LEARNING INFINITE-HORIZON RESTLESS MULTI-ARMED BANDITS WITH NON-STATIONARY DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Online restless multi-armed bandits (RMABs) typically assume that each arm evolves according to a Markov Decision Process (MDP) with stationary state transitions and rewards. However, in real-world applications such as healthcare and recommendation systems, these assumptions are often violated by non-stationary dynamics, posing significant challenges for traditional RMAB algorithms. In this work, we investigate N-armed online infinite-horizon RMABs with non-stationary transitions constrained by a total variation budget B. We propose NS-Whittle, a novel algorithm that integrates arm-specific sliding window reinforcement learning with an upper confidence bound mechanism to simultaneously estimate evolving transition dynamics and manage exploration-exploitation trade-offs. We establish that NS-Whittle achieves a regret bound of $\tilde{\mathcal{O}}(N^2B^{1/4}T^{3/4})$ by leveraging a relaxed definition of regret. This work provides the first foundational theoretical framework for non-stationary RMAB problems.

1 INTRODUCTION

RMABs provide a powerful framework for sequential decision-making in dynamic environments. An RMAB instance consists of N arms, each evolving independently as an MDP, whose state continues to evolve according to its underlying dynamics, irrespective of whether it is chosen. Unlike traditional multi-armed bandits (MABs), RMABs allow for non-i.i.d. reward distributions that depend on time-varying states, making them well-suited for real-world applications such as queuing systems Ansell et al. (2003); Nino-Mora (2002), healthcare interventions Mate et al. (2022; 2020), and wireless networks Wei et al. (2010); Sun et al. (2018). However, in practical scenarios, these dynamics are often unknown a priori, transforming RMABs into an online learning problem requiring a balance between exploration and exploitation with regret guarantees.

Recent work has made strides in developing algorithms for online RMABs under the assumption of stationary dynamics Wang et al. (2023). However, this assumption often fails to hold in real-world settings. For example, in healthcare applications, patient responses to treatments can evolve over time, while external factors like changing user preferences or environmental influences introduce non-stationarity. This necessitates developing algorithms that can handle non-stationary dynamics.

To address the non-stationary dynamics, Cheung et al. (2020) proposed a sliding window approach to handle non-stationary MDPs. A straightforward idea is to treat the entire RMAB problem as a single large MDP and apply the same method. While this approach simplifies the learning framework, it introduces significant computational challenges, as it requires maintaining and updating transition estimates for a joint state space that grows exponentially with the number of arms N. This compounded complexity results in the curse of dimensionality, making such methods intractable for large-scale, high-dimensional RMAB problems.

Most RMAB settings impose budget constraints on the number of selected arms per round, adding another layer of complexity. Under such constraints, the definition of regret becomes less straightforward, accounting for both the non-stationary dynamics and the time-varying nature of the Lagrange multiplier. This means the underlying constrained MDP is no longer directly compatible with the original algorithm proposed in Cheung et al. (2020). This challenge is further compounded by the fact that non-stationary RMABs require a dynamically changing Bellman equation, where the value function, policy, and Lagrange multiplier all adapt over time. At this point, a natural question arises:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Can we design a computationally tractable algorithm that can operate in a non-stationary restless multi-armed bandit environment?

This paper provides a positive answer to this question. We first resolve the lack of a well-defined dynamic regret in non-stationary RMABs by formally characterizing it through non-stationary value functions. By leveraging proxy and stationary subproblems in our theoretical analysis, our framework captures the essential optimality gap while significantly reducing computational complexity, bridging the gap between theory and practice for non-stationary RMABs.

Our approach builds on two key insights:

- **Arm-Specific Sliding Windows**: Instead of treating the entire RMAB as a single large MDP, we maintain transition estimates separately for each arm. This significantly reduces the computational burden, enabling more localized learning and avoiding the exponential blow-up in the joint state space. This architecture further facilitates arm-specific sliding windows, where each window size W_i can be independently optimized to track transition dynamics of its corresponding arm.
- **Non-stationary optimistic Whittle index**: We apply the minimax theorem to compute the optimistic non-stationary transition probabilities in the confidence intervals estimated by sliding windows. The optimistic transition probabilities later are used to compute the non-stationary Whittle index and update the pulling threshold.

These techniques significantly reduce both the dimensionality and computational overhead, making the approach scalable even in large, high-dimensional, non-stationary environments.

Crucially, unlike treating the entire RMAB as a single large MDP with state space $\mathcal{S}^N$, which leads to regret bounds that scale exponentially with N, our approach maintains only a quadratic dependence on N by decoupling the joint optimization across arms. This effectively avoids the curse of dimensionality by leveraging the weak dependency between arms, providing a scalable and theoretically grounded solution to non-stationary RMABs. Furthermore, our approach achieves a regret bound of $\widetilde{\mathcal{O}}(N^2|\mathcal{S}|^{\frac{1}{2}}B^{\frac{1}{4}}T^{\frac{3}{4}}\sqrt{\log(1/\delta)})$ with probability at least $1-\delta$, where N is the number of arms and $|\mathcal{S}|$ represents the shared state space of the independent MDPs. This is a significant improvement over the naive approach of treating the RMAB as a single non-stationary MDP to use the regret bound in Cheung et al. (2020), which would result in a much larger regret bound of $\widetilde{\mathcal{O}}(|\mathcal{S}|^{\frac{2N}{3}}B^{\frac{1}{4}}T^{\frac{3}{4}})$ due to the exponentially large state space.

In particular, when budget $B = 0$, i.e., the transition is stationary, our regret bound reduces to the best known regret bound $\widetilde{O}(T^{\frac{1}{2}})$ for the stationary non-episodic and episodic RMAB settings as shown in Wang et al. (2023); Jung & Tewari (2019); Jung et al. (2019), effectively bridging the gap between the non-stationary case and the stationary case.

We further validate our approach through simulation experiments, demonstrating that incorporating arm-specific sliding windows in non-stationary RMABs provides significant performance gains. Compared to algorithms that use the entire history, such as Online Whittle Wang et al. (2023), NS-Whittle adapts more effectively to changing dynamics and achieves lower cumulative regret.

2 Related Work

Offline and Online RMABs Restless multi-armed bandits (RMABs) Whittle (1988) generalize MABs Slivkins et al. (2019) by modeling N arms as independent MDPs with K concurrent pulls. Offline RMAB optimization is PSPACE-hard Papadimitriou & Tsitsiklis (1999), favoring the Whittle index policy Weber & Weiss (1990), which is asymptotically optimal under indexability. In online stationary RMABs, frequentist regret bounds of $\tilde{O}(N|S|\sqrt{T\log T})$ have been achieved Wang et al. (2023). However, direct application of standard MDP algorithms to RMABs typically incurs exponential regret due to the joint state space.

Non-stationary MDPs and RMABs For non-stationary infinite-horizon MDPs, sliding-window mechanisms achieve $O(T^{3/4})$ regret Cheung et al. (2020). While recent work in episodic non-stationary RMABs Shisher et al. (2025) attains $\mathcal{O}(T^{4/5})$ regret , our research addresses the more rigorous infinite-horizon setting.

We provide a more detail discussion for related work in Appendix I. The primary challenge in infinite-horizon environments is the lack of periodic resets or mixing-time guarantees. In non-stationary settings, underlying transition dynamics often shift before the state distribution can reach equilibrium, rendering standard mixing assumptions inapplicable. Without a mechanism for recovery, agents risk becoming trapped in suboptimal states. We provide the first theoretical regret bound for non-stationary RMABs that scales polynomially with N by applying martingale concentration inequalities directly to value function differences.

3 Problem Statement

We consider an RMAB problem with N restless arms over a horizon T. Each arm $i \in [N]$ is modeled as a non-stationary discounted MDP $\mathcal{M}_i := (\mathcal{S}, \mathcal{A}, \gamma, R_i, \{P_{t,i}\}_{t=1}^T)$, where $\mathcal{S}$ is the state space, $\mathcal{A} = \{0, 1\}$ is the action space (1 for active and 0 for passive), and $\gamma \in [0, 1)$ is the discount factor. The reward $R_i(s_{t,i}, a_{t,i}) \leq 1$ is stationary and known. To capture environmental shifts, we define a variation budget $B := \sum_{t=1}^T \sum_{i=1}^N B_{t,i}$, where $B_{t,i} := \max_{s,a} \|P_{t+1,i}(\cdot|s, a) - P_{t,i}(\cdot|s, a)\|_1$ captures the transition dynamics' change between steps.

To facilitate regret analysis, we represent the system as a unified joint MDP. At each step $t \in [T]$, the learner observes the global state $\boldsymbol{s}_t := (s_{t,1}, \ldots, s_{t,N}) \in \mathcal{S}^N$ and selects a global action $\boldsymbol{a}_t := (a_{t,1}, \ldots, a_{t,N}) \in \{0, 1\}^N$ satisfying the budget constraint $\sum_{i=1}^N a_{t,i} \leq K$. The joint dynamics, rewards, and policies are defined by the aggregation of individual arms: (i) Joint Transition: $\boldsymbol{P}_t(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) := \prod_{i=1}^N P_{t,i}(s_{t+1,i}|s_{t,i}, a_{t,i})$. (ii) Joint Reward: $\boldsymbol{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) := \sum_{i=1}^N R_i(s_{t,i}, a_{t,i})$. (iii) Joint Policy: $\pi_t(\boldsymbol{a}_t \mid \boldsymbol{s}_t) = \prod_{i=1}^N \pi_{t,i}(a_{t,i}|s_{t,i})$. Based on the joint MDP representation, we can further define the constrained optimization objective for RMABs and the relaxed regret proxy.

3.1 Constrained Optimization Objective

Based on the joint MDP formulation, objective of RMAB is to find an optimal joint policy π that maximizes the expected total discounted reward while strictly satisfying the activation budget K at each time step. This can be formally stated as the following constrained optimization problem:

$$\max_{\pi} \mathbb{E}_{\boldsymbol{s}_{t+1} \sim \boldsymbol{P}_t(\boldsymbol{s}_t, \boldsymbol{a}_t), \boldsymbol{a}_t \sim \pi(\boldsymbol{s}_t)} \left[\sum_{t=1}^T \gamma^{t-1} \boldsymbol{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) \mid \boldsymbol{s}_0 = \boldsymbol{s} \right] \text{ s.t. } \|\pi(\boldsymbol{s})\|_1 \leq K, \ \forall \boldsymbol{s} \in \mathcal{S}^N \quad (1)$$

The constraint $\|\pi(\boldsymbol{s})\|_1 \leq K$ couples the actions of all arms, making the problem computationally intractable as the joint state space size $|\mathcal{S}^N|$ grows exponentially with N (Papadimitriou & Tsitsiklis, 1987). To overcome this curse of dimensionality, we employ the Lagrange relaxation technique. By introducing a dual variable λ that penalizes budget violations, we relax the per-step constraint and incorporate it into the reward function. This relaxation transforms the intractable coupled problem into a form that allows for decoupling across arms, motivating the relaxed value functions defined in the following section.

3.2 Optimal Policy and Regret Relaxation

To evaluate sequential decision-making in non-stationary RMABs, we characterize policy performance through value functions under Lagrangian relaxation. Specifically, for an arbitrary policy π, the true non-stationary Lagrangian value function evaluated from time step t at state s under a given sequence of dual multipliers $\{\lambda_\tau\}_{\tau=t}^T$ is defined as:

$$J^{\pi}_{\boldsymbol{P}_{t:T}, \lambda_{t:T}}(\boldsymbol{s}) := \mathbb{E}_{\pi, \{\boldsymbol{P}_\tau\}_{\tau=t}^T} \left[\sum_{\tau=t}^T \gamma^{\tau-t} \boldsymbol{R}_{\lambda_\tau}(\boldsymbol{s}_\tau, \boldsymbol{a}_\tau) \,\middle|\, \boldsymbol{s}_t = \boldsymbol{s} \right], \quad (2)$$

where $R_\lambda(\boldsymbol{s}_t, \boldsymbol{a}_t) := \boldsymbol{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) - \lambda(\sum_{i=1}^N \boldsymbol{a}_{t,i} - K)$ is the penalized Lagrangian reward incorporating the per-step activation budget constraint K through the dual multiplier $\lambda \geq 0$, and $\{\lambda_\tau^*\}$ denotes the sequence of dual variables.

Across the single continuous trajectory $\{\boldsymbol{s}_t\}_{t=1}^T$ generated under the sequence of deployed policies $\{\pi_t\}_{t=1}^T$, the standard dynamic regret measuring cumulative suboptimality under dual multipliers

$\{\lambda_t\}_{t=1}^T$ is formulated as:

$$\mathrm{Reg}_{\mathrm{dyn},\lambda_{t:T}}(T) := \sum_{t=1}^{T} \left(\max_{\pi} J^{\pi}_{\boldsymbol{P}_{t:T},\lambda_{t:T}}(\boldsymbol{s}_t) - J^{\pi_t}_{\boldsymbol{P}_{t:T},\lambda_{t:T}}(\boldsymbol{s}_t) \right), \tag{3}$$

where $\lambda_{t:T} := (\lambda_t, \ldots, \lambda_T)$ denotes the sequence of dual multipliers over the remaining horizon. Rather than relying on immediate reward differences $\sum_{t=1}^{T}(\boldsymbol{R}(\boldsymbol{s}_t^*, \boldsymbol{a}_t^*) - \boldsymbol{R}(\boldsymbol{s}_t, \boldsymbol{a}_t))$, where $\boldsymbol{s}_t^*$ is the state visited under an optimal policy, evaluating regret via value-function gaps is essential in non-episodic discounted settings to penalize myopic decisions that steer the system into low-return trapping states. In this formulation, the dual multiplier sequence $\lambda_{t:T}$ is introduced to directly capture and penalize per-step activation constraint violations at each decision step t, thereby unifying performance optimality and budget feasibility under the Lagrangian framework. Furthermore, the value functions are truncated at the finite terminal horizon T rather than extending to infinity: in a non-stationary environment, defining an infinite-horizon optimal policy would inherently require non-causal foresight into unknown future transition dynamics beyond step T, which is physically unrealizable in online learning. This definition measures the cumulative performance improvement achievable from each visited state $\boldsymbol{s}_t$. Our formulation also aligns with established benchmarks in online reinforcement learning and the sample complexity of exploration, which seek to minimize value function discrepancies. Moreover, in the special case of a single arm ($N = 1$) under stationary dynamics ($B = 0$), our metric in (3) degenerates exactly to the canonical infinite-horizon MDP regret definition established in Zhou et al. (2021) and He et al. (2022).

4 Methodology

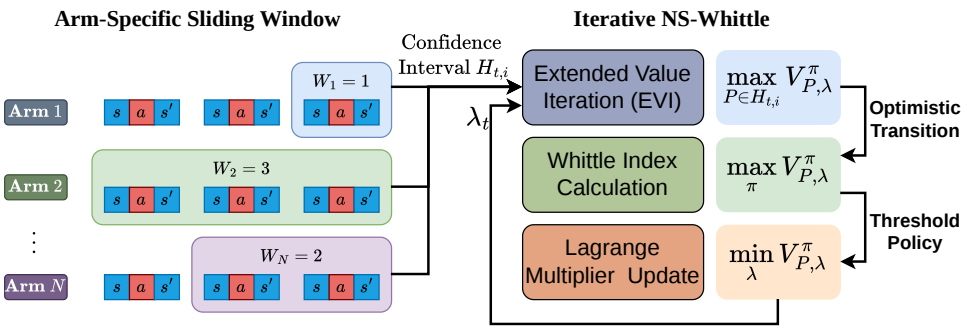


Figure 1: Each arm i maintains an independent sliding window W_i, and using recent transitions to construct arm-specific confidence intervals $H_{t,i}$. The algorithm iteratively solves the min-max-max problem: (i) EVI identifies the optimistic transitions within $H_{t,i}$. (ii) Whittle Index Calculation decouples the joint optimization to determine the optimal policy π_t, and (iii) Lagrange Multiplier Update enforces the global budget constraint K by setting λ_t as the K-th largest index.

The core methodology of NS-Whittle involves solving the Lagrangian relaxation while employ arm-specific sliding windows to control how much historical data is used for estimating transition probabilities. We first detail the sliding-window mechanism in Section 4.1, and introduce a prototype of the algorithm in Section 4.2. However, directly using this prototype algorithm is computationally inefficient due to (i) the exponential growth of the joint MDP state space and (ii) the complexity of the associated minimax problem. To address these challenges, Section 4.2 introduces a decomposition leveraging the indexability property, and Section 4.3 outlines an iterative learning framework to efficiently solve the minimax problem (8). The overall pipeline is illustrated in Figure 1.

4.1 Sliding Windows Estimation with UCB

We adopt the sliding-window framework from Cheung et al. (2020) to estimate optimistic transition probabilities. Each arm i is associated with a window size $W_i \in \mathbb{N}$, which specifies the number of most recent time steps used to estimate the unknown transition dynamics, rather than relying on

the entire history. For arm i at time t, we define the following quantities over window size W_i: the counter, the empirical estimator, and the average transition probability for all $a \in \mathcal{A}, s \in \mathcal{S}$,

$$\widetilde{P}_{t,i}(s'|s,a) = \sum_{q=(t-1-W_i)^+}^{t-1} \mathbb{1}(s_{q,i} = s, a_{q,i} = a, s_{q+1,i} = s')/N_{t,i}^+(s,a), \tag{4}$$

$$\hat{P}_{t,i}(s'|s,a) = \sum_{q=(t-1-W_i)^+}^{t-1} P_{q,i}(s'|s,a)\mathbb{1}(s_{q,i} = s, a_{q,i} = a)/N_{t,i}^+(s,a). \tag{5}$$

where $N_{t,i}^+(s,a) = \sum_{q=(t-1-W_i)^+}^{t-1} \mathbb{1}(s_{q,i} = s, a_{q,i} = a)$ is the visit counter of the state-action pair (s,a) over the sliding window W_i and the operator x^+ denotes $\max\{x, 1\}$. Notably, $\widetilde{P}_{t,i}(s'|s,a)$ is an unbiased estimator of $\hat{P}_{t,i}(s'|s,a)$. We define the confidence interval used in UCB, along with the corresponding event, which ensures that the average transition lies within this interval, as follows:

$$H_{t,i}(s,a,\eta) = \Big\{ P \in \Delta(\mathcal{S}) \;\Big|\; \|P(\cdot|s,a) - \widetilde{P}_{t,i}(\cdot|s,a)\| \le \text{rad}_{t,i}(s,a) + \eta \Big\}, \tag{6}$$

where $\text{rad}_{t,i}(s,a) = \sqrt{2|S|\log(|S||A|T/\delta)/N_{t,i}^+(s,a)}$ is the radius of the confidence interval and $\{\eta_i\}_{i=1}^N \ge 0$ is an exploration parameter introduced to incorporate confidence interval in the transition estimates, ensuring sufficient exploration.

4.2 A Prototype of NS-Whittle

Directly computing the optimal policy to minimize (3) is intractable in an online setting, as it requires non-causal foresight of future transition dynamics across the entire trajectory. To circumvent this difficulty, we introduce a stationary MDP surrogate. For a fixed transition kernel P and Lagrange multiplier λ, the stationary discounted value function under policy π is defined as:

$$V_{\boldsymbol{P},\lambda}^{\pi}(\boldsymbol{s}) := \mathop{\mathbb{E}}_{(\boldsymbol{s},\boldsymbol{a})\sim(P,\pi)} \left[\sum_{t=0}^{T} \gamma^t \boldsymbol{R}_\lambda(\boldsymbol{s}_t, \boldsymbol{a}_t) | \, \boldsymbol{s}_1 = \boldsymbol{s} \right], \tag{7}$$

The detailed approximation error induced by this proxy is formally captured in Lemma 5.1. The optimal primal policy π_t^* and the corresponding dual variable λ_t^* at time step t are obtained by solving the relaxed Lagrangian:

$$\pi_t^*, \lambda_t^* = \arg\max_{\pi} \min_{\lambda} V_{\boldsymbol{P}_t,\lambda}^{\pi}(\boldsymbol{s}). \tag{8}$$

Given the confidence interval defined in Section 4.1, a prototype of NS-Whittle; can be formulated by estimating the optimistic transition within this interval:

$$\pi_t = \operatorname*{argmax}_{\pi} \max_{\boldsymbol{P}\in\{H_{t,i}(s,a,\eta_i)\}_{i=1}^N} \min_{\lambda} V_{\boldsymbol{P},\lambda}^{\pi}(\boldsymbol{s}). \tag{9}$$

Once the outer minimization over λ is fixed, the inner double maximization can be efficiently solved using the Extended Value Iteration (EVI) Auer et al. (2008). Specifically, for all $s \in \mathcal{S}, a \in \mathcal{A}, i \in [1, N]$, we initialize $\bar{Q}_{\lambda,i}^{(0)}(s,a) = 0$ and update for U iterations as

$$\bar{Q}_{\lambda,i}^{(u+1)}(s,a) = -\lambda a + R_i(s,a) + \gamma \max_{P\in H_{t,i}(s,a,\eta_i)} \sum_{s'} P(s'|s,a) \max_{a'} \bar{Q}_{\lambda,i}^{(u)}(s',a'). \tag{10}$$

Whittle index threshold policy Indexability, defined in Definition 4.1, provides a threshold-based relaxation that allows updating λ without directly solving the original minimization problem, which is computationally challenging due to its approximation via EVI. We define $\bar{Q}_{\lambda,i}^{(U)}(s,1)$ as the *Whittle index* for arm i. This relaxation allows each arm to be optimized independently according to its own dynamics, while the penalty term implicitly enforces the overall budget constraint. Consequently, the original high-dimensional constrained MDP is decomposed into a set of lower-dimensional per-arm problems, significantly enhancing computational tractability. Under indexability, we can approximate the policy π_t by selecting the K arms with the highest Whittle indices at each time step.

Definition 4.1 (Indexability Mate et al. (2020); Mittal et al. (2024)). *For arm i, define*

$$B_i(\lambda) = \left\{ s \in \mathcal{S} : Q^{\pi^*_\lambda}_{P,\lambda}(s,0) \geq Q^{\pi^*_\lambda}_{P,\lambda}(s,1) \right\}, \tag{11}$$

where $\pi^*_\lambda := \operatorname{argmax}_\pi Q^\pi_{P,\lambda}$ is the optimal policy for a given subsidy λ. We can rewrite $B(\lambda)$ as: $B_i(\lambda) = \{s \in \mathcal{S} : \pi^*_\lambda(s) = 0\}$. An RMAB problem is indexable if $B_i(\lambda)$ is non-decreasing in λ for all $i \in [1, N]$, i.e. $B_i(\lambda_1) \subseteq B_i(\lambda_2)$ if $\lambda_1 \leq \lambda_2$.

Indexability is essential for the computational tractability of large-scale RMABs Wang et al. (2023). It is rigorously established across diverse domains, including public health interventions , priority queueing , and opportunistic wireless access.

4.3 Iterative Algorithm for Solving the Minimax Problem

In this section, we present a practical algorithm to solve the minimax problem in an iterative manner. The full algorithm is provided in Algorithm 1. The procedure consists of two main steps: Given an initial λ_0, at time t, we do

1. **Inner Double Maximization**: For each arm, we apply EVI(10) to solve the inner double maximization problem. This step computes the optimal action values given a fixed penalty λ_{t-1}, effectively capturing the optimal action value function $\bar{Q}^{(U)}_{\lambda_{t-1},i}$ for each arm i.
2. **Outer Minimization:** After obtaining the Whittle indices:

$$\mathcal{W}_i(s_{t,i}) := \inf_\lambda \left\{ \lambda : \bar{Q}^{(U)}_{\lambda,i}(s_{t,i},1) = \bar{Q}^{(U)}_{\lambda,i}(s_{t,i},0) \right\}, \tag{12}$$

   we update the penalty λ_t by setting it to be the Kth highest Whittle index. This step approximates the original minimization by the property of indexability and enforcing the global budget constraint through the threshold λ_t.

These steps are repeated iteratively for all time step, yielding a computationally efficient approximation to the original minimax problem.

5 Theoretical Analysis

Recalling that NS-Whittle relies on the snapshot-optimal proxy policy and dual variables defined in (8), we introduce the following lemma to bound the gap induced by the regret proxy.

Lemma 5.1. *The minimized dynamic regret is bounded in terms of the proxy regret $\mathit{Reg}_{\mathit{proxy}}(T)$ by:*

$$\min_{\{\lambda_t\}_{t=1}^T} \mathit{Reg}_{\mathit{dyn},\lambda_{1:T}}(T) \leq \mathit{Reg}_{\mathit{proxy}}(T) + \frac{2\gamma^2 N(1+U_\lambda)}{(1-\gamma)^3} B + \frac{4N(1+U_\lambda)}{(1-\gamma)^2}, \tag{13}$$

where $\mathit{Reg}_{\mathit{proxy}}(T) := \sum_{t=1}^T \left(V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) \right)$, and π^*_t is the snapshot-optimal policy under dynamics $\boldsymbol{P}_t$ and multiplier λ^*_t as defined in (8)*. The detailed proof is provided in Appendix B*

Remark 5.2. *Lemma 5.1 establishes that the computationally tractable proxy regret provably serves as an upper bound for the dynamic regret. Notably, in a stationary environment ($B = 0$): (i) The term $(2\gamma^2 N(1+U_\lambda)B)/((1-\gamma)^3)$ naturally vanishes as transition dynamics remain time-invariant. (ii) The evaluation horizon of the value functions does not need to be truncated at T, as stationary policies operate without requiring non-causal foresight, where the terminal boundary discrepancy D_T is eliminated, which implies $D_t \leq 0$ in (38). Consequently, the constant boundary error $4N(1+U_\lambda)/(1-\gamma)^2$ drops out completely.*

To evaluate the optimality of our proposed NS-Whittle algorithm, we first establish a theoretical lower bound for non-stationary RMABs. By reducing the problem to a class of non-stationary MDPs with variation budget B, we can have the following theorem:

Theorem 5.3. *For any learning algorithm, there exists a non-stationary RMAB instance with N arms and total variation B such that the expected regret satisfies:*

$$\mathbb{E}[\min_{\lambda_{1:T}} \mathrm{Reg}_{\mathrm{dyn},\lambda_{1:T}}(T)] = \Omega(N^{2/3} B^{1/3} T^{2/3}). \tag{14}$$

The detailed proof is provided in Appendix C. There remains a theoretical gap between the $\Omega(T^{2/3})$ lower bound and our $\mathcal{O}(T^{3/4})$ upper bound. Closing this gap is currently an open problem in the study of non-stationary RMABs. The primary difficulty stems from the cumulative changes in a non-stationary environment, which cause historical data to become "stale," meaning past observations no longer provide accurate information for estimating current transition dynamics. This effect is particularly challenging in the infinite-horizon discounted setting where there are no periodic resets to clear outdated information. Achieving a regret bound that matches the $\Omega(T^{2/3})$ limit would likely require more advanced techniques to balance the trade-off between information freshness and estimation variance.

5.1 Main Results

We outline the regret analysis through three primary stages: First, in Lemma 5.1, we bridge the theoretical gap between the standard dynamic regret and the proxy regret, bounding their tracking error and establishing a tractable surrogate for the subsequent analysis. Second, in Section 5.2, we deploy a regret substitution technique to isolate the performance discrepancy resulting from differing transition dynamics under the same execution policy. Finally, in Section 5.3, we bound the proxy regret under the high-probability good event by exploiting the optimism inherent in the UCB framework and applying Hoeffding's inequality to control the martingale difference sequences.

Theorem 5.4. *With probability at least $1 - 6\delta' - N/\delta$, we have*

$$\min_{\lambda_{1:T}} Reg_{dyn,\lambda_{1:T}}(T) = \mathcal{O}\Bigg(\frac{U_\lambda N^2 BW}{\eta} + U_\lambda N\sqrt{T\log\frac{1}{\delta'}} + U_\lambda T N^2\eta + \frac{U_\lambda T N^2\,|S|}{\sqrt{W}}\sqrt{\log\left(|S|\,T/\delta\right)} + \frac{2\gamma^2 N(1+U_\lambda)}{(1-\gamma)^3}B + \frac{4N(1+U_\lambda)}{(1-\gamma)^2}\Bigg). \tag{15}$$

Theorem 5.4 establishes a sub-linear regret bound for NS-Whittle. The detailed proof is provided in Appendix F. Notice that U_λ represents a uniform upper bound for the optimal dual variables $\{\lambda_t^*\}_{t=1}^T$. It is critical to establish the boundedness, as a diverging dual variable would lead to the collapse by preventing the convergence of the relaxed value functions. To address this, we provide a formal derivation of this constant bound in Lemma 5.10.

Remark 5.5. *To optimize the regret bound, we can choose $W_i^* = |S|T^{1/2}(\sum_{t=1}^T B_{t,i})^{-\frac{1}{2}}, \eta_i^* = \sqrt{BW_i^*/T}$ for arm i. The regret reduces to $Reg_{\lambda_t^*}(T) = \widetilde{\mathcal{O}}(N^2|\mathcal{S}|^{\frac{1}{2}}B^{\frac{1}{4}}T^{\frac{3}{4}}\sqrt{\log 1/\delta})$. For the case where T is unknown, we can employ the doubling trick Besson & Kaufmann (2018), which adaptively estimates T on a exponatial timescale and increases the regret by at most a factor of $\log(T)$.*

Remark 5.6. *When the environment is stationary, i.e., $B = 0$, we can choose $W = T, \eta = 1/T$ and the regret bound reduces to $\widetilde{\mathcal{O}}(N^2|S|T^{\frac{1}{2}})$, which matches the best known non-episodic and episodic regret in stationary RMABs shown in Wang et al. (2023); Jung & Tewari (2019); Jung et al. (2019).*

Remark 5.7. *In the special case where the environment comprises a single arm (i.e., $N = 1$), the problem reduces to an MDP with a two-dimensional action space. Under this setting, our regret bound aligns with the result of Cheung et al. (2020), achieving a rate of $\mathcal{O}(T^{3/4})$.*

5.2 Regret Substitution

Lemma 5.8. *Given π_t, and the estimated transitions $\bar{P}_t$, $\boldsymbol{P}_t$, λ_t^* are the true transition and optimal dual variable, respectively. With probability at least $1 - N/\delta$, We have*

$$V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t^*}(\boldsymbol{s}_t) \geq V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t). \tag{16}$$

Proof. We have

$$V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t^*}(\boldsymbol{s}_t) \underset{(i)}{\geq} V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t}(\boldsymbol{s}_t) \underset{(ii)}{\geq} V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t}(\boldsymbol{s}_t) \underset{(iii)}{\geq} V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t), \tag{17}$$

where (i) holds due to λ_t is the minimizer of $V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_t)$, (ii) follows from the fact that $\pi_t, \bar{\boldsymbol{P}}_t$ maximize $V^{\pi}_{\boldsymbol{P},\lambda_t}(\boldsymbol{s}), \forall\, \boldsymbol{s} \in \mathcal{S}^N$, and (iii) holds because λ_t^* is the minimizer of $V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_t)$. □

Lemma 5.8 provides a value function substitution that decouples decision error from transition estimation error, allowing us to reformulate the original regret into the form Reg'(T):

$$\mathrm{Reg}'(T) := \sum_{t=1}^{T} \left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t, \lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t, \lambda_t^*}(\boldsymbol{s}_t) \right). \tag{18}$$

By isolating the difference between the optimistic transitions $\bar{\boldsymbol{P}}_t$ and the true dynamics $\boldsymbol{P}_t$, this reformulation significantly simplifies the analysis. This gap can subsequently be bounded using the confidence interval radius, yielding a more tractable expression for the regret bound.

5.3 Upper bound of $\mathrm{Reg}'(T)$

Applying Hoeffding's inequality to $\mathrm{Reg}'(T)$ requires that the underlying random variables, which is the value functions with Lagrange relaxation in this case, be bounded. To guarantee this, Lemma 5.10 ensures that the optimal Lagrange multipliers remain uniformly bounded throughout the time horizon. Without such a bound, the optimal dual variables could potentially diverge, causing the regret to scale uncontrollably and undermining the feasibility of a sub-linear performance guarantee.

The analysis of the dual variables begins with strong duality assumption to the RMAB problem. By ensuring the primal and dual optimal values coincide, we can guarantee that solving the relaxed problem with λ^* yields a policy that is indeed optimal for the original constrained objective (1).

Assumption 5.9. *We have strong duality applied on the RMAB problem (1), the optimal solution is given by* $\pi^* = \mathrm{argmax}_\pi \min_{\lambda \geq 0} L(\pi, \lambda, s) := V^\pi(s) - \sum_{s' \in \mathcal{S}} \lambda(s') \cdot (\|\pi(s')\|_1 - K)$*, where* $L(\pi, \lambda, s)$ *is the corresponding Lagrangian.*

Assumption 5.9 is technically necessary to ensure the Lagrangian relaxation effectively bridges the decoupled subproblems with the global budget constraint. This requirement is consistent with recent state-of-the-art research Chen et al. (2026), which also utilizes strong duality within the context of unichain MDPs.

Lemma 5.10. *Suppose there exists a strictly feasible policy* $\bar{\pi}$ *such that* $\|\bar{\pi}(s')\|_1 < K$ *for all* $s' \in \mathcal{S}$*. Then, for all* $t \in [T]$*, the optimal dual variable* λ_t^* *is uniformly bounded by a constant* $U_\lambda \geq 0$*, such that* $\lambda_t^* \leq U_\lambda$*.*

The detailed proof is provided in Appendix G. By leveraging the Slater's condition, we can derive a uniform upper bound U_λ for the optimal dual variables λ_t^*.

Lemma 5.11. *For all* $t \in [T]$*, and* $\boldsymbol{P} \in \mathcal{P}$*, Then, for all* $\boldsymbol{s} \in \mathcal{S}^N$*, we have*

$$|V^{\pi_t}_{\boldsymbol{P}, \lambda_t^*}(\boldsymbol{s})| \leq N(1 + U_\lambda)/(1 - \gamma), \tag{19}$$

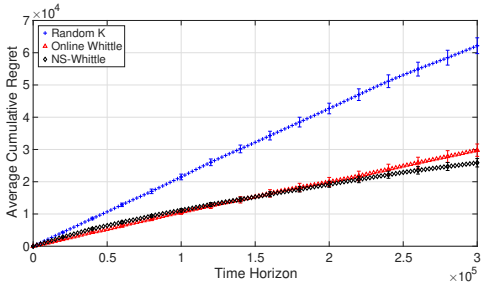
where $\mathcal{P}$ *is the set of joint transitions for all arms. The proof is provided in Appendix A.*

The following lemma provides an upper bound of $\mathrm{Reg}'(T)$ under the good event.

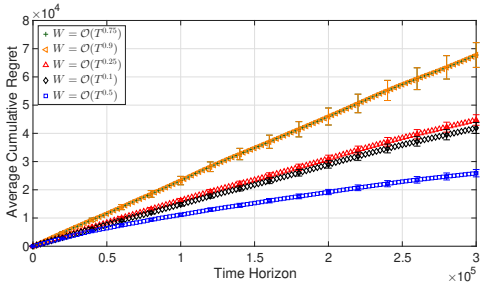
Lemma 5.12. *Given* $P_{t,i}(\cdot|s,a) \in H_{t,i}(s,a,\eta), \forall (s \in \mathcal{S}^N, a \in \mathcal{A}^N, i \in [1,N], t \in [1,T])$*, with probability at least* $1 - 6\delta' - N/\delta$*, we have*

$$\mathrm{Reg}'(T) \leq \mathcal{O}\left(\frac{U_\lambda}{p_{min}} \left(N\sqrt{T \log \frac{1}{\delta'}} + TN^2\eta + \frac{TN^2\,|S|}{\sqrt{W}} \sqrt{\log\left(|S|\,T/\delta\right)} \right) \right), \tag{20}$$

where we assume non-zero transitions are bounded below by a constant $p_{\min} > 0$ (Assumption E.1). The complete derivation is provided in Appendix E. This lemma leverages the Azuma-Hoeffding inequality to bound the martingale difference sequences inherent in the regret formulation. A main technical challenge in non-stationary environments is that the state distribution typically fails to converge to a stationary one, making standard mixing-time assumptions used in episodic RL analysis are not applicable here. By applying martingale techniques directly to the value function differences, our analysis successfully overcomes the absence of periodic resets and mixing guarantees, representing a central theoretical contribution of this work.



(a) Cumulative regret across baseline methods



(b) Cumulative regret across window size settings

Figure 2: Cumulative regret averaged over 50 trials with $T = 3 \times 10^5$, $|\mathcal{S}| = 2$, $K = 2$, and $N = 10$ on non-stationary restless bandits. The perturbation on the logit of transition dynamics is sampled from $\mathcal{N}(0, 0.1 \cdot \mathbf{I}_{|\mathcal{S}|})$, and the error bars indicate the standard error.

6 Experiments

To evaluate the performance of the proposed iterative algorithm in Section 4.3, we conduct experiments on a non-stationary RMAB problem with $N = 10$ arms under the indexability assumption.

Environment The experimental environment utilizes a non-stationary RMAB framework where indexability is maintained through rejection sampling and transition dynamics evolve via periodic Gaussian noise applied to logits. Performance is assessed using surrogate regret relative to an optimal Whittle index policy computed with true transition probabilities, with comprehensive configuration and hyperparameter details provided in Appendix H.

Baselines We compare NS-Whittle against the following baselines: (i) **Online Whittle Index**: Adapted from the episodic restless bandit algorithm in Wang et al. (2023). For a fair comparison in the non-episodic setting, we modify it by concatenating all episodes into a single, continuous sequence. (ii) **Random K arms selection** A simple strategy that randomly selects K arms at each time step, ignoring both the arm states and their history, providing a straightforward baseline for comparison.

Effectiveness. Figure 2a illustrates the effectiveness of NS-Whittle in terms of cumulative regret. We can observe that NS-Whittle outperforms Online Whittle, a strong baseline for stationary online RMABs that uses the entire history to estimate transition probabilities. In non-stationary environments, using all past data can bias the estimates, since early observations are less informative and differ significantly from later transitions, limiting the Effectiveness of Online Whittle. NS-Whittle also demonstrates greater robustness, exhibiting lower standard deviation compared to all other baselines.

Effect of Window Size. Figure 2b compares the performance of NS-Whittle under different choices of W relative to the horizon T. The results indicate that the optimal choice is $W = \mathcal{O}(T^{0.5})$, consistent with the discussion in Remark 5.5. We adopt the exact form of the pre-constant given in Remark 5.5, modifying only the order in T. Note that the constant factor of the window size was not fine-tuned; the big-$\mathcal{O}$ notation is used here for simplicity.

Unavailable Budget Information Figure 3 presents empirical results under misspecified variation-budget scales. In these experiments, NS-Whittle uses $M \times B_i$ as the budget parameter for each arm i i to determine both its window size and η_i. Across the five different choices of $M = 0.25, 0.5, 1.0, 2.0, 4.0$, the regret curves exhibit similar slopes, and all remain below that of Online Whittle once $t \geq 10^6$. These findings indicate that the true budget serves as an effective upper bound, and that using an underestimated value of M can still achieve competitive performance when the true budget is unknown. NS-Whittle therefore remains robust even without accurate budget information, confirming that the algorithm does not rely on additional oracle knowledge. Note that Online Whittle corresponds to the special case $M = 0$.

7 Conclusion

This work establishes a scalable framework for non-stationary RMABs. We address evolving dynamics while significantly reducing dimensionality through decoupled optimization. Our approach provides a robust foundation for resource-constrained decision-making in dynamic environments.

AI USE STATEMENT

In this work, we used generative AI tools to proofread and polish the writing, and to assist in verifying the mathematical proofs. All AI-assisted text and mathematical derivations were carefully reviewed by the authors, and we take full responsibility for the final content of this paper.

ETHICS STATEMENT

This paper presents foundational theoretical research in reinforcement learning and restless multi-armed bandits. To the best of our knowledge, this work does not raise any direct ethical issues or immediate negative societal impacts.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we have provided all detailed mathematical proofs in the Appendix. Additionally, the complete source code, environments, and hyperparameter configurations used to generate the experimental results are included in the supplementary material.

REFERENCES

PS Ansell, Kevin D Glazebrook, José Nino-Mora, and M O'Keeffe. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research*, 57:21–39, 2003.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*, 2021.

Gongpu Chen, Soung Chang Liew, and Deniz Gündüz. Gino-q: Learning an asymptotically optimal index policy for restless multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 20032–20040, 2026.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International conference on machine learning*, pp. 1843–1854. PMLR, 2020.

Jing Fu, Yoni Nazarathy, Sarat Moka, and Peter G Taylor. Towards q-learning the whittle index for restless bandits. In *2019 Australian & New Zealand Control Conference (ANZCC)*, pp. 249–254. IEEE, 2019.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps, 2022. URL https://arxiv.org/abs/2010.00587.

Young Hun Jung and Ambuj Tewari. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Young Hun Jung, Marc Abeille, and Ambuj Tewari. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*, 2019.

Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*, 2014.

P Kumar and Woei Lin. Optimal adaptive controllers for unknown markov chains. *IEEE Transactions on Automatic Control*, 27(4):765–774, 1982.

Haoyang Liu, Keqin Liu, and Qing Zhao. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory*, 59(3):1902–1916, 2012.

Keqin Liu and Qing Zhao. A restless bandit formulation of opportunistic access: Indexablity and index policy. In *2008 5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*, pp. 1–5. IEEE, 2008.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.

Shuang Liu and Hao Su. Regret bounds for discounted mdps. *arXiv preprint arXiv:2002.05138*, 2020.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International conference on machine learning*, pp. 7447–7458. PMLR, 2021.

Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12017–12025, 2022.

Vishesh Mittal, Rahul Meshram, Deepak Dev, and Surya Prakash. Indexability of finite state restless bandits. In *2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pp. 618–622. IEEE, 2024.

José Nino-Mora. Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Mathematical programming*, 93(3):361–413, 2002.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

Francisco Robledo, Vivek Borkar, Urtzi Ayesta, and Konstantin Avrachenkov. Qwi: Q-learning with whittle index. *ACM SIGMETRICS Performance Evaluation Review*, 49(2):47–50, 2022.

Md Kamran Chowdhury Shisher, Vishrant Tripathi, Mung Chiang, and Christopher G. Brinton. Online learning of whittle indices for restless bandits with non-stationary transition kernels, 2025. URL `https://arxiv.org/abs/2506.18186`.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Yao Sun, Gang Feng, Shuang Qin, and Sanshan Sun. Cell association with user behavior awareness in heterogeneous cellular networks. *IEEE Transactions on Vehicular Technology*, 67(5):4589–4601, 2018.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic whittle index policy: Online learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10131–10139, 2023.

Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of applied probability*, 27(3):637–648, 1990.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*, pp. 4300–4354. PMLR, 2021.

Yifei Wei, F Richard Yu, and Mei Song. Distributed optimal relay selection in wireless cooperative networks with finite-state markov channels. *IEEE Transactions on Vehicular Technology*, 59(5): 2149–2158, 2010.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping, 2021. URL https://arxiv.org/abs/2006.13165.

A USEFUL LEMMAS

Lemma A.1 (Azuma-Hoeffding inequality)**.** *Let $(X_i, \mathcal{F}_i)_{i=1,\cdots,t}$ to be a martingale satisfying $X_i < M, \forall i \in [t]$, for some positive constant M. Then, for any $n \in [t]$, we have*

$$\Pr\left(X_n \leq 2M\sqrt{n \log \frac{1}{\delta}}\right) \geq 1 - \delta \tag{21}$$

Lemma A.2 (Lemma 7, Cheung et al. (2020))**.** *For any $t \notin \widetilde{Q}_T$, $\forall (s, a, i)$, we have*

$$P_{t,i}(\cdot|s, a) \in H_{t,i}(s, a, \eta). \tag{22}$$

Proof. We follow a similar proof as of Lemma 7 in Cheung et al. (2020), based on the construction of $\widetilde{Q}_T$ provided in Section D. Suppose there exists a time step $t \notin \widetilde{Q}_T$ such that $P_{t,i}(\cdot \mid s, a) \notin H_{t,i}(s, a, \eta)$ for some state s, action a and arm i. According to the first condition in the construction of Q_T, such a t should have been included in Q_T. Otherwise, we would have $|t - t'| \leq W_i$ for some $t' \in Q_T$, which contradicts the construction of $\widetilde{Q}_T$. □

Lemma A.3 (Lemma 1, Cheung et al. (2020))**.** *For all $i \in [N]$, we have $\Pr(\xi_i) \geq 1 - \delta/2$.*

Lemma 5.11. *For all $t \in [T]$, and $\boldsymbol{P} \in \mathcal{P}$, Then, for all $\boldsymbol{s} \in \mathcal{S}^N$, we have*

$$|V^{\pi_t}_{\boldsymbol{P},\lambda^*_t}(\boldsymbol{s})| \leq N(1 + U_\lambda)/(1 - \gamma), \tag{23}$$

where $\mathcal{P}$ is the set of joint transitions for all arms.

Proof. By definition of $V^{\pi_t}_{\boldsymbol{P},\lambda^*_t}(\boldsymbol{s})$, we have

$$V^{\pi_t}_{\boldsymbol{P},\lambda^*_t}(\boldsymbol{s}) = \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim(\boldsymbol{P},\pi_t)}\left[\sum_{t\in\mathbb{N}} \gamma^{t-1}\left(\boldsymbol{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) - \lambda^*_t\left(\pi_t(\boldsymbol{s}_t)^\top \mathbb{1} - K\right)\right) | \, \boldsymbol{s}_1 = \boldsymbol{s}\right] \tag{24}$$

$$\leq \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a})\sim(\boldsymbol{P},\pi_t)}\left[\sum_{t\in\mathbb{N}} \gamma^{t-1}\left(\boldsymbol{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) + U_\lambda N\right) | \, \boldsymbol{s}_1 = \boldsymbol{s}\right] \tag{25}$$

$$\leq N(1 + U_\lambda)/(1 - \gamma), \tag{26}$$

where (25) holds by $\pi_t(\boldsymbol{s})^\top \mathbb{1} \leq K$ and Lemma 5.10, and (26) caused by $0 \leq R_i(s, a) \leq 1, \forall s \in \mathcal{S}, a \in \mathcal{A}$, and $i \in [N]$. Therefore, the statement holds. □

B PROOF OF LEMMA 5.1

Lemma 5.1. *The minimized dynamic regret is bounded in terms of the proxy regret $\textit{Reg}_{\textit{proxy}}(T)$ by:*

$$\min_{\{\lambda_t\}_{t=1}^T} \textit{Reg}_{\textit{dyn},\lambda_{1:T}}(T) \leq \textit{Reg}_{\textit{proxy}}(T) + \frac{2\gamma^2 N(1 + U_\lambda)}{(1 - \gamma)^3} B + \frac{4N(1 + U_\lambda)}{(1 - \gamma)^2}, \tag{27}$$

where $\textit{Reg}_{\textit{proxy}}(T) := \sum_{t=1}^T \left(V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t)\right)$, and π^*_t is the snapshot-optimal policy under dynamics $\boldsymbol{P}_t$ and multiplier λ^*_t as defined in (8)*.*

Proof. We first bound the discrepancy between the true non-stationary Lagrangian value function and the stationary snapshot value function under the snapshot-optimal dual sequence $\{\lambda^*_t\}_{t=1}^T$. At step t, for a fix policy π, the true non-stationary value function and the snapshot value function satisfy:

$$J^\pi_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}}(\boldsymbol{s}) = \boldsymbol{R}_{\lambda^*_t}(\boldsymbol{s}, \pi(\boldsymbol{s})) + \gamma \sum_{\boldsymbol{s}'} \boldsymbol{P}_t(\boldsymbol{s}' \mid \boldsymbol{s}, \pi(\boldsymbol{s})) J^\pi_{\boldsymbol{P}_{t+1:T},\lambda^*_{t+1:T}}(\boldsymbol{s}'), \tag{28}$$

$$V^\pi_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) = \boldsymbol{R}_{\lambda^*_t}(\boldsymbol{s}, \pi_t(\boldsymbol{s})) + \gamma \sum_{\boldsymbol{s}'} \boldsymbol{P}_t(\boldsymbol{s}' \mid \boldsymbol{s}, \pi(\boldsymbol{s})) V^\pi_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}'). \tag{29}$$

Subtracting (29) from (28), we obtain:

$$
\begin{aligned}
&J^{\pi}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}}(\boldsymbol{s}) - V^{\pi}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) \\
&= \gamma \mathbb{E}_{\boldsymbol{s}'\sim \boldsymbol{P}_t(\cdot|\boldsymbol{s},\pi(\boldsymbol{s}))}\left[J^{\pi}_{\boldsymbol{P}_{t+1:T},\lambda^*_{t+1:T}}(\boldsymbol{s}') - V^{\pi}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}')\right] \\
&= \gamma \mathbb{E}_{\boldsymbol{s}'\sim \boldsymbol{P}_t(\cdot|\boldsymbol{s},\pi(\boldsymbol{s}))}\left[\underbrace{\left(J^{\pi}_{\boldsymbol{P}_{t+1:T},\lambda^*_{t+1:T}}(\boldsymbol{s}') - V^{\pi}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}')\right)}_{(ii)} + \underbrace{\left(V^{\pi}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}') - V^{\pi}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}')\right)}_{(i)}\right].
\end{aligned} \tag{30}
$$

Bounding Term (i): First, by the definition of π^*_t and λ^*_t in (8), we have:

$$
\begin{aligned}
V^{\pi^*_t}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}) - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) &\le V^{\pi^*_{t+1}}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}) - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) && (31)\\
&= \min_{\lambda\in[0,U_\lambda]} \max_{\pi} V^{\pi}_{\boldsymbol{P}_{t+1},\lambda}(\boldsymbol{s}) - \min_{\lambda\in[0,U_\lambda]} \max_{\pi} V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}) && (32)\\
&\le \sup_{\lambda\in[0,U_\lambda]} \left(\max_{\pi} V^{\pi}_{\boldsymbol{P}_{t+1},\lambda}(\boldsymbol{s}) - \max_{\pi} V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s})\right) && (33)\\
&\le \sup_{\lambda\in[0,U_\lambda]} \max_{\pi} \left(V^{\pi}_{\boldsymbol{P}_{t+1},\lambda}(\boldsymbol{s}) - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s})\right), && (34)
\end{aligned}
$$

where (33) holds by $(\min f - \min g) \le \sup(f-g)$ for all functions f and g. Notice that in the difference on the right-hand side of (34), the dual multiplier λ is identical across both terms, which implies that the per-step Lagrangian rewards coincide: $R_\lambda - R_\lambda = 0$. Therefore, applying the standard Simulation Lemma for MDPs with identical reward functions:

$$
\begin{aligned}
V^{\pi}_{\boldsymbol{P}_{t+1},\lambda}(\boldsymbol{s}) - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}) &\le \frac{\gamma}{1-\gamma} \| \boldsymbol{P}_{t+1} - \boldsymbol{P}_t \|_{1,\infty} \| V^{\pi}_{\boldsymbol{P}_t,\lambda} \|_\infty \\
&\le \frac{\gamma N(1+U_\lambda)}{(1-\gamma)^2} \| \boldsymbol{P}_{t+1} - \boldsymbol{P}_t \|_{1,\infty} \le \frac{\gamma N(1+U_\lambda)}{(1-\gamma)^2} B_t, && (35)
\end{aligned}
$$

where the second inequality utilizes the uniform bound $\|V^{\pi}_{\boldsymbol{P}_t,\lambda}\|_\infty \le \frac{N(1+U_\lambda)}{1-\gamma}$ for all $\lambda \in [0, U_\lambda]$. Substituting this back into (34), for all $\boldsymbol{s} \in \mathcal{S}^N$, we obtain:

$$
V^{\pi^*_t}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}) - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) \le V^{\pi^*_{t+1}}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}) - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) \le \frac{\gamma N(1+U_\lambda)}{(1-\gamma)^2} B_t. \tag{36}
$$

Next, for the execution policy π_t deployed by the algorithm, the global budget constraint is strictly satisfied at every time step: $\sum_{i=1}^N \pi_{t,i}(\boldsymbol{s}) = K, \forall s$. Thus, the Lagrangian penalty term vanishes identically: $\lambda(\sum_{i=1}^N \pi_{t,i}(\boldsymbol{s}) - K) = 0$. Consequently, $R_{\lambda^*_{t+1}}(s, \pi_t(\boldsymbol{s})) = R_{\lambda^*_t}(s, \pi_t(\boldsymbol{s})) = R(s, \pi_t(\boldsymbol{s}))$, and the drift bound holds symmetrically:

$$
V^{\pi_t}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(s) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}(s) \le \frac{\gamma N(1+U_\lambda)}{(1-\gamma)^2} B_t. \tag{37}
$$

Bounding Term (ii): Define

$$
D^*_t := \| \left(\max_{\pi} J^{\pi}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}} - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}\right) \|_\infty \text{ and } D^{\text{alg}}_t := \| \left(J^{\pi_t}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}} - V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}\right) \|_\infty \tag{38}
$$

By the Bellman optimality defined in (28)-(29), and subtracting the two equations and using the property of the optimal policy, we obtain

$$
\begin{aligned}
&\max_{\pi} J^{\pi}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}}(\boldsymbol{s}) - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) \\
=&\gamma \sum_{\boldsymbol{s}'} \boldsymbol{P}_t(\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) \left(\max_{\pi} J^{\pi}_{\boldsymbol{P}_{t+1:T},\lambda^*_{t+1:T}}(\boldsymbol{s}') - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}')\right) && (39)\\
=&\gamma \sum_{\boldsymbol{s}'} \boldsymbol{P}_t(\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a}) \| \max_{\pi} J^{\pi}_{\boldsymbol{P}_{t+1:T},\lambda^*_{t+1:T}}(\boldsymbol{s}') - V^{\pi^*_{t+1}}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}') + V^{\pi^*_{t+1}}_{\boldsymbol{P}_{t+1},\lambda^*_{t+1}}(\boldsymbol{s}') - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}') \|_\infty && (40)
\end{aligned}
$$

Based on (36), we obtain:

$$\max_{\pi} J^{\pi}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}}(\boldsymbol{s}) - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) \le D^*_t \le \gamma D^*_{t+1} + \frac{\gamma^2 N(1+U_\lambda)}{(1-\gamma)^2} B_t. \tag{41}$$

Similarly, for the deployed execution policy π_t, subtracting the evaluation Bellman equations and applying (37) yields:

$$J^{\pi_t}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}}(\boldsymbol{s}) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}) \le D^{\text{alg}}_t \le \gamma D^{\text{alg}}_{t+1} + \frac{\gamma^2 N(1+U_\lambda)}{(1-\gamma)^2} B_t. \tag{42}$$

Setting $D_t := \max\{D^*_t, D^{\text{alg}}_t\}$, it satisfies the linear recurrence $D_t \le \gamma D_{t+1} + \frac{\gamma^2 N(1+U_\lambda)}{(1-\gamma)^2} B_t$. Unrolling this linear recurrence forward to the horizon T:

$$D_t \le \frac{\gamma^2 N(1+U_\lambda)}{(1-\gamma)^2} \sum_{\tau=t}^{T-1} \gamma^{\tau-t} B_\tau + \gamma^{T-t} D_T. \tag{43}$$

Summing D_t across all time steps $t \in [T]$ and interchanging the summation order:

$$\sum_{t=1}^{T} D_t \le \frac{\gamma^2 N(1+U_\lambda)}{(1-\gamma)^2} \sum_{t=1}^{T} \sum_{\tau=t}^{T-1} \gamma^{\tau-t} B_\tau + \sum_{t=1}^{T} \gamma^{T-t} D_T \tag{44}$$

$$\le \frac{\gamma^2 N(1+U_\lambda)}{(1-\gamma)^3} \sum_{\tau=1}^{T-1} B_\tau + \frac{D_T}{1-\gamma} \tag{45}$$

$$\le \frac{\gamma^2 N(1+U_\lambda)}{(1-\gamma)^3} B + \frac{2N(1+U_\lambda)}{(1-\gamma)^2}, \tag{46}$$

where (45) holds by the geometric series bound $\sum_{t=1}^{\tau} \gamma^{\tau-t} \le \frac{1}{1-\gamma}$, and (46) holds due to the global budget $\sum_{\tau=1}^{T-1} B_\tau \le B$ alongside the terminal boundary condition $D_T \le \frac{2N(1+U_\lambda)}{1-\gamma}$. Then, we have

$$\min_{\lambda_{1:T}} \text{Reg}_{\text{dyn},\lambda_{1:T}}(T) \le \text{Reg}_{\text{dyn},\lambda^*_{1:T}}(T) \tag{47}$$

$$= \sum_{t=1}^{T} \left(\max_{\pi} J^{\pi}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}}(\boldsymbol{s}_t) - V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) \right) + \sum_{t=1}^{T} \left(V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) \right)$$

$$+ \sum_{t=1}^{T} \left(V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) - J^{\pi_t}_{\boldsymbol{P}_{t:T},\lambda^*_{t:T}}(\boldsymbol{s}_t) \right) \tag{48}$$

$$\le \sum_{t=1}^{T} \left(V^{\pi^*_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda^*_t}(\boldsymbol{s}_t) \right) + \sum_{t=1}^{T} D^*_t + \sum_{t=1}^{T} D^{\text{alg}}_t \tag{49}$$

$$\le \text{Reg}_{\text{proxy}}(T) + 2 \sum_{t=1}^{T} D_t \tag{50}$$

$$\le \text{Reg}_{\text{proxy}}(T) + \frac{2\gamma^2 N(1+U_\lambda)}{(1-\gamma)^3} B + \frac{4N(1+U_\lambda)}{(1-\gamma)^2} \tag{51}$$

where (49) holds by (38), and (51) holds due to (46), which concludes the proof. □

C Proof of Theorem 5.3

Theorem 5.3. *For any learning algorithm, there exists a non-stationary RMAB instance with N arms and total variation B such that the expected regret satisfies:*

$$\mathbb{E}\left[\min_{\{\lambda_t\}_{t=1}^{T}} \text{Reg}_{\text{dyn},\{\lambda_t\}_{t=1}^{T}}(T) \right] = \Omega(N^{2/3} B^{1/3} T^{2/3}) \tag{52}$$

Proof. (i) Structural Decoupling: The primary challenge in establishing a lower bound for RMAB regret lies in the coupled budget constraint K, which forces a joint dependency among the N arms. To address this, we construct a decoupled RMAB instance where this global constraint is rendered non-binding. Specifically, we consider a environment where $K = N$. This reduction ensures that the optimal policy for the joint MDP is simply the concatenation of optimal policies for N separate sub-problems. Consequently, the total dynamic regret decomposes into the summation of N independent regrets, where each arm $n \in [N]$ evolves as a separate non-stationary MDP. (ii) Regret Lower Bound: For each decoupled arm, we apply the lower bound established by Mao et al. (2021), which prove that for a variation budget B and horizon T, the dynamic regret is at least $\Omega(B^{1/3}T^{2/3})$. By accumulating these parallel environments, the total RMAB regret is established as

$$\sum_{i=1}^{N} \Omega\left(\left(\frac{B}{N}\right)^{1/3} T^{2/3}\right) = N \cdot \Omega\left(N^{-1/3}B^{1/3}T^{2/3}\right) = \Omega(N^{2/3}B^{1/3}T^{2/3}). \tag{53}$$

□

D Counting Bad Event Occurrences

We begin by analyzing the sliding window technique to quantify the time steps where bad events occur. We define the bad event set Q_T as the set of time steps satisfying the following two conditions:

- There exist $s \in \mathcal{S}, a \in \mathcal{A}$, and $i \in [N]$ such that $P_{t,i}(\cdot|s,a) \notin H_{t,i}(s,a,\eta)$, indicating that the true transition dynamics fall outside the confidence set.
- For each $t' \in Q_T$, the gap between consecutive bad event occurrences is larger than the window size. *i.e.,* $, t - t' > W$.

Next, we construct an extended set $\widetilde{Q}_T$ that includes all elements of Q_T along with any time steps within a window W of any element in Q_T, defined as $\widetilde{Q}_T = Q_T \cup \{t \in [T] : \exists t' \in Q_T, t - t' \in [0, W]\}$. This extension constructs a pessimistic set connected to Q_T to capture all adverse events. With this setup, we state the following lemma:

Lemma D.1. *Define the event* $\xi_i = \left\{\hat{P}_{t,i}(s'|s,a) \in H_{t,i}(s,a,0), \forall s \in \mathcal{S}, a \in \mathcal{A}, t \in [T]\right\}$. *Conditioned on the intersection of these events* $\{\xi_i\}_{i=1}^N$*, we have any time step* t *where the true transition dynamics fall outside the confidence interval for at least one arm, e.g.* $\{t \mid \exists i \in [N], P_{t,i}(\cdot|s,a) \notin H_{t,i}(s,a,\eta)\} \subseteq \widetilde{Q}_T$. *Furthermore, the cardinality of this set is bounded by:* $|\widetilde{Q}_T| \leq WB/\eta$.

This lemma quantifies the occurrence of bad events periods where transition estimates are inaccurate by leveraging the global variation budget B. The detailed proof is provided in Appendix D.

Lemma D.1. *Conditioned on* $\{\xi_i\}_{i=1}^N$*, we have* $\forall t$ *such that* $\exists i \in [N], P_{t,i}(\cdot|s,a) \notin H_{t,i}(s,a,\eta)$*, is in* $\widetilde{Q}_T$ *and* $|\widetilde{Q}_T| \leq WB/\eta$.

Proof. For all $t \in Q_T$, we claim that exist (s,a,i) and $\exists t' \in [(t-W)^+, t]$, we have

$$\|P_{t,i}(\cdot|s,a) - P_{t',i}(\cdot|s,a)\|_1 \geq \eta. \tag{54}$$

Assume $\forall (s,a,i,t' \in [(t-W)^+, t]), \|P_{t,i}(\cdot|s,a) - P_{t',i}(\cdot|s,a)\|_1 < \eta$, and consider

- Case 1, if $N_{t,i}(s,a) = 0$, we have

$$\left\|P_{t,i}(\cdot|s,a) - \widetilde{P}_{t,i}(\cdot|s,a)\right\|_1 = 0 \leq \text{rad}_{t,i}(s,a) \leq \text{rad}_{t,i}(s,a) + \eta, \tag{55}$$

  which contradicts to the fact $\forall t \in Q_T, P_{t,i}(\cdot|s,a) \notin H_{t,i}(s,a,\eta)$.

- Case 2, if $N_{t,i}(s,a) > 0$, we have

$$\begin{aligned}
&\left\| P_{t,i}(\cdot|s,a) - \hat{P}_{t,i}(\cdot|s,a) \right\|_1 \\
&\leq \frac{\sum_{q=(t-1-W)^+}^{t-1} \left\| P_{t,i}(\cdot|s,a) - P_{q,i}(\cdot|s,a) \right\|_1 \cdot \mathbb{1}(s_{q,i} = s, a_{q,i} = a)}{N_{t,i}^+(s,a)} \qquad (56)\\
&\leq \eta, \qquad (57)
\end{aligned}$$

where (56) holds by triangle inequality, and (57) holds due to the assumption. Then, we have

$$\left\| P_{t,i}(\cdot|s,a) - \widetilde{P}_{t,i}(\cdot|s,a) \right\|_1 \leq \left\| \widetilde{P}_{t,i}(\cdot|s,a) - \hat{P}_{t,i}(\cdot|s,a) \right\|_1 + \left\| P_{t,i}(\cdot|s,a) - \hat{P}_{t,i}(\cdot|s,a) \right\|_1 \qquad (58)$$

$$\leq \text{rad}_{t,i}(s,a) + \eta, \qquad (59)$$

where (58) follows from triangle inequality, and (59) holds due to the condition on $\{\xi_i\}_{i=1}^N$ and (57). Then, we have (59) contradicts the fact $\forall t \in Q_T, P_{t,i}(\cdot|s,a) \notin H_{t,i}(s,a,\eta)$.

Finally, we have

$$\begin{aligned}
B &= \sum_{t=1}^{T} \max_{s,a,i} \left\| P_{t+1,i}(\cdot|s,a) - P_{t,i}(\cdot|s,a) \right\|_1 \qquad (60)\\
&\geq \sum_{t \in Q_T} \sum_{t' \in [(t-1-W)^+, (t-1)^+]} \max_{s,a,i} \left\| P_{t'+1,i}(\cdot|s,a) - P_{t',i}(\cdot|s,a) \right\|_1 \qquad (61)\\
&\geq \sum_{t \in Q_T} \max_{s,a,i,t' \in [(t-1-W)^+, (t-1)^+]} \left\| P_{t'+1,i}(\cdot|s,a) - P_{t',i}(\cdot|s,a) \right\|_1 \qquad (62)\\
&\geq |Q_T|\, \eta, \qquad (63)
\end{aligned}$$

where (60) follows from the budget constrain definition, while (61) holds because the time steps in Q_T are spaced at least W apart. (62) follows by replacing the summation with a maximum operation, and (63) holds by (54). We finished the proof. □

E Proof of Lemma 5.12

Assumption E.1. *For the non-stationary state transitions over time, we assume a uniform lower bound on all non-zero transition probabilities, defined as* $p_{min} := \min_{(s,a,s',t,i) \notin P_0} P_{t,i}(s'|s,a)$, *where* P_0 *is the set of all zero-probability transitions given by*

$$P_0 := \left\{ (s,a,s't,i) | P_{t,i}(s'|s,a) = 0, \forall s, s' \in \mathcal{S}, a \in \mathcal{A}, t \in [T] \text{ and } i \in [N] \right\}. \qquad (64)$$

This assumption ensures that non-zero probabilities in the true state transition have a minimum value, which is critical for controlling the magnitude of transition ratios that arise in importance sampling. Specifically, it only requires that the observed transition probabilities remain above a known constant, without demanding explicit knowledge of the constant $p_{\min}$ or the exact composition of the zero-transition P_0. This makes the assumption applicable, aligning with similar conditions commonly adopted in model-based reinforcement learning literature Kumar & Lin (1982).

Recall that

$$\text{Reg}'(T) := \sum_{t=1}^{T} \left(V^{\pi_t}_{\widetilde{\boldsymbol{P}}_t, \lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t, \lambda_t^*}(\boldsymbol{s}_t) \right). \qquad (65)$$

The following lemma shows the upper bound of $\text{Reg}'(T)$.

Lemma 5.12. *Condition on* $\{P_{t,i}(\cdot|s,a) \in H_{t,i}(s,a,\eta), \forall(s,a,i)\}_{t=1}^{T}$*, with probability at least* $1-6\delta'-N/\delta$*, we have*

$$\begin{aligned}\mathrm{Reg}'(T) \leq \frac{2N(1+U_\lambda)\gamma}{(1-\gamma)^2}\sqrt{T\log\frac{1}{\delta'}} + \frac{\gamma N(U_\lambda+1)}{p_{min}(1-\gamma)^2}\sqrt{T\log\frac{1}{\delta'}} \\ + \frac{\gamma N(U_\lambda+1)}{p_{min}(1-\gamma)^2}\left(TN\eta + N\cdot\sqrt{2|S|\log(|S||A|T/\delta)}\left(\frac{T}{\sqrt{W}}\cdot\left(\sqrt{2}+1\right)\sqrt{|\mathcal{S}|\,|\mathcal{A}|}\right)\right).\end{aligned} \tag{66}$$

Proof. By the definition of $\mathrm{Reg}'(T)$, we have

$$\mathrm{Reg}'(T) = \sum_{t=1}^{T}\left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) \tag{67}$$

$$= \gamma\sum_{t=1}^{T}\left[\mathop{\mathbb{E}}_{s'\sim\bar{\boldsymbol{P}}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}')\right] - \mathop{\mathbb{E}}_{s'\sim\boldsymbol{P}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}')\right]\right] \tag{68}$$

$$\begin{aligned}&= \gamma\sum_{t=1}^{T}\Bigg[\underbrace{\mathop{\mathbb{E}}_{s'\sim\boldsymbol{P}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}') - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}')\right] - \left(V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1}) - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_{t+1})\right)}_{:=B_1} \\ &\quad\underbrace{-\mathop{\mathbb{E}}_{s'\sim\boldsymbol{P}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}')\right] + V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1})}_{:=B_2} \\ &\quad\underbrace{+\mathop{\mathbb{E}}_{s'\sim\bar{\boldsymbol{P}}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}')\right] - V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1})}_{:=B_3}\Bigg]\end{aligned} \tag{69}$$

$$\leq \frac{2N(1+U_\lambda)\gamma}{1-\gamma}\sqrt{T\log\frac{1}{\delta'}} + B_3, \tag{70}$$

where (68) holds due to the immediate rewards for both value function, $V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t^*}(\boldsymbol{s}_t)$ and $V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)$, are the same under the identical policy π_t, and (70) holds with probability at least $1-2\delta'$ by applying Azuma-Hoeffding inequality, presented in Lemma A.1, on the martingale difference sequences, B_1 and B_2, with upper bound $M = (N+U_\lambda N)/(1-\gamma)$. For B_3, we have

$$B_3 = \gamma\sum_{t=1}^{T-1}\left(\mathop{\mathbb{E}}_{s'\sim\bar{\boldsymbol{P}}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}')\right] - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_{t+1})\right) \tag{71}$$

$$= \gamma\sum_{t=1}^{T-1}\left(\mathop{\mathbb{E}}_{s'\sim\boldsymbol{P}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[\frac{\bar{\boldsymbol{P}}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}{\boldsymbol{P}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}')\right] - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_{t+1})\right) \tag{72}$$

$$\begin{aligned}&= \gamma\sum_{t=1}^{T-1}\Bigg(\underbrace{\mathop{\mathbb{E}}_{s'\sim\boldsymbol{P}_t(\cdot|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}\left[\frac{\bar{\boldsymbol{P}}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}{\boldsymbol{P}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}')\right] - \frac{\bar{\boldsymbol{P}}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}{\boldsymbol{P}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1})}_{B_4} \\ &\quad + \underbrace{\frac{\bar{\boldsymbol{P}}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}{\boldsymbol{P}_t(\boldsymbol{s}_{t+1}\,|\,\boldsymbol{s}_t,\boldsymbol{a}_t)}V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1}) - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_{t+1})}_{B_5}\Bigg).\end{aligned} \tag{73}$$

For B_4, the same as B_1 and B_2, by Azuma-Hoeffding inequality (Lemma A.1) with $M = p_{\min}^{-1}\cdot N(1+U_\lambda)/(1-\gamma)$, and probability at least $1-\delta'$, we have

$$B_4 \leq \frac{\gamma N(U_\lambda+1)}{p_{\min}(1-\gamma)}\sqrt{T\log\frac{1}{\delta'}}. \tag{74}$$

For B_5, we have

$$B_5 = \gamma \sum_{t=1}^{T-1} \left(\frac{\bar{\boldsymbol{P}}_t(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)}{\boldsymbol{P}_t(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)} V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1}) - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_{t+1}) \right) \tag{75}$$

$$= \gamma \sum_{t=1}^{T-1} \left(\frac{\bar{\boldsymbol{P}}_t(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)}{\boldsymbol{P}_t(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)} - 1 \right) V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1}) + \gamma \sum_{t=1}^{T-1} \left(V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1}) - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_{t+1}) \right) \tag{76}$$

$$\leq \frac{\gamma N(U_\lambda + 1)}{p_{\min}(1-\gamma)} \underbrace{\sum_{t=1}^{T-1} \left(\bar{\boldsymbol{P}}_t(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) - \boldsymbol{P}_t(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) \right)}_{B_6} + \underbrace{\gamma \sum_{t=1}^{T-1} \left(V^{\pi}_{\bar{\boldsymbol{P}}_t,\lambda}(\boldsymbol{s}_{t+1}) - V^{\pi}_{\boldsymbol{P}_t,\lambda}(\boldsymbol{s}_{t+1}) \right)}_{B_7}, \tag{77}$$

where (77) holds by Assumption E.1 and Lemma 5.11. For B_6, by the property that $\{P_{t,i} \in H_{t,i}\}_{i=1}^N$ for all $t \notin \widetilde{Q}_T$, we further have

$$B_6 \leq \sum_{t=1}^{T} \sum_{s,a} \sum_{i=1}^{N} \left(\text{rad}_{t,i}(s,a) + \eta \right) \mathbb{1}(s_{t,i} = s, a_{t,i} = a) \tag{78}$$

$$\leq TN\eta + \sqrt{2|S| \log(|S||A|T/\delta)} \cdot \sum_{i=1}^{N} \sum_{s,a} \sum_{t=1}^{T} \frac{\mathbb{1}(s_{t,i} = s, a_{t,i} = a)}{\sqrt{N_{t,i}(s,a)}} \tag{79}$$

Then, by partitioning the total time T into $M_i(T)$ episodes, each of width W_i for arm i, we denote the starting time step of the m-th episode as $\tau_i(m)$ for arm i, we have

$$\sum_{s,a} \sum_{t=1}^{T} \frac{\mathbb{1}(s_{t,i} = s, a_{t,i} = a)}{\sqrt{N_{t,i}(s,a)}} = \sum_{s,a} \sum_{m \in M_i(T)} \sum_{t=\tau_i(m)}^{\tau_i(m+1)-1} \frac{\mathbb{1}(s_{t,i} = s, a_{t,i} = a)}{\sqrt{N_{t,i}(s,a)}} \tag{80}$$

$$\leq \sum_{s,a} \sum_{m \in M_i(T)} \sum_{t=\tau_i(m)}^{\tau_i(m+1)-1} \frac{\mathbb{1}(s_{t,i} = s, a_{t,i} = a)}{\sqrt{\hat{N}_{t,i}(s,a)}} \tag{81}$$

$$\leq \sum_{s,a} \sum_{m \in M_i(T)} \left(\sqrt{2} + 1 \right) \sqrt{\hat{N}_{\tau_i(m+1)-1,i}(s,a)} \tag{82}$$

$$\leq \frac{T}{W_i} \cdot \left(\sqrt{2} + 1 \right) \sqrt{|\mathcal{S}|\,|\mathcal{A}| \sum_{s,a} \hat{N}_{\tau_i(m+1)-1,i}(s,a)} \tag{83}$$

$$= \frac{T}{\sqrt{W_i}} \cdot \left(\sqrt{2} + 1 \right) \sqrt{|\mathcal{S}|\,|\mathcal{A}|} \tag{84}$$

where (81) holds by defining $\hat{N}_{t,i}(s,a) := \sum_{q=\max\{(t-1-W_i),\tau_i(m)\}}^{t-1} \mathbb{1}(s_{q,i} = s, a_{q,i} = a)$, if $t \in m$-th episode, (82) holds by Lemma 19 in Auer et al. (2008), and (83) holds by Jenson's inequality. Substituting (84) into (79), and denoting $W := \min_{i \in [N]} W_i$, we have

$$B_6 \leq TN\eta + N \cdot \sqrt{2|S| \log(|S||A|T/\delta)} \left(\frac{T}{\sqrt{W}} \cdot \left(\sqrt{2} + 1 \right) \right) \sqrt{|\mathcal{S}|\,|\mathcal{A}|} \tag{85}$$

To simplify the notation, we aggregate the error terms from equations (70), (73), (74), (77), and (85). With probability at least $1 - 3\delta$, we denote the sum of the error terms as a constant C, defined as:

$$C := \frac{2N(1+U_\lambda)\gamma}{1-\gamma} \sqrt{T \log \frac{1}{\delta'}} + \frac{\gamma N(U_\lambda + 1)}{p_{\min}(1-\gamma)} \sqrt{T \log \frac{1}{\delta'}} + \frac{\gamma N(U_\lambda + 1)}{p_{\min}(1-\gamma)} \left(TN\eta + N \cdot \sqrt{2|S| \log(|S||A|T/\delta)} \left(\frac{T}{\sqrt{W}} \cdot \left(\sqrt{2} + 1 \right) \sqrt{|\mathcal{S}|\,|\mathcal{A}|} \right) \right). \tag{86}$$

Using this definition, with probability at least $1-3\delta$, the regret can be expressed as:

$$\begin{aligned}
&\mathrm{Reg}'(s_t, \bar{P}_t, P_t, \pi_t, \lambda_t^*) \\
&= \sum_{t=1}^{T} \left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t, \lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t, \lambda_t^*}(\boldsymbol{s}_t) \right) && (87) \\
&\leq C + \gamma \sum_{t=1}^{T-1} \left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t, \lambda_t^*}(\boldsymbol{s}_{t+1}) - V^{\pi_t}_{\boldsymbol{P}_t, \lambda_t^*}(\boldsymbol{s}_{t+1}) \right) && (88) \\
&= \frac{2N(1+U_\lambda)\gamma}{(1-\gamma)^2} \sqrt{T \log \frac{1}{\delta'}} + \frac{\gamma N(U_\lambda+1)}{p_{\min}(1-\gamma)^2} \sqrt{T \log \frac{1}{\delta'}} \\
&\quad + \frac{\gamma N(U_\lambda+1)}{p_{\min}(1-\gamma)^2} \left(TN\eta + N \cdot \sqrt{2|S| \log(|S||A|T/\delta)} \left(\frac{T}{\sqrt{W}} \cdot \left(\sqrt{2}+1\right) \sqrt{|\mathcal{S}|\,|\mathcal{A}|} \right) \right), && (89)
\end{aligned}$$

where (88) holds by (77) and (89) follows by repeatedly applying the similar process (68)-(86) to the term $\gamma \sum_{t=1}^{T-1} \left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t, \lambda_t^*}(\boldsymbol{s}_{t+1}) - V^{\pi_t}_{\boldsymbol{P}_t, \lambda_t^*}(\boldsymbol{s}_{t+1}) \right)$ for T iterations, and using the summation formula for the geometric series. We have finished the proof.

□

F Proof of Theorem 5.4

Theorem 5.4. *With probability at least $1-6\delta'-N/\delta$, we have*

$$\begin{aligned}
&\min_{\lambda_{1:T}} \mathit{Reg}_{\mathit{dyn}, \lambda_{1:T}}(T) \\
&\leq \frac{2\gamma^2 N(1+U_\lambda)}{(1-\gamma)^3} B + \frac{4N(1+U_\lambda)}{(1-\gamma)^2} \\
&\quad + \frac{2NBW(N+U_\lambda K)}{\eta(1-\gamma)} + \frac{2N(1+U_\lambda)\gamma}{(1-\gamma)^2} \sqrt{T \log \frac{1}{\delta'}} + \frac{\gamma N(U_\lambda+1)}{p_{\mathit{min}}(1-\gamma)^2} \sqrt{T \log \frac{1}{\delta'}} \\
&\quad + \frac{\gamma N(U_\lambda+1)}{p_{\mathit{min}}(1-\gamma)^2} \left(TN\eta + N \cdot \sqrt{2|S| \log(|S||A|T/\delta)} \left(\frac{T}{\sqrt{W}} \cdot \left(\sqrt{2}+1\right) \sqrt{|\mathcal{S}|\,|\mathcal{A}|} \right) \right) && (90)
\end{aligned}$$

Proof. By the definition of regret in (3) and apply Lemma 5.1, we have

$$\min_{\lambda_{1:T}} \mathrm{Reg}_{\mathrm{dyn}, \lambda_{1:T}}(T) \leq \mathrm{Reg}_{\mathrm{proxy}}(T) + \frac{2\gamma^2 N(1+U_\lambda)}{(1-\gamma)^3} B + \frac{4N(1+U_\lambda)}{(1-\gamma)^2} \tag{91}$$

Then, considering the following definitions

$$A_1 := \frac{2NBW(N+U_\lambda K)}{\eta(1-\gamma)} \tag{92}$$

$$\begin{aligned}
A_2 := &\frac{2N(1+U_\lambda)\gamma}{(1-\gamma)^2} \sqrt{T \log \frac{1}{\delta'}} + \frac{\gamma N(U_\lambda+1)}{p_{\min}(1-\gamma)^2} \sqrt{T \log \frac{1}{\delta'}} \\
&+ \frac{\gamma N(U_\lambda+1)}{p_{\min}(1-\gamma)^2} \left(TN\eta + N \cdot \sqrt{2|S| \log(|S||A|T/\delta)} \left(\frac{T}{\sqrt{W}} \cdot \left(\sqrt{2}+1\right) \sqrt{|\mathcal{S}|\,|\mathcal{A}|} \right) \right)
\end{aligned} \tag{93}$$

we have

$$\begin{aligned}
&\Pr\left(\mathrm{Reg}_{\text{proxy}}(T) > A_1 + A_2\right) \\
&= \underbrace{\Pr\left(\sum_{t=1}^{T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) \geq A_1 + A_2 \mid \{\xi_i\}_{i=1}^N\right)}_{I_1} \cdot \Pr\left(\{\xi_i\}_{i=1}^N\right) \\
&\quad + \underbrace{\Pr\left(\sum_{t=1}^{T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_1 + A_2 \mid \text{not } \{\xi_i\}_{i=1}^N\right)}_{I_2} \cdot \Pr\left(\text{not } \{\xi_i\}_{i=1}^N\right) \qquad (94) \\
&\leq \underbrace{\Pr\left(\sum_{t\in\widetilde{Q}_T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_1 \mid \{\xi_i\}_{i=1}^N\right)}_{I_3} \cdot \Pr\left(\{\xi_i\}_{i=1}^N\right) \\
&\quad + \Pr\left(\sum_{t\notin\widetilde{Q}_T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_2 \mid \{\xi_i\}_{i=1}^N\right) \cdot \Pr\left(\{\xi_i\}_{i=1}^N\right) + \Pr\left(\text{not } \{\xi_i\}_{i=1}^N\right), \qquad (95) \\
&\leq \Pr\left(\sum_{t\notin\widetilde{Q}_T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_2 \mid \{\xi_i\}_{i=1}^N\right) \cdot \Pr\left(\{\xi_i\}_{i=1}^N\right) + \Pr\left(\text{not } \{\xi_i\}_{i=1}^N\right), \qquad (96)
\end{aligned}$$

where (95) follows from the union bound for I_1 and the fact that $I_2 \leq 1$, while (96) holds because $I_3 = 0$ due to Lemma D.1, Lemma 5.11, and the fact that the size of $\widetilde{Q}_T$ is W times $|Q_T|$. Then, we proceed with the RHS of (96). Given that $\Pr\left(\{\xi_i\}_{i=1}^N\right) \leq 1$ and $I_4 \geq 0$, and define $\mathcal{B}_t := \{P_{t,i}(\cdot|s,a) \in H_{t,i}(s,a,\eta), \forall(s,a,i)\}$, we have

$$\begin{aligned}
&\Pr\left(\sum_{t\notin\widetilde{Q}_T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_2 \mid \{\xi_i\}_{i=1}^N\right) \cdot \Pr\left(\{\xi_i\}_{i=1}^N\right) + \Pr\left(\text{not } \{\xi_i\}_{i=1}^N\right) \\
&\leq \Pr\left(\sum_{t\notin\widetilde{Q}_T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_2 \mid \{\xi_i\}_{i=1}^N\right) + \Pr\left(\text{not } \{\xi_i\}_{i=1}^N\right) \\
&\quad + \underbrace{\Pr\left(\sum_{t\in\widetilde{Q}_T}\left(V^{\pi_t^*}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_2 \mid \{\xi_i\}_{i=1}^N \cap \mathcal{B}_t\right)}_{I_4} \qquad (97) \\
&\leq \Pr\left(\sum_{t\notin\widetilde{Q}_T}\left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_2 \mid \{\xi_i\}_{i=1}^N \cap \mathcal{B}_t\right) + \Pr\left(\text{not } \{\xi_i\}_{i=1}^N\right) \\
&\quad + \Pr\left(\sum_{t\in\widetilde{Q}_T}\left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right) > A_2 \mid \{\xi_i\}_{i=1}^N \cap \mathcal{B}_t\right) \qquad (98) \\
&\leq 2 \cdot \Pr\left(\mathrm{Reg}'(1,T) > A_2 \mid \{\xi_i\}_{i=1}^N \cap \mathcal{B}_t\right) + \Pr\left(\text{not } \{\xi_i\}_{i=1}^N\right) \qquad (99) \\
&\leq 6\delta' + N/\delta, \qquad (100)
\end{aligned}$$

where (98) since, for all $t \notin \widetilde{Q}_T$, the event $\mathcal{B}_t$ follows from Lemma A.2, (99) holds by defining $\mathrm{Reg}'(1,T) := \sum_{t=1}^{T}\left(V^{\pi_t}_{\bar{\boldsymbol{P}}_t,\lambda_t^*}(\boldsymbol{s}_t) - V^{\pi_t}_{\boldsymbol{P}_t,\lambda_t^*}(\boldsymbol{s}_t)\right)$, and (100) follows from Lemma A.3 and Lemma 5.12.

□

G PROOF OF LEMMA 5.10

Lemma 5.10. *Suppose there exists a strictly feasible policy $\bar{\pi}$ such that $\|\bar{\pi}(s')\|_1 < K$ for all $s' \in \mathcal{S}$. Then, for all $t \in [T]$, the optimal dual variable λ_t^* is uniformly bounded by a constant $U_\lambda \geq 0$, such that $\lambda_t^* \leq U_\lambda$.*

Proof. Consider the constrained optimization problem for the joint MDP:

$$\max_{\pi} V^{\pi}(s) \quad \text{s.t.} \quad \|\pi(s')\|_1 \leq K, \forall s' \in \mathcal{S}, \tag{101}$$

the optimal solution is given by $\pi^* = \operatorname{argmax}_\pi \min_{\lambda \geq 0} L(\pi, \lambda, s)$ by Assumption 5.9. Let $\lambda^* := \max_\pi \operatorname{argmin}_{\lambda \geq 0} L(\pi, \lambda, s)$ be the optimal dual variable. For any policy, including the strictly feasible policy $\bar{\pi}$, the optimal value function $V^*(s)$ satisfies:

$$V^*(s) = \max_{\pi} L(\pi, \lambda^*, s) \geq L(\bar{\pi}, \lambda^*, s) = V^{\bar{\pi}}(s) - \sum_{s' \in \mathcal{S}} \lambda^*(s') \cdot (\|\bar{\pi}(s')\|_1 - K). \tag{102}$$

Rearranging the terms in (102) yields:

$$\sum_{s' \in \mathcal{S}} \lambda^*(s') \cdot (K - \|\bar{\pi}(s')\|_1) \leq V^*(s) - V^{\bar{\pi}}(s). \tag{103}$$

Let $\epsilon = \min_{s' \in \mathcal{S}} (K - \|\bar{\pi}(s')\|_1) > 0$ represent the feasibility margin. If we select $\bar{\pi}$ to be the "no-pull" policy (i.e., $\|\bar{\pi}(s')\|_1 = 0$ for all s'), we have $\epsilon = K$. It follows that:

$$\sum_{s' \in \mathcal{S}} \lambda^*(s') \leq \frac{1}{K(1-\gamma)} := U_\lambda. \tag{104}$$

Since the rewards are bounded in $[0, 1]$ and $\gamma < 1$, the difference $V^*(s) - V^{\bar{\pi}}(s)$ is naturally bounded by the maximum possible value function magnitude. Thus, there exists a constant U_λ such that $\|\lambda^*\|_1 \leq U_\lambda$ for all $t \in [T]$. □

H EXPERIMENTAL CONFIGURATION

Server specifications. All experiments are conducted on a server equipped with dual Intel Xeon Gold 6226 CPUs (2.7 GHz, 24 cores per node) and 192 GB DDR4-2933 MHz memory.

Environment At the beginning of the experiment, each arm is assigned a deterministic, stationary, and known reward. To satisfy the indexability assumption, we employ rejection sampling to ensure that for every arm, the action of pulling it yields a higher expected value than not pulling it. The environment operates under a constraint where at most $K = 2$ arms can be activated at each time step. To evaluate a policy π, we define the surrogate simple regret as the reward gap relative to the optimal policy at time t:$\boldsymbol{R}(\boldsymbol{s}_t^{\pi^*}, \boldsymbol{a}_t^{\pi^*}) - \boldsymbol{R}(\boldsymbol{s}_t^{\pi}, \boldsymbol{a}_t^{\pi})$. In this context, the optimal policy is the Whittle index policy computed using the true transition probabilities. Here, $\boldsymbol{s}_t^{\pi^*}$ and $\boldsymbol{a}_t^{\pi^*}$ represent the state and action trajectories under the Whittle index policy at time t, while $\boldsymbol{s}_t^{\pi}$ and $\boldsymbol{a}_t^{\pi}$ denote the corresponding state and action observed under policy π.

The relaxed regret proxy is both theoretically grounded and experimentally flexible for the RMAB setting. Inspired by recent dynamic regret definitions in non-stationary MDPs Cheung et al. (2020), we adapted the regret definition by using stationary value functions to quantify suboptimality in the discounted setting. This ensures our performance metric remains consistent with established frameworks for infinite-horizon MDPs. Regarding our experiments, there is a known equivalence in stationary discounted RL between suboptimality and the attained reward difference Zhou et al. (2021); Liu & Su (2020). To ensure complete clarity, the final version will incorporate more discussion of the analyzed regret proxy.

Environment dynamics. We maintain logits that are mapped to probabilities via a softmax function, with Gaussian noise added periodically to introduce non-stationarity. After generating transitions for

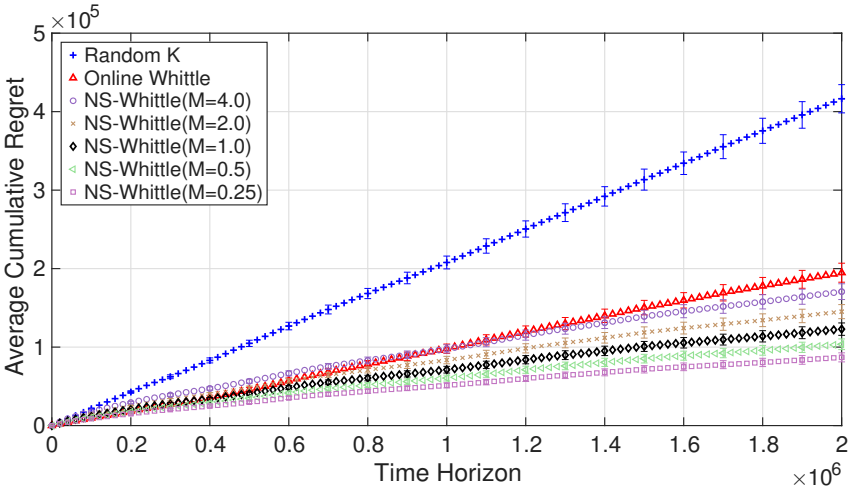


Figure 3: The empirical results for NS-Whittle when using $M \times B_i$ as the budget parameter for each arm i to determine both the window size and η_i.

all time steps, the total budget is computed and made available to NS-Whittle. The environment design has two key properties: (i) To satisfy the indexability assumption, we conduct rejection sampling to ensure that pulling an arm yields a higher expected reward than not pulling it. (ii) To maintain tractability, the frequency of transition changes decays exponentially. Uniform non-stationarity render the problem intractable and lead to linear regret, since only a limited time window of data would remain useful.

At each time step, the environment has a probability ρ that the transition dynamics will change. The value of ρ is halved whenever $\rho = 500, 1000, \ldots$. When a transition change occurs, a random noise sample is drawn from $\mathcal{N}(0, 0.1)$ and added to the underlying transition logic. The resulting vector is then normalized and passed through a softmax function to obtain valid transition probabilities.

To satisfy the indexability constraint, we perform rejection sampling to ensure that, for every arm, the action of pulling it yields a higher expected value than not pulling it.

Experiment setup. At each time step, there are $N = 10$ arms, and at most $K = 2$ arms can be activated. The total reward is the sum of the rewards obtained from the selected arms. The regret is defined as the performance difference between the Whittle index policy and the comparison policy.

Hyperparameters. The following hyperparameters are used for both the Online Whittle and NS-Whittle algorithms:

- $U = 10$: Number of iterations for the EVI
- $\lambda_0 = 0$: Initial Lagrange multiplier
- $M = 10$: Maximum number of iterations for Whittle index computation
- $\epsilon = 10^{-6}$: Convergence tolerance for EVI
- $\gamma = 0.99$: Discount factor used in EVI

For NS-Whittle, we employ the doubling trick Besson & Kaufmann (2018) to dynamically update the total horizon T, which is used to compute both the window size W_i and the exploration parameter η_i for each arm i. The doubling schedule is applied at time steps $t = 10{,}000, 20{,}000, 40{,}000, \ldots$, and is independent of the decay schedule governing the non-stationarity of the environment (e.g., $\rho = 500, 1000, \ldots$). For practical implementation, when no relevant data exist within the corresponding window size, we use the entire historical data to estimate $N^{+}_{t,i}(s, a)$.

For both NS-Whittle and Online Whittle, if no data are available for a specific transition (s, a), we set the estimated transition probability as a uniform distribution, i.e., $\forall s' \in \mathcal{S}$ s.t $\widetilde{P}_{t,i}(s'|s, a) = \frac{1}{|\mathcal{S}|}$ for arm i at time t.

I Related Work

Offline restless multi-armed bandits (RMABs) Restless multi-armed bandits (RMABs) Whittle (1988) generalize multi-armed bandits (MABs) Slivkins et al. (2019); Kuleshov & Precup (2014) by allowing multiple (K) arms to be pulled at the same time, where arms are modeled as independent MDPs. With the MDP parameters given, optimizing the RMAB as an offline optimization problem is known to be PSPACE-hard Papadimitriou & Tsitsiklis (1999). As an approximate solution, the Whittle index policy Weber & Weiss (1990) is shown to be asymptotically optimal under the indexability condition Liu & Zhao (2008; 2010). Many Whittle index-based approaches Robledo et al. (2022); Biswas et al. (2021); Fu et al. (2019) were proposed to solve RMABs in polynomial time.

Online learning in MDPs When the MDP parameters are unknown a priori, the goal is to interact with the environment to learn the MDP parameters and simultaneously maximize the cumulative rewards. In stationary infinite-horizon MDPs, algorithms Auer et al. (2008); Auer & Ortner (2006); Bartlett & Tewari (2012) with $\tilde{O}(S\sqrt{AT})$ upper bound on regret under different conditions and a lower bound $\Omega(\sqrt{SAT})$ were shown, where S and A denotes the state and action space size. In non-stationary infinite-horizon MDPs with limited changes in MDP parameters, Cheung et al. (2020); Wei & Luo (2021) propose the *SWUCRL2-CW* algorithm with $O(T^{\frac{3}{4}})$ regret upper bound and no lower bound is known. When the number of changes is known, Gajane et al. (2018) shows a $O(T^{\frac{2}{3}})$ regret bound. For episodic non-stationary MDPs, Mao et al. (2021) shows a $\tilde{O}(T^{\frac{2}{3}})$ regret with a lower bound $\Omega(T^{\frac{2}{3}})$. A detailed summary and a master approach that handle non-stationary MDPs can be found in Wei & Luo (2021).

Online learning in RMABs For RMABs, due to the exponentially large state and action space in RMABs, directly applying the results from MDPs can incur exponentially large regret bound. For stationary RMABs, when the rewards are collected only by the arms pulled, Liu et al. (2012) demonstrate a $O(\log T)$ regret bound. When rewards are collected from all arms, the problems become significantly hard where Jung & Tewari (2019) use a Thompson sampling approach to achieve a $\tilde{O}(\sqrt{T})$ Bayesian regret bound, while Wang et al. (2023) use UCB to show a $\tilde{O}(N|S|\sqrt{T\log T})$ frequentist regret bound. For non-stationary RMABs, there has been no prior work on the regret analysis. Our work provides the first regret bound that scales polynomially in the number of arms in non-stationary RMABs.

Non-stationary RMABs Shisher et al. (2025) achieves a regret of $\mathcal{O}(T^{2/3}\tilde{V}^{1/3}+T^{4/5})$ in an episodic RMAB setting. However, our work addresses the significantly more challenging infinite-horizon setting. The key difference lies in the dependency on state distributions. In episodic settings, agents can rely on periodic resets or mixing-time assumptions to ensure the state distribution converges to a stationary distribution, thereby obtaining nearly i.i.d. samples for learning. In contrast, our infinite-horizon MDP setting prohibits resets, and the time-varying nature of the learning policy prevents the Markov chain from mixing effectively. Critically, the mixing time assumption is fundamentally difficult to satisfy in non-stationary environments, as the underlying transition dynamics shift before the state distribution can reach equilibrium. Without resets, stationary policies, or mixing guarantees, an agent risks becoming trapped in suboptimal states with no mechanism for recovery. To overcome this, we theoretically bound the regret by applying martingale concentration inequalities directly to the value function differences.

J Pseudo Code of NS-Whittle

Algorithm 1 NS-Whittle

Input: Horizon T, activation budget K, variation budget B.
Initialize: $\lambda_0 \leftarrow 0$, window size $W_i = |S|T^{1/2}(\sum_{t=1}^{T} B_{t,i})^{-\frac{1}{2}}$.
for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, N$ **do**
    Construct confidence interval $H_{t,i}$ using data from $[(t - W_i)^+, t - 1]$.
    Run EVI(10) with λ_{t-1} for U iterations to obtain optimal action value function $\bar{Q}^{(U)}_{\lambda_{t-1},i}$.
    Compute each arm's Whittle index $\mathcal{W}_i(s_{t,i})$ through (12).
  end for
  Select the K arms with the highest Whittle indices to form the action vector $\boldsymbol{a}_t$.
  Set the Lagrange multiplier λ_t to the K-th largest Whittle index.
  Execute $\boldsymbol{a}_t$, observe $\boldsymbol{s}_{t+1}$, update history.
end for