Active Inference-Driven Multi-Armed Bandits: Superior Performance through Dynamic Correlation Adjustments

Xiaoqi Lin*

Faculty of Computer Science and Technology, Qilu University of Technology (Shandong academy of sciences),Jinan, China

Abstract. In recent years, Multi-Armed Bandit (MAB) algorithms have gained substantial attention due to their effectiveness in real-world applications, such as recommendation systems, autonomous systems, and dynamic resource allocation. Traditional MAB algorithms, such as UCB and Thompson Sampling, often lack mechanisms to incorporate correlations between arms, limiting their adaptability and optimality in complex environments. This paper presents a novel MAB framework that integrates Active Inference through a dynamic Adaptive Influence Factor (AIF) mechanism. The AIF mechanism builds correlation matrices to capture inter-arm dependencies and dynamically adjusts exploration strategies through an influence factor, γ, which adapts over time based on pull counts. This adaptive exploration enhances decision-making in sparse and uncertain environments by leveraging correlations. The proposed framework is evaluated on movie recommendation data, with AIF-based algorithms, particularly AIF-TS, significantly outperforming traditional and correlated bandit approaches in settings with high data sparsity. These results demonstrate that dynamically adjusting exploration based on inter-arm relationships substantially improves performance in real-world applications, where data quality and relationships are often variable. The findings suggest that incorporating inter-arm correlations with active inference can lead to more efficient and effective decision-making in adaptive systems, highlighting the potential of AIF-based MAB algorithms in addressing real-world challenges.

1 Introduction

1.1 Formatting the title, authors and affiliations

The Multi-Armed Bandit (MAB) problem is a foundational framework in sequential decision-making, modeling the trade-off between exploration and exploitation. It has significant applications in domains such as recommendation systems, online advertising, and adaptive clinical trials [1]. In the classical MAB setting, an agent selects from a set of arms,

* Corresponding author: 202204370113@stu.qlu.edu.cn

each with an unknown reward distribution, aiming to maximize cumulative rewards over time.

Traditional MAB algorithms like the Upper Confidence Bound (UCB) and Thompson Sampling (TS) have been extensively studied for balancing exploration and exploitation [2]. UCB strategies rely on optimism in the face of uncertainty, selecting arms with the highest upper confidence bounds, while TS employs Bayesian inference to sample from the posterior distribution of each arm's reward [3]. While effective in many settings, these methods often struggle in environments with sparse feedback and correlations between arms [4,5]. The inability to leverage inter-arm correlations can lead to suboptimal performance, particularly in complex or dynamic environments.

Active inference provides a principled approach for autonomous agents operating in dynamic and non-stationary environments [6]. This framework diverges from traditional reinforcement learning by eliminating the reliance on explicit reward signals. Agents infer behaviors based on preference learning even in the absence of external rewards, which is particularly advantageous in real-world applications where the environment is complex and unpredictable [6, 7].

Thompson Sampling, on the other hand, is a Bayesian approach to MAB problems that effectively balances exploration and exploitation. It maintains a probability distribution over expected rewards for each arm and samples from these distributions to select the next arm to pull. This method has been shown to outperform many traditional strategies, especially in scenarios where the reward distributions are dynamic [8]. The adaptability of Thompson Sampling to non-stationary environments and correlated arms further enhances its applicability in real-world settings, such as online advertising and clinical trials [9, 10].

1.2 Active Inference & Adaptive Influence Factor

Active inference elucidates how autonomous agents make decisions in dynamic and uncertain environments by minimizing Expected Free Energy (EFE), which combines both epistemic (uncertainty-reducing) and pragmatic (reward-seeking) exploration [6, 11]. Grounded in Bayesian principles, agents continuously update their beliefs about the world based on incoming sensory data and take actions that minimize their prediction errors. This approach contrasts significantly with traditional reinforcement learning methods, which often rely on explicit reward signals to guide learning [6, 12, 13].

The computational processes involved in active inference include the formulation of a generative model that captures the dynamics of the environment, allowing agents to infer hidden states and make predictions about future observations. Agents sample actions based on their posterior beliefs, integrating both exploration and exploitation into a unified decision-making process [11, 14]. This integration leads to a more nuanced understanding of the exploration-exploitation trade-off by incorporating different types of uncertainty. Agents can dynamically adjust their exploration strategies based on the type of uncertainty they perceive, enhancing their adaptability to changing environments [15, 13].

This paper introduces the Adaptive Influence Factor (AIF) as an extension of active inference principles within MAB frameworks. The AIF mechanism dynamically adjusts based on inter-arm correlations, allowing agents to leverage additional environmental information. By integrating the AIF into traditional algorithms like UCB and Thompson Sampling, AIF-UCB and AIF-TS facilitate more nuanced exploration strategies, particularly in environments with sparse rewards and correlated arms. This incorporation provides a refined understanding of the exploration-exploitation trade-off, enabling agents to dynamically adjust their strategies based on perceived uncertainty, enhancing their adaptability to changing environments.

1.3 Motivation & Contribution

The rapid growth of applications that rely on adaptive decision-making under uncertainty—such as recommendation systems, autonomous vehicles, and resource allocation—demands algorithms that can effectively balance exploration and exploitation in complex, sparse, and dynamic environments. Traditional Multi-Armed Bandit (MAB) algorithms, like UCB and Thompson Sampling, offer foundational approaches to these challenges but often struggle when data is sparse, noisy, or when the environment undergoes frequent changes. Furthermore, while some MAB variants incorporate correlation structures (e.g., C-UCB and C-TS), they still lack mechanisms to dynamically adapt to evolving data patterns, which limits their robustness in real-world applications.

1.3.1 Motivation

The motivation behind this research stems from the need to enhance MAB algorithms' adaptability and robustness in environments characterized by high sparsity, uncertainty, and noise. Existing solutions either suffer from high cumulative regret in noisy settings or lack the flexibility to respond to changing conditions effectively. Additionally, while recent advancements have introduced correlated bandit algorithms to leverage similarity structures, these methods do not dynamically adjust their exploration-exploitation strategies based on the environment's changing states. Active inference principles, which allow for adaptive balance by adjusting exploration dynamically, show potential to fill this gap. By incorporating dynamic gamma adjustments and active inference frameworks, this study seeks to develop MAB algorithms that can adapt more effectively to diverse, challenging environments.

1.3.2 Contributions

1) **Introduction of AIF-TS and AIF-UCB Algorithms**: This study proposes two novel algorithms—AIF-TS (Active Inference-based Thompson Sampling) and AIF-UCB (Active Inference-based Upper Confidence Bound)—that integrate active inference principles with traditional MAB structures. By dynamically adjusting gamma values, these algorithms are designed to balance exploration and exploitation in response to real-time changes in data quality, sparsity, and noise levels.
2) **Dynamic Gamma Adjustment Mechanism**: A core innovation of the proposed algorithms is the dynamic gamma adjustment mechanism. Unlike static exploration parameters, dynamic gamma adjusts based on the cumulative experience of the algorithm, enabling a shift from exploration to exploitation as the environment stabilizes. This approach improves performance in dynamic settings, reducing cumulative regret and enhancing decision-making efficiency.
3) **Extensive Experimental Evaluation**: Through rigorous experiments under varying sparsity levels (p=0.1, 0.3, 0.5, 0.7) and padding values (padval=0.1, 0.3, 0.5, 0.7), the study demonstrates that AIF-TS consistently outperforms both traditional and correlated bandit algorithms across all settings. The experiments highlight the robustness of AIF-TS in maintaining lower cumulative regret in noisy and sparse conditions, suggesting its potential for real-world applications where data quality fluctuates.
4) **Insights into the Limitations of Existing Correlated MAB Algorithms**: By comparing AIF-TS and AIF-UCB to correlated bandit algorithms (C-UCB and C-TS), this study provides insights into the limitations of current correlation-based approaches. While C-UCB and C-TS leverage correlations to some extent, they lack the adaptability and

resilience found in active inference-based algorithms, particularly under high noise and sparsity.

5) **Framework for Future Adaptable MAB Research**: This study contributes a foundation for future research into adaptable MAB algorithms, particularly for settings that require continuous adaptation. By introducing active inference principles to the MAB framework, this work opens avenues for further research in integrating adaptive mechanisms into decision-making algorithms, paving the way for more resilient solutions in uncertain, dynamic environments.

2 Related work

2.1 Classical Multi-Armed Bandit Algorithms

2.1.1 Traditional Multi-Armed Bandits

The traditional MAB problem addresses the exploration-exploitation trade-off in sequential decision-making. Two widely used algorithms are the ε-greedy algorithm and Thompson Sampling (TS). The ε-greedy algorithm selects a random arm with probability ε for exploration and the best-known arm with probability $1-\epsilon$ for exploitation [2]. TS, introduced by Thompson in 1933, operates on a Bayesian framework, selecting arms based on the probability that they are the best option [16, 17]. TS has been shown to achieve logarithmic expected regret, making it powerful in environments where reward distributions change over time [2, 16].

2.1.2 Contextual Multi-Armed Bandits

The contextual MAB (CMAB) framework evolves from the traditional MAB model by incorporating additional context information to enhance decision-making. This allows algorithms to consider the context associated with each arm, leading to more informed choices and improved performance. Adaptations of successful MAB policies, such as UCB, have been developed to operate within the CMAB context, utilizing classification algorithms to predict rewards based on contextual information [3, 5, 18].

2.1.3 Active Inference in Multi-Armed Bandits

Active inference integrates principles from Bayesian inference with decision-making paradigms, positing that agents actively seek to minimize uncertainty about the environment while maximizing expected rewards. In MABs, active inference facilitates exploration through epistemic curiosity, where agents prioritize actions that reduce uncertainty [6, 8]. By employing a belief-based model, active inference adaptively balances exploration and exploitation without relying solely on reward signals, enhancing adaptability in dynamic environments [6, 18, 13].

2.2 Correlated MAB Algorithms

Research has explored leveraging correlations between arms to improve MAB performance. Gupta et al. [5] introduced methods for handling correlated arms, allowing algorithms to leverage these correlations, optimizing the selection process across multiple arms, and reducing sub-optimal choices. Their approach incorporates correlation information but does

not adjust exploration strategies based on evolving correlations, limiting effectiveness in non-stationary environments. Other studies have examined Gaussian Process Bandits [19, 20] and graph-based methods [21, 22] to model correlations.

2.3 Active Inference in Decision-Making

Active inference provides a first-principles account of autonomous agent behavior by minimizing Expected Free Energy (EFE). This framework has been applied in various decision-making contexts, enabling agents to perform epistemic exploration and adapt to changing environments [6, 2]. Tschantz et al. [11] demonstrated the scalability of active inference in high-dimensional tasks, highlighting its potential advantages over traditional reinforcement learning methods. The flexibility inherent in active inference-driven algorithms allows them to thrive in challenging real-world applications, where environmental dynamics are constantly changing, and traditional methods may fall short.

2.4 Performance Improvements

Empirical studies have shown that active inference can provide significant performance enhancements in MAB tasks compared to traditional algorithms. Marković et al. [8] conducted an empirical evaluation comparing an active inference algorithm with two established bandit algorithms, finding that active inference substantially outperformed both in dynamic switching bandit scenarios. Wakayama and Ahmed [23] further demonstrated that active inference strategies, when applied to contextual MABs, require fewer iterations to identify optimal actions and achieve superior cumulative regret compared to conventional methods.

2.5 Adaptation to Dynamic Environments

Active inference algorithms exhibit remarkable flexibility in adapting to dynamic environments, making them suitable for real-world applications where conditions frequently change. Delavari et al. [24] discussed the implementation of active inference in autonomous vehicle control, showcasing its ability to minimize prediction error in rapidly changing environments. Yu et al. [25] highlighted the integration of graphical models with bandit frameworks to address complex dependencies between actions and observations, opening new avenues for adapting to dynamic scenarios.

3 Problem formulation

3.1 Classical MAB Problem

In the classical MAB problem, an agent selects one of K arms at each time step t, aiming to maximize cumulative rewards over T rounds. Each arm k provides rewards drawn from an unknown distribution. The agent seeks to minimize cumulative regret, defined as the difference between the rewards obtained by an optimal strategy and those collected by the agent [18].

3.2 Correlated Arms

In many real-world scenarios, arms are not independent; the rewards of one arm can influence others. This correlation can be captured using a correlation matrix C, where $C(k, j)$ represents the correlation between arms k and j [11]. Leveraging these correlations can enhance exploration efficiency, particularly in sparse environments where feedback is limited.

3.3 Active Inference Framework

Active inference reframes decision-making by minimizing Expected Free Energy (EFE), combining both epistemic and pragmatic exploration [6]. The agent maintains a belief about the environment and updates this belief based on observations. The Adaptive Influence Factor (AIF) dynamically adjusts the impact of correlated arms based on observed rewards and arm selection frequencies, allowing the agent to leverage correlations effectively.

4 Proposed algorithms

4.1 AIF-UCB

The AIF-UCB algorithm modifies the traditional UCB confidence interval to incorporate inter-arm correlations through the Adaptive Influence Factor γ :

$$UCB_k = \hat{u}_k + B_k\sqrt{\frac{2\log t}{n_k}} + \gamma \sum_j C(k,j)\left(\hat{u}_j - \hat{u}_k\right) \quad (1)$$

-$\hat{u}_k$: Estimated mean reward of arm k.
-n_k: Number of times arm k has been pulled.
-B_k: Scaling factor.
-γ : Adaptive influence factor.
-$C(k,j)$: Correlation between arms k and j.

The gamma factor γ adapts dynamically based on exploration phases and arm selection frequencies, enhancing the algorithm's ability to leverage correlations. In early stages, a higher γ encourages exploration based on correlations, while in later stages, γ decreases to focus on exploitation.

4.2 Dynamic Gamma Adjustment Mechanism

The gamma parameter γ is adjusted based on two factors:

4.2.1 Exploration Phase

As time t increases, γ decreases exponentially, indicating reduced reliance on correlations as more data is gathered.

$$\gamma = \gamma_{initial} \cdot \exp\left(-\gamma_{decay} \cdot t\right) \quad (2)$$

4.2.2 Pull Counts

For arms with fewer pulls, γ is increased to strengthen the influence of correlations, promoting exploration.: We develop a mechanism for adjusting the gamma parameter dynamically based on exploration phases and arm selection frequencies, enhancing exploration efficiency.

$$\gamma = \begin{cases} 1.5 \cdot \gamma, if \ \ n_k < threshold \\ 0.5 \cdot \gamma, \ otherwise \end{cases} \quad (3)$$

4.3 AIF-TS

Similarly, the AIF-TS algorithm incorporates correlations into the Thompson Sampling process by adjusting the sampled means:

$$\hat{u}_k^{correlated} = \hat{u}_k + \gamma \sum_j C(k,j)\left(\hat{u}_j - \hat{u}_k\right) \quad (4)$$

This adjustment allows AIF-TS to balance exploration and exploitation more effectively, particularly in environments with sparse rewards. By considering the influence of correlated arms, AIF-TS can make more informed sampling decisions.

5 Experiments

5.1 Datasets

This set of experiments utilizes the MovieLens dataset[26], a widely adopted benchmark dataset containing 1M ratings for 3,883 movies rated by 6,040 users. Each movie is assigned a rating on a 1-5 scale and is categorized into one or more genres. For simplification, a single genre per movie is randomly selected for analysis. The dataset is split into two parts: the first half, comprising the most frequent user ratings, serves as the training set to generate pseudo-reward entries, while the second half is designated as the test set for evaluating algorithm performance. This split ensures that the rating distributions for training and testing are distinct, thus better simulating real-world recommendation environments.

5.2 Experimental Metrics

The goal of these experiments is to assess various algorithms' ability to provide accurate movie recommendations under conditions of data noise and sparsity. In this context, cumulative regret is employed as the primary metric, representing the gap between the reward obtained by consistently selecting the optimal movie and that achieved by the algorithm. Experiments are conducted across a range of parameter settings to understand algorithm sensitivity to varying levels of data quality.

5.2.1 Cumulative Regret

Cumulative regret serves as the primary metric to assess recommendation performance. Regret is defined as the difference between the reward from always selecting the optimal action (i.e., recommending the best movie) and the actual reward obtained by the algorithm. Consequently, cumulative regret represents the total regret accumulated over a given number of rounds.

Lower cumulative regret indicates better performance, suggesting that the algorithm more accurately identifies and recommends the best movies over time.

5.2.2 Impact of Masking (Parameter p)

The parameter p represents the fraction of pseudo-reward entries removed from the training data to simulate incomplete information. By setting p to values 0.1, 0.3, 0.5, and 0.7, the

algorithms' ability to adapt to increasing data sparsity is evaluated. With higher p values, fewer pseudo-reward entries are available, simulating scenarios with greater information loss.

5.2.3 Robustness to Noise (Parameter padval)

The padval parameter represents added noise to the pseudo-reward data, simulating uncertainty in user preferences. By setting padval to 0.1, 0.3, 0.5, and 0.7 in successive experiments, the algorithms' stability in the face of noise is assessed. Each setting reflects a different level of confidence in the pseudo-rewards, with higher values indicating increased noise.

5.2.4 Confidence Intervals

Confidence intervals are plotted around cumulative regret curves for each algorithm. These intervals capture the variability in performance across multiple iterations, reflecting the consistency of each algorithm.
Narrower confidence intervals indicate greater reliability and consistency, as the algorithm exhibits stable performance across different experimental runs.

5.3 Baseline Algorithms

The performance of AIF-UCB and AIF-TS was compared with the following baseline algorithms:

- UCB
- TS
- Correlated UCB (C-UCB)
- Correlated Thompson Sampling (C-TS)

5.4 Parameter Settings

5.4.1 Number of Rounds (T)

Each algorithm executed over 10,000 rounds, where in each round, an arm was selected, and a reward was obtained. This setting allowed for the observation of long-term performance and cumulative regret trends, essential for evaluating an algorithm's adaptability and consistency over extended use.

5.4.2 Iterations

Each algorithm was run for 200 iterations to ensure reliability and statistical significance in the results. Multiple runs across different initial conditions provided a comprehensive understanding of the algorithm's performance variability.

5.4.3 Gamma Decay in AIF Algorithms

The gamma parameter (γ) was adjusted dynamically within the Adaptive Importance Feedback (AIF) algorithms, gradually decreasing over time to favor exploitation over exploration. This decay strategy aimed to optimize performance by encouraging a balance between trying new recommendations and selecting known high-reward options.

5.4.4 Data Masking Parameter (p)

Various masking ratios (e.g., p=0.1,0.3,0.5,0.7) were used to simulate different levels of partial data availability, reflecting real-world situations where data may be incomplete or sparse. This setting tested each algorithm's adaptability to data sparsity, particularly its ability to continue making accurate recommendations with limited pseudo-reward entries.

5.4.5 Padding Values (padval)

Different padding values (0.1, 0.3, 0.5, 0.7) were used to replace missing pseudo-reward entries, simulating varying degrees of noise or uncertainty in user feedback. These values represented different levels of confidence in the pseudo-rewards, allowing an assessment of each algorithm's robustness to noisy data and its capacity to maintain consistent performance when data quality varies.

6 Results and analysis

6.1 Recommending the Best Movie

The figures illustrate the cumulative regret across various algorithms—UCB, TS, C-UCB, C-TS, AIF-UCB, and AIF-TS—under different sparsity levels (p) and padding values (padval) using the MovieLens dataset. Each figure corresponds to a fixed sparsity level, with padding values set at 0.1, 0.3, 0.5, and 0.7 for subfigures (a) through (d), respectively. Cumulative regret generally increases with padval across all scenarios, showcasing the algorithms' performance in handling noise and variability.

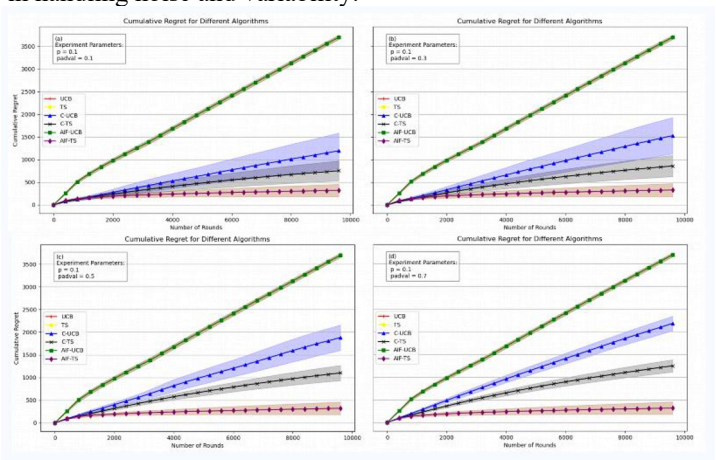

Fig. 1. Cumulative Regret Across Algorithms with Varying Pad Values (Low Sparsity, p=0.1).

Shown as Figure 1, each algorithm's performance under low sparsity is depicted. Despite minimal data sparsity, AIF-TS excels over others across increasing noise levels, maintaining lower cumulative regret and demonstrating significant robustness against added variability.

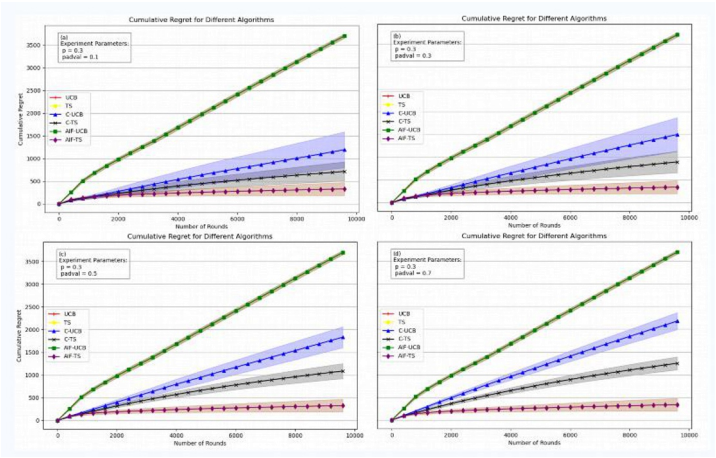

Fig. 2. Cumulative Regret Across Algorithms with Varying Pad Values (Moderate Sparsity, p=0.3).

Shown as Figure 2, this figure explores a moderate sparsity setting. Here, AIF-TS continues to exhibit the lowest cumulative regret, effectively managing rising pad values better than other algorithms. This figure highlights AIF-TS's consistent performance superiority in environments with moderate levels of noise and data uncertainty.

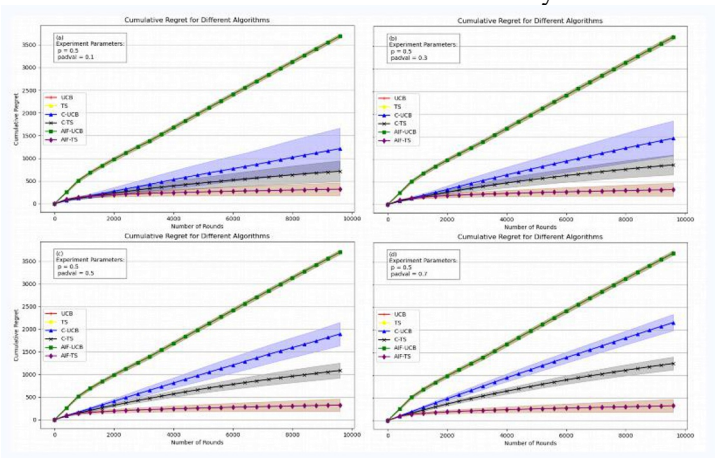

Fig. 3. Cumulative Regret Across Algorithms with Varying Pad Values (High Sparsity, p=0.5).

Shown as Figure 3, the algorithm responses illustrate high sparsity conditions, where AIF-TS's adaptability becomes more pronounced. Under these circumstances, it significantly outperforms its competitors by maintaining the lowest cumulative regret, highlighting its effectiveness in managing higher levels of uncertainty and sparsity.

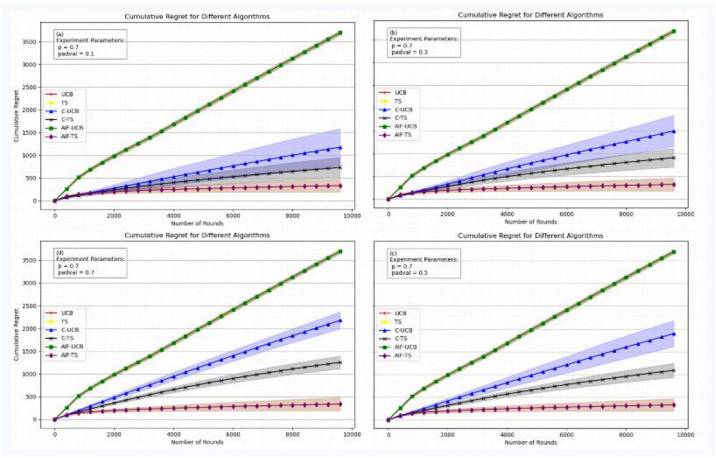

Fig. 4. Cumulative Regret Across Algorithms with Varying Pad Values (Extremely High Sparsity, p=0.7).

Shown as Figure 4, this figure represents the most challenging scenario with extremely high sparsity. AIF-TS not only maintains its performance advantage but also accentuates the limitations of traditional algorithms such as UCB and C-UCB, which exhibit significantly higher regrets as noise levels increase. This underscores AIF-TS's robustness and exceptional adaptability to complex recommendation tasks under conditions of severe data sparsity and noise.

6.2 Analysis

6.2.1 Algorithm Performance Under Different Noise and Sparsity Settings

6.2.1.1 Extremely Low Sparsity (Figure 1: p = 0.1)

Impact of Increasing Pad Values: All algorithms experienced an increase in cumulative regret as pad values rose from 0.1 to 0.7, reflecting the disruptive impact of added noise in the pseudo-reward data. AIF-TS maintained the lowest cumulative regret, proving its robustness against increased noise levels, while UCB and TS showed the largest increases, indicating their sensitivity to noise.

Algorithm Performance Comparison: AIF-TS excelled across all pad values, affirming its stability and resilience under low data sparsity and noise. Conversely, C-UCB and C-TS improved over UCB and TS but struggled more as noise increased, suggesting their lack of resilience compared to AIF-TS. AIF-UCB, although better than baseline UCB, still fell short of AIF-TS, whose additional exploration mechanism enabled superior adaptation to fluctuating data quality.

Sensitivity to Masking and Noise: At a low masking level (p=0.1), minimal data sparsity facilitated reliance on pseudo-reward entries, with AIF-TS showing consistent performance. C-UCB and C-TS, however, exhibited sharper increases in cumulative regret with rising pad values, indicating their vulnerability to the quality of pseudo-reward data and arm correlations.

6.2.1.2 Low Sparsity (Figure 2: p = 0.3)

Impact of Increasing Pad Values: Cumulative regret rose for all algorithms as pad values increased, highlighting the effect of noise on decision-making. AIF-TS consistently had the lowest cumulative regret, showcasing its robustness. In contrast, UCB and TS faced significant challenges, underscoring their limitations in handling data uncertainty.
Algorithm Performance Comparison: AIF-TS maintained superior performance, continuously displaying low cumulative regret and confirming its adaptability in moderately sparse and noisy conditions. C-UCB and C-TS showed some improvements over UCB and TS but were less resilient to noise, as evidenced by their higher cumulative regrets at elevated pad values.
Sensitivity to Masking and Noise: With a masking level of p=0.3, lower sparsity allowed algorithms to better utilize pseudo-reward entries. AIF-TS remained stable, while C-UCB and C-TS were more affected by increased pad values, revealing their sensitivity to data quality and correlations.

6.2.1.3 Moderate Sparsity (Figure 3: p = 0.5)

Impact of Increasing Pad Values: As pad values increased, all algorithms faced rising cumulative regret due to the added noise and data quality issues, complicating accurate arm selection. AIF-TS consistently displayed the most robust performance, maintaining the lowest cumulative regret. Conversely, UCB and TS experienced the most significant increases, indicating their high sensitivity to noise.
Algorithm Performance Comparison: AIF-TS outperformed all other algorithms across pad settings, affirming its effective adaptation in environments with moderate sparsity and noise. C-UCB and C-TS demonstrated moderate improvements but were less effective under noisy conditions, showing higher cumulative regret as noise increased.
Sensitivity to Masking and Noise: At a moderate sparsity level (p=0.5), the algorithms faced a challenging data environment. AIF-TS exhibited stable performance, while C-UCB and C-TS were more vulnerable to fluctuations in pseudo-reward data quality and inter-arm correlations.

6.2.1.4 Extremely High Sparsity (Figure 4: p = 0.7)

Impact of Increasing Pad Values: Cumulative regret increased across all algorithms with rising pad values, reflecting severe noise and uncertainty in pseudo-reward data. AIF-TS consistently showed the lowest cumulative regret, demonstrating its capacity to handle substantial noise effectively. UCB and TS, however, exhibited the steepest increases, underscoring their sensitivity to high noise levels.
Algorithm Performance Comparison: AIF-TS maintained superior performance, reinforcing its robustness in extreme data sparsity and noisy conditions. While C-UCB and C-TS made some gains over UCB and TS, they struggled more under increased noise, showing greater cumulative regret.
Sensitivity to Masking and Noise: The extremely high masking level (p=0.7) posed significant challenges, with AIF-TS maintaining stable performance. C-UCB and C-TS, however, faced sharper increases in cumulative regret, highlighting their susceptibility to the quality of pseudo-reward data and inter-arm correlations.

6.2.2 Overall Summary and Conclusion

Performance Improvements: Empirical analysis shows that AIF-TS and other active inference-driven MAB algorithms significantly enhance performance in both static and dynamic environments, adaptable across various levels of data sparsity. AIF-TS consistently had the lowest cumulative regret, indicating its high decision-making efficiency and resilience, particularly under conditions of high sparsity and noise.
Adaptation to Dynamic Environments: AIF-TS excels in dynamic settings, such as autonomous vehicle control, where it must adjust to rapidly changing conditions. Its flexibility is crucial for applications requiring continuous response to unpredictable factors.
Comparison with Correlated and Traditional Algorithms: AIF-based algorithms (AIF-UCB and AIF-TS) consistently outperformed their correlated (C-UCB and C-TS) and traditional (UCB and TS) counterparts. While C-UCB and C-TS improved over UCB and TS by leveraging correlations, they lacked the adaptive strategy adjustment that AIF algorithms possess, resulting in higher regret in fluctuating environments. Traditional algorithms like UCB and TS, which do not capitalize on correlations or adaptive mechanisms, demonstrated the highest cumulative regret, indicating slower convergence and less effective performance under noisy conditions.
Impact of Dynamic Gamma: Dynamic gamma adjustment was instrumental in enhancing the performance of AIF algorithms. During the initial stages, a higher gamma (γ) value facilitated broad exploration across correlated arms, thereby reducing uncertainty. As data accumulated, gamma decreased, enabling the algorithms to focus on exploiting the most rewarding arms. This dynamic adjustment balanced exploration and exploitation effectively, leading to improved exploration efficiency and overall performance, especially in sparse and noisy environments.

7 Conclusion and future work

7.1 Conclusion and Future Research Directions

This research has demonstrated the efficacy of the AIF-TS algorithm in handling high levels of data sparsity and noise within multi-armed bandit (MAB) environments, revealing that active inference-driven approaches can significantly outperform traditional and correlation-based algorithms. By incorporating dynamic gamma adjustment and leveraging active inference principles, AIF-TS has shown superior adaptability, achieving lower cumulative regret across various settings of sparsity and noise. This advancement addresses the limitations of traditional MAB algorithms, especially in noisy and sparse data scenarios, and offers a robust solution for real-world applications that involve uncertain and evolving information, such as recommendation systems and dynamic resource allocation.

7.2 Implications and Contributions

The findings of this study provide a new benchmark for MAB algorithms by demonstrating that adaptive, active inference-based approaches are better suited to dynamic environments where data quality fluctuates. This research fills a critical gap in MAB studies by offering a framework that adapts well to both sparse and high-noise conditions, which are common in real-world scenarios but have been less explored in previous research. The dynamic gamma adjustment mechanism, which balances exploration and exploitation based on environmental feedback, contributes a novel methodology that can enhance the performance of adaptive algorithms in various fields. For researchers and practitioners, the experimental results and

methodological insights from AIF-TS serve as a foundation to further explore adaptive, inference-based designs that can handle more complex data conditions.

7.3 Limitations and Future Directions

Despite the promising outcomes of the AIF-TS algorithm, there are several limitations to address to enhance its applicability and optimize performance. The computational demands of the dynamic gamma adjustment mechanism may limit real-time performance in large-scale applications where quick responses are essential. Future work could optimize this mechanism through approximate methods or adaptive schedules to reduce continuous recalibration needs.

The performance of AIF-UCB in high sparsity and noisy conditions, although improved over traditional UCB, still falls short of AIF-TS. This suggests the active inference in UCB is less effective at balancing exploration and exploitation under uncertainty. Future research might explore hybrid approaches or refine the active inference strategy in AIF-UCB, potentially incorporating elements from Thompson Sampling for more flexible response to varying data quality.

Additionally, this study was conducted in a controlled setting that may not fully reflect the complexities of real-world scenarios, such as environments with shifting reward distributions. Investigating how AIF-TS and AIF-UCB perform in dynamically evolving environments could greatly extend this work.

Future directions could involve developing more efficient computational frameworks for implementing AIF-TS and AIF-UCB in large-scale, real-time applications. Integrating advanced reinforcement learning or neural network-based inference with active inference could yield more adaptable, resilient models. Further refining AIF algorithms to handle multi-objective optimization problems or environments with overlapping rewards could broaden their applicability in areas like recommendation systems, autonomous navigation, and adaptive resource allocation.

References

1. N. Sajid, P. J. Ball, T. Parr, K. J. Friston, Active inference: demystified and compared. arXiv:1909.10863v3 [cs.AI] (30 Oct 2020)
2. V. Kuleshov, D. Precup, Algorithms for the multi-armed bandit problem. J. Mach. Learn. Res. 1 (2000), pp. 1-48
3. S. Agrawal, N. Goyal, Analysis of Thompson Sampling for the Multi-armed Bandit Problem. JMLR: Workshop and Conference Proceedings 23 (2012), pp. 39.1 – 39.26
4. D. Marković, H. Stojić, S. Schwöbel, S. J. Kiebel, An empirical evaluation of active inference in multi-armed bandits. Neural Networks 144 (2021), pp. 229 – 246
5. S. Gupta, S. Chaudhari, G. Joshi, O. Yağan, Multi-Armed Bandits with Correlated Arms. Carnegie Mellon University (2021)
6. P. A. Ortega, D. A. Braun, Generalized Thompson sampling for sequential decision-making and causal inference. Complex Adaptive Systems Modeling 2 (2014)
7. D. Cortes, Adapting multi-armed bandits policies to contextual bandits scenarios. arXiv:1811.04383v2 (2019)
8. G. Burtini, J. Loeppky, R. Lawrence, A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit. arXiv:1510.00757 (2015)
9. Y. Li, L. Liu, W. Pu, H. Liang, Z.-Q. Luo, Optimistic Thompson Sampling for No-Regret Learning in Unknown Games. arXiv:2402.09456v2 [cs.LG] (2024)

10. W. Qian, C.-K. Ing, J. Liu, Adaptive Algorithm for Multi-armed Bandit Problem with High-dimensional Covariates (2023)
11. A. Tschantz, M. Baltieri, A. K. Seth, C. L. Buckley, Scaling active inference. arXiv:1911.10601v1 [cs.LG] (2019)
12. G. Pezzulo, F. Rigoli, K. Friston, Active Inference, Homeostatic Regulation and Adaptive Behavioural Control. Prog. Neurobiol. (2024)
13. A. Paul, T. Isomura, A. Razi, On Predictive Planning and Counterfactual Learning in Active Inference. Entropy 26 (2024), p. 484
14. M. S. Tomov, V. Q. Truong, R. A. Hundia, S. J. Gershman, Dissociable neural correlates of uncertainty underlie different exploration strategies. Nat. Commun. 11 (2020), p. 2371
15. S. Gijsen, M. Grundei, F. Blankenburg, Active inference and the two-step task. Sci. Rep. 12 (2022), p. 17682
16. W. R. Thompson, On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. Biometrika 25(3/4) (1933), pp. 285-294
17. O. Chapelle, L. Li, An Empirical Evaluation of Thompson Sampling. Yahoo! Research (2023)
18. T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput. Mater. 5 (2019), p. 21
19. N. Srinivas, A. Krause, S. M. Kakade, M. Seeger, Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. (2010)
20. S. R. Chowdhury, A. Gopalan, On Kernelized Multi-armed Bandits. (2017)
21. N. Cesa-Bianchi, C. Gentile, G. Lugosi, Regret minimization for reserve prices in second-price auctions. IEEE Trans. Inf. Theory 59(11) (2013), pp. 7455-7462
22. M. Valko, A. Carpentier, R. Munos, Semi-bandit optimization in the adversarial setting. (2014)
23. S. Wakayama, N. Ahmed, Active Inference for Autonomous Decision-Making with Contextual Multi-Armed Bandits. University of Colorado Boulder (2023)
24. E. Delavari, J. Moore, J. Hong, J. Kwon, Towards Human-Like Driving: Active Inference in Autonomous Vehicle Control. arXiv:2407.07684 (2024)
25. T. Yu, B. Kveton, Z. Wen, R. Zhang, O. J. Mengshoel, Graphical Models Meet Bandits: A Variational Thompson Sampling Approach. In Proc. of the 37th International Conference on Machine Learning, PMLR 119 (2020)
26. F. M. Harper, J. A. Konstan, The movielens datasets: History and context. ACM Trans. Interact. Intell. Syst. (TiiS) 5, 4 (2015), Article 19