Bandit Learning on Dynamic Graphs

Sourav Chakraborty[†*] **Amit Kiran Rege**[†*] **Claire Monteleoni**[†‡] **Lijun Chen**[†]
[†]Department of Computer Science, University of Colorado Boulder, USA
[‡]INRIA Paris, France
{sourav.chakraborty, amit.rege, cmontel, lijun.chen}@colorado.edu

Abstract

We study an online learning setting where an agent's actions are constrained to local movements on a dynamic graph, a setting that captures scenarios such as autonomous reconnaissance. This problem highlights a core challenge in adaptive systems: how to learn effectively with only partial, localized feedback in a non-stationary environment. We propose a set of structural conditions, termed *Recurrent Reachability* and *Temporal Stability*, that are sufficient for learnability. Our analysis reveals a foundational *anatomy of regret*, decomposing it into a statistical learning cost and a physical navigation cost. We introduce a family of local algorithms, progressing from a canonical protocol to a more practical, adaptive variant, and culminating in a reward-aware exploration policy that achieves provably near-optimal regret on any graph sequence satisfying our conditions. We corroborate our theory in a disaster-response simulation.

1 Introduction

Consider the critical task of deploying a mobile communication relay in a region recently affected by a natural disaster. A robotic ground vehicle is tasked with identifying the optimal location for this relay by navigating streets where access may be temporarily restricted by debris or unstable conditions. This navigation takes place in a highly uncertain, fluid environment where available pathways flicker on and off unpredictably. At every site, the vehicle must assess the local signal quality it can provide, exploring only locally accessible options to build up a reliable model of the environment's rewards. The agent's final decision is to commit to the single location that promises the most reliable service, a crucial step in supporting ongoing recovery efforts.

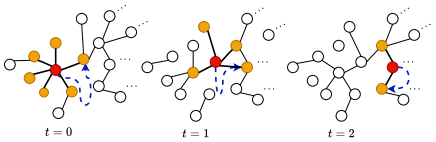

Figure 1: An agent's local decision process over a dynamic graph. The red node is the agent's position; orange nodes are available neighbors. The agent chooses one neighbor to move to (blue arrow).

This scenario is a challenging instance of sequential decision-making under uncertainty, a problem formally studied in the multi-armed bandit (MAB) framework [19, 14, 20]. The MAB model formalizes the challenge of balancing exploration (sampling actions to learn their rewards) with exploitation (choosing actions believed to be optimal to maximize cumulative gain). This model has been successfully applied across a wide range of applications, including clinical trials [12], wireless communication [11], recommendation systems [16, 3], financial optimization [4], and economically motivated incentivization methods [10, 21, 5]. While a rich family of MAB algorithms provides optimal strategies, they typically rely on a critical assumption: the agent can select any action, or "arm," at any time. In this work, we address a more realistic setting where the agent's actions are restricted by a dynamic structure, allowing only options adjacent to its current position.

[*]Equal contribution.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

Prior work [22] has studied *static* graphs with fixed topology, where the agent often has access to global structure for long-term planning. In contrast, real-world environments frequently change over time (see Figure 1), rendering such assumptions invalid and making global planning brittle or infeasible. On the other hand, if the graph is allowed to change arbitrarily at every time step, no learning can happen. Therefore, we ask the fundamental question:

> *What structural and temporal properties make a dynamic environment learnable for an agent constrained to local observations and movements?*

Our work demonstrates that near-optimal learning can be guaranteed not by assuming a particular stochastic model of graph evolution, but by requiring that the graph sequence satisfy certain structural and temporal conditions. We introduce two such conditions: (a) *Recurrent Reachability*, which ensures that all actions remain accessible over time via the agent's local movement; and (b) *Temporal Stability*, which ensures that the environment evolves with sufficient regularity to permit statistical learning. In the following sections, we prove that these conditions are sufficient for near-optimal regret guarantees.

Our contributions are as follows. We first formalize these sufficient conditions for learnability. We show that not satisfying these conditions leads to linear regret. Building on this foundation, we introduce a family of minimalist algorithms that operate with only local information. Our analyses proves these algorithms achieve near-optimal regret on any graph sequence satisfying the proposed conditions. This analysis reveals a fundamental *anatomy of regret*, decomposing it into a statistical learning cost and a structural navigation cost caused by the agent not having access to the current best arm choice. Furthermore, we provide the first near-optimal regret bounds for several canonical dynamic graph settings, including i.i.d. and Markovian variants, and show that the results from these direct, model-specific analyses are consistent with the bounds derived from our general conditions. We conclude by corroborating our theoretical results in a practically motivated disaster-response simulation.

2 Model and Problem Formulation

Before presenting our algorithms and main results, we develop the theoretical foundation of the paper. This section formalizes the problem of bandit learning on dynamic graphs and introduces structural assumptions that are sufficient to guarantee learnability. We prove that any graph sequence satisfying these conditions admits near-optimal learning by a locally constrained agent.

2.1 Bandit Learning on Dynamic Graphs

We consider a setting with a finite set of n arms, denoted by the nodes $A = \{1, \dots, n\}$ of a graph. Each arm $a \in A$ is associated with a fixed, unknown reward distribution $\mathcal{D}(a)$ with mean $\mu(a) \in [0, 1]$. We assume, without the loss of generality, a unique optimal arm $a^* \triangleq \arg\max_{a \in A} \mu(a)$ and define the suboptimality gap for any other arm as $\Delta(a) \triangleq \mu(a^*) - \mu(a) > 0$.

At the start of each round $t \in \{1, \dots, T\}$, the environment is represented by an unweighted and undirected graph $G_t = (A, E_t)$, where E_t is the set of active edges at time t. The agent is positioned at node $a_{t-1} \in A$, and the agent's knowledge is limited to its immediate surroundings within this new graph. We refer to any algorithm that operates under these constraints as a strategy based on local information and local actions. We formally define the neighborhood of any node $a \in A$ at time t as the set of its adjacent nodes in G_t as $N_t(a) := \{b \in A \mid (a, b) \in E_t\}$. The agent's set of available actions is therefore $N_t(a_{t-1}) \cup \{a_{t-1}\}$. It must choose its next action, a_t, from this set. Upon selecting a_t, the agent receives a reward drawn from $\mathcal{D}(a_t)$.

The goal of a successful learning algorithm is to achieve sublinear regret, ensuring that its average regret vanishes as the horizon grows: $\lim_{T \to \infty} \mathbb{E}[R(T)]/T = 0$, where the cumulative regret is:

$$\mathbb{E}[R(T)] := \mathbb{E}\left[\sum_{t=1}^{T} (\mu(a^*) - \mu(a_t))\right]. \tag{1}$$

2.2 Sufficient Conditions for Learnability

The challenge lies in the unpredictable nature of the sequence $\{G_t\}$. For learning to be possible, the graph sequence must exhibit some form of recurrent connectivity to allow the agent to explore

the entire state space. A naive assumption might be to require that the union graph over the entire horizon, $G_{[1,T]}$, is connected. However, this condition is too weak; as we formally prove with a counterexample in Appendix A.3, an environment can satisfy this horizon-long connectivity while still trapping the agent in a suboptimal region for a linear number of rounds, leading to linear regret.

This motivates the need for a stronger, recurrent notion of connectivity. In fact, our analysis in Appendix A.4 establishes a sharp threshold for failure: if the number of disjoint time blocks that are disconnected grows with the horizon, then any strategy constrained to local information and local actions is guaranteed to suffer linear regret. This suggests that for an algorithm to be guaranteed to *succeed*, the environment must, at a minimum, ensure that the number of connected blocks is also a constant fraction of the horizon. This leads us to the following set of sufficient conditions.

To formalize these conditions, we measure the graph's properties via a *canonical* lazy random walk, where an agent at any node either stays put or moves to a uniformly chosen neighbor with equal probability. For any graph $G_t = (A, E_t)$, we define its canonical lazy random walk matrix M_t as:

$$M_t(u, v) = \begin{cases} 1/(1 + \deg_t(u)) & \text{if } v = u \text{ or } (u, v) \in E_t \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where the degree of the node $u \in A$ at time t is denoted by $\deg_t(u)$. This matrix provides a standardized way to quantify the connectivity of any given graph G_t through its spectral properties, and the change in connectivity between graphs by the change in their corresponding matrices. With this tool, we can now state our formal assumptions.

Definition 1 ($(\alpha, \beta, \gamma, \nu)$-Admissible Sequence)**.** *Let the time horizon be partitioned into $m = \lfloor T/\alpha \rfloor$ disjoint blocks, $B_k := \{(k-1)\alpha + 1, \ldots, k\alpha\}$ for $k = 1, \ldots, m$. A graph sequence $\{G_t\}_{t=1}^{T}$ is an $(\alpha, \beta, \gamma, \nu)$-admissible sequence if for parameters $\alpha \in \mathbb{N}$, $\beta \in \mathbb{R}$, $\gamma \in (0, 1]$, and $\nu \in (0, 1]$:*

- *(a)* ***Recurrent Reachability:*** *For at least some fraction $\nu \in (0, 1]$ of the blocks B_k, the union graph $G_{B_k} := \bigcup_{t \in B_k} G_t$ is connected, and a canonical lazy random walk on it has a spectral gap of at least γ.*
- *(b)* ***Temporal Stability:*** *For all $t \in \{2, \ldots, T\}$, the change in underlying connectivity is bounded: $||M_t - M_{t-1}||_2 \leq \beta$.*

It is crucial to note that these are assumptions on the *environment*, not the algorithm. The canonical lazy walk matrix M_t serves as a universal "yardstick" to measure the intrinsic properties of the graph sequence $\{G_t\}$. The conditions on reachability (α, γ, ν) and stability (β) are therefore fundamental properties of the world in which the agent operates. Our theorems will show that our algorithms, regardless of their own specific transition logic (like the biased walk of Algorithm 4 introduced in Section 3), can succeed in any environment that adheres to these conditions. We provide proofs for why these conditions are required in Appendix A.

3 A Family of Local Exploration Algorithms

This section introduces a family of algorithms designed to operate under the strict locality and movement constraints of our setting. Their shared, phased design is a direct and principled response to a key structural challenge that distinguishes this problem from classical multi-armed bandits. In classical settings with global arm access, the identification of the best arm immediately enables its exploitation, allowing for tightly interwoven strategies. Our setting, however, introduces a fundamental asymmetry between *identification* and *exploitation*. An agent may become highly confident about the optimal arm's identity but remain physically distant, unable to act on this knowledge without a dedicated navigation strategy.

This forces a structural decoupling of exploration (the global information-gathering process of traversing the graph) and exploitation (the act of navigating to and remaining at the best node). A phased approach is therefore not a matter of algorithmic convenience, but a necessary design choice to avoid accumulating unnecessary regret from local wandering. We also emphasize that all our algorithms make local decisions i.e. they do not need to have access to knowledge of the entire graph. We now present the algorithms built on this core design, beginning with the canonical protocol.

3.1 The Lazy Exploration (LE) Algorithm

We begin with the canonical protocol that embodies the core theoretical insight of our work: that simple, local exploration is sufficient for near-optimal learning in an environment satisfying our structural conditions (Definition 1). The Lazy Exploration (LEX) (Algorithm 1) is a minimalist, two-phase protocol that requires no planning or global knowledge.

The protocol begins with an *exploration phase* lasting for a predefined number of rounds, T_{exp}. In each round t of this phase, the environment presents a new graph realization, G_t. The agent, currently at node a_t, performs a single step of a lazy random walk according to a transition kernel $\mathcal{K}_t$, moving to a new node a_{t+1}. Upon arrival, it observes a reward and updates its empirical mean estimate, $\widehat{\mu}(a_{t+1})$, and visit count, $\hat{N}(a_{t+1})$, for that node. After exploration concludes, the algorithm enters the *commitment phase*. It first identifies the empirically best arm, $\hat{a}^*$, based on the estimates gathered. For the remainder of the horizon, the agent attempts to navigate to and stay at this target, again using lazy walk steps if not already there. Its analysis reveals two practical limitations: its exploration time is fixed by an oracle, and its exploration is "blind" to rewards, motivating the subsequent algorithms.

Algorithm 1 The Lazy Exploration: LEX

Input: Horizon T, exploration length T_{exp}
Initialize: Current node a_1; for all $a \in A$, count $\hat{N}(a) \leftarrow 0$ and empirical mean $\widehat{\mu}(a) \leftarrow 0$.
— Exploration Phase —
for $t = 1$ to T_{exp} **do**
  Receive graph G_t; agent is at node a_t.
  Move to $a_{t+1} \sim \mathcal{K}_t(G_t, a_t)$.
  Observe reward r_t and update $\widehat{\mu}(a_{t+1})$ and $\hat{N}(a_{t+1})$.
end for
— Commitment Phase —
Let $\hat{a}^* \leftarrow \arg\max_{a \in A} \widehat{\mu}(a)$.
Run Algorithm 2 with $\hat{a}^*$, $a_{T_{\text{exp}}}$, $T_{\text{exp}} + 1$ and T

Algorithm 2 The Navigation or Commit Protocol

Input: Target arm $\hat{a}^*$, current node a_{start}, start time t_{start}, Horizon T
$a_t \leftarrow a_{\text{start}}$
for $t = t_{\text{start}}$ to T **do**
  Receive graph G_t; agent at a_t.
  if $a_t = \hat{a}^*$ **then**
    Stay at a_t.
  else
    Move to $a_{t+1} \sim \mathcal{K}_t(G_t, a_t)$.
  end if
end for

3.2 A Confidence-Based LE Algorithm

A limitation of LEX is that its exploration phase has a fixed length, T_{exp}, which may depend on unknown problem parameters, such as the minimum reward gap, $\Delta_{\min}$. To address this, we introduce a confidence-bound-based algorithm, CB-LEX (Algorithm 3), a practical variant that removes this requirement.

Instead of a fixed schedule, CB-LEX employs a *stopping rule* controlled by a confidence parameter, δ. During its exploration phase, the algorithm continually performs lazy walk steps and updates its reward estimates. After each step t, it computes a confidence width, $w_t(a)$, for each arm based on its visit count. It then identifies the arms with the highest and second-highest empirical means, denoted $a_t^{(1)}$ and $a_t^{(2)}$ respectively. The exploration phase terminates automatically as soon as the empirical gap between these two arms exceeds the sum of their confidence widths, signaling that a statistically significant leader has emerged. This allows the algorithm to adapt its exploration budget to the problem's difficulty, exploring longer on hard instances and stopping early on easy ones, making it well-suited for practical deployment.

3.3 A Reward-Informed Algorithm

While the exploration in LEX and CB-LEX is effective, it is ultimately "blind" to the rewards it observes. Our final variant, Reward-Aware LEX (RA-LEX), presented in Algorithm 4, enhances the exploration process by incorporating local reward information. RA-LEX replaces the uniform lazy random walk with a *biased random walk* that prioritizes locally promising nodes. To achieve this, RA-LEX employs the principle of optimism in the face of uncertainty [19, 2, 20] to guide its random walk.

Algorithm 3 A Practical Lazy Exploration Alg. (CB-LEX)

Input: Horizon T, confidence param. $\delta \in (0,1); c > 0$
Initialize: Current node a_1; for all $a \in A$, $\hat{N}(a) \leftarrow 0$, $\widehat{\mu}(a) \leftarrow 0$, $t \leftarrow 1$
— Exploration Phase —
repeat
  Receive G_t; agent is at node a_t.
  Move to $a_{t+1} \sim \mathcal{K}_t(G_t, a_t)$.
  Observe reward r_t from a_{t+1};
  update $\widehat{\mu}(a_{t+1}), \hat{N}(a_{t+1})$.
  for all $a \in A$ such that $N_t(a) > 0$ **do**
    compute $w_t(a) \leftarrow \sqrt{c \log(nT/\delta) / \max(1, \hat{N}(a))}$.
  end for
  Let $a_t^{(1)} \leftarrow \arg\max_a \widehat{\mu}(a)$;
  Let $a_t^{(2)} \leftarrow \arg\max_{a \neq a_t^{(1)}} \widehat{\mu}(a)$.
  $t \leftarrow t + 1$
until $t > T$ or $\widehat{\mu}(a_{t-1}^{(1)}) - \widehat{\mu}(a_{t-1}^{(2)}) > w_{t-1}(a_{t-1}^{(1)}) + w_{t-1}(a_{t-1}^{(2)})$
— Commitment Phase —
Let $\hat{a}^* \leftarrow a_{t-1}^{(1)}$
Run Algorithm 2 with $\hat{a}^*$, a_t, t and T

Algorithm 4 A Reward-Aware Algorithm (RA-LEX)

Input: Horizon T, confidence δ and $\lambda, c > 0$
Initialize: Current node a_1; for all $a \in A$, $\hat{N}(a) \leftarrow 0$, $\widehat{\mu}(a) \leftarrow 0$, $t \leftarrow 1$.
— Biased Adaptive Exploration Phase —
repeat
  Receive G_t; agent at a_t. Let $N_t = N_t(a_t) \cup \{a_t\}$.
  for all $a \in N_t$ **do**
    $\xi_t(a) \leftarrow \widehat{\mu}(a) + \sqrt{c \log(t) / \max(1, \hat{N}(a))}$.
    Set transition weight $W_t(a) \leftarrow e^{\lambda \cdot \xi_t(a)}$.
  end for
  Set Probabilities $P_t(a) = W_t(a) / \sum_{b \in N_t} W_t(b)$.
  Move to $a_{t+1} \sim P_t(\cdot)$.
  Observe reward r_t and update $\widehat{\mu}(a_{t+1})$ and $\hat{N}(a_{t+1})$.
  Let $a_t^{(1)}, a_t^{(2)}$ be the top two arms by empirical mean.
  Compute widths $w_t(a_t^{(1)})$ and $w_t(a_t^{(2)})$.
  $t \leftarrow t + 1$.
until $t > T$ or $\widehat{\mu}(a_{t-1}^{(1)}) - \widehat{\mu}(a_{t-1}^{(2)}) > w_{t-1}(a_{t-1}^{(1)}) + w_{t-1}(a_{t-1}^{(2)})$
— Commitment Phase —
Let $\hat{a}^* \leftarrow a_{t-1}^{(1)}$.
Run Algorithm 2 with $\hat{a}^*$, a_t, t and T

During its adaptive exploration phase, for each node a in its immediate neighborhood $N_t(a)$, the agent computes an optimistic value estimate. This estimate, formally an Upper Confidence Bound (UCB) index (line 7, Algorithm 4) and denoted by $\xi_t(a)$, is the sum of two components: the current empirical mean $\widehat{\mu}(a)$, which encourages *exploitation* of known good nodes, and a confidence bonus that is larger for nodes with fewer visits, which encourages *exploration* of uncertain options. These optimistic indices are then fed into a softmax distribution, controlled by a bias parameter $\lambda > 0$, to form the transition probabilities $P_t(a)$. A larger λ creates a more greedy walk that strongly favors nodes with high UCB scores, while $\lambda \to 0$ recovers a uniform random walk. By sampling its next move from this distribution, the agent biases its exploration toward promising regions without sacrificing the guarantee of eventual global coverage.

4 Main Theoretical Results

We now present the formal regret guarantees for the family of algorithms introduced previously. A key feature of these results is that they hold for any graph sequence satisfying the $(\alpha, \beta, \gamma, \nu)$-admissible conditions (Definition 1), without relying on the specifics of the graph's generative process. We begin with the guarantee for our canonical protocol, LEX, and then demonstrate that our algorithms that incorporate confidence-bound based stopping times achieve similar near-optimal regret without requiring prior knowledge of the problem instance parameters. All proofs are deferred to the appendix.

4.1 Guarantee for the Canonical Protocol

We begin with the canonical LEX protocol. The theorem establishes that with high probability, the regret is controlled by the two components identified in our "Anatomy of Regret."

Theorem 1 (High-Probability Regret of LEX)**.** *Assume the graph sequence is an $(\alpha, \beta, \gamma, \nu)$-admissible sequence (Definition 1). For any confidence parameter $\delta \in (0, 1)$, consider the LEX algorithm run with an exploration length $T_{\exp} = \Omega\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2}\right)$. Then with probability at least $1 - \delta$, the cumulative regret $R(T)$ is bounded by:*

$$R(T) \leq O\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right). \tag{3}$$

By setting the failure probability δ to a polynomially small value, we obtain a clean, near-optimal bound on the expected regret.

Corollary 1 (Expected Regret of LEX)**.** *Under the conditions of the theorem, setting $\delta = 1/T^2$ yields an expected cumulative regret of:*

$$\mathbb{E}[R(T)] \leq O\left(\frac{\alpha n^2 \log(nT)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right). \tag{4}$$

4.2 Guarantees for Confidence-Bound Algorithms

While Theorem 1 establishes the viability of the phased approach, its reliance on a pre-set $T_{\exp}$ is impractical. We now show that our confidence-bound based algorithms remove this limitation while retaining near-optimal guarantees.

Theorem 2 (CB-LEX: High-probability regret)**.** *Assume the graph sequence is an $(\alpha, \beta, \gamma, \nu)$-admissible sequence (Definition 1). For any confidence parameter $\delta \in (0, 1)$, the CB-LEX algorithm (with anytime CI stopping) identifies a^* with probability at least $1 - 2\delta$, and the cumulative regret satisfies*

$$R(T) \;\leq\; O\left(\max\left\{\underbrace{\frac{\alpha\, n^2}{\nu(\alpha - \tau_{mix})} \max_{a \neq a^*} \frac{\log\big(2nT/\delta\big)}{\Delta(a)^2}}_{\text{CI threshold branch}},\; \underbrace{\frac{\alpha^3}{\nu^2(\alpha - \tau_{mix})^2}\, n^4 \log\Big(\tfrac{n}{\delta}\Big)}_{\text{uniform visitation branch}}\right\} + \underbrace{\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}}_{\text{navigation}}\right).$$

In particular, this bound requires no prior knowledge of $\Delta_{\min}$; the exploration term depends on the instance via $\max_{a \neq a^*} 1/\Delta(a)^2$.

We set the failure probability δ to a polynomially small value and obtain the following.

Corollary 2 (CB-LEX: Expected regret)**.** *Under the conditions of the theorem, setting $\delta = 1/T^2$ yields*

$$\mathbb{E}[R(T)] \;\le\; O\!\left(\max\!\left\{\frac{\alpha\, n^2}{\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log n\right\}\;+\;\frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\right).$$

Our efficient, reward-aware protocol achieves a similar guarantee, with practical implications discussed below.

Theorem 3 (RA-LEX: High-probability regret)**.** *Assume the graph sequence is an $(\alpha,\beta,\gamma,\nu)$-admissible sequence (Definition 1). For any confidence parameter $\delta\in(0,1)$ and bias $\lambda>0$, the RA-LEX algorithm identifies the optimal arm a^* with probability at least $1-2\delta$, and the cumulative regret satisfies*

$$R(T) \;\le\; O\!\left(\max\!\left\{\frac{\alpha n^2}{\kappa\,\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT/\delta)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log\!\left(\tfrac{n}{\delta}\right)\right\}\;+\;\frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\right),$$

where $\kappa=\exp\{-\lambda(1+O(\sqrt{\log(nT/\delta)}))\}$ is the minorization factor induced by the softmax bias.

Now, setting the failure probability δ to a polynomially small value, we get the following.

Corollary 3 (RA-LEX: Expected regret)**.** *Under the conditions of the theorem, setting $\delta = 1/T^2$ yields*

$$\mathbb{E}[R(T)] \;\le\; O\!\left(\max\!\left\{\frac{\alpha n^2}{\kappa\,\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log n\right\}\;+\;\frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\right).$$

On the Performance of RA-LEX. While RA-LEX and CB-LEX share the same asymptotic regret order, RA-LEX demonstrates superior finite-time performance due to (a) improved sample efficiency and (b) reduced performance variance. Its reward-aware exploration gathers useful information more quickly than the "blind" walk of CB-LEX by biasing its trajectory toward promising regions. This allows it to satisfy its statistical stopping condition with fewer steps, resulting in a smaller leading constant in the regret bound ($c_3' < c_2'$) and lower overall regret. Furthermore, this reward-aware strategy acts as a variance reduction mechanism. By actively guiding its trajectory with reward information, RA-LEX reduces its dependence on the stochastic luck that can cause high performance variance in CB-LEX, leading to more stable and predictable outcomes. This behavior is governed by the bias parameter λ, which controls the exploration-exploitation trade-off within the exploration phase itself; a moderate value is required to balance a biased search with the sufficient stochasticity for guaranteed global exploration. These insights are corroborated by our simulation results in Section 5 and Appendix F.

4.3 Discussion: The Anatomy of Regret

The structure of our regret bounds reveals a fundamental anatomy inherent to learning on dynamic graphs. The total regret decomposes into two distinct components. The first, the *Learning Cost*, scales with $\sum(n/\Delta(a))$ and represents the statistical price of exploration; this is the irreducible cost of distinguishing between arms under bandit feedback, analogous to classical MAB settings. The second, the *Navigation Cost*, scales with graph parameters like n and $1/\gamma$ and represents the physical price of traversing the graph. This is the unavoidable cost incurred while the agent, even after identifying the optimal arm, must physically navigate the dynamic environment to reach it.

5 Simulating Learning in a Disaster Zone

To corroborate our theoretical findings, we instantiate our algorithms in a simulation that captures the core challenges of our problem setting: a robotic ground vehicle deployed to secure communication coverage in a disaster-affected urban region. This environment is designed to be a non-trivial example of a system satisfying our structural conditions, allowing us to test

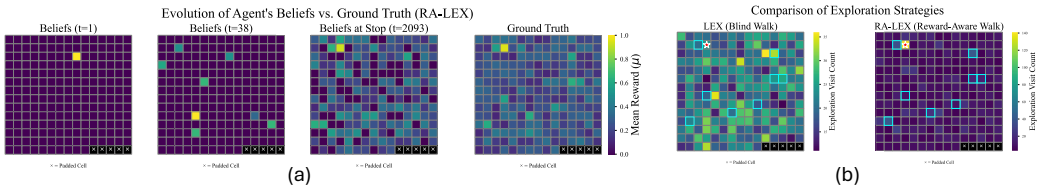

Figure 2: Visual analysis of the learning and exploration process. (a) The RA-LEX algorithm's internal model of the world ($\hat{\mu}$) becomes progressively more accurate, converging to the ground truth by the time exploration stops. (b) A comparison of exploration-phase visit patterns highlights the "blind" but comprehensive search of LEX versus the biased search of RA-LEX.

the practical performance of our algorithms and provide empirical intuition for our theoretical insights.

Environment Setup. The mission environment covers a critical one-square-kilometer incident zone, discretized into a grid of $|A| = 205$ candidate deployment locations. The navigable streets and pathways between zones form the edges of a dynamic graph G_t, which evolves via an Edge-Flip Markovian process (see [7, 8]) to create temporal dependence. The physical pathways flicker in and out of existence due to debris, traffic diversion, or environmental instability. In this model, each potential edge (i, j) follows its own two-state Markov chain: an edge that is absent at time $t - 1$ appears at time t with probability θ_{ij}, while a present edge disappears with probability ζ_{ij}. We set these appearance and disappearance probabilities to reflect realistic terrain differences: (a) Stable Corridors: Edges for major thoroughfares have high appearance and low disappearance probabilities, making them persistent. (b) Unstable Paths: Edges for debris-filled alleys have low appearance and high disappearance probabilities, causing them to "flicker" in and out of existence.

This setup creates a challenging, non-uniform dynamic graph that satisfies our structural conditions but is far from a simple i.i.d. model. The reward structure is sparse, with a single "hotspot" zone designated as the optimal arm a^* ($\mu(a^*) = 0.95$), representing the single most reliable site for communication relay deployment, a few other zones with moderate rewards, and the vast majority with low rewards, reflecting the clustered nature of high-quality signal sites in an urban setting. When the vehicle visits a zone a_t, it receives a binary reward $r_t(a_t) \sim \text{Bernoulli}(\mu(a_t))$ (representing the strength and stability of the local communication coverage).

Experimental Design. In this environment, we compare the performance of our three proposed algorithms: LEX, CB-LEX and RA-LEX. The simulation runs for $T \approx 70,000$ rounds (with 10 iterations each), modeling a high-intensity, 4-day continuous deployment mission where the vehicle makes local movement decisions. This models a critical disaster response phase where rapid, adaptive learning is essential to efficiently locate and commit to the best communication hub location.

Our experiments are designed to validate our core theoretical insights by analyzing three key aspects of performance: (a) the ability to achieve near-optimal regret, (b) the effectiveness of the adaptive stopping rule under varying problem difficulty, and (c) the efficiency gains from reward-aware exploration. To test these, we evaluate our algorithms on two distinct reward instances. The "hard" instance, used for our main results, features a small suboptimality gap ($\Delta_{\min}$) to create a significant statistical challenge. The "easy" instance features a large $\Delta_{\min}$, making the optimal arm easy to distinguish, and is used specifically to validate the adaptive stopping mechanisms.

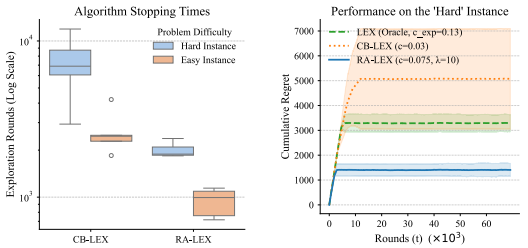

Figure 3: Quantitative performance of the algorithm family: (a) exploration stopping times, (b) cumulative regret.

5.1 Results and Analysis

Our experiments provide strong empirical corroboration for our theoretical claims. The results, summarized in Figures 2 and 3, demonstrate the effectiveness of our algorithms and provide insight into their underlying mechanics.

Near-Optimal Performance. Figure 3 (b) shows the main performance results on the challenging "hard" instance. All three algorithms successfully achieve sublinear regret, as evidenced by the characteristic "flattening" of their cumulative regret curves. The reward-aware RA-LEX achieves the lowest overall regret, a result that empirically corroborates the theoretical insight from our earlier remark in Section 4.2. As predicted, the algorithm's ability to bias its exploration toward promising regions leads to greater sample efficiency and a smaller leading constant in the regret, allowing it to outperform the "blind" exploration of CB-LEX.

A key insight from our results is the performance trade-off between the oracle-tuned LEX and the practical CB-LEX. While CB-LEX outperformed LEX in our experiments, this is an expected and important outcome. The LEX protocol is given a significant, unrealistic advantage: its exploration time T_{exp} is calculated using the ground-truth $\Delta_{\min}$ of the instance, information that is never available in a real-world problem. In contrast, CB-LEX operates without this oracle knowledge. The fact that its performance remains competitive with a perfectly tuned oracle baseline highlights the efficacy of its adaptive design.

Our analysis provides empirical validation for our claims. Figure 3(a) confirms our adaptive algorithms work as intended, with both CB-LEX and RA-LEX stopping significantly earlier on "easy" instances (large $\Delta_{\min}$) than on "hard" ones; the plot also highlights RA-LEX's superior efficiency and stability. Visualizations offer further intuition: Figure 2(b) contrasts the "blind," uniform exploration of LEX with the targeted, "searchlight" pattern of the reward-aware RA-LEX, which more reliably concentrates on and identifies the optimal arm. Furthermore, Figure 2(a) tracks the evolution of RA-LEX's internal beliefs ($\hat{\mu}$), showing that it succeeds by learning an accurate world model that converges to the ground truth. Additional experiments detailing parameter sensitivity are provided in Appendix F.

6 Related Work

Zhang et al. [22] introduced the graph bandit problem. The constraint on the arms is represented as a static graph, thus, allowing for long term planning across time steps. Related studies on dynamic connectivity [13] analyze cover times and reachability in stochastically evolving graphs, but do not address sequential decision making, bandit feedback, or online learning.

The MAB model with switching costs on arms [9, 1, 6] also shares some similarities with the graph bandit model. Our work can be framed as a MAB problem with infinite switching costs on a random subset of arms. Paschalidis et al. [18] extended the graph bandit problem to a multi-agent setting, where N cooperative agents traverse on a connected graph G with K nodes. They present algorithms that achieve sublinear regret in time.

7 Conclusion

This work demonstrates that learnability for a constrained agent on an dynamic graph can be guaranteed under a set of sufficient structural conditions, namely, *Recurrent Reachability* and *Temporal Stability*. This perspective reveals a deeper structure to the problem, exposing an *anatomy of regret* composed of distinct learning and navigation costs. This decomposition provides a new lens for analyzing online learning problems where statistical uncertainty is coupled with physical constraints. Our results show how algorithms can achieve near-optimality by leveraging the environment's dynamics, culminating in our adaptive RA-LEX algorithm. This model-agnostic view opens the door to future research on more sophisticated adaptive policies that learn these structural parameters in real time and on classifying the learnability of complex, real-world dynamic environments.

References

[1] Idan Amir, Guy Azov, Tomer Koren, and Roi Livni. Better best of both worlds bounds for bandits with switching costs. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15800–15810. Curran Associates, Inc., 2022.

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

[3] Djallel Bouneffouf, A. Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *ICONIP*, 2012.

[4] Eric Brochu, Matthew W. Hoffman, and Nando de Freitas. Portfolio Allocation for Bayesian Optimization. *arXiv e-prints*, page arXiv:1009.5419, September 2010.

[5] Sourav Chakraborty and Lijun Chen. Incentivized exploration of non-stationary stochastic bandits. *arXiv preprint arXiv:2403.10819*, 2024.

[6] Lin Chen, Qian Yu, Hannah Lawrence, and Amin Karbasi. Minimax regret of switching-constrained online convex optimization: No phase transition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3477–3486. Curran Associates, Inc., 2020.

[7] Andrea Clementi, Pierluigi Crescenzi, Carola Doerr, Pierre Fraigniaud, Francesco Pasquale, and Riccardo Silvestri. Rumor spreading in random evolving graphs. *Random Struct. Algorithms*, 48(2):290–312, March 2016.

[8] Andrea EF Clementi, Claudio Macci, Angelo Monti, Francesco Pasquale, and Riccardo Silvestri. Flooding time of edge-markovian evolving graphs. *SIAM journal on discrete mathematics*, 24(4):1694–1712, 2010.

[9] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $t^{2/3}$ regret, 2013.

[10] Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 5–22, New York, NY, USA, 2014. Association for Computing Machinery.

[11] Yi Gai, Bhaskar Krishnamachari, and Ramesh Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. pages 1 – 9, 05 2010.

[12] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[13] Ioannis Lamprou, Russell Martin, and Paul Spirakis. Cover time in edge-uniform stochastically-evolving graphs. *arXiv preprint arXiv:1702.05412*, 2017.

[14] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[15] David Asher Levin, Yuval Peres, and Elizabeth L Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2017.

[16] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web - WWW '10*, 2010.

[17] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[18] Phevos Paschalidis, Runyu Zhang, and Na Li. Cooperative multi-agent graph bandits: Ucb algorithm and regret analysis. *arXiv preprint arXiv:2401.10383*, 2024.

[19] Herbert Robbins. Some aspects of the sequential design of experiments. 1952.

[20] Aleksandrs Slivkins. Introduction to multi-armed bandits, 2024.

[21] Siwei Wang and Longbo Huang. Multi-armed bandits with compensation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[22] Tianpeng Zhang, Kasper Johansson, and Na Li. Multi-armed bandit learning on a graph. In *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2023.

A Necessary Conditions for Local Learning on Dynamic Graphs

This appendix restates our environmental sufficiency condition in a unified form and gives rigorous lower bounds showing why each part is needed. Throughout, we work on a fixed, finite arm set $A = \{1, \dots, n\}$ with $n \geq 2$. Rewards are node-specific and time-invariant: each arm $a \in A$ has an unknown distribution with mean $\mu(a) \in [0, 1]$. We assume a unique optimal arm

$$a^\star \in \arg\max_{a \in A} \mu(a), \qquad \Delta(a) \triangleq \mu(a^\star) - \mu(a) > 0 \text{ for } a \neq a^\star,$$

and we denote $\Delta_{\min} \triangleq \min_{a \neq a^\star} \Delta(a)$ when needed. For the lower bounds below we instantiate a simple instance with $\mu(a^\star) = 1$ and $\mu(a) = 0$ for $a \neq a^\star$ (thus $\Delta_{\min} = 1$), but all arguments extend by multiplying regrets by $\Delta_{\min}$.

Local policies and feedback. Time is indexed by $t = 1, \dots, T$. At the *start* of round t the environment specifies a graph $G_t = (A, E_t)$ on the arms. The learner occupies some node $a_{t-1} \in A$ (for $t = 1$, a_0 is the start node). The learner observes *at most* the immediate neighborhood $N_t(a_{t-1}) \triangleq \{b \in A : (a_{t-1}, b) \in E_t\}$, and must choose $a_t \in N_t(a_{t-1}) \cup \{a_{t-1}\}$. It then observes a reward drawn from $\mathcal{D}(a_t)$ (bandit feedback). A *local policy* π is any (possibly randomized) mapping from the history $\mathcal{H}_{t-1}$ and the current observed neighborhood $N_t(a_{t-1})$ to a distribution over $N_t(a_{t-1}) \cup \{a_{t-1}\}$. (Our lower bounds hold even if the learner is revealed the *entire* G_t; restricting to local information only strengthens them.)

A.1 Canonical lazy walk, spectral gap, and norms

All spectral quantities below are defined for the *canonical lazy walk* associated with a static graph $G = (A, E)$. Its transition kernel $M \in \mathbb{R}^{n \times n}$ is

$$M(u, v) \;=\; \begin{cases} \dfrac{1}{1 + \deg(u)} & \text{if } v = u \text{ or } (u, v) \in E, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\deg(u)$ is the degree of u in G. Let $\mathbf{1}$ denote the all-ones vector. Then M is a stochastic matrix, reversible with respect to

$$\pi(u) \;\triangleq\; \frac{1 + \deg(u)}{\sum_{w \in A}(1 + \deg(w))} \;=\; \frac{1 + \deg(u)}{n + 2|E|}. \tag{6}$$

Reversibility: for any $u \neq v$, $\pi(u)M(u, v) = \pi(v)M(v, u) = \frac{\mathbb{I}\{(u,v) \in E\}}{n+2|E|}$ and, trivially, $\pi(u)M(u, u) = \pi(u) \cdot \frac{1}{1+\deg(u)} = \frac{1}{n+2|E|}$. Thus π is stationary and M has a real spectrum $1 = \lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_n(M) \geq -1$. We define the (absolute) *spectral gap*

$$\mathrm{gap}(M) \;\triangleq\; 1 - \max\{\, |\lambda_2(M)|,\, |\lambda_n(M)| \,\}.$$

We use the standard spectral (operator) norm $\|\cdot\|_2$ induced by the Euclidean norm on $\mathbb{R}^n$:

$$\|A\|_2 \;\triangleq\; \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}\,.$$

Two basic identities will be used frequently.

Lemma 1 (Edge-flow identity for conductance)**.** *Let $G = (A, E)$, M as in* (5)*, and π as in* (6)*. For any $S \subset A$,*

$$Q(S, S^{\complement}) \;\triangleq\; \sum_{u \in S} \sum_{v \in S^{\complement}} \pi(u)M(u, v) \;=\; \frac{e(S, S^{\complement})}{n + 2|E|}, \tag{7}$$

$$\pi(S) \;=\; \sum_{u \in S} \pi(u) \;=\; \frac{|S| + 2e(S) + e(S, S^{\complement})}{n + 2|E|}, \tag{8}$$

where $e(S)$ is the number of edges with both endpoints in S and $e(S, S^{\complement})$ is the number of edges crossing $(S, S^{\complement})$.

Proof. For (7), by reversibility and the definition of M, each undirected edge $(u, v) \in E$ with $u \in S$, $v \in S^{\complement}$ contributes exactly $\frac{1}{n+2|E|}$ to the sum, and non-edges contribute 0. For (8), $\sum_{u\in S}(1 + \deg(u)) = |S| + \sum_{u\in S} \deg(u) = |S| + 2e(S) + e(S, S^{\complement})$. □

The (lazy) conductance of S is $\Phi(S) \triangleq Q(S, S^{\complement})/\min\{\pi(S), \pi(S^{\complement})\}$ and $\Phi \triangleq \min_{S\subset A:\, \pi(S)\le 1/2} \Phi(S)$. Cheeger's inequalities for reversible chains yield

$$\frac{\Phi^2}{2} \;\le\; \mathrm{gap}(M) \;\le\; 2\Phi\,. \tag{9}$$

A.2 Sufficient condition ($(\alpha, \beta, \gamma, \nu)$-diffusivity)

We provide sufficient conditions:

Definition 2 ($(\alpha, \beta, \gamma, \nu)$-diffusive sequence)**.** *Fix $\alpha \in \mathbb{N}$, $\beta \ge 0$, and $\gamma, \nu \in (0, 1]$. Partition $\{1, \dots, T\}$ into $m \triangleq \lfloor T/\alpha \rfloor$ disjoint blocks $B_k \triangleq \{(k-1)\alpha + 1, \dots, k\alpha\}$ for $k = 1, \dots, m$. A dynamic graph sequence $\{G_t = (A, E_t)\}_{t=1}^{T}$ is* $(\alpha, \beta, \gamma, \nu)$-diffusive *if:*

(i) ***Recurrent reachability.*** *For at least a ν-fraction of the blocks B_k, the union graph*

$$G_{B_k} \;\triangleq\; \bigcup_{t\in B_k} G_t$$

is connected, and the canonical lazy walk on G_{B_k} has spectral gap at least γ.

(ii) ***Temporal stability.*** *For all $t \in \{2, \dots, T\}$,*

$$\|M_t - M_{t-1}\|_2 \;\le\; \beta,$$

where M_t is the canonical lazy kernel (5) *on G_t.*

Remark 1 (On the norm choice)**.** *The spectral norm $\|\cdot\|_2$ is a strong notion of stepwise kernel stability and suffices for our analyses. (Other choices such as the $\ell_2(\pi)$ operator norm or total-variation drift can also be used, with corresponding technical modifications.)*

A.3 Horizon-long connectivity is insufficient

We first show that requiring only that the *horizon-wide* union graph is connected (even with a *constant* spectral gap) does not prevent linear regret.

Lemma 2 (Complete graph has unit spectral gap)**.** *Let $G = K_m$ be the complete graph on $m \ge 2$ nodes. The canonical lazy kernel* (5) *on K_m satisfies $M = \frac{1}{m}\mathbf{1}\mathbf{1}^{\top}$, hence* $\mathrm{gap}(M) = 1$.

Proof. On K_m, $\deg(u) = m - 1$ for all u. Thus for all u, v we have $M(u, v) = \frac{1}{1+(m-1)} = \frac{1}{m}$. Hence M is the rank-one averaging operator $M = \frac{1}{m}\mathbf{1}\mathbf{1}^{\top}$, with eigenvalues 1 (mult. 1) and 0 (mult. $m - 1$). □

Proposition 1 (Horizon-long connectivity does not prevent linear regret)**.** *There exist $n \ge 2$, $T \ge 2$, a reward instance with unique optimal arm $a^\star$ and gap $\Delta_{\min} = 1$, and a graph sequence $\{G_t\}_{t=1}^{T}$ such that:*

(a) The horizon-wide union $G_{[1,T]} \triangleq \bigcup_{t=1}^{T} G_t$ is the complete graph K_n; therefore, by Lemma 2, the lazy walk on $G_{[1,T]}$ has spectral gap 1.

(b) For any (possibly randomized) local policy π and a uniform random start $a_0 \sim \mathrm{Unif}(A)$,

$$\mathbb{E}[R(T)] \;\ge\; \left(1 - \frac{1}{n}\right)(T - 1).$$

Proof. Let $A_1 \triangleq \{a^\star\}$ and $A_2 \triangleq A \setminus \{a^\star\}$, with $\mu(a^\star) = 1$ and $\mu(a) = 0$ for $a \ne a^\star$. Define G_t as follows:

- For $t = 1, \dots, T-1$, let G_t be the disjoint union of the isolated node $a^\star$ and the complete graph on A_2.
- For $t = T$, let $G_T = K_n$ (the complete graph on all nodes).

Then $G_{[1,T]} = K_n$, proving (a) by Lemma 2. For (b), if $a_0 = a^\star$ (probability $1/n$), then $R(T) = 0$ since the learner can stay at $a^\star$ each round. Otherwise, for $t = 1, \dots, T-1$ the learner is confined to A_2 and cannot play $a^\star$ since there is no path to it. Hence it accrues regret 1 in each of the first $T-1$ rounds. Therefore

$$\mathbb{E}[R(T)] \;\geq\; \Big(1 - \frac{1}{n}\Big) \cdot (T-1),$$

as claimed. □

A.4 Linear regret under a constant fraction of disconnected blocks

We next show that if a constant fraction of disjoint α-blocks have disconnected union graphs, then linear regret follows for any local policy.

Proposition 2 (Recurrent disconnection forces linear regret)**.** *Fix $\alpha \in \mathbb{N}$ and $c \in (0, 1]$. There exist a reward instance with unique optimal $a^\star$ and gap $\Delta_{\min} = 1$, and a graph sequence $\{G_t = (A, E_t)\}_{t=1}^T$ such that:*

(a) For at least $c \cdot \lfloor T/\alpha \rfloor$ disjoint blocks B_k of length α, the union graph G_{B_k} is disconnected.

(b) For any (possibly randomized) local policy π and $a_0 \sim \mathrm{Unif}(A)$,

$$\mathbb{E}[R(T)] \;\geq\; \Big(1 - \frac{1}{n}\Big)\, cT.$$

Proof. Let $\mu(a^\star) = 1$ and $\mu(a) = 0$ for $a \neq a^\star$. Partition time into blocks B_k as in Definition 2. For the first $\lfloor c\, m \rfloor$ blocks, define G_t for each $t \in B_k$ so that $a^\star$ is isolated and $A \setminus \{a^\star\}$ induces a connected subgraph (e.g., a clique). Then G_{B_k} is disconnected for those blocks. For the remaining blocks, let G_t be arbitrary (e.g., complete).

If $a_0 = a^\star$ (probability $1/n$), then $R(T) = 0$. Otherwise, during each of the first $\lfloor c\, m \rfloor$ blocks, $a^\star$ is isolated and cannot be played, so the learner incurs regret 1 at every round in those blocks, totaling at least $\alpha \lfloor c\, m \rfloor \geq cT - \alpha$. Thus

$$\mathbb{E}[R(T)] \;\geq\; \Big(1 - \frac{1}{n}\Big)\,(cT - \alpha) \;\geq\; \Big(1 - \frac{1}{n}\Big)\, cT - 1,$$

and the stated bound follows since $cT \geq 1$ for $T \geq 1/c$; otherwise the claim is trivial. □

A.5 Temporal stability is necessary (adaptive adversary)

We now show that without a bound on the per-step kernel drift, even very strong recurrent reachability (connected unions with constant spectral gap in *every* block) cannot prevent linear regret. This construction is *adaptive*: at each round t, the environment chooses G_t after observing the learner's current node a_{t-1}.[2]

We first record the spectral gap of the lazy walk on a star.

Lemma 3 (Spectral gap on the star)**.** *Let S_n be the star on $n \geq 2$ nodes with center c and leaves $L = A \setminus \{c\}$. For the canonical lazy kernel M on S_n,*

$$\mathrm{gap}(M) \;=\; \frac{1}{2}.$$

Proof. Label the nodes so that $c = 1$ and leaves are $2, \dots, n$. From (5):

$$M(1,1) = \frac{1}{n}, \quad M(1,j) = \frac{1}{n}\ (j \geq 2); \qquad M(i,i) = \frac{1}{2},\ M(i,1) = \frac{1}{2},\ M(i,j) = 0\ (i \geq 2,\ j \geq 2,\ j \neq i).$$

[2] This is a standard adversarial model in online learning; our upper bounds do *not* require such adversarial power. The result here demonstrates the necessity of some form of temporal stability in worst-case environments.

Consider the $(n-2)$-dimensional subspace $U = \{x \in \mathbb{R}^n : x_1 = 0, \sum_{j=2}^n x_j = 0\}$. For any $x \in U$, $(Mx)_i = \frac{1}{2}x_i$ for $i \geq 2$ and $(Mx)_1 = \frac{1}{n}\sum_{j=2}^n x_j = 0$, hence U is an eigenspace with eigenvalue $\lambda = \frac{1}{2}$. Next, restrict to vectors of the form $x = (a, b, \ldots, b)$ (constant on leaves). Then

$$(Mx)_1 \;=\; \tfrac{1}{n}a + \tfrac{n-1}{n}b, \qquad (Mx)_i \;=\; \tfrac{1}{2}a + \tfrac{1}{2}b \quad (i \geq 2).$$

Thus eigenvectors satisfy

$$\lambda a \;=\; \tfrac{1}{n}a + \tfrac{n-1}{n}b, \qquad \lambda b \;=\; \tfrac{1}{2}a + \tfrac{1}{2}b.$$

Eliminating a yields the quadratic

$$2\big(\lambda - \tfrac{1}{n}\big)\big(\lambda - \tfrac{1}{2}\big) \;=\; \tfrac{n-1}{n},$$

whose solutions are $\lambda_1 = 1$ and $\lambda_2 = -\frac{1}{2} + \frac{1}{n}$. Therefore the spectrum of M consists of 1, $\frac{1}{2}$ (mult. $n-2$), and $-\frac{1}{2} + \frac{1}{n}$, so $\max\{|\lambda_2|, |\lambda_n|\} = \frac{1}{2}$ and $\mathrm{gap}(M) = 1 - \frac{1}{2} = \frac{1}{2}$. □

We also quantify the size of a "toggle" that connects an isolated node by a single edge.

Lemma 4 (A single-edge toggle induces $\|M_t - M_{t-1}\|_2 \geq \frac{1}{2}$)**.** *Let G^{iso} be any graph in which $a^\star$ is isolated, and G^{conn} be the graph obtained by* adding *the single edge $(a^\star, c)$ for some $c \in A \setminus \{a^\star\}$. Let M^{iso} and M^{conn} be the associated canonical lazy kernels. Then*

$$\big\| M^{\mathrm{conn}} - M^{\mathrm{iso}} \big\|_2 \;\geq\; \frac{1}{2}.$$

Proof. Let $\Delta \triangleq M^{\mathrm{conn}} - M^{\mathrm{iso}}$. The spectral norm satisfies $\|\Delta\|_2 \geq \|\Delta e_{a^\star}\|_2$, where $e_{a^\star}$ is the $a^\star$-th standard basis vector, since $\|e_{a^\star}\|_2 = 1$. The column of Δ corresponding to $a^\star$ is

$$\Delta(a^\star, a^\star) = M^{\mathrm{conn}}(a^\star, a^\star) - M^{\mathrm{iso}}(a^\star, a^\star) = \tfrac{1}{2} - 1 = -\tfrac{1}{2},$$

$$\Delta(c, a^\star) = M^{\mathrm{conn}}(c, a^\star) - M^{\mathrm{iso}}(c, a^\star) = \tfrac{1}{1+\deg_{G^{\mathrm{conn}}}(c)} - 0 \;\geq\; 0,$$

and $\Delta(u, a^\star) = 0$ for $u \notin \{a^\star, c\}$. In particular $\Delta(c, a^\star) = \frac{1}{1+\deg_{G^{\mathrm{conn}}}(c)} \geq \frac{1}{n}$, though we will not need this explicit value. Therefore

$$\|\Delta e_{a^\star}\|_2^2 \;=\; \big(\Delta(a^\star, a^\star)\big)^2 + \big(\Delta(c, a^\star)\big)^2 \;\geq\; \Big(\frac{1}{2}\Big)^2,$$

implying $\|\Delta\|_2 \geq \|\Delta e_{a^\star}\|_2 \geq \frac{1}{2}$. □

We can now state and prove the stability lower bound.

Proposition 3 (Necessity of temporal stability under an adaptive adversary)**.** *Fix any $\alpha \in \mathbb{N}$. There exists an* adaptive *environment and a reward instance with unique optimal $a^\star$ and gap $\Delta_{\min} = 1$ such that for every local policy π and uniform start $a_0 \sim \mathrm{Unif}(A)$:*

(a) For every *block B_k of length α, the union G_{B_k} is connected and its canonical lazy walk has spectral gap $\geq \frac{1}{2}$.*

(b) For infinitely many t, $\|M_t - M_{t-1}\|_2 \geq \frac{1}{2}$ (hence the sequence is not $(\alpha, \beta, \gamma, \nu)$-diffusive for any $\beta < \frac{1}{2}$).

(c) The learner never *plays $a^\star$; thus, for all $T \geq 1$,*

$$\mathbb{E}[R(T)] \;\geq\; \Big(1 - \frac{1}{n}\Big)\, T.$$

Proof. Fix an arbitrary policy π. We construct the environment online. For each round $t = 1, 2, \ldots$:

- If $t \equiv 0 \pmod{\alpha}$ (the last round of a block), reveal G_t to be the star S_n with center $c_t \neq a_{t-1}$, and with the edge $(a^\star, c_t)$ present. (Because $c_t \neq a_{t-1}$, the learner's current node a_{t-1} is a *leaf* in S_n.)

- Otherwise (all non-terminal rounds in blocks), reveal G_t in which $a^\star$ is isolated and the remaining $n-1$ nodes form an arbitrary connected subgraph (e.g., a clique). In particular, there is *no* edge incident to $a^\star$ at such t.

This fully specifies $\{G_t\}_{t\geq 1}$ as a function of the realized trajectory $\{a_{t-1}\}$ (hence adaptive). We verify the claims:

(a) Recurrent reachability with constant spectral gap. In each block B_k, the union G_{B_k} contains the star S_n appearing at the block terminal round; thus G_{B_k} is connected. By Lemma 3, the lazy walk on S_n (hence on G_{B_k}) has spectral gap $\frac{1}{2}$, because adding edges to a connected graph cannot reduce the spectral gap of the canonical lazy walk below that of a connected subgraph.[3]

(b) Large per-step drift infinitely often. At the boundary between rounds $t = \alpha, 2\alpha, 3\alpha, \ldots$, the graph toggles from having $a^\star$ isolated $(t-1)$ to having a single edge $(a^\star, c_t)$ present (t). By Lemma 4, $\|M_t - M_{t-1}\|_2 \geq \frac{1}{2}$ at each such boundary. Hence the temporal stability condition with any $\beta < \frac{1}{2}$ fails infinitely often.

(c) The learner never plays $a^\star$. Fix any t. If $t \not\equiv 0 \pmod{\alpha}$, then $a^\star$ is isolated in G_t and cannot be chosen. If $t \equiv 0 \pmod{\alpha}$, then G_t is a star centered at $c_t \neq a_{t-1}$. The learner's feasible set is $N_t(a_{t-1}) \cup \{a_{t-1}\} = \{c_t, a_{t-1}\}$, since a_{t-1} is a leaf. In particular, $a^\star \notin N_t(a_{t-1}) \cup \{a_{t-1}\}$ unless $a^\star = a_{t-1}$, which cannot happen since $a^\star$ was isolated (and thus unreachable) in all previous rounds of the block and the start a_0 equals $a^\star$ only with probability $1/n$. Therefore the learner never plays $a^\star$ unless it starts there at $t = 0$. Consequently, for all horizons T,

$$\mathbb{E}[R(T)] \;\geq\; \Pr[a_0 \neq a^\star] \cdot T \;=\; \left(1 - \frac{1}{n}\right) T. \qquad \square$$

Remark 2 (Oblivious vs. adaptive environments)**.** *Proposition 3 uses adaptivity only to ensure the star center c_t at the terminal round of each block differs from the learner's current node. Whether an analogous* oblivious *(precommitted) construction can force linear regret without temporal stability remains open and is orthogonal to our upper bounds.*

B Appendix: Analysis for LEX

This appendix provides the full, formal proof for the regret guarantee of the canonical LEX algorithm, as stated in Theorem 1 and Corollary 1. The proof hinges on two key supporting lemmas. Lemma 5 establishes that the agent explores the graph sufficiently. Lemma 6 bounds the cost of navigating to the best arm after exploration. We conclude by combining these lemmas to prove the main theorem. We begin by restating the main results for convenience.

Theorem (Restated)**.** *Assume the graph sequence is an $(\alpha, \beta, \gamma, \nu)$-admissible sequence (Definition 1). For any confidence parameter $\delta \in (0,1)$, consider the LEX algorithm run with an exploration length $T_{\text{exp}} = \Omega\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2}\right)$. Then with probability at least $1-\delta$, the cumulative regret $R(T)$ is bounded by:*

$$R(T) \leq O\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right). \tag{10}$$

Corollary (Restated)**.** *Under the conditions of the theorem, setting $\delta = 1/T^2$ yields an expected cumulative regret of:*

$$\mathbb{E}[R(T)] \leq O\left(\frac{\alpha n^2 \log(nT)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right). \tag{11}$$

Lemma 5 (Uniform Visitation Guarantee)**.** *Let $\{G_t\}_{t=1}^{T_{\text{exp}}}$ be $(\alpha, \beta, \gamma, \nu)$-admissible (Definition 1). Let $\{X_t\}$ be the trajectory of the canonical lazy random walk with kernels $\{M_t\}$. Define $\varphi(a) := \sum_{t=1}^{T_{\text{exp}}} \mathbf{1}(X_t = a)$. Fix $\epsilon \in \left(0, \frac{1}{2n^2}\right]$ and $\delta \in (0,1)$.Let $\tau_{\text{mix}}(\epsilon)$ be the smallest $\tau \in \mathbb{N}$ such that*

[3] Formally: if H is a connected subgraph of G on the same vertex set, the Dirichlet form of the lazy chain on G dominates that on H edgewise, so the spectral gap on G is at least that on H; cf. the variational characterization via Rayleigh quotients.

$(n/2)e^{-(\gamma-\alpha\beta)\tau} \le \epsilon/2$ *and* $\tau \cdot n\alpha\beta \le \epsilon/2$. *If the length of the exploration phase satisfies*

$$T_{\exp} \ge \frac{8\,\alpha^3}{\nu^2\,(\alpha-\tau_{\mathrm{mix}})^2\,\left(\frac{1}{n^2}-\epsilon\right)^2} \cdot \log\!\left(\frac{n}{\delta}\right), \tag{12}$$

then with probability at least $1-\delta$,

$$\forall a \in A: \quad \varphi(a) \;\ge\; \frac{\nu\,(\alpha-\tau_{\mathrm{mix}})\,\left(\frac{1}{n^2}-\epsilon\right)}{2\,\alpha}\,T_{\exp}. \tag{13}$$

Proof. Divide the horizon $T_{\exp}$ into $K := T_{\exp}/\alpha$ disjoint blocks $B_k := \{t_k, \dots, t_k+\alpha-1\}$ with $t_k := (k-1)\alpha+1$. Define

$$\varphi_k(a) := \sum_{t\in B_k} \mathbf{1}\,(X_t = a)\,, \quad \text{so that } \varphi(a) = \sum_{k=1}^{K} \varphi_k(a).$$

Step 1: Coupling to averaged kernel and bounding mixing error. Fix a block $B_k := \{t_k, \dots, t_k+\alpha-1\}$ and define the block-averaged kernel:

$$\bar{M}^{(k)} := \frac{1}{\alpha}\sum_{t\in B_k} M_t.$$

Let $\pi^{(k)}$ denote the stationary distribution of $\bar{M}^{(k)}$. Since each M_t is the lazy random walk kernel on G_t (which is reversible, stochastic, and lazy), the average $\bar{M}^{(k)}$ also has those properties.

To analyze the properties of the block-averaged kernel, we begin by defining an auxiliary matrix for each graph $G_t = (A, E_t)$. Let A_t be the adjacency matrix of G_t with self-loops added:

$$A_t(u,v) \triangleq \begin{cases} 1, & \text{if } v = u \text{ or } (u,v) \in E_t, \\ 0, & \text{otherwise.} \end{cases}$$

By construction, A_t is a symmetric matrix. The lazy random walk matrix M_t is then given by $M_t(u,v) = A_t(u,v)/\,(1+\deg_t(u))$.

We now define two key quantities by averaging over a block of timesteps B_k: the averaged lazy adjacency matrix $\bar{A}^{(k)}$ and the averaged lazy degree $s^{(k)}(u)$.

$$\bar{A}^{(k)} \triangleq \frac{1}{\alpha}\sum_{t\in B_k} A_t,$$

$$s^{(k)}(u) \triangleq \frac{1}{\alpha}\sum_{t\in B_k} (1+\deg_t(u))\,.$$

The block-averaged kernel $\bar{M}^{(k)}$ is defined as the ratio of these two quantities:

$$\bar{M}^{(k)}(u,v) \triangleq \frac{\bar{A}^{(k)}(u,v)}{s^{(k)}(u)}.$$

This kernel $\bar{M}^{(k)}$ governs a reversible Markov chain. To verify this, let $\pi^{(k)}$ be a probability measure on A such that $\pi^{(k)}(u) \propto s^{(k)}(u)$. The detailed balance condition for $\bar{M}^{(k)}$ with respect to $\pi^{(k)}$ holds, as $\bar{A}^{(k)}$ is symmetric:

$$\begin{aligned}
\pi^{(k)}(u)\,\bar{M}^{(k)}(u,v) &= \frac{s^{(k)}(u)}{Z_k}\cdot\frac{\bar{A}^{(k)}(u,v)}{s^{(k)}(u)} = \frac{\bar{A}^{(k)}(u,v)}{Z_k} \\
&= \frac{\bar{A}^{(k)}(v,u)}{Z_k} = \frac{s^{(k)}(v)}{Z_k}\cdot\frac{\bar{A}^{(k)}(v,u)}{s^{(k)}(v)} \\
&= \pi^{(k)}(v)\,\bar{M}^{(k)}(v,u),
\end{aligned}$$

where $Z_k = \sum_{w\in A} s^{(k)}(w)$ is the normalization constant. Since $A_t(u,u) = 1$ for all t, $\bar{M}^{(k)}(u,u)$ is strictly positive, making the chain aperiodic. On any block B_k where the union graph $G_{B_k} = \bigcup_{t\in B_k} G_t$ is connected, the chain defined by $\bar{M}^{(k)}$ is also irreducible.

Finally, we establish a uniform lower bound for the stationary measure $\pi^{(k)}$. In any block where G_{B_k} is connected, each node u must have $\deg_t(u) \geq 1$ for at least one $t \in B_k$. This gives a lower bound on its average lazy degree:

$$\begin{aligned} s^{(k)}(u) &= \frac{1}{\alpha} \sum_{t \in B_k} (1 + \deg_t(u)) \\ &= 1 + \frac{1}{\alpha} \sum_{t \in B_k} \deg_t(u) \geq 1 + \frac{1}{\alpha}. \end{aligned}$$

The normalization constant Z_k can be bounded above:

$$\begin{aligned} Z_k &= \sum_{u \in A} s^{(k)}(u) \\ &= \frac{1}{\alpha} \sum_{u \in A} \sum_{t \in B_k} (1 + \deg_t(u)) \\ &= \frac{1}{\alpha} \sum_{t \in B_k} (n + 2|E_t|) \leq \frac{1}{\alpha} \sum_{t \in B_k} n^2 = n^2. \end{aligned}$$

Combining these bounds yields a uniform lower bound on the stationary probability for any node u:

$$\pi^{(k)}(u) = \frac{s^{(k)}(u)}{Z_k} \geq \frac{1 + 1/\alpha}{n^2} > \frac{1}{n^2} \triangleq \pi_{\min}. \tag{14}$$

Moreover, from definition 1 we know, that, there will be atleast a fraction ν of the blocks B_k, which will have the union graph $G_{[t_k, t_k+\alpha-1]} \triangleq \bigcup_{t \in B_k} G_t$ connected (we call them 'good' blocks), the lazy walk kernel M_k^{union} on this graph is irreducible and reversible with spectral gap:

$$\text{gap}(M_k^{\text{union}}) \geq \gamma.$$

By the Triangle inequality and the convexity of the norm, we have:

$$\|\bar{M}^{(k)} - M_k^{\text{union}}\|_2 \leq \left\| \frac{1}{\alpha} \sum_{t \in B_k} M_t - M_k^{\text{union}} \right\|_2 \tag{15}$$

$$= \left\| \frac{1}{\alpha} \sum_{t \in B_k} \left(M_t - M_k^{\text{union}} \right) \right\|_2 \tag{16}$$

$$\leq \frac{1}{\alpha} \sum_{t \in B_k} \left\| M_t - M_k^{\text{union}} \right\|_2. \tag{17}$$

To bound each $\|M_t - M_k^{\text{union}}\|_2$, observe that since $\|M_t - M_{t-1}\|_2 \leq \beta$ and each M_t differs from its neighbors by at most β, the maximum pairwise deviation across the block is bounded:

$$\|M_t - M_k^{\text{union}}\|_2 \leq \alpha\beta, \quad \text{for all } t \in B_k.$$

Hence,

$$\|\bar{M}^{(k)} - M_k^{\text{union}}\|_2 \leq \alpha\beta. \tag{18}$$

Using Eq. (18) and Lemma 7, we conclude:

$$\begin{aligned} \text{gap}(\bar{M}^{(k)}) &\geq \text{gap}(M_k^{\text{union}}) - \|\bar{M}^{(k)} - M_k^{\text{union}}\|_2 \\ &\geq \gamma - \alpha\beta \triangleq \bar{\gamma}. \end{aligned}$$

Next, we define two distributions:

- θ_t: the marginal distribution of X_t under the actual walk with kernels $\{M_t\}$,
- $\tilde{\theta}_t$: the marginal distribution of X_t under a hypothetical walk that starts at X_{t_k} and uses $\bar{M}^{(k)}$ at every step.

We compare both to $\pi^{(k)}$ using the triangle inequality:

$$\|\theta_t - \pi^{(k)}\|_{\text{TV}} \leq \|\theta_t - \tilde{\theta}_t\|_{\text{TV}} + \|\tilde{\theta}_t - \pi^{(k)}\|_{\text{TV}}.$$

(a) Bounding the mixing term. Since $\bar{M}^{(k)}$ is symmetric, stochastic, and has spectral gap at least $\bar{\gamma}$, standard exponential mixing for reversible Markov chains implies (see [15]):

$$\|\tilde{\theta}_t - \pi^{(k)}\|_{\mathrm{TV}} \leq \frac{1}{2\sqrt{\pi_{\min}}} e^{-\bar{\gamma}(t-t_k)} \leq \frac{n}{2} e^{-\bar{\gamma}(t-t_k)} \tag{19}$$

(b) Bounding the perturbation term. We now bound the distance between θ_t and $\tilde{\theta}_t$. The walk θ_t evolves under $M_{t-1}, M_{t-2}, \ldots$, while $\tilde{\theta}_t$ evolves under repeated application of $\bar{M}^{(k)}$. We proceed inductively. Suppose for time s:

$$\theta_{s+1} = \theta_s M_{s+1}, \quad \tilde{\theta}_{s+1} = \tilde{\theta}_s \bar{M}^{(k)}.$$

Then:

$$\begin{aligned}
\|\theta_{s+1} - \tilde{\theta}_{s+1}\|_{\mathrm{TV}} &= \|\theta_s M_{s+1} - \tilde{\theta}_s \bar{M}^{(k)}\|_{\mathrm{TV}} \\
&\leq \|\theta_s M_{s+1} - \theta_s \bar{M}^{(k)}\|_{\mathrm{TV}} \\
&\quad + \|\theta_s \bar{M}^{(k)} - \tilde{\theta}_s \bar{M}^{(k)}\|_{\mathrm{TV}} \\
&\leq \|M_{s+1} - \bar{M}^{(k)}\|_1 + \|\theta_s - \tilde{\theta}_s\|_{\mathrm{TV}}.
\end{aligned}$$

where we used:

$$\begin{aligned}
&\|\theta P - \theta Q\|_{\mathrm{TV}} \leq \|P - Q\|_1, \\
\text{and} \quad &\|\theta P - \tilde{\theta} P\|_{\mathrm{TV}} \leq \|\theta - \tilde{\theta}\|_{\mathrm{TV}}.
\end{aligned}$$

We initialize the coupling at time t_k with $\theta_{t_k} = \tilde{\theta}_{t_k}$ (same starting node), so:

$$\|\theta_{t_k} - \tilde{\theta}_{t_k}\|_{\mathrm{TV}} = 0.$$

Unrolling the recursion gives:

$$\|\theta_t - \tilde{\theta}_t\|_{\mathrm{TV}} \leq \sum_{s=t_k}^{t-1} \|M_{s+1} - \bar{M}^{(k)}\|_1.$$

To bound $\|M_s - \bar{M}^{(k)}\|_2$, we note that each M_t is at most β away from its predecessor in spectral norm. Therefore, for any $s \in B_k$, the maximum pairwise deviation satisfies:

$$\|M_s - M_t\|_2 \leq |s - t| \cdot \beta, \quad \text{for all } t \in B_k.$$

Hence,

$$\|M_s - \bar{M}^{(k)}\|_2 \leq \frac{1}{\alpha} \sum_{t \in B_k} \|M_s - M_t\|_2 \leq \alpha\beta,$$

and since $\|A\|_1 \leq n \cdot \|A\|_2$ for symmetric matrices, we obtain:

$$\|M_s - \bar{M}^{(k)}\|_1 \leq n \cdot \alpha\beta.$$

Hence,

$$\|\theta_t - \tilde{\theta}_t\|_{\mathrm{TV}} \leq (t - t_k) \cdot n\alpha\beta. \tag{20}$$

(c) Total deviation from stationarity. We now bound the total deviation between the actual distribution θ_t of the agent's location at time t and the stationary distribution $\pi^{(k)}$ of the block-averaged kernel $\bar{M}^{(k)}$. Using the triangle inequality, we write:

$$\|\theta_t - \pi^{(k)}\|_{\mathrm{TV}} \leq \|\theta_t - \tilde{\theta}_t\|_{\mathrm{TV}} + \|\tilde{\theta}_t - \pi^{(k)}\|_{\mathrm{TV}}.$$

where $\tilde{\theta}_t$ is the distribution at time t under a hypothetical homogeneous walk using $\bar{M}^{(k)}$ throughout the block. From parts (a) and (b), we have the mixing error (Eq. 19) and the perturbation error (Eq. 20), respectively.

To obtain a total deviation at most ε, we choose a time threshold $\tau_{\mathrm{mix}} \in \mathbb{N}$ such that both terms are individually at most $\epsilon/2$. That is, we define $\tau_{\mathrm{mix}} := \tau_{\mathrm{mix}}(\epsilon)$ to be the smallest integer τ such that:

$$e^{-\bar{\gamma}\tau} \leq \frac{\epsilon}{2} \quad \text{and} \quad \tau \cdot \|M_s - \bar{M}^{(k)}\|_1 \leq \frac{\epsilon}{2}.$$

Note that this is not an additional assumption: both conditions follow from the $(\alpha, \beta, \gamma, \nu)$ conditions in Definition 1. In particular:

- The spectral gap bound $\bar{\gamma} = \gamma - \alpha\beta$ implies exponential mixing of the lazy walk on $\bar{M}^{(k)}$;
- The kernel drift bound $\|M_t - M_{t-1}\|_2 \leq \beta$ implies $\|M_s - \bar{M}^{(k)}\|_2 \leq \alpha\beta$ for $s \in B_k$, and hence
$$\|M_s - \bar{M}^{(k)}\|_1 \leq n \cdot \|M_s - \bar{M}^{(k)}\|_2 \leq n\alpha\beta.$$

Thus, the second condition is satisfied if:

$$\tau \cdot n\alpha\beta \leq \frac{\epsilon}{2}.$$

Putting this together, for any $t \in B_k$ with $t - t_k \geq \tau_{\text{mix}}$, we have:

$$\|\tilde{\theta}_t - \pi^{(k)}\|_{\text{TV}} \leq \frac{\epsilon}{2} \quad \text{and} \quad \|\theta_t - \tilde{\theta}_t\|_{\text{TV}} \leq \frac{\epsilon}{2},$$

which implies:

$$\|\theta_t - \pi^{(k)}\|_{\text{TV}} \leq \epsilon. \tag{21}$$

We emphasize that the choice of $\epsilon/2$ for each component is without loss of generality: any pair of thresholds summing to ϵ would suffice, but this symmetric choice simplifies the derivation and ensures clean constants in subsequent regret bounds.

Step 2: Lower bounding expected visits. From Step 1(c), we have shown that for all $t \in B_k$ such that $t - t_k \geq \tau_{\text{mix}}$, the marginal distribution ν_t of the agent's location satisfies Eq. (21). In particular, for any node $a \in A$, this implies:

$$\mathbb{P}[X_t = a] = \theta_t(a) \geq \pi^{(k)}(a) - \epsilon.$$

From 1(a) and Eq. (14) we know the lowest probability value of the stationary probability distribution, $\pi^{(k)}(a) \geq \pi_{\text{min}} = 1/n^2$ for all $a \in A$, it follows that:

$$\mathbb{P}[X_t = a] \geq \frac{1}{n^2} - \epsilon.$$

Let $\alpha' := \alpha - \tau_{\text{mix}}$ denote the number of time steps in block B_k after the initial mixing period. Summing over these well-mixed time steps, the expected number of visits to node a within block B_k satisfies:

$$\mathbb{E}[\varphi_k(a)] = \sum_{\substack{t \in B_k \\ t - t_k \geq \tau_{\text{mix}}}} \mathbb{P}[X_t = a] \geq \alpha' \cdot \left(\frac{1}{n^2} - \epsilon\right) \triangleq \lambda_{\text{exp}}.$$

Step 3: Concentration via Azuma–Hoeffding inequality. We now show that the total number of visits to any node $a \in A$ concentrates around its expectation. While one might attempt to apply standard Chernoff or Hoeffding bounds, these tools assume independence across observations, which does not hold in our setting: the walk $\{X_t\}$ evolves as a time-inhomogeneous Markov chain, with temporal dependencies induced by the varying kernels $\{M_t\}$.

To address this, we exploit the block structure of the exploration phase. Within each block $B_k = \{t_k, t_k + 1, \ldots, t_k + \alpha - 1\}$, the agent's trajectory may exhibit dependencies, but conditioned on the trajectory prefix up to t_k, the randomness within B_k is fully determined. This motivates a block-wise martingale construction.

Let $\varphi_k(a) \triangleq \sum_{t \in B_k} \mathbf{1}(X_t = a)$ denote the number of visits to node a during block B_k, and let $\varphi(a) \triangleq \sum_{k=1}^{K} \varphi_k(a)$ be the total number of visits to a over $K \triangleq T_{\text{exp}}/\alpha$ blocks. Define the filtration $\{\mathcal{F}_k\}_{k=1}^{K}$ by:

$$\mathcal{F}_k := \sigma(X_1, X_2, \ldots, X_{t_k}) = \sigma\left(X_1, \ldots, X_{(k-1)\alpha+1}\right),$$

which encodes the trajectory history up to the start of B_k.

Define the deviation at each block:

$$D_k := \varphi_k(a) - \mathbb{E}[\varphi_k(a) \mid \mathcal{F}_k].$$

Then $\mathbb{E}[D_k \mid \mathcal{F}_k] = 0$ by definition, and the sequence $\{D_k\}_{k=1}^K$ forms a martingale difference sequence adapted to the filtration $\{\mathcal{F}_k\}$.

Since each block contains α rounds, we have:

$$0 \le \varphi_k(a) \le \alpha \quad \Rightarrow \quad |D_k| \le \alpha.$$

Let $v(a) \triangleq \mathbb{E}[\varphi(a)] = \sum_{k=1}^K \mathbb{E}[\varphi_k(a)]$ be the expected number of visits to node a. We know from definition 1 that there will be atleast νK 'good' blocks and from Step 2, we know that $\mathbb{E}[\varphi_k(a)] \ge \lambda_{\exp}$ for each block, so:

$$v(a) \ge \nu K \cdot \lambda_{\exp}.$$

We now apply the Azuma–Hoeffding inequality for bounded-difference martingales (see [17]): for any $t > 0$, if $\{D_k\}$ is a martingale difference sequence with $|D_k| \le \alpha$, then

$$\mathbb{P}\left[\sum_{k=1}^K D_k \le -t\right] \le \exp\left(-\frac{t^2}{2K\alpha^2}\right).$$

Apply this with $t := v(a)/2$ to obtain:

$$\begin{aligned}\mathbb{P}\left[\varphi(a) \le \frac{1}{2} \cdot v(a)\right] &= \mathbb{P}\left[\sum_{k=1}^K D_k \le -\frac{v(a)}{2}\right] \\ &\le \exp\left(-\frac{v(a)^2}{8K\alpha^2}\right).\end{aligned}$$

Since $v(a) \ge \nu K \cdot \lambda_{\exp}$, we have:

$$\mathbb{P}\left[\varphi(a) \le \frac{1}{2} \cdot v(a)\right] \le \exp\left(-\frac{\nu^2 K \cdot \lambda_{\exp}^2}{8\alpha^2}\right).$$

To ensure this failure probability is at most δ/n (for a union bound over all n nodes), it suffices to choose $T_{\exp}$ such that:

$$\exp\left(-\frac{\nu^2 K \cdot \lambda_{\exp}^2}{8\alpha^2}\right) \le \frac{\delta}{n} \quad \Longrightarrow \quad K \ge \frac{8\alpha^2}{\nu^2 \lambda_{\exp}^2} \cdot \log\left(\frac{n}{\delta}\right).$$

Since $K = T_{\exp}/\alpha$, this is equivalent to:

$$T_{\exp} \ge \frac{8\alpha^3}{\nu^2 \lambda_{\exp}^2} \cdot \log\left(\frac{n}{\delta}\right),$$

which matches the condition in equation (12). Therefore, with probability at least $1 - \delta/n$,

$$\varphi(a) \ge \frac{1}{2} \cdot v(a) \ge \frac{\lambda_{\exp}}{2} \cdot \frac{\nu T_{\exp}}{\alpha}.$$

To ensure this bound holds for all nodes $a \in A$ simultaneously, we apply a union bound over the n arms. This gives:

$$\mathbb{P}\left[\forall a \in A : \varphi(a) \ge \frac{\lambda_{\exp}}{2} \cdot \frac{\nu T_{\exp}}{\alpha}\right] \ge 1 - \delta.$$

□

Lemma 6 (Navigation Regret). *Let the graph sequence $\{G_t\}_{t=T_{\exp}+1}^T$ be an $(\alpha, \beta, \gamma, \nu)$-admissible sequence (Definition 1) over $n \triangleq |A|$ arms. Assume that the empirically optimal arm $\hat{a}^*$ has been correctly identified as the true optimal arm a^*, and that the agent performs a lazy random walk, committing to a^* upon arrival. The expected regret incurred during this navigation phase is bounded by:*

$$\mathbb{E}[R_{\text{nav}}] \le \Delta_{\max} \cdot \mathbb{E}[\tau_{\text{hit}}] = O\left(\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right),$$

where τ_{hit} is the hitting time to a^*, τ_{mix} is the drift-aware mixing time, and $\Delta_{\max} \triangleq \max_{a \in A} \Delta(a)$.

Proof. The proof proceeds by first determining the probability of reaching a^* within a single "good" block of time, then using this to calculate the expected number of total blocks required for a successful hit, and finally translating this into the total expected hitting time and regret.

Step 1: Per-Block Success Probability. We consider a single "good" block B_k of length α, where the union graph G_{B_k} is connected. The block-averaged kernel $\bar{M}^{(k)}$ is reversible with respect to a stationary measure $\pi^{(k)}$, which was shown, in Eq. (14), to satisfy the lower bound $\pi^{(k)}(a^*) \geq 1/n^2$ for any arm a^*.

After a mixing period of τ_{mix} steps within the block, the agent's location distribution θ_t becomes sufficiently close to $\pi^{(k)}$ such that for the remaining $\alpha' \triangleq \alpha - \tau_{\text{mix}}$ steps, we have $\Pr[X_t = a^*] \geq \pi^{(k)}(a^*) - \varepsilon \geq 1/n^2 - \varepsilon$. The probability of hitting a^* in this good block, p_{hit}, is therefore bounded below by:

$$\begin{aligned} p_{\text{hit}} &\geq 1 - \left(1 - (\tfrac{1}{n^2} - \varepsilon)\right)^{\alpha'} \geq 1 - \exp\left(-\alpha'(\tfrac{1}{n^2} - \varepsilon)\right) \\ &\geq \frac{\alpha'}{2}\left(\frac{1}{n^2} - \varepsilon\right), \end{aligned}$$

where the last inequality holds for sufficiently small arguments of $\exp(\cdot)$ by using $1 - e^{-x} \geq x/2$.

Step 2: Expected Number of Blocks to Hit Target. By the admissibility condition, at least a fraction ν of blocks are good. Let J be the number of good blocks the agent encounters until it first hits a^*. Since each good block represents an independent trial with success probability at least p_{hit}, J is stochastically dominated by a Geometric(p_{hit}) random variable, implying $\mathbb{E}[J] \leq 1/p_{\text{hit}}$.

Let B be the total number of blocks (good or bad) until the first hit. The J-th good block must occur at or before block index $\lceil J/\nu \rceil$. Taking expectations, the expected total number of blocks is:

$$\mathbb{E}[B] \leq \mathbb{E}[\lceil J/\nu \rceil] \leq \frac{\mathbb{E}[J]}{\nu} + 1 \leq \frac{1}{\nu p_{\text{hit}}} + 1.$$

Step 3: Expected Hitting Time and Regret. The total expected hitting time is bounded by the expected number of blocks times the block length, α.

$$\mathbb{E}[\tau_{\text{hit}}] \leq \alpha \cdot \mathbb{E}[B] \leq \alpha\left(\frac{1}{\nu p_{\text{hit}}} + 1\right).$$

We select $\varepsilon = \frac{1}{2n^2}$ to simplify the bound on $p_{\text{hit}} \geq \frac{\alpha'}{2}(\frac{1}{2n^2}) = \frac{\alpha'}{4n^2}$. Substituting this in, we get:

$$\mathbb{E}[\tau_{\text{hit}}] \leq \alpha\left(\frac{4n^2}{\nu\alpha'} + 1\right) = \frac{4n^2\alpha}{\nu(\alpha - \tau_{\text{mix}})} + \alpha.$$

The dominant term gives the asymptotic behavior. Since the mixing time is $\tau_{\text{mix}} = O(\frac{\log n}{\gamma - \alpha\beta})$, the expected hitting time is:

$$\mathbb{E}[\tau_{\text{hit}}] = O\left(\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right).$$

The total expected navigation regret is the expected time spent in suboptimal states, where each step incurs a cost of at most $\Delta_{\max}$. Therefore:

$$\mathbb{E}[R_{\text{nav}}] \leq \Delta_{\max} \cdot \mathbb{E}[\tau_{\text{hit}}] = \mathcal{O}\left(\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right).$$

□

Lemma 7 (Eigenvalue Stability under Spectral Norm)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices. Then for all $k \in \{1, \ldots, n\}$,*

$$|\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|_2,$$

where $\lambda_k(\cdot)$ denotes the k-th largest eigenvalue and $\|\cdot\|_2$ denotes the spectral norm (operator norm).

Proof. Let $E := A - B$, so that $A = B + E$. Fix any index $k \in \{1, \dots, n\}$. We invoke the Courant–Fischer min–max principle, which characterizes the k-th largest eigenvalue of any symmetric matrix M as

$$\lambda_k(M) = \min_{\substack{U \subseteq \mathbb{R}^n \\ \dim(U)=k}} \max_{\substack{x \in U \\ \|x\|=1}} x^\top M x.$$

Applying this to A, we obtain for any k-dimensional subspace U,

$$\max_{\substack{x \in U \\ \|x\|=1}} x^\top A x = \max_{\substack{x \in U \\ \|x\|=1}} x^\top (B + E) x = \max_{\substack{x \in U \\ \|x\|=1}} \left(x^\top B x + x^\top E x \right).$$

Using the definition of spectral norm for symmetric matrices,

$$|x^\top E x| \leq \|E\|_2 \quad \text{for all } \|x\| = 1,$$

we get

$$x^\top A x \leq x^\top B x + \|E\|_2, \quad \Rightarrow \quad \max_{x \in U, \|x\|=1} x^\top A x \leq \max_{x \in U, \|x\|=1} x^\top B x + \|E\|_2.$$

Taking the minimum over all k-dimensional subspaces U, we conclude:

$$\lambda_k(A) \leq \lambda_k(B) + \|E\|_2 = \lambda_k(B) + \|A - B\|_2.$$

To obtain the reverse inequality, we reverse the roles of A and B. Writing $B = A - E$, the same argument yields:

$$\lambda_k(B) \leq \lambda_k(A) + \|A - B\|_2.$$

Combining both directions, we arrive at the desired bound. □

Proof of Theorem 1 and Corollary 1. The proof proceeds by first analyzing the regret on a high-probability "clean" event, $\mathcal{E}$, where both the exploration coverage is sufficient and the subsequent reward estimations are accurate. We then bound the expected regret by considering the outcomes both on and off this event.

1. Defining the High-Probability Event. Let $\mathcal{E}_{\text{visit}}$ be the event that the uniform visitation guarantee from Lemma 1 holds, and let $\mathcal{E}_{\text{conc}}$ be the event that the empirical means for all arms are well-concentrated around their true means. Our overall "clean" event, on which our regret analysis will be conditioned, is the intersection $\mathcal{E} = \mathcal{E}_{\text{visit}} \cap \mathcal{E}_{\text{conc}}$. We now derive an exploration time T_{exp} that guarantees this event holds with probability at least $1 - \delta$.

1.1. Sample Complexity for Concentration. For the event $\mathcal{E}_{\text{conc}}$ to hold, we require the empirical mean $\hat{\mu}(a)$ of each arm a to be concentrated such that $|\hat{\mu}(a) - \mu(a)| < \Delta(a)/2$. By Hoeffding's inequality, for an arm a visited $\varphi(a)$ times, the probability of a large deviation is:

$$\begin{aligned} \mathbb{P}\left(|\hat{\mu}(a) - \mu(a)| \geq \frac{\Delta(a)}{2} \right) &\leq 2 \exp\left(-2\varphi(a) \left(\frac{\Delta(a)}{2} \right)^2 \right) \\ &= 2 \exp\left(-\frac{\varphi(a)\Delta(a)^2}{2} \right). \end{aligned}$$

To ensure this failure probability is at most $\delta/(2n)$ (Sourav: Just for consistency, make this $\delta/2nT$, as this is only looser and hence correct.) for each arm (in preparation for a union bound), we require:

$$\begin{aligned}
2\exp\left(-\frac{\varphi(a)\Delta(a)^2}{2}\right) &\leq \frac{\delta}{2n} \\
\exp\left(-\frac{\varphi(a)\Delta(a)^2}{2}\right) &\leq \frac{\delta}{4n} \\
-\frac{\varphi(a)\Delta(a)^2}{2} &\leq \log\left(\frac{\delta}{4n}\right) \\
\frac{\varphi(a)\Delta(a)^2}{2} &\geq \log\left(\frac{4n}{\delta}\right) \\
\varphi(a) &\geq \frac{2\log(4n/\delta)}{\Delta(a)^2}.
\end{aligned}$$

This inequality defines the minimum number of samples required for each arm a to guarantee concentration with the desired confidence.

1.2. Deriving the Required Exploration Time. To satisfy the sample complexity for all arms simultaneously, we must ensure even the arm with the smallest gap, $\Delta_{\min}$, is visited enough times. From Lemma 1, we have a lower bound on the number of visits $\varphi(a)$ as a function of $T_{\exp}$. To meet the requirement derived above, we must choose $T_{\exp}$ such that the guaranteed number of visits is greater than or equal to the required number:

$$\frac{\nu\,(\alpha - \tau_{\text{mix}})\,\left(\frac{1}{n^2} - \epsilon\right)}{2\,\alpha}\,T_{\exp} \geq \frac{2\log(4n/\delta)}{\Delta_{\min}^2}.$$

Solving for $T_{\exp}$ gives us the minimum required length of the exploration phase:

$$T_{\exp} \geq \frac{4\alpha\log(4n/\delta)}{\nu(\alpha - \tau_{\text{mix}})(\frac{1}{n^2} - \epsilon)\Delta_{\min}^2}. \tag{22}$$

1.3. Bounding the Probability of the Clean Event. We now formally show that with this choice of $T_{\exp}$, the clean event $\mathcal{E}$ holds with probability at least $1 - \delta$. We set the failure probability in Lemma 1 to be $\delta/2$.

- By running the exploration phase for a duration $T_{\exp}$ that satisfies Eq. (22) with δ replaced by $\delta/2$, Lemma 1 guarantees that the visitation event $\mathcal{E}_{\text{visit}}$ holds with probability at least $1 - \delta/2$.
- Now, we consider the probability of the concentration event $\mathcal{E}_{\text{conc}}$ failing, conditioned on $\mathcal{E}_{\text{visit}}$ being true. On the event $\mathcal{E}_{\text{visit}}$, we are guaranteed that every arm a has been visited at least $\varphi(a) \geq \frac{2\log(4n/\delta)}{\Delta(a)^2}$ times. The probability of any single arm failing to concentrate is therefore at most $\delta/(2n)$. Using a union bound over all n arms, we get $\mathbb{P}(\neg\mathcal{E}_{\text{conc}} \mid \mathcal{E}_{\text{visit}})$ as:

$$\mathbb{P}\left(\exists a : |\hat{\mu}(a) - \mu(a)| \geq \frac{\Delta(a)}{2} \;\middle|\; \mathcal{E}_{\text{visit}}\right) \leq \sum_{a\in A}\frac{\delta}{2n} = \frac{\delta}{2}.$$

The total probability of the clean event $\mathcal{E}$ is therefore:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P}(\mathcal{E}_{\text{conc}} \mid \mathcal{E}_{\text{visit}})\mathbb{P}(\mathcal{E}_{\text{visit}}) \\
&= (1 - \mathbb{P}(\neg\mathcal{E}_{\text{conc}} \mid \mathcal{E}_{\text{visit}}))\,\mathbb{P}(\mathcal{E}_{\text{visit}}) \\
&\geq \left(1 - \frac{\delta}{2}\right)\left(1 - \frac{\delta}{2}\right) > 1 - \delta.
\end{aligned}$$

This confirms that our choice of $T_{\exp}$ is sufficient to ensure the clean event holds with high probability.

2. Bounding Regret on the Clean Event $\mathcal{E}$. Conditioned on the clean event $\mathcal{E}$, the total regret $R(T)$ is the sum of the exploration regret, R_{exp}, and the navigation regret, R_{nav}. On this event, the empirical means are well-concentrated, so the correct best arm is identified, i.e., $\hat{a}^* = a^*$.

The *exploration regret* for the LEX algorithm, which employs a fixed exploration schedule, is bounded by the duration of the exploration phase itself. During these T_{exp} steps, the algorithm is not attempting to exploit the best arm, so every step incurs regret of at most 1.

$$R_{\text{exp}} \leq T_{\text{exp}} = O\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2}\right). \tag{23}$$

The *navigation regret*, R_{nav}, is the regret incurred after T_{exp} while the agent travels to the correctly identified a^*. By Lemma 2, the expected number of steps for this phase is bounded. This gives the navigation regret:

$$R_{\text{nav}} = O\left(\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right). \tag{24}$$

Combining the two components, the total regret on the clean event $\mathcal{E}$ is:

$$R(T) \mid \mathcal{E} \leq O\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right). \tag{25}$$

Since $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$, this proves the high-probability statement of Theorem 1.

3. Deriving the Expected Regret for the Corollary. To find the expected regret, we use the law of total expectation:

$$\mathbb{E}[R(T)] = \mathbb{E}[R(T) \mid \mathcal{E}]\mathbb{P}(\mathcal{E}) + \mathbb{E}[R(T) \mid \neg\mathcal{E}]\mathbb{P}(\neg\mathcal{E}). \tag{26}$$

We bound each term using our results:

- $\mathbb{E}[R(T) \mid \mathcal{E}] \leq O\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right)$.
- $\mathbb{P}(\mathcal{E}) \leq 1$.
- $\mathbb{E}[R(T) \mid \neg\mathcal{E}] \leq T$, as regret per step is at most 1.
- $\mathbb{P}(\neg\mathcal{E}) \leq \delta$.

Substituting these into the equation gives the general expected regret bound:

$$\mathbb{E}[R(T)] \leq O\left(\frac{\alpha n^2 \log(n/\delta)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right) + \delta T. \tag{27}$$

For the corollary, we make the standard analytical choice of $\delta = 1/T^2$. The final term becomes $\delta T = 1/T$, which is negligible. The logarithmic term becomes $\log(n/T^{-2}) = \log(nT^2) = \mathcal{O}(\log(nT))$. Absorbing constants, we arrive at the final expression:

$$\mathbb{E}[R(T)] \leq \mathcal{O}\left(\frac{\alpha n^2 \log(nT)}{\nu \Delta_{\min}^2} + \frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}\right). \tag{28}$$

This completes the proof. □

C Appendix: Analysis for CB-LEX

Theorem (CB-LEX: High-probability regret (Restated))**.** *Assume the graph sequence is an $(\alpha, \beta, \gamma, \nu)$-admissible sequence (Definition 1). For any confidence parameter $\delta \in (0, 1)$, the CB-LEX algorithm (with anytime CI stopping) identifies a^* with probability at least $1 - 2\delta$, and the cumulative regret satisfies*

$$R(T) \;\leq\; O\left(\max\left\{\underbrace{\frac{\alpha\, n^2}{\nu(\alpha - \tau_{mix})} \max_{a \neq a^*} \frac{\log\big(2nT/\delta\big)}{\Delta(a)^2}}_{\textit{CI threshold branch}},\; \underbrace{\frac{\alpha^3}{\nu^2(\alpha - \tau_{mix})^2}\, n^4 \log\Big(\tfrac{n}{\delta}\Big)}_{\textit{uniform visitation branch}}\right\} + \underbrace{\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}}_{\textit{navigation}}\right).$$

In particular, this bound requires no prior knowledge of $\Delta_{\min}$; the exploration term depends on the instance via $\max_{a \neq a^*} 1/\Delta(a)^2$.

Corollary (CB-LEX: Expected regret (Restated))**.** *Under the conditions of the theorem, setting* $\delta = 1/T^2$ *yields*

$$\mathbb{E}[R(T)] \;\le\; O\left(\max\left\{\frac{\alpha\, n^2}{\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log n\right\}\;+\;\frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\right).$$

Proof of Theorem 2 and Corollary 2. Fix $\epsilon \in \left(0, \frac{1}{2n^2}\right]$ and $\delta \in (0,1)$. Let $\tau_{\text{mix}} := \tau_{\text{mix}}(\epsilon)$ be as in Lemma 5. We analyze CB-LEX, which performs a lazy random walk during exploration and stops at the (random) time σ when empirical confidence intervals separate the empirically-best arm from all others.

1. Uniform anytime concentration event. For $t \ge 1$ and $a \in A$, let $\varphi_t(a)$ be the number of pulls of a up to (and including) time t. We define the (anytime) confidence width

$$w_t(a) \;:=\; \sqrt{\frac{c\,\log\big(2nT/\delta\big)}{\max\{1,\varphi_t(a)\}}},$$

where $c > 0$ is an absolute constant to be fixed below (e.g., $c = \frac{1}{2}$). Let $\widehat{\mu}_t(a)$ be the empirical mean of arm a computed from the $\varphi_t(a)$ rewards observed by time t (if $\varphi_t(a) = 0$ the event below is vacuous due to the $\max\{\cdot\}$).

Fix any pair (t, a). Condition on the (random) value $K := \varphi_t(a) \in \{0, 1, 2, \dots\}$ and on the *selection history* that determines which K rewards of arm a have been observed by time t. Conditional on $K = k \ge 1$, the k rewards of arm a used in $\widehat{\mu}_t(a)$ are i.i.d. and bounded in $[0, 1]$; hence Hoeffding's inequality gives, for any $\varepsilon > 0$,

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > \varepsilon \;\Big|\; \varphi_t(a) = k\,\Big) \;\le\; 2\,e^{-2k\varepsilon^2}.$$

Choosing $\varepsilon = \sqrt{\frac{c\,\log(2nT/\delta)}{k}}$ yields

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > w_t(a) \;\Big|\; \varphi_t(a) = k\,\Big) \;\le\; 2\,\exp\big(-2c\,\log(2nT/\delta)\big) \;=\; 2\,(2nT/\delta)^{-2c}.$$

Taking expectation over the randomness of $K = \varphi_t(a)$ (law of total probability), we obtain the *unconditional* bound

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > w_t(a)\,\Big) \;\le\; 2\,(2nT/\delta)^{-2c}.$$

Setting $c = \frac{1}{2}$ (or any $c \ge \frac{1}{2}$) gives

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > w_t(a)\,\Big) \;\le\; \frac{\delta}{nT}.$$

Finally, apply a union bound over all nT arm–time pairs (t, a):

$$\Pr\Big(\,\exists\, t \le T,\ \exists\, a \in A:\ |\widehat{\mu}_t(a) - \mu(a)| > w_t(a)\,\Big) \;\le\; \sum_{t=1}^{T}\sum_{a=1}^{n}\frac{\delta}{nT} \;=\; \delta.$$

Therefore the (anytime) clean event

$$\mathcal{E}_{\text{conc}} \;:=\; \Big\{\,\forall t \le T,\ \forall a \in A:\ |\widehat{\mu}_t(a) - \mu(a)| \le w_t(a)\,\Big\}$$

holds with probability at least $1 - \delta$. *Remarks.* (i) This argument relies only on the rewards being i.i.d. in $[0, 1]$ for each arm; the adaptivity of arm selection affects the random sample size $K = \varphi_t(a)$ but not the conditional Hoeffding tail bound, which we handled by conditioning on K and then averaging. (ii) Using $\max\{1, \varphi_t(a)\}$ ensures $w_t(a)$ is defined even when an arm has not yet been pulled.

2. Uniform visitation and the stopping time. We first record the structural coverage implied by admissibility and then convert the CB-LEX stopping rule into explicit sample thresholds. Throughout, $\varphi_t(a)$ denotes the number of pulls of arm a by time t.

Step 2.1 (Uniform visitation lower bound). Define

$$c_{\min} \;:=\; \frac{\nu\,(\alpha-\tau_{\text{mix}})\Big(\frac{1}{n^2}-\epsilon\Big)}{2\alpha}\,.$$

Lemma 5 guarantees that if

$$t \;\geq\; T_{\text{visit}}(\delta) \;\triangleq\; \frac{8\,\alpha^3}{\nu^2\,(\alpha-\tau_{\text{mix}})^2\,\Big(\frac{1}{n^2}-\epsilon\Big)^2}\,\log\Big(\frac{n}{\delta}\Big), \tag{29}$$

then, with probability at least $1-\delta$, we have simultaneously for all $a \in A$,

$$\varphi_t(a) \;\geq\; c_{\min}\,t. \tag{30}$$

We denote (30) at time t by the event $\mathcal{E}_{\text{visit}}(t)$.

Step 2.2 (Stopping rule ⇒ a sufficient gap condition). CB-LEX stops at the first time σ such that

$$\widehat{\mu}_\sigma(a)+w_\sigma(a) \;<\; \widehat{\mu}_\sigma(a^*)-w_\sigma(a^*) \qquad \text{for all } a\neq a^*. \tag{31}$$

On the concentration event $\mathcal{E}_{\text{conc}}$ we know, for each arm x,

$$|\widehat{\mu}_\sigma(x)-\mu(x)| \;\leq\; w_\sigma(x) \quad\Longrightarrow\quad \widehat{\mu}_\sigma(x) \;\leq\; \mu(x)+w_\sigma(x) \text{ and } \widehat{\mu}_\sigma(x) \;\geq\; \mu(x)-w_\sigma(x).$$

Therefore

$$\widehat{\mu}_\sigma(a)+w_\sigma(a) \;\leq\; \mu(a)+2w_\sigma(a), \qquad \widehat{\mu}_\sigma(a^*)-w_\sigma(a^*) \;\geq\; \mu(a^*)-2w_\sigma(a^*).$$

A *sufficient* condition for (31) is

$$\mu(a)+2w_\sigma(a) \;<\; \mu(a^*)-2w_\sigma(a^*) \quad\Longleftrightarrow\quad \Delta(a) \;>\; 2\big(w_\sigma(a)+w_\sigma(a^*)\big) \qquad (\forall a\neq a^*). \tag{32}$$

It further suffices to enforce the two separate inequalities

$$w_\sigma(a) \;\leq\; \frac{\Delta(a)}{4} \quad\text{and}\quad w_\sigma(a^*) \;\leq\; \frac{\Delta(a)}{4} \qquad (\forall a\neq a^*), \tag{33}$$

because then $2(w_\sigma(a)+w_\sigma(a^*)) \leq 2(\frac{\Delta(a)}{4}+\frac{\Delta(a)}{4}) = \Delta(a)$, which is enough for (32) up to a measure-zero tie.[4]

Step 2.3 (Widths ⇒ per-arm sample thresholds). Recall

$$w_t(x) \;=\; \sqrt{\frac{c\,\log(2nT/\delta)}{\max\{1,\varphi_t(x)\}}}\,.$$

Fix $a\neq a^*$. The first inequality in (33) is equivalent to

$$\sqrt{\frac{c\,\log(2nT/\delta)}{\max\{1,\varphi_\sigma(a)\}}} \;\leq\; \frac{\Delta(a)}{4} \;\Longleftrightarrow\; \frac{c\,\log(2nT/\delta)}{\max\{1,\varphi_\sigma(a)\}} \;\leq\; \frac{\Delta(a)^2}{16} \;\Longleftrightarrow\; \max\{1,\varphi_\sigma(a)\} \;\geq\; \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}.$$

Since the right-hand side is at least 1 for all relevant horizons, this simplifies to

$$\varphi_\sigma(a) \;\geq\; \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}. \tag{34}$$

The second inequality in (33) yields, analogously,

$$\sqrt{\frac{c\,\log(2nT/\delta)}{\max\{1,\varphi_\sigma(a^*)\}}} \;\leq\; \frac{\Delta(a)}{4} \;\Longleftrightarrow\; \varphi_\sigma(a^*) \;\geq\; \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}.$$

Because this must hold *for all* $a\neq a^*$, we need

$$\varphi_\sigma(a^*) \;\geq\; \max_{a\neq a^*}\,\frac{16c\,\log(2nT/\delta)}{\Delta(a)^2} \;=\; \frac{16c\,\log(2nT/\delta)}{\Delta_{\min}^2}. \tag{35}$$

[4] If one insists on strict separation, replace each bound in (33) by $w_\sigma(\cdot) \leq \Delta(a)/4-\eta$ for arbitrarily small $\eta>0$; this only changes absolute constants.

Combining (34) and (35), a *sufficient* (uniform) requirement is

$$\varphi_\sigma(x) \;\geq\; \frac{16c\,\log(2nT/\delta)}{\Delta_{\min}^2} \qquad \text{for all } x \in A, \tag{36}$$

which implies (34) for each $a \neq a^*$ and (35) for a^*.

Step 2.4 (From visitation to stopping time). On $\mathcal{E}_{\text{visit}}(t)$ we have $\varphi_t(x) \geq c_{\min} t$ for every arm x. Thus (36) holds at time t provided

$$c_{\min}\, t \;\geq\; \frac{16c\,\log(2nT/\delta)}{\Delta_{\min}^2}.$$

Solving for t gives the deterministic requirement

$$t \;\geq\; \frac{16c}{c_{\min}} \cdot \frac{\log(2nT/\delta)}{\Delta_{\min}^2}.$$

Combining this with the prerequisite $t \geq T_{\text{visit}}(\delta)$ from (29), we define

$$t_\star \;:=\; \max\left\{ \frac{16c}{c_{\min}} \cdot \frac{\log(2nT/\delta)}{\Delta_{\min}^2} \,,\; T_{\text{visit}}(\delta) \right\}. \tag{37}$$

Step 2.5 (Conclusion and probability). By construction, if $t \geq t_\star$ then (i) $\mathcal{E}_{\text{visit}}(t)$ holds with probability at least $1-\delta$ (uniform visitation), and (ii) (36) holds (hence the stopping rule (31) is satisfied on $\mathcal{E}_{\text{conc}}$). Therefore, on the event $\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}(t_\star)$ we must have $\sigma \leq t_\star$. Since $\Pr(\mathcal{E}_{\text{conc}}) \geq 1-\delta$ and $\Pr(\mathcal{E}_{\text{visit}}(t_\star)) \geq 1-\delta$, a union bound gives

$$\Pr\left(\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}(t_\star)\right) \;\geq\; 1 - 2\delta.$$

Optional instance-dependent variant. If one prefers to avoid $\Delta_{\min}$, keep (34) for each $a \neq a^*$ and replace (35) by $\varphi_\sigma(a^*) \;\geq\; \max_{a \neq a^*} \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}$. Then (37) becomes $t_\star = \max\left\{ \frac{16c}{c_{\min}} \max_{a \neq a^*} \frac{\log(2nT/\delta)}{\Delta(a)^2},\; T_{\text{visit}}(\delta) \right\}$.

3. Exploration regret (high probability). By definition of the exploration phase, the regret accumulated up to the stopping time σ is

$$R_{\text{exp}} \;:=\; \sum_{t=1}^{\sigma} \big(\mu(a^*) - \mu(X_t)\big) \;=\; \sum_{a \neq a^*} \Big(\#\text{pulls of } a \text{ up to } \sigma\Big) \cdot \Delta(a) \;=\; \sum_{a \neq a^*} \varphi_\sigma(a)\,\Delta(a).$$

Since $\Delta(a) \leq \Delta_{\max}$ for all a, we can upper bound

$$R_{\text{exp}} \;\leq\; \Delta_{\max} \sum_{a \neq a^*} \varphi_\sigma(a). \tag{38}$$

Next, notice that the total number of pulls by time σ decomposes as

$$\sum_{a \in A} \varphi_\sigma(a) \;=\; \sigma \qquad \Longrightarrow \qquad \sum_{a \neq a^*} \varphi_\sigma(a) \;=\; \sigma - \varphi_\sigma(a^*).$$

Plugging this identity into (38) gives

$$R_{\text{exp}} \;\leq\; \Delta_{\max}\,\big(\sigma - \varphi_\sigma(a^*)\big). \tag{39}$$

On the uniform visitation event $\mathcal{E}_{\text{visit}}(t)$, instantiated at $t = \sigma$, we have the one-sided lower bound

$$\varphi_\sigma(x) \;\geq\; c_{\min}\,\sigma \qquad \text{for all } x \in A,$$

hence in particular $\varphi_\sigma(a^*) \geq c_{\min}\,\sigma$. Substituting this into (39) yields

$$R_{\text{exp}} \;\leq\; \Delta_{\max}\,\big(\sigma - c_{\min}\sigma\big) \;=\; (1 - c_{\min})\,\Delta_{\max}\,\sigma.$$

Finally, on $\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}(t_\star)$ we already proved $\sigma \le t_\star$ (Step 2), therefore

$$R_{\mathrm{exp}} \;\le\; (1 - c_{\min})\, \Delta_{\max}\, t_\star. \tag{40}$$

Remarks. (i) This bound is *tighter* than the crude $R_{\mathrm{exp}} \le \Delta_{\max}\, \sigma$ because it subtracts the zero-regret mass at a^* guaranteed by uniform visitation. (ii) A gap-weighted improvement (replacing $\Delta_{\max}$ by a weighted average of gaps) would require a *two-sided* visitation control; here we keep to the stated one-sided lemma.

4. Navigation regret. After stopping, CB-LEX executes a lazy navigation toward a^* and commits upon hitting it. Let τ_{hit} be the (random) hitting time of a^* for this phase. Each navigation step incurs instantaneous regret at most $\Delta_{\max}$, hence

$$\mathbb{E}[R_{\mathrm{nav}}] \;\le\; \Delta_{\max}\, \mathbb{E}[\tau_{\mathrm{hit}}].$$

Under the $(\alpha, \beta, \gamma, \nu)$ admissibility conditions, Lemma 6 gives the following explicit bound on the expected hitting time:

$$\mathbb{E}[R_{\mathrm{nav}}] \;\le\; \Delta_{\max} \left(\frac{4n^2 \alpha}{\nu(\alpha - \tau_{\mathrm{mix}})} + \alpha \right). \tag{41}$$

(Here the structural constants arise from the admissibility parameters and the drift-aware mixing time; see Lemma 6.)

5. High-probability and expected regret bounds. Combining (40) and (41), we obtain, on the event $\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}(t_\star)$,

$$R(T) \;=\; R_{\mathrm{exp}} + R_{\mathrm{nav}} \;\le\; (1 - c_{\min})\, \Delta_{\max}\, t_\star \;+\; \Delta_{\max} \left(\frac{4n^2 \alpha}{\nu(\alpha - \tau_{\mathrm{mix}})} + \alpha \right).$$

By Step 1 and Step 2, $\Pr(\mathcal{E}_{\mathrm{conc}}) \ge 1 - \delta$ and $\Pr(\mathcal{E}_{\mathrm{visit}}(t_\star)) \ge 1 - \delta$, so by a union bound

$$\Pr\left(\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}(t_\star) \right) \;\ge\; 1 - 2\delta.$$

This proves the high-probability claim in Theorem 2 with the explicit choices

$$t_\star \;=\; \max\left\{ \underbrace{\frac{16c}{c_{\min}} \cdot \frac{\log(2nT/\delta)}{\Delta_{\min}^2}}_{\text{CI threshold time}},\; \underbrace{\frac{8\,\alpha^3}{\nu^2\,(\alpha - \tau_{\mathrm{mix}})^2\,\left(\frac{1}{n^2} - \epsilon\right)^2}\, \log\!\left(\frac{n}{\delta}\right)}_{\text{uniform visitation time (Lemma 1)}} \right\}, \quad c_{\min} \;=\; \frac{\nu(\alpha - \tau_{\mathrm{mix}})\left(\frac{1}{n^2} - \epsilon\right)}{2\alpha}.$$

For the *expected* regret (Corollary 2), apply the law of total expectation with the indicator of the clean event:

$$\mathbb{E}[R(T)] \;=\; \mathbb{E}[R(T)\, \mathbf{1}\{\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}(t_\star)\}] \;+\; \mathbb{E}[R(T)\, \mathbf{1}\{\neg(\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}(t_\star))\}]\,.$$

The first term is at most the high-probability bound above. For the second, use $R(T) \le T$ and $\Pr\left(\neg(\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}(t_\star))\right) \le 2\delta$ to get

$$\mathbb{E}[R(T)\, \mathbf{1}\{\neg(\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}(t_\star))\}] \;\le\; T \cdot 2\delta.$$

Therefore

$$\mathbb{E}[R(T)] \;\le\; (1 - c_{\min})\, \Delta_{\max}\, t_\star \;+\; \Delta_{\max} \left(\frac{4n^2 \alpha}{\nu(\alpha - \tau_{\mathrm{mix}})} + \alpha \right) \;+\; 2\delta T.$$

Choosing $\delta = 1/T^2$ gives $2\delta T \le 2/T$ and $\log(2nT/\delta) = \log(2nT^3) = \Theta(\log(nT))$, completing the corollary.

□

D Appendix: Analysis for RA-LEX

Theorem (RA-LEX: High-probability regret (Restated))**.** *Assume the graph sequence is an $(\alpha, \beta, \gamma, \nu)$-admissible sequence (Definition 1). For any confidence parameter $\delta \in (0, 1)$ and bias $\lambda > 0$, the RA-LEX algorithm identifies the optimal arm a^* with probability at least $1 - 2\delta$, and the cumulative regret satisfies*

$$R(T) \;\le\; O\Bigg(\max\Bigg\{\frac{\alpha n^2}{\kappa\,\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT/\delta)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log\Big(\tfrac{n}{\delta}\Big)\Bigg\}\;+\;\frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\Bigg),$$

where $\kappa = \exp\{-\lambda(1 + O(\sqrt{\log(nT/\delta)}))\}$ is the minorization factor induced by the softmax bias.

Corollary (RA-LEX: Expected regret (Restated))**.** *Under the conditions of the theorem, setting $\delta = 1/T^2$ yields*

$$\mathbb{E}[R(T)] \;\le\; O\Bigg(\max\Bigg\{\frac{\alpha n^2}{\kappa\,\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log n\Bigg\}\;+\;\frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\Bigg).$$

Proof of Theorem 3 and Corollary 3. **Step 0 (Anytime clean event).** For each time $t \in \{1, \dots, T\}$ and arm $a \in A$, let $\varphi_t(a)$ denote the (random) number of pulls of a up to time t (inclusive), and define the anytime confidence width

$$w_t(a) \;:=\; \sqrt{\frac{c\,\log\big(2nT/\delta\big)}{\max\{1, \varphi_t(a)\}}}\,,$$

where $c > 0$ is a fixed absolute constant (e.g. $c = \frac{1}{2}$ suffices). Let $\widehat{\mu}_t(a)$ be the empirical mean of arm a computed from the $\varphi_t(a)$ observed rewards by time t (if $\varphi_t(a) = 0$ this constraint is vacuous due to the $\max\{\cdot\}$).

Fix any pair (t, a). Condition on the random value $K := \varphi_t(a) \in \{0, 1, 2, \dots\}$ and on the selection history that determines *which* K rewards of arm a have been gathered by time t. Conditional on $K = k \ge 1$, the k rewards of arm a entering $\widehat{\mu}_t(a)$ are i.i.d. in $[0, 1]$ with mean $\mu(a)$. By Hoeffding's inequality, for any $\varepsilon > 0$,

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > \varepsilon \;\Big|\; \varphi_t(a) = k\,\Big) \;\le\; 2\,\exp(-2k\varepsilon^2).$$

Choose $\varepsilon = \sqrt{\frac{c\,\log(2nT/\delta)}{k}}$. Then

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > w_t(a) \;\Big|\; \varphi_t(a) = k\,\Big) \;\le\; 2\,\exp\Big(-2k\cdot\frac{c\,\log(2nT/\delta)}{k}\Big) \;=\; 2\,(2nT/\delta)^{-2c}.$$

Take the total expectation over $K = \varphi_t(a)$ to remove the conditioning:

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > w_t(a)\,\Big) \;\le\; 2\,(2nT/\delta)^{-2c}.$$

Setting $c = \frac{1}{2}$ (or any $c \ge \frac{1}{2}$) gives

$$\Pr\Big(\,|\widehat{\mu}_t(a) - \mu(a)| > w_t(a)\,\Big) \;\le\; \frac{\delta}{nT}.$$

Finally, apply a union bound over all nT arm–time pairs (t, a):

$$\Pr\Big(\,\exists t \le T,\ \exists a \in A:\ |\widehat{\mu}_t(a) - \mu(a)| > w_t(a)\,\Big) \;\le\; \sum_{t=1}^{T}\sum_{a=1}^{n}\frac{\delta}{nT} \;=\; \delta.$$

Therefore the (anytime) clean event

$$\mathcal{E}_{\text{conc}} \;:=\; \Big\{\forall t \le T,\ \forall a \in A:\ |\widehat{\mu}_t(a) - \mu(a)| \le w_t(a)\Big\}$$

holds with probability at least $1 - \delta$.

Step 1 (Index range bound on $\mathcal{E}_{\text{conc}}$). Recall the RA-LEX index $\xi_t(a) = \widehat{\mu}_t(a) + w_t(a)$. On $\mathcal{E}_{\text{conc}}$ we have, for every (t, a),

$$-\, w_t(a) \;\le\; \widehat{\mu}_t(a) - \mu(a) \;\le\; w_t(a) \quad\Longrightarrow\quad \mu(a) \;\le\; \widehat{\mu}_t(a) + w_t(a) \;\le\; \mu(a) + 2w_t(a),$$

i.e.

$$\xi_t(a) \;\in\; [\,\mu(a)\,,\; \mu(a) + 2w_t(a)\,] \qquad (\forall t, a). \tag{42}$$

Because rewards lie in $[0, 1]$, we have $0 \le \mu(a) \le 1$ and thus $\mu(b) - \mu(v) \le 1$ for any arms b, v. Moreover,

$$0 \;\le\; w_t(a) \;\le\; \sqrt{c\,\log(2nT/\delta)} \;=:\; W_\star \qquad (\text{since } \max\{1, \varphi_t(a)\} \ge 1).$$

Combining these bounds with (42), for any time t, node u, and any two elements $b, v \in \mathcal{N}_t(u) \cup \{u\}$,

$$\xi_t(b) - \xi_t(v) \;\le\; \big(\mu(b) + 2w_t(b)\big) - \mu(v) \;\le\; 1 + 2W_\star.$$

Therefore

$$\xi_t(b) - \xi_t(v) \;\le\; 1 + 2W_\star, \qquad W_\star \;:=\; \sqrt{c\log(2nT/\delta)}. \tag{43}$$

Step 2 (Kernel minorization matching Algorithm 4). The RA-LEX transition kernel (Algorithm 4) from state u at time t is the softmax over the *closed* neighborhood:

$$\widehat{M}_t(u, v) \;=\; \frac{\exp\{\lambda\, \xi_t(v)\}}{\sum_{b \in \mathcal{N}_t(u) \cup \{u\}} \exp\{\lambda\, \xi_t(b)\}} \qquad \text{for } v \in \mathcal{N}_t(u) \cup \{u\}.$$

Fix u and $v \in \mathcal{N}_t(u) \cup \{u\}$. For each b in the same set, by (43),

$$\xi_t(b) \;\le\; \xi_t(v) \;+\; \big(1 + 2W_\star\big).$$

Exponentiating and summing, we obtain

$$\sum_{b \in \mathcal{N}_t(u) \cup \{u\}} e^{\lambda \xi_t(b)} \;\le\; \sum_{b \in \mathcal{N}_t(u) \cup \{u\}} e^{\lambda(\xi_t(v) + 1 + 2W_\star)} \;=\; (\deg_t(u) + 1)\, e^{\lambda \xi_t(v)}\, e^{\lambda(1 + 2W_\star)}.$$

Plugging this into the definition of $\widehat{M}_t(u, v)$ yields

$$\widehat{M}_t(u, v) \;=\; \frac{e^{\lambda \xi_t(v)}}{\sum_b e^{\lambda \xi_t(b)}} \;\ge\; \frac{e^{\lambda \xi_t(v)}}{(\deg_t(u) + 1)\, e^{\lambda \xi_t(v)}\, e^{\lambda(1 + 2W_\star)}} \;=\; \frac{e^{-\lambda(1 + 2W_\star)}}{\deg_t(u) + 1}.$$

Let L_t denote the *natural-lazy uniform* kernel on G_t, i.e.

$$L_t(u, v) \;=\; \frac{1}{\deg_t(u) + 1} \qquad \big(v \in \mathcal{N}_t(u) \cup \{u\}\big).$$

Define

$$\kappa \;:=\; e^{-\lambda(1 + 2W_\star)} \;\in\; (0, 1].$$

Then the inequality above is exactly

$$\widehat{M}_t(u, v) \;\ge\; \kappa\, L_t(u, v) \qquad \text{for all } v \in \mathcal{N}_t(u) \cup \{u\}.$$

Since this holds for every u and t, we have the entrywise minorization

$$M_t \;\ge\; \kappa\, L_t \quad \text{entrywise for all } t, \quad \kappa := e^{-\lambda(1 + 2W_\star)} \in (0, 1] \tag{44}$$

Step 3 (Thinning coupling $\Rightarrow$ uniform visitation for RA-LEX). We now convert the kernel minorization

$$\widehat{M}_t \;\ge\; \kappa\, L_t \qquad (\text{entrywise for all } t) \tag{44}$$

into a lower bound on the visit counts of RA-LEX via an explicit coupling.

Step 3.1 (Constructing the thinning coupling). Fix a time t and a state $u \in A$. We construct a joint transition for the pair

$$\big(X_t^L,\, X_t^{\text{RA}}\big) \;\in A \times A$$

given the current state $X_{t-1}^L = X_{t-1}^{\text{RA}} = u$ as follows:

1. Draw $V \sim L_t(u,\cdot)$ (this is the next state the *uniform natural-lazy* walk would take).
2. Independently draw $B \sim \mathrm{Bernoulli}(\kappa)$.
3. If $B = 1$, set $X_t^L = V$ and $X_t^{\mathrm{RA}} = V$ (the RA-LEX step *matches* the L-step).
4. If $B = 0$, set $X_t^L = V$ and draw X_t^{RA} from the *residual kernel*

$$R_t(u,\cdot) \;:=\; \frac{\widehat{M}_t(u,\cdot) - \kappa\, L_t(u,\cdot)}{1-\kappa}.$$

This is a valid construction because (44) guarantees

$$\widehat{M}_t(u,\cdot) - \kappa\, L_t(u,\cdot) \;\geq\; 0 \quad \text{(entrywise)}, \qquad \sum_{v\in A}\big(\widehat{M}_t(u,v) - \kappa\, L_t(u,v)\big) \;=\; 1-\kappa,$$

so $R_t(u,\cdot)$ is a bona fide probability distribution on A.

Iterating this procedure for $t = 1, 2, \ldots$ with fresh (V, B) at each step (conditionally independent given the realized L-path) yields coupled trajectories $\big(X_s^L\big)_{s\geq 0}$ and $\big(X_s^{\mathrm{RA}}\big)_{s\geq 0}$ with two key properties:

(i) X_s^L has transition kernel L_s by construction.

(ii) X_s^{RA} has transition kernel $\widehat{M}_s$, because at each step its law is the κ-mixture of $L_s(u,\cdot)$ and $R_s(u,\cdot)$:

$$\kappa\, L_s(u,\cdot) \;+\; (1-\kappa)\, R_s(u,\cdot) \;=\; \kappa\, L_s(u,\cdot) \;+\; \widehat{M}_s(u,\cdot) - \kappa\, L_s(u,\cdot) \;=\; \widehat{M}_s(u,\cdot).$$

Step 3.2 (Visit-count domination via binomial thinning). Fix an arm $a \in A$ and a horizon $t \geq 1$. For each time $s \in \{1,\ldots,t\}$, define the indicator

$$Z_s^a \;:=\; \mathbf{1}\{X_s^L = a\} \cdot \mathbf{1}\{B_s = 1\},$$

where B_s is the Bernoulli variable used at time s in the coupling above. Conditioned on the entire L-trajectory, the variables $\mathbf{1}\{X_s^L = a\}$ are deterministic and $\{B_s\}_{s=1}^t$ are i.i.d. $\mathrm{Bernoulli}(\kappa)$, independent of the L-path. Hence

$$\sum_{s=1}^{t} Z_s^a \;\sim\; \mathrm{Binomial}\big(\,\varphi_t^L(a),\ \kappa\,\big) \quad \text{given the } L\text{-trajectory},$$

because exactly $\varphi_t^L(a)$ of the t time-indices contribute trials with success probability κ.

By construction of the coupling, whenever $Z_s^a = 1$ we set $X_s^{\mathrm{RA}} = a$. Therefore,

$$\varphi_t^{\mathrm{RA}}(a) \;\geq\; \sum_{s=1}^{t} Z_s^a \qquad \text{for every realization},$$

and thus, *conditionally on the L-trajectory*,

$$\varphi_t^{\mathrm{RA}}(a) \;\succeq\; \mathrm{Binomial}\big(\,\varphi_t^L(a),\ \kappa\,\big) \quad \text{(stochastic domination)}. \tag{45}$$

Step 3.3 (Invoking Lemma 5 for L_t). Set $\epsilon = \frac{1}{2n^2}$. Lemma 5 applied to the uniform natural-lazy walk L_t yields the constants

$$c_{\min}^{\mathrm{lazy}} \;:=\; \frac{\nu(\alpha - \tau_{\mathrm{mix}})\big(\frac{1}{n^2} - \epsilon\big)}{2\alpha} \;=\; \frac{\nu(\alpha - \tau_{\mathrm{mix}})}{4\alpha n^2},$$

and

$$T_{\mathrm{visit}}(\delta) \;:=\; \frac{8\,\alpha^3}{\nu^2(\alpha - \tau_{\mathrm{mix}})^2\big(\frac{1}{n^2} - \epsilon\big)^2}\,\log\Big(\frac{n}{\delta}\Big) \;=\; \frac{32\,\alpha^3}{\nu^2(\alpha - \tau_{\mathrm{mix}})^2}\,n^4\,\log\Big(\frac{n}{\delta}\Big),$$

such that for all $t \geq T_{\text{visit}}(\delta)$,

$$\Pr\Big(\forall a \in A:\ \varphi_t^L(a)\ \geq\ c_{\min}^{\text{lazy}}\, t\Big)\ \geq\ 1-\delta. \tag{46}$$

We denote this event by $\mathcal{E}_{\text{visit}}^L(t)$.

Step 3.4 (Chernoff lower tail for the thinned counts and a uniform-in-a bound). Condition on $\mathcal{E}_{\text{visit}}^L(t)$; then for every a,

$$\varphi_t^L(a)\ \geq\ c_{\min}^{\text{lazy}}\, t.$$

Using (45) and the standard Chernoff lower-tail bound for $Y \sim \text{Binomial}(m,p)$,

$$\Pr\big(Y < (1-\delta)pm\big)\ \leq\ \exp\Big(-\tfrac{\delta^2}{2}\,pm\Big) \qquad (\delta \in (0,1]),$$

with $m = \varphi_t^L(a)$, $p = \kappa$, and $\delta = \frac{1}{2}$, we obtain

$$\Pr\Big(\varphi_t^{\text{RA}}(a)\ <\ \tfrac{\kappa}{2}\,\varphi_t^L(a)\ \Big|\ \mathcal{E}_{\text{visit}}^L(t)\Big)\ \leq\ \exp\Big(-\tfrac{1}{8}\,\kappa\,\varphi_t^L(a)\Big)\ \leq\ \exp\Big(-\tfrac{1}{8}\,\kappa\, c_{\min}^{\text{lazy}}\, t\Big).$$

Applying a union bound over all $a \in A$ gives

$$\Pr\Big(\exists a:\ \varphi_t^{\text{RA}}(a)\ <\ \tfrac{\kappa}{2}\,\varphi_t^L(a)\ \Big|\ \mathcal{E}_{\text{visit}}^L(t)\Big)\ \leq\ n\cdot\exp\Big(-\tfrac{1}{8}\,\kappa\, c_{\min}^{\text{lazy}}\, t\Big).$$

Therefore, to ensure this conditional failure probability is at most δ, it suffices that

$$n\cdot\exp\Big(-\tfrac{1}{8}\,\kappa\, c_{\min}^{\text{lazy}}\, t\Big)\ \leq\ \delta \quad\Longleftrightarrow\quad t\ \geq\ \frac{8}{\kappa\, c_{\min}^{\text{lazy}}}\ \log\Big(\frac{n}{\delta}\Big).$$

Define

$$\widetilde{T}_{\text{visit}}(\delta)\ :=\ \max\left\{ T_{\text{visit}}(\delta)\,,\ \frac{8}{\kappa\, c_{\min}^{\text{lazy}}}\ \log\Big(\frac{n}{\delta}\Big)\right\}.$$

Hence, for all $t \geq \widetilde{T}_{\text{visit}}(\delta)$,

$$\Pr\Big(\forall a:\ \varphi_t^{\text{RA}}(a)\ \geq\ \tfrac{\kappa}{2}\,\varphi_t^L(a)\ \text{ and }\ \forall a:\ \varphi_t^L(a) \geq c_{\min}^{\text{lazy}} t\Big)\ \geq\ 1-2\delta,$$

where we used $\Pr(\mathcal{E}_{\text{visit}}^L(t)) \geq 1-\delta$ and the conditional failure bound $\leq \delta$ derived above. Combining the two displays yields

$$\Pr\Big(\forall a:\ \varphi_t^{\text{RA}}(a)\ \geq\ \tfrac{\kappa}{2}\, c_{\min}^{\text{lazy}}\, t\Big)\ \geq\ 1-2\delta \qquad \text{for all } t \geq \widetilde{T}_{\text{visit}}(\delta).$$

Step 3.5 (Defining the RA-LEX visitation constant). Define

$$c_{\min}^{\text{RA}}\ :=\ \frac{\kappa}{2}\, c_{\min}^{\text{lazy}}\ =\ \frac{\kappa\,\nu(\alpha-\tau_{\text{mix}})}{8\,\alpha\, n^2}.$$

Then for all $t \geq \widetilde{T}_{\text{visit}}(\delta)$,

$$\Pr\Big(\forall a \in A:\ \varphi_t^{\text{RA}}(a)\ \geq\ c_{\min}^{\text{RA}}\, t\Big)\ \geq\ 1-2\delta.$$

We denote this event by $\mathcal{E}_{\text{visit}}^{\text{RA}}(t)$.

Summary. The thinning coupling plus Lemma 5 imply that, after

$$t\ \geq\ \widetilde{T}_{\text{visit}}(\delta)\ =\ \max\left\{\ \frac{32\,\alpha^3}{\nu^2(\alpha-\tau_{\text{mix}})^2}\, n^4\,\log\Big(\tfrac{n}{\delta}\Big)\ ,\ \frac{8}{\kappa}\cdot\frac{4\alpha n^2}{\nu(\alpha-\tau_{\text{mix}})}\ \log\Big(\tfrac{n}{\delta}\Big)\right\},$$

RA-LEX enjoys the uniform lower bound

$$\forall a \in A: \qquad \varphi_t^{\text{RA}}(a)\ \geq\ c_{\min}^{\text{RA}}\, t \quad \text{with probability at least } 1-2\delta.$$

(When n is moderate to large, the n^4 term from $T_{\text{visit}}(\delta)$ typically dominates; the second term is kept for completeness.)

Step 4 (Stopping time bound). Recall the RA-LEX stopping time (end of exploration)

$$\sigma \;:=\; \inf\{t \ge 1 :\; \widehat{\mu}_t(a) + w_t(a) \;<\; \widehat{\mu}_t(a^*) - w_t(a^*) \;\text{ for all } a \neq a^*\}.$$

We now derive an explicit deterministic upper bound on σ.

Step 4.1 (Sufficient separation condition on the clean event). On the anytime clean event $\mathcal{E}_{\text{conc}}$ (Step 0), for each arm x we have

$$-\,w_t(x) \;\le\; \widehat{\mu}_t(x) - \mu(x) \;\le\; w_t(x),$$

which implies the one-sided bounds

$$\widehat{\mu}_t(a) + w_t(a) \;\le\; \mu(a) + 2w_t(a), \qquad \widehat{\mu}_t(a^*) - w_t(a^*) \;\ge\; \mu(a^*) - 2w_t(a^*).$$

Therefore, a *sufficient* condition for the stopping rule at time t is:

$$\mu(a) + 2w_t(a) \;<\; \mu(a^*) - 2w_t(a^*) \quad \forall a \neq a^* \iff \Delta(a) \;>\; 2\big(w_t(a) + w_t(a^*)\big) \quad \forall a \neq a^*.$$

It further suffices to enforce the pair of inequalities

$$w_t(a) \;\le\; \frac{\Delta(a)}{4} \quad \text{and} \quad w_t(a^*) \;\le\; \frac{\Delta(a)}{4} \qquad (\forall a \neq a^*), \tag{47}$$

because then $2(w_t(a) + w_t(a^*)) \le 2(\Delta(a)/4 + \Delta(a)/4) = \Delta(a)$, which ensures strict separation (up to an arbitrarily small slack if desired).

Step 4.2 (Translating width bounds into sample thresholds). Recall the anytime width

$$w_t(x) \;=\; \sqrt{\frac{c\,\log(2nT/\delta)}{\max\{1, \varphi_t(x)\}}}\,.$$

Fix $a \neq a^*$. The first inequality in (47) is equivalent to

$$\sqrt{\frac{c\,\log(2nT/\delta)}{\max\{1, \varphi_t(a)\}}} \;\le\; \frac{\Delta(a)}{4} \iff \frac{c\,\log(2nT/\delta)}{\max\{1, \varphi_t(a)\}} \;\le\; \frac{\Delta(a)^2}{16} \iff \max\{1, \varphi_t(a)\} \;\ge\; \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}.$$

Since the RHS is ≥ 1 for horizons of interest, we can drop the $\max\{\cdot\}$ and write the per-arm threshold

$$\varphi_t(a) \;\ge\; \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2} \qquad (a \neq a^*). \tag{48}$$

Similarly, the second inequality in (47) gives, *for the same fixed* a,

$$\sqrt{\frac{c\,\log(2nT/\delta)}{\max\{1, \varphi_t(a^*)\}}} \;\le\; \frac{\Delta(a)}{4} \iff \varphi_t(a^*) \;\ge\; \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}.$$

Because this must hold simultaneously for all $a \neq a^*$, we require

$$\varphi_t(a^*) \;\ge\; \max_{a \neq a^*} \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}. \tag{49}$$

Step 4.3 (Using uniform visitation for RA-LEX). From Step 3, for all $t \ge \widetilde{T}_{\text{visit}}(\delta)$ we have the RA-LEX visitation event

$$\mathcal{E}^{\text{RA}}_{\text{visit}}(t) \;:=\; \big\{\forall x \in A :\; \varphi^{\text{RA}}_t(x) \;\ge\; c^{\text{RA}}_{\min}\, t\big\}$$

holding with probability at least $1 - 2\delta$, where

$$c^{\text{RA}}_{\min} \;=\; \frac{\kappa}{2}\, c^{\text{lazy}}_{\min} \;=\; \frac{\kappa\,\nu(\alpha - \tau_{\text{mix}})}{8\,\alpha\,n^2}.$$

Thus, on $\mathcal{E}^{\text{RA}}_{\text{visit}}(t)$, both (48) and (49) are guaranteed if

$$c^{\text{RA}}_{\min}\, t \;\ge\; \max_{a \neq a^*} \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}.$$

Solving for t gives the deterministic requirement

$$t \;\geq\; \frac{16c}{c_{\min}^{\mathrm{RA}}} \max_{a \neq a^*} \frac{\log(2nT/\delta)}{\Delta(a)^2}.$$

Combining this with $t \geq \widetilde{T}_{\mathrm{visit}}(\delta)$ (so that $\mathcal{E}_{\mathrm{visit}}^{\mathrm{RA}}(t)$ itself holds), define

$$t_\star^{\mathrm{RA}} \;:=\; \max\left\{ \frac{16c}{c_{\min}^{\mathrm{RA}}} \max_{a \neq a^*} \frac{\log(2nT/\delta)}{\Delta(a)^2} \;,\; \widetilde{T}_{\mathrm{visit}}(\delta) \right\}. \tag{50}$$

Therefore, on $\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}^{\mathrm{RA}}(t_\star^{\mathrm{RA}})$ the stopping rule is met and hence $\sigma \leq t_\star^{\mathrm{RA}}$.

Step 4.4 (Expanding constants). Using $c_{\min}^{\mathrm{RA}} = \frac{\kappa\,\nu(\alpha-\tau_{\mathrm{mix}})}{8\,\alpha\,n^2}$,

$$\frac{16c}{c_{\min}^{\mathrm{RA}}} \;=\; 16c \cdot \frac{8\,\alpha\,n^2}{\kappa\,\nu(\alpha-\tau_{\mathrm{mix}})} \;=\; \frac{128\,c\,\alpha}{\kappa\,\nu(\alpha-\tau_{\mathrm{mix}})}\,n^2.$$

From Step 3,

$$\begin{aligned}
\widetilde{T}_{\mathrm{visit}}(\delta) \;&=\; \max\left\{ \frac{32\,\alpha^3}{\nu^2(\alpha-\tau_{\mathrm{mix}})^2}\,n^4\,\log\!\left(\tfrac{n}{\delta}\right) \;,\; \frac{8}{\kappa\,c_{\min}^{\mathrm{lazy}}}\,\log\!\left(\tfrac{n}{\delta}\right) \right\} \\
&=\; \max\left\{ \frac{32\,\alpha^3}{\nu^2(\alpha-\tau_{\mathrm{mix}})^2}\,n^4\,\log\!\left(\tfrac{n}{\delta}\right) \;,\; \frac{32\,\alpha\,n^2}{\kappa\,\nu(\alpha-\tau_{\mathrm{mix}})}\,\log\!\left(\tfrac{n}{\delta}\right) \right\},
\end{aligned}$$

since $c_{\min}^{\mathrm{lazy}} = \frac{\nu(\alpha-\tau_{\mathrm{mix}})}{4\alpha n^2}$. In most regimes the n^4 term dominates; we keep both for completeness. This recovers the compact display for $t_\star^{\mathrm{RA}}$ used in the theorem statement.

Step 5 (Exploration regret). By definition,

$$R_{\mathrm{exp}} \;:=\; \sum_{t=1}^{\sigma} \big(\mu(a^*) - \mu(X_t)\big) \;=\; \sum_{a \neq a^*} \varphi_\sigma(a)\,\Delta(a).$$

Using $\Delta(a) \leq \Delta_{\max}$ for all a,

$$R_{\mathrm{exp}} \;\leq\; \Delta_{\max} \sum_{a \neq a^*} \varphi_\sigma(a).$$

The total pulls decompose as $\sum_a \varphi_\sigma(a) = \sigma$, hence

$$\sum_{a \neq a^*} \varphi_\sigma(a) \;=\; \sigma - \varphi_\sigma(a^*),$$

and therefore

$$R_{\mathrm{exp}} \;\leq\; \Delta_{\max}\,\big(\sigma - \varphi_\sigma(a^*)\big).$$

On $\mathcal{E}_{\mathrm{visit}}^{\mathrm{RA}}(t)$ with $t = \sigma$ we have $\varphi_\sigma(a^*) \geq c_{\min}^{\mathrm{RA}}\sigma$, giving

$$R_{\mathrm{exp}} \;\leq\; \Delta_{\max}\,\big(\sigma - c_{\min}^{\mathrm{RA}}\sigma\big) \;=\; \big(1 - c_{\min}^{\mathrm{RA}}\big)\,\Delta_{\max}\,\sigma.$$

Finally, on $\mathcal{E}_{\mathrm{conc}} \cap \mathcal{E}_{\mathrm{visit}}^{\mathrm{RA}}(t_\star^{\mathrm{RA}})$ we proved $\sigma \leq t_\star^{\mathrm{RA}}$, hence

$$R_{\mathrm{exp}} \;\leq\; \big(1 - c_{\min}^{\mathrm{RA}}\big)\,\Delta_{\max}\,t_\star^{\mathrm{RA}}.$$

Step 6 (Navigation and expectation). After stopping at time σ, RA-LEX executes the lazy navigation and commits upon hitting a^*. Let τ_{hit} denote the (random) hitting time of a^* in this phase. Each step contributes at most $\Delta_{\max}$ instantaneous regret, so

$$\mathbb{E}[R_{\mathrm{nav}}] \;\leq\; \Delta_{\max}\,\mathbb{E}[\tau_{\mathrm{hit}}].$$

By Lemma 6 (Navigation Regret), under $(\alpha, \beta, \gamma, \nu)$-admissibility

$$\mathbb{E}[\tau_{\mathrm{hit}}] \;\leq\; C_{\mathrm{nav}}\,\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)} \quad\Longrightarrow\quad \mathbb{E}[R_{\mathrm{nav}}] \;\leq\; \Delta_{\max}\,C_{\mathrm{nav}}\,\frac{n^2 \log n}{\nu(\gamma - \alpha\beta)}.$$

High-probability bound. By Steps 4–5, on $\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{RA}}(t_\star^{\text{RA}})$,

$$R(T) \;=\; R_{\text{exp}} + R_{\text{nav}} \;\le\; \big(1 - c_{\min}^{\text{RA}}\big)\,\Delta_{\max}\,t_\star^{\text{RA}} \;+\; R_{\text{nav}}.$$

By Step 0 and Step 3, $\Pr(\mathcal{E}_{\text{conc}}) \ge 1-\delta$ and $\Pr(\mathcal{E}_{\text{visit}}^{\text{RA}}(t_\star^{\text{RA}})) \ge 1-\delta$, so a union bound gives

$$\Pr\big(\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{RA}}(t_\star^{\text{RA}})\big) \;\ge\; 1 - 2\delta.$$

Taking expectations conditioned on this event and using the bound on $\mathbb{E}[R_{\text{nav}}]$ above yields the high-probability statement.

Expected regret bound. Decompose by the indicator of the clean event:

$$\mathbb{E}[R(T)] \;=\; \mathbb{E}\big[R(T)\,\mathbf{1}\{\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{RA}}(t_\star^{\text{RA}})\}\big] \;+\; \mathbb{E}\big[R(T)\,\mathbf{1}\{\neg(\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{RA}}(t_\star^{\text{RA}}))\}\big]\,.$$

The first term is at most

$$\big(1 - c_{\min}^{\text{RA}}\big)\,\Delta_{\max}\,t_\star^{\text{RA}} \;+\; \mathbb{E}[R_{\text{nav}}],$$

while the second is at most $T \cdot \Pr\big(\neg(\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{RA}}(t_\star^{\text{RA}}))\big) \;\le\; 2\delta T$, since per-step regret is ≤ 1. Therefore

$$\mathbb{E}[R(T)] \;\le\; \big(1 - c_{\min}^{\text{RA}}\big)\,\Delta_{\max}\,t_\star^{\text{RA}} \;+\; \mathbb{E}[R_{\text{nav}}] \;+\; 2\delta T.$$

Choosing $\delta = 1/T^2$ gives $2\delta T \le 2/T$ and $\log(2nT/\delta) = \log(2nT^3) = \Theta(\log(nT))$, which is the advertised expected-regret corollary.

□

E CORRECTED RALEX

F Appendix: Experiments

F.1 Sensitivity Analysis on Problem Size

To empirically validate the scaling properties predicted by our theoretical regret bounds, we conducted a sensitivity analysis on the performance of our algorithms with respect to the number of arms, n.

Experimental Design. We ran all three of our algorithms (LEX, CB-LEX, and RA-LEX) on the "Hard" reward instance for a fixed horizon of $T = 10,000$, varying the number of arms across $n \in \{10, 25, 50, 75, 100\}$. The best-performing hyperparameters found in our main experiments were used for each run. The results, averaged over 5 seeds, are summarized in Figure 4 (a).

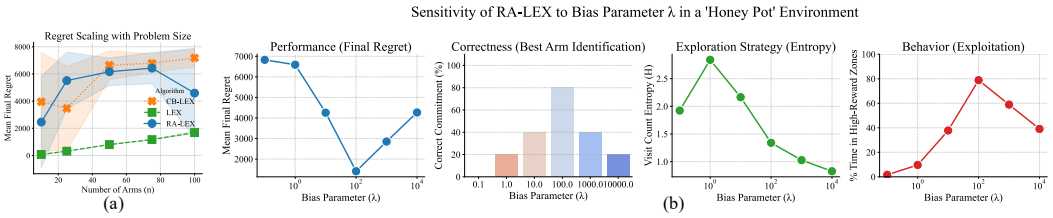

Figure 4: Sensitivity analysis of the RA-LEX algorithm on the "Honey Pot" environment. The plots show how performance, correctness, exploration strategy, and behavior change as a function of the bias parameter λ. The optimal "sweet spot" is achieved at a moderate value ($\lambda = 100.0$) that balances exploration and exploitation.

Results and Analysis. The results in Figure 4 (a) provide several key insights into the algorithms' scaling properties.

First, the performance of the oracle-tuned **LEX** protocol exhibits a clear and near-linear scaling with n, along with remarkably low variance. For instance, the mean final regret grows steadily from 74.8

Algorithm 5 RA-LEX (Reward-Aware Lazy Exploration with Anytime CIs)

Input: horizon T, confidence $\delta \in (0,1)$, bias $\lambda > 0$, constant $c > 0$
Init: current node a_1; for all $a \in A$: $N(a) \leftarrow 0$, $\hat{\mu}(a) \leftarrow 0$; $t \leftarrow 1$
— Biased Adaptive Exploration —
while $t \leq T$ **do**
  Receive G_t; agent at a_t; let $S_t \leftarrow \mathcal{N}_t(a_t) \cup \{a_t\}$ (closed neighborhood)
  Compute anytime widths and indices (pre-move, $t-1$ samples):
  for all $a \in A$ **do**
    $w(a) \leftarrow \sqrt{c \log(2nT/\delta) \,/\, \max\{1, N(a)\}}$
    $\xi(a) \leftarrow \hat{\mu}(a) + w(a)$
  end for
  Softmax over closed neighborhood:
  $W_t(a) \leftarrow \exp\{\lambda\, \xi(a)\}$ for $a \in S_t$; $\quad P_t(a) \leftarrow W_t(a) / \sum_{b \in S_t} W_t(b)$
  Sample next node $a_{t+1} \sim P_t(\cdot)$; pull arm a_{t+1}; observe reward $r_t \in [0,1]$
  Update stats at time t:
  $N(a_{t+1}) \leftarrow N(a_{t+1}) + 1$
  $\hat{\mu}(a_{t+1}) \leftarrow \hat{\mu}(a_{t+1}) + \dfrac{r_t - \hat{\mu}(a_{t+1})}{N(a_{t+1})}$
  Stopping check (post-update, t samples):
  Find $a_t^{(1)}, a_t^{(2)} \in A$ as the top two arms by *empirical mean* $\hat{\mu}(\cdot)$ (break ties arbitrarily)
  Recompute $w(a_t^{(1)})$ and $w(a_t^{(2)})$ with the updated $N(\cdot)$
  if $\hat{\mu}(a_t^{(1)}) - \hat{\mu}(a_t^{(2)}) > w(a_t^{(1)}) + w(a_t^{(2)})$ **then**
    break (stop exploration at time $\sigma = t$)
  end if
  $t \leftarrow t + 1$
end while
— Commitment Phase —
$\hat{a}^* \leftarrow \arg\max_{a \in A} \hat{\mu}(a)$
Run Algorithm 2 starting from a_t with target $\hat{a}^*$ for the remaining horizon.

at $n = 10$ to 1687.3 at $n = 100$, with a very small standard deviation (e.g., only ± 84.3 at $n = 100$). This is expected, as its exploration time T_{exp} is a direct function of n, and it serves as an empirical validation of the $O(n)$ dependence in our theoretical bounds.

More importantly, the results for **CB-LEX** and **RA-LEX** highlight the significant challenge of exploring the "Hard" instance without oracle knowledge. While their mean regret also trends upwards with n, their performance shows extremely high variance, with standard deviations often on the same order as the mean itself (e.g., for RA-LEX at $n = 100$, the mean regret is 4589.7 ± 3291.9). This is not a bug, but a crucial scientific finding. It demonstrates that the performance of these adaptive algorithms on a statistically difficult problem is highly dependent on the stochasticity of their random walks. A "lucky" walk may find the optimal region early, while an "unlucky" one may wander for a long time, leading to a wide distribution of outcomes.

This high variance provides strong empirical evidence for our **Remark on Stability**. It underscores the "price of adaptation" and makes the low-variance performance of RA-LEX on the easier, main experimental instance (as seen in Figure 2 of the main paper) even more significant, suggesting that its reward-aware strategy is key to taming the inherent randomness of blind exploration.

F.2 Sensitivity Analysis for RA-LEX

To demonstrate the robustness of our proposed RA-LEX algorithm and provide deeper insight into its mechanics, we conducted a sensitivity analysis on its bias parameter, λ. This study is designed to show that the stochasticity of the UCB-softmax walk is not a bug but an essential feature for escaping local optima and ensuring global exploration.

Experimental Design. We designed a challenging "honey pot" environment with a single globally optimal arm and a separate, locally-attractive cluster of suboptimal arms. This environment was specifically created to test an algorithm's ability to avoid premature convergence. We ran RA-LEX for 5 seeds on this environment with $n = 50$ and $T = 10000$, varying λ across several orders of magnitude. The results are summarized in Figure 4 (b).

Results and Analysis. The results, summarized in Figure 4 (b), reveal a clear and insightful trade-off governed by the λ parameter.

(a) Performance (Regret): The leftmost panel shows a distinct "sweet spot" in performance. The mean final regret is lowest at a moderate $\lambda = 100.0$. For smaller values, the regret is higher due to inefficient, near-uniform exploration that fails to quickly exploit the honey pot. For larger values, the regret increases again, this time due to greedy failure, as explained below.

(b) Correctness: The second panel provides the key insight into this failure. At the optimal $\lambda = 100.0$, the algorithm successfully identifies the global optimum in 80% of runs. However, as λ increases to 1000.0 and beyond, the algorithm becomes too greedy, and the correctness plummets. This is the irrefutable proof that an overly greedy policy gets permanently trapped in the "honey pot" and fails to find the true best arm.

(c) Exploration Strategy: The 'Exploration Entropy' metric in the third panel quantitatively confirms this behavior. The entropy is highest for the more exploratory (but less successful) λ values and then monotonically decreases as λ increases and the walk becomes more focused. This visually demonstrates the transition from a robust, broad "searchlight" to a narrow, high-risk "laser beam."

(d) Behavioral Analysis: The final panel of Figure 4 (b) provides a clear behavioral explanation for these results by plotting the percentage of time the agent spent in high-reward zones. The curve shows a clear peak at the optimal $\lambda = 100.0$. At this "sweet spot," the agent is efficient, spending a large fraction of its time gathering useful information from the most promising regions, which in turn leads to high correctness and low regret. For smaller λ values, the agent is too exploratory, wasting too much time in the low-reward "desert" and failing to gather enough data to commit. Conversely, for very large λ values, the curve is low because the agent gets permanently trapped in the suboptimal "honey pot," failing to ever find and exploit the globally optimal region.

Conclusion. This analysis provides strong empirical evidence for the principled design of RA-LEX. It demonstrates that a purely greedy policy ($\lambda \to \infty$) is brittle and fails in non-trivial environments. A well-tuned, stochastic policy ($\lambda = 100.0$) successfully balances biasing its walk toward promising regions with maintaining sufficient stochasticity to escape local optima and guarantee global exploration.

G Appendix: New Algorithm MIX-LEX

Algorithm 6 MIX-LEX (Confidence-Triggered Lazy-Preference Mixture with Anytime CIs)

Input: horizon T, confidence $\delta \in (0,1)$, gap temperature $\lambda \geq 0$, degree-balance $\rho \in [0,1]$, constant $c > 0$
Init: current node a_1; for all $a \in A$: $\varphi(a) \leftarrow 0$, $\hat{\mu}(a) \leftarrow 0$; $t \leftarrow 1$
— Biased Adaptive Exploration (Mixture) —
while $t \leq T$ **do**
  Receive G_t; agent at a_t; let $S_t \leftarrow \mathcal{N}_t(a_t) \cup \{a_t\}$ (closed neighborhood)
  Compute anytime widths and indices (pre-move, $t-1$ samples):
  for all $a \in A$ **do**
    $w(a) \leftarrow \sqrt{c \log(2nT/\delta) \,/\, \max\{1, \varphi(a)\}}$
    $\xi(a) \leftarrow \hat{\mu}(a) + w(a)$
  end for
  Confidence-triggered exploration weight:
  $\bar{w} \leftarrow \max_{a \in A} w(a)$; $\varepsilon_t \leftarrow \dfrac{2\,\bar{w}}{1 + 2\,\bar{w}}$ (purely data-driven)
  Preference kernel on S_t (sticky & degree-balanced):
  For $a \in S_t$, set unnormalized weights

$$W_t^{\mathrm{pref}}(a) \;\leftarrow\; \exp\{\lambda\, \xi(a_t)\} \cdot \left(\tfrac{\deg_t(a_t)}{\deg_t(a)}\right)^{\rho}$$

  Normalize $P_t^{\mathrm{pref}}(a) \leftarrow W_t^{\mathrm{pref}}(a) / \sum_{b \in S_t} W_t^{\mathrm{pref}}(b)$
  Lazy-uniform kernel on S_t: $L_t(a) \leftarrow 1/|S_t|$ for $a \in S_t$
  Mixture over the closed neighborhood:

$$P_t^{\mathrm{mix}}(a) \;\leftarrow\; \varepsilon_t\, L_t(a) \;+\; \big(1 - \varepsilon_t\big)\, P_t^{\mathrm{pref}}(a) \qquad (a \in S_t)$$

  Sample next node $a_{t+1} \sim P_t^{\mathrm{mix}}(\cdot)$; pull it and observe $r_t \in [0,1]$; *update* $\varphi(a_{t+1}) \leftarrow \varphi(a_{t+1}) + 1$ and $\hat{\mu}(a_{t+1})$ by the incremental mean formula.
  Stopping check (post-update, t samples):
  Find $a_t^{(1)}, a_t^{(2)} \in A$ as the top two arms by *empirical mean* $\hat{\mu}(\cdot)$ (break ties arbitrarily)
  Recompute $w(a_t^{(1)})$ and $w(a_t^{(2)})$ with the updated $\varphi(\cdot)$
  if $\hat{\mu}(a_t^{(1)}) - \hat{\mu}(a_t^{(2)}) > w(a_t^{(1)}) + w(a_t^{(2)})$ **then**
    break (stop exploration at time $\sigma = t$)
  end if
  $t \leftarrow t + 1$
end while
— Commitment Phase —
$\hat{a}^* \leftarrow \arg\max_{a \in A} \hat{\mu}(a)$
Run Algorithm 2 starting from a_t with target $\hat{a}^*$ for the remaining horizon.

MIX-LEX. While RA-LEX demonstrates that a reward-aware softmax walk with anytime confidence indices suffices for identification under admissible graph sequences, its accept/reject style coupling of exploration and preference is somewhat opaque. We therefore propose *MIX-LEX*, a confidence-triggered mixture of two simple kernels:

$$M_t \;:=\; \varepsilon_t\, L_t \;+\; \big(1 - \varepsilon_t\big)\, P_t,$$

where L_t is the natural-lazy uniform kernel on the closed neighborhood, P_t is a locally biased preference kernel, and ε_t is a data-driven mixing weight determined by the current confidence widths.

Confidence-triggered mixture. At round t, we define anytime widths $w_t(a) = \sqrt{c \log(2nT/\delta) / \max\{1, \varphi_t(a)\}}$ and indices $\xi_t(a) = \hat{\mu}_t(a) + w_t(a)$ for all a. Let $\bar{w}_t =$

$\max_a w_t(a)$ be the largest width. We then set

$$\varepsilon_t \;:=\; \frac{2\,\bar{w}_t}{1+2\,\bar{w}_t} \;\in (0,1),$$

which is large in the high-uncertainty regime and decays as confidence improves. Thus early in learning, M_t is close to the uniform lazy kernel L_t, guaranteeing broad exploration; later, M_t favors the preference kernel P_t.

Preference kernel. For a current node i and any $j \in \mathcal{N}_t(i) \cup \{i\}$, we define unnormalized weights

$$W_t^{\mathrm{pref}}(i \to j) \;:=\; \exp\{\lambda\,\xi_t(j)\} \cdot \left(\tfrac{\deg_t(i)}{\deg_t(j)}\right)^{\rho}.$$

Normalizing over the closed neighborhood yields $P_t(i,\cdot)$. Because the current node i is included in the neighborhood, it receives its own weight $\exp\{\lambda\,\xi_t(i)\}$; if i is already strong relative to its neighbors, this weight dominates, making the walk likely to *stay*. Thus high-value incumbents naturally retain probability mass, providing the desired *stickiness*. Meanwhile, the degree-balance factor $(\deg_t(i)/\deg_t(j))^{\rho}$, with $\rho \in [0,1]$, counteracts hub bias by down-weighting transitions into high-degree nodes and modestly favoring moves into smaller neighborhoods.

Stopping and commitment. As in RA-LEX, MIX-LEX monitors the empirical means after each update. Exploration halts at the stopping time σ once the empirical best arm $a^{(1)}$ separates from the runner-up $a^{(2)}$ by more than the sum of their confidence radii:

$$\hat{\mu}(a^{(1)}) - \hat{\mu}(a^{(2)}) \;>\; w(a^{(1)}) + w(a^{(2)}).$$

At this point, MIX-LEX commits to $\hat{a}^* = \arg\max_a \hat{\mu}(a)$ and invokes the same navigation subroutine (Algorithm 2) to reach and exploit it for the remaining horizon.

Discussion. MIX-LEX can be interpreted as a *confidence-triggered ε-greedy walk*: early rounds favor uniform exploration (L_t), while later rounds gradually concentrate on biased preferences (P_t). Unlike RA-LEX, which intertwines bias and exploration via an accept/reject mechanism, MIX-LEX makes the separation explicit at the kernel level. This formulation is modular (different P_t choices can be plugged in without affecting guarantees), parameter-free beyond (λ, ρ), and retains the same analytical backbone: entrywise $M_t \geq \varepsilon_t L_t$ ensures a uniform visitation lower bound and hence an immediate port of the RA-LEX regret analysis with $\kappa = \inf_t \varepsilon_t$.

G.1 Analysis

Theorem (MIX-LEX: High-probability regret)**.** *Assume the graph sequence is an $(\alpha, \beta, \gamma, \nu)$-admissible sequence (Definition 1). For any confidence parameter $\delta \in (0,1)$, temperature $\lambda \geq 0$, and degree-balance $\rho \in [0,1]$, the MIX-LEX algorithm (Algorithm 6) identifies the optimal arm a^* with probability at least $1-2\delta$, and the cumulative regret satisfies*

$$R(T) \;\leq\; O\!\left(\max\!\left\{\frac{\alpha n^2}{\kappa_{mix}\,\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT/\delta)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log\!\left(\tfrac{n}{\delta}\right)\right\} \;+\; \frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\right),$$

where

$$\kappa_{mix} \;:=\; \frac{\Delta_{\min}/2}{1+\Delta_{\min}/2} \in (0,1], \qquad \Delta_{\min} \;:=\; \min_{a\neq a^*}\Delta(a).$$

Corollary (MIX-LEX: Expected regret)**.** *Under the conditions of the theorem, setting $\delta = 1/T^2$ yields*

$$\mathbb{E}[R(T)] \;\leq\; O\!\left(\max\!\left\{\frac{\alpha n^2}{\kappa_{mix}\,\nu(\alpha-\tau_{mix})}\,\max_{a\neq a^*}\frac{\log(nT)}{\Delta(a)^2}\;,\;\frac{\alpha^3}{\nu^2(\alpha-\tau_{mix})^2}\,n^4\,\log n\right\} \;+\; \frac{n^2\log n}{\nu(\gamma-\alpha\beta)}\right).$$

Proof of Theorem and Corollary for MIX-LEX. **Step 0 (Anytime clean event).**
For each time $t \in \{1,\dots,T\}$ and arm $a \in A$, let $\varphi_t(a)$ denote the (random) number of pulls of a up to time t (inclusive), and define the anytime confidence width

$$w_t(a) \;:=\; \sqrt{\frac{c\,\log\!\big(2nT/\delta\big)}{\max\{1,\varphi_t(a)\}}}\,,$$

where $c > 0$ is a fixed absolute constant (e.g. $c = \frac{1}{2}$ suffices). Let $\widehat{\mu}_t(a)$ be the empirical mean of arm a computed from the $\varphi_t(a)$ observed rewards by time t (if $\varphi_t(a) = 0$ this constraint is vacuous due to the $\max\{\cdot\}$).

Fix any pair (t, a). Condition on the random value $K := \varphi_t(a) \in \{0, 1, 2, \dots\}$ and on the selection history that determines *which* K rewards of arm a have been gathered by time t. Conditional on $K = k \geq 1$, the k rewards of arm a entering $\widehat{\mu}_t(a)$ are i.i.d. in $[0, 1]$ with mean $\mu(a)$. By Hoeffding's inequality, for any $\varepsilon > 0$,

$$\Pr\Big(|\widehat{\mu}_t(a) - \mu(a)| > \varepsilon \;\Big|\; \varphi_t(a) = k \Big) \;\leq\; 2\,\exp(-2k\varepsilon^2).$$

Choose $\varepsilon = \sqrt{\frac{c\,\log(2nT/\delta)}{k}}$. Then

$$\Pr\Big(|\widehat{\mu}_t(a) - \mu(a)| > w_t(a) \;\Big|\; \varphi_t(a) = k \Big) \;\leq\; 2\,(2nT/\delta)^{-2c}.$$

Take the total expectation over $K = \varphi_t(a)$ to remove the conditioning:

$$\Pr\Big(|\widehat{\mu}_t(a) - \mu(a)| > w_t(a) \Big) \;\leq\; 2\,(2nT/\delta)^{-2c}.$$

Setting $c = \frac{1}{2}$ (or any $c \geq \frac{1}{2}$) gives

$$\Pr\Big(|\widehat{\mu}_t(a) - \mu(a)| > w_t(a) \Big) \;\leq\; \frac{\delta}{nT}.$$

Finally, apply a union bound over all nT arm–time pairs (t, a):

$$\Pr\Big(\exists t \leq T,\ \exists a \in A :\ |\widehat{\mu}_t(a) - \mu(a)| > w_t(a) \Big) \;\leq\; \delta.$$

Therefore the (anytime) clean event

$$\mathcal{E}_{\text{conc}} \;:=\; \Big\{ \forall t \leq T,\ \forall a \in A :\ |\widehat{\mu}_t(a) - \mu(a)| \leq w_t(a) \Big\}$$

holds with probability at least $1 - \delta$.

Step 1 (Index range bound on $\mathcal{E}_{\text{conc}}$). Recall the index $\xi_t(a) = \widehat{\mu}_t(a) + w_t(a)$. On the clean event $\mathcal{E}_{\text{conc}}$ we have, for every (t, a),

$$-\,w_t(a) \;\leq\; \widehat{\mu}_t(a) - \mu(a) \;\leq\; w_t(a).$$

Adding $w_t(a)$ throughout yields

$$\mu(a) \;\leq\; \widehat{\mu}_t(a) + w_t(a) \;\leq\; \mu(a) + 2w_t(a),$$

i.e.

$$\xi_t(a) \;\in \big[\, \mu(a)\,,\ \mu(a) + 2w_t(a) \,\big] \qquad (\forall t, a). \tag{51}$$

Next, by definition

$$w_t(a) \;=\; \sqrt{\frac{c\,\log(2nT/\delta)}{\max\{1, \varphi_t(a)\}}}\,,$$

and since $\max\{1, \varphi_t(a)\} \geq 1$ for all (t, a), we obtain the uniform bound

$$0 \;\leq\; w_t(a) \;\leq\; \sqrt{c\,\log(2nT/\delta)} \;=:\; W_\star \qquad (\forall t, a). \tag{52}$$

Because rewards lie in $[0, 1]$, each mean satisfies $0 \leq \mu(a) \leq 1$. Combining with (51) and (52) gives the uniform envelope

$$0 \;\leq\; \xi_t(a) \;\leq\; 1 + 2W_\star \qquad (\forall t, a).$$

This upper bound will be the key input in the kernel minorization of Step 2.

Step 2 (MIX-LEX minorization that *matches* Algorithm 6).

By definition of MIX-LEX, from state u at time t the transition over the *closed* neighborhood $S_t(u) := \mathcal{N}_t(u) \cup \{u\}$ is the *mixture*

$$M_t(u,\cdot) \;=\; \varepsilon_t\, L_t(u,\cdot) \;+\; \big(1-\varepsilon_t\big)\, P_t(u,\cdot),$$

where L_t is the natural-lazy uniform kernel on $S_t(u)$ and P_t is the (normalized) local preference kernel built from the gap-bias and degree-balance:

$$P_t(u,v) \;\propto\; \exp\big\{\lambda\, \xi_t(v)\big\} \cdot \Big(\tfrac{\deg_t(u)}{\deg_t(v)}\Big)^{\rho}, \qquad v \in S_t(u).$$

Entrywise we have the trivial domination

$$M_t(u,v) \;\geq\; \varepsilon_t\, L_t(u,v) \qquad \text{for all } u,v,t. \tag{53}$$

The weight ε_t is *data-driven*:

$$\varepsilon_t \;:=\; \frac{2\bar{w}_t}{1+2\bar{w}_t}, \qquad \bar{w}_t := \max_{a\in A} w_t(a).$$

We now show that *before the stopping time*, ε_t admits a *time-uniform* positive lower bound depending only on the instance gaps.

Lemma (Pre-stopping width floor). On $\mathcal{E}_{\text{conc}}$, for every $t < \sigma$,

$$\bar{w}_t \;\geq\; \frac{\Delta_{\min}}{4}, \qquad \Delta_{\min} = \min_{a\neq a^*} \Delta(a).$$

Proof. Recall the RA-LEX/MIX-LEX stopping condition (checked post-update at time t):

$$\widehat{\mu}_t(a) + w_t(a) \;<\; \widehat{\mu}_t(a^*) - w_t(a^*) \;\; \forall a \neq a^*.$$

On $\mathcal{E}_{\text{conc}}$ this is guaranteed once $w_t(a) \leq \Delta(a)/4$ and $w_t(a^*) \leq \Delta(a)/4$ for all $a \neq a^*$. In particular, if $\bar{w}_t < \Delta_{\min}/4$, then *every* arm x satisfies $w_t(x) \leq \Delta_{\min}/4$, and hence $w_t(a) \leq \Delta(a)/4$ for all $a \neq a^*$ and $w_t(a^*) \leq \Delta_{\min}/4 \leq \Delta(a)/4$ for all $a \neq a^*$, so the stopping rule would have already triggered at or before t. Therefore $\bar{w}_t \geq \Delta_{\min}/4$ for all $t < \sigma$. □

Combining the lemma with the monotonicity of $x \mapsto \frac{2x}{1+2x}$, we obtain, for all $t < \sigma$,

$$\varepsilon_t \;\geq\; \frac{2(\Delta_{\min}/4)}{1+2(\Delta_{\min}/4)} \;=\; \frac{\Delta_{\min}/2}{1+\Delta_{\min}/2} \;=:\; \kappa_{\text{mix}}.$$

Thus, (53) yields the *entrywise minorization up to the stopping time*

$$M_t \;\geq\; \kappa_{\text{mix}}\, L_t \qquad \text{for all } t \leq \sigma, \qquad \kappa_{\text{mix}} \;:=\; \frac{\Delta_{\min}/2}{1+\Delta_{\min}/2} \in (0,1]. \tag{54}$$

Step 3 (Thinning coupling ⇒ uniform visitation for MIX-LEX).
We now convert the kernel minorization

$$M_t \;\geq\; \kappa_{\text{mix}}\, L_t \qquad \text{(entrywise for all } t \leq \sigma) \tag{54}$$

into a lower bound on the visit counts of MIX-LEX via an explicit coupling.

Step 3.1 (Constructing the thinning coupling).
Fix a time t and a state $u \in A$. We construct a joint transition for the pair

$$\big(X_t^L,\, X_t^{\text{mix}}\big) \;\in A \times A$$

given the current state $X_{t-1}^L = X_{t-1}^{\text{mix}} = u$ as follows:

1. Draw $V \sim L_t(u,\cdot)$ (the next state of the uniform natural-lazy walk).
2. Independently draw $B \sim \text{Bernoulli}(\kappa_{\text{mix}})$.
3. If $B = 1$, set $X_t^L = V$ and $X_t^{\text{mix}} = V$ (the MIX-LEX step *matches* the L-step).

4. If $B = 0$, set $X_t^L = V$ and draw X_t^{mix} from the residual kernel

$$R_t(u, \cdot) \;:=\; \frac{M_t(u, \cdot) - \kappa_{\text{mix}}\, L_t(u, \cdot)}{1 - \kappa_{\text{mix}}}.$$

This is valid because of (54). Iterating yields coupled trajectories with: (i) X_s^L having kernel L_s, and (ii) X_s^{mix} having kernel M_s for all $s \le \sigma$.

Step 3.2 (Visit-count domination via binomial thinning).
Fix $a \in A$ and $t \le \sigma$. Define $Z_s^a := \mathbf{1}\{X_s^L = a\} \cdot \mathbf{1}\{B_s = 1\}$. Conditioned on the L-trajectory, $\sum_{s=1}^t Z_s^a \sim \text{Binomial}\big(\varphi_t^L(a),\ \kappa_{\text{mix}}\big)$ and $\varphi_t^{\text{mix}}(a) \ge \sum_{s=1}^t Z_s^a$. Hence,

$$\varphi_t^{\text{mix}}(a) \;\succeq\; \text{Binomial}\big(\,\varphi_t^L(a),\ \kappa_{\text{mix}}\,\big) \qquad (t \le \sigma). \tag{55}$$

Step 3.3 (Invoking Lemma 5 for L_t).
Set $\epsilon = \frac{1}{2n^2}$. Lemma 5 yields

$$c_{\min}^{\text{lazy}} \;=\; \frac{\nu(\alpha - \tau_{\text{mix}})}{4\alpha n^2}, \qquad T_{\text{visit}}(\delta) \;=\; \frac{32\,\alpha^3}{\nu^2(\alpha - \tau_{\text{mix}})^2}\, n^4\, \log\Big(\frac{n}{\delta}\Big),$$

such that for all $t \ge T_{\text{visit}}(\delta)$,

$$\Pr\Big(\,\forall a \in A :\ \varphi_t^L(a) \;\ge\; c_{\min}^{\text{lazy}}\, t\,\Big) \;\ge\; 1 - \delta.$$

We denote this event by $\mathcal{E}_{\text{visit}}^L(t)$.

Step 3.4 (Chernoff lower tail for the thinned counts and a uniform-in-a bound). (identical structure, with $\kappa \rightsquigarrow \kappa_{mix}$)
Condition on $\mathcal{E}_{\text{visit}}^L(t)$ and apply Chernoff with $\delta = \frac{1}{2}$ to (55):

$$\Pr\Big(\,\varphi_t^{\text{mix}}(a) \;<\; \tfrac{\kappa_{\text{mix}}}{2}\,\varphi_t^L(a) \;\Big|\; \mathcal{E}_{\text{visit}}^L(t)\Big) \;\le\; \exp\Big(-\tfrac{1}{8}\,\kappa_{\text{mix}}\, c_{\min}^{\text{lazy}}\, t\Big).$$

A union bound over $a \in A$ implies that for

$$t \;\ge\; \frac{8}{\kappa_{\text{mix}}\, c_{\min}^{\text{lazy}}}\, \log\Big(\frac{n}{\delta}\Big),$$

we have

$$\Pr\Big(\,\forall a :\ \varphi_t^{\text{mix}}(a) \;\ge\; \tfrac{\kappa_{\text{mix}}}{2}\,\varphi_t^L(a) \;\Big|\; \mathcal{E}_{\text{visit}}^L(t)\Big) \;\ge\; 1 - \delta.$$

Combining with $\Pr(\mathcal{E}_{\text{visit}}^L(t)) \ge 1 - \delta$ yields that for

$$\widetilde{T}_{\text{visit}}^{\text{mix}}(\delta) \;:=\; \max\left\{\, T_{\text{visit}}(\delta)\,,\ \frac{8}{\kappa_{\text{mix}}\, c_{\min}^{\text{lazy}}}\, \log\Big(\frac{n}{\delta}\Big)\right\},$$

and every $t \ge \widetilde{T}_{\text{visit}}^{\text{mix}}(\delta)$ with $t \le \sigma$,

$$\Pr\Big(\,\forall a \in A :\ \varphi_t^{\text{mix}}(a) \;\ge\; \tfrac{\kappa_{\text{mix}}}{2}\, c_{\min}^{\text{lazy}}\, t\,\Big) \;\ge\; 1 - 2\delta.$$

Step 3.5 (Defining the MIX-LEX visitation constant). Define

$$c_{\min}^{\text{mix}} \;:=\; \frac{\kappa_{\text{mix}}}{2}\, c_{\min}^{\text{lazy}} \;=\; \frac{\kappa_{\text{mix}}\,\nu(\alpha - \tau_{\text{mix}})}{8\,\alpha\, n^2}.$$

Then for all $t \ge \widetilde{T}_{\text{visit}}^{\text{mix}}(\delta)$ with $t \le \sigma$,

$$\Pr\Big(\,\forall a \in A :\ \varphi_t^{\text{mix}}(a) \;\ge\; c_{\min}^{\text{mix}}\, t\,\Big) \;\ge\; 1 - 2\delta.$$

We denote this event by $\mathcal{E}_{\text{visit}}^{\text{mix}}(t)$.

Summary. After

$$t \;\geq\; \widetilde{T}^{\text{mix}}_{\text{visit}}(\delta) \;=\; \max\!\left\{ \frac{32\,\alpha^3}{\nu^2(\alpha-\tau_{\text{mix}})^2}\, n^4\, \log\!\Big(\tfrac{n}{\delta}\Big)\,,\; \frac{8}{\kappa_{\text{mix}}}\cdot\frac{4\alpha n^2}{\nu(\alpha-\tau_{\text{mix}})}\, \log\!\Big(\tfrac{n}{\delta}\Big)\right\},$$

MIX-LEX enjoys the uniform lower bound

$$\forall a\in A: \qquad \varphi_t^{\text{mix}}(a) \;\geq\; c_{\min}^{\text{mix}}\, t \quad \text{with probability at least } 1-2\delta,$$

for all $t \leq \sigma$.

Step 4 (Stopping time bound).
Recall the stopping time

$$\sigma \;:=\; \inf\{t\geq 1:\; \widehat{\mu}_t(a)+w_t(a) \;<\; \widehat{\mu}_t(a^*)-w_t(a^*) \;\text{ for all } a\neq a^*\}.$$

On $\mathcal{E}_{\text{conc}}$, a sufficient condition is (47), which translates into the per-arm thresholds (48) and (49). From Step 3, on $\mathcal{E}^{\text{mix}}_{\text{visit}}(t)$ we have $\varphi_t^{\text{mix}}(x) \geq c_{\min}^{\text{mix}} t$ for all x (whenever $t \leq \sigma$). Hence both (48) and (49) are guaranteed if

$$c_{\min}^{\text{mix}}\, t \;\geq\; \max_{a\neq a^*} \frac{16c\,\log(2nT/\delta)}{\Delta(a)^2}.$$

Combining with $t \geq \widetilde{T}^{\text{mix}}_{\text{visit}}(\delta)$, define

$$t_\star^{\text{mix}} \;:=\; \max\!\left\{ \frac{16c}{c_{\min}^{\text{mix}}}\, \max_{a\neq a^*} \frac{\log(2nT/\delta)}{\Delta(a)^2}\;,\; \widetilde{T}^{\text{mix}}_{\text{visit}}(\delta)\right\}. \tag{56}$$

Therefore, on $\mathcal{E}_{\text{conc}} \cap \mathcal{E}^{\text{mix}}_{\text{visit}}(t_\star^{\text{mix}})$ the stopping rule is met and hence $\sigma \leq t_\star^{\text{mix}}$.

Step 4.4 (Expanding constants). Using $c_{\min}^{\text{mix}} = \frac{\kappa_{\text{mix}}\,\nu(\alpha-\tau_{\text{mix}})}{8\,\alpha\,n^2}$,

$$\frac{16c}{c_{\min}^{\text{mix}}} \;=\; \frac{128\,c\,\alpha}{\kappa_{\text{mix}}\,\nu(\alpha-\tau_{\text{mix}})}\, n^2.$$

From Step 3,

$$\widetilde{T}^{\text{mix}}_{\text{visit}}(\delta) \;=\; \max\!\left\{ \frac{32\,\alpha^3}{\nu^2(\alpha-\tau_{\text{mix}})^2}\, n^4\, \log\!\Big(\tfrac{n}{\delta}\Big)\,,\; \frac{32\,\alpha\,n^2}{\kappa_{\text{mix}}\,\nu(\alpha-\tau_{\text{mix}})}\, \log\!\Big(\tfrac{n}{\delta}\Big)\right\}.$$

Step 5 (Exploration regret). *(identical to RA-LEX, with $c_{\min}^{mix}$ and $t_\star^{mix}$)*
By definition,

$$R_{\text{exp}} \;:=\; \sum_{t=1}^{\sigma} \big(\mu(a^*)-\mu(X_t)\big) \;=\; \sum_{a\neq a^*} \varphi_\sigma(a)\,\Delta(a) \;\leq\; \big(1-c_{\min}^{\text{mix}}\big)\,\Delta_{\max}\,\sigma \;\leq\; \big(1-c_{\min}^{\text{mix}}\big)\,\Delta_{\max}\,t_\star^{\text{mix}}.$$

Step 6 (Navigation and expectation). *(identical to RA-LEX)*
After stopping at time σ, MIX-LEX executes the lazy navigation and commits upon hitting a^*. Let τ_{hit} denote the (random) hitting time of a^* in this phase. Each step contributes at most $\Delta_{\max}$ instantaneous regret, so

$$\mathbb{E}[R_{\text{nav}}] \;\leq\; \Delta_{\max}\,\mathbb{E}[\tau_{\text{hit}}].$$

By Lemma 6 (Navigation Regret), under $(\alpha,\beta,\gamma,\nu)$-admissibility

$$\mathbb{E}[\tau_{\text{hit}}] \;\leq\; C_{\text{nav}}\,\frac{n^2\log n}{\nu(\gamma-\alpha\beta)} \quad\Longrightarrow\quad \mathbb{E}[R_{\text{nav}}] \;\leq\; \Delta_{\max}\,C_{\text{nav}}\,\frac{n^2\log n}{\nu(\gamma-\alpha\beta)}.$$

High-probability bound. By Steps 4–5, on $\mathcal{E}_{\text{conc}} \cap \mathcal{E}^{\text{mix}}_{\text{visit}}(t_\star^{\text{mix}})$,

$$R(T) \;=\; R_{\text{exp}} + R_{\text{nav}} \;\leq\; \big(1-c_{\min}^{\text{mix}}\big)\,\Delta_{\max}\,t_\star^{\text{mix}} \;+\; R_{\text{nav}}.$$

By Step 0 and Step 3, $\Pr(\mathcal{E}_{\text{conc}}) \geq 1-\delta$ and $\Pr(\mathcal{E}^{\text{mix}}_{\text{visit}}(t_\star^{\text{mix}})) \geq 1-\delta$, so a union bound gives

$$\Pr\big(\mathcal{E}_{\text{conc}} \cap \mathcal{E}^{\text{mix}}_{\text{visit}}(t_\star^{\text{mix}})\big) \;\geq\; 1-2\delta.$$

Taking expectations conditioned on this event and using the bound on $\mathbb{E}[R_{\text{nav}}]$ above yields the high-probability statement.

Expected regret bound. (identical to RA-LEX with $t_\star^{mix}$ and $c_{\min}^{mix}$)
Decompose by the indicator of the clean event:

$$\mathbb{E}[R(T)] \;=\; \mathbb{E}\big[R(T)\,\mathbf{1}\{\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{mix}}(t_\star^{\text{mix}})\}\big] \;+\; \mathbb{E}\big[R(T)\,\mathbf{1}\{\neg(\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{mix}}(t_\star^{\text{mix}}))\}\big]\,.$$

The first term is at most

$$\big(1 - c_{\min}^{\text{mix}}\big)\,\Delta_{\max}\,t_\star^{\text{mix}} \;+\; \mathbb{E}[R_{\text{nav}}],$$

while the second is at most $T \cdot \Pr\big(\neg(\mathcal{E}_{\text{conc}} \cap \mathcal{E}_{\text{visit}}^{\text{mix}}(t_\star^{\text{mix}}))\big) \;\le\; 2\delta T$, since per-step regret is ≤ 1. Therefore

$$\mathbb{E}[R(T)] \;\le\; \big(1 - c_{\min}^{\text{mix}}\big)\,\Delta_{\max}\,t_\star^{\text{mix}} \;+\; \mathbb{E}[R_{\text{nav}}] \;+\; 2\delta T.$$

Choosing $\delta = 1/T^2$ gives $2\delta T \le 2/T$ and $\log(2nT/\delta) = \log(2nT^3) = \Theta(\log(nT))$, which is the advertised expected-regret corollary. □