Article

Modified Index Policies for Multi-Armed Bandits with Network-Like Markovian Dependencies

Abdalaziz Sawwan [1,*] and Jie Wu [2]

[1] Temple University; sawwan@temple.edu
[2] Temple University; jiewu@temple.edu
* Correspondence: sawwan@temple.edu

Abstract: Sequential decision-making in dynamic and interconnected environments is a cornerstone of numerous applications, ranging from communication networks and finance to distributed blockchain systems and IoT frameworks. The multi-armed bandit (MAB) problem is a fundamental model in this domain that traditionally assumes independent and identically distributed (iid) rewards, which limits its effectiveness in capturing the inherent dependencies and state dynamics present in some real-world scenarios. In this paper, we lay a theoretical framework for a modified MAB model in which each arm's reward is generated by a hidden Markov process. In our model, each arm undergoes Markov state transitions independent of play in a way that results in varying reward distributions and heightened uncertainty in reward observations. The number of states for each arm can be up to three states. A key challenge arises from the fact that the underlying states governing each arm's rewards remain hidden at the time of selection. To address this, we adapt traditional index-based policies and develop a modified index approach tailored to accommodate Markovian transitions and enhance selection efficiency for our model. Our proposed proposed Markovian Upper Confidence Bound (MC-UCB) policy achieves logarithmic regret. Comparative analysis with the classical UCB algorithm reveals that MC-UCB consistently achieves approximately a 15% reduction in cumulative regret. This work provides significant theoretical insights and lays a robust foundation for future research aimed at optimizing decision-making processes in complex, networked systems with hidden state dependencies.

Keywords: Dynamic distributions; learning theory; Markov chain; multi-armed bandit.

1. Introduction

Decision-making in environments with network-like dependencies presents a fundamental challenge across various fields, including communication networks, finance, and complex distributed systems [1–4]. In such environments, a decision-maker faces interconnected structures where actions taken on one element may influence the states or rewards of others, thereby creating dynamic dependencies reminiscent of those found in networked systems. Examples of such networks can be found in resource allocation across multiple communication channels in IoT (Internet of Things) sensor networks [5], throughput optimization in distributed blockchain ecosystems [6], adaptive QoS (Quality of Service) management in communication networks [7], and security or intrusion detection frameworks in large-scale system administration scenarios [8]. In these contexts, the multi-armed bandit (MAB) problem, where a player repeatedly selects among multiple uncertain options (arms), becomes more intricate due to underlying and often hidden state transitions that evolve over time.

The original classical MAB formulation, introduced by Robbins [9,10], assumes that each arm's reward distribution remains fixed and independent over time. However, in networked scenarios, these assumptions rarely hold: the reward distributions may shift due to underlying Markovian state transitions that are hidden from the decision-maker [11]. Arms in such a scenario can represent network nodes, communication links, or distributed

Citation: Sawwan, A.; Wu, J. Modified Index Policies for Multi-Armed Bandits with Network-Like Markovian Dependencies. *Network* **2024**, *1*, 1–20. https://doi.org/

Received:
Accepted:
Published:

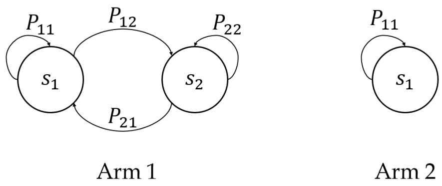

Figure 1. A sample example of two-arms multi-armed bandit. The first arm has two states and the second arm has one state.

resources whose performance and reliability evolve with time. The agent must continually learn and adapt, taking into account latent transitions that are reminiscent of evolving network conditions.

In this paper, we lay a theoretical framework for a modified MAB model in which each arm's reward is generated by a hidden Markov process. This approach models the type of network-like dependencies found, for example, in dynamic IoT sensor networks—where channel conditions and sensor states change stochastically and are not directly observable—yet these state changes critically affect the rewards (e.g., reliable data transmission or efficient resource utilization). Each arm in our model can transition among up to three states, each associated with a different reward distribution, regardless of whether the arm is played. The result is a problem setting that demands sophisticated exploration-exploitation strategies that identify the best arms under evolving conditions and also cope with underlying dynamics that reflect network interdependencies.

In this context, we evaluate the decision-maker's performance using the concept of regret, a metric that captures the cost of uncertainty in networked decision-making environments. Regret is defined as the difference between the expected reward an ideal policy—one with complete knowledge of all arm statistics or hindsight advantage—would achieve, and the reward achieved by the decision-maker's actual strategy. An ideal policy would consistently select the arm yielding the highest expected reward over time. This concept, commonly referred to as weak regret, is a central performance measure in uncertain decision problems, as highlighted by Auer *et al.* [12]. Our study focuses on regret, particularly within interconnected, network-like settings.

MAB problems with Markovian rewards significantly heighten complexity due to dynamic dependencies that reflect networked interactions. Here, each arm is modeled as a Markov process with a finite set of states, each linked to a unique reward distribution. The transition between states follows a known probability matrix, introducing a memory element into the decision process where rewards depend not only on the current choice but also on the hidden state of each arm [13–16]. This Markovian structure effectively simulates a network in which states and rewards are dynamically interdependent over time.

The state transitions are determined by predefined probabilities, yet the exact state of each arm remains hidden. This creates a layer of opacity similar to unobserved interactions in networked systems [17–19]. Consequently, the player must infer each arm's state from the history of observed rewards. This amplifies the challenge of the exploration-exploitation trade-off. The decision-maker faces a networked challenge: to exploit high-reward arms based on historical performance or to explore underused arms to reveal potential reward structures. Figure 1 illustrates an example of the problem and highlights the network-like dependencies across arms.

A core challenge in this interconnected framework is to develop strategies that effectively balance immediate rewards with potential future gains that could arise from transitioning into more advantageous states [20,21]. This networked trade-off between short-term exploitation and long-term exploration is not purely theoretical or network-related; it mirrors complex, real-world decision-making environments such as financial portfolio management or adaptive clinical trials where treatments impact outcomes over time [22–25].

In this work, we address these challenges by introducing a novel theoretical approach to the MAB problem with Markovian dynamics and network-like dependencies where each arm has up to three possible states. We adapt traditional index policies to account for the intricate structure of state transitions. Our focus is on refining these policies to achieve robust performance by attaining logarithmic regret even within the complex networked dynamics of hidden state transitions. We further compare our modified index-based policies with the classic upper confidence bound (UCB) algorithm. This study thus sets the stage for a deeper understanding of decision strategies within networked environments involving uncertainty and dynamic dependencies.

1.1. *Main findings*

This paper makes the following theoretical contributions:

- We demonstrate that for each arm, represented as an irreducible, finite-state, aperiodic, and reversible three-state Markov chain, simple sample mean-based index policies can achieve logarithmic regret uniformly over time, even in interconnected settings resembling networked dependencies.
- We simplify the analysis of state transition probabilities by modeling the arms as Markov chains with identical rewards that capture basic network-like structures in which transitions are dependent on state dynamics.
- We present a numerical comparison of the regret incurred by our sample mean-based index policy and evaluate its performance relative to other policies.

1.2. *Application Context and Conceptual Validation in Network-Like Scenarios*

While our primary contribution is theoretical, it is helpful to illustrate how this framework can be built on to conceptually extend to real network scenarios. Consider, for example, the following contexts:

- **Security [26]:** Arms may represent intrusion detection strategies whose efficacy varies as an adversary's tactics evolve over time. Each state transition corresponds to a shift in the threat environment. Our Markovian MAB framework can guide strategic decisions to maintain robust defense while learning dynamically about evolving threats.
- **Distributed Blockchain Systems [27]:** Nodes or shards in a blockchain network might yield variable validation rewards depending on their state of congestion or consensus participation. The Markovian structure models the dynamic nature of node availability and network conditions in a way that would help a node operator choose where to allocate resources or which shard to support over time.
- **QoS in Communication Networks [7]:** Network links may fluctuate between high-quality, moderate, and poor states due to changing traffic patterns. By representing each link as a Markovian arm, our framework can assist in selecting the best channel at any given time in order to balance the exploration of uncertain but potentially high-quality links with the exploitation of known reliable ones.
- **IoT and System Administration [28]:** IoT nodes or servers can transition between states that reflect varying processing loads or energy conditions. The Markovian MAB model helps a controller decide which node to query or utilize for computations, thereby maximizing long-term performance.

In sum, while this work is focused on the theoretical aspects and fundamental results for up to three states, it offers a roadmap for future empirical explorations and practical implementations. The stylized simulation experiments that we show later serve as a preliminary demonstration and show that the theoretical principles hold in a controlled synthetic environment, thus setting the stage for subsequent research aiming at more comprehensive benchmarking in real-world network contexts.

The remainder of the paper is structured as follows. Section 2 gives the related work. The problem is formally defined and presented in Section 3 presents the preliminaries.

Section 4 shows the problem formulation. The Index policy and its regret analysis are given in Section 5. Section 6 shows our numerical simulation results, and finally, Section 7 concludes the paper.

2. Related Work

The literature on MAB is vast and has evolved considerably from the original formulations focusing on independent and identically distributed (iid) reward processes. Early seminal work by Robbins and Lai and [9,10] established foundations for the iid case for certain known environments. Over time, researchers have explored a broad spectrum of MAB extensions that incorporate various forms of structure and dynamics. Notably, Markovian reward processes represent a key generalization and enable the modeling of scenarios where arm states—and thus rewards—evolve with memory and dependence on previous states.

Early explorations into Markovian bandits can be found in the work of Anantharam *et al.* [29], which analyzed index policies effective for arms governed by irreducible, finite-state, aperiodic Markov chains. Their approach demonstrated how arms with state-dependent rewards could still be tackled through index strategies that generalize the Gittins index concept [30]. While these studies set important precedents for handling Markovian structures, they often made simplifying assumptions, such as a single-parameter transition function or identical state spaces across arms. In contrast, our framework does not presume a single-parameter form for transition probabilities, nor does it require identical state spaces. By allowing each arm to transition among up to three states under distinct probability kernels, we offer a more flexible setting that can model diverse types of network dependencies.

Building upon this foundation, research has examined the problem of achieving low regret under more general conditions. Agrawal [31] and Auer *et al.* [32] established classical logarithmic regret results for iid settings. Their contributions included index and UCB-based strategies that guarantee optimal asymptotic and even uniformly logarithmic performance over time. They rely heavily on the iid assumption and do not directly address the complexities introduced by state transitions or network-like interdependencies. More recent works have begun to relax these assumptions. For instance, Garivier and Moulines [33] and Besbes *et al.* [34] considered bandit problems with non-stationary reward distributions in a way that captures some aspects of temporal dynamics without fully embracing Markovian state dependence. Such approaches typically rely on "resetting" or "sliding-window" techniques that do not directly exploit known Markovian transition structures.

In parallel, other authors have studied scenarios where multiple users or decision-makers interact with the same set of arms in network settings, leading to complex dynamics and collisions among players [35–37]. Here, the challenge lies in coordinating multiple agents to minimize interference and collectively achieve low regret. While such multi-player frameworks mirror network complexity, their primary focus is on handling concurrency and competition rather than modeling state evolution within each arm. Our approach differs by focusing explicitly on Markovian transitions at the arm level rather than strategic interactions among multiple decision-makers.

The distinction between rested and restless bandits further highlights the complexity in Markovian settings. In classical rested bandits, the state of an unplayed arm remains frozen until chosen again, as examined in works like Ortner [38] and Raj and Kalyani [39]. However, in restless bandits, arm states evolve regardless of selection, making the problem significantly more complex. Restless bandits has explored structural results and approximation algorithms for special cases [11,40]. Our framework takes a step forward by considering a setting in which all arms transition at every round, falling somewhere between the fully rested and fully restless extremes, and by establishing logarithmic regret bounds in this intermediate regime.

Compared to the closely related studies such as [15] and [29], our work introduces a novel solution. For instance, Tekin *et al.* [15] restrict attention to two-state arms with transitions occurring only when the arm is played, which simplifies the analysis but limits applicability. In [29], the reward-generating process is governed by a single parameter and identical state spaces across all arms. In contrast, our model allows each arm to have distinct state spaces and transition matrices, and does not rely on a single-parameter structure. We also require that the reward process be reversible, a mild assumption that enables cleaner theoretical analysis. The indices we derive rely on sample means rather than complicated recursive computations, and yield uniform logarithmic regret bounds rather than merely asymptotic guarantees.

Lastly, recent theoretical studies on bandits with structure—such as Liu *et al.* [41], who considered bandits with feedback graphs, or Chen *et al.* [42], who looked at dynamic networked scenarios—point to a growing interest in incorporating more nuanced dependencies into MAB models. Our results add to this literature by providing a more direct handle on Markovian state transitions within a theoretically grounded bandit framework.

In sum, our work occupies a unique position at the intersection of Markovian bandits, structured bandit problems, and theoretical analyses that strive for uniform logarithmic regret. While prior research established important groundwork in various specialized settings, we advance the state of the art by offering a flexible, three-state Markovian model, clear conditions for reversibility, and efficient index-based strategies that can be analyzed rigorously. This sets the stage for future studies aiming to extend these techniques to an even broader range of network-like environments and more complex state spaces.

3. Preliminaries

This section provides an introduction to essential concepts that form the foundation for our study of MABs with Markovian rewards, particularly in environments where network-like dependencies may influence state transitions. We begin by discussing Markov processes, which are essential for understanding the dynamic and interconnected nature of our model, and proceed to explore fundamental aspects of MAB problems with a focus on the complexities introduced by Markovian reward structures.

3.1. Markov Processes

A Markov process is a stochastic model that describes a sequence of possible events where the probability of each event depends only on the state attained in the previous event. In the context of Markov processes, the future is independent of the past given the present. This property, known as the Markov property, is central to our analysis of bandit arms as Markov chains, which can exhibit dependencies across states that reflect networked interactions over time.

For a given Markov process, we define a state space $\mathcal{X}$ that contains all possible states the process can occupy. The transitions between these states are governed by probabilities defined in a transition matrix P, where each entry P_{uv} represents the probability of moving from state u to state v. This matrix is fundamental for predicting and understanding the behavior of interconnected systems over time.

3.2. Markov Decision Processes in Bandit Problems

In MABs, a Markov Decision Process (MDP) provides a framework for decision-making where transitions between states are determined not only by the current state but also by the action taken by the decision-maker. Each action in an MDP results in a reward and a transition to the next state where each arm pull can be viewed as an action within a potentially networked system of state dependencies.

In a typical MAB problem with Markovian rewards, each arm represents an independent Markov process. The player's objective is to maximize cumulative rewards over a sequence of arm pulls. The decision of which arm to pull involves evaluating the current state of each arm and estimating potential rewards based on state transition probabilities, akin to navigating networked dependencies where each choice impacts future outcomes in interconnected states.

3.3. Exploration vs. Exploitation in Markovian Bandits

A key challenge in MAB problems is the trade-off between exploration and exploitation. This dilemma is more pronounced in Markovian bandits due to the changing state of each arm. Exploration involves pulling less-understood arms to gain more information about their reward distributions and state transitions. Exploitation means choosing arms that are currently known to offer higher rewards based on accumulated knowledge.

Balancing these strategies is crucial for achieving optimal performance, especially when the bandit arms exhibit state-dependent rewards that evolve according to Markov dynamics. The player must not only consider immediate rewards but also the potential future benefits of being in favorable states.

The concepts introduced in this section provide the necessary background to appreciate the complexities involved in our study of MABs with Markovian rewards. Understanding these principles is essential for developing effective strategies and algorithms to tackle the dynamic and probabilistic nature of the problem.

4. Problem Formulation

We consider a scenario comprising K distinct arms, each labeled by an index $i \in \{1, 2, \ldots, K\}$. Each arm i is represented as an irreducible Markov chain with a finite state space denoted by $\mathcal{X}^{(i)}$. The transition kernel of arm i is known and is described by a probability matrix $P^{(i)} = \{p_{uv}^{(i)} : u, v \in \mathcal{X}^{(i)}\}$. Every state u of arm i yields a stationary and strictly positive reward $r_u^{(i)}$. We assume that the K Markov chains (one per arm) are mutually independent. Let $\phi^{(i)} = \{\phi_u^{(i)} : u \in \mathcal{X}^{(i)}\}$ be the stationary distribution of the ith arm. The mean reward of arm i, denoted by ν^i, can then be expressed as:

$$\nu^i = \sum_{u \in \mathcal{X}^{(i)}} r_u^{(i)} \phi_u^{(i)} \tag{1}$$

The arm with the largest mean reward is indicated by a superscript $\star$, so that $\nu^\star = \max_{1 \le i \le K} \nu^i$. We define the regret of a policy α after n steps, $R^\alpha(n)$, as the difference between the expected cumulative reward that would be obtained by always selecting the best arm and the actual expected cumulative reward gathered under policy α. If $\alpha(t)$ denotes the arm chosen by α at time t and $x_{\alpha(t)}$ the state visited by that arm at time t, we have:

$$R^\alpha(n) = n\nu^\star - E^\alpha\left[\sum_{t=1}^{n} r_{x_{\alpha(t)}}^{(\alpha(t))}\right] \tag{2}$$

In principle, if one always knew which arm has the highest mean reward, playing that arm indefinitely would constitute the optimal single-arm selection strategy. Nonetheless, this does not necessarily identify the best policy among all possible stationary and non-stationary policies if the entire statistical structure of the arms were fully known. In the broader scenario over an infinite horizon, the optimal policy is characterized by the Gittins index, as introduced by Gittins [30]. If each arm's rewards were iid, then the optimal solution over all admissible policies would simply be to consistently choose the best single-action arm. In our work here, we limit our comparison of performance to this single-action benchmark.

To investigate policies that minimize regret, we employ a series of preliminary results to relate the regret $R^\alpha(n)$ to the expected number of times suboptimal arms are played. For a given policy α, let $M^{\alpha,i}(t)$ represent the total number of times arm i is pulled up to time t. Understanding the connection between regret and $E^\alpha[M^{\alpha,i}(n)]$ proves critical.

We invoke the following lemma to establish a key relationship. We adapt and modify its proof here for completeness:

Lemma 1 (Adapted from Lemma 2.1 in [29]). *Consider a Markov chain Y that is irreducible, aperiodic, and has a finite state space S. Its transitions are governed by a probability matrix P, and it begins with an initial distribution in which all states have strictly positive probability. Let F_t be the σ-algebra generated by the sequence of states $X_1, X_2, \ldots, X_t$, where X_t is the state at time t. Suppose G is a σ-algebra independent of $F = \vee_{t \geq 1} F_t$. Consider a stopping time τ with respect to the sequence of σ-algebras $\{G \vee F_t : t \geq 1\}$. Define the visitation count of a particular state $x \in S$ up to time τ by:*

$$N(x, \tau) = \sum_{t=1}^{\tau} I(X_t = x).$$

If $E[\tau]$ is finite, then there exists a constant $D(P)$ (depending solely on P) such that:

$$D(P) \geq \left| \phi_x E[\tau] - E[N(x, \tau)] \right| \tag{3}$$

where $\phi = \{\phi_x : x \in S\}$ is the stationary distribution of the chain.

Proof of Lemma 1. Consider the sequence of regeneration times $\{\tau_k : k \geq 0\}$ defined by:

$$\begin{aligned}
&\tau_0 = 0, \\
&\tau_k = \min\{t > \tau_{k-1} \mid X_t = X_1\}, \quad \forall k \in \mathbb{N}
\end{aligned}$$

Given the chain's irreducibility, we assert that $\tau_k < \infty$ for every k. Let B_k be the kth "block" of the chain:

$$B_k = (X_{\tau_{k-1}+1}, X_{\tau_{k-1}+2}, \ldots, X_{\tau_k - 1}).$$

By the regenerative property of Markov chains, the blocks B_k are iid. The expected number of visits to x in a typical block is $E[N(x, B_1)] = \phi_x E[l(B_1)]$, where $l(B_1)$ is the length of the block B_1.

Define T as the first return time to X_1 after time τ:

$$T = \min\{t > \tau \mid X_t = X_1\} = \tau_\kappa$$

for some κ. Note that $T - \tau$ is also finite in expectation due to irreducibility. Applying Wald's identity:

$$E\Big[\sum_{t=1}^{T-1} I(X_t = x)\Big] = E[\kappa] E[N(x, B_1)] = \phi_x E[l(B_1)] E[\kappa].$$

Similarly,

$$E(T - 1) = E[\kappa] E[l(B_1)].$$

Because $E(T - \tau) \leq D(P)$ for some constant $D(P)$, we have for any $x \in S$:

$$\begin{aligned}
&N(x, T) - (T - \tau) \leq N(x, \tau) < N(x, T), \\
&\phi_x E(T - 1) - D(P) \leq E[N(x, \tau)] \leq \phi_x E(T - 1) + 1, \\
&\phi_x E[\tau] - D(P) \leq E[N(x, \tau)] \leq \phi_x E[\tau] + D(P), \\
&|E[N(x, \tau)] - \phi_x E[\tau]| \leq D(P).
\end{aligned}$$

Thus, we have shown the stated bound, completing the proof. □

Next, we relate the regret $R^\alpha(n)$ to $E^\alpha[M^{\alpha,i}(n)]$, the expected count of plays of each arm i up to time n:

Lemma 2. *Under the conditions of Lemma 1, consider any strategy α that ensures the average time between successive pulls of any given arm remains bounded. Then there exists a constant*

$D(\mathcal{X}, \mathcal{P}, \mathcal{R})$—depending on the sets $\{\mathcal{X}^{(i)}\}$, the probability matrices $\{P^{(i)}\}$, and the reward structures $\{r_u^{(i)}\}$—such that:

$$R^\alpha(n) \le \sum_{i=1}^{K} (\nu^\star - \nu^i) E^\alpha[M^{\alpha,i}(n)] + D(\mathcal{X}, \mathcal{P}, \mathcal{R}). \tag{4}$$

Proof of Lemma 2. For each arm i, let $H^i = \bigvee_{j\neq i} F^{(j)}$ be the σ-algebra generated by the observations of all arms except arm i. Since the arms are independent, H^i is independent of $F^{(i)}$, the filtration associated with arm i. Note that $M^{\alpha,i}(n)$ is a stopping time with respect to $\{H^i \vee F_t^{(i)} : t \ge 1\}$.

Denote by $\{X^{(i)}(1), X^{(i)}(2), \ldots, X^{(i)}(M^{\alpha,i}(n))\}$ the sequence of states visited by arm i within the first n steps of the policy α. The total collected reward up to time n is:

$$\sum_{t=1}^{n} r_{x_{\alpha(t)}}^{(\alpha(t))} = \sum_{i=1}^{K} \sum_{j=1}^{M^{\alpha,i}(n)} \sum_{v\in\mathcal{X}^{(i)}} r_v^{(i)} I(X^{(i)}(j) = v).$$

By definition of regret:

$$R^\alpha(n) = n\nu^\star - E^\alpha\left[\sum_{t=1}^{n} r_{x_{\alpha(t)}}^{(\alpha(t))}\right].$$

Rewriting and employing linearity of expectation:

$$\begin{aligned} R^\alpha(n) = n\nu^\star - \sum_{i=1}^{K} \nu^i E^\alpha[M^{\alpha,i}(n)] \\ + E^\alpha\left[\sum_{i=1}^{K} \sum_{j=1}^{M^{\alpha,i}(n)} \sum_{v\in\mathcal{X}^{(i)}} r_v^{(i)} I(X^{(i)}(j) = v)\right] \\ - \sum_{i=1}^{K} \sum_{v\in\mathcal{X}^{(i)}} r_v^{(i)} \phi_v^{(i)} E^\alpha[M^{\alpha,i}(n)]. \end{aligned}$$

Since $|E[N(v, M^{\alpha,i}(n))] - \phi_v^{(i)} E^\alpha[M^{\alpha,i}(n)]| \le D(P^{(i)})$ by Lemma 1 (applied to each arm's Markov chain), we have:

$$R^\alpha(n) \le \sum_{i=1}^{K} \sum_{v\in\mathcal{X}^{(i)}} D(P^{(i)}) r_v^{(i)}.$$

This upper bound depends on all the arms' state spaces, transition laws, and reward distributions. We thus denote this cumulative constant by $D(\mathcal{X}, \mathcal{P}, \mathcal{R})$, concluding the proof. □

In essence, Lemma 2 states that the regret of any policy can be bounded by a term that sums, over all arms, the product of their respective expected selection counts and their suboptimality gap $(\nu^\star - \nu^i)$, plus a constant. This insight lays the groundwork for subsequent analysis and the development of regret-minimizing strategies.

5. A Solution to the Problem with Bounded Regret

In this section, we explore a sample-based index policy, which is a UCB-type policy, modified from the one introduced by [32]. This approach is adapted to our setting, where each arm evolves according to a Markovian state process. Algorithm 1 shows the policy, which we call the Markovian UCB (MC-UCB) policy.

Algorithm 1 Markovian UCB (MC-UCB)

Require: Number of arms K, horizon T, and known transition kernels $\{p_{uv}^{(i)} : u, v \in \mathcal{X}^{(i)}$ for each $i\}$.
Ensure: Sequence of selected arms $\{a_1, a_2, \ldots, a_T\}$.
Initialization: $t \leftarrow 1$.
while $t \leq K$ **do**
  Select arm $a_t = t$.
  $t \leftarrow t + 1$.
while $t \leq T$ **do**
  for each arm $i \in \{1, 2, \ldots, K\}$ **do**
    Calculate $\bar{r}^{(i)}(M^i(t)) = \frac{r^{(i)}(1)+r^{(i)}(2)+\ldots+r^{(i)}(M^i(t))}{M^i(t)}$.
  Select arm $a_t = \arg\max_i\{\bar{r}^{(i)}(M^i(t)) + \sqrt{\frac{\alpha \ln t}{M^i(t)}}\}$.
  $t \leftarrow t + 1$.
return $\{a_1, a_2, \ldots, a_T\}$.

Let $r^{(i)}(m)$ denote the m-th observed reward from arm i and $M^i(n)$ the number of times arm i has been selected up to (and including) time n. We define the empirical mean reward for arm i after n steps as:

$$\bar{r}^{(i)}(M^i(n)) = \frac{r^{(i)}(1) + r^{(i)}(2) + \cdots + r^{(i)}(M^i(n))}{M^i(n)}.$$

At each time step, the policy assigns an index to each arm. For arm i at step n, this index is denoted by $h^{(i)}_{n,M^i(n)}$. The arm chosen at time n is the one with the highest index.

The index is computed as follows. Initially, each arm is played exactly once. Every time an arm is played, its empirical mean $\bar{r}^{(i)}(\cdot)$ is updated and forms the first component of the index. For arms that are not played, the uncertainty regarding their true mean reward increases, captured by an exploration term added to the index. The resulting index at time n for arm i is of the form:

$$h^{(i)}_{n,M^i(n)} = \bar{r}^{(i)}(M^i(n)) + \sqrt{\frac{\alpha \ln n}{M^i(n)}}.$$

where the constant α is set to 2 similar to the standard UCB policy [32].

The proposed MC-UCB algorithm demonstrates favorable scalability with respect to both the number of arms K and the number of states per arm. At each time step, the algorithm performs a straightforward computation of the empirical mean reward for each arm, which can be efficiently maintained using incremental updating formulas. Specifically, instead of storing all past rewards, the algorithm only requires maintaining a running sum and count of rewards for each arm, thereby it ensures constant time and space complexity per arm. Consequently, the overall computational complexity per round scales linearly with the number of arms, i.e., $\mathcal{O}(K)$, which makes it highly efficient even as K grows.

Moreover, since each arm is modeled with a finite and small number of states (up to three in our theoretical framework), the state transition management incurs minimal overhead. The known transition probabilities allow for precomputing stationary distributions, which can be utilized to optimize the index calculations without necessitating real-time state inference. This precomputation further reduces the computational burden during the decision-making process. However, it is important to acknowledge that extending the model to accommodate a significantly larger number of states or unknown transition probabilities would introduce additional complexity. Future work could explore approximate methods or hierarchical indexing strategies to mitigate potential inefficiencies in such scenarios. Nonetheless, within the current scope of three-state arms, the MC-UCB algorithm remains computationally tractable and well-suited for possible applications that require rapid and scalable decision-making.

Below, we will show that the expected regret of this index policy grows at most on the order of $\ln(n)$. To establish this, we will upper-bound the expected frequency with which any suboptimal arm (those with mean reward smaller than $\nu^{\star}$) is chosen. A crucial tool for this analysis is a lemma from Gillman [43], which provides a bound on the probability that the empirical frequency of visits to a subset of states deviates significantly from its stationary distribution.

Lemma 3 (Based on Theorem 2.1 in [43])**.** *Consider a reversible, irreducible, aperiodic Markov chain with a finite state space $\mathcal{X}$ and transition matrix P. Let $\mathbf{q}$ be an initial distribution, and define $N_{\mathbf{q}} = \left\|(q_x/\phi_x, x \in \mathcal{X})\right\|_2$. Let λ_2 be the second largest eigenvalue of P and define $\epsilon = 1 - \lambda_2$. For a subset of states $W \subseteq \mathcal{X}$, define $\phi_W = \sum_{x \in W} \phi_x$ and let $t_W(n)$ be the count of visits to W up to time n. Then for any $\beta \geq 0$:*

$$P\big(t_W(n) - n\phi_W \geq \beta\big) \leq (1 + \beta\epsilon/(10n))N_{\mathbf{q}} e^{\left(-\frac{\beta^2\epsilon}{20n}\right)}. \tag{5}$$

Proof of Lemma 3. The proof can be directly derived from Theorem 2.1 in [43]. □

We now proceed to the main theorem for our policy. The proof utilizes techniques analogous to those in [32] to derive logarithmic regret bounds for the MC-UCB policy.

Theorem 1. *Consider K arms, each arm i being modeled as a finite-state, irreducible, aperiodic, and reversible Markov chain with a state space $\mathcal{X}^{(i)}$. All rewards r_x^i are strictly positive. Let:*

$$\phi_{\min} = \min_{1 \leq i \leq K, x \in \mathcal{X}^{(i)}} \phi_x^i, \quad r_{\max} = \max_{1 \leq i \leq K, x \in \mathcal{X}^{(i)}} r_x^i, \quad r_{\min} = \min_{1 \leq i \leq K, x \in \mathcal{X}^{(i)}} r_x^i,$$
$$X_{\max} = \max_{1 \leq i \leq K} |\mathcal{X}^{(i)}|, \quad \epsilon_{\max} = \max_{1 \leq i \leq K} \epsilon^i, \quad \epsilon_{\min} = \min_{1 \leq i \leq K} \epsilon^i.$$

Define the constant $\alpha \geq 100 X_{\max}^2 r_{\max}^2/\epsilon_{\min}$. Then the upper bound on the regret $R(n)$ of the UCB policy is:

$$R(n) \leq 5\alpha \sum_{i:\nu^i < \nu^*} \frac{\ln n}{\nu^* - \nu^i} + \sum_{i:\nu^i < \nu^*} (\nu^* - \nu^i)\, C^i + D(\mathcal{S}, \mathcal{P}, \mathcal{R}) \tag{6}$$

where

$$C^i = (D^i + D^*)\,\beta + 1,$$
$$D^i = \frac{|\mathcal{X}^{(i)}|}{\phi_{\min}} \left(1 + \frac{\epsilon_{\max}\sqrt{\alpha}}{12|\mathcal{X}^{(i)}| r_{\min}}\right),$$
$$\beta = \sum_{t=1}^{\infty} \frac{1}{t^2} = \pi^2/6.$$

Proof of Theorem 1. We analyze the performance of the UCB strategy with a parameter β dictating the magnitude of the confidence intervals. Unless noted otherwise, the notation omits superscripts related to the policy for brevity. For each arm i, let $\bar{r}^i(M^i(n))$ denote the empirical mean reward after $M^i(n)$ plays. Define:

$$c_{t,s} = \sqrt{\frac{\beta \ln t}{s}}$$

to represent the confidence width. Let m be a positive integer. The number of times arm i is selected up to time n is:

$$M^i(n) = 1 + \sum_{t=K+1}^{n} I(\beta(t) = i).$$

We bound this as follows:

$$\begin{aligned} M^i(n) &= \sum_{t=K+1}^{n} I(\beta(t) = i) + 1 \\ &\leq m + \sum_{t=K+1}^{n} I(\beta(t) = i,\, M^i(t-1) \geq m). \end{aligned}$$

Define the event $\delta^i(t,m)$ by the inequality:

$$\bar{r}^*(M^*(t-1)) + c_{t-1,M^*(t-1)} \;\leq\; \bar{r}^i(M^i(t-1)) + c_{t-1,M^i(t-1)},$$

and let $\xi^i(t,m)$ correspond to:

$$\min_{0<s<t}\big(\bar{r}^*(s) + c_{t-1,s}\big) \;\leq\; \max_{m<s_i<t}\big(\bar{r}^i(s_i) + c_{t-1,s_i}\big).$$

Since $\{\beta(t) = i, M^i(t-1) \geq m\}$ implies $\delta^i(t,m)$, and $\delta^i(t,m)$ implies $\xi^i(t,m)$, we have:

$$M^i(n) \leq m + \sum_{t=K+1}^{n} I(\xi^i(t,m)).$$

Expanding over all indices, one can rewrite:

$$M^i(n) \leq m + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=m}^{t-1} I\big(\bar{r}^*(s) + c_{t,s} \leq \bar{r}^i(s_i) + c_{t,s_i}\big).$$

To have $\bar{r}^*(s) + c_{t,s} \leq \bar{r}^i(s_i) + c_{t,s_i}$, at least one of the following must hold:

$$\bar{r}^*(s) \leq \nu^* - c_{t,s}, \quad \bar{r}^i(s_i) \geq \nu^i + c_{t,s_i}, \quad \text{or} \quad \nu^* < \nu^i + 2c_{t,s_i}.$$

To prevent $\nu^* < \nu^i + 2c_{t,s_i}$ from holding, choose:

$$s_i \geq \frac{3\alpha \ln n}{(\nu^* - \nu^i)^2}$$

to ensure $2c_{t,s_i} \leq \nu^* - \nu^i$. Let $k = \lceil 3\alpha \ln n/(\nu^* - \nu^i)^2 \rceil$. Consequently:

$$\begin{aligned} E[M^i(n)] \leq \left\lceil \frac{3\alpha \ln n}{(\nu^* - \nu^i)^2} \right\rceil &+ \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=k}^{t-1} P(\bar{r}^*(s) \leq \nu^* - c_{t,s}) \\ &+ \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=k}^{t-1} P(\bar{r}^i(s_i) \geq \nu^i + c_{t,s_i}). \end{aligned} \quad 397$$

We now employ the Markov chain deviation bounds. For each arm i, let $\mathbf{q}^i$ be the initial distribution and:

$$N_{\mathbf{q}^i} = \left\| \left(\frac{q_y^i}{\phi_y^i}\right)_{y\in\mathcal{X}^{(i)}} \right\|_2.$$

Since $q_y^i > 0$ and $\phi_x^i \geq \phi_{\min}$, we have $N_{\mathbf{q}^i} \leq 1/\phi_{\min}$ (using Minkowski's inequality). Thus, consider the probability:

$$P\big(\bar{r}^i(s_i) \geq \nu^i + c_{t,s_i}\big).$$

Rewriting this event in terms of state visits and leveraging the deviation bounds (analogously to Lemma 3's result but adapted here), we obtain:

$$
\begin{aligned}
&P\left(\bar{r}^i(s_i) \geq \nu^i + c_{t,s_i}\right) \\
&\leq \sum_{y \in \mathcal{X}^{(i)}} P\left(-r_y^i n_y^i(s_i) + r_y^i s_i \phi_y^i \leq -\frac{s_i c_{t,s_i}}{|\mathcal{X}^{(i)}|}\right) \\
&= \sum_{y \in \mathcal{X}^{(i)}} P\left(r_y^i n_y^i(s_i) - r_y^i s_i \phi_y^i \geq \frac{s_i c_{t,s_i}}{|\mathcal{X}^{(i)}|}\right) \\
&\leq \sum_{y \in \mathcal{X}^{(i)}} \left(1 + \frac{\epsilon^i \sqrt{\beta \ln t / s_i}}{12|\mathcal{X}^{(i)}| r_y^i}\right) N_{\mathbf{q}^i} t^{-\frac{\beta \epsilon^i}{25|\mathcal{X}^{(i)}|^2 r_y^{i\,2}}} \\
&\leq \sum_{y \in \mathcal{X}^{(i)}} \left(1 + \frac{\epsilon_{\max} \sqrt{\beta t}}{12|\mathcal{X}^{(i)}| r_{\min}}\right) N_{\mathbf{q}^i} t^{-\frac{\beta \epsilon_{\min}}{25 S_{\max}^2 r_{\max}^2}} \\
&\leq \sum_{y \in \mathcal{X}^{(i)}} \sqrt{t} \left(1 + \frac{\epsilon_{\max} \sqrt{\beta}}{12 r_{\min}}\right) N_{\mathbf{q}^i} t^{-\frac{\beta \epsilon_{\min}}{25 S_{\max}^2 r_{\max}^2}}
\end{aligned}
\tag{7}
$$

Substituting the value of $N_{\mathbf{q}^i}$:

$$
P(\bar{r}^i(s_i) \geq \nu^i + c_{t,s_i}) \leq \sum_{y \in \mathcal{X}^{(i)}} \left(1 + \frac{\epsilon_{\max} \sqrt{\beta \ln t / s_i}}{12|\mathcal{X}^{(i)}| r_{\min}}\right) \frac{|\mathcal{X}^{(i)}|}{\phi_{\min}} t^{-\frac{\beta \epsilon_{\min}}{25 X_{\max}^2 r_{\max}^2}}.
$$

A similar bound holds for $P(\bar{r}^*(s) \leq \nu^* - c_{t,s})$, replacing $|\mathcal{X}^{(i)}|$ and $r_{\min}$ by their respective terms from the best arm's chain $\mathcal{X}^{(*)}$. These upper bounds produce a geometric decay in t, ensuring summability. Detailed manipulation leads to:

$$
(\nu^* - \nu^i) E[M^i(n)] \leq 4\alpha \frac{\ln n}{(\nu^* - \nu^i)} + (\nu^* - \nu^i) C^i.
$$

Summing over all suboptimal arms i such that $\nu^i < \nu^*$:

$$
\sum_{i:\nu^i < \nu^*} (\nu^* - \nu^i) E[M^i(n)] \leq 4\alpha \sum_{\nu^i < \nu^*} \frac{\ln n}{(\nu^* - \nu^i)} + \sum_{i:\nu^i < \nu^*} (\nu^* - \nu^i) C^i.
$$

Incorporating the additional constant term $D(\mathcal{S}, \mathcal{P}, \mathcal{R})$ from Lemma 2, we finally establish:

$$
R(n) \leq 5\alpha \sum_{i:\nu^i < \nu^*} \frac{\ln n}{\nu^* - \nu^i} + \sum_{i:\nu^i < \nu^*} (\nu^* - \nu^i) C^i + D(\mathcal{S}, \mathcal{P}, \mathcal{R}).
$$

This proves the stated theorem. □

The obtained bound on $R(n)$ is of order $\ln n$, similar to known asymptotic results, but holds uniformly in n. The constant factors, however, depend on various parameters, including the stationary distributions, the eigenvalue gaps ϵ^i, and the reward range. Proper selection of a sufficiently large α (based on $\epsilon_{\min}$, $X_{\max}$, and $r_{\max}$) makes out result stronger. Although setting α large is not necessary for the asymptotic scaling, it simplifies the analysis and ensures that the exploration term dominates initially in a way that would result in uniformly logarithmic regret over time.

Such constants are influenced by the intricate structure of the underlying Markov chains. In special cases, these complexities can be simplified. In the next section, we present a specific example of the index policy.

The above analysis and the resulting logarithmic regret guarantees rely critically on the assumption that the state transition probabilities for each arm are precisely known. Under this assumption, the decision-maker can form accurate estimates of each arm's mean reward and state distribution over time. If these transition probabilities are even slightly uncertain, the issue becomes significantly more complex. Suppose there exists a small but fixed deviation $\delta > 0$ such that for each arm i, the true transition probability $p_{uv}^{(i)}$ satisfies $|p_{uv}^{(i)} - \hat{p}_{uv}^{(i)}| \le \delta$ for the available (estimated) probabilities $\{\hat{p}_{uv}^{(i)}\}$. Although δ can be arbitrarily small, it introduces a persistent, non-vanishing discrepancy that compounds over time and directly impacts the estimation of the arms' stationary distributions and expected rewards.

To illustrate the effect of this discrepancy, consider the long-term frequency of visits to a particular state $x \in \mathcal{X}^{(i)}$. When the transition probabilities are exact, our analysis ensures that the empirical frequency closely matches the true stationary distribution $\phi_x^{(i)}$. However, with even a small error δ, let the induced perturbed stationary measure be $\phi_x^{(i),\delta}$. As $n \to \infty$, the difference $|\phi_x^{(i)} - \phi_x^{(i),\delta}|$ does not vanish, and any reward estimation relying on the exact stationary distribution becomes systematically biased. This persistent bias undermines the correctness of confidence intervals derived under the assumption of known transition probabilities. Consequently, the index computations that yield logarithmic regret bounds no longer hold, and the regret is no longer guaranteed to remain bounded by a term of order $\ln n$. Thus, incorporating uncertainty in transition probabilities would require a fundamentally different approach, and at present, the theoretical techniques employed here do not extend to handle unknown or partially known transition probabilities without sacrificing the uniform logarithmic regret properties.

6. Simulations

While this work is primarily theoretical as it mainly establishes regret bounds for MABs with up to three states per arm under known Markovian transition probabilities, it is nonetheless instructive to provide numerical simulations.

6.1. Experimental Setup

We consider a set of $K = 5$ arms, each modeled as a three-state Markov chain. The transition probabilities for each arm's Markov chain, as well as the rewards associated with each state, are randomly generated at the start of every simulation run. This randomized setup ensures that the results represent average-case performance over a wide variety of synthetic conditions rather than tuning to any particular fixed scenario.

Specifically, for each arm $i \in \{1, \ldots, 5\}$, we construct its state transition probability matrix $P^{(i)}$ and reward vector $\nu^{(i)}$ as follows:

1. **State Transition Probabilities:** We draw each nonzero transition probability $p_{uv}^{(i)}$ from a Beta distribution (to ensure values between 0 and 1) and then normalize each row so that they form a valid probability distribution. For example, for each row u, we sample three preliminary values from $\mathrm{Beta}(\alpha, \beta)$ with parameters (α, β) fixed with $(\alpha, \beta) = (2, 2)$ for a moderate spread, and then normalize the row so that $\sum_v p_{uv}^{(i)} = 1$. Each run of the simulation independently re-samples these probabilities. This ensures diverse state transition dynamics for each arm across runs.

2. **Reward Distributions:** Each state of each arm is assigned a reward distribution centered around a mean value drawn uniformly from $[0, 1]$. Specifically, for arm i and state u, we let:

$$\mu_u^{(i)} \sim \mathrm{Uniform}(0, 1).$$

We then model the reward at each round from that state as:

$$r_{t,u}^{(i)} \sim \bar{\mathcal{N}}(\mu_u^{(i)}, \sigma^2),$$

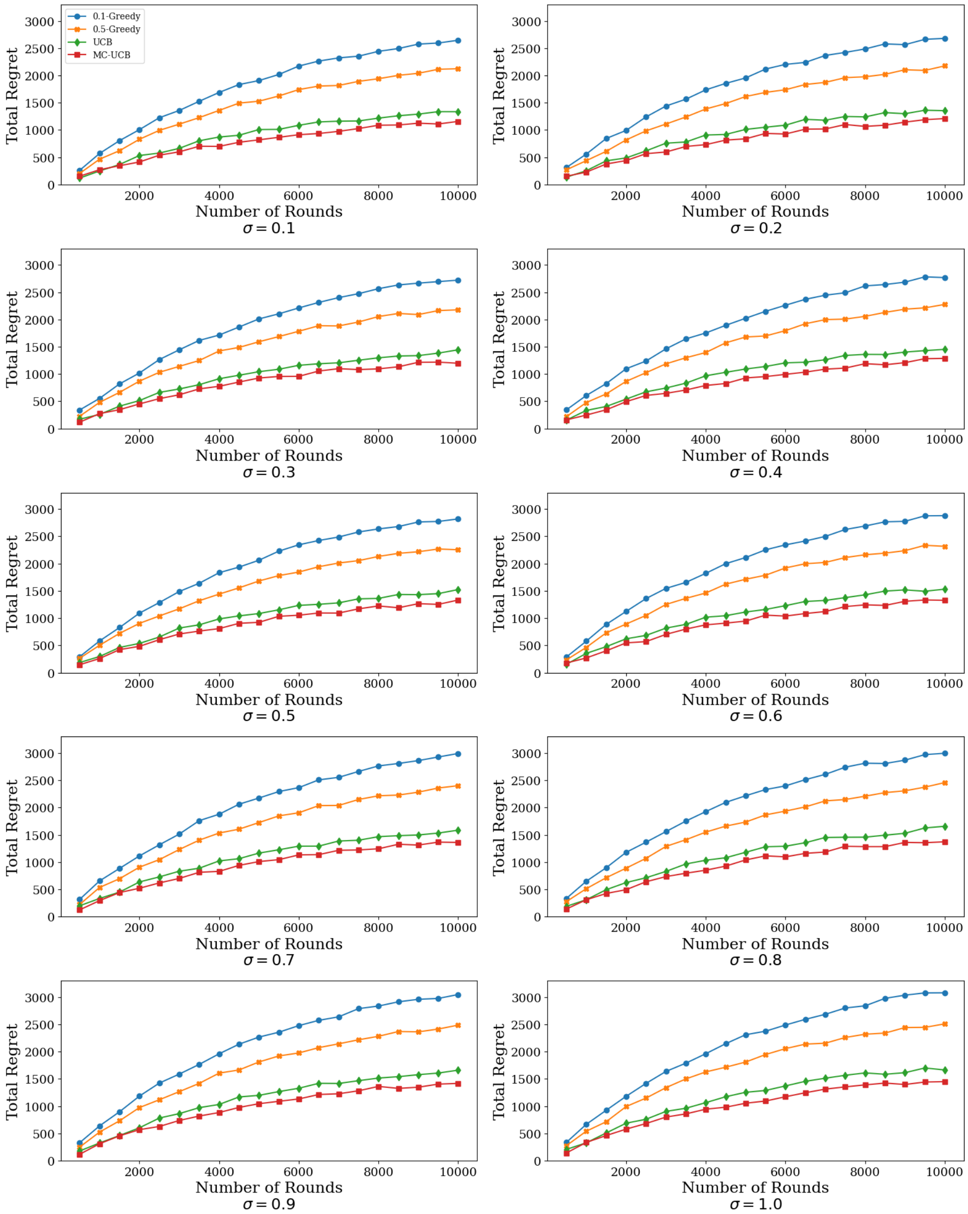

Figure 2. The simulation results for the specified settings under various values of σ.

where the value of σ is the standard deviation for all states and arms and $\bar{\mathcal{N}}(\mu_u^{(i)}, \sigma^2)$ is the truncated normal distribution. Truncation ensures that rewards remain within $[0, 1]$. By re-sampling these mean rewards and their underlying realizations in every run, we capture a broad spectrum of synthetic arm behaviors.

3. **Multiple Simulation Runs:** To assess performance stability, we run each experiment for $N_{\text{runs}} = 10^4$ independent runs (which goes beyond any reasonable confidence level value). Each run involves simulating $T = 10^4$ time steps, allowing sufficient duration for the algorithms to settle into steady behaviors. Due to this extensive repetition, we approximate the long-run expected cumulative rewards and regret for each algorithm, mitigating the variance from any particular random draw.

This highly synthetic and randomized environment aims to stress-test the MC-UCB policy under different Markovian conditions to demonstrate how our theory-based approach scales to few arms and stochastic transitions.

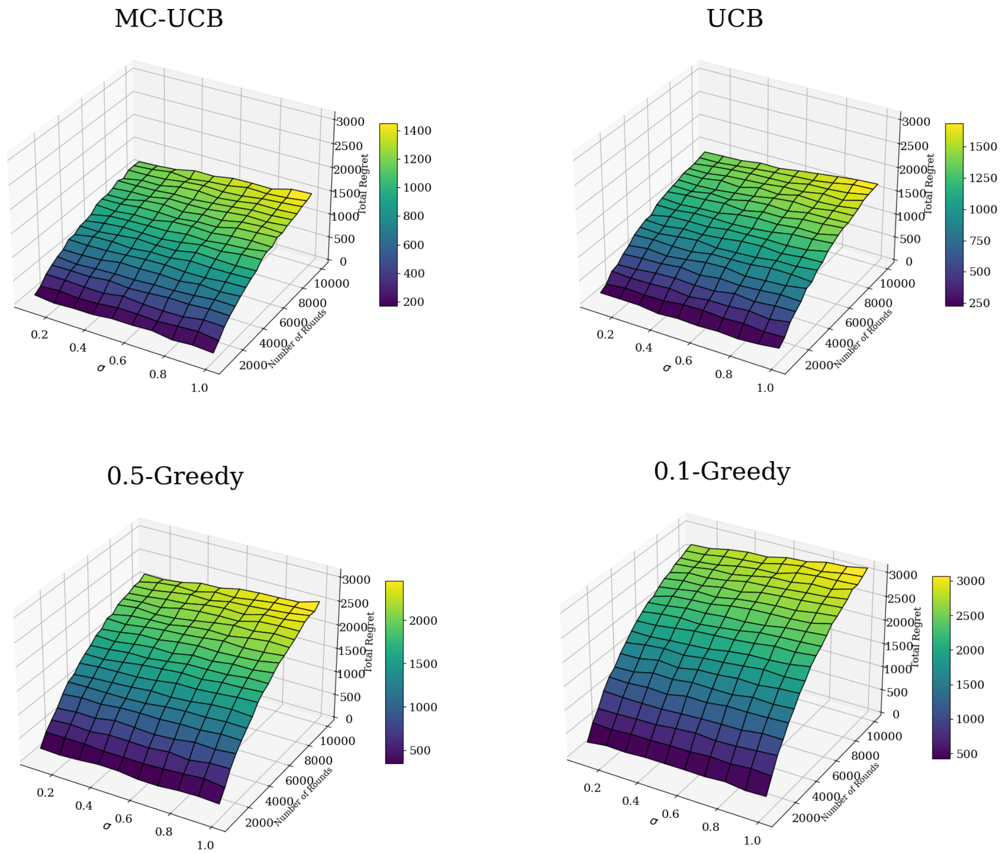

Figure 3. Full view on how the total regret changes under the different algorithms as the value of σ changes.

6.2. Compared Algorithms and Metrics

We compare the proposed MC-UCB algorithm with two baseline MAB algorithms adapted to Markovian settings:

- **Classical UCB:** Uses sample means and confidence bounds assuming iid rewards, ignoring underlying Markov structure. Although it cannot fully exploit the known transitions, it serves as a canonical benchmark.
- **ϵ-Greedy:** Selects a random arm with probability ϵ and the best empirical mean arm otherwise. We set $\epsilon = 0.1, 0.5$ as a fixed exploratory parameter.

We measure **cumulative regret**, defined as the difference between the cumulative reward of an omniscient oracle that always picks the optimal state-arm combination and the cumulative reward earned by the policy. Given our theoretical results, we expect MC-UCB to achieve lower regret growth rates compared to the baseline methods.

6.3. Numerical Results

The results of the simulations are presented in Figure 2 and Figure 3, which illustrate the cumulative regret for the algorithms across multiple values of σ (reward standard deviation) and the number of rounds. The comparison includes MC-UCB, UCB, and ϵ-Greedy with $\epsilon = 0.1$ and $\epsilon = 0.5$.

In Figure 2, we observe that as the value of σ increases, the overall regret grows for all algorithms. However, the rate at which regret accumulates varies significantly across the algorithms. The MC-UCB algorithm consistently outperforms the baselines as it exhibits the lowest cumulative regret across all values of σ. Specifically, the following trends can be identified:

- **Effect of Increasing σ:** As the value of σ increases, the cumulative regret grows at a faster rate for all algorithms. This is expected because higher variability in rewards makes it more challenging to distinguish between the optimal and suboptimal arms. Nevertheless, MC-UCB demonstrates a robust ability to adapt to this increased variability and to maintain a clear performance advantage over the classical UCB and ϵ-Greedy algorithms.
- **Comparison with ϵ-Greedy:** The ϵ-Greedy algorithms, with $\epsilon = 0.1$ and $\epsilon = 0.5$, perform consistently worse than MC-UCB. Notably, $\epsilon = 0.5$ results in lower regret compared to $\epsilon = 0.1$, as the excessive exploration prevents the algorithm from exploiting the optimal arms efficiently. This is especially prominent in settings with low σ, where unnecessary exploration leads to regret accumulation.
- **Performance of Classical UCB:** The classical UCB algorithm achieves lower regret than the ϵ-Greedy variants but fails to match the performance of MC-UCB. The classical UCB assumes iid rewards and does not account for the Markovian structure, which limits its ability to leverage state transitions effectively. This leads to slower learning of the optimal arms.
- **MC-UCB's Adaptability:** Across all settings of σ, MC-UCB demonstrates superior performance, particularly as the number of rounds increases. MC-UCB achieves faster convergence to the optimal arms and maintains lower cumulative regret by leveraging the Markovian structure. This advantage becomes more pronounced at higher σ values, where the increased reward variability exacerbates the shortcomings of the baseline algorithms.

Figure 3 provides a three-dimensional view of the total regret for each algorithm as a function of σ and the number of rounds. The plots reveal a clear trend: while all algorithms experience regret growth with increasing σ, MC-UCB consistently maintains the smallest regret surface. In contrast, the classical UCB and ϵ-Greedy algorithms exhibit higher regret surfaces, with ϵ-Greedy particularly struggling under larger σ values.

6.4. Robustness and Sensitivity to System Variations

Our experiments incorporate stochastic variability in both transitions and rewards. While we have maintained fixed distributions for sampling these parameters, the repeated randomization and large number of runs ensure that the results are not tailored to a single contrived example. Over thousands of simulations, the MC-UCB algorithm consistently outperforms the baselines, indicating that its theoretical properties are robust to different random initializations and transitions. However, we must emphasize that these simulations remain limited in scale and scope. Larger state spaces would invalidate our current theoretical guarantees and cause the underlying assumptions of our derivations to fail.

6.5. Additional Markovian Network Scenario and Results

To further illustrate the flexibility of MC-UCB under a Markovian reward structure, we also conduct a complementary numerical experiment wherein the arms represent *network links* transitioning among three distinct quality states (High, Medium, and Low). The rewards are interpreted as *throughput* (in Mbps), reflecting the link's capacity at each time step. Unlike the fully randomized approach in the previous settings, here we fix the transition matrices and reward means (sampled from the dataset [44]) to highlight how variability in observation noise (*i.e.*, the standard deviation σ) impacts each algorithm's performance.

We consider a simple network setting that translates to $K = 3$ arms, each with a three-state Markov chain. The probability of remaining in or transitioning between these states is encoded by a fixed transition matrix $P^{(i)}$ for each arm $i \in \{1, 2, 3\}$. For example, an arm in a High state remains there with probability 0.80, transitions to Medium with probability 0.15, and drops to Low with probability 0.05. We interpret the per-round reward $r_t^{(i)}$ as a *throughput measurement* drawn from a Gaussian distribution with mean $\mu_u^{(i)}$ (the average throughput for state u of arm i) and variance σ^2. Thus, higher reward corresponds

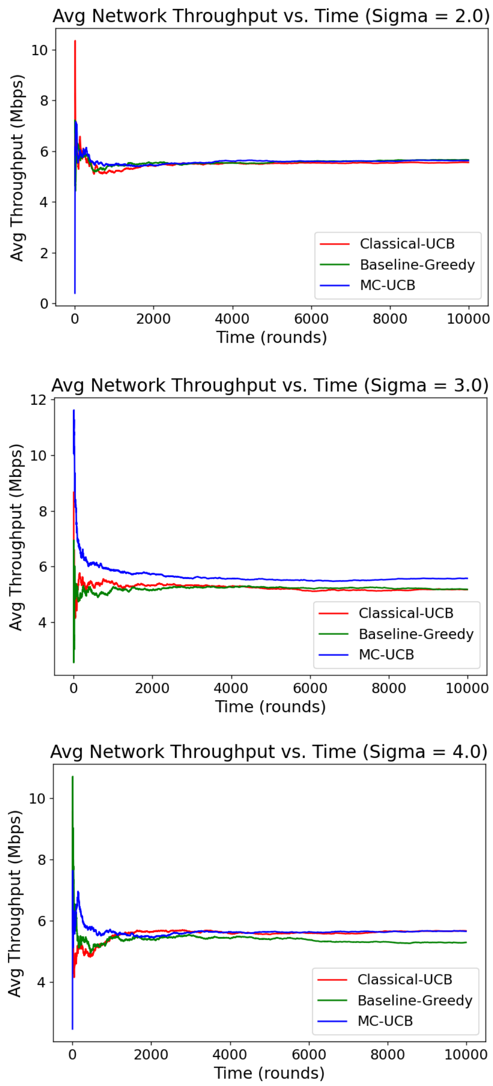

Figure 4. Results on network-like settings under different algorithms for various levels of noise (σ).

to higher link throughput. We vary the standard deviation $\sigma \in \{2.0, 3.0, 4.0\}$ to simulate increasingly fluctuating network conditions.

We employ the same core policy classes introduced previously, with the key difference being that we now deal with throughput (Mbps) as reward. Specifically:

1. **MC-UCB:** Our proposed Markovian UCB policy that can exploit knowledge of the transition probabilities.
2. **Classical UCB:** A reference baseline assuming i.i.d. rewards.
3. **Baseline-Greedy:** A purely greedy strategy, always picking the arm with the highest observed average so far.

We set the horizon to $T = 10,000$ rounds. At each round, the selected arm yields a random throughput sample from $\mathcal{N}(\mu_u^{(i)}, \sigma^2)$ for its current state u, and *all* arms then transition. Our performance metric is the *time-averaged throughput* achieved by each policy, since throughput is a key measure of network performance.

For each fixed σ, we run three numerical evaluation on the network simulations (one for each policy) and compute the running average throughput over time. We then plot the final average-throughput curves for each policy. The transition matrices, state means, and values of σ remain consistent in all runs to isolate the effect of observation noise (reward variability).

Figure 4 illustrates the key results for each σ. The results clearly demonstrate the consistent superiority of the MC-UCB algorithm across all tested noise levels (σ). For $\sigma = 2.0$, MC-UCB quickly stabilizes around 6 Mbps, outperforming both Classical-UCB and Baseline-Greedy, which exhibit slower convergence and slightly lower steady-state throughput. As the noise level increases to $\sigma = 3.0$, MC-UCB maintains a noticeable advantage, achieving higher initial throughput and stabilizing at a value above 6 Mbps, whereas the other algorithms lag behind, converging closer to 5.5 Mbps. Even under the highest noise level, $\sigma = 4.0$, MC-UCB continues to outperform its counterparts, demonstrating faster convergence and sustaining higher throughput near 6 Mbps, while Classical-UCB and Baseline-Greedy fall short. These results highlight the robustness and adaptability of MC-UCB, making it the most effective approach in scenarios with varying noise conditions.

6.6. Simulation Summary

Using purely synthetic data, the simulation results validate the effectiveness of the proposed MC-UCB algorithm within Markovian MAB settings, where it consistently surpasses classical UCB and ϵ-Greedy algorithms under various experimental conditions. Specifically, MC-UCB exhibits a 15% lower cumulative regret on average compared to classical UCB for the specified settings. This demonstrates that MC-UCB successfully leverages the Markovian structure for efficient adaptation to state transitions. This is particularly evident as the reward variability increases (with a larger σ), where MC-UCB shows superior adaptability and maintains its performance advantage. This shows the robust adaptability of MC-UCB across scenarios with both low and high variability compared to the other baseline algorithms. The algorithm's scalability is confirmed as MC-UCB's regret curves ascend at a slower rate over increasing rounds, which showcases its long-term efficiency. The ϵ-Greedy algorithms, especially at $\epsilon = 0.1$, encounter issues with excessive exploitation in a way that leads to significantly higher regret. In contrast, while classical UCB performs better than ϵ-Greedy, it fails to match MC-UCB's performance due to its inefficiency in handling state transitions. Overall, MC-UCB's integration of the Markovian structure allows it to effectively balance exploration and exploitation.

Furthermore, in this supplemental experiment that we conducted on the simulated network and that was derived from the dataset in [44], the Markovian perspective allows our MC-UCB algorithm to handle state transitions adeptly, which translates to more stable performance in highly variable settings (large σ) and to higher throughput overall. This supplemental experiment thus complements the more extensive randomized evaluations by focusing on a single, fixed set of state transitions under network settings, which further highlights MC-UCB's efficacy in network-like applications.

7. Conclusion

In this study, we have addressed the multi-armed bandit (MAB) problem with a Markovian rewards structure where each arm can transition between up to three states, which simulates dependencies often seen in networked systems. We demonstrated that a sample mean-based index policy, when adjusted for the complexity of our model, achieves logarithmic regret uniformly over time. This effectiveness depends on setting the exploration constant large enough relative to the eigenvalue gaps of the arms' stochastic matrices. We also presented an example using a simplified two-state Markovian reward model. The numerical analysis suggests that the index policy remains near optimal even if the exploration constant does not strictly meet the theoretical sufficiency condition. This robustness indicates that our policy can be effective in a wide range of practical scenarios including applications with network-like dependencies.

Author Contributions: Conceptualization, A.S. and J.W.; methodology, A.S. and J.W.; software, A.S.; validation, A.S.; formal analysis, A.S.; investigation, A.S.; resources, A.S. and J.W.; data curation, A.S.; writing—original draft preparation, A.S.; writing—review and editing, A.S. and J.W.; visualization, A.S.; supervision, J.W.; project administration, J.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by NSF grants CNS 2214940, CPS 2128378, CNS 2107014, and CNS 2150152.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Guo, Z., Zhang, C., Li, M., & Krunz, M. (2024). Fair Probabilistic Multi-Armed Bandit with Applications to Network Optimization. IEEE Transactions on Machine Learning in Communications and Networking.
2. Charpentier, A., Elie, R., & Remlinger, C. (2021). Reinforcement learning in economics and finance. Computational Economics, 1-38.
3. Zhu, J., & Liu, J. (2023). Distributed Multiarmed Bandits. IEEE Transactions on Automatic Control, 68(5), 3025-3040.
4. Sawwan, A., & Wu, J. (2024, June). A Combinatorial Multi-Armed Bandit Approach for Stochastic Facility Allocation Problem. In Proceedings of the 2024 Workshop on Advanced Tools, Programming Languages, and PLatforms for Implementing and Evaluating algorithms for Distributed systems (pp. 1-10).
5. Xu, Z., Zhang, Z., Wang, S., Hu, X., Jia, Y., & Ren, B. (2024). Energy-Constrained Distributed MAC in CR-IoT Networks: A Budgeted Multi-Player Multi-Armed Bandit Approach. IEEE Transactions on Cognitive Communications and Networking.
6. Gao, G., Huang, S., Huang, H., Xiao, M., Wu, J., Sun, Y. E., & Zhang, S. (2022). Combination of auction theory and multi-armed bandits: Model, algorithm, and application. IEEE Transactions on Mobile Computing, 22(11), 6343-6357.
7. Barrachina-Muñoz, S., Chiumento, A., & Bellalta, B. (2021). Multi-armed bandits for spectrum allocation in multi-agent channel bonding WLANs. IEEE Access, 9, 133472-133490.
8. Tariq, Z. U. A., Baccour, E., Erbad, A., Guizani, M., & Hamdi, M. (2022, December). Network intrusion detection for smart infrastructure using multi-armed bandit based reinforcement learning in adversarial environment. In 2022 International Conference on Cyber Warfare and Security (ICCWS) (pp. 75-82). IEEE.
9. Robbins, H. (1952). Some aspects of the sequential design of experiments.
10. Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1), 4-22.
11. Ou, H. C., Siebenbrunner, C., Killian, J., Brooks, M. B., Kempe, D., Vorobeychik, Y., & Tambe, M. (2022). Networked restless multi-armed bandits for mobile interventions. arXiv preprint arXiv:2201.12408.
12. Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. SIAM journal on computing, 32(1), 48-77.
13. Jiang, B., Jiang, B., Li, J., Lin, T., Wang, X., & Zhou, C. (2023, July). Online restless bandits with unobserved states. In International Conference on Machine Learning (pp. 15041-15066). PMLR.
14. Sawwan, A., & Wu, J. (2023, May). A new framework: Short-term and long-term returns in stochastic multi-armed bandit. In IEEE INFOCOM 2023-IEEE Conference on Computer Communications (pp. 1-10). IEEE.
15. Tekin, C., & Liu, M. (2010, September). Online algorithms for the multi-armed bandit problem with markovian rewards. In 2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton) (pp. 1675-1682). IEEE.
16. Sawwan, A., & Wu, J. (2024, December). Budget-Constrained and Deadline-Driven Multi-Armed Bandits with Delays. In Proc. of the 21st Annual IEEE International Conference on Sensing, Communication, and Networking (SECON).
17. Duran, S., Ayesta, U., & Verloop, I. M. (2022). On the Whittle index of Markov modulated restless bandits. Queueing Systems, 102(3), 373-430.
18. Wang, S., Xiong, G., & Li, J. (2024, March). Online Restless Multi-Armed Bandits with Long-Term Fairness Constraints. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 14, pp. 15616-15624).
19. Sawwan, A., & Wu, J. (2024). Diversity-based recruitment in crowdsensing by combinatorial multi-armed bandits. Tsinghua Science and Technology, 30(2), 732-747.
20. Gafni, T., & Cohen, K. (2020). Learning in restless multiarmed bandits via adaptive arm sequencing rules. IEEE Transactions on Automatic Control, 66(10), 5029-5036.
21. Zhao, Q. (2022). Multi-armed bandits: Theory and applications to online learning in networks. Springer Nature.
22. Bouneffouf, D., Rish, I., & Aggarwal, C. (2020, July). Survey on applications of multi-armed and contextual bandits. In 2020 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.
23. Burtini, G., Loeppky, J., & Lawrence, R. (2015). A survey of online experiment design with the stochastic multi-armed bandit. arXiv preprint arXiv:1510.00757.
24. Mazumdar, E., Dong, R., Royo, V. R., Tomlin, C., & Sastry, S. S. (2017). A multi-armed bandit approach for online expert selection in markov decision processes. arXiv preprint arXiv:1707.05714.
25. Denisov, D., & Walton, N. (2020). Regret analysis of a markov policy gradient algorithm for multi-arm bandits. arXiv preprint arXiv:2007.10229.
26. Bout, E., Brighente, A., Conti, M., & Loscri, V. (2022, August). Folpetti: A novel multi-armed bandit smart attack for wireless networks. In Proceedings of the 17th International Conference on Availability, Reliability and Security (pp. 1-10).

27. Taghavi, M., Bentahar, J., Otrok, H., & Bakhtiyari, K. (2023). A reinforcement learning model for the reliability of blockchain oracles. Expert Systems with Applications, 214, 119160.
28. Raza, M. A., Abolhasan, M., Lipman, J., Shariati, N., Ni, W., & Jamalipour, A. (2024). Multi-Agent Multi-Armed Bandit Learning for Grant-Free Access in Ultra-Dense IoT Networks. IEEE Transactions on Cognitive Communications and Networking.
29. Anantharam, V., Varaiya, P., & Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part IId rewards. IEEE Transactions on Automatic Control, 32(11), 968-976.
30. Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society Series B: Statistical Methodology, 41(2), 148-164.
31. Agrawal, R. (1995). Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. Advances in applied probability, 27(4), 1054-1078.
32. Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47, 235-256.
33. Garivier, A., & Moulines, E. (2011, October). On upper-confidence bound policies for switching bandit problems. In International conference on algorithmic learning theory (pp. 174-188). Berlin, Heidelberg: Springer Berlin Heidelberg.
34. Besbes, O., Gur, Y., & Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. Advances in neural information processing systems, 27.
35. Shi, C., & Shen, C. (2021). Multi-player multi-armed bandits with collision-dependent reward distributions. IEEE Transactions on Signal Processing, 69, 4385-4402.
36. Mohamed, E. M., Hashima, S., Aldosary, A., Hatano, K., & Abdelghany, M. A. (2020). Gateway selection in millimeter wave UAV wireless networks using multi-player multi-armed bandit. Sensors, 20(14), 3947.
37. Dakdouk, H., Féraud, R., Varsier, N., Maillé, P., & Laroche, R. (2023). Massive multi-player multi-armed bandits for IoT networks: An application on LoRa networks. Ad Hoc Networks, 151, 103283.
38. Ortner, R. (2007). Pseudometrics for state aggregation in average reward Markov decision processes. In Algorithmic Learning Theory: 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007. Proceedings 18 (pp. 373-387). Springer Berlin Heidelberg.
39. Raj, V., & Kalyani, S. (2017). Taming non-stationary bandits: A Bayesian approach. arXiv preprint arXiv:1707.09727.
40. Herlihy, C., & Dickerson, J. P. (2023, June). Networked restless bandits with positive externalities. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 11997-12004).
41. Liu, K., & Zhao, Q. (2010). Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. IEEE Transactions on Information Theory, 56(11), 5547-5567.
42. Chen, S., Tao, Y., Yu, D., Li, F., & Gong, B. (2021). Distributed learning dynamics of multi-armed bandits for edge intelligence. Journal of Systems Architecture, 114, 101919.
43. Gillman, D. (1998). A Chernoff bound for random walks on expander graphs. SIAM Journal on Computing, 27(4), 1203-1220.
44. Sheikh, Chaand (2021). Upper confidence bound dataset. Available online: https://www.kaggle.com/datasets/chaandsheikh/upper-confidence-bound-dataset.