Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

RISK-AVERSE CONTEXTUAL MULTI-ARMED BANDIT PROBLEM WITH LINEAR PAYOFFS

Yifan Lin
Yuhao Wang
Enlu Zhou

H. Milton School of Industrial and Systems Engineering
Georgia Institute of Technology
755 Ferst Drive NW
Atlanta, GA 30332, USA

ABSTRACT

In this paper we consider the contextual multi-armed bandit problem for linear payoffs under a risk-averse mean-variance criterion. 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 contexts, we prove a regret bound of $O((1+\rho+\frac{1}{\rho})d\ln T\ln\frac{K}{\delta}\sqrt{dKT^{1+2\varepsilon}\ln\frac{K}{\delta}\frac{1}{\varepsilon}})$ that holds with probability $1-\delta$ under the mean-variance criterion with risk tolerance ρ, for any $0<\varepsilon<\frac{1}{2}$, $0<\delta<1$. The empirical performance of the algorithms is demonstrated via a portfolio selection problem.

1 INTRODUCTION AND PROBLEM SETTING

The multi-armed bandit (MAB) problem is a classical online decision-making problem with limited feedback. In this paper, we consider the MAB problem with contexts (also known as covariates or side information): at each round $t=1,2,\cdots,T$, a context $x_i(t)\in\mathbb{R}^d$ is revealed for each arm $i\in K$. After observing the contexts, the decision maker plays one of the K arms $a(t)$ and receives a reward $r_{a(t)}(t)$ (also called payoff) of that arm. We assume the reward for arm i at round t is generated from an unknown distribution ν_i with mean $x_i(t)^\top\mu_i$ linear in the context and variance σ_i^2, where $\mu_i\in\mathbb{R}^d$ is the mean parameter and σ_i^2 is the variance parameter. This model is called *disjoint* since the mean and variance parameters are not shared among different arms. In traditional MAB or contextual MAB, the best arm is usually the one with the largest expected reward. 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 and choose the mean-variance criterion given its advantages in interpretability, computation, and popularity among practitioners. Let the mean-variance of arm i at round t be $\mathrm{MV}_i(t):=x_i(t)^\top\mu_i-\rho\sigma_i^2$, where $\rho\geq 0$ is the risk tolerance that reflects the risk attitude of the decision maker. The goal of the risk-averse decision maker is to minimize the cumulative regret, defined as $\mathscr{R}(T):=\sum_{t=1}^{T}\mathrm{MV}_{a^*(t)}(t)-\mathrm{MV}_{a(t)}(t)$, where $a(t)$ is the action chosen by the algorithm at round t. and $a^*(t):=\arg\max_{i\in[K]}x_i(t)^\top\mu_i-\rho\sigma_i^2$.

2 ALGORITHM AND IMPLEMENTATION

To solve risk-averse contextual MAB, we propose an algorithm based on Thompson Sampling (TS). 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. We assume a Gaussian likelihood for the reward, and use the normal-gamma conjugate prior for the mean and variance parameters. Algorithm 1 gives the full mean-variance TS algorithm for the disjoint model.

initialization:
pull each arm i once at round 0 and observe rewards $r_i(0)$; set $A_i(1) = \mathbf{I}_d + x_i(0)x_i(0)^\top$,
$b_i(1) = x_i(0)r_i(0)$, $C_i(1) = \frac{1}{2}$, $D_i(1) = \frac{1}{2}(r_i(0)^2 - x_i(0)^\top A_i(1)^{-1}x_i(0))$, $T_i(1) = \{0\}$, for all $i \in [K]$;
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 T_i, A_i, b_i, C_i, D_i only for $i = a(t)$: $T_{a(t)}(t+1) = T_{a(t)}(t) \bigcup \{t\}$;
  $A_i(t+1) = A_i(t) + x_i(t)x_i(t)^\top$; $b_i(t+1) = b_i(t) + x_i r_i(t)$; $C_i(t+1) = C_i(t) + \frac{1}{2}$;
  $D_i(t+1) = D_i(t) + \frac{1}{2}[b_i(t)^T A_i(t)^{-1}b_i(t) - b_i(t+1)^\top A_i(t+1)^{-1}b_i(t+1) + r_i(t)^2]$.
end

Algorithm 1: Mean-variance Thompson sampling for the disjoint model (MVTS-D).

In the numerical experiment, we apply our proposed TS algorithms to a portfolio selection problem. We empirically evaluate the following algorithms in the portfolio selection problem: (1) our proposed MVTS-D algorithm; (2) a variant MVTS-DN used in our regret analysis that samples the variance from a Gaussian distribution instead of the Gamma distribution; (3) a TS algorithm originally designed for the risk-neutral setting from Agrawal and Goyal (2013), namely TS-A; (4) a mean variance TS algorithm that makes no use of the contexts from Zhu and Tan (2020), namely MVTS; (5) a uniform sampling algorithm that randomly chooses an arm to pull at each round. Figure 1 shows that our proposed MVTS-D and MVTS-DN algorithms achieve better regrets compared to the three benchmarks in all cases, as we take into consideration the risk of the reward and learn the parameters over time making use of the contexts.

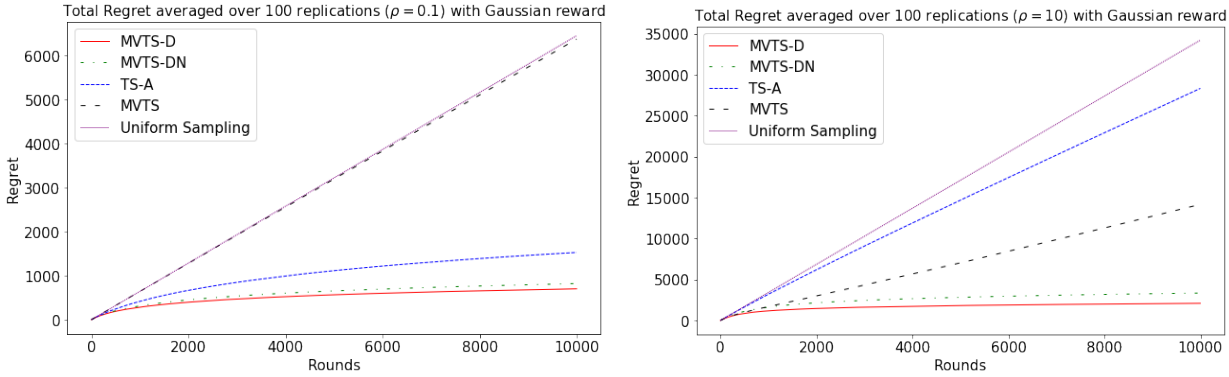

Figure 1: Total regrets comparison with different risk tolerances, averaged over 100 replications.