward distributions. Yu et al. (2018) investigated the problem on pure exploration of MAB with heavy-tailed payoffs by relaxing the assumption of payoffs with sub-Gaussian noises. These works show that it is possible to achieve meaningful regret bounds when the reward distribution is not sub-Gaussian. For example, square of Gaussian has a χ^2 distribution, which is not sub-Gaussian; hence, the analyses of these works are not applicable.

1.2. Contributions

In this paper, we focus on the MABs under the mean-variance risk criterion. Our contributions are as follows:

- **Four algorithms:** We propose three Thompson Sampling-based algorithms for Gaussian bandits—MTS, VTS, and MVTS—each suitable for use in different regimes. We also demonstrate flexibility and generality of our approach by proposing BMVTS, another Thompson-Sampling-based algorithm, but this time, for Bernoulli bandits.

- **Comprehensive regret analyses:** We provide theoretical analyses of the algorithms and show that in some regimes, the regrets either generalize existing results (Agrawal & Goyal, 2012) or meet the information-theoretic lower bound by Vakili & Zhao (2016). The analyses are novel because previous methods for risk-averse MAB problems impose a sub-Gaussian assumption on the variance of the reward distributions so are not applicable to our Gaussian setting. Thus, we need to derive new anti-concentration bounds (cf. Lemmas 3 and 4). The regret is also analyzed for BMVTS.

- **Extensive set of simulations:** We provide extensive sets of simulations for both Gaussian and Bernoulli bandits to show that our algorithms outperform state-of-the-art algorithms for mean-variance bandits. In particular, MVTS dominates LCB-based algorithms over all risk tolerances ρ.

In the majority of the paper, we use Gaussian bandits as an example to illustrate our algorithms and proof techniques, but the same method (albeit with different concentration bounds) can be use to prove regret bounds for Bernoulli bandits (cf. Theorem 4).

The rest of this paper is organized as follows. We introduce mean-variance MABs and some notations in Section 2. In Section 3, we present Thompson Sampling algorithms for mean-variance Gaussian bandits. Some regret analyses are provided in Section 4. A set of numerical simulations is reported to validate the theoretical results in Section 5. In Section 6, we conclude the discussions. Detailed/full proofs are deferred to the supplementary material.

2. Problem formulation

In this section we introduce the main notations and define mean-variance MABs. Consider a MAB ν with K arms and a single player. The problem is defined over a time horizon of length n. The mean and the variances of the reward distributions are fixed and unknown (the frequentist setting). At each time $t \in \{1, \ldots, n\}$, the player chooses one arm to play. Playing arm i at time t yields a random reward $X_{i,t}$ drawn from ν_i. All reward samples are independent conditionally to the choice of the arm. A *policy* $\pi(\cdot) : (t, A_1, X_1, \ldots, A_{t-1}, X_{t-1}) \to [K]$, is a function that specifies the action of the player at each time. The policy depends on the history $(A_1, X_1, \ldots, A_{t-1}, X_{t-1})$. Let $T_{i,n}$ denote the number of times that the player pulls arm i during the time periods $\{1, \ldots, n\}$.

In the standard MAB problem, the objective of the player is to minimize the expected cumulative regret. Here, instead, we focus on finding the arm which effectively balances its expected reward and variability. Although there are a large number of models for the trade-off between return and risk, such as *Sharpe ratio* from Sharpe (1966) and the *Knightian uncertainty* from Knight (1921), here we focus on the most popular and simple model, namely, the mean-variance model proposed by Markowitz (1952).

Definition 1 *The* mean-variance *of an arm i with mean μ_i, variance σ_i^2 and risk tolerance ρ is $\mathrm{MV}_i = \rho\mu_i - \sigma_i^2$.*

Let arm 1 be the best arm, i.e., $\mathrm{MV}_1 = \max_{i \in [K]} \mathrm{MV}_i$. Based on Definition 1, we can recover two extreme cases by considering the extremal values of the risk tolerance ρ. When $\rho = 0$, the mean-variance MAB is a variance minimization problem. As $\rho \to \infty$, the problem reduces to standard MAB in which one seeks to maximize the reward.

Given i.i.d. samples $\{X_{i,s}\}_{s=1}^{t}$ drawn from distribution ν_i, we define the empirical mean-variance as

$$\widehat{\mathrm{MV}}_{i,t} = \rho\hat{\mu}_{i,t} - \hat{\sigma}_i^2, \qquad \text{where} \tag{1}$$

$$\hat{\mu}_{i,t} = \frac{1}{t}\sum_{s=1}^{t} X_{i,s}, \quad \hat{\sigma}_{i,t}^2 = \frac{1}{t}\sum_{s=1}^{t}\left(X_{i,s} - \hat{\mu}_{i,t}\right)^2. \tag{2}$$

Let $\Gamma_{i,j} = \mu_i - \mu_j$ and $\Delta_i = \mathrm{MV}_1 - \mathrm{MV}_i$ denote, respectively, the difference between the means of arms i and j and the difference between the mean-variances of i and 1. Let $T_{i,j}$ be the number of times that arm i is pulled during first j periods. Assume throughout that $\Delta_i > 0$. We remark that using the unbiased estimate of the sample variance $\hat{\sigma}_i^2 = \frac{1}{t-1}\sum_{s=1}^{t}\left(X_{i,s} - \hat{\mu}_{i,t}\right)^2$ does not change our regret bound and conclusions. Hence, we use the definition of the sample variance in (2) for simplicity.

Given a learning policy $\pi(\cdot)$ and the reward process $\{X_{\pi(t),t}\}_{t=1}^{n}$, we define the *empirical mean-variance* as

$$\widehat{\mathrm{MV}}_n\left(\pi\right) = \rho\hat{\mu}_n\left(\pi\right) - \hat{\sigma}_n^2(\pi), \quad \text{where} \tag{3}$$

$$\hat{\mu}_n(\pi) = \frac{1}{n}\sum_{t=1}^{n} X_{\pi(t),t}, \quad \text{and} \tag{4}$$

$$\hat{\sigma}_n^2(\pi) = \frac{1}{n}\sum_{t=1}^{n}\left(X_{\pi(t),t} - \hat{\mu}_n\left(\pi\right)\right)^2. \tag{5}$$

Obviously, the optimal policy should choose arm 1 for all $t \in \{1, \dots, n\}$. For each policy π, this leads to the definition of the regret, which is the difference of the empirical mean-variance of the policy and the optimal mean-variance.

Definition 2 *The* expected regret *of a policy $\pi(\cdot)$ over n rounds is defined as*

$$\mathbb{E}\left[\mathcal{R}_n\left(\pi\right)\right] = n\big(\mathrm{MV}_1 - \mathbb{E}\big[\widehat{\mathrm{MV}}_n(\pi)\big]\big). \tag{6}$$

We remark the expectation in (6) is taken over the sample path of the rewards $\{X_{\pi(t),t}\}_{t=1}^{n}$. The expected regret can alternatively be written as the expectation of

$$\mathcal{R}_n\left(\pi\right) = \mathcal{R}_n^{(1)}\left(\pi\right) + \mathcal{R}_n^{(2)}\left(\pi\right), \quad \text{where}$$

$$\mathcal{R}_n^{(1)}\left(\pi\right) := n\sum_{i=1}^{K}\left(\rho\frac{T_{i,n}}{n}(\mu_1 - \hat{\mu}_{i,T_{i,n}}) + \frac{T_{i,n}}{n}(\hat{\sigma}_{i,T_{i,n}}^2 - \sigma_1^2)\right),$$

$$\mathcal{R}_n^{(2)}\left(\pi\right) := n\sum_{i=1}^{K}\frac{T_{i,n}}{n}(\hat{\mu}_{i,T_{i,n}} - \hat{\mu}_n(\pi))^2.$$

This definition of expected regret leads to a natural objective—to design an algorithm whose regret increases as slowly as possible as n increases. The objective of our problem is to balance the tradeoff between risk and return, but this definition of regret does not give us a view of how the components of the regret (i.e., the regret related to the risk and regret related to the return) influences the overall regret $\mathbb{E}\left[\mathcal{R}_n\left(\pi\right)\right]$. This motivates the following quantity.

Definition 3 *The* expected pseudo-regret *for a policy $\pi(\cdot)$ over n rounds is defined as*

$$\mathbb{E}\big[\widetilde{\mathcal{R}}_n(\pi)\big] = \sum_{i=2}^{K}\mathbb{E}\left[T_{i,n}\right]\Delta_i + \frac{1}{n}\sum_{i=1}^{K}\sum_{j\neq i}\mathbb{E}\left[T_{i,n}T_{j,n}\right]\Gamma_{i,j}^2. \tag{7}$$

Let us relate the expected regret with the expected pseudo-regret. The expected pseudo-regret can be divided into two parts. The first term in (7) is the regret of the expected mean-variance that results from choosing suboptimal arms, which is exactly the expectation of $\mathcal{R}_n^{(1)}\left(\pi\right)$. Given a policy $\pi(\cdot)$, the variance of the reward process $\{X_{\pi(t),t}\}_{t=1}^{n}$ also influences $\mathbb{E}[\mathcal{R}_n^{(1)}\left(\pi\right)]$, but as can be seen from the definition of $\mathcal{R}_n^{(1)}\left(\pi\right)$, the regret of the variance $n(\hat{\sigma}_n^2(\pi) - \sigma_1^2)$ is only one of the two terms that comprises $\mathcal{R}_n^{(1)}\left(\pi\right)$. We also have to take into account the regret of variance which arises due to the switching between different arms; this is the term $\mathcal{R}_n^{(2)}\left(\pi\right)$. The second term in (7) is an upper bound of $\mathbb{E}[\mathcal{R}_n^{(2)}\left(\pi\right)]$ (see the proof of Lemma 1). Because the number pulls of suboptimal arms $T_{i,n}$ for $i \neq 1$ is explicitly delineated in the expected pseudo-regret, this version of the regret is easier to work with in the sequel.

Lemma 1 *The difference between the expectations of these two expected regrets can be bounded as follows:*

$$\mathbb{E}\left[\mathcal{R}_n(\pi)\right] \leq \mathbb{E}\big[\widetilde{\mathcal{R}}_n(\pi)\big] + 3\sum_{i=1}^{K}\sigma_i^2. \tag{8}$$

This lemma shows that we can obtain a bound on the expected regret by proving a bound on the expected pseudo-regret. Hence, we focus on the analysis of $\widetilde{\mathcal{R}}_n(\pi)$. The second term in (7) can be upper bounded as follows:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{K}\sum_{j\neq i}\mathbb{E}\left[T_{i,n}T_{j,n}\right]\Gamma_{i,j}^2 \\
&= \frac{1}{n}\left(\sum_{j\neq 1}\mathbb{E}\left[T_{1,n}T_{j,n}\right]\Gamma_{1,j}^2 + \sum_{i=2}^{K}\sum_{j\neq i}\mathbb{E}\left[T_{i,n}T_{j,n}\right]\Gamma_{i,j}^2\right) \\
&\leq \frac{1}{n}\left(\sum_{j\neq 1}n\mathbb{E}\left[T_{j,n}\right]\Gamma_{1,j}^2 + \sum_{i=2}^{K}n\mathbb{E}\left[T_{i,n}\right]\Gamma_{i,\max}^2\right) \\
&\leq 2\sum_{i=2}^{K}\mathbb{E}\left[T_{i,n}\right]\Gamma_{i,\max}^2,
\end{aligned} \tag{9}$$

Algorithm 1 Update $(\hat{\mu}_{i,t-1}, T_{i,t-1}, \alpha_{i,t-1}, \beta_{i,t-1})$

```
Input: Prior parameters $(\hat{\mu}_{i,t-1}, T_{i,t-1}, \alpha_{i,t-1}, \beta_{i,t-1})$ and new sample $X_{i,t}$
Update the mean: $\hat{\mu}_{i,t} = \frac{T_{i,t-1}}{T_{i,t-1}+1}\hat{\mu}_{i,t-1} + \frac{1}{T_{i,t-1}+1}X_{i,t}$
Update the number of samples and the shape parameter: $T_{i,t} = T_{i,t-1} + 1, \alpha_{i,t} = \alpha_{i,t-1} + \frac{1}{2}$
Update the rate parameter: $\beta_{i,t} = \beta_{i,t-1} + \frac{T_{i,t-1}}{T_{i,t-1}+1} \cdot \frac{(X_{i,t}-\hat{\mu}_{i,t-1})^2}{2}$
```

where $\Gamma^2_{i,\max} = \max\{(\mu_i - \mu_j)^2 : j = 1, \ldots, K\}$.

By applying (9), the pseudo-regret can be written as

$$\mathbb{E}\big[\widetilde{\mathcal{R}}_n(\pi)\big] \leq \sum_{i=2}^{K} \mathbb{E}\left[T_{i,n}\right] \left(\Delta_i + 2\Gamma^2_{i,\max}\right). \tag{10}$$

This inequality shows that it suffices to bound the number of pulls of each suboptimal arm $i \neq 1$.

3. Algorithms for mean-variance MAB

In this section, we introduce three Thompson Sampling-based risk-averse MAB algorithms. To illustrate the design of the algorithms, we consider the Gaussian bandits in detail, i.e., $\nu \in \mathcal{E}^K_{\mathcal{N}}(1) = \{\nu = (\nu_1, \ldots, \nu_K) : \nu_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \sigma_i^2 \leq 1\}$. We provide short discussions concerning Bernoulli bandits in Sections 3.3 and 4.4.

An important step of Thompson Sampling algorithms is the updating of parameters based on Bayes rule. Consider the general prior for the Gaussian with unknown means and precisions, i.e., the Normal-Gamma prior. Let μ and τ be the mean and precision of the Gaussian respectively. If $(\mu, \tau) \sim \text{Normal-Gamma}(\mu, T, \alpha, \beta)$, then $\tau \sim \text{Gamma}(\alpha, \beta)$, and $\mu|\tau \sim \mathcal{N}(\mu, 1/(\tau T))$. Since the Normal-Gamma distribution is the conjugate prior for the Gaussian with unknown mean and variance, we use Algorithm 1 to update these parameters.

3.1. Thompson Sampling algorithms for mean learning and variance learning

We propose a variant of Thompson Sampling algorithm to solve the mean-variance Gaussian MAB problem when ρ is large; we call this Mean Thompson Sampling (MTS). In each period, the variance of each arm will be estimated and the algorithm sequentially updates the posterior mean of each arm. We propose another algorithm to handle the case in which ρ is small. We call this algorithm Variance Thompson Sampling (VTS). VTS estimates the mean of each arm and updates the posterior precision of each arm. At the beginning of each period, VTS samples a precision from the posterior and then selects the optimal arm according to the estimates of the mean and the Thompson samples of the precision. Pseudocodes of the MTS and VTS algorithms are shown in Algorithm 2.

Algorithm 2 Thompson Sampling for Mean Learning (MTS) and Variance Learning (VTS)

```
Input: $\hat{\mu}_{i,0} = 0, T_{i,0} = 0, \alpha_{i,0} = \frac{1}{2}, \beta_{i,0} = \frac{1}{2}$.
for each $t = 1, 2, \ldots, K$ do
    Play arm $t$ and observe the reward $X_{t,t}$
    Update$(\hat{\mu}_{t,t-1}, T_{t,t-1}, \alpha_{t,t-1}, \beta_{t,t-1})$
end for
for each $t = K+1, \ldots,$ do
    MTS:
        • Sample $\theta_{i,t}$ from $\mathcal{N}\left(\hat{\mu}_{i,t-1}, 1/T_{i,t-1}\right)$.
        • Play arm $i\left(t\right) = \arg\max_i \rho\theta_{i,t} - 2\beta_{i,t-1}$ and observe reward $X_{i(t),t}$.
    VTS:
        • Sample $\tau_{i,t}$ from $\text{Gamma}\left(\alpha_{i,t-1}, \beta_{i,t-1}\right)$.
        • Play arm $i(t) = \arg\max_{i\in[K]} \rho\hat{\mu}_{i,t-1} - 1/\tau_{i,t}$ and observe reward $X_{i(t),t}$.
    Update$(\hat{\mu}_{i(t),t-1}, T_{i(t),t-1}, \alpha_{i(t),t-1}, \beta_{i(t),t-1})$
end for
```

3.2. A Thompson Sampling algorithm for the Gaussian mean-variance MAB

The proposed algorithms can effectively solve the risk-averse MAB problem in two extreme scenarios (e.g., large or small ρ). However, it is difficult to decide on a suitable threshold for a player to choose the algorithm she needs because she does not know the true means and variances (and neither does she know ρ). In this section, we propose a *combined* or *unified* Thompson Sampling algorithm to address this problem. The player chooses a prior over the set of feasible bandits parameters for both the mean and precision. In each round, the player samples a pair of parameters from each posterior and plays an arm according to the optimal action under these parameters.

The Mean-Variance Thompson Sampling (MVTS) algorithm, shown in Algorithm 3, uses the conjugate prior of a Gaussian which is parametrized by the mean and precision. When ρ is small, the mean-variance MAB problem reduces to a variance minimization problem. In this regime, MVTS is consistent with VTS. On the other hand, when $\rho \to \infty$, the problem reduces to the standard reward maximization problem. In this case, MVTS is consistent with MTS. Hence, we expect that MVTS performs well over all $\rho \in \mathbb{R}_+$. We show that this is indeed the case in Theorem 3.

Algorithm 3 Thompson Sampling for Gaussian mean-variance bandits (MVTS)

```
Input: $\hat{\mu}_{i,0} = 0, T_{i,0} = 0, \alpha_{i,0} = \frac{1}{2}, \beta_{i,0} = \frac{1}{2}$.
for each $t = 1, 2, \ldots, K$ do
    Play arm $t$ and update $\hat{\mu}_{t,t} = X_{t,t}$.
    $\mathrm{Update}(\hat{\mu}_{t,t-1}, T_{t,t-1}, \alpha_{t,t-1}, \beta_{t,t-1})$
end for
for each $t = K+1, K+2, \ldots,$ do
    Sample $\tau_{i,t}$ from $\mathrm{Gamma}(\alpha_{i,t-1}, \beta_{i,t-1})$.
    Sample $\theta_{i,t}$ from $\mathcal{N}(\hat{\mu}_{i,t-1}, 1/T_{i,t-1})$
    Play arm $i(t) = \arg\max_{i \in [K]} \rho\theta_{i,t} - 1/\tau_{i,t}$ and observe reward $X_{i(t),t}$
    $\mathrm{Update}(\hat{\mu}_{i(t),t-1}, T_{i(t),t-1}, \alpha_{i(t),t-1}, \beta_{i(t),t-1})$
end for
```

Algorithm 4 Thompson Sampling for Bernoulli mean-variance bandits (BMVTS)

```
Input: $\alpha_{i,1} = 1, \beta_{i,1} = 1$.
for each $t = 1, 2, \ldots,$ do
    Sample $\theta_{i,t}$ from $\mathrm{Beta}(\alpha_{i,t}, \beta_{i,t})$.
    Play arm $i(t) = \arg\max_{i \in [K]} \rho\theta_{i,t} - \theta_{i,t}(1 - \theta_{i,t})$ and observe reward $X_{i(t),t}$
    Update parameters: $\alpha_{i,t+1} = \alpha_{i,t} + X_{i(t),t}$,
        $\beta_{i,t+1} = \beta_{i,t} + (1 - X_{i(t),t})$
end for
```

3.3. A Thompson sampling algorithm for Bernoulli mean-variance bandits

In this section, we present the BMVTS algorithm for Bernoulli mean-variance bandit problem, which is BMVTS. Under Bernoulli bandits setting, the reward of arm i is generated from Bernoulli distribution with success probability p_i. Hence, we use the Beta distribution as the prior of each arm's reward distribution. Pseudocode of the BMVTS algorithm is shown in Algorithm 4.

4. Regret analysis

We present our regret bounds and a sketch of the proof for MVTS. There are some non-trivial technical details that are required for MVTS because of the random variables involved in some error events; see the dependencies of the random variables in Figure 2, Lemmas 3 and 4 and their accompanying discussions. The *asymptotic* regret bounds that we present here are derived from the *finite-horizon* regret bounds, which are available in the supplementary materials (see Theorem S-1, S-2, S-3).

4.1. Regret analysis for MTS

Theorem 1 *If $\rho > \max\left\{\sigma_1^2/\Gamma_i : i = 1, 2, \ldots, K\right\}$, the asymptotic expected regret produced by MTS for mean-variance Gaussian bandits satisfies*

$$\varlimsup_{n\to\infty} \frac{\mathbb{E}\big[\widetilde{\mathcal{R}}_n(\mathrm{MTS})\big]}{\log n} \le \sum_{i=2}^{K} \frac{2\rho^2}{(\rho\Gamma_{1,i} - \sigma_1^2)^2}\left(\Delta_i + 2\Gamma_{i,\max}^2\right). \tag{11}$$

By Lemma 1, the same bound holds for $\mathbb{E}[\mathcal{R}_n(\mathrm{MTS})]$. This remark also applies to Theorems 2 and 3.

Remark 1 (The assumption) The reason for the assumption on ρ is that the mean-variance has the same order as the mean of each arm. Thus the mean is the dominant term. However, in our numerical simulations, we can observe that MTS still performs well even this condition is not met in practice (see Figure 4).

4.2. Regret analysis for VTS

Theorem 2 *Let $h(x) = \frac{1}{2}(x - 1 - \log x)$. If $\rho \le \min\left\{\Delta_i/\Gamma_i : \Delta_i/\Gamma_i > 0\right\}$ and $\Gamma_i^2 > 2\sigma_1^2 h\left(\sigma_i^2/\sigma_1^2\right)$ for all i, the asymptotic expected regret of VTS for mean-variance Gaussian bandits satisfies*

$$\varlimsup_{n\to\infty} \frac{\mathbb{E}\big[\widetilde{\mathcal{R}}_n(\mathrm{VTS})\big]}{\log n} \le \sum_{i=2}^{K} \frac{1}{h\left(\sigma_i^2/\sigma_1^2\right)}\left(\Delta_i + 2\Gamma_{i,\max}^2\right). \tag{12}$$

Remark 2 (Assumptions) Bubeck et al. (2013) show that the only condition we need to achieve the same form of regret bound as the sub-Gaussian case is that the reward distributions have finite variance. Here, we consider Gaussian bandits; hence this assumption is fulfilled, leading to the bound $\mathbb{E}[\mathcal{R}_n(\mathrm{VTS})] = O(\log n)$. However, other works on risk-averse MABs require more stringent conditions on ν, e.g., Vakili & Zhao (2015) assume that the empirical variance of ν_i is sub-Gaussian and Sani et al. (2012) use the assumption that the reward is bounded almost surely.

Remark 3 (Scale invariance) The bound in (12) depends only on the ratio of the variances, which is similar to the fact that the regret for the standard MAB depends only on the differences of the means. This justifies our assumption that $\nu \in \mathcal{E}_{\mathcal{N}}^K(1)$ since we can rescale the variances.

4.3. Regret analysis for MVTS

Theorem 3 *The asymptotic expected regret of MVTS for mean-variance Gaussian bandits satisfies*

$$\varlimsup_{n\to\infty} \frac{\mathbb{E}\big[\widetilde{\mathcal{R}}_n(\mathrm{MVTS})\big]}{\log n} \le \sum_{i=2}^{K} \max\left\{\frac{2}{\Gamma_{1,i}^2}, \frac{1}{h(\sigma_i^2/\sigma_1^2)}\right\}\left(\Delta_i + 2\Gamma_{i,\max}^2\right). \tag{13}$$

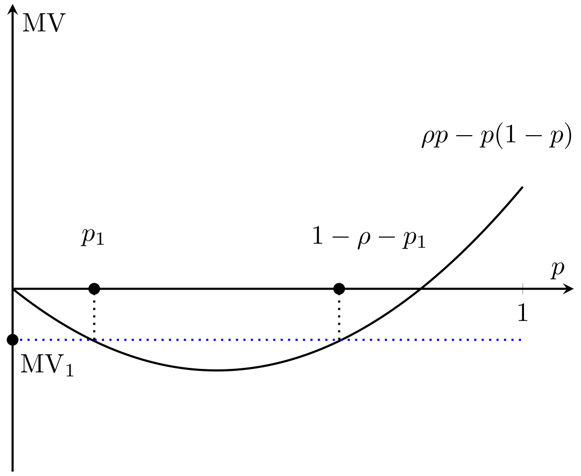

Figure 1. Mean-variance of Bernoulli bandit

Remark 4 (The extreme cases) When take ρ to its extremal values, we can elaborate on discussion after Definition 1 in terms of the regrets. First, as $\rho \to \infty$, the mean will dominate the mean-variance. Hence, $\mathbb{P}_t(\widehat{\text{MV}}_{i,t} \ge \text{MV}_1 - (\rho+1)\varepsilon) \to \mathbb{P}_t(\theta_{i,t} \ge \mu_1 - \varepsilon)$. By applying the same proof technique as that for MVTS, so we have

$$\varlimsup_{n\to\infty}\ \varlimsup_{\rho\to\infty} \frac{\mathbb{E}\big[\widetilde{\mathcal{R}}_n\,(\text{MVTS})\big]}{\rho\log n} \le \sum_{i=2}^{K} \frac{2}{\Gamma_{1,i}}.$$

This shows us that MVTS has a similar behavior as MTS when ρ is large. When $\rho \to 0$, we have $\Delta_i \to \sigma_i^2 - \sigma_1^2$, $\mathbb{P}_t(\widehat{\text{MV}}_{i,t} \ge \text{MV}_1 - (1+\rho)\varepsilon) \to \mathbb{P}_t(-\frac{1}{\tau_{i,t}} \ge -\sigma_1^2 - \varepsilon)$, so that

$$\varlimsup_{n\to\infty}\ \varlimsup_{\rho\to 0} \frac{\mathbb{E}\big[\widetilde{\mathcal{R}}_n\,(\text{MVTS})\big]}{\log n} \le \sum_{i=2}^{K} \frac{\sigma_i^2 - \sigma_1^2 + 2\Gamma_{i,\max}^2}{h(\sigma_i^2/\sigma_1^2)}.$$

Thus, MVTS can learn the variances. Compare this to (12) and note that $\Delta_i \to \sigma_i^2 - \sigma_1^2$ so when $\rho \to 0$, the problem reduces to the variance minimization problem. These conclusions are corroborated by our numerical simulations.

4.4. Regret analysis for BMVTS

Theorem 4 *If $\rho \in (0,1)$, then the asymptotic expected regret of BMVTS for mean-variance Bernoulli bandits satisfies*

$$\varlimsup_{n\to\infty} \frac{\mathbb{E}\big[\widetilde{\mathcal{R}}_n\,(\text{BMVTS})\big]}{\log n} \le \sum_{i=2}^{K} \max\left\{\frac{1}{2\Gamma_i^2}, \frac{1}{2\,(1-\rho-p_1-p_i)^2}\right\}\left(\Delta_i + 2\Gamma_{i,\max}^2\right). \tag{14}$$

Remark 5 (The asssumption on ρ) If $\rho \ge 1$, the mean-variance $\text{MV}(p) = \rho p - p(1-p)$ is increasing in $p \in [0,1]$, reducing to a standard MAB. Hence, we consider $\rho \in (0,1)$.

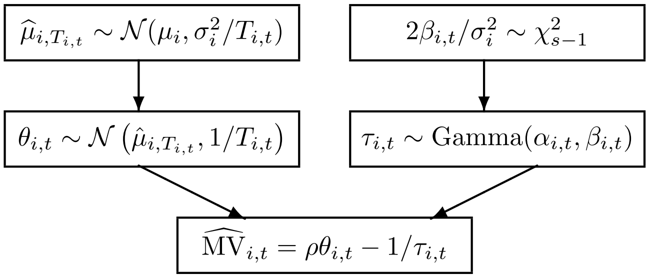
$\widehat{\mu}_{i,T_{i,t}} \sim \mathcal{N}(\mu_i, \sigma_i^2/T_{i,t})$
$2\beta_{i,t}/\sigma_i^2 \sim \chi_{s-1}^2$
$\theta_{i,t} \sim \mathcal{N}\left(\hat{\mu}_{i,T_{i,t}}, 1/T_{i,t}\right)$
$\tau_{i,t} \sim \text{Gamma}(\alpha_{i,t}, \beta_{i,t})$
$\widehat{\text{MV}}_{i,t} = \rho\theta_{i,t} - 1/\tau_{i,t}$

Figure 2. Hierarchical structure of the mean-variance Thompson samples in MVTS.

Remark 6 (The bound) The mean-variance of the best arm is $\text{MV}_1 = \rho p_1 - p_1(1-p_1)$. Hence, this value of the mean-variance MV_1 corresponds to *two* different p_1's (i.e., p_1 and $1-\rho-p_1$). See Fig. 1. This introduces two error events in which one suboptimal arm performs better than the best arm; this is the reason for the appearance of the "max" in (14).

4.5. Comparisons with lower bounds

Regret bound in Theorem 1

Recall that $\Delta_i = \sigma_i^2 - \sigma_1^2 + \rho\Gamma_{1,i}$, so that in the regime in which $\rho \to +\infty$, $\Delta_i = \Theta(\rho\Gamma_{1,i})$ and we observe that

$$\lim_{\rho\to\infty}\ \varlimsup_{n\to\infty} \frac{\mathbb{E}\big[\widetilde{\mathcal{R}}_n\,(\text{MTS})\big]}{\rho\log n} \le \sum_{i=2}^{K} \frac{2}{\Gamma_{1,i}}. \tag{15}$$

This is consistent with the fact that for $\rho \to \infty$, the mean-variance problem reduces to the maximization of the reward (without the risk aspect), for which Thompson Sampling is already known to be nearly-optimal (Agrawal & Goyal (2012)). The bound in (15) coincides with Theorem 36.3 in Lattimore & Szepesvári (2020). Hence Theorem 1 generalizes the analysis to all $\rho < \infty$.

Order optimality of Theorem 2

Vakili & Zhao (2015) proved that the expected regret of any consistent algorithm for mean-variance MAB is $\Omega\big(\frac{\log n}{\Delta^2}\big)$ where $\Delta = \min_{i\neq 1}\Delta_i$. In Theorem 2, by Taylor's theorem, $h(x) = (x-1)^2/4 + o((x-1)^2)$ as $x \to 1$. Consider the regime in which for all $i \neq 1$, $\sigma_i^2 \to \sigma_1^2$ and $\rho = o(\sigma_i^2 - \sigma_1^2)$ as $\sigma_i^2 - \sigma_1^2 \to 0$. In this case, $\Delta_i = \Theta(\sigma_i^2 - \sigma_1^2)$ and one has $h(\sigma_i^2/\sigma_1^2) = \Theta(\Delta_i^2)$ as $\Delta_i \to 0$. Further, since $\Gamma_{i,\max} = \Theta(1)$ for all i, the upper bound in (12) reduces to $O(\frac{1}{\Delta^2})$ and so the expected regret scales as $O\big(\frac{\log n}{\Delta^2}\big)$, asymptotically matching the lower bound in Vakili & Zhao (2015). Thus, in this particular regime, VTS is *information-theoretically optimal*. This regime can be thought of as a "hard instance" for variance learning/minimization since all the σ_i's are close to one another and ρ is asymptotically smaller than $\sigma_i^2 - \sigma_1^2$; the latter implying that the importance of the mean part of the mean-variance objective is de-emphasized.

By Remark 4, MVTS particularizes to MTS and VTS when $\rho \to \infty$ and $\rho \to 0^+$ respectively. According to the above discussions, we conclude that MVTS is order optimal when ρ assumes these extremal values.

4.6. Proof Sketch of Theorem 3

We use MVTS as an example to demonstrate the key steps of the proofs. There are two main difficulties in proving an upper bound of the regret for MVTS. First, the Thompson sample of precision is not sub-Gaussian. Hence, we need to derive sufficiently tight lower and upper bounds on the tail probability of the posterior distribution (which is not sub-Gaussian). The other difficulty is the hierarchical structure of the parameters in MVTS, which is more complicated than standard MAB (see Figure 2). We note that the Thompson samples of the mean and precision are from different distributions. Deriving upper and lower tail bounds for $\widehat{\mathrm{MV}}_{i,t}$ is a major challenge because the normal distribution is sub-Gaussian, but the Gamma is only sub-exponential.

From (10), we know that it suffices to prove an upper bound of $\mathbb{E}\left[T_{i,n}\right]$. Let the Thompson sample of the mean-precision pair of arm i for time t be $(\theta_{i,t}, \tau_{i,t})$. Let the sample mean-variance be $\widehat{\mathrm{MV}}_{i,t} = \rho\theta_{i,t} - 1/\tau_{i,t}$. Define

$$E_i(t) := \left\{\widehat{\mathrm{MV}}_i(t) \le \mathrm{MV}_1 - (1+\rho)\varepsilon\right\}$$

which is the event that the Thompson sample of arm i is $(1+\rho)\varepsilon$-smaller than the optimal arm at period t. Event $E_i(t)$ is highly likely to occur when the algorithm has explored sufficiently since arm 1 is optimal. However, the algorithm does not choose arm i when $E_i(t)$ occurs with high probability because it chooses the arm with the maximal mean-variance in each period. The algorithm chooses arm i when $E_i^c(t)$, an event with small probability (under Thompson sampling), occurs. The expectation can be divided into two parts as follows.

Lemma 2 (Lattimore & Szepesvári (2020)) *Let* $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot|A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ *be the probability measure conditioned on the history up to time* $t-1$ *and* $G_{is} = \mathbb{P}_t\left(E_i(t)^c | T_{i,t} = s\right)$. *Then,*

$$\mathbb{E}[T_{i,n}] \le \mathbb{E}\left[\sum_{s=0}^{n-1}\left(\frac{1}{G_{1s}} - 1\right)\right] + \mathbb{E}\left[\sum_{s=0}^{n-1}\mathbb{I}\left\{G_{is} > \frac{1}{n}\right\}\right] + 1. \tag{16}$$

Similar to the standard MAB formulation, the first term can be controlled as the Gamma distribution has a heavy tail. Specifically,

$$\frac{1}{G_{1s}} - 1 \ge \frac{2}{\mathbb{P}_t(\widehat{\sigma}_{i,s}^2 \ge \sigma_1^2, \widehat{\mu}_{i,s} \le \mu_1)} - 1.$$

Hence, to bound the first term in (16), we need a lower bound on the tail of the distribution of the empirical mean-variance. Note that given $T_{1,t} = s$, the random variables $\widehat{\mu}_{1,s} = \mu, \widehat{\sigma}_{1,s}^2 = \sigma^2$, $\theta_{1,t}$ and $\tau_{1,t}$ are independent because we sample them from different distributions independently.

Lemma 3 (Tail Lower Bound) *We have*

$$\begin{aligned}
&\mathbb{P}_t\big(\widehat{\mathrm{MV}}_{1,t} \ge \mathrm{MV}_1 - (1+\rho)\varepsilon \,\big|\, T_{1,t} = s, \widehat{\mu}_{1,s} = \mu, \widehat{\sigma}_{1,s}^2 = \sigma^2\big) \\
&\ge \begin{cases} \mathbb{P}_t\left(\frac{1}{\tau_{1,t}} - \sigma_1^2 \le \varepsilon\right) \cdot \mathbb{P}_t\left(\theta_{1,t} - \mu_1 \ge -\varepsilon\right) & \\ & \text{if } \ \sigma^2 \ge \sigma_1^2, \mu \le \mu_1 \\ \frac{1}{2}\mathbb{P}_t\left(\frac{1}{\tau_{1,t}} - \sigma_1^2 \le \varepsilon\right) & \text{if } \ \sigma^2 \ge \sigma_1^2, \mu > \mu_1 \\ \frac{1}{2}\mathbb{P}_t\left(\theta_{1,t} - \mu_1 \ge -\varepsilon\right) & \text{if } \ \sigma^2 < \sigma_1^2, \mu \le \mu_1 \\ \frac{1}{4} & \text{if } \ \sigma^2 < \sigma_1^2, \mu > \mu_1 \end{cases}.
\end{aligned}$$

Lemma 3 helps us to split the lower bounding into two parts. Now we can deal with mean and variance separately. The tail probability bound of Gaussian distribution is standard, but for Gamma distribution, we devise a new and non-standard anti-concentration bound that is crucial in the analysis of Thompson Sampling for mean-variance MABs.

Lemma 4 *For a Gamma random variable* $X \sim \mathrm{Gamma}\left(\alpha, \beta\right)$ *with shape* $\alpha \ge 2$ *and rate* $\beta > 0$,

$$\mathbb{P}\left(X \ge x\right) \ge \frac{1}{\Gamma(\alpha)}\exp\left(-\beta x\right)\left(1 + \beta x\right)^{\alpha - 1}, \text{ for } x > 0.$$

We obtain an upper bound of the first term in (16) by plugging the bound in Lemma 4 into the terms in Lemma 3 and integrating the conditional probability over (μ, σ^2).

For the second term, essentially, we need to bound $\mathbb{P}_t\left(G_{is} > 1/n\right)$. By direct computation,

$$\begin{aligned}
&\mathbb{P}_t\big(\widehat{\mathrm{MV}}_{i,s} \ge \mathrm{MV}_1 - (1+\rho)\varepsilon \,\big|\, T_{i,t} = s, \widehat{\mu}_{i,s} = \mu, \widehat{\sigma}_{i,s}^2 = \sigma^2\big) \\
&= \mathbb{P}_t\left(\rho\theta_{i,t} - \frac{1}{\tau_{i,t}} \ge \rho\mu_1 - \sigma_1^2 - (1+\rho)\varepsilon\right) \\
&= \mathbb{P}_t\left(\rho(\theta_{i,t} - \mu_1) + \left(\sigma_1^2 - \frac{1}{\tau_{i,t}}\right) \ge -(1+\rho)\varepsilon\right) \\
&\le \mathbb{P}_t\left(\theta_{i,t} - \mu_1 \ge -\varepsilon\right) + \mathbb{P}_t\left(\frac{1}{\tau_{i,t}} - \sigma_1^2 \le \varepsilon\right).
\end{aligned}$$

Hence, we have following key relationship

$$\left\{G_{is} > \frac{1}{n}\right\} \implies \begin{cases} \mathbb{P}_t\left(\tau_i(t) \ge \frac{1}{\sigma_i^2 + \varepsilon}\right) \ge \frac{1}{2n} \\ \text{or} \\ \mathbb{P}_t\left(\theta_{i,t} - \mu_1 \ge -\varepsilon\right) \ge \frac{1}{2n} \end{cases}. \tag{17}$$

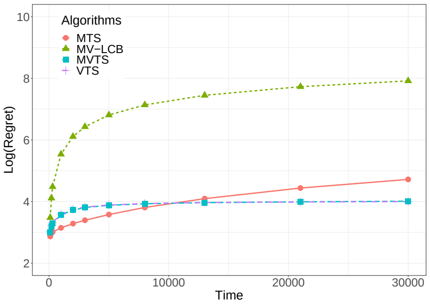

Figure 3. Regrets for $\rho = 10^{-3}$

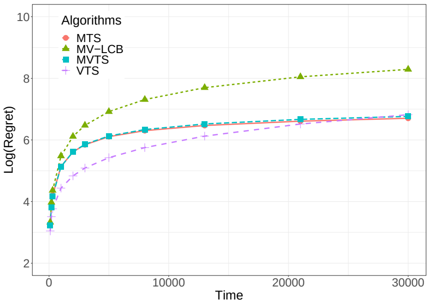

Figure 4. Regrets for $\rho = 1$

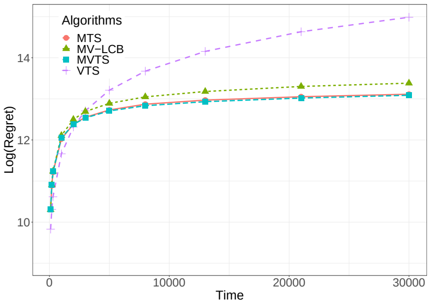

Figure 5. Regrets for $\rho = 1000$

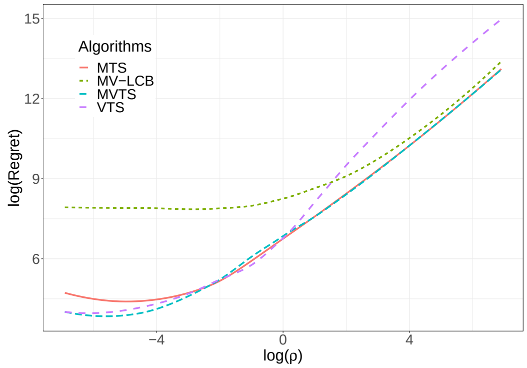

Figure 6. Regret of Gaussian MV MAB with $K = 15$.

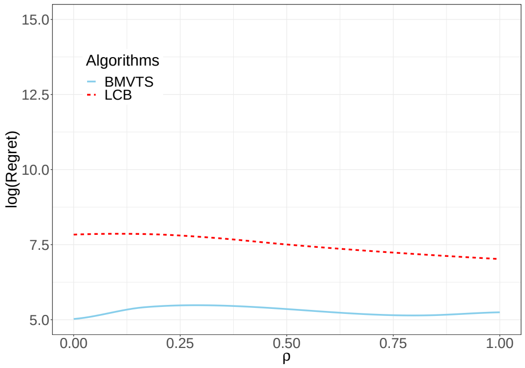

Figure 7. Regret of Bernoulli MV MAB with $K = 15$..

This relation presents a method to bound the probability of

$$\left\{G_{is} > \frac{1}{n}\right\} = \left\{\mathbb{P}_t\left(\rho\theta_{i,t} - \frac{1}{\tau_{i,t}} \geq \mathrm{MV}_1 - (1+\rho)\varepsilon\right) > \frac{1}{n}\right\}.$$

The probability of this event is *a priori* not straightforward to bound because $\theta_{i,t}$ and $\tau_{i,t}$ are random variables and so are the parameters that define them (cf. Figure 2). However, equipped with (17), we can decouple the Thompson samples representing the mean and variance. To bound the second term on right of (16), we use an upper bound on the tail of the Gamma distribution.

Lemma 5 (Harremoës (2016)) *Under the conditions of Lemma 4,*

$$\mathbb{P}\left(X \geq x\right) \leq \exp\left(-2\alpha h\left(\frac{\beta x}{\alpha}\right)\right), \quad \text{for } x > \frac{\alpha}{\beta}. \quad (18)$$

Plugging the bounds in Lemmas 3 and 4 into (16) in Lemma 2 allows us to bound the regret of MVTS. We present complete proofs of the regrets of MTS, VTS and MVTS in the supplementary material.

5. Numerical Simulations

There are other algorithms that achieve the optimal regret bound, such as MV-LCB (Vakili & Zhao, 2016). Due to the complexity of the problem and the differences in the assumptions on the reward distributions (e.g., MV-LCB in Vakili & Zhao (2016) assumes the variances of the arm distributions are sub-Gaussian), it is difficult to perform a fair comparison of their theoretical regret bounds. We emphasize though that our analyses require less stringent assumptions. Hence, we compare these algorithms via extensive numerical simulations in this section. The R code for all our experiments is provided along with this submission.

We report numerical simulations to validate our theoretical results in the previous sections. We consider the variance minimization problem ($\rho = 10^{-3}$), the expected reward maximization problem ($\rho = 1000$), and an intermediate case ($\rho = 1$). The $K = 15$ Gaussian arms are set to the same as the experiments from Sani et al. (2012) (i.e. $\mu = (0.1, 0.2, 0.23, 0.27, 0.32, 0.32, 0.34, 0.41, 0.43, 0.54, 0.55, 0.56, 0.67, 0.71, 0.79)$, $\sigma_i^2 = (0.05, 0.34, 0.28, 0.09, 0.23, 0.72, 0.19, 0.14, 0.44, 0.53, 0.24, 0.36, 0.56, 0.49, 0.85)$). We run MV-LCB, MTS, VTS and MVTS.

In Figures 3, 4 and 5 we present the expected regret $\mathcal{R}_n(\pi)$, which is averaged over 500 runs. The standard deviations of the regrets are small compared to the averages, and therefore are omitted from the all plots. For $\rho = 10^{-3}$, VTS outperforms all the other algorithms. For $\rho = 1000$, MTS has the smallest regret compared to all the other algorithms. For

$\rho = 1$, MTS, VTS and MVTS have similar performances, all of which are better than MV-LCB.

In order to validate our algorithms further and to observe how they perform as functions of ρ, we ran our algorithms with different $\rho \in [10^{-3}, 1000]$. The time horizon $n = 30,000$ is fixed and the regret is averaged over 500 runs. We report the regrets in Fig. 6. As expected, the regret of MVTS coincides with VTS when ρ is small and with MTS when ρ is large. We see also a significant performance improvement of MVTS compared to MV-LCB for all $\rho \in \mathbb{R}_{+}$.

The experiments on Bernoulli mean-variance bandits are presented in Fig. 7. Here the arm distributions are Bernoulli distributions with success probabilities $(0.1, 0.2, 0.23, 0.27, 0.32, 0.32, 0.34, 0.41, 0.43, 0.54, 0.55, 0.56, 0.67, 0.71, 0.79)$. The regret is averaged over 500 runs with a fixed time horizon $n = 30,000$. We also designed and implemented an LCB-based algorithm (analogous to those designed by Sani et al. (2012) and Vakili & Zhao (2016)) for Bernoulli mean-variance bandits. However, Fig. 4 clearly show that BMVTS significantly outperforms the LCB-based algorithm over all $\rho \in (0, 1)$.

6. Conclusion

To the best of our knowledge, this is the first work applying Thompson sampling to solve risk-averse MAB problems. We proved regret bounds that are asymptotically tight in certain regimes and recover known results in other regimes. Experimental results show that our algorithms, particularly MVTS beats the state-of-the-art LCB-style MAB algorithms for Bernoulli and Gaussian mean-variance bandits over all risk tolerance parameters ρ.

There are many different methods to model the risk-return trade-off, such as CVaR (Kolla et al. (2019), Xu et al. (2018). Galichet et al. (2013)) and *entropy risk* (Maillard (2013)) among others. Hence, more work is needed to explore the performance of Thompson Sampling, or indeed other algorithms, using different risk measures. We leave the regret analyses for other risk measures and comparisons to the methodologies herein for future work.

Acknowledgements

We would like to express our gratitude to the anonymous reviewers for their valuable comments which have helped to improve the presentation of this paper. We also thank Joel Chang for pointing out some typos.

This work is supported in part by a Singapore National Research Foundation (NRF) Fellowship under grant number R-263-000-D02-281.

References

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1965.

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

Audibert, J.-Y., Munos, R., and Szepesvári, C. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

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

Cassel, A., Mannor, S., and Zeevi, A. A general approach to multi-armed bandits under risk criteria. *arXiv preprint arXiv:1806.01380*, 2018.

Csiszár, I. and Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

David, Y. and Shimkin, N. Pure exploration for max-quantile bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 556–571. Springer, 2016.

David, Y., Szörényi, B., Ghavamzadeh, M., Mannor, S., and Shimkin, N. Pac bandits with risk constraints. In *ISAIM*, 2018.

Even-Dar, E., Kearns, M., and Wortman, J. Risk-sensitive online learning. In *International Conference on Algorithmic Learning Theory*, pp. 199–213. Springer, 2006.

Galichet, N., Sebag, M., and Teytaud, O. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pp. 245–260, 2013.

Harremoës, P. Bounds on tail probabilities in exponential families. *arXiv preprint arXiv:1601.05179*, 2016.

Knight, F. H. Risk, uncertainty and profit, 1921.

Kolla, R. K., Jagannathan, K., et al. Risk-aware multi-armed bandits using conditional value-at-risk. *arXiv preprint arXiv:1901.00997*, 2019.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

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

Maillard, O.-A. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 218–233. Springer, 2013.

Markowitz, H. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Salomon, A. and Audibert, J.-Y. Deviations of stochastic bandit regret. In *International Conference on Algorithmic Learning Theory*, pp. 159–173. Springer, 2011.

Sani, A., Lazaric, A., and Munos, R. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pp. 3275–3283, 2012.

Sharpe, W. F. Mutual fund performance. *The Journal of Business*, 39(1):119–138, 1966.

Thompson, W. R. 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, S. and Zhao, Q. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1330–1335. IEEE, 2015.

Vakili, S. and Zhao, Q. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

Xu, J., Haskell, W. B., and Ye, Z. Index-based policy for risk-averse multi-armed bandit. *arXiv preprint arXiv:1809.05385*, 2018.

Yu, X., Shao, H., Lyu, M. R., and King, I. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 937–946, 2018.