A Reduction Algorithm for Markovian Contextual Linear Bandits

Kaan Buyukkalayci [1] **Osama Hanna** [2] **Christina Fragouli** [1]

Abstract

Recent work shows that when contexts are drawn i.i.d., linear contextual bandits can be reduced to single-context linear bandits. This "contexts are cheap" perspective is highly advantageous, as it allows for sharper finite-time analyses and leverages mature techniques from the linear bandit literature, such as those for misspecification and adversarial corruption. Motivated by applications with temporally correlated availability, we extend this perspective to Markovian contextual linear bandits, where the action set evolves via an exogenous Markov chain. Our main contribution is a reduction that applies under uniform geometric ergodicity. We construct a stationary surrogate action set to solve the problem using a standard linear bandit oracle, employing a delayed-update scheme to control the bias induced by the non-stationary conditional context distributions. We further provide a phased algorithm for unknown transition distributions that learns the surrogate mapping online. In both settings, we obtain a high-probability worst-case regret bound matching that of the underlying linear bandit oracle, with only lower-order dependence on the mixing time.

1. Introduction

Contextual linear bandits form a central paradigm in online learning due to their ability to elegantly formalize a broad range of sequential decision-making problems, where context (side information) serves a crucial role in guiding decisions. Their modeling versatility has enabled extensive real-world adoption, driving advances in applications such as autonomous systems, personalized online recommendations, online controlled experiments, and diagnostic tools in healthcare (Chacun et al., 2024; Wakayama & Ahmed, 2023; Varatharajah & Berry, 2022; Zhang & Yuan, 2023; Bojinov & Gupta, 2022). Motivated by practical applications, in this paper we develop algorithms and regret bounds for what we term Context-Markovian Linear Bandits, a setting in which temporal dependence in the context process introduces fundamental challenges that do not arise in the i.i.d setting.

Context-Markovian linear bandits extend the classical framework by allowing contexts to evolve according to a Markov chain. In the traditional setting, each round assumes that the learner observes a context drawn independently from some distribution, representing side information such as environmental conditions, user preferences, or patient characteristics. In many practical scenarios, however, the context distribution is better modeled as dynamic: the current location and sensor readings of an autonomous robot may depend on its previous states, user behavior in recommendation and online controlled experimentation systems may exhibit temporally correlated interests and population drift, and the progression of a cancer cell may evolve over time. The classical i.i.d. assumption fails to capture such temporal dependencies, which can significantly influence the decision-making process.

Our work focuses on understanding how ergodic Markovian correlations in the context process affect worst-case regret guarantees. In the i.i.d. contextual linear bandit setting, the tightest known upper bounds were obtained by Hanna et al. (2023b), who showed that by carefully constructing a surrogate expected action set, the problem can be reduced to an ordinary linear bandit instance. This reduction yields regret bounds of order $O(d\sqrt{T \log T})$ for both known and unknown context distributions, where T is the time horizon and d is the dimension of the unknown parameter, matching the classical $\Omega(d\sqrt{T})$ lower bound for linear bandits up to logarithmic factors. Importantly, this framework also opened the door to tackling more complex variants of the problem by demonstrating that the presence of contexts need not inherently worsen regret guarantees. However, this reduction crucially relies on independence across rounds and breaks down in the presence of Markovian dependence, where induced bias invalidates standard linear bandit analyses.

In this work, we take a step beyond the i.i.d. assumption and establish that analogous reduction-based guarantees extend

[1]University of California, Los Angeles [2]Meta, Superintelligence Lab. Correspondence to: Kaan Buyukkalayci <kaanbkalayci@g.ucla.edu>.

Preprint. February 11, 2026.

to the contextual linear bandit setup where contexts evolve with an ergodic Markovian process. A key technical insight underlying our approach is the use of delayed feedback: by introducing a delay in the rewards supplied to the underlying linear bandit algorithm, we allow the context process sufficient time to mix, which reduces the bias in the actions observed by the linear bandit oracle and enables the linear bandit to operate on effectively unbiased feedback despite temporal dependence in the contexts. Controlling this vanishing bias and showing that it can be safely absorbed by UCB-style and elimination-based linear bandit algorithms requires a delicate analysis, as the bias interacts nontrivially with confidence bounds and arm elimination criteria.

Our contributions can be summarized as follows.

(i) Reduction Framework: We derive a reduction that establishes an equivalence between context-Markovian linear bandits and linear bandits with a carefully crafted surrogate action set.

(ii) Known Distribution: When the stationary distribution of the context process is known, we prove a high-probability regret bound of order $O\big(d\sqrt{T\log T}\big)$ for sufficiently large d, matching the best known rates for standard linear bandit algorithms.

(iii) Unknown Distribution: When the stationary distribution is unknown, we provide a phased elimination algorithm that learns the surrogate mapping online. We establish a high-probability regret bound of $O\Big(d\sqrt{T\log T/(1-\beta)}\Big)$ for sufficiently large d, where β is a parameter that governs the mixing rate.

(iv) Optimality: In both settings, our bounds match the best known results for linear bandits up to mixing-time factors, demonstrating that temporal dependence does not incur a regret penalty beyond constant factors.

(v) Empirical Validation: We further provide numerical experiments illustrating favorable performance compared to LinUCB.

The paper is organized as follows. Section 2 reviews related work, and Section 3 introduces the problem setup and notation. An overview of the provided regret bounds is given in Section 4. Sections 5 and 6 address the cases of known and unknown stationary context distributions, respectively. Numerical results are presented in Section 7, and Section 8 concludes the paper.

2. Related Work

Linear and contextual linear bandits. The ordinary (single-context) linear bandit problem is known to admit worst-case regret of order $O(d\sqrt{T\log T})$ (e.g., Lattimore et al., 2020). Such rates are also achieved with high probability by optimism-based algorithms, such as OFUL, and are minimax-optimal up to logarithmic factors according to existing lower bounds (e.g., Dani et al., 2008; Abbasi-Yadkori et al., 2011; Lattimore & Szepesvári, 2020). In linear contextual bandits, each round reveals a context-dependent action set, and rewards are linear in an unknown parameter; canonical algorithms such as LinUCB/OFUL and linear Thompson sampling achieve $\tilde{O}(d\sqrt{T})$ regret under sub-Gaussian noise (e.g., Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013; Lattimore & Szepesvári, 2020). While these worst-case rates match those of linear bandits in order, reductions from stochastic contextual models to linear bandits are highly advantageous because they inherit sharper finite-time guarantees from linear-bandit solvers (often improving constants and logarithmic factors), simplify the algorithms and analysis, and because they enable the use of linear-bandit techniques in settings where the corresponding contextual analyses are substantially more involved, such as model misspecification, batched interaction, and robustness to adversarial corruptions (Hanna et al., 2023b).

Asymptotically optimal exploration in contextual linear bandits. Beyond worst-case optimality, a substantial line of work aims to achieve asymptotically optimal (problem-dependent) regret by designing exploration policies guided by regret lower bounds, often via primal–dual or saddle-point reformulations. In the contextual linear bandit setting, Hao et al. (2020) develop exploration that adapts to benign context distributions and improves asymptotic behavior compared to standard optimistic methods. Building on lower-bound–driven optimization viewpoints and online learning ideas, Tirinzoni et al. (2020) propose SOLID, an asymptotically optimal algorithm with improved computational and finite-time properties under i.i.d. contexts. These works are complementary to ours: they refine exploration when contexts are independent, whereas our focus is on temporal dependence in the context sets.

Stochastic contextual linear bandits with random action sets and reduction. The closest starting point for our results is the recent reduction framework of Hanna et al. (2023b), that study stochastic contextual linear bandits where each context is a *random set of actions*. They show that, when the context distribution is known, one can reduce the problem exactly to an ordinary linear bandit by constructing a surrogate action set from the expected argmax action. When the context distribution is unknown, they reduce to a sequence of *misspecified* linear bandit instances and retain $\tilde{O}(d\sqrt{T})$-type worst-case regret. The follow up work in Hanna et al. (2023a), proposed a computationally efficient implementation of the reduction algorithms. Our work extends this reduction to a strictly more dependent regime in which the random action sets evolve according to a Markov chain rather than i.i.d. sampling.

Adversarial losses, stochastic availability, and reduction-based approaches. A different but conceptually related thread studies *stochastic decision sets* (sleeping/stochastically available actions) under adversarial losses, beginning with online combinatorial optimization with stochastic decision sets (Neu & Valko, 2014). In adversarial linear contextual bandits where per-round action sets are drawn from a fixed distribution, recent work has focused on whether one can bypass access to a context simulator while retaining near-optimal regret and polynomial-time efficiency (Liu et al., 2023; Olkhovskaya et al., 2023). Most recently, van Erven et al. (2025) extends (Hanna et al., 2023b)'s reduction to reduce adversarial linear contextual bandits with stochastic action sets to misspecification-robust adversarial linear bandits with fixed action sets, obtaining $\mathrm{poly}(d)\sqrt{T}$-type regret in polynomial time without knowing the context distribution. While our setting is stochastic (sub-Gaussian) rather than adversarial in rewards, these results reinforce the broader theme that *random availability* can often be handled via carefully constructed surrogates; our contribution is to show this remains true even under Markovian dependence.

Bandits and reinforcement learning with Markov structure. There is extensive literature on bandits with Markovian dynamics, including rested/restless Markovian bandits where each arm's state evolves according to a Markov chain (e.g., Tekin & Liu, 2012; Ortner et al., 2012; Weber & Weiss, 1990; Bertsimas & Niño-Mora, 2000) and variants with hidden Markov states and side information (e.g., Yemini et al., 2019). These models differ fundamentally from ours: in our setting rewards remain linear and the dependence enters through the *context/action set process* itself, which enables a reduction to linear bandits rather than requiring index policies or MDP-style value learning. On the RL side, structured exploration in MDPs (e.g., Ok et al., 2018) studies how known structure in transitions/rewards can reduce exploration costs; our work instead leverages mixing of the observed context process to control bias in a reduction-based bandit algorithm.

Our setting can also be interpreted as a linear Markov decision process with *exogenous actions*: the context process $(\mathcal{A}_t)$ evolves autonomously according to a Markov kernel P, independent of the learner's actions, while rewards are linear in the selected action features. In principle, this allows the use of algorithms for linear MDPs, such as LSVI–UCB (Jin et al., 2020). However, generic MDP regret bounds are poorly suited to this regime, as they incur unnecessary pessimism by accounting for transition uncertainty that is irrelevant when the dynamics are exogenous. By avoiding a full MDP treatment and instead reducing the problem to a linear bandit with a carefully controlled bias via delayed feedback, our analysis recovers regret guarantees that match the best known rates for linear bandits whenever the context process mixes sufficiently fast.

Latent or partially observed context dynamics and non-stationarity. Another related line considers hidden or partially observed state processes that generate contexts and/or rewards. For example, Nelson et al. (2022) "linearize" contextual bandits with latent state dynamics via online EM/HMM ideas, and Zeng et al. (2024) study partially observed, temporally correlated contexts with linear payoffs using filtering and system-identification techniques. These works address *partial observability*; by contrast, our contexts are fully observed but temporally dependent. Finally, non-stationary contextual bandits (e.g., Luo et al., 2018) target drifting distributions and optimize dynamic or switching benchmarks; our Markovian model is stationary but dependent, and our guarantees quantify how mixing controls the reduction error.

3. Setup and Notation

We use the shorthand notation $[i] = \{1, 2, \ldots, i\}$ for any $i \in \mathbb{N}$ with $i > 0$, where $\mathbb{N}$ denotes the set of natural numbers. We write $y = O(f(x))$ if there exist constants $c > 0$ and $x_0 \in \mathbb{R}$ such that $y \leq cf(x), \quad \forall x > x_0$. We use the notation $\tilde{O}(f(x))$ to suppress logarithmic factors. We use $\|\cdot\|_2$ to denote the Euclidean norm, and $\mathrm{TV}(\cdot, \cdot)$ to denote the total variation distance, which is defined by

$$\mathrm{TV}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \frac{1}{2} \int |dP - dQ|.$$

where P and Q are probability measures on a common measurable space $(\Omega, \mathcal{F})$. A *δ-net* of a set $\mathcal{A} \subseteq \mathbb{R}^d$ (with respect to the ℓ_2 norm) is a set $\mathcal{B}$ such that for every $a \in \mathcal{A}_t$ there exists $b \in \mathcal{B}$ with $\|a - b\|_2 \leq \delta$ where $\delta > 0$ and where $\mathbb{R}$ denotes the set of real numbers.

At each round $t \in [T]$, the learner plays an action a_t and observes a reward

$$r_t = \langle a_t, \theta_\star \rangle + \eta_t$$

where the action $a_t \in \mathcal{A}_t$ is chosen from the available action set $\mathcal{A}_t$, which is termed the *context*. $\theta_\star \in \Theta \subseteq \mathbb{R}^d$ is an unknown parameter, and η_t is sub-Gaussian noise satisfying $\mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0$ and $\mathbb{E}[\exp(\lambda \eta_t) \mid \mathcal{F}_t] \leq \exp(\lambda^2/2)$ for all $\lambda \in \mathbb{R}$. Here, $\mathcal{F}_t = \sigma\{\mathcal{A}_1, a_1, r_1, \ldots, \mathcal{A}_t, a_t\}$ denotes the filtration representing the history up to time t, with $\sigma(X)$ being the σ-algebra generated by X. We assume each context set $\mathcal{A}_t$ is compact. Whenever the maximizer $\arg\max \langle a, \theta \rangle$ is not unique, ties are broken according to a fixed measurable rule, for any $a \in \mathcal{A}_t$, $\theta \in \Theta$. At every round t, the context set $\mathcal{A}_t$ is revealed to the learner before selecting an action. We adopt the standard boundedness assumptions $\|a\|_2 \leq 1$ for all $a \in \mathcal{A}_t$ and $\|\theta\|_2 \leq 1$ for all

$\theta \in \Theta$ almost surely. The goal of the learner is to minimize regret defined as

$$R_T = \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle a, \theta_\star \rangle - \langle a_t, \theta_\star \rangle$$

In addition, we consider the setting in which the sequence of context sets $(\mathcal{A}_t)_{t=1}^T$ evolves as a Markov chain on a measurable state space $(\mathsf{S}, \mathcal{F}_\mathsf{S})$, where S is a Polish space and the dynamics are governed by a transition kernel $P : \mathsf{S} \times \mathcal{F}_\mathsf{S} \to [0, 1]$. For any probability measure ν on S, we define its pushforward measure under P by

$$(P\nu)(B) \triangleq \int_\mathsf{S} P(a, B)\, d\nu(a), \qquad \forall B \in \mathcal{F}_\mathsf{S},$$

where $P(a, B)$ denotes the probability that $\mathcal{A}_{t+1} \in B$ given $\mathcal{A}_t = a$, and denote by $P^t\nu$ the t-step pushforward measure obtained by t successive applications of P. The Markov chain is assumed to be uniformly geometrically ergodic, with unique stationary distribution π. We formally state this classical assumption as follows.

Definition 3.1 (Uniform Geometric Ergodicity)**.** The context process $(\mathcal{A}_t)_{t\geq 1}$ is uniformly geometrically ergodic if there exist constants $C_{\text{mix}} < \infty$ and $\beta \in (0, 1)$, depending only on the transition kernel P, such that for any initial distribution μ on S and all $t \geq 0$:

$$\text{TV}(P^t\mu, \pi) \;\leq\; C_{\text{mix}}\, \beta^t, \tag{1}$$

where π is the unique stationary distribution.

While our results apply to general (possibly infinite) state spaces, in the finite-state case irreducibility and aperiodicity already imply a unique stationary distribution and uniform geometric ergodicity, and therefore this assumption does not impose an additional restriction. We also note that although the total variation distance in (1) vanishes when the chain is initialized from the stationary distribution, our bounds hold uniformly over all initial distributions. In particular, the analysis reveals no substantive advantage to stationary initialization.

4. Summary of Regret Guarantees

We propose two provably efficient algorithms for the context–Markovian linear bandits problem. Under the assumption that the context process is uniformly geometrically ergodic, the algorithms achieve a high-probability regret bound of the same order as the best known bounds for standard (single-context) linear bandits, up to a multiplicative factor $O\Big(\sqrt{1/(1-\beta)}\Big)$, where β denotes the convergence rate in (1). This bound is obtained via a simple reduction-based approach first outlined in (Hanna et al., 2023b). In particular, it implies that the best known regret bound can be recovered up to order when the mixing time is not excessively slow. Although our analysis focuses on Markovian contexts, the approach extends naturally to all ergodic processes governing the context distribution, provided that convergence to the stationary distribution is sufficiently fast.

Algorithm 1 Reduction from Markovian contexts to Single Context under Known Transitions

Input: Confidence Parameter δ, Single-context linear bandit algorithm Λ, Feedback delay τ
for $t = 1 : \tau$ **do**
  Pick θ_t randomly such that $g_\pi(\theta_t) \in \mathcal{X}_\pi$
  Play $a_t = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_t \rangle$, obtaining r_t.
end for
for $t = (\tau + 1) : T$ **do**
  Let the action that Λ selects be $g_\pi(\theta_t) \in \mathcal{X}_\pi$ after observing action-reward pairs $(g_\pi(\theta_1), r_1), \ldots, (g_\pi(\theta_{t-\tau-1}), r_{t-\tau-1})$
  Play $a_t = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_t \rangle$ and receive reward r_t.
  Provide $(g_\pi(\theta_{t-\tau}), r_{t-\tau})$ to Λ.
end for

We show that both when (i) the stationary distribution of the context process is known and when (ii) it is unknown, a linear bandit oracle can be employed to obtain high probability regret bounds of order $O\big(d\sqrt{T \log T}\big)$ and $O\Big(d\sqrt{(T \log T)/(1-\beta)}\Big)$, respectively, when the associated term dominates.

5. Known Stationary Distribution

We construct a surrogate action set $\mathcal{X}_\pi = \{g_\pi(\theta) : \theta \in \Theta\}$, where $g_\pi(\theta) := \mathbb{E}_{\mathcal{A}\sim\pi}[\arg\max_{a \in \mathcal{A}} \langle a, \theta \rangle]$ is the expected greedy action under the stationary distribution π. This set is supplied to a linear bandit algorithm Λ, which treats $\mathcal{X}_\pi$ as its fixed arm set.

The interaction proceeds as follows: At round t, Λ selects a proxy action $g_\pi(\theta_t) \in \mathcal{X}_\pi$. This proxy is *not* played. Instead, the learner observes the current Markovian context $\mathcal{A}_t$ and plays the greedy action $a_t := \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_t \rangle$, receiving reward r_t. We enforce a *delayed feedback*: the pair $(g_\pi(\theta_{t-\tau}), r_{t-\tau})$ is revealed to Λ only after a delay of τ rounds.

The delay τ allows the Markov chain to mix. While $\mathbb{E}[r_t|\theta_t]$ corresponds to the value of the specific context $\mathcal{A}_t$, the linear bandit oracle Λ expects rewards corresponding to the stationary average $g_\pi(\theta_t)$. By delaying the update, we ensure that the dependence between the action selection (based on history up to $t - \tau$) and the specific context $\mathcal{A}_t$ decays. This effectively reduces the discrepancy between the observed reward and the stationary expectation to a

vanishing bias, which standard linear bandit algorithms can tolerate. We formalize this in the following proposition.

Proposition 1. When $\tau = \lceil c_\tau \log T/(1-\beta)\rceil$, where $c_\tau > 1$ is a fixed constant, the reward r_t can be written as

$$r_t = \langle g_\pi(\theta_t), \theta_\star\rangle + \Delta_t + \eta'_t,$$

where $|\Delta_t| \le 2C_{\text{mix}}T^{-c_\tau}$ almost surely, $\mathbb{E}[\eta'_t \mid \mathcal{F}'_t] = 0$ and $\mathbb{E}[\exp(\lambda\eta'_t) \mid \mathcal{F}'_t] \le \exp((17/2)\lambda^2)$ for all $\lambda \in \mathbb{R}$, with $\mathcal{F}'_t = \sigma\{\theta_{\tau+1}, r_{\tau+1}, \dots, \theta_{t-\tau}\}$ denoting the filtration of the delayed history up to time t.

Proof. Let ρ be a probability distribution over contexts and let $\theta \in \Theta$. Define

$$g_\rho(\theta) := \mathbb{E}_{\mathcal{A}\sim\rho}\Big[\arg\max_{a\in\mathcal{A}}\langle a, \theta\rangle \mid \theta\Big].$$

We first show that, for any fixed θ, the Euclidean distance between the corresponding $g_\rho(\theta)$ induced by any two distributions over the contexts is upper bounded by the total variation distance between these two distributions.

Lemma 5.1. *For any two probability measures ρ, ρ' on the measurable space of contexts $(\mathsf{S}, \mathcal{F}_\mathsf{S})$ and any $\theta \in \Theta$. It holds that*

$$\|g_\rho(\theta) - g_{\rho'}(\theta)\|_2 \;\le\; 2\,\mathrm{TV}(\rho, \rho'),$$

We defer the proof of this lemma to Appendix A.

Letting $a_t := \arg\max_{a\in\mathcal{A}_t}\langle a, \theta_t\rangle$, define,

$$b_t := \langle a_t, \theta_\star\rangle - \langle g_\pi(\theta_t), \theta_\star\rangle, \qquad \Delta_t := \mathbb{E}[\,b_t \mid \mathcal{F}'_t\,]$$

For $t > \tau$,

$$r_t = \langle a_t, \theta_\star\rangle + \eta_t = \langle g_\pi(\theta_t), \theta_\star\rangle + b_t + \eta_t.$$

Let $\rho_t(\cdot) := \mathbb{P}(\mathcal{A}_t \in \cdot \mid \mathcal{F}'_t)$ denote the conditional distribution over the contexts at time t with respect to $\mathcal{F}'_t$. Then

$$\begin{aligned}\Delta_t &= \mathbb{E}\left[\left\langle \arg\max_{a\in\mathcal{A}_t}\langle a, \theta_t\rangle - g_\pi(\theta_t), \theta_\star\right\rangle \,\middle|\, \mathcal{F}'_t\right] \\ &= \left\langle \mathbb{E}\left[\arg\max_{a\in\mathcal{A}_t}\langle a, \theta_t\rangle \,\middle|\, \mathcal{F}'_t\right] - g_\pi(\theta_t), \theta_\star\right\rangle \\ &= \langle g_{\rho_t}(\theta_t) - g_\pi(\theta_t), \theta_\star\rangle,\end{aligned}$$

Using $\|\theta_\star\|_2 \le 1$ and Cauchy-Schwarz inequality,

$$|\Delta_t| \le \|g_{\rho_t}(\theta_t) - g_\pi(\theta_t)\|_2 \le 2\,\mathrm{TV}(\rho_t, \pi).$$

where the second equality follows from Lemma 5.1.

For all measurable $A \in \mathcal{F}_\mathsf{S}$, since $\mathcal{F}'_t$ is measurable with respect to the history of the context process up to time $t-\tau$, the Markov property implies,

$$\mathbb{P}(\mathcal{A}_t \in A \mid \mathcal{A}_{t-\tau}, \mathcal{F}'_t) = \mathbb{P}(\mathcal{A}_t \in A \mid \mathcal{A}_{t-\tau}) \tag{2}$$

Let

$$\rho_t(A) = \mathbb{E}[P^\tau(\mathcal{A}_{t-\tau}, A) \mid \mathcal{F}'_t]\,.$$

and

$$\mu_{t-\tau}(\cdot) := \mathbb{P}(\mathcal{A}_{t-\tau} \in \cdot \mid \mathcal{F}'_t).$$

Then $\rho_t = \mu_{t-\tau}P^\tau$.

Uniform geometric ergodicity gives, almost surely,

$$\mathrm{TV}(\rho_t, \pi) = \mathrm{TV}(\mu_{t-\tau}P^\tau, \pi) \le C_{\text{mix}}\beta^\tau \tag{3}$$

Therefore it holds almost surely,

$$|\Delta_t| \le 2\mathrm{TV}(\rho_t, \pi) \le 2C_{\text{mix}}\beta^\tau,$$

Since $-\log\beta \ge 1-\beta$ for $0 < \beta < 1$,

$$\begin{aligned}|\Delta_t| &\le 2C_{\text{mix}}\beta^\tau \le 2C_{\text{mix}}\exp\left(-\tau(1-\beta)\right) \\ &\le 2C_{\text{mix}}\exp\left(-c_\tau\log T\right) = 2C_{\text{mix}}T^{-c_\tau}\end{aligned}$$

Setting

$$\eta'_t := \eta_t + b_t - \Delta_t.$$

By construction,

$$\mathbb{E}[\eta'_t \mid \mathcal{F}'_t] = \mathbb{E}[\eta_t \mid \mathcal{F}'_t] + \mathbb{E}[b_t \mid \mathcal{F}'_t] - \Delta_t = 0,$$

Finally, η'_t is conditionally sub-Gaussian since η_t is conditionally sub-Gaussian given $\mathcal{F}_t \supseteq \mathcal{F}'_t$ and $b_t - \Delta_t$ is $\mathcal{F}'_t$-conditionally mean-zero and bounded. Indeed, $\|a_t\|_2 \le 1$ and $\|g_\pi(\theta_t)\|_2 \le 1$ imply $|b_t| = |\langle a_t - g_\pi(\theta_t), \theta_\star\rangle| \le 2$, and hence $|b_t - \Delta_t| \le 4$. By Hoeffding's lemma, $b_t - \Delta_t$ is conditionally sub-Gaussian with proxy variance at most 16, and since η_t is conditionally sub-Gaussian with proxy variance 1, their sum η'_t is conditionally sub-Gaussian with proxy variance at most 17, yielding $\mathbb{E}[\exp(\lambda\eta'_t) \mid \mathcal{F}'_t] \le \exp((17/2)\lambda^2)$ for all $\lambda \in \mathbb{R}$. □

Since $c_\tau > 1$, the bias term Δ_t satisfies $\sup_{t\le T}|\Delta_t| = o(1)$ as $T \to \infty$, and therefore does not affect the regret order of standard linear bandit algorithms. In particular, for UCB-style methods such as (Abbasi-Yadkori et al., 2011) specifies, the contribution of Δ_t results only in a vanishing additive term in the cumulative regret. We give a more detailed analysis of this behavior in Appendix B.1. For elimination-style algorithms, such as the phased elimination algorithm of Lattimore et al. (2020) used in Section 6, the effect of Δ_t can be absorbed by enlarging the elimination radius to account for a misspecification level $\varepsilon_T = 2C_{\text{mix}}T^{-c_\tau}$, as in Remark D.1 of Lattimore et al. (2020), or by any larger

vanishing radius when C_{mix} is unknown, so that the final regret stays in the order of $O(d\sqrt{T \log T})$.

It remains to bound the difference between the regret incurred by the linear bandit algorithm Λ and that incurred by Algorithm 1.

Theorem 5.2. *Consider a context-Markovian linear bandit instance M as defined in Section 3, and let L denote the associated linear bandit instance with action set $\mathcal{X}_\pi = \{g_\pi(\theta) : \theta \in \Theta\}$. For any linear bandit algorithm Λ, the regret of Algorithm 1 satisfies, with probability at least $1-\delta$,*

$$\left|R_T^\Gamma(M) - R_T^\Lambda(L)\right| \le c\left(\sqrt{T\tau \log \frac{\tau}{\delta}} + \tau\right)$$

where τ is set to $\tau = \lceil c_\tau \log T/(1-\beta)\rceil$. $R_T^\Lambda(L)$ is the regret of Λ on L, $R_T^\Gamma(M)$ is the regret of Algorithm 1 on M, $\beta \in (0,1)$ is the geometric mixing rate of the underlying Markov chain, and $c > 0$ is a universal constant.

Proof Sketch. We express the regret difference as a sum of per-round deviations between the stationary feature map $g_\pi(\cdot)$ and the greedy action induced by the realized context, together with an initial τ-round truncation. The analysis proceeds via a τ-spaced blocking argument that yields martingale difference sequences, to which the Azuma–Hoeffding inequality is applied. The complete proof is deferred to Appendix B.2. □

This result implies that Algorithm 1 incurs an additional regret difference of order $O\left(\sqrt{T\tau \log(\tau T)} + \tau\right)$ with probability at least $1 - 1/T$. Since the tightest known high-probability regret bound for linear bandit algorithms is $O(d\sqrt{T \log T})$, it follows that whenever the mixing-dependent term is dominated by the linear bandit regret (e.g., in sufficiently fast-mixing regimes), Algorithm 1 achieves the same regret order $O(d\sqrt{T \log T})$ with high probability. Next, we show that in expectation the regret of Algorithm 1 differs from that of the underlying linear bandit oracle by a lower order additive mixing-dependent term.

Corollary 5.3. *Consider a context-Markovian linear bandit instance M as defined in Section 3, and let L denote the associated linear bandit instance with action set $\mathcal{X}_\pi = \{g_\pi(\theta) : \theta \in \Theta\}$. The regret of Algorithm 1 satisfies,*

$$\left|\mathbb{E}\left[R_T^\Gamma(M)\right] - \mathbb{E}\left[R_T^\Lambda(L)\right]\right| \;\le\; 2\tau \;+\; 4T\, C_{\text{mix}} \beta^\tau$$

where τ is set to $\tau = \lceil c_\tau \log T/(1-\beta)\rceil$ with $c_\tau > 1$ thus $\beta^\tau \le T^{-c_\tau}$, and $R_T^\Lambda(L)$ is the regret of Λ on the linear bandit instance L, $R_T^\Gamma(M)$ is the regret of Algorithm 1 on the context-Markovian bandit instance M.

Proof Sketch. Expressing the regret difference in a manner analogous to Theorem 5.2, we use a similar result to Proposition 1 to characterize the bias induced by mixing as an additive term. The full proof is provided in Appendix B.3. □

Corollary 5.3 shows that, in expectation, the dependence on the mixing rate enters additively through the delay parameter τ, whereas high-probability control of the realized regret difference $R_T^\Gamma(M) - R_T^\Lambda(L)$ necessarily incurs a multiplicative dependence on the mixing rate β. The underlying reason is that, after algebraic cancellation, this difference involves additive functionals of the Markov context process of the form $S_T = \sum_{t=1}^T \big(h(\mathcal{A}_t) - \mathbb{E}_\pi[h(\mathcal{A})]\big)$ for bounded functions h. For a geometrically ergodic Markov chain with mixing parameter β, the autocovariance decays as $\mathrm{Cov}(h(\mathcal{A}_0), h(\mathcal{A}_k)) = O(\beta^k)$, which implies $\mathrm{Var}(S_T) = \Theta(T/(1-\beta))$. As a consequence, Bernstein–Freedman–type concentration inequalities for additive functionals of Markov chains yield $|S_T| = O\left(\sqrt{T \log(1/\delta)/(1-\beta)}\right)$ with probability at least $1-\delta$ (Paulin, 2015), explaining why a $(1-\beta)^{-1/2}$ factor on the $\sqrt{T}$ scale is unavoidable when controlling the realized regret difference with high probability, despite being avoidable in expectation in the known stationary distribution case. The corresponding concentration inequality is stated explicitly in Section 6.

6. Unknown Stationary Distribution

We now consider the setting in which the stationary distribution π of the Markov context process is unknown. In this case, the surrogate map $g_\pi(\cdot)$ is unavailable and must be estimated online. We propose Algorithm 2, which proceeds in epochs, maintaining at each phase m an empirical surrogate map $g^{(m)}(\cdot)$ constructed from past observations and defined only on a finite $1/T$-net $\Theta' \subseteq \Theta$. The induced surrogate action set $\mathcal{X}_m = \{g^{(m)}(\theta) : \theta \in \Theta'\}$ is supplied to a misspecification-robust linear bandit algorithm Λ_{ϵ_m}, where ϵ_m accounts for the approximation error between $g^{(m)}$ and the stationary surrogate g_π. As the number of samples used to construct $g^{(m)}$ increases across epochs, this misspecification level decays at a controlled rate, allowing Λ_{ϵ_m} to operate with diminishing bias.

Theorem 6.1. *Consider Algorithm 2 with epoch lengths $T_m = t^{(m+1)} - t^{(m)} = \tau + 2^{m-1}$ and delay $\tau = \lceil c_\tau \log T/(1-\beta)\rceil$. Let Λ_{ϵ_m} be the PE algorithm described in Lattimore et al. (2020). Then, the regret of Algorithm 2 $R_T^\Gamma(M)$ on the context-Markovian instance M specified in Section 3 satisfies with probability at least $1 - 1/T$,*

$$R_T^\Gamma(M) \le C\left(d\sqrt{\frac{T \log T}{1-\beta}}\right),$$

for sufficiently large d [1], where $C > 0$ is a universal constant.

[1] The full bound is provided in Appendix C.4

Algorithm 2 Reduction from Markovian Contexts to Single Context (Unknown Stationary Distribution)

Input: δ, phase lengths $\{t^{(m)}\}_{m=1}^{M+1}$, misspecification-robust linear bandit Λ_ϵ, delay τ, $1/T$-net Θ' over Θ
Initialize: $g^{(1)} : \Theta' \to \mathbb{R}^d$ randomly, $\epsilon_1 = 1$, $\mathcal{X}_1 = \{g^{(1)}(\theta) : \theta \in \Theta'\}$
for $m = 1 : M$ **do**
  for $t = (t^{(m)} + 1) : (t^{(m)} + \tau)$ **do**
    Pick θ_t randomly such that $g^{(m)}(\theta_t) \in \mathcal{X}_\pi$
    Play $a_t = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_t \rangle$, obtaining r_t.
  end for
  for $t = (t^{(m)} + \tau + 1) : t^{(m+1)}$ **do**
    Let $g^{(m)}(\theta_t) \in \mathcal{X}_m$ be the arm selected by Λ_{ϵ_m} after observing rewards $r_{t^{(m)}+1}, \ldots, r_{t-\tau-1}$
    Play $a_t = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_t \rangle$, observe r_t
    Feed $(g^{(m)}(\theta_{t-\tau}), r_{t-\tau})$ to Λ_{ϵ_m}
  end for
  $g^{(m+1)}(\cdot) = \frac{1}{t^{(m+1)}} \sum_{t=1}^{t^{(m+1)}} \arg\max_{a \in \mathcal{A}_t} \langle a, \cdot \rangle$
  $\mathcal{X}_{m+1} = \{g^{(m+1)}(\theta) : \theta \in \Theta'\}$
  Pick ϵ_{m+1} as the upper bound stated in Lemma 6.2
end for

Theorem 6.1 quantifies the price of not knowing the stationary distribution π: estimating $g_\pi(\cdot)$ from a single Markovian trajectory requires high-probability uniform control of Markov additive functionals, whose fluctuations scale as the aforementioned $\sqrt{\mathrm{Var}(S_T)} = \Theta(\sqrt{T/(1-\beta)})$ under uniform geometric ergodicity, so the regret scales as the best known linear bandit rate with an additional inflation factor of $(1-\beta)^{-1/2}$ induced by the Markovian dependence.

We establish the claim of Theorem 6.1 via three intermediate lemmas, which directly yield the stated regret bound.

1. We characterize, for each epoch, the model misspecification induced by approximating $g_\pi(\cdot)$ with $g^{(m+1)}(\cdot)$. (Lemma 6.2).

2. We bound, with high probability, the difference between the cumulative regret of the Λ_{ϵ_m} instances and that of Algorithm 2. (Lemma 6.3)

3. We bound, with high probability, the cumulative regret incurred by all Λ_{ϵ_m} instances invoked within Algorithm 2 (Lemma 6.4).

Before stating our first lemma in this section, we first provide a relevant result from Paulin (2015).

Concentration of Additive Functionals in Uniformly Ergodic Markov Chains. (Paulin, 2015) Let $(A_t)_{t \geq 1}$ be a uniformly ergodic Markov chain parametrized by $(C_{\mathrm{mix}}, \beta)$ on a Polish state space S and let $h : \mathsf{S} \to \mathbb{R}$ be a function that satisfies $|h(x)| \leq 1$, $\forall x \in \mathsf{S}$. Then for any $T \in \mathbb{N}$, any $\delta \in (0,1)$, and for any initial distribution of A_1, it holds with probability at least $1-\delta$,

$$\left| \sum_{t=1}^{T} (h(A_t) - \mathbb{E}[h(A_t)]) \right| \leq \sqrt{18\, T\, \frac{\log(4C_{\mathrm{mix}})}{1-\beta} \log\left(\frac{2}{\delta}\right)}. \tag{4}$$

This result follows from Corollary 2.10, together with Remark 2.11 in Paulin (2015), specialized to $\|h\|_\infty \leq 1$.

Lemma 6.2. *For each $m \in [M]$, and $\forall \theta \in \Theta', \forall \theta' \in \Theta$, with probability at least $1 - \delta/M$,*

$$\left| \langle g_\pi(\theta), \theta' \rangle - \left\langle g^{(m)}(\theta), \theta' \right\rangle \right| \leq \sqrt{\frac{36\, \log(4C_{\mathrm{mix}})}{(1-\beta)\, t^{(m)}} \log\left(\frac{2M|\Theta'|}{\delta}\right)} + \frac{C_{\mathrm{mix}}}{(1-\beta)\, t^{(m)}} + \frac{2}{T}.$$

Proof Sketch. After fixing $\theta \in \Theta'$ and $\theta' \in \Theta$, we decompose the error $\langle g_\pi(\theta) - g^{(m)}(\theta), \theta' \rangle$ into a stochastic fluctuation term, in which we use the borrowed result from Paulin (2015), and a bias term induced by the nonstationarity of the Markov context process. We extend the bound from Θ' to all $\theta' \in \Theta$ via a covering argument, introducing only a negligible $O(1/T)$ error. The full proof is covered in Appendix C.1 □

Lemma 6.3. *Let Λ_ϵ be an algorithm for linear bandits with ϵ misspecification, and let $R_{T_m}^{\Lambda_{\epsilon_m}}(L_m)$ be the regret incurred by Λ_{ϵ_m} in Algorithm 2 in epoch m on the misspecified linear bandit instance L_m with $T_m = t^{(m+1)} - t^{(m)}$. With probability at least $1-\delta$ it holds that,*

$$\left| R_T^{\Gamma}(M) - R_T^{\Lambda_\epsilon}(L_\epsilon) \right| \leq c \left(\sqrt{T \frac{\log T}{1-\beta} \log \frac{\log T/(1-\beta)}{\delta}} + \frac{(\log T)^2}{1-\beta} \right)$$

where $c > 0$ is a universal constant and $R_T^{\Lambda_\epsilon}(L_\epsilon) = \sum_{m=1}^{M} R_{T_m}^{\Lambda_{\epsilon_m}}(L_m)$ and $R_T^{\Gamma}(M)$ is the regret of Algorithm 2 on the contextual bandit instance M defined in Section 3.

Proof sketch. The difference is upper bounded via the triangle inequality with the term,

$$\left| R_T^{\Gamma}(M) - R_T^{\Lambda}(L) \right| + \left| R_T^{\Lambda}(L) - R_T^{\Lambda_\epsilon}(L_\epsilon) \right|,$$

where $R_T^{\Lambda}(L)$ effectively denotes the regret of a hypothetical algorithm that has access to the stationary distribution. The first term is readily bounded in Section 5. The second term corresponds to the discrepancy between accumulated regrets over Θ and its discretization Θ', which amounts to a constant when summed over T. Full proof is deferred to Appendix C.2. □

Lemma 6.4. *Let Algorithm 2 be run with epoch lengths* $T_m := t^{(m+1)} - t^{(m)} = \tau + 2^{m-1}$, *and delay* $\tau = \lceil c_\tau \log T/(1-\beta) \rceil$. *Let* Λ_ϵ *be the PE algorithm specified in Lattimore et al. (2020) with* ϵ_m *being input as the upper bound in Lemma 6.2. Then, with probability at least* $1-\delta$, *the cumulative regret incurred by all instances of* Λ_{ϵ_m} *in Algorithm 2 satisfies*

$$R_T^{\Lambda_\epsilon} \le c\left(\sqrt{\frac{dT}{1-\beta}}\sqrt{d\log T + \log\tfrac{1}{\delta}} + \frac{\sqrt{d}\,\log T}{1-\beta}\right),$$

where $c > 0$ *is a universal constant.*

Proof sketch. The result follows by plugging the misspecification bound from Lemma 6.2 into the regret guarantee of the PE algorithm given in Lattimore et al. (2020) and summing the resulting bounds over all epochs. The full derivation is provided in Appendix C.3. □

7. Numerical Results

We illustrate our reduction framework through a numerical experiment on a synthetic Context-Markovian linear bandit environment. The primary goal of this experiment is to illustrate how the proposed delayed feedback mechanism effectively stabilizes learning in the presence of temporally dependent contexts. In particular, the experiment aims to highlight the qualitative behavior predicted by our theory and to contrast it with a standard linear bandit baseline that does not explicitly account for Markovian dependence.

We simulate the context process $(\mathcal{A}_t)_{t\ge1}$ using a Markov chain defined over a state space S, with transition kernel $P = \beta Q + (1-\beta)\mathbf{1}\pi^\top$. For each state $s \in \mathsf{S}$, the context set $\mathcal{A}_s = \{a_{s,1}, \dots, a_{s,K}\} \subset \mathbb{R}^d$ consists of randomly initialized unit-norm vectors. Here, Q is a local transition kernel on a ring graph, which creates temporal correlation across successive contexts, while the mixture with the stationary distribution π ensures uniform ergodicity.

Figure 1 reports the cumulative regret over time, averaged over 20 independent runs, comparing the reduction-based approach in Algorithm 1 with an OFUL-style linear UCB baseline with a fixed exploration coefficient. While Algorithm 1 incurs higher regret in the early stages due to the random starting phase and the delayed updates, it is able to leverage information across contexts more effectively. Consequently, it identifies the optimal action more rapidly, and its cumulative regret eventually falls below that of the baseline. Moreover, the delayed feedback mechanism leads to more stable action selection, as reflected in reduced variance across runs at later stages of learning.

Finally, when the same UCB baseline is employed as the linear bandit subroutine Λ within Algorithm 1, we observe that increasing the regularization parameter by multiple orders of magnitude is necessary to mitigate the influence of early exploratory actions during the random warm-start phase, which can otherwise lead to poor conditioning of the estimator. Empirically, we also find that treating the delay coefficient c_τ as a tunable hyperparameter results in improved performance, with the most favorable behavior observed for values around $c_\tau \approx 1$. These observations are consistent with our theoretical analysis, which predicts that an appropriately chosen delay reduces the effective bias induced by Markovian dependence while preserving the benefits of standard linear bandit algorithms.

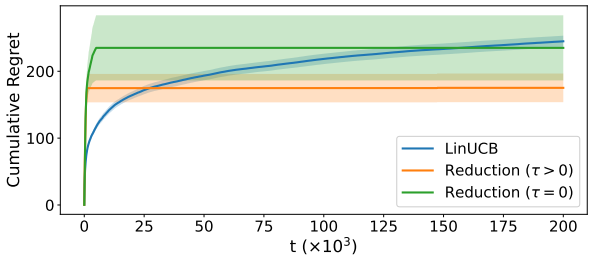

Figure 1. Mean cumulative regret over 20 runs; shaded regions indicate ± one standard error across runs around the mean.

8. Conclusion

Learning in interactive systems rarely takes place in static or independently generated environments. In many applications, the set of available actions, features, or contexts evolves gradually over time, shaped by latent dynamics, system inertia, or external processes. This work studies such *ergodically evolving context structures* and asks whether temporal dependence fundamentally alters the statistical limits of contextual bandit learning.

Our results suggest that it does not. We show that when the context process is sufficiently ergodic, the learning problem retains the same essential difficulty as in the classical i.i.d. setting. The dominant regret term continues to be governed by the intrinsic complexity of linear bandit learning, while the effect of temporal dependence enters only through an explicit and interpretable mixing penalty.

Conceptually, this work extends the "contexts can be cheap" perspective beyond static distributions to dynamical environments. Rather than requiring independence, we show that it suffices for the context process to forget its past at an appropriate rate. A delayed-feedback reduction converts dependence into a vanishing misspecification error, allowing standard linear bandit algorithms to operate effectively without resets, blocking, or episodic assumptions. This highlights a general principle: ergodicity can play a role analogous to independence in online decision-making, provided it is explicitly incorporated into the algorithmic design.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011. URL http://papers.nips.cc/paper_files/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.html.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 127–135. PMLR, 2013. URL https://proceedings.mlr.press/v28/agrawal13.html.

Bertsimas, D. and Niño-Mora, J. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

Bojinov, I. and Gupta, S. Online experimentation: Benefits, operational and methodological challenges, and scaling guide. 2022. URL https://hdsr.mitpress.mit.edu/pub/aj31wj81.

Chacun, G. et al. Dronebandit: Multi-armed contextual bandits for collaborative edge-to-cloud inference in resource-constrained nanodrones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3):1–23, 2024.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214. PMLR, 2011. URL https://proceedings.mlr.press/v15/chu11a.html.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In Servedio, R. A. and Zhang, T. (eds.), *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Hanna, O., Yang, L., and Fragouli, C. Efficient batched algorithm for contextual linear bandits with large action space via soft elimination. *Advances in Neural Information Processing Systems*, 36:56772–56783, 2023a.

Hanna, O., Yang, Y., and Fragouli, C. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In *Proceedings of The 36th Annual Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1791–1821. PMLR, 2023b. URL https://proceedings.mlr.press/v195/hanna23a.html.

Hao, B., Lattimore, T., and Szepesvári, C. Adaptive exploration in linear contextual bandit. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*. PMLR, 2020. URL https://proceedings.mlr.press/v108/hao20a.html.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828. URL https://banditalgs.com.

Lattimore, T., Szepesvári, C., and Weisz, G. Learning with good feature representations in bandits and in rl with a generative model. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Liu, H., Wei, C., and Zimmert, J. Bypassing the simulator: Near-optimal adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a3a661eb3308d0bb686f6a4bac521032-Abstract-Conference.html.

Luo, H., Wei, C., Agarwal, A., and Langford, J. Efficient contextual bandits in non-stationary worlds. In *Proceedings of The 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*. PMLR, 2018. URL https://proceedings.mlr.press/v75/luo18a.html.

Nelson, E., Bhattacharjya, D., Gao, T., Liu, M., Bouneffouf, D., and Poupart, P. Linearizing contextual bandits with latent state dynamics. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1477–1487. PMLR, 2022. URL https://proceedings.mlr.press/v180/nelson22a.html.

Neu, G. and Valko, M. Online combinatorial optimization with stochastic decision sets and adversarial losses. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://papers.nips.cc/paper/5381-online-combinatorial-optimization-with-stochastic-decision-sets-and-adversarial-losses.

Ok, J., Proutière, A., and Tranos, V. Exploration in structured reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Olkhovskaya, J., Mayo, J. J., van Erven, T., Neu, G., and Wei, C. First- and second-order bounds for adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c2201e444d2b22a10ca50116a522b9a9-Abstract-Conference.html.

Ortner, R., Ryabko, D., and Auer, P. Regret bounds for restless markov bandits. In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pp. 214–228. Springer, 2012. doi: 10.1007/978-3-642-34106-9_18.

Paulin, D. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic journal of probability*, 20, September 2015. ISSN 1083-6489. doi: 10.1214/EJP.v20-4039.

Tekin, C. and Liu, M. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58 (8):5588–5611, 2012.

Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

van Erven, T., Mayo, J. J., Olkhovskaya, J., and Wei, C. An improved algorithm for adversarial linear contextual bandits via reduction. *arXiv preprint arXiv:2508.11931*, 2025. URL https://arxiv.org/abs/2508.11931.

Varatharajah, Y. and Berry, B. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8), 2022. ISSN 2075-1729. doi: 10.3390/life12081277. URL https://www.mdpi.com/2075-1729/12/8/1277.

Wakayama, S. and Ahmed, N. Observation-augmented contextual multi-armed bandits for robotic search and exploration. *arXiv preprint arXiv:2312.12583*, 2023.

Weber, R. R. and Weiss, G. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

Yemini, M., Leshem, A., and Somekh-Baruch, A. The restless hidden markov bandit with linear rewards and side information. *arXiv preprint arXiv:1910.10271*, 2019. URL https://arxiv.org/abs/1910.10271.

Zeng, S., Bhatt, S., Koppel, A., and Ganesh, S. Partially observable contextual bandits with linear payoffs. *arXiv preprint arXiv:2409.11521*, 2024. URL https://arxiv.org/abs/2409.11521.

Zhang, Z. and Yuan, T. Evaluating online bandit exploration in large-scale recommender system. 2023.

A. Proof of Lemma 5.1

Proof. Let $f_\theta(\mathcal{A}) := \arg\max_{a\in\mathcal{A}} \langle a, \theta\rangle$.

$$g_\rho(\theta) - g_{\rho'}(\theta) = \int_{\mathsf{S}} f_\theta(\mathcal{A})\,(d\rho - d\rho').$$

Using the fact that for any $v \in \mathbb{R}^d$, $\|v\|_2 = \sup_{\|w\|_2\le 1}\langle w, v\rangle$,

$$\begin{aligned}
\|g_\rho(\theta) - g_{\rho'}(\theta)\|_2 &= \sup_{\|w\|_2\le 1} \left\langle w, \int_{\mathsf{S}} f_\theta(\mathcal{A})\,(d\rho - d\rho')\right\rangle \\
&= \sup_{\|w\|_2\le 1} \int_{\mathsf{S}} \langle w, f_\theta(\mathcal{A})\rangle\,(d\rho - d\rho') \\
&= \sup_{\|w\|_2\le 1} \left|\int_{\mathsf{S}} \langle w, f_\theta(\mathcal{A})\rangle\,(d\rho - d\rho')\right| \\
&\le \sup_{\|w\|_2\le 1} \int_{\mathsf{S}} |\langle w, f_\theta(\mathcal{A})\rangle|\,|d\rho - d\rho'| \\
&\le \int_{\mathsf{S}} |d\rho - d\rho'| = 2\mathrm{TV}(\rho, \rho')
\end{aligned}$$

where the last inequality follows from $\|f_\theta(\mathcal{A})\|_2 \le 1$, since $\|a\|_2 \le 1$ for all $a \in \mathcal{A}$ for all $\mathcal{A} \in \mathsf{S}$ and the Cauchy–Schwarz inequality. □

B. Full Proofs Relating to the Regret Bound of Algorithm 1

B.1. Performance of OFUL in Abbasi-Yadkori et al. (2011) with Bias of Order T^{-c_τ}

For $t > \tau$, the rewards satisfy

$$r_t = \langle x_t, \theta_\star\rangle + \Delta_t + \eta'_t, \qquad \sup_{t\le T}|\Delta_t| \le \varepsilon_T, \qquad \varepsilon_T := 2C_{\mathrm{mix}}T^{-c_\tau},$$

where $c_\tau > 1$, (x_t) is $\mathcal{F}_{t-1}$-measurable, and η'_t is conditionally sub-Gaussian given $\mathcal{F}_{t-1}$. Following Abbasi-Yadkori et al. (2011), define

$$V_t = \lambda I + \sum_{s=\tau+1}^{t} x_s x_s^\top, \qquad \hat\theta_t = V_t^{-1}\sum_{s=\tau+1}^{t} x_s r_s.$$

Then

$$\hat\theta_t - \theta_\star = V_t^{-1}\sum_{s=\tau+1}^{t} x_s\eta'_s + V_t^{-1}\sum_{s=\tau+1}^{t} x_s\Delta_s.$$

By Theorem 1 and Eq. (5) of Abbasi-Yadkori et al. (2011), with probability at least $1-\delta$,

$$\left\|V_t^{-1}\sum_{s=\tau+1}^{t} x_s\eta'_s\right\|_{V_t} \le \beta_t(\delta),$$

where $\beta_t(\delta)$ is the standard self-normalized confidence radius. Moreover,

$$\left\|V_t^{-1}\sum_{s=\tau+1}^{t} x_s\Delta_s\right\|_{V_t} = \left\|\sum_{s=\tau+1}^{t} x_s\Delta_s\right\|_{V_t^{-1}} \le \varepsilon_T\sqrt{\sum_{s=\tau+1}^{t} x_s^\top V_t^{-1} x_s} \le \varepsilon_T\sqrt{\sum_{s=\tau+1}^{t} x_s^\top V_{s-1}^{-1} x_s} \le \varepsilon_T\sqrt{2\log\det(V_t/\lambda I)},$$

where the second inequality uses $V_t^{-1} \preceq V_{s-1}^{-1}$ and the last follows from Lemma 11 of Abbasi-Yadkori et al. (2011). Since $\|x_t\|_2 \le 1$ and $\lambda \ge 1$,

$$\log\det(V_t/\lambda I) \le d\log\left(1 + \frac{t}{\lambda d}\right) \le d\log(1+T).$$

Hence,

$$\|\hat{\theta}_t - \theta_\star\|_{V_t} \le \beta_t(\delta) + \varepsilon_T \sqrt{2d \log T}.$$

Running OFUL with this confidence radius and repeating the regret analysis of Theorem 3 in Abbasi-Yadkori et al. (2011) yields

$$R_T^{\mathrm{OFUL}} \le c\Big(\beta_T(\delta)\sqrt{T} + \varepsilon_T \sqrt{dT \log T}\Big)$$

for a universal constant $c > 0$. Since $\varepsilon_T = O(T^{-c_\tau})$ with $c_\tau > 1$, the second term is $o(\sqrt{T})$, and the regret order remains $O(d\sqrt{T}\log T)$.

B.2. Proof of Theorem 5.2

Proof. The regret of the linear bandit algorithm is defined as

$$R_T^{\Lambda}(L) = \sum_{t=\tau}^{T} \max_{\theta\in\Theta} \langle g_\pi(\theta), \theta_\star\rangle - \langle g_\pi(\theta_t), \theta_\star\rangle, \tag{5}$$

while the regret of the contextual bandit algorithm is defined as

$$R_T^{\Gamma}(M) = \sum_{t=1}^{T} \max_{a\in\mathcal{A}_t} \langle a, \theta_\star\rangle - \left\langle \arg\max_{a\in\mathcal{A}_t} \langle a, \theta_t\rangle, \theta_\star \right\rangle \le 2\tau + \sum_{t=\tau}^{T} \max_{a\in\mathcal{A}_t} \langle a, \theta_\star\rangle - \left\langle \arg\max_{a\in\mathcal{A}_t} \langle a, \theta_t\rangle, \theta_\star \right\rangle \tag{6}$$

where the inequality follows from $\|a\|_2 \le 1, \|\theta\|_2 \le 1$ for all $a \in \mathcal{A}_t$ and $\theta \in \Theta$. We first show that, for any context distribution ρ,

$$\langle g_\rho(\theta'), \theta'\rangle = \max_{\theta\in\Theta} \langle g_\rho(\theta), \theta'\rangle, \qquad \forall \theta' \in \Theta. \tag{7}$$

Indeed, for any $\theta', \theta'' \in \Theta$,

$$\max_{\theta\in\Theta} \langle g_\rho(\theta), \theta'\rangle \ge \langle g_\rho(\theta'), \theta'\rangle = \mathbb{E}\left[\max_{a\in\mathcal{A}_t} \langle a, \theta'\rangle\right] \ge \mathbb{E}\left[\left\langle \arg\max_{a\in\mathcal{A}_t} \langle a, \theta''\rangle, \theta'\right\rangle\right] = \langle g_\rho(\theta''), \theta'\rangle.$$

Taking $\theta'' = \arg\max_\theta \langle g_\rho(\theta), \theta'\rangle$ forces equality.

Since $\max_{\theta\in\Theta} \langle g_\pi(\theta), \theta_\star\rangle = \langle g_\pi(\theta_\star), \theta_\star\rangle$, the regret difference can be written as,

$$\left|R_T^{\Gamma}(M) - R_T^{\Lambda}(L)\right| \le \left|\sum_{t=\tau}^{T} b_t(\theta_t) - b_t(\theta_\star)\right| + 2\tau \le \left|\sum_{t=\tau}^{T} b_t(\theta_t)\right| + \left|\sum_{t=\tau}^{T} b_t(\theta_\star)\right| + 2\tau \tag{8}$$

where

$$b_t(\theta) := \langle g_\pi(\theta), \theta_\star\rangle - \left\langle \arg\max_{a\in\mathcal{A}_t} \langle a, \theta\rangle, \theta_\star \right\rangle$$

Defining the filtration $\mathcal{G}_t := \sigma(\theta_1, \mathcal{A}_1, \dots \theta_{t-\tau}, \mathcal{A}_{t-\tau})$, we prove that τ subsequences of $b_t(\theta_t) - \mathbb{E}\left[b_t(\theta_t)|\mathcal{G}_t\right]$ exist whose sums satisfy the martingale property with respect to subsequences of $\mathcal{G}_t$ shifted by τ. For ease of notation, we drop the argument θ_t and write b_t in the subsequent part to denote $b(\theta_t)$.

Let $t_{r,1} < t_{r,2} < \cdots < t_{r,m_r}$ for $r \in \{0, \dots, \tau - 1\}$ be the time indices of a subsequence that satisfy $t_{r,k} = t_{r,k-1} + \tau$ for $k \in [m_r]$.

Defining $\Sigma_{t_{r,k}} := \sum_{j=1}^{k} \left(b_{t_{r,j}} - \mathbb{E}\left[b_{t_{r,j}}|\mathcal{G}_{t_{r,j}}\right]\right)$,

$$\mathbb{E}\left[\Sigma_{t_{r,k}}|\mathcal{G}_{t_{r,k-1}+\tau}\right] = \mathbb{E}\left[\Sigma_{t_{r,k}}|\mathcal{G}_{t_{r,k}}\right] = \Sigma_{t_{r,k-1}} + \mathbb{E}\left[b_{t_{r,k}} - \mathbb{E}\left[b_{t_{r,k}}|\mathcal{G}_{t_{r,k}}\right]|\mathcal{G}_{t_{r,k}}\right] = \Sigma_{t_{r,k-1}}$$

Since $\Sigma_{t_{r,k}}$ is $\mathcal{G}_{t_{r,k}+\tau}$-measurable and has bounded increments, by Azuma-Hoeffding inequality, it holds with probability at least $1 - \delta/2\tau$ that,

$$|\Sigma_{r,m_r}| \le c' \sqrt{m_r \log \frac{\tau}{\delta}}$$

for a fixed constant c'. By the union bound, it holds for all $r \in \{0, \dots, \tau - 1\}$, with probability at least $1 - \delta/2$ that,

$$\left|\sum_{r=0}^{\tau-1} \Sigma_{r,m_r}\right| \overset{(i)}{\leq} \sum_{r=0}^{\tau-1} |\Sigma_{r,m_r}| \leq c' \sqrt{\log \frac{\tau}{\delta}} \sum_{r=0}^{\tau-1} \sqrt{m_r} \overset{(ii)}{\leq} c' \sqrt{\log \frac{\tau}{\delta}} \sqrt{\tau \sum_{r=0}^{\tau-1} m_r}$$
$$= c' \sqrt{T\tau \log \frac{\tau}{\delta}} \leq c'' \sqrt{T \frac{\log T}{1-\beta} \log \frac{\log T/(1-\beta)}{\delta}}.$$

for a fixed constant $c'' > 0$ where (i) follows from the triangle inequality and (ii) follows from the Cauchy–Schwarz inequality.

Then, with probability at least $1 - \delta/2$,

$$\left|\sum_{t=\tau}^{T} b_t(\theta_t)\right| \leq \left|\sum_{t=\tau}^{T} b_t(\theta_t) - \mathbb{E}\left[b_t(\theta_t)|\mathcal{G}_t\right]\right| + \left|\sum_{t=\tau}^{T} \mathbb{E}\left[b_t(\theta_t)|\mathcal{G}_t\right]\right| \leq \left|\sum_{r=0}^{\tau-1} \Sigma_{r,m_r}\right| + \sum_{t=\tau}^{T} |2C_{\text{mix}}\beta^\tau|$$
$$\leq c'' \sqrt{T \frac{\log T}{1-\beta} \log \frac{\log T/(1-\beta)}{\delta}} + \frac{2C_{\text{mix}}}{T^{c_\tau - 1}}.$$

where the inequality $\left|\sum_{t=\tau}^{T} \mathbb{E}\left[b_t(\theta_t)|\mathcal{G}_t\right]\right| \leq \sum_{t=\tau}^{T} |2C_{\text{mix}}\beta^\tau|$ follows since $\mathcal{G}_t$ does not contain any information for the context beyond $\mathcal{A}_{t-\tau}$ given $\mathcal{A}_{t-\tau}$, and we can apply the same argument behind equations (2) and (3). Bounding $\left|\sum_{t=\tau}^{T} b_t(\theta_\star)\right|$ similarly using the filtration $\mathcal{G}'_t := \sigma(\mathcal{A}_1, \dots, \mathcal{A}_{t-\tau})$ and combining with (8) concludes the proof.

□

B.3. Proof of Corollary 5.3

Proof. Similar to (8), (5) and (6) suggest that,

$$\left|\mathbb{E}\left[R_T^\Gamma(M)\right] - \mathbb{E}\left[R_T^\Lambda(L)\right]\right| \leq \left|\sum_{t=\tau}^{T} \mathbb{E}\left[b_t(\theta_t)\right] - \mathbb{E}\left[b_t(\theta_\star)\right]\right| + 2\tau \leq \left|\sum_{t=\tau}^{T} \mathbb{E}\left[b_t(\theta_t)\right]\right| + \left|\sum_{t=\tau}^{T} \mathbb{E}\left[b_t(\theta_\star)\right]\right| + 2\tau \tag{9}$$

where,

$$b_t(\theta) := \langle g_\pi(\theta), \theta_\star \rangle - \left\langle \arg\max_{a \in \mathcal{A}_t} \langle a, \theta \rangle, \theta_\star \right\rangle$$

It follows from the tower property and Jensen's inequality that,

$$|\mathbb{E}\left[b_t(\theta_t)\right]| = |\mathbb{E}\left[\mathbb{E}\left[b_t(\theta_t) \mid \mathcal{G}_t\right]\right]| \leq \mathbb{E}\left[|\mathbb{E}\left[b_t(\theta_t) \mid \mathcal{G}_t\right]|\right]$$

where $\mathcal{G}_t := \sigma(\theta_1, \mathcal{A}_1, \dots \theta_{t-\tau}, \mathcal{A}_{t-\tau})$

Since $\mathcal{G}_t$ does not contain any information for the context beyond $\mathcal{A}_{t-\tau}$ given $\mathcal{A}_{t-\tau}$, applying the same argument behind equations (2) and (3) shows,

$$\mathbb{E}\left[|\mathbb{E}\left[b_t(\theta_t) \mid \mathcal{G}_t\right]|\right] \leq 2\text{TV}(\mu_{t-\tau}P^\tau, \pi) \leq 2C_{\text{mix}}\beta^\tau$$

where the inner expectation is bounded by a constant almost surely and $\mu_{t-\tau}$ is the probability measure over the contexts, conditioned on $\mathcal{G}_t$. Hence,

$$\left|\sum_{t=\tau}^{T}\mathbb{E}\left[b_t(\theta_t)\right]\right| \le \sum_{t=\tau}^{T}\left|\mathbb{E}\left[b_t(\theta_t)\right]\right| \le \sum_{t=\tau}^{T}\mathbb{E}\left[\left|\mathbb{E}\left[b_t(\theta_t)\mid \mathcal{G}_t\right]\right|\right] \le 2TC_{\text{mix}}\beta^{\tau}$$

Bounding $|\mathbb{E}[b_t(\theta_\star)]|$ using $\mathcal{G}'_t = \sigma(\mathcal{A}_1,\dots,\mathcal{A}_{t-\tau})$ with the same argument and summing over T gives the final result. □

C. Full Proofs Relating to the Regret Bound of Algorithm 2

C.1. Proof of Lemma 6.2

Proof. For fixed $\theta,\theta' \in \Theta'$, let,

$$h(\mathcal{A}_t) := \left\langle \arg\max_{a\in\mathcal{A}_t}\langle a,\theta\rangle, \theta'\right\rangle$$

Then,

$$\begin{aligned}\left|\langle g_\pi(\theta),\theta'\rangle - \left\langle g^{(m)}(\theta),\theta'\right\rangle\right| &= \frac{1}{t^{(m)}}\left|\sum_{t=1}^{t^{(m)}}\langle g_\pi(\theta),\theta'\rangle - h(\mathcal{A}_t)\right| \\ &\le \frac{1}{t^{(m)}}\left|\sum_{t=1}^{t^{(m)}} h(\mathcal{A}_t) - \mathbb{E}\left[h(\mathcal{A}_t)\right]\right| + \frac{1}{t^{(m)}}\sum_{t=1}^{t^{(m)}}\left|\mathbb{E}\left[h(\mathcal{A}_t)\right] - \langle g_\pi(\theta),\theta'\rangle\right|\end{aligned} \tag{10}$$

The first sum in (10) can be bounded as described in (4) since $h(\mathcal{A}_t)\le 1$ for all $a\in\mathcal{A}_t, \theta'\in\Theta$. Hence, with probability at least $1-\delta/(M|\Theta'|^2)$,

$$\left|\sum_{t=1}^{t^{(m)}} h(\mathcal{A}_t) - \mathbb{E}\left[h(\mathcal{A}_t)\right]\right| \le \sqrt{36\,t^{(m)}\,\frac{\log(4C_{\text{mix}})}{1-\beta}\log\left(\frac{2M|\Theta'|}{\delta}\right)} \tag{11}$$

where, using the union bound, (11) holds for $\forall\theta,\theta'\in\Theta'$ with probability at least $1-\delta/M$. The second sum in (10) can be bounded with Lemma 5.1:

$$\begin{aligned}\sum_{t=1}^{t^{(m)}}\left|\mathbb{E}\left[h(\mathcal{A}_t)\right] - \langle g_\pi(\theta),\theta'\rangle\right| &\le \sum_{t=1}^{t^{(m)}}\left|\left\langle \mathbb{E}\left[\arg\max_{a\in\mathcal{A}_t}\langle a,\theta\rangle\right] - g_\pi(\theta),\theta'\right\rangle\right| \\ &\le \sum_{t=1}^{t^{(m)}}\left\|g_t(\theta) - g_\pi(\theta)\right\|_2 \le \sum_{t=1}^{t^{(m)}} 2C_{\text{mix}}\beta^t \le \frac{2C_{\text{mix}}}{1-\beta}\end{aligned} \tag{12}$$

Where $g_t(\cdot)$ denotes the $g_\rho(\cdot)$ function evaluated at the marginal law of $\mathcal{A}_t$ at time t. Combining (10), (11) and (12) gives for all $\theta,\theta'\in\Theta'$, with probability at least $1-\delta/M$,

$$\left|\langle g_\pi(\theta),\theta'\rangle - \left\langle g^{(m)}(\theta),\theta'\right\rangle\right| \le \sqrt{\frac{36\,\log(4C_{\text{mix}})}{(1-\beta)\,t^{(m)}}\log\left(\frac{2M|\Theta'|}{\delta}\right)} + \frac{C_{\text{mix}}}{(1-\beta)\,t^{(m)}}$$

Finally, picking any $\theta\in\Theta'$ and $\theta'\in\Theta$ guarantees there exists $\phi'\in\Theta'$ such that $\|\theta'-\phi'\|_2\le 1/T$. By Cauchy–Schwarz and the triangle inequality,

$$\left|\langle g_\pi(\theta),\theta'\rangle - \left\langle g^{(m)}(\theta),\theta'\right\rangle\right| \le \left|\langle g_\pi(\theta),\phi'\rangle - \left\langle g^{(m)}(\theta),\phi'\right\rangle\right| + \frac{2}{T},\ \forall\theta'\in\Theta,\forall\theta\in\Theta'.$$

This concludes the proof.

□

C.2. Proof of Lemma 6.3

Proof. The regret $R_T^{\Gamma}(M)$ can be decomposed as,

$$\left|R_T^{\Gamma}(M) - R_T^{\Lambda_\epsilon}(L_\epsilon)\right| \leq \left|R_T^{\Gamma}(M) - R_T^{\Lambda}(L)\right| + \left|R_T^{\Lambda}(L) - R_T^{\Lambda_\epsilon}(L_\epsilon)\right| \tag{13}$$

where R_T^{Λ} is defined as,

$$R_T^{\Lambda}(L) = \sum_{t=1}^{T} \max_{\theta\in\Theta} \langle g_\pi(\theta), \theta_\star\rangle - \langle g_\pi(\theta_t), \theta_\star\rangle ,$$

By definition, the regret incurred by Λ_ϵ operated in Algorithm 2 is,

$$R_T^{\Lambda}(L_\varepsilon) = \sum_{t=1}^{T} \max_{\theta\in\Theta'} \langle g_\pi(\theta), \theta_\star\rangle - \langle g_\pi(\theta_t), \theta_\star\rangle .$$

We first prove that $\left|R_T^{\Gamma}(M) - R_T^{\Lambda_\epsilon}(L_\epsilon)\right|$ is bounded by a constant. We have,

$$\left|R_T^{\Lambda}(L) - R_T^{\Lambda}(L_\varepsilon)\right| = \sum_{t=1}^{T} \left| \max_{\theta\in\Theta} \langle g_\pi(\theta), \theta_\star\rangle - \max_{\theta\in\Theta'} \langle g_\pi(\theta), \theta_\star\rangle \right|. \tag{14}$$

We now bound the inner difference uniformly in t. It is derived in (7) that,

$$\langle g_\pi(\theta'), \theta'\rangle = \max_{\theta\in\Theta} \langle g_\pi(\theta), \theta'\rangle , \qquad \forall \theta' \in \Theta.$$

Fix $\theta_\star \in \Theta$. Since Θ' is a $1/T$-net of Θ, there exists $\phi \in \Theta'$ such that $\|\theta_\star - \phi\|_2 \leq 1/T$. Using (7), boundedness $\|g_\pi(\theta)\|_2 \leq 1$, and the triangle inequality,

$$\begin{aligned}
\max_{\theta\in\Theta} \langle g_\pi(\theta), \theta_\star\rangle &= \langle g_\pi(\theta_\star), \theta_\star\rangle \\
&\leq \langle g_\pi(\theta_\star), \phi\rangle + \frac{1}{T} \\
&\leq \max_{\theta\in\Theta} \langle g_\pi(\theta), \phi\rangle + \frac{1}{T} \\
&= \max_{\theta\in\Theta'} \langle g_\pi(\theta), \phi\rangle + \frac{1}{T} \\
&\leq \max_{\theta\in\Theta'} \langle g_\pi(\theta), \theta_\star\rangle + \frac{2}{T}.
\end{aligned}$$

By symmetry, the reverse inequality also holds, hence

$$\left| \max_{\theta\in\Theta} \langle g_\pi(\theta), \theta_\star\rangle - \max_{\theta\in\Theta'} \langle g_\pi(\theta), \theta_\star\rangle \right| \leq \frac{2}{T}. \tag{15}$$

Substituting (15) into (14) yields

$$\left|R_T^{\Lambda}(L) - R_T^{\Lambda}(L_\varepsilon)\right| \leq \sum_{t=1}^{T} \frac{2}{T} = 2, \tag{16}$$

For the difference $\left|R_T^{\Gamma}(M) - R_T^{\Lambda}(L)\right|$, we use the result of Theorem 5.2 after accounting for the τ-length intervals where actions are selected randomly. We have, by the triangle inequality,

$$\left|R_T^{\Gamma}(M) - R_T^{\Lambda}(L)\right| \le \sum_{t\in I_r} \left|R_t^{\Gamma}(M) - R_t^{\Lambda}(L)\right| + \left|\sum_{t\in I_t} \left(R_t^{\Gamma}(M) - R_t^{\Lambda}(L)\right)\right| \tag{17}$$

where $I_r = \bigcup_{m=1}^{M}\{t^{(m)}+1,\dots,t^{(m)}+\tau\}$ denotes the set of time indices where θ_t was picked randomly and $I_t = \bigcup_{m=1}^{M}\{t^{(m)}+\tau+1,\dots,t^{(m+1)}\}$ denotes the remaining time indices. $R_t^{\Gamma}(M)$ and $R_t^{\Lambda}(L)$ denote the per-round regret of Algorithm 2 and Λ, respectively. Note that when $t \in I_t$, the definition of R_t^{Λ} is identical to the regret of the linear bandit algorithm Λ that is aware of the stationary distribution used in Algorithm 1.

By Theorem 5.2, we have, with probability at least $1-\delta/2$,

$$\left|\sum_{t\in I_t} \left(R_t^{\Gamma}(M) - R_t^{\Lambda}(L)\right)\right| \le c'\left(\sqrt{T\frac{\log T}{1-\beta}\log\frac{\log T/(1-\beta)}{\delta}} + \frac{\log T}{1-\beta}\right) \tag{18}$$

where $c' > 0$ is a universal constant. For the second term in (17), since there are $M = O(logT)$ epochs by the epoch schedule $t^{(m)} = 2^{m-1}+\tau$ and the per-step regret is bounded by 1, it holds almost surely,

$$\sum_{t\in I_r} \left|R_t^{\Gamma}(M) - R_t^{\Lambda}(L)\right| \le 2M\tau \le c''\left(\frac{(logT)^2}{1-\beta}\right) \tag{19}$$

for a constant $c'' > 0$. Combining (19) and (18) with (17) completes the proof. □

C.3. Proof of Lemma 6.4

Proof. Let $T_m = \tau + 2^{m-1}$, $H_m = 2^{m-1}$, and $t^{(m)} = (m-1)\tau + (2^{m-1}-1)$. Throughout the proof, let each appearance of the symbol c denote a (possibly different) positive constant. For each epoch $m \in [M]$, define the misspecification event

$$\mathcal{E}_m \;:=\; \Big\{\forall\theta\in\Theta',\ \forall\theta'\in\Theta:\ \left|\langle g_\pi(\theta),\theta'\rangle - \langle g^{(m)}(\theta),\theta'\rangle\right| \le \epsilon_m\Big\},$$

where $\epsilon_1 := 1$ and for $m \ge 2$,

$$\epsilon_m := \sqrt{\frac{36\,\log(4C_{\mathrm{mix}})}{(1-\beta)\,t^{(m)}}\log\!\left(\frac{2M|\Theta'|}{\delta}\right)} + \frac{C_{\mathrm{mix}}}{(1-\beta)\,t^{(m)}} + \frac{2}{T}.$$

By Lemma 6.2, $\Pr(\mathcal{E}_m) \ge 1-\delta/M$ for each m, hence by a union bound,

$$\Pr\Big(\bigcap_{m=1}^{M}\mathcal{E}_m\Big) \;\ge\; 1-\delta.$$

In epoch m, Algorithm 2 runs a fresh instance of Λ_{ϵ_m} and feeds it exactly H_m action–reward pairs. The guarantee of the PE algorithm specified in Lattimore et al. (2020) then gives, with probability at least $1-\delta/(2M)$,

$$R_{H_m}^{\Lambda_{\epsilon_m}} \le c\Big(\sqrt{dH_m\log\!\left(\tfrac{2M|\Theta'|}{\delta}\right)} + \sqrt{d}\,H_m\,\epsilon_m\Big).$$

Taking a union bound over $m\in[M]$ for these bandit-regret events and intersecting with $\bigcap_{m=1}^{M}\mathcal{E}_m$, we obtain that, with probability at least $1-3\delta/2$,

$$R_T^{\Lambda_\epsilon} = \sum_{m=1}^{M} R_{H_m}^{\Lambda_{\epsilon_m}} \le c\sqrt{d\log\!\left(\tfrac{2M|\Theta'|}{\delta}\right)}\sum_{m=1}^{M}\sqrt{H_m} + c\sqrt{d}\sum_{m=1}^{M}H_m\,\epsilon_m. \tag{20}$$

Since $H_m = 2^{m-1}$ and $\sum_{m=1}^{M} H_m = 2^M - 1 \le T$, we have

$$\sum_{m=1}^{M} \sqrt{H_m} = \sum_{m=1}^{M} 2^{(m-1)/2} \le c\, 2^{M/2} \le c\sqrt{T}.$$

Next, $\epsilon_1 = 1$ so $H_1\epsilon_1 \le 1$. For $m \ge 2$, $t^{(m)} \ge 2^{m-1} - 1 \ge \frac{1}{2}H_m$, hence $H_m\sqrt{1/t^{(m)}} \le \sqrt{2}\sqrt{H_m}$ and $H_m/t^{(m)} \le 2$, so

$$\sum_{m=2}^{M} H_m \sqrt{\frac{1}{t^{(m)}}} \le c \sum_{m=2}^{M} \sqrt{H_m} \le c\sqrt{T}, \qquad \sum_{m=2}^{M} \frac{H_m}{t^{(m)}} \le cM, \qquad \sum_{m=1}^{M} H_m \frac{1}{T} \le 1.$$

Therefore, applying the result of Lemma 6.2

$$\sum_{m=1}^{M} H_m\, \epsilon_m \le c\sqrt{T}\sqrt{\frac{\log(4C_{\mathrm{mix}})}{1-\beta}\log\Big(\tfrac{2M|\Theta'|}{\delta}\Big)} + c\,\frac{C_{\mathrm{mix}}}{1-\beta}\,M + c,$$

Moreover, it is well known that if $\Theta \subseteq \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}$, the $1/T$-net $\Theta' \subseteq \Theta$ satisfies $|\Theta'| \le (6T)^d$, so that

$$\log\Big(\tfrac{2M|\Theta'|}{\delta}\Big) = O\big(d\log T + \log\tfrac{1}{\delta}\big).$$

With $M = \Theta(\log T)$ and absorbing C_{mix} into constants, this yields

$$R_T^{\Lambda_\epsilon} = O\left(\sqrt{\frac{dT}{1-\beta}}\sqrt{d\log T + \log\tfrac{1}{\delta}} \;+\; \frac{\sqrt{d}\,\log T}{1-\beta}\right).$$

□

C.4. High Probability Regret Bound of Algorithm 2

Summing the terms obtained in Lemmas 6.3 and 6.4 give the final high probability regret bound,

$$R_T^{\Gamma}(M) \le C\left(d\sqrt{\frac{T\log T}{1-\beta}} + \frac{\sqrt{d}\,\log T}{1-\beta} + \sqrt{\frac{T\log T}{1-\beta}\log\left(\frac{T\log T}{1-\beta}\right)} + \frac{(\log T)^2}{1-\beta}\right),$$

where $C > 0$ is a universal constant.

D. Details on Numerical Results

Environmental Parameters. Table 1 describes the parameters utilized when generating the the environment specified in Section 7. Let $\mathsf{S} = \{0, 1, \ldots, S-1\}$ denote the finite state space, endowed with the stationary distribution $\pi(s)$ over $s \in \mathsf{S}$. The transition dynamics are controlled by parameters $\beta \in (0,1)$, $P_{\text{loop}} \in (0,1)$, and an integer $N_{\text{neighbors}} \ge 1$, with degree $\deg = 2m$.

For each $s \in \mathsf{S}$, define the ring-neighborhood multiset

$$\mathcal{N}(s) \;:=\; \{(s+k) \bmod S : k = 1, \ldots, N_{\text{neighbors}}\} \;\cup\; \{(s-k) \bmod S : k = 1, \ldots, N_{\text{neighbors}}\},$$

so that $|\mathcal{N}(s)| = \deg$.

The local kernel $Q \in \mathbb{R}^{S\times S}$ is defined row-wise by

$$Q_{s,s} \;:=\; P_{\text{loop}}, \qquad Q_{s,j} \;:=\; \frac{1 - P_{\text{loop}}}{\deg} \;\text{ for } j \in \mathcal{N}(s), \qquad Q_{s,j} \;:=\; 0 \;\text{ otherwise},$$

Table 1. Environment parameters.

Quantity	Value
Horizon	$T = 200{,}000$
Feature dimension	$d = 20$
Number of states	$S = 40$
Actions per state	$K = 20$
Parameter bank size	$M = 256$
Noise std. dev.	$\sigma = 0.5$
Mixture parameter	$\beta = 0.75$
Self-loop prob. in Q	$P_{\text{loop}} = 0.20$
Ring neighbors each side	$N_{\text{neighbors}} = 2$
Stationary distribution	$\pi(s) = 1/S$ (uniform)
Number of Runs	N = 20

Table 2. Algorithm hyperparameters.

Method	Hyperparameter	Value
Contextual LinUCB	Regularization	$\lambda_{\text{base}} = 10^{-2}$
Contextual LinUCB	UCB multiplier	$\alpha_{\text{base}} = 2.0$
Reduction Algorithm	Regularization	$\lambda_{\text{alg1}} = 100.0$
Reduction Algorithm	UCB multiplier	$\alpha_{\text{alg1}} = 2.0$
Reduction Algorithm	Bonus cap	$\texttt{bonus\_cap} = 2.5$
Warm start	Delay coefficient	$c_\tau = 1.0$
Warm start	Delay	$\tau = \left\lceil c_\tau \frac{\log T}{1-\beta} \right\rceil$

so that $\sum_{j=0}^{S-1} Q_{s,j} = 1$ for all s.

The final transition kernel is,

$$P \;:=\; (1-\varepsilon)Q + \varepsilon \mathbf{1}\pi^\top,$$

or equivalently,

$$P_{s,s'} \;=\; \beta Q_{s,s'} + (1-\beta)\,\pi(s').$$

The resulting Markov chain is uniformly ergodic with stationary distribution π.

Algorithm Hyperparameters. The hyperparameters used in each algorithm are described in Table 2. LinUCB was used as the linear bandit oracle Λ and the exploration bonus was capped by `bonus_cap` to ensure the first τ random actions do not destabilize the reduction algorithm.