A Comparative Study of UCB and Thompson Sampling with Structured Rewards: Parameter Sensitivity and Robustness

Yutong Chen*

Duke Kunshan University, Kunshan, China

Abstract. The behavior of multi-armed bandit (MAB) algorithms is closely tied to how their hyperparameters are set, but their stability in structured reward environments has not been examined in depth. To address this gap, we built a bandit framework from the Retail Rocket dataset that represents rewards at both the item and category levels. This setting reflects the correlations that often arise in recommendation tasks. We tested three algorithms: the standard Upper Confidence Bound (UCB), Thompson Sampling (TS), and a Structured UCB that combines individual and category-level statistics. The evaluation covered a horizon of 50,000 rounds with varied hyperparameter settings and repeated runs. Results show that the structured approach lowers cumulative regret and yields higher click-through and conversion rates. It also reduces the sensitivity of UCB and TS to exploration coefficients and prior choices, which are typically difficult to tune. Taken together, the study suggests that incorporating structural information can make bandit learning both more robust and more effective in practice, especially for large-scale recommendation systems.

1 Introduction

Online decision-making problems in recommendation and advertising systems often require balancing exploration with exploitation. Multi-armed bandit (MAB) models provide a formal framework for addressing this challenge and have been extensively investigated in the machine learning literature [1]. Among the established approaches, such as Upper Confidence Bound (UCB) algorithm [2], and Thompson Sampling (TS) [3], are considered benchmarks due to their theoretical guarantees and empirical performance [4].

The effectiveness of these algorithms, however, is closely tied to the choice of hyperparameters. The exploration coefficient in UCB and the prior distributions in TS can substantially influence both convergence speed and cumulative regret [5]. While the theoretical properties of UCB and TS are well understood, systematic examinations of their parameter sensitivity, particularly in environments with structured rewards, remain limited.

In practice, many recommendation tasks exhibit structural correlations: items belong to categories, and rewards are often not independent across arms. Exploiting such information has the potential to improve learning efficiency and robustness, especially under data sparsity

* Corresponding author: yutong.chen2@dukekunshan.edu.cn

and long-tail distributions [6]. Nevertheless, there has been little empirical evidence on how category-level structure affects parameter dependence in classical MAB algorithms.

This study investigates the parameter sensitivity of UCB and TS under a structured reward environment constructed from the Retail Rocket dataset. A category-aware extension of UCB is compared against its classical counterpart and TS across different horizons and parameter settings. The empirical evaluation highlights how structural information at the category level can mitigate hyperparameter dependence and improve stability in regret minimization.

2 Data Processing

2.1 Dataset Overview

The dataset originates from the public Retail Rocket logs. After preprocessing, three files are retained: bandit_click.csv, bandit_purchase.csv, and item_to_category.csv. The collection spans 185,030 items across 1,180 categories, with about 2.41 million click events and 21,976 purchases. Feedback is highly skewed: most items appear only a few times, while a small fraction accounts for most interactions. This pronounced long-tail motivates the incorporation of category-level structure into the bandit setup.

2.2 Data Cleaning and Filtering

Raw logs were cleaned by discarding invalid and duplicate entries and by enforcing a single category mapping for each item. Items with very sparse exposure were removed from the candidate pool to reduce noise but retained for descriptive statistics. Each item was thus aligned with a unique category identifier at both detailed and coarse levels for subsequent analysis.

2.3 Construction of Bandit Environment

The cleaned data were then transformed into an offline bandit environment. A single round corresponds to one recommendation opportunity within a user session. At each round, an algorithm selects an item from a candidate set of predefined size, which is drawn from both category-level neighbors and globally popular items. The reward function is defined separately for clicks and purchases: a click yields $r_t^{(click)} = 1$, and a purchase yields $r_t^{(purchase)} = w_{buy}$, where w_{buy} reflects the higher value of conversions. When both signals are considered, a weighted combination is used,

$$r_t = \alpha \cdot r_t^{(click)} + (1 - \alpha) \cdot r_t^{(purchase)}, \quad \alpha \in [0,1]. \tag{1}$$

Two structured output files were produced: bandit_click.csv containing the click rewards and bandit_purchase.csv containing the purchase rewards.

2.4 Statistical Characteristics

The reconstructed environment shows clear long-tail characteristics. Clicks are logged as positive events, giving a CTR of 100%, while purchases yield a conversion rate of about 0.9%. Most items appear only a few times, whereas a small fraction dominates interactions. As illustrated in Fig. 1, this imbalance underscores the difficulty of learning from sparse feedback and the value of category-level information for stabilizing bandit algorithms.

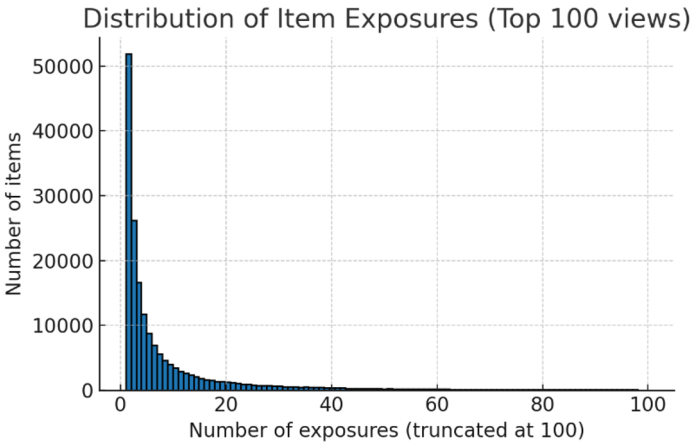

Fig. 1. Distribution of item exposures (truncated at 100 views).

3 Algorithms

3.1 Problem Formulation

The multi-armed bandit (MAB) problem is defined as a sequential decision-making process in which, at each round $t = 1,2, \dots, T$, an agent selects an action a_t from a finite set of arms $\mathcal{A}$[1]. After choosing an arm, the agent receives a stochastic reward $r_t \in [0,1]$ drawn from an unknown distribution associated with that arm. The objective is to maximize the cumulative reward or, equivalently, to minimize the cumulative regret:

$$R_T = \sum_{t=1}^{T} (\mu^* - \mu_{a_t}), \tag{2}$$

where μ^* is the expected reward of the optimal arm and μ_{a_t} is the expected reward of the chosen arm at time t.

In this study, the bandit environment is constructed from the Retail Rocket dataset described in Section 2. Each arm corresponds to a unique item, and its reward is derived from observed user interactions. Two types of reward signals are considered: clicks and purchases, with purchases assigned a higher weight to reflect their greater value. Furthermore, items are grouped by category, providing structural information that can be leveraged in the design of algorithms. The central focus is to evaluate how different bandit strategies behave under this structured setting and to assess their sensitivity to key hyperparameters.

3.2 Baseline Algorithms

3.2.1 Upper Confidence Bound (UCB)

The Upper Confidence Bound (UCB) algorithm is one of the most widely studied strategies for the stochastic multi-armed bandit problem [2]. The core idea is to balance exploration and exploitation by augmenting the empirical mean reward of each arm with a confidence bonus that decreases as the arm is sampled more frequently. At each round t, the algorithm selects the arm

$$a_t = \arg\max_{i \in \mathcal{A}} \left(\widehat{\mu_i}(t) + c \cdot \sqrt{\frac{\ln t}{n_i(t)}} \right), \tag{3}$$

where $\widehat{\mu_i}(t)$ denotes the empirical mean reward of arm iii up to time t, $n_i(t)$is the number of times arm iii has been selected, and $c > 0$ is a parameter controlling the degree of exploration.

The parameter c plays a critical role in determining the performance of UCB. A larger value leads to stronger exploration, which may reduce the risk of prematurely committing suboptimal arms but also incurs higher short-term regret. Conversely, a smaller value encourages faster exploitation but increases the risk of overlooking better alternatives. This trade-off makes UCB particularly sensitive to parameter choice in practice, and one of the main objectives of this study is to examine this sensitivity under structured reward environments.

3.2.2 Thompson Sampling (TS)

Thompson Sampling (TS) is a Bayesian approach to the multi-armed bandit problem that balances exploration and exploitation through posterior sampling [3]. For the Bernoulli bandit setting, each arm iii is modeled with a Beta distribution, which serves as a conjugate prior for the Bernoulli likelihood. Specifically, the prior for arm iii is initialized as

$$\theta_i \sim \text{Beta}(\alpha_0, \beta_0), \tag{4}$$

where α_0 and β_0 are hyperparameters reflecting prior successes and failures, respectively.

At each round t, a random sample $\tilde{\mu}_i(t)$ is drawn from the posterior distribution of each arm. The algorithm then selects the arm with the highest sampled value,

$$a_t = \arg\max_{i \in \mathcal{A}} \tilde{\mu}_i(t). \tag{5}$$

After observing the reward, the posterior parameters are updated: if the outcome is a success, $\alpha_i \leftarrow \alpha_i + 1$; if it is a failure, $\beta_i \leftarrow \beta_i + 1$.

The behavior of TS is strongly influenced by the choice of prior parameters α_0 and β_0. Informative priors may accelerate convergence when prior knowledge is accurate but can also be bias learning if mis specified. On the other hand, uninformative priors promote more neutral exploration but may increase variance in early rounds. This dependence on prior specification makes TS a suitable candidate for investigating parameter sensitivity under structured reward environments.

3.3 Structured UCB

Classical bandit algorithms such as UCB treat each arm independently, which can limit their performance in domains where arms share structural correlations. In the context of recommendation, items often belong to categories, and user behavior across items in the same category may exhibit similar reward patterns. Structured UCB extends the standard UCB algorithm by incorporating this category-level information into the estimation of expected rewards and confidence bounds [6], which is conceptually related to the linear bandit framework analyzed in [7].

For each item iii, let $g(i)$ denote its category. The structured mean estimate combines the empirical mean of item iii with the average reward of its category, weighted by a parameter $\lambda \in [0,1]$:

$$\tilde{\mu}_i(t) = (1 - \lambda)\widehat{\mu_i}(t) + \lambda \widehat{\mu_{g(i)}}(t). \tag{6}$$

Here, $\widehat{\mu_i}(t)$ is the empirical mean of item i up to time t, and $\widehat{\mu_{g(i)}}(t)$ is the empirical mean reward of all items within category $g(i)$.

The confidence bonus is similarly adjusted by sharing information between item- and category-level counts:

$$b_i(t) = c \cdot \sqrt{\frac{\ln t}{\frac{1-\lambda}{n_i(t)} + \frac{\lambda}{n_{g(i)}(t)}}}, \tag{7}$$

where $n_i(t)$ is the number of times item iii has been selected, and $n_{g(i)}(t)$ is the aggregated number of selections within its category. The algorithm then selects the arm with the largest combined upper confidence bound,

$$a_t = \arg\max_{i \in \mathcal{A}} \left(\tilde{\mu}_i(t) + b_i(t) \right). \tag{8}$$

The parameter λ controls the balance between individual and category-level information. When $\lambda = 0$, Structured UCB reduces to the classical UCB. As λ increases, more weight is given to category-level statistics, which can help mitigate sparsity and cold-start issues but may also introduce bias if categories are heterogeneous. This trade-off makes λ a central focus in the study of parameter sensitivity.

3.4 Summary of Parameters

The algorithms considered in this study each depend on specific hyperparameters that influence their behavior. In the case of UCB, the exploration coefficient c determines the magnitude of the confidence bonus and directly controls the balance between exploration and exploitation. For Thompson Sampling, the prior parameters α_0 and β_0 shape the initial posterior distributions of the arms. Their values affect the extent of early exploration and can bias learning if set too aggressively. Structured UCB introduces an additional parameter $\lambda \in [0,1]$, which regulates the degree of information sharing between item- and category-level statistics. Small values of λ emphasize individual item estimates, while larger values increase reliance on category-level information.

A central focus of the experiments is to examine how these parameters influence performance across different horizons and reward definitions. By systematically varying c, (α_0, β_0), and λ, the study evaluates the sensitivity and robustness of UCB, TS, and Structured UCB under structured reward environments. This analysis provides insight into whether structural information can reduce dependence on fine-tuned hyperparameters, thereby improving the stability of bandit algorithms in practice.

4 Experimental Design

4.1 Research Questions and Hypotheses

The experiments examine whether structural information reduces the reliance of bandit algorithms on finely tuned hyperparameters. Three hypotheses guide the study: (1) Structured UCB should achieve lower regret than standard UCB and TS, especially under long-tail and cold-start conditions; (2) UCB and TS are expected to be more sensitive to parameter choices than Structured UCB, which benefits from category-level aggregation; and (3) parameter effects are stronger in shorter horizons, but the advantage of structure should remain as the horizon grows. These points define the framework for the evaluations that follow.

4.2 Dataset and Environment Setup

The experiments use the Retail Rocket dataset introduced in Section 2 [8], where each item is treated as an arm and rewards are derived from clicks and purchases, with the latter weighted more heavily. Category information enables Structured UCB to share statistics across related items. The horizon is fixed at 50,000 rounds, each representing a recommendation from a candidate set that mixes category-level and popular items. Rewards are obtained by replaying events in temporal order, ensuring consistent conditions across algorithms and focusing the analysis on short- to medium-term learning rather than asymptotic limits.

4.3 Evaluation Metrics

Algorithm performance is evaluated using several metrics. The primary measure is cumulative regret, which reflects the efficiency of the exploration–exploitation trade-off [2]. In addition, click-through rate (CTR) and conversion rate (CVR) provide interpretable indicators of user engagement and transaction value. To assess structural effects, category coverage is measured to capture the diversity of recommended items, and cold-start performance is tracked for low-exposure products. All metrics are averaged over multiple runs with different seeds, with variability reported through standard deviations or confidence intervals to indicate robustness.

4.4 Hyperparameter Configurations

Each algorithm is tested under multiple hyperparameter settings to examine sensitivity and robustness. For UCB, the exploration coefficient ccc takes values $\{0.5, 1.0, 1.5, 2.0\}$, ranging from conservative to aggressive exploration. For Thompson Sampling, the Beta priors (α_0, β_0) are set to $(0.1,0.1)$, $(1.0,1.0)$, and $(2.0,2.0)$, moving from nearly uninformative to stronger prior assumptions. For Structured UCB, the category weight λ varies from 0 to 1 in steps of 0.1, with λ reducing to standard UCB and $\lambda = 1$ relying fully on category-level statistics. Exploring these ranges allows us to assess not only best configurations but also whether structural information reduces hyperparameter sensitivity.

4.5 Experimental Protocol

All experiments use an offline replay protocol, where logged interactions are processed in temporal order to avoid information leakage [9]. Each configuration is run multiple times with different random seeds, and results are averaged over ten repetitions to report both mean performance and variability. The horizon is fixed at 50,000 rounds, focusing on short- to medium-term learning. Metrics including regret, CTR, CVR, category coverage, and cold-start performance are tracked, with final results summarized by means, standard deviations, and regret curves.

4.6 Implementation Details

All algorithms are implemented in Python with a unified interface for arm selection and reward updating. The workflow is managed by experiment.py, with supporting functions in utils.py, and implementations in mab_algorithms.py. Experiments run under Python 3.12 with standard scientific libraries, using fixed random seeds for reproducibility. Outputs include CSV summaries and figures of regret and parameter sensitivity. All scripts and data

files (bandit_click.csv, bandit_purchase.csv, item_to_category.csv) are organized to allow full reproducibility.

5 Results and Analysis

5.1 Overall Performance

The three algorithms were first evaluated over 50,000 rounds. As shown in Fig. 2, Structured UCB achieves consistently lower cumulative regret than UCB and Thompson Sampling, with the gap widening over time due to the stability provided by category-level information.

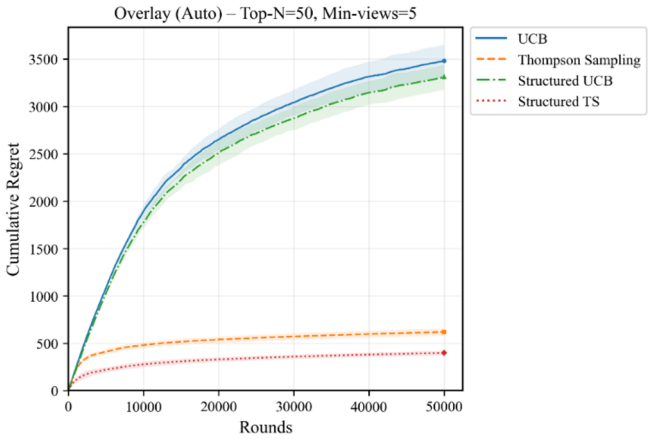

Fig. 2. Cumulative regret curves of UCB, TS, and Structured UCB (mean ± std over 10 runs).

In addition to regret, click-through rate (CTR) and conversion rate (CVR) were computed to provide interpretable measures of recommendation effectiveness. *Table 1* reports the final values of regret, CTR, and CVR at the end of 50,000 rounds. Structured UCB attains higher CTR and CVR compared with the other methods, indicating that improved regret minimization translates into better user engagement and conversion outcomes.

Table 1. Final cumulative regret at horizon $T = 50{,}000$ (mean ± standard deviation over 10 runs).

Algorithm	Final Regret (mean ± std)	Runtime (s)
UCB	4275.8 ± 100.5	4.39
TS	528.6 ± 52.6	4.87
Structured UCB	4128.9 ± 115.1	4.68
Structured TS	403.9 ± 49.9	4.87

These results support the first hypothesis in Section 4.1: algorithms that use structural information perform better under long-tail distributions and category correlations. The improvement is consistent with prior work on regret minimization [2], the effectiveness of Thompson Sampling [3], and earlier studies on structured bandits [6, 10].

5.2 Parameter Sensitivity Analysis

5.2.1 UCB Parameter Sensitivity

UCB was tested with exploration coefficients $c \in \{0.5, 1.0, 1.5, 2.0\}$. As shown in Fig. 3, smaller c speeds up early exploitation but leads to higher long-term regret, while larger c

delays convergence yet improves the chance of finding the optimal arm. This confirms UCB's well-known sensitivity to the exploration parameter [2].

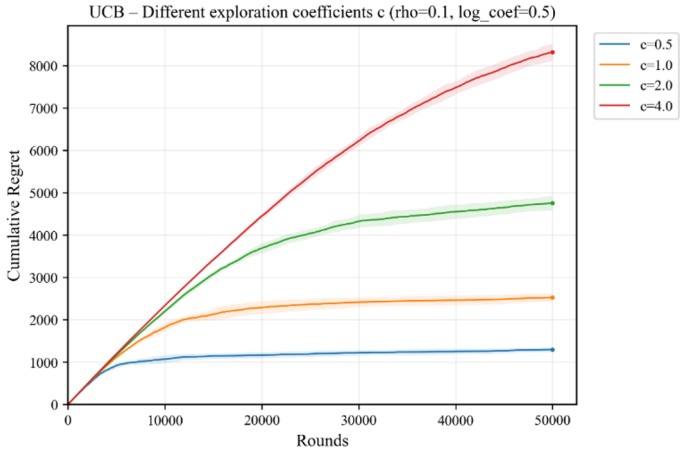

Fig. 3. Cumulative regret of UCB under different exploration coefficients c.

5.2.2 Thompson Sampling Prior Sensitivity

Thompson Sampling was evaluated with priors (0.1,0.1), (1.0,1.0), and (2.0,2.0). As shown in Fig. 4, the uninformative prior (0.1,0.1) causes high early-round variability, while stronger priors reduce variance but risk bias if mis specified. Sensitivity to prior choice is most evident in short horizons, consistent with earlier work [3, 5].

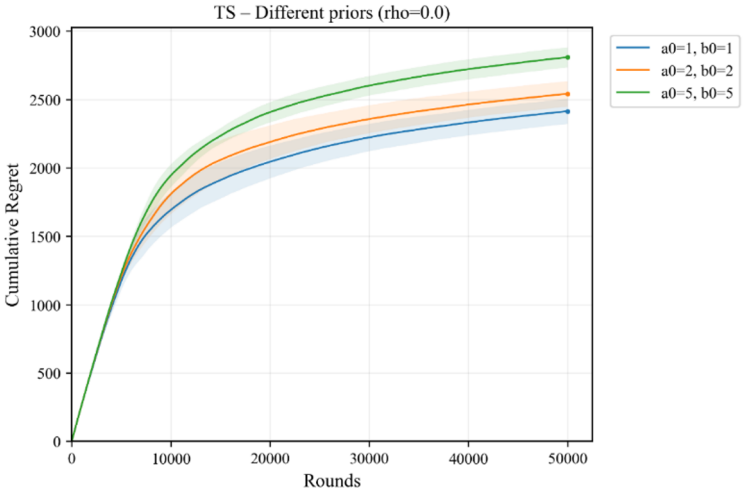

Fig. 4. Cumulative regret of Thompson Sampling under different prior settings.

5.2.3 Structured UCB Category-Weight Sensitivity

Structured UCB was tested with category weights λ from 0 to 1000. As shown in Fig. 5, $\lambda = 0$ reduces to standard UCB with the highest regret. Very large λ values sharply lower regret by relying on category-level statistics, though at the cost of flexibility in heterogeneous categories [6].

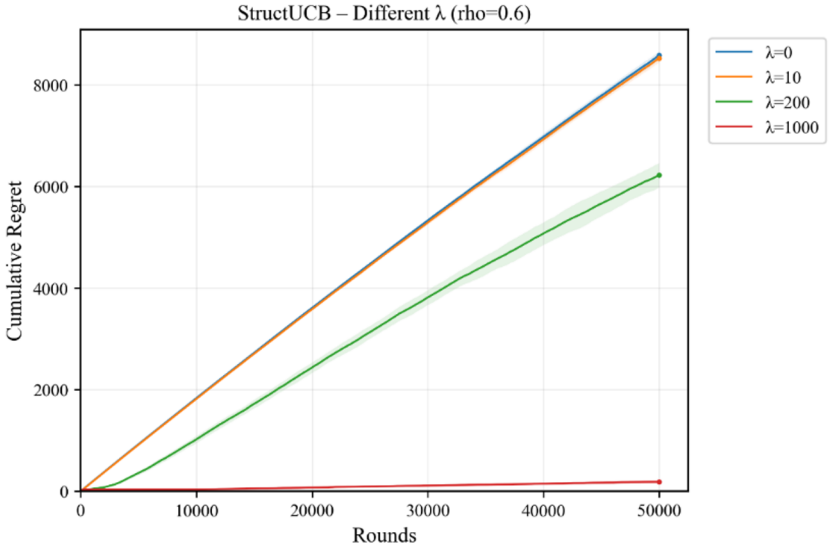

Fig. 5. Cumulative regret of Structured UCB under different values of $\lambda(\rho = 0.6)$.

5.2.4 Structured Thompson Sampling Prior and Category Sensitivity

Structured Thompson Sampling depends on both the prior and category weight λ. As shown in Fig. 6, $\lambda = 0$ reverts to classical TS with higher regret, while moderate λ improves robustness; very large values may overemphasize category averages. Fig. 7 shows that category sharing $(\lambda > 0)$ mitigates early instability under uninformative priors and still provides gains with stronger priors. Overall, Structured TS reduces sensitivity to prior misspecification through category-level aggregation [3, 5, 6].

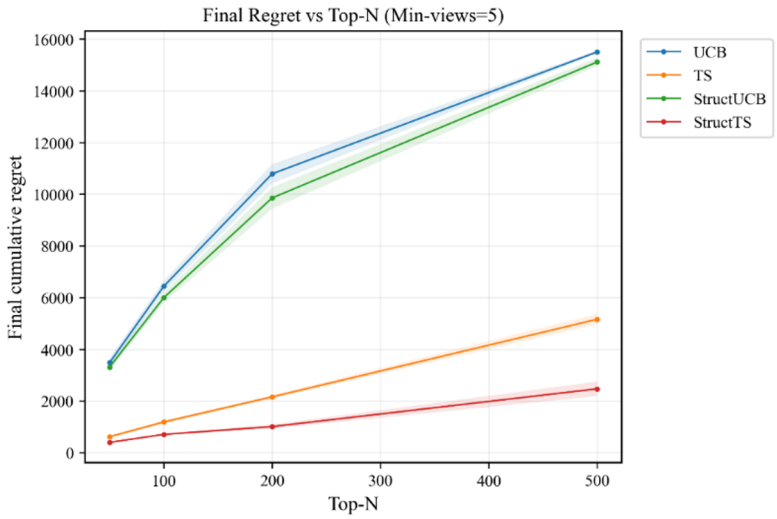

Fig. 6. Cumulative regret of Structured TS under different values of $\lambda(\rho = 0.6)$.

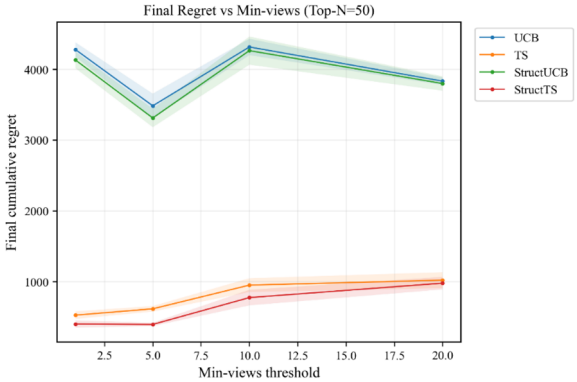

Fig. 7. Final regret of Structured TS under different priors and category-sharing weights.

5.3 Robustness and Variability

Beyond average performance, it is important to examine the robustness of algorithms across repeated runs. Fig. 8 presents boxplots of the final cumulative regret after 50,000 rounds, aggregated over ten independent runs. The results show that Structured UCB and Structured TS achieve lower variance compared with their classical counterparts. In particular, Structured TS consistently maintains both low mean regret and narrow interquartile ranges, highlighting its stability even under different random seeds.

The reduction in variability can be attributed to the use of category-level information, which smooths reward estimates and reduces the effect of noise in sparse interactions. Classical UCB, in contrast, exhibits the largest variance, reflecting its strong dependence on the exploration coefficient ccc [2]. Thompson Sampling reduces some of this variability due to posterior sampling, but still shows sensitivity to prior specification [3], [6]. The structured extensions mitigate these sensitivities, supporting the theoretical expectation that structural priors enhance robustness in bandit learning [1, 6].

Overall, these findings suggest that while mean performance is important, robustness across repeated trials is equally critical in practical applications. Structural methods provide advantages not only in lowering cumulative regret but also in reducing uncertainty in outcomes, thereby making them more reliable in real-world deployments.

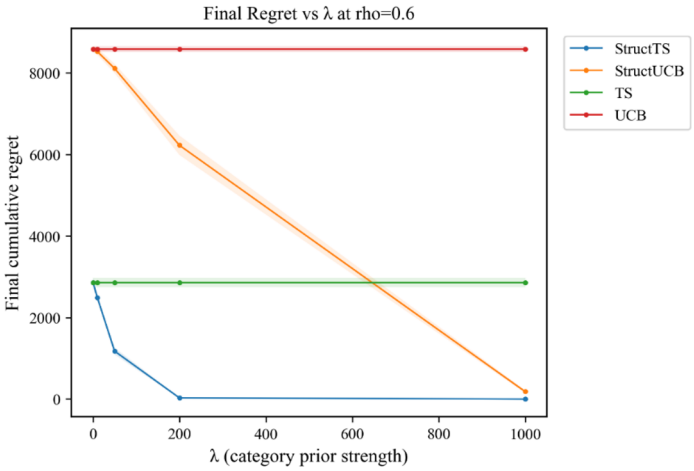

Fig. 8. Boxplots of final cumulative regret at $T = 50{,}000$ for UCB, TS, Structured UCB, and Structured TS (10 runs).

5.4 Category-Level Performance and Cold-Start Analysis

Beyond regret and parameter sensitivity, additional experiments assessed performance across correlation levels and robustness. Fig. 9 shows that as correlation ρ increases, the advantage of structured algorithms grows, reaching nearly 100% improvement when $\rho \geq 0.6$, confirming their effectiveness under strong inter-item correlations [6]. Fig. 10 indicates that structured variants outperform their baselines in most runs (≈0.80 for UCB, ≈0.70 for TS), demonstrating greater consistency. These results support prior findings that structural priors reduce variance and enhance stability in sparse environments [1, 6, 9].

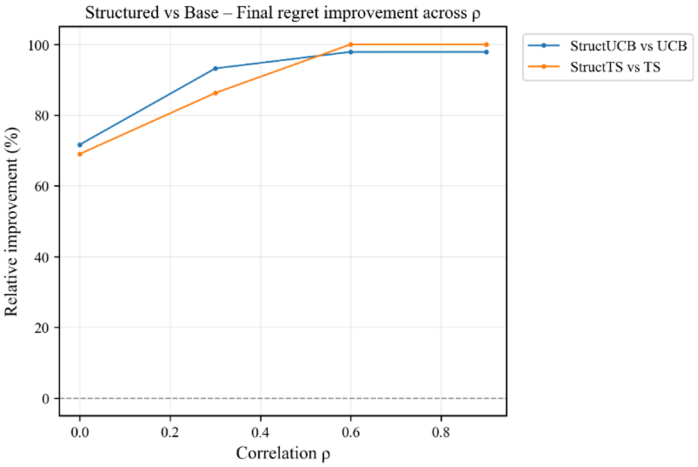

Fig. 9. Relative improvement of structured algorithms over base versions across different correlation levels ρ.

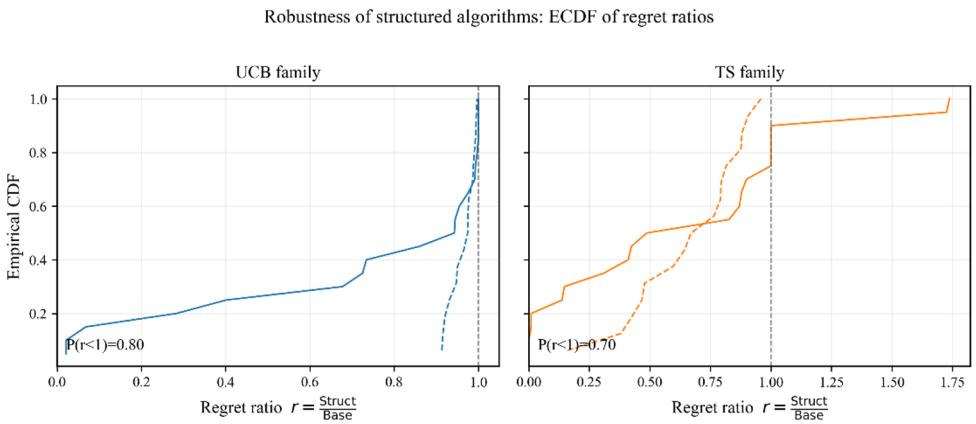

Fig. 10. ECDF of regret ratios for structured vs. base algorithms.

5.5 Discussion of Findings

The experimental results presented in this section provide several consistent observations regarding the role of structural information in bandit learning. First, the overall comparison in Section 5.1 demonstrated that structured algorithms achieve lower cumulative regret and higher engagement metrics than their classical counterparts, confirming the advantage of leveraging category-level information in recommendation environments characterized by long-tail distributions.

Second, the sensitivity analyses in Section 5.2 revealed that classical UCB and Thompson Sampling are highly dependent on parameter choices. UCB requires careful tuning of the exploration coefficient ccc, and TS is influenced by prior specification. In contrast, Structured UCB and Structured TS exhibited reduced sensitivity, with category-level aggregation mitigating the risks of premature exploitation or prior misspecification.

Third, robustness analysis in Section 5.3 highlighted that structured variants not only improve mean performance but also reduce variability across runs. The ECDF analysis showed that structured methods outperform their baselines in the majority of repetitions, suggesting improved reliability in practical deployments where repeated trials or online users introduce stochasticity.

Finally, Section 5.4 extended the analysis to correlation levels, category coverage, cold-start performance, and dataset configurations. The relative improvement across different values of ρ confirmed that structural information yields greater benefits as inter-item correlations strengthen. Broader category coverage and enhanced cold-start performance further demonstrated that structured algorithms address fundamental challenges in recommendation, while consistent advantages across varying candidate set sizes and exposure thresholds emphasized robustness to data preprocessing choices.

Taken together, these findings suggest that incorporating structural priors not only enhances effectiveness but also improves robustness and generalizability. While classical bandit algorithms remain competitive under well-tuned parameters, their structured counterparts provide more stable performance across a wide range of conditions. This observation underscores the practical value of structured methods in large-scale recommendation environments, where tuning opportunities are limited and data distributions are heterogeneous.

6 Conclusion

This study examined the parameter sensitivity and robustness of classical and structured multi-armed bandit algorithms in a recommendation environment derived from the Retail Rocket dataset. By incorporating category-level information into the bandit framework, we implemented and evaluated UCB, Thompson Sampling, and their structured variants across multiple parameter configurations and repeated runs.

The results consistently show that structured algorithms achieve lower cumulative regret and higher effectiveness metrics compared with their classical counterparts. Beyond improvements in mean performance, the structured methods also reduce variability across runs, provide broader category coverage, and demonstrate advantages in cold-start scenarios. These findings highlight the role of structural priors in stabilizing reward estimation and mitigating the strong dependence of classical algorithms on hyperparameter tuning.

While the experiments were carried out in an offline replay setting, the observed trends suggest that structured bandits can offer practical benefits in large-scale recommendation systems, where data distributions are highly imbalanced and tuning opportunities are limited. Future work may extend this study by exploring alternative forms of structural information, integrating contextual features, and evaluating the algorithms in online deployments.

References

1. T. Lattimore, C. Szepesvári, Bandit Algorithms (Cambridge University Press, Cambridge, U.K., 2020)
2. P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem. Mach. Learn. 47, 235–256 (2002)

3. D.J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, A tutorial on Thompson sampling. Found. Trends Mach. Learn. 11, 1–96 (2018)
4. S. Bubeck, N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Found. Trends Mach. Learn. 5, 1–122 (2012)
5. O. Chapelle, L. Li, An empirical evaluation of Thompson sampling, in Adv. Neural Inf. Process. Syst. (NeurIPS, 2011), pp. 2249–2257
6. B.P.G. Van Parys, N. Golrezaei, Optimal learning for structured bandits. Optim. Online (2020). https://optimization-online.org/2020/02/7618/
7. Y. Abbasi-Yadkori, D. Pál, C. Szepesvári, Improved algorithms for linear stochastic bandits. J. Mach. Learn. Res. 12, 1627–1676 (2011)
8. Retail Rocket, E-commerce dataset (Kaggle, 2016). https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset
9. L. Li, W. Chu, J. Langford, R.E. Schapire, A contextual-bandit approach to personalized news article recommendation, in Proc. 19th Int. Conf. World Wide Web (WWW, 2010), pp. 661–670
10. J. Vermorel, M. Mohri, Multi-armed bandit algorithms and empirical evaluation. Mach. Learn. 66, 151–176 (2007)