Sharpe Ratio-Optimized Thompson Sampling for Risk-Aware Online Learning

Sabrina Khurshid
Indian Institue of Technology Delhi
New Delhi, India
eez218683@iitd.ac.in

Mohammad Taha Shah
Indian Institue of Technology Delhi
New Delhi, India
tahashah@dbst.iitd.ac.in

Gourab Ghatak
Indian Institue of Technology Delhi
New Delhi, India
gghatak@ee.iitd.ac.in

Abstract

We investigate the problem of sequential decision-making for Sharpe ratio (SR) maximization in a stochastic bandit setting. We focus on the Thompson Sampling (TS) algorithm, a Bayesian approach celebrated for its empirical performance and exploration efficiency, under the assumption of Gaussian rewards with unknown parameters. Unlike conventional bandit objectives focusing on maximizing cumulative reward, Sharpe ratio (SR) optimization instead introduces an inherent tradeoff between achieving high returns and controlling risk, demanding careful exploration of both mean and variance. Our theoretical contribution is a novel regret decomposition specifically designed for the SR, highlighting the role of information acquisition about the reward distribution in driving learning efficiency. Empirical simulations show that our algorithm significantly outperform existing algorithms.

1 Introduction

Reinforcement Learning (RL) algorithms have traditionally focused on maximizing cumulative expected rewards. While this objective is theoretically sound, it often overlooks crucial aspects of risk, particularly the variability in returns. In real-world applications such as finance, healthcare, and robotics, where decision-making under uncertainty is critical, practitioners seek algorithms that not only aim for high returns but also manage the risk associated with those returns.

The SR, defined as the ratio of expected return and the standard deviation of returns i.e $\frac{\mu}{\sigma}$, offers a classical framework for evaluating risk-adjusted performance. In contrast, Thompson sampling (TS), a Bayesian algorithm that selects actions based on posterior samples of expected rewards, is widely known for its empirical efficiency and theoretical guarantees. However, standard TS optimizes for expected returns and does not directly incorporate risk into its decision-making process.

In this paper, we propose a novel algorithm called *Sharpe Ratio-Thompson Sampling* `SR-TS`. Our approach extends the TS framework by considering both the mean and variance of the reward distributions. SR-TS samples from the joint posterior of mean and variance, computes SRs for each arm, and selects the arm with the highest sampled SR. `SR-TS` offers a simple yet principled way to infuse risk-awareness into a online learning algorithm, with applications in domains where controlling variability is as important as achieving high rewards.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

2 Background and Motivation

TS is a Bayesian algorithm for online decision-making in multi-armed bandit (MAB) problems. At each time step, it samples a mean reward from the posterior distribution for each arm and selects the arm with the highest sample. This strategy balances exploration and exploitation naturally and achieves logarithmic regret in many settings. Despite its strengths, standard TS does not account for the variability in the rewards. In high-stakes environments, variability can result in undesirable outcomes even if the mean reward is high. For instance, in finance, a portfolio with a high expected return but large volatility might not be preferable to one with a lower expected return but much smaller variance. This motivates the need for integrating risk-awareness into the action-selection strategy.

2.1 Literature

Among Bayesian approaches, TS Thompson [1933] has emerged as a particularly effective algorithm, combining strong empirical performance with theoretical guarantees. Its regret bounds are often expressed in terms of expected information gain about the reward distributions Kaufmann et al. [2012], Honda and Takemura [2014]. It has been extensively studied in Korda et al. [2013], Agrawal and Goyal [2017, 2012], Russo and Van Roy [2014], Dong and Van Roy [2018].

Deriving online algorithms for optimizing the SR is particularly challenging since even offline policies experience constant regret with respect to the best expert Even-Dar et al. [2006]. Thus, the theoretical foundations for SR optimization within the bandit framework remains limited to Khurshid et al. [2025] and Cassel et al. [2018], wherein they propose UCB-based algorithms. Some recent studies have applied TS heuristically to SR-based strategies Zhu et al. [2019], Morariu, Shen and Wang [2016], but they lack formal regret analysis. Moreover, unlike traditional regret, where information gain about the mean suffices, minimizing regret with respect to the SR inherently requires learning the full distributional structure of each arm. This necessitates a novel regret decomposition that explicitly accounts for uncertainty to both mean and variance parameters.

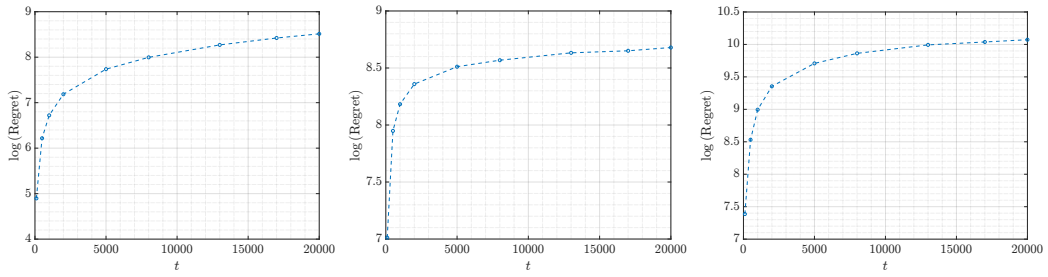
Existing literature on risk-aware bandits has explored mean-variance trade-offs and conditional value-at-risk (CVaR) approaches. However, few methods directly optimize the SR. Moreover, most such algorithms rely on upper-confidence-bound (UCB) techniques. Our approach aims to fill this gap.

3 Algorithm Description

We propose the SR-TS algorithm, which integrates variance information into TS to optimize for SR. Each arm i is modeled using a Normal-Inverse-Gamma prior, which jointly estimates the mean μ_i and variance σ_i^2 of the reward distribution. When arm i is pulled and reward r_i is observed, we update the posterior parameters accordingly. The key distinction from standard TS is the use of both sampled mean and sampled standard deviation to compute a risk-adjusted performance measure. This method is simple to implement and retains the computational efficiency of standard TS. It can be readily extended to non-Gaussian settings using more robust posteriors.

Algorithm 1 Sharpe Ratio Thompson Sampling - SRTS

Initialization: $\hat{\mu}_{i,0} = 0, \alpha_{i,0} = \frac{1}{2}, \beta_{i,0} = \frac{1}{2}, s_{i,0} = 0$
for each $t = 1, 2, ..., K$ **do**
  Play arm t and update $\hat{\mu}_{t,t} = X_{t,t}$
  Update $(\hat{\mu}_{t,t-1}, \alpha_{t,t-1}, \beta_{t,t-1}, s_{t,t-1})$
end for
for each $t = K+1, K+2, ..., n$ **do**
  Sample $\tau_{i,t}$ from Gamma($\alpha_{i,t-1}, \beta_{i,t-1}$).
  Sample $\theta_{i,t}$ from $\mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{s_{i,t-1}})$.
  Play arm $i(t) = \arg\max_{i \in [K]} \frac{\theta_{i,t}}{L_0 + \frac{\rho}{\tau_{i,t}}}$ and observe reward $X_{i(t),t}$
  Update $(\hat{\mu}_{i(t),t-1}, \alpha_{i(t),t-1}, \beta_{i(t),t-1}, s_{i(t),t-1})$
end for



(a) (b) (c)

Figure 1: (a) Regret for $\rho = 1$, (b) Regret for $\rho = 1$ and $\mu_i = 1$, $i \in \{1, 2, \ldots, K\}$, (c) Regret for $\rho = 0$

4 Theoretical Analysis

Definition 1. *The* Sharpe ratio *of an arm i with mean μ_i, variance σ_i^2, and risk tolerance ρ is*

$$\xi_i = \frac{\mu_i}{L_0 + \rho\sigma_i^2},$$

where L_0 is the regularization term and is same for all the arms.

Definition 2. *The expected regret of a policy π over n rounds is defined as*

$$\mathbb{E}[\mathcal{R}_n(\pi)] = n\left(\xi_1 - \mathbb{E}[\bar{\xi}_n(\pi)]\right).$$

where $\bar{\xi}_n(\pi)$ is the empirical SR under π. We derive a tractable decomposition:

- Each sub-optimal arm i contributes proportionally to $\mathbb{E}[s_{i,n}]$.
- The coefficient of $\mathbb{E}[s_{i,n}]$ is a Sharpe gap Δ_i plus correction terms capturing variance heterogeneity.

where $s_{i,n}$ is the number of times arm i is selected by a policy π up to round n.

Theorem 1. *The expected regret of a policy π over n rounds for SR is given as,*

$$\mathbb{E}\left[\mathcal{R}_n(\pi)\right] \leq \sum_{i=1}^{K} \mathbb{E}[s_{i,n}] \left(\Delta_i + \frac{\xi_i \rho \left(\frac{\Lambda_{\max}^2}{2} + (\sum_{i=1}^{K} \sigma_i^2) - \sigma_i^2 \right)}{L_0 + \frac{\rho}{2}\Lambda_{\max}^2 + \rho \sum_{i=1}^{K} \sigma_i^2} \right) + A_7$$

where $\Lambda_{\max} = \max(\mu_i - \mu_j)$, and A_7 is a constant independent of $s_{i,n}$ and n.

5 Experimental Evaluation

We validate `SR-TS` through simulations. We generate $K = 10$ arms with $\mu = \big(0.10, 0.27, 0.34, 0.41, 0.43, 0.55, 0.56, 0.67, 0.71, 0.79\big)$ and $\sigma^2 = \big(0.05, 0.09, 0.19, 0.14, 0.44, 0.24, 0.36, 0.56, 0.49, 0.85\big)$ taken from Sani et al. [2012]. In Fig. 1, we plot the expected regret, which is averaged over 500 realizations. The standard deviations of the regrets are small compared to the averages, and thus are omitted from the all plots. In Fig. 1a we plot the expected regret of `SRTS` w.r.t to time for $\rho = 1$. Fig. 1b plots regret for variance minimization as we fix the mean of all arms to 1, likewise Fig. 1c plots regret for mean maximization as we take $\rho = 0$. In Fig. 1b, we see that as ρ increases regret first increases slightly and then decreases. We then compare `SR-TS` with risk-averse UCB variant of SR such as `U-UCB` Cassel et al. [2018] and `UCB-RSSR` Khurshid et al. [2025]. Fig. 2 show that `SR-TS` achieves superior performance in comparison to its baselines.

6 Discussion and Conclusion

`SR-TS` introduces a risk-aware perspective to the widely adopted TS algorithm. By optimizing for SR rather than raw expected return, `SR-TS` addresses the needs of applications where variability

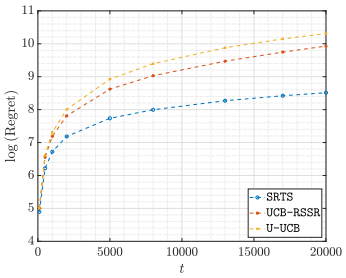


Figure 2: SRTS v/s U-UCB v/s UCB-RSSR for $\rho = 1$ and $L_0 = 1$.

and downside risk must be controlled. Our approach offers a practical solution for experimentalists and invites theoretical investigation from the learning theory community. We provide a simple, implementable algorithm and a foundation for regret analysis in terms of risk-adjusted performance. Future work includes extending SR-TS to non-stationary settings.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Conference on learning theory*, pages 1295–1306. PMLR, 2018.

Shi Dong and Benjamin Van Roy. An information-theoretic analysis for thompson sampling with many actions. *Advances in Neural Information Processing Systems*, 31, 2018.

Eyal Even-Dar, Michael Kearns, and Jennifer Wortman. Risk-sensitive online learning. In *Algorithmic Learning Theory: 17th International Conference, ALT 2006, Barcelona, Spain, October 7-10, 2006. Proceedings 17*, pages 199–213. Springer, 2006.

Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors. In *Artificial Intelligence and Statistics*, pages 375–383. PMLR, 2014.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR, 2012.

Sabrina Khurshid, Mohammed Shahid Abdulla, and Gourab Ghatak. Optimizing sharpe ratio: risk-adjusted decision-making in multi-armed bandits. *Machine Learning*, 114(2):32, 2025.

Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26, 2013.

Alin Morariu. *Financial Bandits-Development of Thompson Sampling for Financial Data*. PhD thesis, Toronto Metropolitan University.

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

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. *Advances in neural information processing systems*, 25, 2012.

Weiwei Shen and Jun Wang. Portfolio blending via thompson sampling. In *IJCAI*, pages 1983–1989, 2016.

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

Mengying Zhu, Xiaolin Zheng, Yan Wang, Yuyuan Li, and Qianqiao Liang. Adaptive portfolio by solving multi-armed bandit via thompson sampling. *arXiv preprint arXiv:1911.05309*, 2019.