Decision-Making in Evolving Environments: A Bayesian Multi-Agent Bandit Framework

Extended Abstract

Mohammad Essa Alsomali
Lancaster University
Lancaster, United Kingdom
m.alsomali@lancaster.ac.uk

Leandro Soriano Marcolino
Lancaster University
Lancaster, United Kingdom
l.marcolino@lancaster.ac.uk

Barry Porter
Lancaster University
Lancaster, United Kingdom
b.f.porter@lancaster.ac.uk

Roberto Rodrigues-Filho
Federal University of Santa Catarina
Santa Catarina, Brazil
roberto.filho@ufsc.br

ABSTRACT

We introduce *DAMAS (Dynamic Adaptation through Multi-Agent Systems)*, a novel framework for decision-making in non-stationary environments characterized by varying reward distributions and dynamic constraints. Our framework integrates a multi-agent system with Multi-armed Bandits (MAB) algorithms and Bayesian updates. Each agent in DAMAS specializes in a particular environmental state. The system employs Bayesian estimation to continuously update the probabilities of being in each environmental state, enabling rapid adaptation to changing conditions. Our evaluation of DAMAS included both synthetic environments and real-world web server workloads.

KEYWORDS

Multi-Armed Bandits; Decision-Making; Multi-Agent Systems; Bayesian Inference

ACM Reference Format:
Mohammad Essa Alsomali, Leandro Soriano Marcolino, Barry Porter, and Roberto Rodrigues-Filho. 2025. Decision-Making in Evolving Environments: A Bayesian Multi-Agent Bandit Framework: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

In today's technology-driven world, adaptive decision-making systems face the critical challenge of maintaining optimal performance in non-stationary environments. From industrial automation to financial trading and web services [8, 11], these systems must process data and respond quickly while adapting to changing conditions. For instance, web servers experience dynamic and unpredictable real-time workloads, where static configurations often fail to maintain optimal performance under varying demand patterns [1].

The exploration-exploitation trade-off remains fundamental in managing such systems, particularly in sequential decision-making and reinforcement learning [6, 12]. Recent advances in this field have produced various approaches to address non-stationary environments. Cavenaghi et al. [3] proposed a concept drift-aware algorithm for non-stationary multi-armed bandits, demonstrating improved performance in detecting and adapting to changes. Studies have demonstrated adaptive decision-making using techniques such as chaotic semiconductor lasers for dynamically changing reward environments [9] and distributed consensus algorithms for multi-agent multi-armed bandits in dynamic settings [5].

While existing solutions show promise, they often face limitations in managing the combination of dynamic constraints and fluctuating reward distributions. For example, many relevant algorithms for non-stationary environments, such as Sliding-Window UCB [7] and Sliding-Window TS [13], assume uniform rewards between 0 and 1, or -1 and 1. These approaches struggle when rewards fall outside these ranges or when the range itself changes over time. Bayesian optimization has emerged as a powerful method of adaptive learning [4], particularly in scenarios where uncertainty plays an important role, but its integration with multi-agent systems for dynamic environments remains underexplored.

To address these challenges, we propose DAMAS, a novel framework that integrates multi-armed bandits (MAB) algorithms with Bayesian updates through a multi-agent system. Our approach seamlessly combines the MAB algorithm's ability to balance exploration and exploitation with Bayesian updates for environmental state estimation. The key contributions of this paper are: (i) A framework for dynamic environments that integrates MAB algorithms with Bayesian updates to estimate the current environment and update the Q-values based on the current uncertainty. (ii) An effective mechanism for handling varying reward ranges through multiple Q-values and specialized agents. (iii) A Bayesian Optimization approach for hyper-parameter adjustment that accounts for uncertainty in agent performance across different environments.

2 METHODOLOGY

DAMAS addresses decision-making in dynamic environments where conditions change over time, represented by a set of environments $\mathbf{E} = \{e_1, e_2, e_3, ..., e_n\}$. Each environment e_i is characterized by means $\mu_i(a)$ and standard deviations $\sigma_i(a)$ for rewards associated with different actions $a \in \mathbf{A}$, where $\mathbf{A}$ is the set of possible actions.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

Framework Overview: The core of DAMAS consists of a multi-agent system $\Phi = \{\phi_1, \phi_2, ..., \phi_n\}$, where each agent ϕ_i corresponds to environment e_i. Each agent maintains its Q-values $Q(\phi, a)$ and employs a modified UCB1 algorithm for action selection: $a^* = \underset{a}{\text{argmax}} \left(Q(\phi, a) + c\sqrt{\frac{2\log(t)}{N(\phi,a)}} \right)$, where a^* is the selected action, $Q(\phi, a)$ is the estimated Q-value for action a by agent ϕ, t is the total number of trials, $N(\phi, a)$ is the number of times action a has been selected by agent ϕ, and c controls exploration. The final action is sampled based on the probability $P(e_i)$.

Dynamic Adaptation Process: The system operates through three key mechanisms: (i) Action Selection and Q-value Updates: where actions are sampled based on environmental probabilities $P(e_i)$, and Q-values are updated for all agents using:

$$\begin{aligned} S(\phi, a_t) &\leftarrow S(\phi, a_t) + P(e_\phi) \cdot r_t \\ N(\phi, a_t) &\leftarrow N(\phi, a_t) + P(e_\phi) \\ Q(\phi, a_t) &\leftarrow S(\phi, a_t)/N(\phi, a_t) \end{aligned} \tag{1}$$

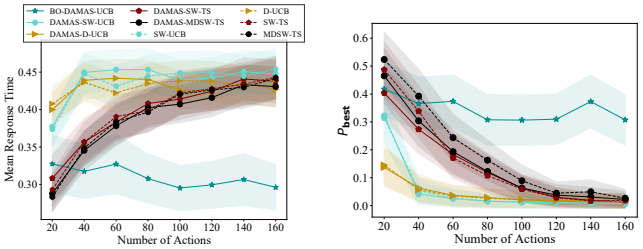
where $S(\phi_i, a_t)$ maintains the weighted sum of rewards, and $N(\phi_i, a_t)$ tracks the effective number of times action a_t has been selected by agent ϕ_i. The weighting by $P(e_i)$ ensures that updates are proportional to the system's confidence in each environment. (ii) Environmental Probability Updates: The environmental probability update mechanism employs a Bayesian approach. For each environment e_i, the likelihood of observing reward r_t is computed using a Gaussian probability density function: $P(r_t|e_i) = \mathcal{N}(r_t; \tilde{\mu}_i(a_t), \tilde{\sigma}_i(a_t)^2)$, and then using the Bayes' theorem to obtain the posterior probabilities of being in each environment at iteration t, given the observed reward r_t: $P(e_i|r_t) \propto P(r_t|e_i) \cdot P(e_i)$. (iii) Bayesian Optimization for Parameter Tuning: Employs a Gaussian process model for reward dynamics, uses the Lower Confidence Bound (LCB) acquisition to minimize cost, and updates parameters across all agents.

3 EXPERIMENTAL RESULTS

We evaluated DAMAS through both synthetic and real-world experiments, comparing it against state-of-the-art approaches including *UCB1* [2], Sliding-Window UCB (*SW-UCB*) and Discounted UCB (*D-UCB*) [7], and Thompson Sampling variants: Sliding-Window (*SW-TS*), and Mean Discounted Sliding-Window (*MDSW-TS*) [3, 13]. Moreover, we compare multiple DAMAS-enhanced configurations directly against their original MAB algorithms. The DAMAS variants include: (i) *DAMAS-UCB* a multi-agent system that uses *UCB1* with fixed hyper-parameters (c), (ii) *BO-DAMAS-UCB* extends *DAMAS-UCB* by incorporating Bayesian optimisation to adapt hyper-parameters dynamically, and (iii) *DAMAS-SW-UCB*, *DAMAS-D-UCB*, *DAMAS-SW-TS*, and *DAMAS-MDSW-TS*, which applies the DAMAS framework to different MAB algorithms.

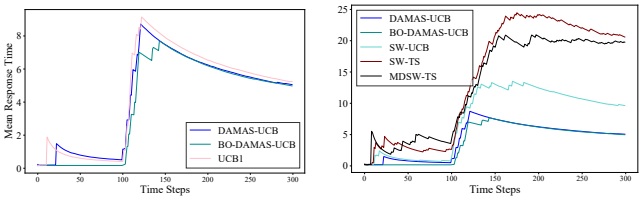
Experimental Setup: Our experiments were tested under different scenarios such as abrupt and incremental changes, here we present a sample of our results for abrupt transition scenarios (Figure 1).

Key Results: Performance Metrics: (i) Mean Response Time (MRT): DAMAS achieved around 30% reduction (Figure 1a) compared to baseline methods; (ii) Probability of Best Action Selection (P_{best}): Up to 50% improvement (Figure 1b) over traditional approaches.



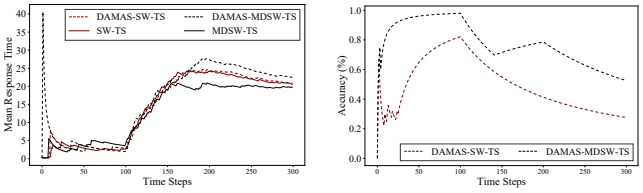
(a) Average performance **(b) Probability of selecting best action**

Figure 1: Comparing best-performing approach for Sudden Change across multi-agent MAB algorithms, including BO-DAMAS-UCB.



(a) Average performance (our methods and *UCB1*) **(b) Average performance (our methods and other MAB)**

Figure 2: Average response time of Multi-Agent Approaches and baseline agent under three real workloads.



(a) Average performance **(b) Probability of accurately identifying the current environment**

Figure 3: Average response time of *DAMAS-SW-TS* vs *SW-TS* and *DAMAS-MDSW-TS* vs *MDSW-TS* and their accuracy.

Real-world web server experiments: For real-world web server environments [10] with diverse request types, including large image files unsuitable for caching or compression, text files amenable to compression, and cacheable image files, featuring sudden changes between distinct workload scenarios, each lasting 100 time steps. *DAMAS-UCB* and *BO-DAMAS-UCB* consistently maintain the lowest average response times in all workload scenarios (Figure 2). In addition, as shown in Figure 3, there are mixed results for DAMAS enhancements, for example, *DAMAS-SW-TS* shows similar performance compared to *SW-TS*, while *DAMAS-MDSW-TS* outperformed *MDSW-TS* for the first 100 steps (first environment), which may be based on its effectiveness in environmental identification (around 95%).

REFERENCES

[1] Rodrigo Araújo and Reid Holmes. 2021. Lightweight self-adaptive configuration using machine learning. In *Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering*. 133–142.

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2 (2002), 235–256.

[3] Emanuele Cavenaghi, Gabriele Sottocornola, Fabio Stella, and Markus Zanker. 2021. Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm. *Entropy* 23, 3 (2021), 380.

[4] Lei Cheng, Feng Yin, Sergios Theodoridis, Sotirios Chatzis, and Tsung-Hui Chang. 2022. Rethinking Bayesian learning for data analysis: The art of prior and inference in sparsity-aware modeling. *IEEE Signal Processing Magazine* 39, 6 (2022), 18–52.

[5] Xiaotong Cheng and Setareh Maghsudi. 2024. Distributed consensus algorithm for decision-making in multi-agent multi-armed bandit. *IEEE Transactions on Control of Network Systems* (2024).

[6] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. 2020. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189* (2020).

[7] Aurélien Garivier and Eric Moulines. 2008. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415* (2008).

[8] Zhongming Liu, Hang Luo, Peng Chen, Qibin Xia, Zhihao Gan, and Wenyu Shan. 2022. An efficient isomorphic CNN-based prediction and decision framework for financial time series. *Intelligent Data Analysis* 26, 4 (2022), 893–909.

[9] Akihiro Oda, Takatomo Mihana, Kazutaka Kanno, Makoto Naruse, and Atsushi Uchida. 2022. Adaptive decision making using a chaotic semiconductor laser for multi-armed bandit problem with time-varying hit probabilities. *Nonlinear Theory and Its Applications, IEICE* 13, 1 (2022), 112–122.

[10] Barry Porter, Matthew Grieves, Roberto Rodrigues Filho, and David Leslie. 2016. *REX*: A development platform and online learning approach for runtime emergent software systems. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 333–348.

[11] Claudio Scordino, Ida Maria Savino, Luca Cuomo, Luca Miccio, Andrea Tagliavini, Marko Bertogna, and Marco Solieri. 2020. Real-time virtualization for industrial automation. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Vol. 1. IEEE, 353–360.

[12] Haobin Shi and Meng Xu. 2019. A multiple-attribute decision-making approach to reinforcement learning. *IEEE Transactions on Cognitive and Developmental Systems* 12, 4 (2019), 695–708.

[13] Francesco Trovo, Stefano Paladino, Marcello Restelli, and Nicola Gatti. 2020. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research* 68 (2020), 311–364.