Trend-aware Thompson sampling algorithm for solving multi-armed bandit problems in e-commerce environment

Ke Xu*

School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi Province, 710129, China

Abstract. Multi-armed bandit algorithms (MAB) have been widely used to make sequential decisions in e-commerce environment such as recommendation systems, online advertising and dynamic pricing strategy adjustments. Among these traditional algorithms, Thompson Sampling (TS) has a superior performance. However, in non-stationary MAB problems, traditional TS algorithm lacks the ability to perceive recent performance changes and differentiate the timeliness of information acquisition. Aiming at compensating these shortcomings, the author proposes an optimized version of traditional TS algorithm: Trend-aware TS algorithm. Through dynamically adjusting variance with 3 mechanisms: trend tracking, ranking adaptation and temporal decay, it avoids premature abandonment of potentially high-quality options which provide stable selection for arms with good recent performance while giving more exploration opportunities to poorly performing arms. In order to verify the effectiveness of this method, the author set up the experiment on the E-Commerce Sales Dataset, comparing the cumulative regret, moving average regret per round and optimal arm selection rate with traditional TS. The results demonstrated that the overall performance of the Trend-aware TS algorithm improves 18.85%, but the improvement declined as the number of rounds increased.

1 Introduction

In e-commerce environments, effective sequential decision-making strategies are crucial for business operations and development. These decision-making challenges can be effectively modeled as multi-armed bandit (MAB) problems, where each action (arm) represents a different choice and the goal is to maximize cumulative rewards while balancing exploration and exploitation trade-offs. MAB algorithms have been widely applied in e-commerce scenarios including recommendation systems [1], online advertising and dynamic pricing strategy adjustments [2, 3].

Thompson Sampling (TS) has demonstrated superior performance compared to traditional algorithms such as ε-greedy and UCB [4, 5]. However, Traditional TS algorithm struggles to solve the problems in non-stationary MAB problem which relies solely on cumulative statistics for decision-making and unable to perceive recent performance changes.

* Corresponding author: kexu@mail.nwpu.edu.cn

Based on these challenges in non-stationary MAB problem, the author proposes an optimization of TS algorithm: Trend-aware TS algorithm. It dynamically adjusts sampling variance to provide stable selection for arms with good recent performance while giving more exploration opportunities to poorly performing arms, avoiding premature abandonment of potentially high-quality choices.

The remainder of this paper is organized as follows: Section 2 introduces the data set and preprocessing; Section 3 reviews traditional TS and its limitations; Section 4 presents the Trend-aware algorithm; Section 5 provides experimental evaluation; and Section 6 concludes the paper.

2 Data set and data preprocessing

2.1 Data set

The data set used in this experiment is the E-Commerce Sales Dataset, which was originally published by ANi on the data world platform. For this study, the author used the version of this data set which was reprinted by the user "Thedevastator" on the Kaggle platform. This data set provides e-commerce sales data from various product categories and through different channels, containing 128,975 products with 22 relevant attributes such as product categories and whether a product was cancelled. It offers sufficient data support for analyses and research related to e-commerce sales.

This data set can be directly accessed via the official Kaggle link *https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data/data.*

2.2 Data preprocessing

To adapt this data set to the context of the MAB problem, the author appropriately processes the data from the E-Commerce Sales Dataset, retaining only two list of data: category of the product and whether the product was canceled. Specifically, the 9 product categories were encoded as integers from 0 to 8, which are treated as the arms in the MAB framework. Meanwhile, whether the product was canceled was encoded as binary values, where 0 indicates that the product was canceled and 1 indicates that the product was not canceled, which is regarded as the reward for each arm pull. After such preprocessing, the data set can be successfully applied to the scenario of the MAB. All experimental results presented in this study are entirely based on this data set.

3 Traditional Thompson sampling algorithm and its limitations

TS was first proposed by Thompson in 1933 which uses Bayesian learning to make a solution to the MAB problem. In each decision round, the algorithm maintains a posterior reward distribution for each arm, samples a reward value from each arm's current posterior distribution, selects the arm with the highest sampled value, and subsequently updates the posterior distribution of the chosen arm. Through this process, the posterior distributions approximate the true reward distribution, naturally achieving a balance between exploration and exploitation via probabilistic sampling.

Numerous studies have demonstrated that the overall performance of the TS algorithm is significantly superior to that of other traditional MAB algorithms. Research conducted by Singh (2021) demonstrated that UCB and TS algorithms significantly outperform the Epsilon-Greedy algorithm in terms of performance, with the former two algorithms not only

converging faster but also achieving the highest rewards more efficiently [4]. Research conducted by Z Wang, C Zhang, and K Chaudhuri (2022) demonstrated that TS algorithms show superior empirical performance in comparison with UCB algorithms and outperform baseline algorithms that do not employ transfer learning [5].

However, traditional TS algorithms still remain limitations. In non-stationary MAB problems, traditional TS algorithms have been applied to such problems, but there is a lack of experimental papers with theoretical analysis, and TS using passive methods has limited theoretical analysis regarding regret [6]. Moreover, there is insufficient research on the optimal arm selection rate of TS algorithms currently. Finally, the decision-making process of traditional TS algorithms relies entirely on cumulative statistical data, thus lacking the capability to perceive recent performance fluctuations. Therefore, this experiment aims to comprehensively improve the performance of traditional TS algorithms from three aspects: cumulative regret and the variance among different experiments, moving average regret per round, and optimal arm selection rate. To this end, based on existing research findings, the author proposes a feasible optimized version of the traditional TS algorithm: the Trend-aware TS algorithm.

4 Trend-aware Thompson sampling algorithm

4.1 Trend-aware mechanism

The core idea of Trend-aware mechanism lies in dynamically adjusting variance to provide stable selection for arms with good recent performance while giving more exploration opportunities to the arms with poor recent performance, avoiding premature abandonment of potentially high-quality options [7]. This compensates for the shortcoming of traditional TS algorithms in being unable to perceive recent performance changes and unable to differentiate the timeliness of information acquisition in non-stationary environments. The variance adjustment formula is as follows:

$$\sigma_{adj}^2 = \sigma_{base}^2 \cdot f_{trend} \cdot r_{adj} \cdot (d_{exp} \cdot 0.7 + 0.3), \quad (1)$$

where σ_{adj}^2 is the adjusted variance, σ_{base}^2 is the base variance, f_{trend} is the trend factor, r_{adj} is the rank adjustment factor, and d_{exp} is the exploration decay factor. It can be seen that the adjusted variance is influenced by three factors: recent performance, ranking, and exploration time. The specific implementation mechanisms will be introduced one by one in the following sections.

4.2 Trend tracking

The specific calculation formula for the trend factor f_{trend} is as follows:

$$f_{trend} = 1.0 + 0.5 \times (\bar{\mu}_{overall} - \bar{\mu}_{recent}), \quad (2)$$

During each decision-making round, the recent average reward $\bar{\mu}_{recent}$ for each arm is recorded through sliding window and compared against the overall average reward $\bar{\mu}_{overall}$ [6]. This mechanism detects a change when the empirical mean deviates from the mean estimate [8]. When $\bar{\mu}_{recent} > \bar{\mu}_{overall}$, this implies that the arm has a superior performance recently, resulting in $f_{trend} < 1$, which consequently reduces the variance and enhances the exploitation of that particular arm. Otherwise, when $\bar{\mu}_{recent} < \bar{\mu}_{overall}$, this indicates poor recent performance, resulting in $f_{trend} > 1$, thereby increasing the variance and providing more exploration opportunities to the arm which performs poorly.

4.3 Ranking adaptation

Ranking and selection has been actively studied in simulation such as top-two TS [9]. Based on the existing research findings, the author proposes a new method of Ranking and selection. Before each decision round, arms are ranked according to their historical average rewards in descending order. Higher-ranked arms essentially maintain their variance, while lower-ranked arms gradually increase their variance, thereby preventing the premature selection of sub-optimal arms. The specific calculation formula for the rank adjustment factor r_{adj} is as follows:

$$r_{adj} = 1.0 + 0.5 \times \left(1.0 - exp\left(-\frac{rank(\hat{\mu}_i)}{K}\right)\right), \quad (3)$$

where $rank(\hat{\mu}_i)$ is the rank of arm i determined by its estimated mean reward $\hat{\mu}_i$, and K represents the total number of arms in the MAB setting.

4.4 Temporal decay

The specific calculation formula for the exploration decay factor d_{exp} is as follows:

$$d_{exp} = exp\left(-1.5 \cdot \frac{t-K}{T-K}\right), \quad (4)$$

where t is the current time step, K is the number of arms, T is the total number of rounds.

Drawing inspiration from the time decay strategies for ε -greedy algorithms discussed by R. Sutton and A. Barto (2018) [10], the author aims to optimize the performance of traditional TS algorithms through temporal decay strategies. As the number of exploration rounds increases, the exploration decay factor gradually decreases, resulting in a gradual reduction in variance. This mechanism provides more exploration opportunities in the early stages of exploration to enable faster discovery of the optimal arm, while reducing exploration opportunities in the later stages to focus on known high-quality choices. As the number of exploration rounds increases, the exploration decay factor gradually decreases, resulting in a gradual reduction in variance. This mechanism provides more exploration opportunities in the early stages of exploration to enable faster discovery of the optimal arm, while reducing exploration opportunities in the later stages to focus on known high-quality choices.

5 Experimental evaluation

5.1 Experimental setup

The experiment is divided into two groups, with the total number of rounds T set to 5,000 and 50,000 respectively. Each experimental group conducts simulations of both the traditional TS algorithm and the Trend-aware TS algorithm on the preprocessed E-Commerce Sales data set. Each simulation is repeated 20 times. Each group plots the images of the cumulative regret as functions of the number of rounds along with error bars, as well as the images of the moving average regret per round and optimal arm selection rate as functions of the number of rounds. Finally, the percentage of the improvement of the trend-aware TS algorithm compare with the traditional TS algorithm from these three aspects is calculated for each experimental group.

5.2 Performance comparison and analysis

When T = 5,000, the experimental results are presented in the Fig.1, Fig.2 and Fig.3.

Fig. 1. Comparison of algorithm cumulative regret when T=5000

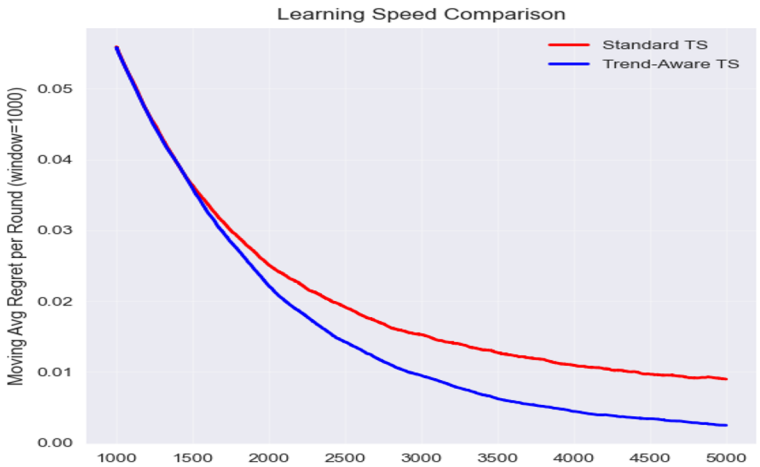

Fig. 2. Comparison of moving average regret per round when T = 5000

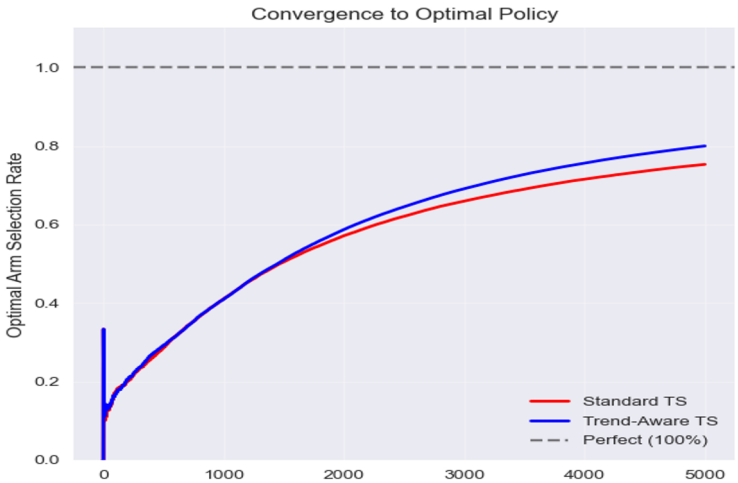

Fig. 3. Comparison of optimal arm selection rate when T = 5000

The figures demonstrate that when T = 5,000, the trend-aware TS algorithm has a superior performance compared to the traditional TS algorithm in total. Specifically, the curves of the cumulative regret and moving average regret per round of the trend-aware algorithm are consistently lower than those of the traditional algorithm, while its optimal arm selection rate is consistently higher. Through calculation, the trend-aware TS algorithm improved 18.85% on overall performance.

When T = 5,0000, the experimental results are presented in the Fig.4, Fig.5 and Fig.6.

Fig. 4. Comparison of algorithm cumulative regret when T = 50000

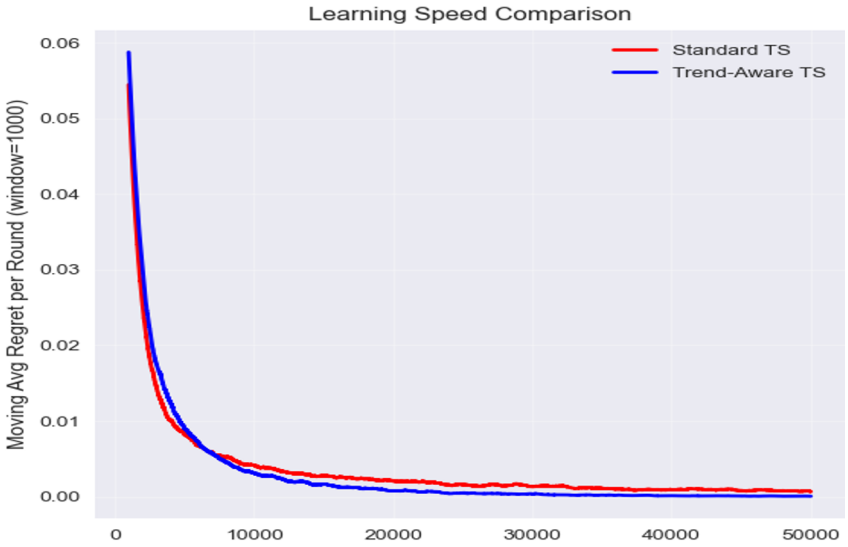

Fig. 5. Comparison of moving average regret per round when T = 50000

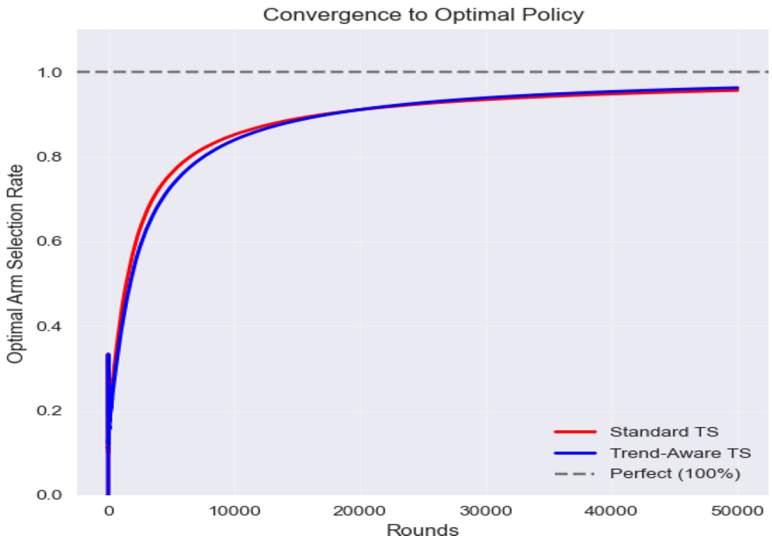

Fig. 6. Comparison of optimal arm selection rate when T = 5000

The figures demonstrate that when T = 50,000, the overall performance improvement effect of the trend-aware TS algorithm decreases. As illustrated in Fig. 4, during the early exploration phase, the cumulative regret of the trend-aware TS algorithm is higher than that of the traditional TS algorithm. Only when T exceeds 20,000 does its cumulative regret gradually become lower than that of the traditional TS algorithm. Furthermore, as illustrated in Fig. 5 and Fig. 6, the curves of moving average regret per round and optimal arm selection rate of the trend-aware TS algorithm nearly coincide with those of the traditional TS algorithm. Through calculation, the overall performance improvement of the trend-aware TS algorithm declined to 13.25%.

6 Conclusion

Given that traditional TS algorithms lack the capability to perceive recent performance in e-commerce scenarios, this study proposed an optimized version — the Trend-aware TS algorithm. The core idea of this algorithm lies in dynamically adjusting variance using three important mechanisms: trend tracking, ranking adaptation, and temporal decay. Through this mechanism, the algorithm can adjust the variance in real-time based on current performance, thereby achieving an improved balance between exploring new options and exploiting known high-quality choices, while preventing the premature abandonment of potentially high-quality arms.

The experiments were conducted on the E-Commerce Sales Dataset to assess the performance of the trend-aware TS algorithm. The results demonstrated a significant effect in reducing cumulative regret and moving average regret per round, as well as improving optimal arm selection rate, thereby proving the effectiveness of the Trend-aware TS algorithm and demonstrating its practical application value in e-commerce environments, such as recommendation systems, online advertising and dynamic pricing strategy adjustments.

However, this study exists limitations. The method lacks sufficient datasets to prove its effectiveness deeply, and as the total number of rounds increases, the performance improvement decreases. Future research will focus on proving the effectiveness of this

method across more e-commerce datasets and exploring the integration of multi-dimensional signals to improve the decision-making strategies in the non-stationary environment.

References

1. E.N. Mambou, I. Woungang, Bandit algorithms applied in online advertisement to evaluate click-through rates, in Proceedings of the 2023 IEEE AFRICON, Nairobi, Kenya 1-5(2023).
2. V. Avadhanula, R. Colini Baldeschi, S. Leonardi, K.A. Sankararaman, O. Schrijvers, Stochastic bandits for multi-platform budget optimization in online advertising, in Proceedings of the Web Conference 2021, 2805-2817(2021).
3. K.J. Ferreira, D. Simchi-Levi, H. Wang, Online network revenue management using Thompson sampling. Oper. Res. **66**, 1586-1602 (2018).
4. A. Singh, Reinforcement learning based empirical comparison of UCB, epsilon-greedy, and Thompson sampling. Int. J. Aquat. Sci. **12**, 2961-2969 (2021).
5. Z. Wang, C. Zhang, K. Chaudhuri, Thompson sampling for robust transfer in multi-task bandits. arXiv preprint arXiv:2206.08556 (2022).
6. H. Qi, F. Guo, L. Zhu, Thompson sampling for non-stationary bandit problems. Entropy **27**, 51 (2025).
7. A. Saha, B. Kveton, Only pay for what is uncertain: variance-adaptive Thompson sampling. arXiv preprint arXiv:2303.09033 (2023).
8. G. Ghatak, A change-detection-based Thompson sampling framework for non-stationary bandits. IEEE Trans. Comput. **70**, 1670-1676 (2020).
9. Y. Peng, G. Zhang, Thompson sampling meets ranking and selection, in Proceedings of the 2022 Winter Simulation Conference, Singapore, 3075-3086 (2022).
10. R. Sutton, A. Barto, Reinforcement learning: an introduction (MIT Press, Cambridge, 2018).