Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Advancing Decision-Making in Dynamic Environments

Xinyan Hu*

College of computer science and technology, Zhejiang University, Hangzhou, China

* Corresponding Author Email: 3220101845@zju.edu.cn

Abstract. This paper introduces the Adaptive Thompson Sampling (AdaTS) algorithm, specifically designed for the challenges posed by opportunistic bandit problems. Unlike conventional multi-armed bandit scenarios where the optimality of actions remains relatively static, opportunistic bandit environments are significantly influenced by fluctuating external conditions. These variations necessitate a highly adaptive strategy for decision-making. The AdaTS algorithm meets this requirement through a novel integration of real-time system load assessments, which dynamically adjust the balance between exploration and exploitation. This method not only ensures more effective adaptation to changing conditions but also considerably enhances performance metrics. By incorporating system load into the decision-making process, AdaTS effectively reduces regret and maximizes reward, particularly under scenarios of variable load intensities. This is particularly evident in environments characterized by binary-valued loads and stochastic reward distributions. The results highlight AdaTS's robustness and efficiency, making it a promising approach for complex environments where adaptability is crucial. This research marks a significant advancement in the field of adaptive algorithms for opportunistic bandit problems.

Keywords: Thompson Sampling algorithm; Opportunistic bandit; Adaptive method.

1. Introduction

In the context of multi-armed bandits (MABs), the stationary stochastic bandit model serves as a fundamental framework. Within this model, each arm of the bandit yields a reward derived from a fixed probability distribution that remains constant over time [1]. These distributions are unknown to the player, who is tasked with navigating the trade-off between exploring various arms to ascertain their reward distributions (exploration) and leveraging the arm known to provide the highest returns (exploitation)..

In stationary stochastic bandits, each arm k of the bandit has an associated reward distribution with a constant expected value u_k. When an arm is pulled, the reward received is a sample from this stationary distribution [2]. The goal of the player is to maximize cumulative rewards over a series of trials, or alternatively, to minimize the regret associated with not always pulling the optimal arm, where the optimal arm k^* is the one with the highest expected reward u^*.

While the assumption of constant regret for pulling suboptimal arms is valid in many traditional applications of stationary stochastic bandits, there are numerous real-world scenarios where the actual regret can vary depending on external conditions. For example, in dynamic market environments, the opportunity cost of choosing a less profitable business strategy might increase during peak demand periods [3]. Similarly, in automated trading systems, the financial regret of not executing the optimal trade could vary with fluctuating market conditions.

By incorporating these dynamic aspects, we can better understand and develop strategies for MABs that are more reflective of complex, real-world decision-making scenarios [4]. These adaptations make the foundational concept of stationary stochastic bandits applicable to a broader range of practical situations where external factors play a crucial role in determining the actual regret of decisions.

Contribution: Adaptive Thompson Sampling Algorithm. We introduce a novel variant of the traditional Thompson Sampling algorithm that dynamically adjusts its exploration-exploitation

balance based on real-time load assessments [5]. This adaptation is designed to respond to the fluctuating demands and constraints inherent in dynamic systems, making it particularly suited for environments where load levels directly impact decision efficiency and outcomes.

At the core of our approach is a method for adjusting the variance of the sampling distribution based on the current system load. This method significantly deviates from traditional static approaches by introducing a load-sensitive modulation factor, which enhances the algorithm's responsiveness to environmental changes. This is a crucial advancement for applications in which operational conditions can vary unpredictably.

Through simulations and real-world case studies, we demonstrate that Adaptive Thompson Sampling outperforms conventional algorithms like UCB and traditional Thompson sampling algorithm in scenarios characterized by variable loads [6]. The results highlight not only improvements in reward maximization but also reductions in regret, validating the algorithm’s effectiveness in leveraging available information under changing conditions.

2. Related Work

2.1. Detailed examination of Multi-Armed Bandit (MAB) Problems

The Multi-Armed Bandit (MAB) framework, originally conceptualized by Auer, Cesa- Bianchi, and Fischer in 2002, tackles the sequential decision-making challenge inherent in balancing the exploration-exploitation dilemma. Within this framework, a learner must choose one action from a limited set of possible actions—referred to as arms—during each round within a predefined finite period [7]. The learner only sees the outcome or reward of the selected action. The primary objective is to identify and consistently select the best-performing arm to maximize the overall expected reward and minimize the regret. Regret is defined as the difference between the rewards that would have been achieved by an ideal strategy, which always selects the best arm throughout the time period, and the rewards actually accrued by the MAB strategy employed.

2.2. Analysis of Traditional Thompson Sampling

Thompson Sampling, also known as posterior sampling and probability matching, was originally proposed by W. R. Thompson in 1933 to tackle two-armed bandit problems in clinical trials [8]. Despite its initial introduction, the algorithm received little attention in academic literature for decades, only being rediscovered sporadically as an effective heuristic (Wyatt, 1997; Strens, 2000). It wasn't until the early 21st century that Thompson Sampling began to gain significant recognition, spurred by influential articles in 2010 and 2011 by Scott and Chapelle & Li, respectively, which demonstrated its robust empirical performance. This renewed interest has led to a rapid expansion of literature exploring its capabilities and enhancements over the past decade.

In recent years, Thompson Sampling has evolved significantly, particularly in its application to dynamic and complex environments. Recent adaptations have focused on improving the algorithm's efficiency and effectiveness in such settings [9]. Notably, the introduction of the Feel-Good TS by Zhang in 2022, which adopts a more aggressive exploration strategy, and the geometry-aware TS by Luo and Bayati in 2023, which achieves a minimax optimal regret, illustrate the ongoing advancements in refining the algorithm for more sophisticated scenarios. These developments highlight the algorithm's adaptability and its enhanced ability to handle the exploration-exploitation trade-off in dynamically changing environments.

The versatility of Thompson Sampling has led to its wide application across various industries and domains [10]. It has been successfully implemented in fields ranging from revenue management and marketing to more technical areas like Monte Carlo tree search and hyperparameter tuning. Major companies such as Adobe, Amazon, Facebook, Google, LinkedIn, Microsoft, Netflix, and Twitter have utilized Thompson Sampling to optimize a range of functions, including internet advertising, website optimization, and recommendation systems [11]. The broad adoption underscores the

algorithm's practical utility and effectiveness in real-world applications, solving complex, stochastic problems across the digital landscape.

2.3. Opportunistic bandit

In the realm of sequential decision-making, the concept of opportunistic bandits has emerged as an influential extension of the traditional stationary stochastic multi-armed bandit (MAB) framework. It was raised by Wu, Guo, Liu in 2018. Distinguished by its focus on environments where the optimality of actions can vary dynamically with external conditions, the opportunistic bandit model captures real-world scenarios where decision-making is subject to shifting contexts [12]. Unlike classical MAB problems, opportunistic bandits consider the variability in potential rewards that are influenced by exogenous factors, often modeled as a "load" of the system. Also, unlike many dynamic scenarios, the opportunistic bandit combines stationary factors and nonstationary factors. The reward distribution of pulling a specific arm remains static, but the actual reward depends on dynamic environment determined by load [13]. Which means the opportunistic bandit distinguishes the actual reward from the instant reward of pulling an arm. This variation introduces a layer of complexity, as the algorithms need to adapt not only to the inherent uncertainty of the reward distributions of the arms but also to the external changes that affect these distributions. The introduction of this model addresses a gap in the traditional MAB setup by acknowledging that in many practical applications. By incorporating the dynamics of environmental states, opportunistic bandit algorithms aim to optimize decision-making in a more nuanced and realistically variable context, thereby enhancing the applicability and effectiveness of bandit solutions in dynamic settings.

3. The Adaptive Thompson Sampling Framework

3.1. System model of Opportunistic bandit

We explore a K-armed stochastic bandit model where the exploration cost is influenced by a fluctuating external factor termed "load". In this setup, each of the K arms offers a random reward $X_{k,t}$ at time t, with $X_{k,t}$ ranging between 0 and 1. These rewards are independent across the arms and identically distributed over time, each with an expected mean u_k. The arm that consistently yields the highest average reward, u^*, is identified as the optimal arm k^*.

The model introduces a dynamic component, L_t, representing the load at each time t, with values also constrained between 0 and 1. Before deciding which arm to pull, the agent is informed of the current load L_t The decision A_t on which arm to pull is based not only on L_t but also on past observations, encapsulated in the history H_{t-1}, which includes sequences of past loads, chosen arms, and the rewards received.

The actual reward $\widetilde{X_t}$ obtained from pulling an arm depends on both the load and the reward of the arm.

$$\widetilde{X_t} = L_t * X_{A_t,t} \tag{1}$$

Although the load does not affect the intrinsic value of $X_{A_t,t}$ once an arm is chosen, it directly impacts the actual reward received. After pulling an arm, the agent gains access to the nominal reward $X_{A_t,t}$ for that pull, allowing them to update their strategy based on both the reward outcomes and the variations in load. This introduces a layer of complexity as the agent must adapt their strategy to optimally balance the immediate cost of exploration influenced by the load with the potential long-term gain from the rewards.

3.2. Adaptive thompson sampling

A basic version of the traditional Thompson Sampling (TS) algorithm employs a Bayesian approach to address the exploration-exploitation dilemma, allowing for effective sequential decision-making. This method models the uncertainty of each action's rewards through probability distributions and

selects the action associated with the highest sample in each iteration. This strategy ensures that actions with higher uncertainty and potentially higher rewards are explored, thus balancing the need to exploit known rewards with the exploration of lesser-known options. As shown in Table 1 and 2. The choice to use Gaussian (normal) distributions in Thompson Sampling is driven by several of their statistical properties, which align well with the requirements of decision-making under uncertainty:

Central Limit Theorem: Gaussian distributions naturally arise as the limit of many random processes, which makes them a natural choice for modeling a wide range of real-world processes. The exact distribution of outcomes is often unknown, but can be approximated as normally distributed due to the aggregation of many small independent effects.

Conjugacy: In Bayesian statistics, the Gaussian distribution is a conjugate prior for many types of likelihoods, including other Gaussians. This conjugacy ensures that the posterior distributions also remain Gaussian when the prior and likelihood are Gaussian, simplifying the analytical computation of the posterior.

Computational Simplicity: The mathematical properties of Gaussian distributions, such as having a well-defined mean and variance, facilitate efficient computation. In a Bayesian framework, updates involve simple arithmetic on the parameters (mean and variance), which is computationally efficient and numerically stable.

Table 1. Thompson Sampling

Algorithm 1: Thompson Sampling

1: Init:
prior cumulative distribution function $\boldsymbol{F_1(1), \dots, F_k(1)}$**;**
counts $\boldsymbol{C_k(t) = 1}$**;**
variance $\boldsymbol{\sigma_0^2}$**;**
2: For t = 1 to T do
3: **for each arm i do**
4: $\boldsymbol{\sigma_i^2 = \frac{\sigma_0^2}{C_i(t)}}$;
5: $F_i(t) \sim N(\hat{u}_i(t), \sigma_i^2)$**;**
6: **Sample** $\boldsymbol{\theta_i(t) \sim F_i(t)}$ **independently;**
7: **end for**
8: **Pull the arm with the largest** $\boldsymbol{\theta_i(t)}$
$\boldsymbol{A_t = argmax\, \theta_i(t)}$**;**
$\boldsymbol{1 \le k \le K}$
9: **Observe** $\boldsymbol{X_t}$ **and update** $\boldsymbol{C_t}$ **,** $\boldsymbol{X_t}$ **, and** $\boldsymbol{u(t)}$ **for** $\boldsymbol{A_t}$**:**
10: end for

Table 2. Adaptive Thompson Sampling for opportunistic bandit

Algorithm 2: Adaptive Thompson Sampling for opportunistic bandit

1: Init:
prior cumulative distribution function $\boldsymbol{F_1(1), \dots, F_k(1)}$**;**

counts $C_k(t) = 1$;
variance $\boldsymbol{\sigma_0^2}$**;**

2: For t = 1 to T do
3: Observe $\boldsymbol{L_t}$**;**
4: for each arm i do
5: $\boldsymbol{\sigma_i^2 = \frac{\sigma_0^2}{C_i(t)*(1+\alpha*\widetilde{L_t})}}$**;**
6: $F_i(t) \sim N(\hat{u}_i(t), \sigma_i^2)$**;**
7: Sample $\boldsymbol{\theta_i(t) \sim F_i(t)}$ **independently;**
8: end for
9: Pull the arm with the largest $\boldsymbol{\theta_i(t)}$
$\boldsymbol{A_t = argmax\ \theta_i(t)}$**;**
$$\boldsymbol{1 \le k \le K}$$
10: Observe $\boldsymbol{X_t}$ **and update** $\boldsymbol{C_t}$ **,** $\boldsymbol{X_t}$ **, and** $\boldsymbol{u(t)}$ **for** $\boldsymbol{A_t}$**:**
11: end for

In this study, we upgrade the traditional TS algorithm and introduce a novel AdaTS algorithm specifically tailored for opportunistic bandits. To effectively handle the variations in the load level L_t, we initially standardize L_t to fall within the range [0, 1] using the formula:

$$\widetilde{L_t} = \frac{[L_t]_{l(-)}^{l(+)} - l(-)}{l(+) - l(-)} \tag{2}$$

In the formula, $l(+)$ and $l(-)$ represent the lower and upper bounds used to truncate the load level, respectively. This normalization of L_t makes it easier to tune and set thresholds for algorithm parameters that depend on load levels, such as the coefficients controlling the exploration-exploitation trade-off. Without normalization, these parameters would need constant adjustment to cater to the fluctuating scale of load values, complicating the algorithm's design and its operational efficiency.

The algorithm begins with the initialization of prior cumulative distribution functions $F_i(1)$.
for each arm, setting initial counts $C_k(t)$ to 1, and defining a baseline variance σ_0^2. This setup ensures that every arm starts with a uniform state of knowledge and exploration opportunity, which is crucial for unbiased initial assessment.

At the core of the algorithm's adaptability is the dynamic adjustment of the sampling variance for each arm based on both the count of selections $C_i(t)$ and the normalized $\widetilde{L_t}$. The variance for each arm is calculated as:

$$\sigma_i^2 = \frac{\sigma_0^2}{C_i(t)*(1+\alpha*\widetilde{L_t})} \tag{3}$$

This formulation implies that the variance decreases with the number of times an arm is pulled, which is typical in Thompson Sampling to reduce exploration as more is learned about each arm. However, uniquely, the variance is further modulated by the load level $\widetilde{L_t}$, scaled by a factor α. The intuition of this method is very simple. Higher load levels, indicating possibly more critical or busy system states, reduce the variance more significantly, pushing the algorithm towards exploitation of known good arms to ensure stability and efficiency under stress.

For each round and each arm, the algorithm samples from a normal distribution characterized by the current estimate of the mean reward $\hat{u}_i(t)$ and the dynamically adjusted variance σ_i^2 . This step embodies the exploration-exploitation trade-off, as the sampled values $\theta_i(t)$ may encourage trying

less frequently chosen arms if their sampled values turn out high despite fewer selections. The arm with the highest sample $\theta_i(t)$ is selected for pulling, which aligns with the core principle of Thompson Sampling to choose based on potential rather than just past rewards.

In conclusion, the Adaptive Thompson Sampling algorithm for Opportunistic Bandits offers a robust framework for decision-making in dynamic environments. It enhances traditional Thompson Sampling by incorporating load-dependent variance adjustments, thereby aligning decision-making more closely with current system states and operational demands. This method not only optimizes the immediate reward outcomes but also contributes to long-term system stability and responsiveness, making it a valuable strategy in practical applications facing variable loads.

4. Performance Analysis

In our analysis, we delve into the complexities presented by environments with random binary-valued loads and random rewards. We define the load L_t to take on values from the set $\{\epsilon_0, 1-\epsilon_1\}$, with a probability $P\{L_t = \epsilon_0\} = \rho$, where ρ, lies strictly within the open interval (0, 1). Here, the rewards $X_{k,t}$ are assumed to be independently and identically distributed (i.i.d) random variables within the interval [0, 1], with the expected value $E[X_k, t] = u_k$.

As shown in figure 1, in this performance analysis, to simplify the situation, we set $\epsilon_0 = \epsilon_1 = 0$ and $\rho = 0.5$.

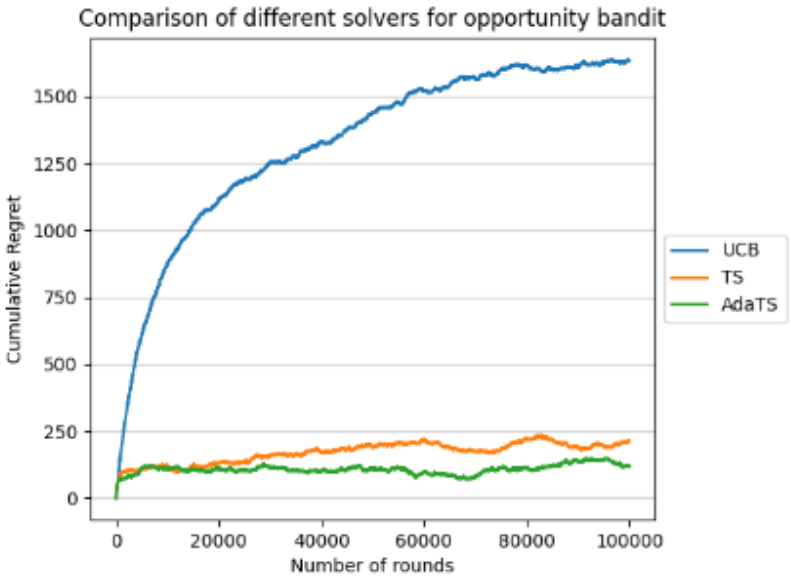

Figure 1. (Photo/Picture credit: Original).

Upon examining the performance graph, which compares the cumulative regret over time for three different solvers—UCB, traditional Thompson Sampling (TS), and the proposed AdaTS—several key observations emerge:

UCB Performance: The UCB algorithm shows a cumulative regret that increases significantly more than TS and AdaTS, suggesting that it is less effective in adapting to the dynamic conditions modeled by the binary-valued load and random rewards.

TS vs. AdaTS: Traditional Thompson Sampling maintains a lower cumulative regret than UCB, indicating a better balance of exploration and exploitation. However, it still accumulates regret steadily over time.

AdaTS Advantage: The Adaptive Thompson Sampling (AdaTS) demonstrates a distinct advantage, maintaining the lowest cumulative regret across the number of rounds. This is indicative of its superior ability to adapt to the varying load conditions. The relatively flat curve of the AdaTS in

comparison with the other two methods suggests that it quickly converges to the optimal arm and makes better decisions under uncertainty, reflecting the benefits of dynamically adjusting the algorithm's parameters based on the observed load.

Stability and Efficiency: Notably, the AdaTS curve shows the least variability, implying that it is the most stable and efficient among the tested algorithms, particularly in an environment with binary-valued loads. Its ability to perform well in such a setting aligns with the proposed methodology, which is designed to handle random load fluctuations and reward distributions effectively.

In summary, the plot clearly illustrates the efficacy of AdaTS in minimizing regret in opportunistic bandit scenarios, outperforming traditional methods under the conditions of random binary-valued load and random rewards. The results reinforce the potential of AdaTS as a robust solver for dynamic and uncertain environments, validating the theoretical framework presented in this paper.

5. Application to Real-World Problems

The versatility of the AdaTS algorithm makes it applicable to a myriad of real-world problems where system conditions dynamically influence decision-making efficacy. From optimizing traffic flow in smart cities during fluctuating volumes to managing inventory during sudden changes in consumer demand, AdaTS proves particularly useful. Its real-time adaptability also extends to online platforms for personalizing content delivery based on user activity spikes and dips, ensuring that the optimal choice aligns with current system states and user engagement levels.

6. Discussion

Our findings suggest that AdaTS offers substantial improvements over traditional methods, adapting seamlessly to environmental changes. Future research may explore the potential of AdaTS in more complex scenarios, including those with non-binary load states and multi-dimensional reward structures and analyze its regret under different load structure. Additionally, while AdaTS demonstrates remarkable stability, the exploration of parameter optimization and algorithmic tuning in line with specific industry demands could further enhance its practical application.

7. Conclusion

The Adaptive Thompson Sampling (AdaTS) algorithm marks a substantial innovation in decision-making strategies for dynamic and stochastic environments. By adeptly incorporating system load metrics into the exploration-exploitation calculus, AdaTS demonstrates superior performance over traditional algorithms in opportunistic bandit scenarios. This enhancement confirms its viability as a robust tool for both academic inquiry and practical application. The adaptability and resilience of AdaTS are particularly suited to the evolving demands of contemporary applications, heralding a promising avenue for future advancements in adaptive decision-making methodologies. Its effectiveness in adapting to variable system conditions showcases the potential to drive significant improvements in a variety of fields, including finance, healthcare, and automated systems, where decision-making must adjust swiftly to changing environmental factors. This algorithm not only extends the theoretical framework of multi-armed bandits but also sets a new benchmark for the practical deployment of learning algorithms in complex real-world scenarios.

References

[1] H. Wu, X. Guo, and X. Liu, “Adaptive Exploration-Exploitation Tradeoff for Opportunistic Bandits,”.

[2] Chapelle and L. Li, “An Empirical Evaluation of Thompson Sampling,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2011. Accessed: Apr. 19, 2024. [Online]. Available: https:// proceedings. neurips.cc/paper/2011/hash/e53a0a2978c28872a4505bdb51db06dc-Abstract.html.

[3] S. Agrawal and N. Goyal, “Analysis of Thompson Sampling for the Multi-armed Bandit Problem,” in Proceedings of the 25th Annual Conference on Learning Theory, JMLR Workshop and Conference Proceedings, Jun. 2012, p. 39.1-39.26. Accessed: Apr. 19, 2024. [Online]. Available: https://proceedings.mlr.press/v23/agrawal12.html.

[4] Gopalan, S. Mannor, and Y. Mansour, “Thompson Sampling for Complex Online Problems,” in Proceedings of the 31st International Conference on Machine Learning, PMLR, Jan. 2014, pp. 100–108. Accessed: Apr. 19, 2024. [Online]. Available: https://proceedings.mlr.press/v32/gopalan14.html.

[5] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, “A Tutorial on Thompson Sampling,” MAL, vol. 11, no. 1, pp. 1–96, Jul. 2018, doi: 10.1561/2200000070.

[6] F. Trovo, S. Paladino, M. Restelli, and N. Gatti, “Sliding-Window Thompson Sampling for Non-Stationary Settings,” Journal of Artificial Intelligence Research, vol. 68, pp. 311–364, May 2020, doi: 10.1613/jair.1.11407.

[7] T. Lattimore and C. Szepesvári, Bandit Algorithms, 1st ed. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

[8] Q. Zhu and V. Tan, “Thompson Sampling Algorithms for Mean-Variance Bandits,” in Proceedings of the 37th International Conference on Machine Learning, PMLR, Nov. 2020, pp. 11599–11608. Accessed: Apr. 11, 2024. [Online]. Available: https://proceedings.mlr.press/v119/zhu20d.html

[9] E. Cavenaghi, G. Sottocornola, F. Stella, and M. Zanker, “Non Stationary Multi-Armed Bandit: Empirical Evaluation of a New Concept Drift-Aware Algorithm,” Entropy, vol. 23, no. 3, p. 380, Mar. 2021, doi: 10.3390/e23030380.

[10] B. Kveton et al., “Meta-Thompson Sampling,” in Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, pp. 5884–5893. Accessed: Apr. 11, 2024. [Online]. Available: https://proceedings.mlr.press/v139/kveton21a.html.

[11] Y. Liu, B. V. Roy, and K. Xu, “Nonstationary Bandit Learning via Predictive Sampling,” in Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, PMLR, Apr. 2023, pp. 6215–6244. Accessed: Apr. 11, 2024. [Online]. Available: https://proceedings.mlr.press/v206/liu23e.html.

[12] T. Jin, X. Yang, X. Xiao, and P. Xu, “Thompson Sampling with Less Exploration is Fast and Optimal,” in Proceedings of the 40th International Conference on Machine Learning, PMLR, Jul. 2023, pp. 15239–15261. Accessed: Apr. 11, 2024. [Online]. Available: https://proceedings.mlr.press/v202/jin23b.html.

[13] R. Xu, Y. Min, and T. Wang, “Noise-Adaptive Thompson Sampling for Linear Contextual Bandits,” Advances in Neural Information Processing Systems, vol. 36, pp. 23630–23657, Dec. 2023.