J SYST SCI SYST ENG
Vol. 32, No. 3, June 2023, pp. 267–288
DOI: https://doi.org/10.1007/s11518-022-5541-9
ISSN: 1004-3756 (paper), 1861-9576 (online)
CN 11-2983/N

Risk-averse Contextual Multi-armed Bandit Problem with Linear Payoffs

Yifan Lin*, Yuhao Wang*, Enlu Zhou
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
ylin429@gatech.edu, yuhaowang@gatech.edu, enlu.zhou@isye.gatech.edu (✉)

Abstract. In this paper we consider the contextual multi-armed bandit problem for linear payoffs under a risk-averse criterion. At each round, contexts are revealed for each arm, and the decision maker chooses one arm to pull and receives the corresponding reward. In particular, we consider mean-variance as the risk criterion, and the best arm is the one with the largest mean-variance reward. We apply the Thompson sampling algorithm for the disjoint model, and provide a comprehensive regret analysis for a variant of the proposed algorithm. For T rounds, K actions, and d-dimensional feature vectors, we prove a regret bound of $O((1+\rho+\frac{1}{\rho})d\ln T\ln\frac{K}{\delta}\sqrt{dKT^{1+2\epsilon}\ln\frac{K}{\delta}\frac{1}{\epsilon}})$ that holds with probability $1-\delta$ under the mean-variance criterion with risk tolerance ρ, for any $0<\epsilon<\frac{1}{2}$, $0<\delta<1$. The empirical performance of our proposed algorithms is demonstrated via a portfolio selection problem.

Keywords: Multi-armed bandit, context, risk-averse, Thompson sampling

1. Introduction

The multi-armed bandit (MAB) problem is a classic online decision-making problem with limited feedback. In the standard MAB problem, in each of T rounds, a decision maker plays one of the K arms and receives a reward (also called payoff) of that arm. It has a wide variety of real-world applications, such as clinical trials, online advertisement, and portfolio selection. In certain situations, the decision maker may also be provided with contexts (also known as covariates or side information). For example, in personalized web services, the decision maker also knows the demographic, geographic, and behavioral information of the user (Li et al. 2010), which may be useful to infer the conditional average reward of an arm and allows the decision maker to personalize decisions for every situation and even improve the average reward over time. In this paper, we consider the contextual MAB problem. As opposed to the standard MAB problem, before making the choice of which arm to play, the decision maker observes a d-dimensional context x_i associated with each arm i, and chooses an arm to play in the current round based on the rewards of the arms played in the past along with the contexts. In this paper, we assume the expected reward of each arm is linear in the context, i.e., we assume there is an underlying mean parameter $\mu_i \in \mathbb{R}^d$ for each arm i, such that the expected reward for each arm i takes the form $x_i^\top \mu_i$. A class of predictors is said to be linear if each predictor predicts which arm gives the best expected reward that is linear in the observed context. The linear assumption leads to a succinct and tractable representation and is enough for good real-world performance (Li et al. 2010). The goal of the decision maker is to minimize the so-called regret, with

respect to the best linear predictor in hindsight, who predicts exactly μ_i and pulls the arm with the largest expected reward $x_i^\top \mu_i, i \in [K]$.

The MAB (or contextual MAB) problem essentially seeks a trade-off between exploitation (of the current information by playing the arm with the highest estimated reward) and exploration (by playing other arms to collect reward information). The majority of the literature balance this trade-off by designing algorithms that maximize the expected total reward (or equivalently, minimize the expected total regret). However, in many real-world problems, maximizing the expected reward is not always the most desirable. For example, in the portfolio selection problem, some portfolio managers are risk-averse and prefer less risky portfolios with low expected return rather than highly risky portfolios with high expected return. In this case, the risk of the reward should also be taken into consideration. Motivated by such risk consideration in real-world problems, we take a risk-averse perspective on the stochastic contextual MAB problem. Although many risk measures have been used in the risk-averse MAB problems, we focus on the mean-variance paradigm (Markowitz 1952), given its advantages in interpretability, computation, and popularity among practitioners. To the best of our knowledge, we are among the first to consider the risk-averse contextual MAB problem.

To solve risk-averse contextual MAB, we propose algorithms based on Thompson sampling (TS, see Thompson (1933)). TS is one of the earliest heuristics for the MAB problems via a Bayesian perspective. Intuitively speaking, TS assumes a prior distribution on the underlying parameters of the reward distribution for each arm and updates the posterior distributions after pulling the arms. At each round, it samples from the posterior distribution for each arm, and plays the arm that produces the best sampled reward. Most of the existing literature that apply TS to the MAB problem does not care about the variance of the reward distribution, as they intend to maintain a low expected regret. However, in the risk-averse setting, the variance also plays a vital role in determining the best arm. Hence, in addition to sampling the reward mean, we also need to sample the reward variance. It poses great challenges to the Bayesian updating of the parameters as well as the regret analysis, as one also has to bound the deviation of the sampled reward variance. We then provide the theoretical analysis of a variant of the proposed algorithms and show a $O((1+\rho+\frac{1}{\rho})d \ln T \ln \frac{K}{\delta} \sqrt{dKT^{1+2\epsilon} \ln \frac{K}{\delta} \frac{1}{\epsilon}})$ regret bound with high probability $1-\delta$, under some mild assumptions on the reward. The variant has the same posterior updating rule but carries out the sampling process differently, due to the technical difficulty in the regret analysis.

1.1 Literature Review

The contextual MAB problem is widely studied for online decision making. Two principled approaches that balance the trade-off between exploitation and exploration in the contextual MAB problem are upper confidence bound (UCB) and TS. The UCB algorithm addresses the problem from a frequentist perspective, which captures the uncertainty of the reward distribution by the width of the confidence bound. Most notably, Auer (2002) considers the contextual MAB problem with lin-

ear reward, and presents LinRel, a UCB-type algorithm that achieves an $\tilde{O}(\sqrt{Td})$ high probability regret bound. Li et al. (2010) present LinUCB, which utilizes ridge regression to estimate the mean parameters and is much easier to solve than LinRel. Chu et al. (2011) later show an $O\left(\sqrt{Td\ln^3(KT\ln(T)/\delta)}\right)$ regret bound that holds with probability $1-\delta$ for a variant of LinUCB. Abbasi et al. (2011) modify the UCB algorithm in Auer (2002) and improve the regret bound by a logarithmic factor.

On the other hand, TS is Bayesian approach to the same problem and is shown to be competitive to or better than the UCB algorithms (Chapelle and Li 2011). Most notably, Agrawal and Goyal (2013) provide the first theoretical guarantee for the TS algorithm on the contextual MAB problem, and shows an $\tilde{O}(d^{3/2}\sqrt{T})$ (or $\tilde{O}(d\sqrt{T\log(K)})$) high probability regret bound. Note that in the regret analysis of Agrawal and Goyal (2013), the reward distribution is only assumed to be sub-Gaussian, while the Bayesian updating of the posterior parameters assumes a Gaussian likelihood with an unknown mean but known variance. In this paper, we assume a Gaussian likelihood, but with (unknown mean and) unknown variance. It is well known that the conjugate prior (Schlaifer and Raiffa 1961) for the Gaussian likelihood is normal-gamma. Later in the paper we will show that the in case the reward distribution is not Gaussian, we can still choose to use the normal-gamma posterior updating rule.

The aforementioned literature on the contextual MAB problem all assume the best arm is chosen according to its expected performance. In some applications, however, the risk of the reward should also be taken into consideration. The risk-aware (or risk-averse) MAB problems have drawn increasing attention recently. Sani et al. (2012) consider the mean-variance risk criterion and present a UCB-type algorithm (MV-UCB). Later, Vakili and Zhao (2015) and Vakili and Zhao (2016) further complete the regret analysis of the MV-UCB algorithm. It should be noted that their definition of the regret is with respect to the mean-variance over a given horizon, which does not generalize to the contextual setting because the mean of each arm also depends on the context given at each round. Galichet et al. (2013) consider the conditional value at risk (CVaR, see Rockafellar and Uryasev (2000)) criterion and present the multi-armed risk-aware bandit (MaRaB), a UCB-type algorithm. However, while they choose to pull the arm with the largest empirical CVaR reward, they consider the expected regret (rather than CVaR) and analyze the regret under the assumption that the confidence level $\alpha = 0$ and that the CVaR and average criteria coincide. Maillard (2013) consider entropic risk measure and present RA-UCB algorithm that achieves logarithmic regret. Zimin et al. (2014) consider risk measures as a continuous function of reward mean and reward variance, and present a UCB-type algorithm that achieves logarithmic regret. We extend their regret definition to the contextual setting, while one should note that the optimal arm at each round may vary, depending on the given contexts. Cassel et al. (2018) provide a more systematic approach to analyzing general risk criteria within the stochastic MAB formulation, and design a UCB-type algorithm that reaches $O\left(\sqrt{T}\right)$ regret bound. Liu et al.

(2020) design a UCB-type algorithm for the mean-variance MAB problem with Gaussian rewards. Khajonchotpanya et al. (2021) develop a UCB-based algorithm that maximizes the expected empirical CVaR. On the other hand, even though TS is empirically shown in Baudry et al. (2021) to outperform UCB-type algorithms that often suffer from non-optimal confidence bounds, it has been considered only very recently for the risk-averse setting, due to its lack of theoretical understanding and difficulty of analyzing the regret bound. Chang et al. (2020) consider to minimize the expected cost with CVaR constraint on the cost (threshold constraint) using CVaR-TS algorithm, which achieves improvements over the compared UCB-based algorithms for Gaussian bandits with finite variances. Zhu and Tan (2020) design TS algorithms for Gaussian bandits and Bernoulli bandits, with the optimization objective being the mean-variance. Ang et al. (2021) present a Thompson sampling algorithm to minimize the entropic risk for Gaussian MABs, using similar tricks in Zhu and Tan (2020).

1.2 Contributions and Outlines

In this paper, we consider contextual MAB problems with linear rewards under the mean-variance risk criterion. Our contributions are summarized as follows:

- We develop a TS algorithm, namely mean-variance TS disjoint (MVTS-D) for the disjoint model.
- We present the theoretical analysis of a variant of the proposed algorithm under some mild assumptions. The regret analysis is a non-trivial extension of the existing results in Agrawal and Goyal (2013) to the risk-averse setting, as the considered mean-variance criterion greatly complicates the regret analysis.
- We carry out extensive sets of simulations on different reward distributions to demonstrate the performance of our proposed algorithms in a portfolio selection example.

The rest of the paper is organized as follows. Section 2 introduces the contextual MAB problem under the mean-variance paradigm. Section 3 presents the Thompson sampling algorithm for the disjoint model. We then conduct the regret analysis for a variant of the proposed algorithm in Section 4. Section 5 demonstrates the empirical performance of the proposed algorithms with a portfolio selection example. Section 6 concludes the paper.

2. Problem Setting

Consider a contextual MAB problem with K arms. At each round $t = 1, 2, \cdots, T$, a context $x_i(t) \in \mathbb{R}^d$ is revealed for each arm $i \in K$. The contexts can be chosen arbitrarily. After observing the contexts, the decision maker plays one of the K arms $a(t)$ and receives the reward $r_{a(t)}(t)$. We assume the reward for arm i at round t is generated from an unknown distribution $\nu_i(\mu_i, \sigma_i^2)$ with mean $x_i(t)^\top \mu_i$ and variance σ_i^2, where $\mu_i \in \mathbb{R}^d$ is a fixed but unknown mean parameter for the arm i, and σ_i^2 is a fixed but unknown variance parameter for the arm i. This model is called *disjoint* since the mean parameter and variance parameter are not shared among different arms. Define the history $\mathcal{H}_t = \{x_i(\tau), a(\tau), r_{a(\tau)}(\tau), i = 1, \cdots, K; \tau = 1, \cdots, t\}$, which summarizes all the information of contexts, pulled arms, and corresponding rewards up to round t. All re-

ward samples are independent conditioned on the choice of the arm and the context, and thus

$$\begin{aligned}&\mathbb{E}[r_{a(t)}(t) \mid \{x_i(t)\}_{i=1}^K, a(t), \mathcal{H}_{t-1}]\\ =&\mathbb{E}[r_{a(t)}(t) \mid a(t), x_{a(t)}(t)]\\ =&x_{a(t)}(t)^\top \mu_{a(t)}\end{aligned}$$

and similarly

$$\begin{aligned}&\mathrm{Var}[r_{a(t)}(t) \mid \{x_i(t)\}_{i=1}^K, a(t), \mathcal{H}_{t-1}]\\ =&\mathrm{Var}[r_{a(t)}(t) \mid a(t), x_{a(t)}(t)]\\ =&\sigma^2_{a(t)}\end{aligned}$$

Definition 1 *The mean-variance of arm i at round t with mean $x_i(t)^\top \mu_i$, variance σ_i^2, referred to as regret mean and regret variance, and risk tolerance ρ, is denoted by* $\mathrm{MV}_i(t) := x_i(t)^\top \mu_i - \rho\sigma_i^2$.

It should be noted that different from the traditional contextual MAB problem, the criterion of choosing an optimal arm is based on the mean-variance performance. The risk tolerance parameter reflects the risk attitude of the decision maker. When $\rho = 0$, the decision maker is risk-neutral, and our problem is reduced to the traditional contextual MAB with a risk-neutral criterion. When $\rho \to \infty$, the decision maker hates the risk so much that our problem turns to a variance minimization problem, which can also be easily handled by only sampling regret variance and choosing the arm with the smallest sampled regret variance to pull. Hence, we focus on the more interesting case $0 < \rho < \infty$. As a final note, our algorithm and analysis can also be easily extended to the case $\rho < 0$, which means the decision maker is risk-seeking.

A policy, or allocation strategy π, is an algorithm that chooses at each round t, an arm $a(t)$ to pull, based on the history $\mathcal{H}_{t-1}$ and the context $x_i(t)$ for $i \in [K]$. Let $a^*(t)$ denote the optimal arm to pull at round t under the mean-variance criterion, i.e., $a^*(t) := \arg\max_{i\in[K]} x_i(t)^\top \mu_i - \rho\sigma_i^2$. Let $\Delta_i(t)$ denote the difference between the mean-variances of the optimal arm $a^*(t)$ and arm i, i.e.,

$$\Delta_i(t) = \mathrm{MV}_{a^*(t)}(t) - \mathrm{MV}_i(t)$$

The regret at round t is defined as $R(t) = \Delta_{a(t)}(t)$, where $a(t)$ is the arm to pull at round t, determined by the algorithm π. The goal is to minimize the total regret $\mathcal{R}(T) = \sum_{t=1}^T R(t)$, or in other words, design an algorithm whose regret increases as slowly as possible as T increases. Note that the optimal action may not be a single-arm action, since the optimal arm at each round t is determined by the given contexts.

3. Thompson Sampling

In this section, we consider TS for the risk-averse contextual MAB problem under the disjoint model. In case the true reward distribution is Gaussian, we use Gaussian likelihood (for the rewards) and normal-gamma conjugate prior (for the mean and variance parameters) to design our Thompson sampling algorithm. As for a more general reward distribution that may not be Gaussian, usually we can take two approaches. On the one hand, we can choose the same likelihood as the true reward distribution and update the posterior distribution accordingly. The resulting posterior, however, is usually intractable. One can choose to use Markov Chain Monte Carlo (MCMC) to sample from the posterior, and use those sampled parameters to choose the optimal arm. The latest work along this line can be found in Xu et al. (2022). On the other hand, we can still

choose the Gaussian likelihood, even though it is different from the true reward distribution, which is often refereed to as model mis-specification. Then we apply the variational Bayes (VB) technique to obtain a tractable approximate posterior. We show the details of VB in the subsection below.

3.1 Variational Bayes under Model Mis-specification

For ease of exposition, we first introduce some notations and a simplified problem setting. Denote the mean and variance parameters of the true reward distribution by $\theta = (\mu, \sigma^2)$. The reward $y_1, \cdots, y_n$ are n data points independently and identically distributed (i.i.d.) with a true density $p_0(\cdot)$. The (Gaussian) likelihood is denoted by $p(y \mid \theta)$. The mean field variational Bayes (MFVB) approximates the exact posterior distribution (updated using the likelihood) by a probability distribution with density $q(\theta)$ belonging to some tractable family of distributions $\mathcal{Q}$ that are factorizable, i.e., $\mathcal{Q} = \left\{q(\theta) : q(\theta) = q_1\left(\mu\right) q_2\left(\sigma^2\right)\right\}$. The (optimal) VB posterior $q^*(\theta)$ is then found by minimizing the Kullback-Leibler (KL, see Kullback and Leibler (1951)) divergence from the exact posterior distribution $p(\theta \mid y)$, i.e.,

$$q^*(\theta) = \arg\min_{q(\theta)\in\mathcal{Q}} \left\{ \mathrm{KL}(q \| p(\theta \mid y)) := \int q(\theta) \log \frac{q(\theta)}{p(\theta \mid y)} \mathrm{d}\theta \right\}$$

It is shown in Tran et al. (2021) that the VB posterior takes the form of normal-gamma by choosing a normal-gamma prior and the Gaussian likelihood. It should be noted that the VB posterior is an approximation to the exact posterior. According to the Bernstein-Von Mises theorem under the model mis-specification, the exact posterior converges in distribution to a point mass at θ^* (Kleijn et al. 2021), where θ^* is the value of θ that minimizes the KL divergence between the assumed likelihood and the true reward distribution, i.e.,

$$\theta^* = \arg\min_{\theta} \mathrm{KL}\left(p_0(y) \| p(y \mid \theta)\right)$$

and Wang and Blei (2019) later show that the VB posterior also converges in distribution to a point mass at θ^*, and the VB posterior mean converges almost surely to θ^*. For example, when the true reward is generated according to a uniform distribution with lower bound a and upper bound b, the VB posterior mean $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ converges almost surely to $\theta^* = (\frac{a+b}{2}, \frac{(b-a)^2}{12})$, which is the mean and variance of the uniform distribution. Since we consider the mean-variance objective, we only care about the accuracy of the mean and variance estimates of the reward distribution. Hence, the normal-gamma posterior updating is fair and reasonable, even for the non-Gaussian reward distribution. Hence, in the rest of the paper, we consider the Gaussian likelihood with a normal-gamma prior on the mean and variance parameters.

3.2 TS for the Disjoint Model

Suppose the likelihood of reward $r_i(t)$ for arm i at round t, given the context $x_i(t)$, the mean parameter μ_i and the precision parameter λ_i (reciprocal of the variance parameter σ_i^2), were given by the probability density function (p.d.f.) of the Gaussian distribution $\mathcal{N}(x_i(t)^\top \mu_i, \lambda_i^{-1})$. Let $T_i(t)$ be the set of the rounds that arm i has been pulled during the first $t-1$ rounds, whose cardinality is denoted by $\#T_i(t)$. We show the Bayesian updating of the parameters in the next proposition.

Proposition 1 *Suppose the prior for λ_i at round t is given by Gamma$(C_i(t), D_i(t))$, and conditioned*

Algorithm 1: Posterior Updating for the Disjoint Model at Round t

input : prior parameters $(A_i(t), b_i(t), C_i(t), D_i(t))$, context $x_i(t)$, set of rounds that arm i has been pulled during the first $t-1$ rounds $T_i(t)$; arm to play $a(t)$; and reward sample $r_{a(t)}(t)$.

output: posterior parameters $(A_i(t+1), b_i(t+1), C_i(t+1), D_i(t+1))$ for each arm i.

for $i = 1, 2, \cdots, K, i \neq a(t)$ **do**

  $A_i(t+1) = A_i(t), b_i(t+1) = b_i(t), C_i(t+1) = C_i(t), D_i(t+1) = D_i(t)$;

end

$A_{a(t)}(t+1) = A_{a(t)}(t) + x_{a(t)}(t)x_{a(t)}(t)^\top$;

$b_{a(t)}(t+1) = b_{a(t)}(t) + x_{a(t)}r_{a(t)}(t)$;

$C_{a(t)}(t+1) = C_{a(t)}(t) + \frac{1}{2}$;

$D_{a(t)}(t+1) = D_{a(t)}(t) + \frac{1}{2}[b_{a(t)}(t)^\top A_{a(t)}(t)^{-1} b_{a(t)}(t) - b_{a(t)}(t+1)^\top A_{a(t)}(t+1)^{-1} b_{a(t)}(t+1) + r_{a(t)}(t)^2]$.

on λ_i, the prior for μ_i at round t is given by $\mathcal{N}(A_i(t)^{-1}b_i(t), (\lambda_i A_i(t))^{-1})$. Here $C_i(t)$ is the shape parameter and $D_i(t)$ is the rate parameter of the Gamma distribution. Let

$$A_i(t) = \mathbf{I}_d + \sum_{s \in T_i(t)} x_i(s)x_i(s)^\top$$
$$b_i(t) = \mathbf{0}_{d\times 1} + \sum_{s \in T_i(t)} x_i(s)r_i(s)$$

where $\mathbf{I}_d$ is a d-dimensional identity matrix, $\mathbf{0}_{d\times 1}$ is a d-dimensional zero vector. Then the posterior for λ_i is given by Gamma$(C_i(t+1), D_i(t+1))$, and conditioned on λ_i, the posterior for μ_i is given by $\mathcal{N}(A_i(t+1)^{-1}b_i(t+1), (\lambda_i A_i(t+1))^{-1})$, where

$$C_i(t+1) = C_i(t) + \frac{1}{2}$$
$$D_i(t+1) = D_i(t) + \frac{1}{2}[-b_i(t+1)^\top A_i(t+1)^{-1} b_i(t+1) + b_i(t)^\top A_i(t)^{-1} b_i(t) + r_i(t)^2]$$

Please refer to Appendix C for detailed proof of Proposition 1. We can also obtain the desired posterior distribution by applying variational Bayes to Lasso regression model (see Algorithm 2 in Tran et al. (2021)). Algorithm 1 summarizes the posterior updating for the disjoint model. We now present the Thompson sampling algorithm for the disjoint model in Algorithm 2. At each round t, we generate a sample $\widetilde{\lambda}_i(t)$ from the distribution Gamma$(C_i(t), D_i(t))$, set $\widetilde{\sigma}_i^2(t) = \frac{1}{\widetilde{\lambda}_i(t)}$, and generate a sample $\widetilde{\mu}_i(t)$ from the distribution $\mathcal{N}(A_i(t)^{-1}b_i(t), (\widetilde{\lambda}_i A_i(t))^{-1})$ for each arm i. Then we play the arm i that maximizes

$$\widetilde{\mathrm{MV}}_i(t) = x_i(t)^\top \widetilde{\mu}_i(t) - \rho\widetilde{\sigma}_i^2(t)$$

4. Regret Analysis

In this section, we present our regret bounds and its derivation for a variant of the proposed MVTS-D algorithm. We first make the following assumptions.

Assumption 1 *(i) $\eta_i(t) := r_i(t) - x_i(t)^\top \mu_i$ is R-sub-Gaussian, i.e.,*

$$\mathbb{E}[\lambda \exp(\eta_i(t))] \leq \exp(\frac{\lambda^2 R^2}{2}), \forall \lambda \in \mathbb{R}$$

(ii) $\eta_i(t)^2 - \sigma_i^2$ is R-sub-Gaussian, i.e., for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[\lambda \exp(\eta_i(t)^2 - \sigma_i^2)] \leq \exp(\frac{\lambda^2 R^2}{2})$$

(iii) $\|x_i(t)\| \leq 1, \|\mu\| \leq 1, \mid x_i(t)^\top \mu_i - x_j(t)^\top \mu_j \mid \leq 1, \mid \sigma_i^2 - \sigma_j^2 \mid \leq 1$, for $i, j \in [K], i \neq j$, for all t.

The first and second assumption in Assumption 1 are satisfied when the reward distribution is bounded. In case the positive constants R in (i) and (ii) are different, we take R to be the maximum of the two. The

Algorithm 2: Mean-Variance Thompson Sampling for the Disjoint Model (MVTS-D)

initialization:
pull each arm i once at round 0 and observe rewards $r_i(0)$;
$A_i(1) = \mathbf{I}_d + x_i(0)x_i(0)^\top$, $b_i(1) = x_i(0)r_i(0)$, $C_i(1) = \frac{1}{2}$, $D_i(1) = \frac{1}{2}(r_i(0)^2 - x_i(0)^\top A_i(1)^{-1}x_i(0))$, $T_i(1) = \{0\}$;
for $t = 1, 2, \cdots, T$ **do**
  observe K contexts $x_1(t), \cdots, x_K(t) \in \mathbb{R}^d$;
  for $i = 1, 2, \cdots, K$ **do**
    sample $\widetilde{\lambda}_i(t)$ from distribution Gamma$(C_i(t), D_i(t))$, set $\widetilde{\sigma}_i^2(t) = \frac{1}{\widetilde{\lambda}_i(t)}$;
    sample $\widetilde{\mu}_i(t)$ from distribution $\mathcal{N}\left(A_i(t)^{-1}b_i(t), (\widetilde{\lambda}_i(t)A_i(t))^{-1}\right)$;
    set $\widetilde{\mathrm{MV}}_i(t) = x_i(t)^\top \widetilde{\mu}_i(t) - \rho\widetilde{\sigma}_i^2(t)$;
  end
  play arm $a(t) = \arg\max_{i\in[K]} \widetilde{\mathrm{MV}}_i(t)$ with ties broken arbitrarily;
  observe reward $r_{a(t)}(t) \sim \nu_{a(t)}\left(x_{a(t)}(t)^\top \mu_{a(t)}, \sigma^2_{a(t)}\right)$;
  update $(A_i(t), b_i(t), C_i(t), D_i(t))$ according to Algorithm 1 for each arm i;
  set $T_{a(t)}(t+1) = T_{a(t)}(t) \bigcup \{t\}$.
end

third assumption is required to make the regret bound scale-free, and is standard in the literature (Agrawal and Goyal 2013). The norms $\|\cdot\|$, unless stated otherwise, are l_2-norms. In case $\|x_i(t)\| \le c_1, \|\mu\| \le c_2, \mid x_i(t)^\top \mu_i - x_j(t)^\top \mu_j \mid \le c_3, \mid \sigma_i^2 - \sigma_j^2 \mid \le c_4$ for some constants $c_1, c_2, c_3, c_4 > 0$, our regret bound would increase by a factor of $c = \max\{c_1, c_2, c_3, c_4\}$.

Due to the technical difficulty, instead of sampling the regret variance from the posterior Gamma distribution, we propose to sample the regret variance from a Gaussian distribution with a decaying variance term, where the mean of the Gaussian distribution corresponds to the mean of the Gamma distribution. Gaussian sampling enables us to derive the desired concentration and anti-concentration bounds, which are crucial in the regret analysis. Also, similar to Zhu and Tan (2020), we sample the regret mean and regret variance from different distributions independently. We summarize this variant of the MVTS algorithm in Algorithm 3 (see Appendix C for full algorithm) and name it as MVTS-DN, since it samples the regret variance from a normal distribution. Compared to Algorithm 2, Algorithm 3

- replaces Line 7 by: sample $\widetilde{\sigma}_i^2(t)$ from distribution $\mathcal{N}\left(\frac{D_i(t)}{C_i(t)}, \frac{u^2}{\#T_i(t)}\right)$;
- replaces Line 8 by: sample $\widetilde{\mu}_i(t)$ from distribution $\mathcal{N}\left(A_i(t)^{-1}b_i(t), v^2A_i(t)^{-1}\right)$.

In Algorithm 3, $v = R\sqrt{\frac{4}{\epsilon}d\ln\frac{4K}{\delta}}$, $u = 8R^2d\ln\frac{4K}{\delta}\sqrt{\frac{1}{\epsilon}}$, where $0 < \delta < 1$ is the parameter for confidence level $(1-\delta)$, and $0 < \epsilon < \frac{1}{2}$ is the parameter that controls the prior variance in the sampling process. A smaller ϵ leads to a larger prior variance, which encourages more exploration. We first show the main result of the theoretical analysis and discuss the proof of the result later.

Theorem 1 *Suppose Assumption 1 holds. For the contextual MAB problem with T rounds, K arms, d-dimensional contexts and linear reward under the mean-variance criterion, the MVTS-DN algorithm achieves a total regret of $O((1+\rho+\frac{1}{\rho})d\ln T\ln\frac{K}{\delta}\sqrt{dKT^{1+2\epsilon}\ln\frac{K}{\delta}\frac{1}{\epsilon}})$ that holds with probability $1-\delta$, for any $0 < \epsilon < \frac{1}{2}$, $0 < \delta < 1$.*

Remark 1 *Treating all parameters as constants except the number of rounds T and ϵ, we achieve a regret bound of $O(\sqrt{T^{1+2\epsilon}} \ln T)$, which essentially is the same as that in Agrawal and Goyal (2013). However, when considering the number of arms K, confidence level δ and dimension d, compared with Agrawal and Goyal (2013), here the regret bound has additional terms $\ln \frac{K}{\delta}$ and d. This is caused by controlling the estimation error of regret variance, which is more difficult than that of regret mean. This can be seen in Lemma 2 and 3, in which we obtain the constants $l(T)$ and $h(T)$ with different orders of $\ln \frac{K}{\delta}$ and d.*

To prove Theorem 1, we follow a similar approach as in Agrawal and Goyal (2013). Compared with the risk-neutral case in Agrawal and Goyal (2013), the main difficulty of regret analysis for this risk-averse case arises in the estimation error control of the variance, which appears in our mean-variance objective for arm selection. This difficulty is overcome by sampling the regret variance $\widetilde{\sigma}_i^2(t)$ from a normal distribution instead of the Gamma distribution, as we have argued in the beginning of this section. For ease of exposition, we introduce and summarize some notations below that are relevant to the proofs.

Let the mean parameter estimate be $\hat{\mu}_i(t) = A_i(t)^{-1} b_i(t)$, which is the weighted average of the historical rewards for arm i up to rounds $t-1$. Let the standard deviation of the estimate $x_i(t)^\top \hat{\mu}_i(t)$ be $s_i(t) = \sqrt{x_i(t)^\top A_i(t)^{-1} x_i(t)}$. In the variance sampling, $C_i(t) = \frac{1}{2}\#T_i(t)$, $D_i(t) = \frac{1}{2}[\sum_{s\in T_i(t)} r_i(s)^2 - b_i(t)^\top A_i(t)^{-1} b_i(t)]$. Let the variance parameter estimate be $\hat{\sigma}_i^2(t) = \frac{D_i(t)}{C_i(t)} = \frac{1}{\#T_i(t)}(\sum_{s\in T_i(t)} r_i(s)^2 - b_i(t)^\top A_i(t)^{-1} b_i(t))$.

Definition 2 *Define the following constants in terms of T:*

$$\ell(T) = R\sqrt{d \ln T \ln \frac{4K}{\delta}} + 1$$
$$h(T) = 4R^2 d \ln \frac{4K}{\delta} \sqrt{\ln T}$$
$$g(T) = \sqrt{4d \ln T \sqrt{Kd}} \cdot v + \ell(T)$$
$$q(T) = u\sqrt{2 \ln T} + h(T)$$

These constants are used throughout the proof.

Definition 3 *The saturated set $S(t)$,*

$$S(t) := \Big\{ i \in [K] : g(T) s_i(t) + \rho q(T) \frac{1}{\sqrt{\#T_i(t)}} \le \ell(T) s_{a^*(t)}(t) + \rho h(T) \frac{1}{\sqrt{\#T_i(t)}} \Big\}$$

An arm i is called saturated at round t if $i \in S(t)$, and unsaturated if $i \notin S(t)$.

For an saturated arm i, the standard deviation $s_i(t)$ is small and the number of pulls $\#T_i(t)$ is large. Hence, the estimates of the mean and variance parameters constructed using the previous rewards are quite accurate. The algorithm can easily tell whether it is optimal arm or not. At last, let the filtration $\mathcal{F}_{t-1}$ be the σ-algebra generated by $\mathcal{H}_{t-1} \bigcup \{x_i(t)\}_{i\in[K]}$.

4.1 Proof Outline

We present the proof outline here. We first derive confidence bands for mean and variance parameter estimates $\hat{\mu}_i(t)$, $\hat{\sigma}_i^2(t)$, for all i. Then we derive confidence bands for sampled regret mean and sampled regret variance $\widetilde{\mu}_i(t)$ and $\widetilde{\sigma}_i^2(t)$, for all i. Using these bands and the triangle inequality, we have $MV_{a^*(t)}(t) - MV_{a(t)}(t) \le \widetilde{MV}_{a^*(t)}(t) - \widetilde{MV}_{a(t)}(t) + g(T)(s_{a^*(t)}(t) + s_{a(t)}(t)) + \rho q(T) \frac{1}{\sqrt{\#T_{a^*(t)}(t)}}$. Since $a(t)$ is the arm with largest $\widetilde{MV}$, the regret at any time t can be bounded by $g(T)(s_{a^*(t)}(t) + s_{a(t)}(t)) + \rho q(T)(\frac{1}{\sqrt{\#T_{a^*(t)}(t)}} + \frac{1}{\sqrt{\#T_{a(t)}(t)}})$, where the four terms represent the confidence bands

for arm $a^*(t)$ and $a(t)$. Then, we can bound the total regret if we can bound $\sum_{t=1}^{T} s_{a(t)}(t)$, $\sum_{t=1}^{T} \frac{1}{\sqrt{\#T_{a(t)}(t)}}$, $\sum_{t=1}^{T} s_{a^*(t)}(t)$ and $\sum_{t=1}^{T} \frac{1}{\sqrt{\#T_{a^*(t)}(t)}}$, respectively. For the first two terms, we have $\sum_{t=1}^{T} s_{a(t)}(t) = O(\sqrt{Td \ln T})$ and $\sum_{t=1}^{T} \frac{1}{\sqrt{\#T_{a(t)}(t)}} = O(\sqrt{T \ln T})$. The challenge is left to bound $\sum_{t=1}^{T} s_{a^*(t)}(t)$ and $\sum_{t=1}^{T} \frac{1}{\sqrt{\#T_{a^*(t)}(t)}}$.

For this purpose, we define the saturated and unsaturated arms at any time as in Definition 3. Then, if an arm $a(t) \notin S(t)$ is played at time t, we can bound $s_{a^*(t)}(t)$ and $\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}$ by $s_{a(t)}(t)$ and $\frac{1}{\sqrt{\#T_{a(t)}(t)}}$ multiplied by some factors according to the definition of unsaturated arms. For saturated arms in $S(t)$, we bound the probability of playing such arms at any time t by the probability of playing the optimal arm at time t, $a^*(t)$ multiplied by some factor, given the filtration $\mathcal{F}_{t-1}$. This is helpful since again we can shift those terms indexed with $a^*(t)$ to terms indexed with $a(t)$.

With all the observations, we establish a super-martingale difference, Y_t, with respect to the regret as shown in Lemma 9. Applying the Azuma-Hoeffding inequality for martingales and along with $\sum_{t=1}^{T} s_{a(t)}(t) = O(\sqrt{Td \ln T})$ and $\sum_{t=1}^{T} \frac{1}{\sqrt{\#T_{a(t)}(t)}} = O(\sqrt{T \ln T})$, we obtain the high probability regret bound in Theorem 1.

4.2 Formal Proof of Theorem 1

Lemma 1 (Abbasi et al. (2011), Theorem 1) *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$. Let $\{m_t\}_{t=1}^{\infty}$ be $\mathbb{R}^d$-valued stochastic process such that m_t is $\mathcal{F}_{t-1}$-measurable. For any $t \geq 0$, define*

$$\bar{M}_t = I_d + \sum_{s=1}^{t} m_s m_s^\top, \quad \xi_t = \sum_{s=1}^{t} \eta_s m_s$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

$$\|\xi_t\|^2_{\bar{M}_t^{-1}} \leq 2R^2 \log\left(\frac{\det(\bar{M}_t)^{\frac{1}{2}}}{\delta}\right)$$

where $\|\xi_t\|_{\bar{M}_t^{-1}} = \sqrt{\xi_t^T \bar{M}_t^{-1} \xi_t}$.

The first two lemmas, Lemma 2 and Lemma 3, upper bound the probability of estimation error of mean and variance around their true value.

Lemma 2 (Agrawal and Goyal (2013), Lemma 1) *Define $E^{\mu}(t)$ as the event that $x_i(t)^\top \hat{\mu}_i(t)$ is concentrated around its mean for any arm i, i.e.,*

$$E^{\mu}(t) := \{ \mid x_i(t)^\top \hat{\mu}_i(t) - x_i(t)^\top \mu_i(t) \mid \leq \ell(T) s_i(t), \forall i \in [K]\}$$

Then with probability at least $1-\frac{\delta}{4}$, $E^{\mu}(t)$ holds true for all t and $0 < \delta < 1$.

Lemma 3 *Define $E^{\sigma}(t)$ as the event that $\hat{\sigma}_i^2(t)$ is concentrated around the true variance σ_i^2 for any arm i, i.e.,*

$$E^{\sigma}(t) := \{\mid \hat{\sigma}_i^2(t) - \sigma_i^2 \mid \leq h(T)\frac{1}{\sqrt{\#T_i(t)}}, \forall i \in [K]\}$$

Then with probability at least $1-\frac{\delta}{4}$, $E^{\sigma}(t)$ holds true for all t and $0 < \delta < 1$.

Proof. Recall that $b_i(t) = \sum_{s \in T_i(t)} x_i(s) r_i(s)$. We have

$$\begin{aligned}
&\hat{\sigma}_i^2(t) - \sigma_i^2 \\
&= \frac{1}{\#T_i(t)}\Big[\sum_{s \in T_i(t)} r_i(s)^2 - b_i(t)^\top A_i(t)^{-1} b_i(t)\Big] - \sigma_i^2 \\
&= \frac{1}{\#T_i(t)}\Big[\sum_{s \in T_i(t)} r_i(s)^2 - \sum_{s \in T_i(t)} x_i(s)^\top r_i(s) A_i(t)^{-1} \\
&\quad \sum_{s \in T_i(t)} x_i(s) r_i(s)\Big] - \sigma_i^2
\end{aligned}$$

$$
\begin{aligned}
&=\frac{1}{\#T_i(t)}\Big[\sum_{s\in T_i(t)}(r_i(s)-x_i(s)^\top\mu_i)^2-\sigma_i^2\Big]\\
&+\frac{1}{\#T_i(t)}\Big[2\sum_{s\in T_i(t)}r_i(s)x_i(s)^\top\mu_i\\
&-\sum_{s\in T_i(t)}(x_i(s)^\top\mu_i)^2\Big]\\
&-\frac{1}{\#T_i(t)}\Big[\sum_{s\in T_i(t)}x_i(s)(r_i(s)-x_i(s)^\top\mu_i)\Big]^\top\\
&A_i(t)^{-1}\Big[\sum_{s\in T_i(t)}x_i(s)(r_i(s)-x_i(s)^\top\mu_i)\Big]\\
&-\frac{2}{\#T_i(t)}\Big[\mu_i^\top\sum_{s\in T_i(t)}x_i(s)x_i(s)^\top A_i(t)^{-1}\\
&\sum_{s\in T_i(t)}x_i(s)r_i(s)\Big] \qquad (1)\\
&+\frac{1}{\#T_i(t)}\Big[\mu_i^\top\sum_{s\in T_i(t)}x_i(s)x_i(s)^\top A_i(t)^{-1}\\
&\sum_{s\in T_i(t)}x_i(s)x_i(s)^\top\mu_i\Big] \qquad (2)
\end{aligned}
$$

Recall that $A_i(t)=\mathbf{I}_d+\sum_{s\in T_i(t)}x_i(s)x_i(s)^\top$. Then

$$
\begin{aligned}
(1)&=\frac{2}{\#T_i(t)}\Big[\mu_i^\top\sum_{s\in T_i(t)}x_i(s)r_i(s)-\mu_i^\top A_i(t)^{-1}\\
&\sum_{s\in T_i(t)}x_i(s)r_i(s)\Big]
\end{aligned}
$$

$$
\begin{aligned}
(2)&=\frac{1}{\#T_i(t)}\Big[\mu_i^\top\sum_{s\in T_i(t)}(x_i(s)^\top\mu_i)^2-\mu_i^\top A_i(t)^{-1}\\
&\sum_{s\in T_i(t)}x_i(s)x_i(s)^\top\mu_i\Big]
\end{aligned}
$$

Then we obtain

$$
\begin{aligned}
&\hat{\sigma}_i^2(t)-\sigma_i^2\\
&=\frac{1}{\#T_i(t)}\Big[\sum_{s\in T_i(t)}(r_i(s)-x_i(s)^\top\mu_i)^2-\sigma_i^2\Big] \qquad (3)\\
&-\frac{1}{\#T_i(t)}\Big[\sum_{s\in T_i(t)}x_i(s)(r_i(s)-x_i(s)^\top\mu_i)\Big]^\top\\
&A_i(t)^{-1}\Big[\sum_{s\in T_i(t)}x_i(s)(r_i(s)-x_i(s)^\top\mu_i)\Big] \qquad (4)\\
&+\frac{2}{\#T_i(t)}\Big[\mu_i^\top A_i(t)^{-1}\\
&\sum_{s\in T_i(t)}x_i(s)(r_i(s)-x_i(s)^\top\mu_i)\Big] \qquad (5)\\
&+\frac{1}{\#T_i(t)}\mu_i^\top\mu_i-\frac{1}{\#T_i(t)}\mu_i^\top A_i(t)^{-1}\mu_i
\end{aligned}
$$

To bound (3), we apply Lemma 1. Let the filtration $\mathcal{F}'_{t-1}$ be the σ-algebra generated by $\mathcal{H}_{t-1}\bigcup\{x_i(t)\}_{i\in[K]}\bigcup a(t)$. Let

$$
\eta_i(t)=\begin{cases}(r_i(t)-x_i(t)^\top\mu_i)^2-\sigma_i^2, & \text{if} \quad a(t)=i\\ 0, & \text{if} \quad a(t)\neq i\end{cases}
$$

Then $\eta_i(t)$ is $\mathcal{F}'_t$-measurable. Let

$$
m_i(t)=\begin{cases}1, & \text{if} \quad a(t)=i\\ 0, & \text{if} \quad a(t)\neq i\end{cases}
$$

Then $m_i(t)$ is $\mathcal{F}'_{t-1}$-measurable. Let

$$
\begin{aligned}
\zeta_i(t)&=\sum_{s=1}^{\top}m_i(t)\eta_i(t)\\
&=\sum_{s\in T_i(t)}[(r_i(s)-x_i(s)^\top\mu_i)^2-\sigma_i^2]
\end{aligned}
$$

Lemma 1 implies that with probability at least $1-\delta'$,

$$
\frac{1}{\sqrt{\#T_i(t)}}\mid\zeta_i(t)\mid\leq R\sqrt{\ln\frac{\#T_i(t)}{\delta'^2}}
$$

therefore we have

$$
\mid(3)\mid\leq\frac{1}{\sqrt{\#T_i(t)}}R\sqrt{\ln\frac{\#T_i(t)}{\delta'^2}}
$$

(4) is bounded similarly using Lemma 1. Let

$$
\eta_i(t)=\begin{cases}r_i(t)-x_i(t)^\top\mu_i, & \text{if} \quad a(t)=i\\ 0, & \text{if} \quad a(t)\neq i\end{cases}
$$

$$
m_i(t)=\begin{cases}x_i(t), & \text{if} \quad a(t)=i\\ 0, & \text{if} \quad a(t)\neq i\end{cases}
$$

$$
\begin{aligned}
\xi_i(t)&=\sum_{s=1}^{t}m_i(t)\eta_i(t)\\
&=\sum_{s\in T_i(t)}x_i(s)(r_i(s)-x_i(t)^\top\mu_i)
\end{aligned}
$$

Note that $\det(A_i(t))\leq(T_i(t))^d$. For $d\geq 2$, Lemma 1 implies that with probability at least $1-\delta'$,

$$
\begin{aligned}
\mid(4)\mid&=\frac{1}{\#T_i(t)}\|\xi_i(t)\|^2_{A_i(t)^{-1}}\\
&\leq\frac{1}{\#T_i(t)}R^2d\ln\left(\frac{\#T_i(t)}{\delta'}\right)
\end{aligned}
$$

Similarly, we have

$$\begin{aligned} \mid (5) \mid &\leq \frac{2}{\#T_i(t)} \left\|\mu_i(t)\right\|^2_{A_i(t)^{-1}} \|\xi_i(t)\|^2_{A_i(t)^{-1}} \\ &\leq \frac{2}{\#T_i(t)} R\sqrt{d \ln \frac{\#T_i(t)}{\delta'}} \end{aligned}$$

For the last two terms, we have

$$\begin{aligned} \mid \frac{1}{\#T_i(t)} \mu_i^\top \mu_i \mid &\leq \frac{1}{\#T_i(t)} \\ \mid \frac{1}{\#T_i(t)} \mu_i^\top A_i(t)^{-1} \mu_i \mid &\leq \frac{1}{\#T_i(t)} \end{aligned}$$

Assume $R \geq 1$. Then with probability at least $1 - 2\delta'$, we have

$$\begin{aligned} &\mid \hat{\sigma}_i^2(t) - \sigma_i^2 \mid \\ \leq &\frac{1}{\sqrt{\#T_i(t)}} R\sqrt{\ln \frac{\#T_i(t)}{\delta'^2}} + \frac{1}{\#T_i(t)} R^2 d \ln \frac{\#T_i(t)}{\delta'} \\ &+ \frac{2}{\#T_i(t)} R\sqrt{d \ln \frac{\#T_i(t)}{\delta'}} + \frac{2}{\#T_i(t)} \\ \leq &\frac{1}{\sqrt{\#T_i(t)}} 4R^2 d \ln \frac{1}{\delta'} \sqrt{\ln \#T_i(t)} \end{aligned}$$

Taking $\delta' = \frac{\delta}{4K}$, we have

$$\begin{aligned} &\mid \hat{\sigma}_i^2(t) - \sigma_i^2 \mid \\ \leq &\frac{1}{\sqrt{\#T_i(t)}} 4R^2 d \ln \frac{4K}{\delta} \sqrt{\ln \#T_i(t)} \\ \leq &\frac{1}{\sqrt{\#T_i(t)}} 4R^2 d \ln \frac{4K}{\delta} \sqrt{\ln \mid T \mid} \\ := &h(T) \frac{1}{\sqrt{\#T_i(t)}} \end{aligned}$$

Then with probability at least $1 - \frac{\delta}{4K}$, $\mid \hat{\sigma}_i^2(t) - \sigma_i^2 \mid \leq h(T) \frac{1}{\sqrt{\#T_i(t)}}$ holds $\forall t \geq 1$. Using a union bound we obtain with probability at least $1 - \frac{\delta}{4}$, $E^\sigma(t)$ holds $\forall t \geq 1$. ∎

Lemma 4 and Lemma 5 provide concentration bounds for the posterior samples of regret mean and regret variance around their estimates, respectively.

Lemma 4 (Agrawal and Goyal (2013), Lemma 1) *Define $E^{\tilde{\mu}}(t)$ as the event that $x_i(t)^\top \tilde{\mu}_i(t)$ is concentrated around $x_i(t)^\top \hat{\mu}_i(t)$ for any arm i, i.e.,*

$$\begin{aligned} E^{\tilde{\mu}}(t) := \Big\{ &\mid x_i(t)^\top \tilde{\mu}_i(t) - x_i(t)^\top \hat{\mu}_i(t) \mid \\ &\leq \sqrt{4d \ln T\sqrt{Kd}} \cdot v \cdot s_i(t), \forall i \in [K] \Big\} \end{aligned}$$

Then $\mathbb{P}(E^{\tilde{\mu}}(t) \mid \mathcal{F}_{t-1}) \geq 1 - \frac{1}{T^2}$.

Lemma 5 *Define $E^{\tilde{\sigma}}(t)$ as the event that $\tilde{\sigma}_i^2(t)$ is concentrated around $\hat{\sigma}_i^2(t)$ for any arm i, i.e.,*

$$\begin{aligned} E^{\tilde{\sigma}}(t) := \{ &\mid \tilde{\sigma}_i^2(t) - \hat{\sigma}_i^2(t) \mid \\ &\leq 2\sqrt{\ln T\sqrt{K}} u \frac{1}{\sqrt{\#T_i(t)}}, \forall i \in [K] \} \end{aligned}$$

Then $\mathbb{P}(E^{\tilde{\sigma}}(t) \mid \mathcal{F}_{t-1}) \geq 1 - \frac{1}{T^2}$.

Proof. A direct application of Lemma 5 in Agrawal and Goyal (2013) gives:

$$\begin{aligned} &\mathbb{P}\Big(\mid \frac{\sqrt{\#T_i(t)}}{u} (\tilde{\sigma}_i^2(t) - \hat{\sigma}_i^2(t)) \mid \geq 2\sqrt{\ln T\sqrt{K}} \mid \mathcal{F}_{t-1} \Big) \\ \leq &\frac{1}{\sqrt{\pi}} \cdot \frac{1}{2\sqrt{\ln T\sqrt{K}}} \exp(-2 \ln T\sqrt{K}) \leq \frac{1}{KT^2} \end{aligned}$$

Hence $E^{\tilde{\sigma}}(t)$ holds with probability at least $1 - K \cdot \frac{1}{KT^2} = 1 - \frac{1}{T^2}$. ∎

With Lemma 2-5, we can derive the concentration bounds for $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ around the true value μ_i and σ_i^2, which is useful to bound the regret in terms of $s_{a(t)}(t)$, $s_{a^*(t)}(t)$, $\frac{1}{\sqrt{\#T_{a(t)}(t)}}$ and $\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}$.

The next step is to bound the probability of pulling an arm in the set $S(t)$ as shown in Lemma 8 by the probability of pulling an optimal arm. To prove lemma 8, we first present Lemma 6 and Lemma 7, which lower bound the probability of sampling $\tilde{\mu}_i$ such that $x_i(t)^\intercal \tilde{\mu}_i(t)$ exceeding $x_i^\intercal(t)\mu_i$ by $\ell(t)s_i(t)$ and sampling $\tilde{\sigma}_i^2$ less than $\sigma_i^2 - h(T)\frac{1}{\sqrt{\#T_i(t)}}$.

Lemma 6 *Conditioned on $E^\mu(t)$, we have*

$$\begin{aligned} &\mathbb{P}\Big(x_i(t)^\top \tilde{\mu}_i(t) \geq x_i(t)^\top \mu_i + \ell(T)s_i(t) \mid \mathcal{F}_{t-1}\Big) \\ \geq &\frac{1}{2\sqrt{\pi\epsilon \ln T \cdot T^\epsilon}} \end{aligned}$$

Proof. Given the event $E^{\mu}(t)$, we have

$$| x_i(t)^\top \hat{\mu}_i(t) - x_i(t)^\top \mu_i(t) | \leq \ell(T) s_i(t)$$

Since $x_i(t)^\top \widetilde{\mu}_i(t)$ is a Gaussian random variable that has mean $x_i(t)^\top \hat{\mu}_i(t)$ and standard deviation $v s_i(t)$. Using the anti-concentration inequality in Lemma 5 in Agrawal and Goyal (2013), we have

$$\begin{aligned}
&\mathbb{P}\Big(x_i(t)^\top \widetilde{\mu}_i(t) \geq x_i(t)^\top \mu_i + \ell(T) s_i(t) \mid \mathcal{F}_{t-1}\Big) \\
&= \mathbb{P}\Big(\frac{x_i(t)^\top \widetilde{\mu}_i(t) - x_i(t)^\top \hat{\mu}_i(t)}{v s_i(t)} \geq \\
&\quad \frac{x_i(t)^\top \mu_i - x_i(t)^\top \hat{\mu}_i(t) + \ell(T) s_i(t)}{v s_i(t)} \mid \mathcal{F}_{t-1}\Big) \\
&\geq \mathbb{P}\Big(\frac{x_i(t)^\top \widetilde{\mu}_i(t) - x_i(t)^\top \hat{\mu}_i(t)}{v s_i(t)} \geq Z_t \mid \mathcal{F}_{t-1}\Big) \\
&\geq \frac{1}{\sqrt{\pi}} \frac{1}{Z_t + 1/Z_t} \exp(-\frac{Z_t^2}{2})
\end{aligned}$$

where

$$\begin{aligned}
| Z_t | &= \sqrt{\epsilon \ln T} = \frac{2\ell(T)}{v} \\
&\geq | \frac{x_i(t)^\top \mu_i - x_i(t)^\top \hat{\mu}_i(t) + \ell(T) s_i(t)}{v s_i(t)} |
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\Big(x_i(t)^\top \widetilde{\mu}_i(t) \geq x_i(t)^\top \mu_i + \ell(T) s_i(t) \mid \mathcal{F}_{t-1}\Big) \\
&\geq \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{\epsilon \ln T} + 1/\sqrt{\epsilon \ln T}} \exp(-\frac{\epsilon \ln T}{2})
\end{aligned}$$

Without loss of generality, we can set $\epsilon \ln T \geq 1$ for large T. Thus

$$\begin{aligned}
&\mathbb{P}\Big(x_i(t)^\top \widetilde{\mu}_i(t) \geq x_i(t)^\top \mu_i + \ell(T) s_i(t) \mid \mathcal{F}_{t-1}\Big) \\
&\geq \frac{1}{2\sqrt{\pi \epsilon \ln T \cdot T^{\epsilon}}}
\end{aligned}$$

■

Lemma 7 *Conditioned on $E^{\sigma}(t)$, we have*

$$\begin{aligned}
&\mathbb{P}\Big(\widetilde{\sigma}_i^2(t) \leq \sigma_i^2 - h(T) \frac{1}{\sqrt{\#T_i(t)}} \mid \mathcal{F}_{t-1}\Big) \\
&\geq \frac{1}{2\sqrt{\pi \epsilon \ln T \cdot T^{\epsilon}}}
\end{aligned}$$

Proof. The proof is similar to Lemma 6. ■

Define shorthand notations

$$\begin{aligned}
\omega(t) &= x_{a^*(t)}(t)^\top \mu_{a^*(t)} - x_i(t)^\top \mu_i \\
\Gamma_i(t) &= \sigma_i^2 - \sigma_{a^*(t)}^2 \\
\Lambda_i(t) &= x_i(t)^\top \mu_i - \rho \sigma_i^2
\end{aligned}$$

Replacing μ_i, σ_i by their estimates $\hat{\mu}_i(t)$, $\hat{\sigma}_i(t)$, we have corresponding shorthand notations for $\hat{\omega}(t)$, $\hat{\Gamma}_i(t)$ and $\hat{\Lambda}_i(t)$. Replacing $\hat{\mu}_i(t)$, $\hat{\sigma}_i(t)$ by their samples $\widetilde{\mu}_i(t)$, $\widetilde{\sigma}_i(t)$, we have corresponding shorthand notations for $\widetilde{\omega}(t)$, $\widetilde{\Gamma}_i(t)$ and $\widetilde{\Lambda}_i(t)$.

Lemma 8 *Given any filtration $\mathcal{F}_{t-1}$ such that event $E^{\mu}(t)$ and $E^{\sigma}(t)$ hold, we have*

$$\mathbb{P}\big(a(t) \in S(t) \mid \mathcal{F}_{t-1}\big) \leq \frac{1}{p} \mathbb{P}\big(a(t) = a^*(t) \mid \mathcal{F}_{t-1}\big) + \frac{2}{pT^2}$$

where $p = \frac{1}{4\pi\epsilon \ln T \cdot T^{\epsilon}}$.

Proof. Note that the algorithm chooses arm $a^*(t)$ to pull at round t if the following event happens:

$$\widetilde{\omega}_j(t) + \rho \widetilde{\Gamma}_j(t) \geq 0, \forall j \neq a^*(t)$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\big(a(t) = a^*(t) \mid \mathcal{F}_{t-1}\big) \\
&\geq \mathbb{P}\big(\widetilde{\omega}_j(t) + \rho \widetilde{\Gamma}_j(t) \geq 0, \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\big) \\
&\geq \mathbb{P}\big(\exists i \in S(t) : \widetilde{\Lambda}_{a^*(t)}(t) \geq \widetilde{\Lambda}_i(t), \\
&\quad \widetilde{\Lambda}_i(t) \geq \widetilde{\Lambda}_j(t), \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\big) \\
&\geq \mathbb{P}\big(\forall i \in S(t), \widetilde{\Lambda}_{a^*(t)}(t) \geq \Lambda_{a^*(t)}(t) \\
&\quad + \ell(T) s_{a^*(t)}(t) + \rho \frac{h(T)}{\sqrt{\#T_i(t)}} \geq \widetilde{\Lambda}_i(t), \\
&\quad \exists i \in S(t), \widetilde{\Lambda}_i(t) \geq \widetilde{\Lambda}_j(t), \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\big) \\
&\geq \mathbb{P}\big(\widetilde{\Lambda}_{a^*(t)}(t) \geq \Lambda_{a^*(t)}(t) + \ell(T) s_{a^*(t)}(t) \\
&\quad + \frac{\rho h(T)}{\sqrt{\#T_{a^*(t)}(t)}}, \exists i \in S(t), \widetilde{\Lambda}_i(t) \geq \widetilde{\Lambda}_j(t), \\
&\quad \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\big) \\
&\quad - \mathbb{P}\big(\{\forall i \in S(t), \Lambda_{a^*(t)}(t) + \ell(T) s_{a^*(t)}(t)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\rho h(T)}{\sqrt{\#T_{a^*(t)}(t)}} \geq \widetilde{\Lambda}_i(t)\}^c\big) \\
&= \mathbb{P}\big(\widetilde{\Lambda}_{a^*(t)}(t) \geq \Lambda_{a^*(t)}(t) + \ell(T)s_{a^*(t)}(t) \\
&\quad + \frac{\rho h(T)}{\sqrt{\#T_{a^*(t)}(t)}} \mid \mathcal{F}_{t-1}\big) \\
&\quad \cdot \mathbb{P}\big(\exists i \in S(t), \widetilde{\Lambda}_i(t) \geq \widetilde{\Lambda}_j(t), \\
&\quad \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\big) \\
&\quad - \mathbb{P}\big(\exists i \in S(t), \widetilde{\Lambda}_i(t) \geq \Lambda_{a^*(t)}(t) \\
&\quad + \ell(T)s_{a^*(t)}(t) + \frac{\rho h(T)}{\sqrt{\#T_{a^*(t)}(t)}} \mid \mathcal{F}_{t-1}\big)
\end{aligned}
$$

Since

$$
\begin{aligned}
&\mathbb{P}\big(\widetilde{\Lambda}_{a^*(t)}(t) \geq \Lambda_{a^*(t)}(t) + \ell(T)s_{a^*(t)}(t) \\
&\quad + \frac{\rho h(T)}{\sqrt{\#T_{a^*(t)}(t)}} \mid \mathcal{F}_{t-1}\big) \\
&\geq \mathbb{P}\big(x_{a^*(t)}(t)^\top \widetilde{\mu}_{a^*(t)} \geq x_{a^*(t)}(t)^\top \mu_{a^*(t)} \\
&\quad + \ell(T)s_{a^*(t)}(t) \mid \mathcal{F}_{t-1}\big) \\
&\quad \cdot \mathbb{P}\big(\widetilde{\sigma}^2_{a^*(t)}(t) \geq \sigma^2_{a^*(t)} + \frac{h(T)}{\sqrt{\#T_{a^*(t)}(t)}} \mid \mathcal{F}_{t-1}\big) \\
&\geq \frac{1}{4\pi\epsilon \ln T \cdot T^\epsilon} \\
&:= p
\end{aligned}
$$

Also note that when $E^{\widetilde{\mu}}(t) \bigcap E^{\widetilde{\sigma}}(t)$ holds true, we have that $\forall i \in S(t)$,

$$
\begin{aligned}
\widetilde{\Lambda}_i(t) &\leq \Lambda_i(t) + g(T)s_i(t) + \rho h(T)\frac{1}{\sqrt{\#T_i(t)}} \\
&\leq \Lambda_{a^*(t)}(t) + \ell(T)s_{a^*(t)}(t) + \rho q(T)\frac{1}{\sqrt{\#T_i(t)}}
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
&\mathbb{P}\big(a(t) = a^*(t) \mid \mathcal{F}_{t-1}\big) \\
&\geq p \cdot \mathbb{P}\big(\exists i \in S(t), \widetilde{\Lambda}_i(t) \geq \widetilde{\Lambda}_j(t), \\
&\qquad \forall j \neq a^*(t) \mid \mathcal{F}_{t-1}\big) - \frac{2}{T^2} \\
&\geq p \cdot \mathbb{P}\big(a(t) \in S(t) \mid \mathcal{F}_{t-1}\big) - \frac{2}{T^2}
\end{aligned}
$$

Finally, we have

$$
\mathbb{P}\big(a(t) \in S(t) \mid \mathcal{F}_{t-1}\big) \leq \frac{1}{p}\mathbb{P}\big(a(t) = a^*(t) \mid \mathcal{F}_{t-1}\big) + \frac{2}{pT^2}
$$

■

We construct a super-martingale with respect to the regret in Lemma 9, which is used to bound the total regret later using Azuma-Hoeffding inequality.

Lemma 9 *Recall that the regret at round t is $\Delta_{a(t)}(t)$. Denote by $\mathbf{1}\{\cdot\}$ the indicator function. Let $\Delta'_{a(t)}(t) = \Delta_{a(t)}(t) \cdot \mathbf{1}\{E^\mu(t)\} \cdot \mathbf{1}\{E^\sigma(t)\}$. Let*

$$
\begin{aligned}
Y_t =& \Delta'_{a(t)}(t) - s_{a(t)}(t)\big(g(T) + \frac{g(T)^2}{\ell(T)} + \frac{g(T)q(T)}{\rho h(T)}\big) \\
&- \frac{1}{\sqrt{\#T_{a(t)}(t)}}\big(\rho q(T) + \rho\frac{g(T)q(T)}{\ell(T)} \\
&+ \rho\frac{q(T)^2}{h(T)}\big) - s_{a^*}(t)\frac{g(T)}{p}\mathbf{1}\{a(t) = a^*(t)\} \\
&- \frac{1}{\sqrt{\#T_{a^*(t)}(t)}}\rho\frac{q(T)}{p}\mathbf{1}\{a(t) = a^*(t)\} \\
&- \frac{g(T) + \rho q(T)}{pT^2} - (1+\rho)\frac{2}{T^2}
\end{aligned}
$$

Then $\sum_{s=1}^t Y_s$ is a super-martingale process with respect to the filtration $\mathcal{F}_t$.

Proof. To prove $\sum_{s=1}^t Y_s$ is a super-martingale process, we need to show that for all $1 \leq t \leq T$ and a given filtration $\mathcal{F}_{t-1}$, $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \leq 0$. Conditioned on $E^\mu(t)$ and $E^\sigma(t)$, if both $E^{\widetilde{\mu}}(t)$ and $E^{\widetilde{\sigma}}(t)$ hold true, we have

$$
\begin{aligned}
\omega_i(t) &\leq \widetilde{\omega}_i(t) + g(T)(s_{a^*(t)}(t) + s_{a(t)}(t)) \\
\Gamma_i(t) &\leq \widetilde{\Gamma}_i(t) + \rho q(T)\big(\frac{1}{\sqrt{\#T_i(t)}} + \frac{1}{\sqrt{\#T_{a^*(t)}(t)}}\big)
\end{aligned}
$$

Observe that

$$
\begin{aligned}
&\mathbb{E}[\Delta'_{a(t)}(t) \mid \mathcal{F}_{t-1}] \\
&\leq g(T)\mathbb{E}[s_{a(t)}(t) \mid \mathcal{F}_{t-1}] \\
&\quad + q(T)\mathbb{E}[\frac{1}{\sqrt{\#T_{a(t)}(t)}} \mid \mathcal{F}_{t-1}] \\
&\quad + g(T)s_{a^*(t)}(t) + \rho q(T)\frac{1}{\sqrt{\#T_{a^*(t)}(t)}} \\
&\quad \cdot \mathbb{E}[\mathbf{1}\{a(t) \in S(t)\} + \mathbf{1}\{a(t) \notin S(t)\} \mid \mathcal{F}_{t-1}] \\
&\quad + (1+\rho)(1 - \mathbb{P}(E^{\widetilde{\mu}})) + (1+\rho)(1 - \mathbb{P}(E^{\widetilde{\sigma}})) \\
&\leq g(T)\mathbb{E}[s_{a(t)}(t) \mid \mathcal{F}_{t-1}]
\end{aligned}
$$

$$+\rho q(T)\mathbb{E}[\frac{1}{\sqrt{\#T_{a(t)}(t)}}\mid\mathcal{F}_{t-1}]$$
$$+\big(g(T)s_{a^*(t)}(t)+\rho q(T)\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}\big)$$
$$\cdot\mathbb{P}(a(t)\in S(t)\mid\mathcal{F}_{t-1}) \tag{6}$$
$$+\mathbb{E}\Big[\big(g(T)s_{a^*(t)}(t)+\rho q(T)\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}\big)$$
$$\cdot\mathbf{1}\{a(t)\notin S(t)\}\mid\mathcal{F}_{t-1}\Big]+(1+\rho)\frac{2}{T^2} \tag{7}$$

Note that for (6), we have

$$(6)\le\big(g(T)s_{a^*(t)}(t)+\rho q(T)\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}\big)\cdot\Big[\frac{1}{p}\mathbb{P}(a(t)=a^*(t)\mid\mathcal{F}_{t-1})+\frac{1}{pT^2}\Big]$$
$$\le s_{a^*(t)}(t)\frac{g(T)}{p}\mathbb{P}(a(t)=a^*(t)\mid\mathcal{F}_{t-1})+\rho\frac{q(T)}{p}\mathbb{P}(a(t)=a^*(t)\mid\mathcal{F}_{t-1})\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}+\frac{g(T)}{pT^2}+\frac{\rho g(T)}{pT^2}$$

For (7), notice that when $a(t)\notin S(t)$, we have

$$g(T)s_{a(t)}(t)+\rho q(T)\frac{1}{\sqrt{\#T_{a(t)}(t)}}>\ell(T)s_{a^*(t)}(t)+\rho h(T)\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}$$

Rearrange the above inequality, we have

$$s_{a^*(t)}(t)<\frac{g(T)}{\ell(T)}s_{a(t)}(t)+\frac{\rho q(T)}{\ell(T)}\frac{1}{\sqrt{\#T_{a(t)}(t)}}$$

Also note that

$$\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}<\frac{g(T)}{\rho h(T)}s_{a(t)}(t)+\frac{q(T)}{h(T)}\frac{1}{\sqrt{\#T_{a(t)}(t)}}$$

Hence

$$(7)\le\mathbb{E}\Big[\big(\frac{g(T)^2}{\ell(T)}+\frac{g(T)q(T)}{\rho h(T)}\big)s_{a(t)}(t)+\big(\frac{\rho g(T)q(T)}{\ell(T)}+\frac{\rho q(T)^2}{h(T)}\big)\cdot\frac{1}{\sqrt{\#T_{a(t)}(t)}}\mid\mathcal{F}_{t-1}\Big]$$

Putting all these together, we have $\mathcal{F}_{t-1}$, $\mathbb{E}[Y_t\mid\mathcal{F}_{t-1}]\le 0$, thus $\sum_{s=1}^{t}Y_s$ is a super-martingale process. ■

Now we start to prove the main result Theorem 1.

Proof. First observe that we can bound the absolute value of Y_t by $5(1+\rho+\frac{1}{\rho})\frac{q(T)^2}{\ell(T)}$. Therefore, by the Azuma-Hoeffding inequality, we have

$$\mathbb{P}(\sum_{t=1}^{T}Y_t\ge w)\le\exp\Big(-\frac{w^2\ell(T)^2}{(5(1+\rho+1/\rho))^2q(T)^4}\Big):=\frac{\delta}{2}$$

Thus we set $w=5(1+\rho+1/\rho)\frac{q(T)^2}{\ell(T)}\sqrt{2T\ln\frac{2}{\delta}}$. Then with probability at least $1-\frac{\delta}{2}$, we have

$$\sum_{t=1}^{T}\Delta'_{a(t)}(t)\le\big(g(T)+\frac{g(T)^2}{\ell(T)}+\frac{g(T)q(T)}{\rho h(T)}+\frac{g(T)}{p}\big)\cdot\sum_{t=1}^{T}s_{a(t)}(t)+\rho\big(g(T)+\frac{g(T)q(T)}{\ell(T)}+\frac{g(T)^2}{h(T)}+\frac{q(T)}{p}\big)\cdot\sum_{t=1}^{T}\frac{1}{\sqrt{\#T_{a^*(t)}(t)}}+\frac{g(T)+\rho q(T)}{pT}+5(1+\rho+1/\rho)\frac{q(T)^2}{\ell(T)}\sqrt{2T\ln\frac{2}{\delta}}$$

Using Lemma 3 in Chu et al. (2011), we have

$$\sum_{t=1}^{T}s_{a(t)}(t)=\sum_{i=1}^{K}\sum_{s\in T_i(T)}s_i(t)\le\sum_{i=1}^{K}5\sqrt{d\#T_i(t)\ln\#T_i(t)}\le 5\sqrt{dKT\ln T}$$
$$\sum_{t=1}^{T}\frac{1}{\sqrt{\#T_{a(t)}(t)}}=\sum_{i=1}^{K}\sum_{s\in T_i(T)}s_i(t)\frac{1}{\sqrt{\#T_{a(t)}(t)}}=\sum_{i=1}^{K}\sum_{s=1}^{\#T_i(t)}\frac{1}{\sqrt{s}}\le K\frac{1}{K}\sum_{i=1}^{K}2\sqrt{\#T_i(t)}\le 2K\sqrt{\sum_{i=1}^{K}\frac{\#T_i(t)}{K}}=2\sqrt{KT}$$

Hence, with probability at least $1-\frac{\delta}{2}$, we have

$$\sum_{t=1}^{T}\Delta'_{a(t)}(t)=O((1+\rho+\frac{1}{\rho})d\ln T\ln\frac{K}{\delta}\sqrt{dKT^{1+2\epsilon}\ln\frac{K}{\delta}\frac{1}{\epsilon}})$$

Since $E^{\mu}(t)$ does not hold with probability at most $\frac{\delta}{T^2}T = \frac{\delta}{T} \leq \frac{\delta}{4}$ for $T \geq 4$, and $E^{\sigma}(t)$ does not hold with probability at most $\frac{\delta}{4}$. Therefore, both $E^{\mu}(t)$ and $E^{\sigma}(t)$ holds for all t with probability at least $1 - \frac{\delta}{2}$. Thus $\Delta_{a(t)}(t) = \Delta'_{a(t)}(t)$ for all t with probability at least $1 - \frac{\delta}{2}$. Hence with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{T} \Delta_{a(t)}(t) = O((1 + \rho + \frac{1}{\rho})d \ln T \ln \frac{K}{\delta} \sqrt{dKT^{1+2\epsilon} \ln \frac{K}{\delta} \frac{1}{\epsilon}})$$

■

5. Numerical Experiments

In the numerical experiment, we apply our proposed TS algorithms to a portfolio selection problem.

5.1 Contextual MAB Application to Finance

Application of the bandit algorithm to the portfolio selection problem is not new. To name a few, Shen et al. (2015) apply a UCB-type bandit algorithm to derive the optimal portfolio strategy that represents the combination of passive and active investments according to a risk-adjusted reward function. Huo and Fu (2017) apply a UCB-type bandit algorithm to the portfolio selection problem, under a risk-averse criterion. Zhu et al. (2020) propose an online portfolio selection method that also incorporates contextual information, based on the Exp4 algorithm presented in Auer et al. (2002). We adapt the portfolio selection model in Huo and Fu (2017) to our contextual setting and formally describe the problem setting below.

Consider a financial market with a large set of assets (for example, bonds, stocks and other financial derivatives), from which the portfolio manager selects to construct K portfolios. Each portfolio consists of different assets with different weights. The industries are roughly divided into eleven sectors, namely energy, materials, industrials, communication services, consumer discretionary, consumer staples, healthcare, financials, information technology, real estate, and utilities (Nagy and Ormos 2018). At each round t, the manager collects information about industrial prosperity in those sectors, which makes up the contexts $x_i(t) \in [-1, 1]^d$ for each portfolio $i \in [K]$, where $d \leq 11$ due to the possibility of being incapable to collect information for every sector. A larger context $x_i(t)^j$ indicates a better market condition for the sector $j \in [d]$. After observing the contexts, the manager chooses one portfolio to invest and receives the corresponding reward. For simplicity, we assume the reward of portfolio i follows an unknown distribution ν_i with mean $x_i(t)^{\top}\mu_i$ and variance σ_i^2. The unknown mean parameter $\mu_i \in \mathbb{R}^d$ can be viewed as the sensitivity of the return to the industrial prosperity. The manager is risk-averse with a risk tolerance ρ. The goal is to minimize the cumulative regret over T rounds under the mean-variance criterion.

5.2 Algorithms for Comparison

We empirically evaluate the following algorithms in the portfolio selection problem.

- Our proposed MVTS-D algorithm (Algorithm 2).
- A variant of the TS algorithm MVTS-DN used in our regret analysis (Algorithm 3).
- TS algorithm originally designed for the risk-neutral setting. We compare with the TS algorithm from Agrawal and Goyal (2013), referred to as TS-A.
- Algorithms that make no use of the contexts. In particular, we compare with the Thompson sampling algorithm with

mean-variance criterion for the context-free MAB setting (Algorithm MVTS in Zhu and Tan (2020)).

- A uniform sampling algorithm that randomly chooses an arm to pull at each round.

To illustrate the necessity of taking into account the risk of the reward, we compare with the TS-A algorithm that works for a risk-neutral setting. At each round t, the TS-A algorithm samples $\widetilde{\mu}_i(t)$ from the Gaussian distribution $\mathcal{N}(\hat{\mu}_i(t), v^2 A_i(t)^{-1})$ for each arm $i \in [K]$, and plays the arm $a(t) := \arg\max x_i(t)^\top \widetilde{\mu}_i(t)$. Here $\hat{\mu}_i(t) = A_i(t)^{-1} b_i(t)$, the reward is assumed to follow a R-sub-Gaussian distribution, parameter $v = R\sqrt{\frac{24}{\epsilon} d \ln\left(\frac{1}{\delta}\right)}$, where $\delta \in (0,1)$ and $\epsilon \in (0,1)$ are two parameters used by the algorithm. To illustrate the necessity of making use of contexts that enables to learn the mean and variance parameters over time, we compare with the context-free MVTS algorithm. We include the details of the TS-A algorithm from Agrawal and Goyal (2013) and the context-free MVTS algorithm in Zhu and Tan (2020) in Appendix C.

All the algorithms are tested on the portfolio selection problem over 100 replications. In each replication, we execute the algorithms and collect the total regrets over T rounds. Parameters setting are summarized as follows: $K = 10$, $d = 8$, $T = 10000$. All the implementing details are included in Appendix B.

5.3 Experimental Results

Experiment 1: evaluation of total regrets with different risk tolerances. In this experiment, we evaluate the total regrets of different algorithms associated with different risk tolerances $\rho = 0.1, 1, 10$. The reward distribution is Gaussian. Results are reported in Figure 1.

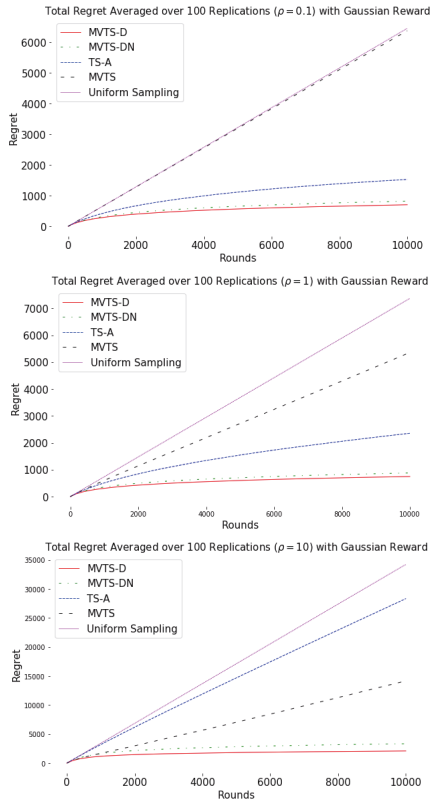

Figure 1 Total Regrets Comparison with Different Risk Tolerances in the Portfolio Selection Problem Averaged over 100 Replications

Figure 1 shows the mean (over 100 replications) of the total regrets over time under different risk tolerances for our proposed MVTS-D and MVTS-DN algorithms, along with three benchmarks, namely TS-A, MVTS, and Uniform Sampling. The (95%) confidence band across different replications is very narrow, hence it is not reported in the figure. From Figure 1, we have the following observations:

- Our proposed MVTS-D and MVTS-DN algorithms achieve better regrets compared to the three benchmarks in all the cases ($\rho = 0.1, 1, 10$).
- As ρ approaches 0 (i.e., risk-neutral case), our proposed MVTS-D and MVTS-

DN algorithms behave similarly to TS-A, which corresponds to the risk-neutral case. As ρ increases, our proposed MVTS-D and MVTS-DN algorithms have similarly steady performance. As for TS-A, it chooses the optimal arm according to its mean performance while overlooking the variance. As ρ increases, variance tends to dominate the choice of the optimal arm. Hence, the performance of TS-A deteriorates.

- In all cases ($\rho = 0.1, 1, 10$), MVTS and Uniform Sampling behave much worse than our proposed MVTS-D and MVTS-DN algorithms, as they make no use of the contexts and thus do not learn over time.

Experiment 2: evaluation of total regrets under different reward distributions. The reward distributions are chosen to be Gaussian, truncated normal, and uniform, respectively. Results are reported in Figure 2.

Figure 2 shows the mean (over 100 replications) of the total regrets over time under different reward distributions for our proposed MVTS-D and MVTS-DN algorithms, along with three benchmarks, namely TS-A, MVTS, and Uniform Sampling. From Figure 2, we have the following observations:

- Our proposed MVTS-D and MVTS-DN algorithms are robust to model misspecification, i.e., the assumed reward distribution is different from the true one.
- Even though our regret analysis does not work for the Gaussian distribution (as the squared reward is sub-exponential instead of sub-Gaussian), in practice our proposed MVTS-DN algorithm still works well.

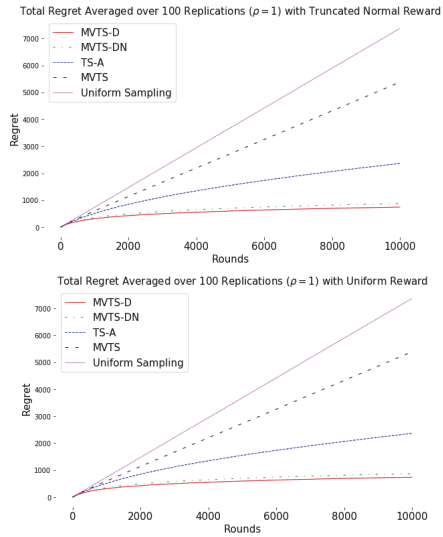

Figure 2 Total Regrets Comparison under Different Reward Distributions in the Portfolio Selection Problem Averaged over 100 Replications. $\rho = 1$

6. Conclusion

In this paper, we apply the Thompson sampling algorithm to the contextual MAB problem under the mean-variance criterion. We show a high probability regret bound for a variant of the proposed TS algorithm. The performances of the proposed algorithm and its variant are empirically shown via a portfolio selection example, with a wide range of reward distributions. It could be an interesting future direction to apply TS algorithm, or UCB-type algorithm, to contextual MAB problem under other risk measures, such as CVaR.

Appendix A Algorithms

See page 286-287.

Appendix B Implementation Details

For the portfolio selection problem, the true mean and variance parameters are summarized in Table 1. In each replication, a new

Table 1 True Mean Parameter $\{\mu_i\}_{i=1}^8$ and Variance Parameter σ for all Ten Portfolios in the Portfolio Selection Problem. True Means are Generated from Uniform Distribution with Low = -0.1 and High = 0.5. True Variances are Generated from Uniform Distribution with Low = 0.1 and High = 1

	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	σ^2
Portfolio 1	0.15	0.33	−0.10	0.08	−0.01	−0.04	0.01	0.11	0.89
Portfolio 2	0.14	0.22	0.15	0.31	0.02	0.43	−0.08	0.30	0.66
Portfolio 3	0.15	0.24	−0.02	0.02	0.38	0.48	0.09	0.32	0.78
Portfolio 4	0.43	0.44	−0.05	−0.08	0.00	0.43	−0.04	0.15	0.41
Portfolio 5	0.47	0.22	0.32	0.09	0.31	0.40	−0.09	0.35	0.34
Portfolio 6	0.49	0.35	0.07	0.37	−0.04	0.17	0.45	0.08	0.91
Portfolio 7	0.07	−0.02	−0.09	0.31	0.03	0.06	0.19	−0.07	0.49
Portfolio 8	0.24	−0.01	0.25	0.32	−0.04	0.15	0.32	0.15	0.97
Portfolio 9	−0.07	0.22	0.30	0.21	0.47	0.25	0.44	−0.02	0.70
Portfolio 10	−0.02	0.38	0.14	−0.00	0.46	0.11	0.35	0.33	0.66

set of contexts are generated and used by all the algorithms. When the reward distribution is truncated normal, we assume it is truncated above -5 and below 5. It is worth noting that when executing the TS-A algorithm in the risk-neutral case, for a small ϵ and δ, the parameter v is computed so large that the total regret grows linearly. This is due to the overly large variance in the mean sampling. For meaningful experiment, we set $v = 1$. For our proposed MVTS-DN algorithm, we face the same situation. Therefore, in the experiments, we set $u = 1$ and $v = 1$, which is different from their theoretical values.

Appendix C Proof of Proposition 1

Proof. Let $m_i(t) = A_i(t)^{-1}b_i(t)$. The prior distribution is given by:

$$\begin{aligned}
&\mathbb{P}(\mu_i, \lambda_i) = \mathbb{P}(\mu_i \mid \lambda_i)\mathbb{P}(\lambda_i) \\
&= \mathcal{N}\left(m_i(t), (\lambda_i A_i(t))^{-1}\right) \cdot \text{Gamma}\left(C_i(t), D_i(t)\right) \\
&\propto \sqrt{\lambda_i}\exp\Big(-\frac{\lambda_i}{2}(\mu_i - m_i(t))^\top A_i(t) \\
&\quad (\mu_i - m_i(t))\Big) \cdot \lambda_i^{C_i(t)-1}\exp\left(-D_i(t)\lambda_i\right)
\end{aligned}$$

Similarly, the likelihood of reward $r_i(t)$ is given by:

$$\mathbb{P}(r_i(t) \mid \mu_i, \lambda_i) \propto \sqrt{\lambda_i}\exp\left(-\frac{\lambda_i}{2}\left(\mu_i^\top x_i(t) - r_i(t)\right)^2\right)$$

Then the posterior distribution is computed as:

$$\begin{aligned}
&\mathbb{P}(\mu_i, \lambda_i \mid r_i(t)) \propto \mathbb{P}(\mu_i, \lambda_i) \cdot \mathbb{P}(r_i(t) \mid \mu_i, \lambda_i) \\
&\propto \lambda_i^{C_i(t)}\exp\Big(-\frac{\lambda_i}{2}\big((\mu_i - m_i(t))^\top A_i(t) \\
&\quad (\mu_i - m_i(t)) + \mu_i^\top x_i(t)x_i(t)^\top\mu_i \\
&\quad - 2\mu_i^\top x_i(t)r_i(t) + r_i(t)^2 + 2D_i(t)\big)\Big) \\
&= \lambda_i^{C_i(t)}\exp\Big(-\frac{\lambda_i}{2}\Big(\mu_i^\top\left(A_i(t) + x_i(t)x_i(t)^\top\right)\mu_i \\
&\quad - 2\mu_i^\top\left(b_i(t) + x_i(t)r_i(t)\right) \\
&\quad + b_i(t)^\top A_i(t)^{-1}b_i(t) + 2D_i(t) + r_i(t)^2\Big)\Big) \\
&= \exp\Big(-\frac{\lambda_i}{2}\big(\mu_i^\top A_i(t+1)\mu_i - 2\mu_i^\top A_i(t+1) \\
&\quad m_i(t+1) + b_i(t)^\top A_i(t)^{-1}b_i(t) + 2D_i(t) + r_i(t)^2\big)\Big)\lambda_i^{C_i(t)} \\
&= \sqrt{\lambda_i}\exp\Big(-\frac{\lambda_i}{2}\big((\mu_i - m_i(t+1))^\top A_i(t+1) \\
&\quad (\mu_i - m_i(t+1))\big)\Big) \cdot \lambda_i^{C_i(t+1)-1} \cdot \exp(-D_i(t+1)\lambda_i) \\
&\propto \mathcal{N}(A_i(t+1)^{-1}b_i(t+1), (\lambda_i A_i(t+1))^{-1}) \\
&\quad \cdot \text{Gamma}(C_i(t+1), D_i(t+1))
\end{aligned}$$

■

Endnote

* These authors contributed equally to this work.

Acknowledgments

The authors gratefully acknowledge the support by the Air Force Office of Scientific

Algorithm 3: Mean-Variance Thompson Sampling for the Disjoint Model with Variance of the Reward Sampled from Normal Distribution (MVTS-DN)

initialization:
for $i = 1, 2, \cdots, K$ **do**
  pull each arm i once at round 0 and observe rewards $r_i(0)$;
  $A_i(1) = \mathbf{I}_d + x_i(0)x_i(0)^\top$, $b_i(1) = x_i(0)r_i(0)$, $C_i(1) = \frac{1}{2}$, $D_i(1) = \frac{1}{2}(r_i(0)^2 - x_i(0)^\top A_i(1)^{-1}x_i(0))$, $T_i(1) = \{0\}$;
end
for $t = 1, 2, \cdots, T$ **do**
  observe K contexts $x_1(t), \cdots, x_K(t) \in \mathbb{R}^d$;
  for $i = 1, 2, \cdots, K$ **do**
    sample $\widetilde{\sigma}_i^2(t)$ from distribution $\mathcal{N}\left(\frac{D_i(t)}{C_i(t)}, \frac{u^2}{\#T_i(t)}\right)$;
    sample $\widetilde{\mu}_i(t)$ from distribution $\mathcal{N}\left(A_i(t)^{-1}b_i(t), v^2A_i(t)^{-1}\right)$;
    set $\widetilde{\mathrm{MV}}_i(t) = x_i(t)^\top\widetilde{\mu}_i(t) - \rho\widetilde{\sigma}_i^2(t)$;
  end
  play arm $a(t) = \arg\max_{i\in[K]} \widetilde{\mathrm{MV}}_i(t)$ with ties broken arbitrarily;
  observe reward $r_{a(t)}(t) \sim \nu_{a(t)}\left(x_{a(t)}(t)^\top\mu_{a(t)}, \sigma_{a(t)}^2\right)$;
  update $(A_i(t), b_i(t), C_i(t), D_i(t))$ according to Algorithm 1 for each arm i;
  set $T_{a(t)}(t+1) = T_{a(t)}(t)\bigcup\{t\}$.
end

Algorithm 4: TS-A Algorithm from Agrawal and Goyal (2013)

initialization:
pull each arm i once at round 0 and observe rewards $r_i(0)$;
$A_i(1) = \mathbf{I}_d + x_i(0)x_i(0)^\top$, $b_i(1) = x_i(0)r_i(0)$, $T_i(1) = \{0\}$;
for $t = 1, 2, \cdots, T$ **do**
  observe K contexts $x_1(t), \cdots, x_K(t) \in \mathbb{R}^d$;
  for $i = 1, 2, \cdots, K$ **do**
    compute $\hat{\mu}_i(t) = A_i(t)^{-1}b_i(t)$;
    sample $\widetilde{\mu}_i(t)$ from distribution $\mathcal{N}\left(\hat{\mu}_i(t), v^2A_i(t)^{-1}\right)$;
  end
  play arm $a(t) = \arg\max_{i\in[K]} x_i(t)^\top\widetilde{\mu}_i(t)$ with ties broken arbitrarily;
  observe reward $r_{a(t)}(t) \sim \nu_{a(t)}\left(x_{a(t)}(t)^\top\mu_{a(t)}, \sigma_{a(t)}^2\right)$;
  update $(A_i(t), b_i(t))$ according to Line 4 and Line 5 in Algorithm 1 for each arm i;
  set $T_{a(t)}(t+1) = T_{a(t)}(t)\bigcup\{t\}$.
end

Algorithm 5: Posterior Updating in the MVTS Algorithm

input : prior parameters $(\hat{\mu}_i(t-1), \hat{T}_i(t-1), \hat{\alpha}_i(t-1), \hat{\beta}_i(t-1))$ and new reward sample $r_i(t)$.
output: posterior parameters $(\hat{\mu}_i(t), \hat{T}_i(t), \hat{\alpha}_i(t), \hat{\beta}_i(t))$.
update the mean: $\hat{\mu}_i(t) = \frac{\hat{T}_i(t-1)}{\hat{T}_i(t-1)+1}\hat{\mu}_i(t-1) + \frac{1}{\hat{T}_i(t-1)+1}r_i(t)$;
update the number of samples: $\hat{T}_i(t) = \hat{T}_i(t-1) + 1$;
update the shape parameter: $\hat{\alpha}_i(t) = \hat{\alpha}_i(t-1) + \frac{1}{2}$;
update the rate parameter: $\hat{\beta}_i(t) = \hat{\beta}_i(t-1) + \frac{\hat{T}_i(t-1)}{\hat{T}_i(t-1)+1} \cdot \frac{(r_i(t)-\hat{\mu}_i(t-1))^2}{2}$.

Algorithm 6: MVTS Algorithm in Zhu and Tan (2020)

initialization:
for $i = 1, 2, \cdots, K$ **do**
  pull each arm i once at round 0 and observe rewards $r_i(0)$;
  $\hat{\mu}_i(0) = r_i(0), \hat{T}_i(0) = 1, \hat{\alpha}_i(0) = \frac{1}{2}, \hat{\beta}_i(0) = \frac{1}{2}$;
end
for $t = 1, 2, \cdots, T$ **do**
  for $i = 1, 2, \cdots, K$ **do**
    sample $\tau_i(t)$ from Gamma$(\hat{\alpha}_i(t-1), \hat{\beta}_i(t-1))$;
    sample $\theta_i(t)$ from $\mathcal{N}(\hat{\mu}_i(t-1), \frac{1}{\hat{T}_i(t-1)})$;
  end
  play arm $a(t) = \arg\max_{i \in [K]} \theta_i(t) - \rho/\tau_i(t)$ and observe reward $r_{a(t)}(t)$;
  update $(\hat{\mu}_{a(t)}(t-1), \hat{T}_{a(t)}(t-1), \hat{\alpha}_{a(t)}(t-1), \hat{\beta}_{a(t)}(t-1))$ according to Algorithm 5.
end

Research under Grant FA9550-19-1-0283 and Grant FA9550-22-1-0244, and National Science Foundation under Grant DMS2053489. The authors would also like to thank the anonymous reviewers for their valuable comments.

Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Abbasi-yadkori Y, Pál D, Reyzin L, Szepesvári C (2011). Improved algorithms for linear stochastic bandits. *Proceedings of the 25th International Conference on Neural Information Processing Systems.* Granada, Spain, December 12-14, 2011.

Agrawal S, Goyal N (2013). Thompson sampling for contextual bandits with linear payoffs. *Proceedings of the 30th International Conference on Machine Learning.* Atlanta, USA, June 16-21, 2013.

Ang M L, Lim E Y, Chang J Q (2021). Thompson sampling for Gaussian entropic risk bandits. *arXiv preprint arXiv:2105.06960*.

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

Auer P, Cesa-Bianchi N, Freund Y, Schapire R E (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* 32(1): 48-77.

Baudry D, Gautron R, Kaufmann E, Maillard O (2021). Optimal Thompson sampling strategies for support-aware cvar bandits. *Proceedings of the 38th International Conference on Machine Learning*. Virtual Event, July 18-24, 2021.

Cassel A, Mannor S, Zeevi A (2018). A general approach to multi-armed bandits under risk criteria. *Proceedings of the 31st Conference on Learning Theory.* Singapore, October 6-9, 2013.

Chang J Q, Zhu Q, Tan V (2020). Risk-constrained Thompson sampling for cvar bandits. *arXiv preprint arXiv:2011.08046*.

Chapelle O, Li L (2011). An empirical evaluation of Thompson sampling. *Proceedings of the 25th International Conference on Neural Information Processing Systems.* Granada, Spain, December 12-14, 2011.

Chu W, Li L, Reyzin L, Schapire R (2011). Contextual bandits with linear payoff functions. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics.* Fort Lauderdale, USA, April 11-13, 2011.

Galichet N, Sebag M, Teytaud O (2013). Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. *Proceedings of the 5th Asian Conference on Machine Learning.* Canberra, ACT, Australia, November 13-15, 2013.

Huo X, Fu F (2017). Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science* 4(11): 171377.

Khajonchotpanya N, Xue Y, Rujeerapaiboon N (2021). A revised approach for risk-averse multi-armed bandits under cvar criterion. *Operations Research Letters* 49(4): 465-472.

Kleijn B, Bas J K, Van d V A W (2012). The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics* 6: 354-381.

Kullback S, Leibler R A (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 22(1): 79-86.

Li L, Chu W, Langford J, Schapire R E (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web.* Raleigh, North Carolina, USA, April 26-30, 2010.

Liu X, Derakhshani M, Lambotharan S, Van d S M (2020). Risk-aware multi-armed bandits with refined upper confidence bounds. *IEEE Signal Processing Letters* 28: 269-273.

Maillard O (2013). Robust risk-averse stochastic multi-armed bandits. *Proceedings of the 24th International Conference on Algorithmic Learning Theory.* Singapore, October 6-9, 2013.

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

Nagy L, Ormos M (2018). Review of Global Industry Classification. *Proceedings of the 32nd European Conference on Modelling and Simulation.* Wilhelmshaven, Germany, May 22-25, 2018.

Rockafellar R T, Uryasev S (2000). Optimization of conditional value-at-risk. *Journal of Risk* 2: 21-41.

Sani A, Lazaric A, Munos R (2012). Risk-aversion in multi-armed bandits. *Proceedings of the 26th International Conference on Neural Information Processing Systems.* Lake Tahoe, USA, December 3-6, 2012.

Schlaifer R, Raiffa H (1916). *Applied statistical decision theory*. Harvard University, Cambridge.

Shen W, Wang J, Jiang Y, Zha H (2015). Portfolio choices with orthogonal bandit learning. *Proceedings of the 24th International Joint Conference on Artificial Intelligence.* Buenos Aires, Argentina, July 25-31, 2015.

Thompson W R (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3-4): 285-294.

Tran M, Nguyen T, Dao V (2021). A practical tutorial on variational Bayes. *arXiv preprint arXiv:2103.01327*.

Vakili S, Zhao Q (2015). Mean-variance and value at risk in multi-armed bandit problems. *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing.* Illinois, USA, September 29 - October 2, 2015.

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

Wang Y, Blei D (2019). Variational bayes under model misspecification. *Proceedings of the 33rd International Conference on Neural Information Processing Systems.* Vancouver, BC, Canada, December 8-14, 2019.

Xu P, Zheng H, Mazumdar E, Azizzadenesheli K, Anandkumar A (2022). Langevin monte carlo for contextual bandits. *Proceedings of the 39th International Conference on Machine Learning.* Maryland, USA, July 17-23, 2022.

Zhu M, Zheng X, Wang Y, Liang Q, Zhang W (2020). Online portfolio selection with cardinality constraint and transaction costs based on contextual bandit. *Proceedings of the 30th International Joint Conference on Artificial Intelligence.* Virtual Event, January 7-15, 2021.

Zhu Q, Tan V (2020). Thompson sampling algorithms for mean-variance bandits. *Proceedings of the 37th International Conference on Machine Learning*. Virtual Event, July 13-18, 2020.

Zimin A, Ibsen-Jensen R, Chatterjee K (2014). Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*.

Yifan Lin is a Ph.D. student in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology, USA. He received his B.Econ. in financial engineering and B.E. in computer science from Wuhan University, China, in 2017, and M.S. in operations research from Columbia University, in 2018. His research interest is in simulation and optimization.

Yuhao Wang is a Ph.D. student in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology, USA. He received his B.S. in mathematics from Nanjing University, China, in 2021. His research interest is in simulation and optimization.

Enlu Zhou is an associate professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. She received the B.S. degree with highest honors in electrical engineering from Zhejiang University, China, in 2004, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 2009. Prior to joining Georgia Tech in 2013, she was an assistant professor in the Industrial & Enterprise Systems Engineering Department at the University of Illinois Urbana-Champaign from 2009-2013. She is a recipient of the Best Theoretical Paper award at the Winter Simulation Conference, AFOSR Young Investigator award, NSF CAREER award, and INFORMS Outstanding Simulation Publication Award. She has served as an associate editor for Journal of Simulation, IEEE Transactions on Automatic Control, and Operations Research.