

Modeling User Autonomy in Recommender Systems Using Markov Perturbation-based Multi-armed Bandits

Yuxuan Yang^{1,a,*}

¹*School of Statistic, East China Normal University, Shanghai Province, 200062, China
a. 10215300424@stu.ecnu.edu.cn*

**corresponding author*

Abstract: Recommender systems play a critical role in e-commerce, streaming media, and other fields, where user engagement is vital to platform success. These systems, however, face challenges like the cold start problem, which limits recommendation accuracy for new users and items due to sparse data. Although multi-armed bandit (MAB) algorithms are considered effective for balancing exploration and exploitation, existing studies lack simulations of algorithm performance in scenarios where recommender systems indirectly influence user choices, limiting insights for performance improvements in real-world applications. This study explores the performance of key MAB algorithms—Explore-Then-Commit (ETC), Upper Confidence Bound (UCB), and Thompson Sampling (TS)—under a novel framework that models user behavior as an unreliable intermediate using Markov state transitions. This Markov perturbation-based approach provides a foundation for indirectly influencing user decisions. Two algorithmic frameworks are examined: one considering only the recommendation system's influence on user choices, and another also considering user feedback as a dynamic influence on state transition probabilities. Results show that, while considering user feedback increases adaptability, it also causes greater volatility. TS achieves the lowest cumulative regret with strong stability, while Asymptotic-UCB achieves a better balance between exploration and exploitation than standard UCB, with consistent adaptability under varying feedback. In contrast, ETC demonstrates significant volatility, highlighting its limitations in feedback-driven scenarios. These findings provide valuable insights for enhancing MAB algorithms in real-world cold start recommendation applications, especially where systems influence user choices indirectly.

Keywords: multi-armed bandit, cold start problem, recommender systems, unreliable intermediate, Markov perturbation

1. Introduction

Recommender systems are information filtering systems designed to predict user ratings and preferences for items, providing personalized recommendations. They play a significant role in sectors such as e-commerce, streaming media, and news portals, greatly enhancing user experience and platform profitability. Existing recommender system methods are mainly divided into three categories: Collaborative Filtering, Content-based Filtering, and hybrid approaches. Content-based Filtering makes recommendations based on user profiles or product descriptions, and CF-based methods recommend based on user behavioral history or similarities between users and items. Hybrid

methods combine the advantages of both to achieve better recommendations. Despite their success, recommender systems still face challenges such as cold start, data sparsity, scalability, diversity, and interpretability, which directly impact system performance and user experience [1].

The cold start problem is a core challenge in recommender systems, especially when there is insufficient historical data for new users or items. The cold start problem significantly impacts the ability of the system to provide personalized recommendations, reducing user experience and overall platform efficiency. Schein et al highlighted the profound impact of the cold start problem on recommender systems, particularly in e-commerce and social media platforms, where it decreases recommendation accuracy, affecting user acceptance and retention rates [2]. Raza et al. further analyzed the cold start problem, emphasizing the adaptability required when introducing new users and items, posing higher demands on the real-time accuracy of recommender systems. The challenges posed by data scarcity and the need for quick user adaptation increase the difficulty in maintaining recommendation precision and timeliness [3].

To address this issue, multi-armed bandit (MAB) algorithms have been introduced to recommender systems to balance exploration and exploitation, especially in cold start scenarios [4]. Fundamental MAB algorithms include Explore-Then-Commit (ETC), Upper Confidence Bound (UCB) and Thompson Sampling (TS) [5-7]. The ETC algorithm first explores each option sufficiently and then selects the best-performing option in the exploitation phase. The UCB algorithm dynamically balances exploration and exploitation by calculating the upper confidence bound of each option and selecting the one with the highest bound. The TS algorithm performs probabilistic sampling based on current beliefs about each option. Any recommender system algorithm, including MABs, cannot directly determine user choices but can only influence decisions through some intermediate intervention.

Ravi et al. points out the challenge faced by recommender systems when trying to maximize rewards, given the presence of the user as an 'unreliable intermediary.' For example, in streaming platforms, users have autonomy in selecting content to watch, and the platform can only influence these choices through targeted ads. In this context, the platform's recommender system resembles a bandit attempting to maximize rewards, while the user acts as an intermediary with autonomy [8]. Current research on MAB algorithms lacks consideration of the practical situation where recommender systems can only indirectly influence user choices.

To simulate real recommender systems, this study uses Markov chains to model user choices as an uncertain intermediary. User states transition through a Markov chain, with the algorithm influencing choices by perturbing the state transition matrix. This reflects autonomous user choices and algorithmic influence. This paper examines multi-armed bandit algorithms under two frameworks: one considering only recommender influence, and the other integrating user feedback. Section 2 covers dataset and design, Section 3 presents results, and Section 4 discusses limitations and future improvements.

2. Research design

2.1. Dataset

The MovieLens 1M dataset, created by the GroupLens research team at the University of Minnesota, USA, is widely used for research in recommender systems and social computing. It contains 1 million ratings from 6,000 users for 4,000 movies, with ratings given as integers from 1 to 5 to reflect user preferences. The dataset also includes detailed user information such as gender, age, and occupation, as well as specific movie attributes like release year, director, and genre. This extensive movie rating data makes it a valuable resource for studying personalized recommendation algorithms and is commonly used for simulating the performance of recommender system algorithms. After merging

the movie data with the rating data, the format of the movie genre-rating pairs was extracted. The ratings for each genre were then stored, with each genre treated as a 'bandit'.

2.2. Algorithm design

In research, referring to Ravi, the user is treated as an unreliable intermediate, and changes in their behavior can be modeled as a Markov chain with a transition matrix represented by $P = [p_{i,j}]_{|S| \times |S|}$. In the experiment dataset, considering the state represents the movie genre currently chosen by the user, P_{ij} is the probability of transitioning from state i to state j and the state space S represents both the movie genres and the set of arms in the MAB, with a cardinality $|S| = K$. The MAB algorithm perturbs the Markov transition matrix to make the system more inclined towards a low-regret state, without directly intervening in the determination of the states themselves.

To simulate the cold start problem, assume the initial state transition matrix is uniform, meaning the transition probability between states is $1/K$, and the user's initial state is random, without considering any initial preference. This assumption is reasonable since that could assume users have equal transition probabilities among different states initially. In a recommendation system with sufficient user behavior data, this matrix can be predicted using user behavior data.

To simulate the perturbation of the MAB algorithm on the transition matrix, algorithm introduce a perturbation variable δ , which describes the influence of the recommendation system on user behavior, such as increasing the probability of selecting a particular movie genre through "targeted recommendation." Therefore, δ is used to perturb the transition probability of a specific state. After applying the perturbation, normalization is performed on the transition matrix to ensure that the sum of each row's probabilities equals 1.

Based on this foundation, two algorithmic frameworks will be established for simulating the basic MAB algorithm in subsequent experiments:

- Algorithm 1: Considering only the influence of the recommendation system on transition probabilities

This framework models the scenario where the recommendation system's effect is solely on perturbing the transition probability matrix without direct consideration of user feedback.

The transition matrix is updated by adding a perturbation δ to the probability of transitioning from the current state s to the chosen arm a . This ensures that the chosen action becomes more likely in the future.

For selecting the next arm, the algorithm uses the updated transition matrix to determine the probability of moving to each possible state. Based on these probabilities, the next actual arm r is chosen probabilistically, and the state is updated accordingly.

- Algorithm 2: Considering both the recommendation system's influence and user feedback on transition probabilities

This framework simulates the influence of both the recommendation system and user feedback, allowing the transition probability matrix to evolve based on user responses in addition to system recommendations.

The idea is to dynamically adjust the transition matrix based on the rewards received for the recommended content. When a user provides feedback (in the form of ratings), the probability of staying in or leaving the current state is adjusted accordingly. Higher ratings increase the probability of staying in the current state (indicating a positive experience), while lower ratings decrease it (indicating dissatisfaction). The changes are made linearly with respect to the reward value, using a parameter β to control the degree of change.

The update to the transition matrix happens after each recommendation, based on the user's feedback. The transition probability to remain in the current state s is increased or decreased depending on whether the reward is above or below a baseline (e.g., a rating of 3). This reflects the adaptive nature of the recommendation system in response to user preferences. The update for p_{ss} can be expressed as follows:

$$p_{ss} = f(\beta, p_{ss}, \text{reward}) = \max(0, \widetilde{p}_{ss}) \quad (1)$$

$$\widetilde{p}_{ss} = \begin{cases} p_{ss} - 2\beta & , \text{reward} = 1 \\ p_{ss} - \beta & , \text{reward} = 2 \\ p_{ss} + \beta & , \text{reward} = 4 \\ p_{ss} + 2\beta & , \text{reward} = 5 \end{cases} \quad (2)$$

The framework of Algorithm 1 and Algorithm 2 is applicable for simulating basic MAB algorithms.

- Evaluation criteria

The performance is evaluated based on cumulative regret $R(T)$, where μ^* is obtained by calculating the maximum average score across different categories, and r_t is the reward at time step t . The cumulative regret is used to measure the performance of the algorithms. Performance comparisons are made by observing the convergence and final value of the line $R(T)$ vs. T .

3. Results

3.1. Performance simulation and comparison of MAB algorithms under algorithm 1

For $\delta = 0.1$ and $\delta = 0.9$, the author conducted 50 experiments, each running for 20,000 rounds, for four algorithms—ETC, UCB, Asymptotic UCB (A-UCB), and TS, which the error bars represent one standard deviation above and below the mean. The results are as Figure 1:

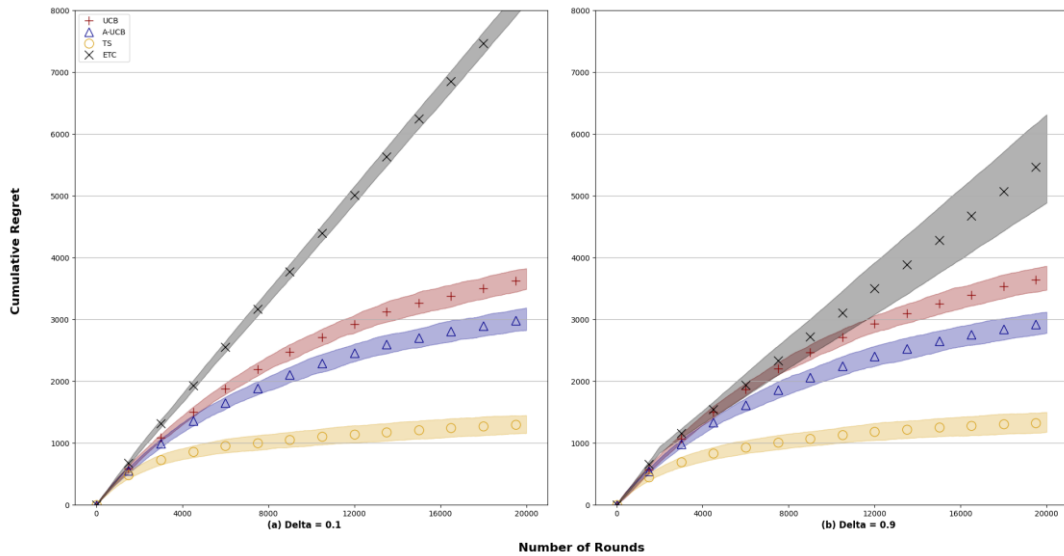


Figure 1: Cumulative regret comparison of MAB algorithms under algorithm 1 (Photo/Picture credit: Original).

3.2. Performance simulation and comparison of MAB algorithms under algorithm 2

In this section, the author introduced a new parameter $\gamma = \beta/\delta$, which represents the ratio of the influence of user feedback to the influence of the recommendation system. γ is set within the range

(0,2].The author evaluated the convergence behavior of the cumulative regret curves for each algorithm over the long term.

For each experiment, a δ value is randomly chosen within the range (0,0.5], and β is calculated as $\beta = \gamma\delta$. This approach ensures that β remains within the range (0,1], preserving a controlled yet flexible ratio between user feedback and system influence. The author used this randomized parameter combination to run simulations, examining the variations in average cumulative regret under different γ values.

For $\gamma = 0.1, 0.5$ and $\gamma = 1, 1.5$, the author conducted 50 experiments, each running for 20,000 rounds. The visualization approach follows the same format as in Section 3.1. The simulation results are as Figure 2:

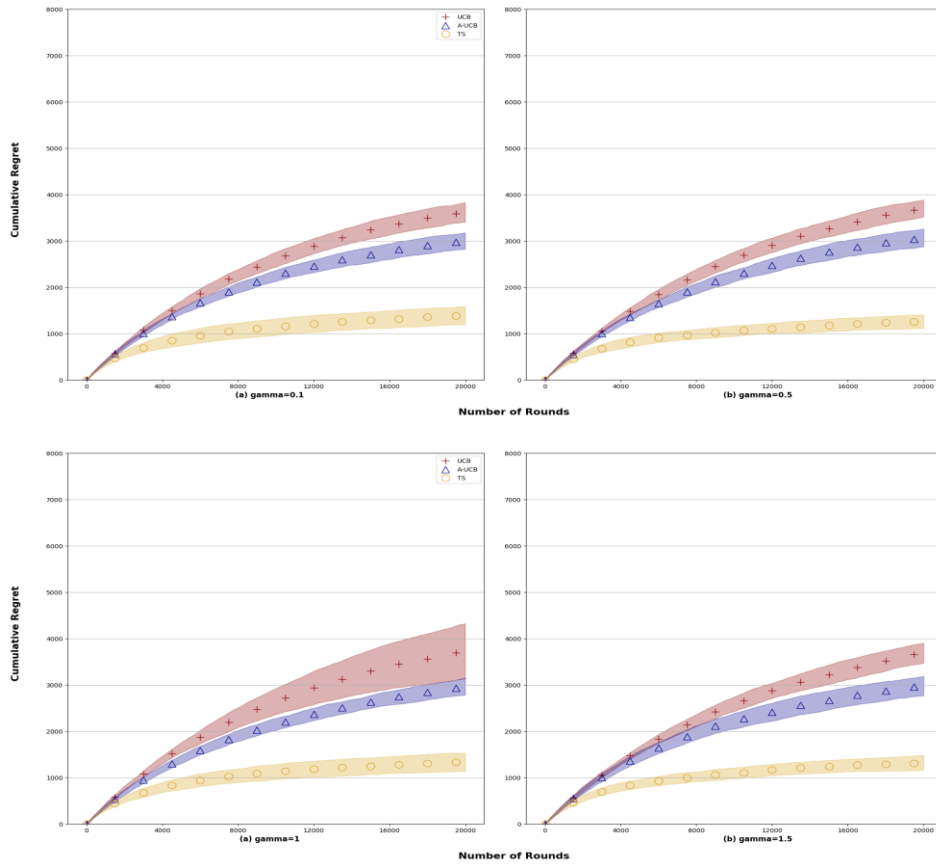


Figure 2: Cumulative regret comparison of MAB algorithms under algorithm 2 (Photo/Picture credit: Original).

4. Discussion

4.1. Analysis of results under algorithm framework 1

Under Algorithm Framework 1, which considers only the influence of the recommendation system's transition matrix, the parameter δ represents the strength of this influence, with larger δ indicating greater impact. The average cumulative regret across different δ values reveals distinct patterns among the algorithms. Figure 1 illustrates the long-term performance, showing that all algorithms improve as δ increases, with cumulative regret decreasing and convergence accelerating. Asymptotic UCB outperforms standard UCB due to its dynamic adjustment over time, while TS achieves the best and most stable performance, leveraging its Bayesian framework for effective

balance. In contrast, ETC performs poorly and becomes more unstable with higher δ , highlighting its rigidity and limited adaptability to dynamic conditions.

4.2. Analysis of results under algorithm framework 2

Figure 2 show cumulative regret trends for different γ values. As γ increases, cumulative regret decreases across all algorithms, indicating changes in performance and adaptability when user feedback is considered. TS consistently achieves the lowest regret with stable convergence and narrower error bands, reflecting its robustness in adapting to user influence. A-UCB also shows lower regret compared to standard UCB, benefiting from improved exploration-exploitation balance, though with some fluctuations. UCB exhibits higher regret and greater sensitivity to feedback variations, evident from wider error bands. ETC is not included due to its extreme volatility, as large fluctuations in cumulative regret indicate poor performance stability, making it unsuitable for scenarios with significant user feedback influence.

4.3. Limitations of the algorithm framework

One of the key limitations of the model lies in the oversimplified mechanism used to influence the transition matrix. Currently, the approach directly increases the transition probability to the algorithm-chosen state by δ from the current state, which does not adequately capture the effectiveness of the recommendation system. This simplistic adjustment does not fully reflect how recommendations should work in real-world scenarios. For instance, repeatedly recommending the same item without considering diminishing effects is unrealistic; in practice, repeated exposure should lead to diminishing returns. Moreover, the influence on unchosen states is not properly accounted for, potentially overlooking the subtler dynamics of user behavior.

Another limitation is the high recommendation frequency assumed in the simulations. In reality, recommendations are not made so frequently, and there should be considerations for the decay in effectiveness during intervals between recommendations, as well as changes in user exploration behavior during these gaps. The model could benefit from incorporating concepts from restless multi-armed bandit (RMAB) models, where user engagement may fluctuate between recommendations, reflecting a more realistic pattern of exploration and engagement.

Furthermore, the influence mechanism of user feedback is too simplistic, particularly when changes in probability drop directly to zero, which can be too absolute and rigid. In real situations, user feedback can decay or increase in response to various factors, indicating that both recommendation and user influence mechanisms should incorporate more complex, nonlinear functions to better capture these dynamics. Additionally, the model currently treats transition matrix changes as independent for each instance, without considering long-term effects. A more realistic approach would account for cumulative impacts over time, reflecting a persistent influence rather than just momentary adjustments.

5. Conclusion

This study investigates the performance of multi-armed bandit algorithms—specifically ETC, UCB, and TS—within recommender system frameworks that model user behavior as an intermediary, rather than a direct outcome of system influence. By introducing Markov chains to model user state transitions, this research captures the complexities faced by recommender systems in conditions such as cold start, where limited historical data restricts the system's influence. Key findings reveal distinct performance characteristics among algorithms. Under conditions where only the recommendation system's influence is considered, the ETC algorithm demonstrates rapid regret reduction as influence strength (δ) increases, although it remains constrained by its two-phase structure, limiting

adaptability to dynamic conditions. UCB and TS show gradual improvement with increased influence, with TS achieving the most stable and effective balance due to its probabilistic approach.

When user feedback is integrated, algorithms generally exhibit improved adaptability. TS achieves the lowest cumulative regret and maintains stable performance as feedback strength (γ) increases, while asymptotic UCB (A-UCB) outperforms standard UCB by better balancing exploration and exploitation under changing user influences. In contrast, ETC shows high sensitivity to feedback fluctuations, with substantial regret volatility, making it less suitable for scenarios involving significant user feedback.

The application of MAB algorithms in cold start and user-driven recommendation systems provides valuable insights for enhancing user engagement and retention in fields such as e-commerce, streaming, and social media, especially when systems influence user choices indirectly. Future research should focus on optimizing the algorithmic framework by incorporating more dynamic, non-linear feedback mechanisms and considering long-term user influence effects. Refining MAB algorithms based on these simulations could help bridge the gap between theoretical models and practical applications, making MAB approaches more effective for real-world cold start recommendation scenarios.

References

- [1] Hasan, E., Rahman, M., Ding, C., Huang, J. X., & Raza, S. (2024). *Based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives*. *arXiv preprint arXiv:2405.05562*.
- [2] Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2023). *Addressing Cold Start in Recommender Systems: A Comprehensive Review*. *ACM Transactions on Information Systems*, 41(3), 1-32
- [3] Raza, S., Rahman, M., Kamawal, S., Toroghi, A., Raval, A., Navah, F., & Kazemeini, A. (2024). *A Comprehensive Review of Recommender Systems: Transitioning from Theory to Practice*. *arXiv:2407.13699*.
- [4] Zhao, X., Xia, L., Tang, J., & Yin, D. (2019). *Deep reinforcement learning for search, recommendation, and online advertising: a survey*. *ACM SIGWEB Newsletter*, (Spring), 1–15.
- [5] Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). *Finite-time analysis of the multiarmed bandit problem*. *Machine Learning*, 47(2–3), 235–256.
- [6] Khamaru, K., & Zhang, C. H. (2024). *Inference with the Upper Confidence Bound Algorithm*. *arXiv preprint arXiv:2408.04595*.
- [7] Riou, C., & Honda, J. (2020). *Bandit algorithms based on thompson sampling for bounded reward distributions*. In *Algorithmic Learning Theory* (pp. 777-826). PMLR.
- [8] Ravi, A. N., Poduval, P., & Moharir, S. (2020). *Unreliable multi-armed bandits: A novel approach to recommendation systems*. In *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)* (pp. 650-653). IEEE.