This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2025.3533299

ON RESTLESS LINEAR BANDITS

Abstract

A more general formulation of the linear bandit problem is considered to allow for dependencies over time. Specifically, it is assumed that there exists an unknown $\mathbb{R}^d$-valued stationary φ-mixing sequence of parameters $(\theta_t,\ t \in \mathbb{N})$ which gives rise to payoffs. This instance of the problem can be viewed as a generalization of both the classical linear bandits with iid noise, and the finite-armed restless bandits. In light of the well-known computational hardness of optimal policies for restless bandits, an approximation is proposed whose error is shown to be controlled by the φ-dependence between consecutive θ_t. An optimistic algorithm, called LinMix-UCB, is proposed for the case where θ_t has an exponential mixing rate. The proposed algorithm is shown to incur a sub-linear regret of $\mathcal{O}\left(\sqrt{dn}\,\mathrm{polylog}(n)\right)$ with respect to an oracle that always plays a multiple of $\mathbb{E}\,\theta_t$. The main challenge in this setting is to ensure that the exploration-exploitation strategy is robust against long-range dependencies. The proposed method relies on Berbee's coupling lemma to carefully select near-independent samples and construct confidence ellipsoids around empirical estimates of $\mathbb{E}\theta_t$.

I. INTRODUCTION

The problem of sequential decision making in the presence of uncertainty is prevalent in a variety of modern applications such as online advertisement, recommendation systems, network routing and dynamic pricing. Through the framework of *stochastic bandits* one can model this task as a repeated game between a *learner* and a stochastic *environment*: at every time-step, the learner chooses an *action* from a pre-specified set of actions $\mathcal{A}$ and receives a (random) real-valued payoff. The objective is to maximize the expected cumulative payoff over time. In a stochastic *linear* bandit model the payoff Y_t received at each time-step t is the inner product between an $\mathbb{R}^d$-valued action X_t and an unknown parameter vector $\theta \in \mathbb{R}^d$. That is,

$$Y_t = \langle \theta, X_t \rangle + \eta_t$$

with random noise η_t; see e.g. [1], [2], as we well as [3], [4] and references therein. Let us point out that in the case where the action space $\mathcal{A}$ is restricted to a set of standard unit vectors in $\mathbb{R}^d$ the problem is reduced to that of finite-armed bandits. The noise sequence η_t is typically assumed to be independently and identically distributed (iid). However, this assumption does not usually hold in practice.

In this paper, we consider a more general formulation of the linear bandit problem which allows for dependencies over time. More specifically, we assume that there exists an unknown $\mathbb{R}^d$-valued stationary sequence of parameters $(\theta_t,\ t \in \mathbb{N})$ giving rise to the payoffs $Y_t = \langle \theta_t, X_t \rangle,\ t \in \mathbb{N}$ with the actions X_t taking values in a unit ball in $\mathbb{R}^d$. As with any bandit problem, the task of balancing the trade-off between exploration and exploitation

calls for finite-time analysis, which in turn relies on concentration inequalities for empirical averages. For this reason, we opt for a natural approach in time-series analysis which is to require that the process satisfy a form of asymptotic independence. More specifically, we assume that $(\theta_t,\ t \in \mathbb{N})$ is φ-mixing so that its sequence of φ-mixing coefficients $\varphi_m,\ m \in \mathbb{N}$ defined by

$$\varphi_m := \sup_{j \in \mathbb{N}} \sup_{\substack{U \in \sigma(\{\theta_t : t=1,\dots,j\}) \\ V \in \sigma(\{\theta_t : t \geq j+m\})}} |P(V) - P(V|U)|$$

converges to 0 as m approaches infinity, see, e.g. [5], [6]. Observe that while θ_t are identically distributed here, they are not independent.

To compare with the classical formulation, we can write

$$Y_t = \langle \theta^*, X_t \rangle + \eta_t$$

where $\theta^* := \mathbb{E}\theta_t$ is the (unknown) stationary mean of θ_t and

$$\eta_t := \langle \theta_t - \theta^*, X_t \rangle$$

is the noise process, which is clearly non-iid. This instance of the problem can be viewed as a generalization of the classical linear bandit problem with iid η_t [2] and the finite-armed restless Markovian and φ-mixing bandits considered in [7] and [8] respectively. Observe that in much the same way as with the examples given in [7, Section 3], a strategy that leverages temporal dependencies can accumulate a higher expected payoff than that which can be obtained by playing a fixed action in $\mathcal{A}$ that is most aligned with θ^*. In the finite-armed restless bandit problem, this is called the *optimal switching strategy* whose exact computation in certain related instances of the problem is known to be intractable [9]. In [10], an algorithm is proposed for finite-armed restless bandits in the case where the payoff distributions follow an AR model. An approximation of the optimal switching strategy for finite-armed restless φ-mixing bandits is provided in [8].

We build upon [8] to approximate the optimal restless bandit strategy using the Optimism in the Face of Uncertainty principal. We show in Proposition 1 that the approximation error of an oracle which always plays a multiple of θ^*, is controlled by φ_1 and the ℓ_2 norm of θ_t, provided that the latter is almost surely bounded. The proof relies on Vitali's covering theorem [11] to address the technical challenges presented by the infinite action space $\mathcal{A} \subseteq \mathbb{R}^d$. We propose an algorithm, namely LinMix-UCB, which aims to minimize the regret in the case where the process $(\theta_t,\ t \in \mathbb{N})$ has an exponential mixing rate. The proposed algorithm is inspired by the UCB-based approach of [2] which is in turn designed for the simpler setting where the noise η_t is iid. The main challenge in our setting is to devise an exploration-exploitation strategy that is robust against the dependencies present in the process $(\theta_t,\ t \in \mathbb{N})$. This problem is similar to the that considered in [8]. However, their setting involves a finite number of arms, allowing the regret analysis to rely on a Hoeffding-type inequality for φ-mixing processes. This approach does not carry over to the linear bandit setting considered in the present paper. We rely on Berbee's coupling lemma [12] to carefully select near-independent samples and construct confidence ellipsoids around empirical estimates of θ^*.

We demonstrate that in the case where that $(\theta_t,\ t \in \mathbb{N})$ has an exponential mixing rate, LinMix-UCB incurs a sub-linear regret of $\mathcal{O}\left(\sqrt{dn\,\mathrm{polylog}(n)}\right)$ with respect to an oracle which always plays a multiple of θ^*. While our results are not confined to Markovian processes, we would like to point out that a natural example of a process with an exponential φ-mixing rate is any stationary ergodic aperiodic Markov chain, see, e.g. [5, Theorem 21.22 - vol. II pp. 329].

II. Preliminaries and Problem Formulation

Let $\Theta \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$. Suppose that $\boldsymbol{\theta} := (\theta_t,\ t \in \mathbb{N})$ is a stationary sequence of Θ-valued random variables defined on a probability space $(\Omega, \mathcal{B}, P)$ such that $\theta_t,\ t \in \mathbb{N}$ takes values in the space $\mathcal{L}_\infty(\Omega, \mathcal{B}, P; \mathbb{R}^d)$ equipped with its Euclidean norm $\|\cdot\|_2$. This means that

$$\|\theta_t\|_{\mathcal{L}_\infty} := \sup_{\omega\in\Omega} \|\theta_t(\omega)\|_2 < \infty.$$

We may sometimes use $\mathcal{L}_\infty$ or $\mathcal{L}_\infty(\Omega; \mathbb{R}^d)$ for $\mathcal{L}_\infty(\Omega, \mathcal{B}, P; \mathbb{R}^d)$ when the remaining parameters are clear from the context. Recall that the φ-dependence (see, e.g. [5]) between any pair of σ-subalgebras $\mathcal{U}$ and $\mathcal{V}$ of $\mathcal{B}$ is given by $\varphi(\mathcal{U}, \mathcal{V}) := \sup\{|P(V) - P(V|U)| : U \in \mathcal{U}, P(U) > 0, V \in \mathcal{V}\}$. This notion gives rise to the sequence of φ-mixing coefficients $\varphi_m,\ m \in \mathbb{N}$ of $\boldsymbol{\theta}$ where for each $m \in \mathbb{N}$

$$\varphi_m := \sup_{j\in\mathbb{N}} \varphi(\sigma(\{\theta_t : 1 \le t \le j\}), \sigma(\{\theta_t : t \ge j + m\})).$$

We further assume that the process $\boldsymbol{\theta}$ is φ-mixing so that $\lim_{m\to\infty} \varphi_m = 0$. Let $\mathcal{A}$ be the unit ball in $\mathbb{R}^d$, which we call the *action space*. The linear bandit problem considered in this paper can be formulated as the following repeated game. At every time-step $t \in \mathbb{N}$, the player chooses an *action* from $\mathcal{A}$ according to a mapping $X_t : \Omega \to \mathcal{A}$ and receives as *payoff* the inner product $\langle \theta_t, X_t\rangle$ between θ_t and X_t. The objective is to maximize the expected sum of the accumulated payoffs. Let $\mathcal{F}_t,\ t \ge 0$ be a filtration that tracks the payoffs $\langle \theta_t, X_t\rangle$ obtained in the past t rounds, i.e. $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and

$$\mathcal{F}_t = \sigma(\{\langle \theta_1, X_1\rangle, \ldots, \langle \theta_t, X_t\rangle\})$$

for $t \ge 1$. Each mapping $X_t,\ t \ge 1$ is assumed to be measurable with respect to $\mathcal{F}_{t-1}$; this is equivalent to stating that X_t for each $t \ge 1$ can be written as a function of the past payoffs up to $t - 1$. The sequence

$$\boldsymbol{\pi} := (X_t,\ t \in \mathbb{N})$$

is called a *policy*. Let $\Pi = \{\boldsymbol{\pi} := (X_t^{(\boldsymbol{\pi})},\ t \ge 1) : X_t^{(\boldsymbol{\pi})}$ is $\mathcal{F}_{t-1}$-measurable for all $t \ge 1\}$ denote the space of all possible policies and define

$$\nu_n = \sup_{\boldsymbol{\pi}\in\Pi} \sum_{t=1}^{n} \mathbb{E}\langle \theta_t, X_t^{(\boldsymbol{\pi})}\rangle. \tag{1}$$

To simplify notation, we may sometimes write X_t for $X_t^{(\boldsymbol{\pi})}$ when the policy $\boldsymbol{\pi}$ is clear from the context.

III. Main Results

Consider the restless linear bandit problem formulated in Section II. Let $\theta^* = \mathbb{E}\,\theta_1$ be the (stationary) mean of the process $(\theta_t,\ t \in \mathbb{N})$. In light of the well-known computational hardness of optimal switching strategies for restless bandits [9], we aim to approximate ν_n via the following relaxation

$$\begin{aligned}\widetilde{\nu}_n &= \sup_{\boldsymbol{\pi}\in\Pi} \sum_{t=1}^{n} \mathbb{E}\langle \theta^*, X_t^{(\boldsymbol{\pi})}\rangle \\ &= n\,\|\theta^*\| \end{aligned} \tag{2}$$

since $\mathcal{A}$ is the unit ball in $\mathbb{R}^d$. A natural first objective is thus to quantify the error $\nu_n - \widetilde{\nu}_n$ incurred by this approximation. We present the following result which can be considered as a slightly more technical analogue of [8, Proposition 9].

Proposition 1. *Let φ_1 be the first φ-mixing coefficient of the process $(\theta_t,\ t \in \mathbb{N})$. For every $n \geq 1$ it holds that*

$$\nu_n - \widetilde{\nu}_n \leq 2n\varphi_1 \left\|\theta_t\right\|_{\mathcal{L}_\infty}\ .$$

Proof. Consider an arbitrary policy $\boldsymbol{\pi} = (X_t^{(\boldsymbol{\pi})},\ t \in \mathbb{N})$ and any $t \in \mathbb{N}$; note that from this point on the notation X_t will be used to denote $X_t^{(\boldsymbol{\pi})}$. Let $\widetilde{P}_t := P \circ X_t^{-1}$ be the pushforward measure on the action space $\mathcal{A}$ under X_t. Fix an $\epsilon > 0$. As follows from Vitali's covering theorem [11, Theorem 2.8] there exists a set of disjoint closed balls $\{B_i : i \in \mathbb{N}\}$ of radii at most ϵ, that covers the action space $\mathcal{A}$ (i.e. the unit ball in $\mathbb{R}^d$), such that $\widetilde{P}_t(\mathcal{A} \setminus \cup_{i\in\mathbb{N}} B_i) = 0$. Note that by assumption $\theta_t \in \mathcal{L}_\infty(\Omega;\mathbb{R}^d)$ so that $\|\theta_t\|_{\mathcal{L}_\infty} := \sup_{\omega\in\Omega} \|\theta_t(\omega)\|_2 < \infty$. Consider a ball B_i for some $i \in \mathbb{N}$ and denote its center by $\bar{x}_i \in \mathbb{R}^d$. Let $E_i := \{X_t \in B_i\}$ denote the pre-image of B_i under X_t. Since B_i is of radius at most ϵ, it holds that,

$$\begin{aligned}\mathbb{E}\langle \theta_t, X_t - \bar{x}_i\rangle \mathbb{I}_{B_i}(X_t) &\leq \mathbb{E} \sup_{x\in B_i} \langle \theta_t, x - \bar{x}_i\rangle \mathbb{I}_{B_i}(X_t) \\ &\leq \int_{E_i} \epsilon \left\|\theta_t\right\|_2 dP \\ &= \epsilon P(E_i) \left\|\theta_t\right\|_{\mathcal{L}_\infty}. \end{aligned} \tag{3}$$

Similarly, for $\theta^* = \mathbb{E}\,\theta_t$ we have,

$$\mathbb{E}\langle \theta^*, X_t - \bar{x}_i\rangle \mathbb{I}_{B_i}(X_t) \leq \epsilon P(E_i) \left\|\theta_t\right\|_{\mathcal{L}_\infty}. \tag{4}$$

Moreover, noting that $\mathcal{A}$ is the unit ball in $\mathbb{R}^d$, by Cauchy-Schwarz inequality, for each $x \in \mathcal{A}$ we have,

$$\begin{aligned}\left\|\langle \theta_t, x\rangle\right\|_{\mathcal{L}_\infty} &= \sup_{\omega\in\Omega} |\langle \theta_t(\omega), x\rangle| \\ &\leq \sup_{\omega\in\Omega} \left\|\theta_t(\omega)\right\|_2 \\ &= \left\|\theta_t\right\|_{\mathcal{L}_\infty} \\ &< \infty. \end{aligned} \tag{5}$$

Hence, as follows from [5, Theorem 4.4(c2) - vol. I pp. 124] it holds that

$$\begin{aligned}
&\frac{\left\|\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle|\mathcal{F}_{t-1})-\mathbb{E}\langle\theta_t,\bar{x}_i\rangle\right\|_{\mathcal{L}_\infty}}{\left\|\langle\theta_t,\bar{x}_i\rangle\right\|_{\mathcal{L}_\infty}}\\
&\leq \sup_{Y\in\mathcal{L}_\infty(\Omega,\sigma(\theta_t),P)}\frac{\left\|\mathbb{E}(Y|\mathcal{F}_{t-1})-\mathbb{E}Y\right\|_{\mathcal{L}_\infty}}{\left\|Y\right\|_{\mathcal{L}_\infty}} &(6)\\
&= 2\varphi(\mathcal{F}_{t-1},\sigma(\theta_t)) &(7)\\
&\leq 2\varphi_1 &(8)
\end{aligned}$$

where $\mathcal{L}_\infty(\Omega,\sigma(\theta_t),P)$ denotes the set of all $\sigma(\theta_t)$-measurable, real-valued random variables such that $\sup_{\omega\in\Omega}|Y(\omega)|<\infty$ almost surely. We have,

$$\begin{aligned}
&|\mathbb{E}((\langle\theta_t,\bar{x}_i\rangle-\langle\theta^*,\bar{x}_i\rangle)\mathbb{I}_{B_i}(X_t))|\\
&=|\mathbb{E}\left(\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle\mathbb{I}_{B_i}(X_t)|\mathcal{F}_{t-1})\right)-\mathbb{E}\langle\theta^*,\bar{x}_i\rangle\mathbb{I}_{B_i}(X_t)|\\
&=|\mathbb{E}\left(\mathbb{I}_{B_i}(X_t)\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle|\mathcal{F}_{t-1})\right)-\mathbb{E}\langle\theta^*,\bar{x}_i\rangle\mathbb{I}_{B_i}(X_t)|\\
&\leq\mathbb{E}\left(\mathbb{I}_{B_i}(X_t)\left|\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle|\mathcal{F}_{t-1})-\mathbb{E}\langle\theta_t,\bar{x}_i\rangle\right|\right)\\
&=\int_{E_i}\left|\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle|\mathcal{F}_{t-1})-\mathbb{E}\langle\theta_t,\bar{x}_i\rangle\right|dP\\
&\leq\int_{E_i}\left\|\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle|\mathcal{F}_{t-1})-\mathbb{E}\langle\theta_t,\bar{x}_i\rangle\right\|_{\mathcal{L}_\infty}dP\\
&=P(E_i)\left\|\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle|\mathcal{F}_{t-1})-\mathbb{E}\langle\theta_t,\bar{x}_i\rangle\right\|_{\mathcal{L}_\infty}\\
&\leq 2\varphi_1P(E_i)\left\|\theta_t\right\|_{\mathcal{L}_\infty} &(9)
\end{aligned}$$

where the second equality follows from noting that X_t is $\mathcal{F}_{t-1}$-measurable, and (9) follows from (5) and (8). Thus,

$$\begin{aligned}
&|\mathbb{E}\ ((\langle\theta_t,X_t\rangle-\langle\theta^*,X_t\rangle)\mathbb{I}_{B_i}(X_t))|\\
&\quad\leq|\mathbb{E}\ \langle\theta_t,X_t-\bar{x}_i\rangle\mathbb{I}_{B_i}(X_t)|\\
&\qquad+|\mathbb{E}\ \langle\theta^*,X_t-\bar{x}_i\rangle\mathbb{I}_{B_i}(X_t)|\\
&\qquad+|\mathbb{E}(\langle\theta_t,\bar{x}_i\rangle-\langle\theta^*,\bar{x}_i\rangle)\mathbb{I}_{B_i}(X_t)|\\
&\quad\leq 2(\epsilon+\varphi_1)P(E_i)\left\|\theta_t\right\|_{\mathcal{L}_\infty} &(10)
\end{aligned}$$

where (10) follows from (3), (4) and (9). By (10), noting that the events $E_i,\ i\in\mathbb{N}$ partition Ω, and by applying the dominated convergence theorem, we have,

$$\begin{aligned}
&|\mathbb{E}\ (\langle\theta_t,X_t\rangle-\langle\theta^*,X_t\rangle)|\\
&\quad\leq\sum_{i\in\mathbb{N}}|\mathbb{E}\ ((\langle\theta_t,X_t\rangle-\langle\theta^*,X_t\rangle)\mathbb{I}_{B_i}(X_t))| &(11)\\
&\quad\leq\sum_{i\in\mathbb{N}}2P(E_i)(\epsilon+\varphi_1)\left\|\theta_t\right\|_{\mathcal{L}_\infty} &(12)\\
&\quad=2(\epsilon+\varphi_1)\left\|\theta_t\right\|_{\mathcal{L}_\infty}. &(13)
\end{aligned}$$

Finally, since this holds for every ϵ, we obtain,

$$\begin{aligned}\nu_n - \widetilde{\nu}_n &\leq \sup_{\boldsymbol{\pi}\in\Pi} \sum_{t=1}^{n} |\mathbb{E}(\langle \theta_t, X_t\rangle - \langle \theta^*, X_t\rangle)| \\ &\leq 2n\varphi_1 \left\|\theta_t\right\|_{\mathcal{L}_\infty} \end{aligned} \tag{14}$$

and the result follows. □

Denote by $\mathcal{R}_{\boldsymbol{\pi}}(n)$ the regret with respect to (2) incurred by a policy $\boldsymbol{\pi} = (X_t^{(\boldsymbol{\pi})},\ t \in \mathbb{N})$ after n rounds, i.e.

$$\mathcal{R}_{\boldsymbol{\pi}}(n) := \widetilde{\nu}_n - \sum_{t=1}^{n} \mathbb{E}\langle \theta_t, X_t^{(\boldsymbol{\pi})}\rangle. \tag{15}$$

In this paper, we consider a subclass of the problem where the process $(\theta_t,\ t \in \mathbb{N})$ has an exponential φ-mixing rate, so that there exists some $a_0, \gamma_0 \in (0,\infty)$ such that for all $m \in \mathbb{N}$,

$$\varphi_m \leq a e^{-\gamma m}. \tag{16}$$

We focus on the case where a and γ in (16) are known. In fact, knowing the exact values of a and γ is not crucial; an upper bound on a and a lower bound on γ would suffice. For simplicity, we retain the parameters a and γ, treating them as bounds on the true rate parameters. We propose LinMix-UCB, outlined in Algorithms 1 and 2 for the cases of finite and infinite horizon respectively.

Algorithm 1 LinMix-UCB (finite horizon)

Input: horizon n; regularization parameter λ; φ-mixing rate parameters: $a, \gamma \in (0,\infty)$

Specify block-length k given by (17)

for $m = 0, 1, 2, \ldots, \lfloor n/k \rfloor$ **do**
  for $\ell = 1, 2, \ldots, k$ **do**
    $t \leftarrow mk + \ell$
    if $m = 0$ **then**
      $X_t \leftarrow \mathbf{x}_0$ ▷ $\mathbf{x}_0$ *is a fixed unit vector in* $\mathcal{A}$
    else
      $X_t \leftarrow \operatorname{argmax}_{x\in\mathcal{A}} \max_{\theta\in C_{\max\{0,m-1\}}} \langle x, \theta\rangle$
    Play action X_t to obtain reward Y_t
    if $\ell = 1$ **then**
      Calculate confidence ellipsoid C_m (20)

The proposed algorithm is inspired by such UCB-type approaches as LinUCB and its variants, including LinRel [1] and OFUL [2], all of which are designed for linear bandits in the simpler iid noise setting, see [4, Chapter 19] and references therein. The main challenge in our setting is to devise an exploration-exploitation strategy that is robust against the long-range dependencies in the process $(\theta_t,\ t \in \mathbb{N})$. We construct confidence ellipsoids around the empirical estimates of θ^* obtained via "near-independent" samples, and similarly to LinUCB and its variants, we rely on the principle of Optimism in the Face of Uncertainty. More specifically, the algorithm works as follows. A finite horizon n is divided into intervals of length

$$k = \max\left\{1, \bar{k}\right\} \tag{17}$$

with $\bar{k}$ defined as

$$\left\lceil \frac{1}{\gamma} \log\left(\frac{6a\gamma n^2}{1 + 4\sqrt{n}\left\|\theta_t\right\|_{\mathcal{L}_\infty} + \sqrt{\frac{8d\times n\log(n(1+\frac{n}{\lambda d}))}{\lambda}}} \right) \right\rceil$$

where λ is a regularization parameter used in the estimation step. At every time-step $t = mk+1,\ m = 0, 1, \ldots, \lfloor n/k \rfloor$ which marks the beginning of each interval of length k, the payoffs $Y_s := \langle \theta_s, X_s \rangle$ collected every k-steps at $s = m'k + 1,\ m' = 0, 1, \ldots, \max\{0, m-1\}$, are used to generate a regularized least-squares estimator $\widehat{\theta}_m$ of $\theta^* = \mathbb{E}\theta_t$, and in turn, produce a confidence ellipsoid C_m. That is, for each $m = 0, 1, \ldots, \lfloor n/k \rfloor$ we have

$$\widehat{\theta}_m := \operatorname*{argmin}_{\theta\in\Theta} \left(\sum_{\substack{m'=0 \\ s=m'k+1}}^{\max\{0,m-1\}} (Y_s - \langle \theta, X_s \rangle)^2 + \lambda \left\|\theta\right\|_2^2 \right) \tag{18}$$

where the regularisation parameter $\lambda > 0$ ensures a unique solution given by

$$\widehat{\theta}_m = (\lambda I + V_m)^{-1} \sum_{\substack{m'=0 \\ s=m'k+1}}^{\max\{0,m-1\}} Y_s X_s \tag{19}$$

with $V_m := \sum_{\substack{m'=0 \\ s=m'k+1}}^{\max\{0,m-1\}} X_s X_s^\top$ takes values in $\mathbb{R}^{d\times d}$ and I is the identity matrix in $\mathbb{R}^{d\times d}$. This gives rise to the following confidence ellipsoid

$$C_m := \left\{ \theta \in \Theta : \left\| \theta - \widehat{\theta}_m \right\|^2_{\zeta^2(\lambda I + V_m)} \le b_n \right\} \tag{20}$$

where $\|x\|_A^2 := x^\top A x$ for $x \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$, $\zeta := 2\left\|\theta_t\right\|_{\mathcal{L}_\infty}$, and $b_n > 0$ is chosen such that $\sqrt{b_n}$ is equal to,

$$2\sqrt{\lambda}\left\|\theta_t\right\|_{\mathcal{L}_\infty} + \sqrt{2\log n + d\log\left(1 + \frac{n}{k_n \lambda d}\right)}. \tag{21}$$

We are now in a position to analyze the regret of the proposed algorithm.

Theorem 1. *Suppose that the stationary φ-mixing process $(\theta_t,\ t\in\mathbb{N})$ has an exponential mixing rate, so that there exists some $a,\gamma\in(0,\infty)$ such that $\varphi_m\le ae^{-\gamma m}$ for all $m\in\mathbb{N}$. The regret (with respect to $\widetilde{\nu}_n$) of LinMix-UCB (finite horizon) played for $n\ge\left\lceil\frac{3a\gamma\sqrt{\lambda}}{2\sqrt{\lambda}\|\theta_t\|_{\mathcal{L}_\infty}+\sqrt{2}}\right\rceil$ rounds satisfies*

$$\frac{\mathcal{R}_{\boldsymbol{\pi}}(n)}{\|\theta_t\|_{\mathcal{L}_\infty}}\le\frac{1}{n}+C\log(n)\sqrt{2dn\log(n(1+\frac{n}{\lambda d}))}$$

where

$$C:=\left(\frac{12(\sqrt{2\lambda}+4\sqrt{2\lambda}\,\|\theta_t\|_{\mathcal{L}_\infty}+1)}{\gamma\sqrt{2\lambda}}\right),$$

and $\lambda>0$ is the regularization parameter.

Proof. For a fixed $k\in\mathbb{N}$, consider the sub-sequence $\theta_{ik+1},\ i=0,1,2,\dots$ of the stationary sequence of $\mathbb{R}^d$-valued random variables $\theta_t,\ t\in\mathbb{N}$. Let $U_j,\ j\in\mathbb{N}$ be a sequence of iid random variables uniformly distributed over $[0,1]$ such that each U_j is independent of $\sigma(\{\theta_t:t\in\mathbb{N}\})$. Set $\widetilde{\theta}_0=\theta_1$. As follows from Berbee's coupling lemma [12] - see also [6, Lemma 5.1, pp. 89] - there exists a random variable

$$\widetilde{\theta}_1=g_1(\widetilde{\theta}_0,\theta_{k+1},U_1)$$

where g_1 is a measurable function from $\mathbb{R}^d\times\mathbb{R}^d\times[0,1]$ to $\mathbb{R}^d$ such that $\widetilde{\theta}_1$ is independent of $\widetilde{\theta}_0$, has the same distribution as θ_{k+1} and

$$\Pr(\widetilde{\theta}_1\ne\theta_{k+1})=\beta(\sigma(\widetilde{\theta}_0),\sigma(\theta_{k+1})).$$

Similarly, there exists a random variable

$$\widetilde{\theta}_2=g_2((\widetilde{\theta}_0,\widetilde{\theta}_1),\theta_{2k+1},U_2)$$

where g_2 is a measurable function from $(\mathbb{R}^d)^2\times\mathbb{R}^d\times[0,1]$ to $\mathbb{R}^d$ such that $\widetilde{\theta}_2$ is independent of $(\widetilde{\theta}_0,\widetilde{\theta}_1)$, has the same distribution as θ_{2k+1} and

$$\Pr(\widetilde{\theta}_2\ne\theta_{2k+1})=\beta(\sigma(\widetilde{\theta}_0,\widetilde{\theta}_1),\sigma(\theta_{2k+1})).$$

Continuing inductively in this way, at each step $j=3,4,\dots$, by Berbee's coupling lemma [12], there exists a random variable

$$\widetilde{\theta}_j=g_j((\widetilde{\theta}_0,\widetilde{\theta}_1,\dots,\widetilde{\theta}_{j-1}),\theta_{jk+1},U_j)$$

where $g_j:(\mathbb{R}^d)^j\times\mathbb{R}^d\times[0,1]\to\mathbb{R}^d$ is a measurable function such that

1) $\widetilde{\theta}_j$ is independent of $(\widetilde{\theta}_0,\widetilde{\theta}_1,\dots,\widetilde{\theta}_{j-1})$

2) $\widetilde{\theta}_j$ has the same distribution as θ_j and,

$$\Pr(\widetilde{\theta}_j\ne\theta_j)=\beta(\sigma(\widetilde{\theta}_0,\widetilde{\theta}_1,\dots,\widetilde{\theta}_{j-1}),\sigma(\theta_{jk+1})).\tag{22}$$

Following a standard argument (see, e.g. [13, Lemma 6]) it can be shown that,

$$\beta(\sigma(\widetilde{\theta}_0,\widetilde{\theta}_1,\dots,\widetilde{\theta}_{j-1}),\sigma(\theta_{jk+1}))\le\beta_k\tag{23}$$

for all $j \in \mathbb{N}$. The sequence of random variables $\widetilde{\theta}_j,\ j = 0, 1, 2, \dots$ can be used to construct a "ghost" payoff process $(\bar{\theta}_t,\ t \in \mathbb{N})$ as follows. Let $\bar{\theta}_{ik+1} = \widetilde{\theta}_i$ for all $i = 0, 1, 2, \dots$ and take $\bar{\theta}_t$ to be an independent copy of θ_1 for all $t = ik + 2, \dots, (i+1)k,\ i = 0, 1, 2, \dots$. Denote by $\boldsymbol{\pi} := (X_t,\ t \in \mathbb{N})$ the policy induced by Algorithm 1 when the process $(\theta_t,\ t \in \mathbb{N})$ is used to generate the payoffs $Y_t := \langle X_t, \theta_t \rangle$. Similarly let $\bar{\boldsymbol{\pi}} := (\bar{X}_t,\ t \in \mathbb{N})$ be the policy generated by Algorithm 1 when the sequence $(\bar{\theta}_t,\ t \in \mathbb{N})$ produces the payoffs $\bar{Y}_t := \langle \bar{X}_t, \bar{\theta}_t \rangle$ at each $t \in \mathbb{N}$. For a fixed $n \in \mathbb{N}$, define the event

$$E_n := \{\exists i \in 0, 1, \dots, (\lfloor n/k \rfloor) - 1 : \theta_{ik+1} \neq \bar{\theta}_{ik+1}\} \tag{24}$$

and observe that as follows from the above coupling argument, i.e. by (22) and (23), it holds that

$$\Pr(E_n) \leq n\beta_k/k. \tag{25}$$

Let $\mathcal{G}_0 = \bar{\mathcal{G}}_0 := \{\emptyset, \Omega\}$. Define the filtrations

$$\mathcal{G}_m := \sigma(\{\theta_{ik+1} : i = 0, 1, \dots, m-1\})$$

and

$$\bar{\mathcal{G}}_m := \sigma(\{\bar{\theta}_{ik+1} : i = 0, 1, \dots, m-1\})$$

for $m = 1, 2, \dots, \lfloor n/k \rfloor$. By design, for $t = 1, \dots, k$, the action X_t is set to a pre-specified constant $\mathbf{x}_0 \in \mathcal{A}$ (independent of the data), and is thus simply $\mathcal{G}_0$-measurable. Observe that the first confidence ellipsoid C_0 which is generated at $t = 1$ is not used directly, but only after k steps. For each time-step $t = mk + \ell$ with $m \in 1, \dots, \lfloor n/k \rfloor$ and $\ell = 1, 2, \dots, k$, the action X_t depends on the confidence ellipsoid $C_{\max\{0,m-1\}}$, which is in turn updated at least k time-steps prior to t. More specifically, X_t is measurable with respect to $\mathcal{G}_m$. As a result, there exists a measurable function

$$f_t : (\mathbb{R}^d)^m \to \mathcal{A}$$

such that

$$X_t = f_t(\theta_1, \theta_{k+1}, \theta_{2k+1}, \dots, \theta_{(m-1)k+1}).$$

In words, f_t is a mathematical representation of Algorithm 1 at time t. Similarly, noting that the same algorithm is applied to $(\bar{\theta}_t, ,\ t \in \mathbb{N})$, it holds that

$$\bar{X}_t = f_t(\bar{\theta}_1, \bar{\theta}_{k+1}, \bar{\theta}_{2k+1}, \dots, \bar{\theta}_{(m-1)k+1}).$$

As a result, for each $i = 1, \dots, \lfloor n/k \rfloor$, we have,

$$\begin{aligned} Y_{ik+1}\mathbb{I}_{E_n^c} &= \langle \theta_{ik+1}, X_{ik+1} \rangle \mathbb{I}_{E_n^c} \\ &= \langle \theta_{ik+1}, f_{ik+1}(\theta_1, \dots, \theta_{(i-1)k+1}) \rangle \mathbb{I}_{E_n^c} \\ &= \langle \bar{\theta}_{ik+1}, f_{ik+1}(\bar{\theta}_1, \dots, \bar{\theta}_{(i-1)k+1}) \rangle \mathbb{I}_{E_n^c} \\ &= \bar{Y}_{ik+1}\mathbb{I}_{E_n^c}, \end{aligned} \tag{26}$$

almost surely. On the other hand, by a simple application of Cauchy-Schwarz and Hölder inequalities, it is straightforward to verify that for each $t = 1, \ldots, n$ we have,

$$\begin{aligned}\mathbb{E}|Y_t \mathbb{I}_{E_n}| &= \int_{E_n} |\langle \theta_t, X_t \rangle| dP \\ &\leq P(E_n) \left\|\theta_t\right\|_{\mathcal{L}_\infty} \end{aligned} \tag{27}$$

and similarly,

$$\mathbb{E}|\bar{Y}_t \mathbb{I}_{E_n}| \leq P(E_n) \left\|\theta_t\right\|_{\mathcal{L}_\infty}. \tag{28}$$

It follows that

$$\begin{aligned}&\sum_{i=0}^{\lfloor n/k \rfloor} \left|\mathbb{E}Y_{ik+1} - \mathbb{E}\bar{Y}_{ik+1}\right| \\ &\quad = \sum_{i=0}^{\lfloor n/k \rfloor} \left|\mathbb{E}(Y_{ik+1}\mathbb{I}_{E_n}) - \mathbb{E}(\bar{Y}_{ik+1}\mathbb{I}_{E_n})\right| && (29) \\ &\quad \leq 2n \left\|\theta_t\right\|_{\mathcal{L}_\infty} P(E_n)/k && (30) \\ &\quad \leq 2n^2 \left\|\theta_t\right\|_{\mathcal{L}_\infty} \beta_k/k^2 && (31)\end{aligned}$$

where (29) follows from (26), (30) follows from (27) and (28), and (31) is due to (25). Next, let us consider the time-steps within each segment. Fix any $t = mk + \ell$ for some $m \in 1, \ldots, \lfloor n/k \rfloor$ and some $\ell \in 2, \ldots, k$. It is straightforward to verify that [5, Theorem 4.4(c2) - vol. I pp. 124] can be extended to the case of vector-valued random variables, by an analogous argument based on simple functions where absolute values of constants are replaced by norms. This leads to,

$$\begin{aligned}&\left\|\mathbb{E}(\theta_t | \bar{\mathcal{G}}_m \vee \mathcal{G}_m) - \mathbb{E}\theta_t\right\|_{\mathcal{L}_\infty} && (32) \\ &\quad \leq 2\varphi(\bar{\mathcal{G}}_m \vee \mathcal{G}_m, \sigma(\theta_t)) \left\|\theta_t\right\|_{\mathcal{L}_\infty} && (33) \\ &\quad \leq 2\varphi_k \left\|\theta_t\right\|_{\mathcal{L}_\infty}. && (34)\end{aligned}$$

Define the event

$$U_m := \{\exists i \in 0, 1, \ldots, m-1 : \theta_{ik+1} \neq \bar{\theta}_{ik+1}\}.$$

Observe that as with (24) we have

$$\Pr(U_m) \leq m\beta_k \tag{35}$$

so that similarly to (27), it holds that,

$$\max\{\mathbb{E}|Y_t \mathbb{I}_{U_m}|, \mathbb{E}|\bar{Y}_t \mathbb{I}_{U_m}|\} \leq m\beta_k \left\|\theta_t\right\|_{\mathcal{L}_\infty}. \tag{36}$$

for $t = mk + \ell$ with some $m \in 1, \ldots, \lfloor n/k \rfloor$ and $\ell \in 2, \ldots, k$ fixed above. Moreover,

$$\mathbb{E}(\langle \bar{\theta}_t, \bar{X}_t \rangle \mathbb{I}_{U_m^c}) = \mathbb{E}(\mathbb{I}_{U_m^c} \mathbb{E}(\langle \bar{\theta}_t, X_t \rangle | \bar{\mathcal{G}}_m \vee \mathcal{G}_m)) \tag{37}$$

$$= \mathbb{E}(\mathbb{I}_{U_m^c} \langle X_t, \mathbb{E}(\bar{\theta}_t | \bar{\mathcal{G}}_m \vee \mathcal{G}_m) \rangle) \tag{38}$$

$$= \mathbb{E}(\mathbb{I}_{U_m^c} \langle X_t, \mathbb{E}\, \theta_t \rangle) \tag{39}$$

where (37) follows from the definition of U_m^c and the fact that $\mathbb{I}_{U_m^c}$ is measurable with respect to $\bar{\mathcal{G}}_m \vee \mathcal{G}_m$, and (38) follows from noting that X_t is measurable with respect $\mathcal{G}_t$. Finally, (39) is due to the stationarity of θ_t together with the fact that by construction $\bar{\theta}_t$ is an independent copy of θ_1 for this choice of t (within the segments). Similarly, we have,

$$\mathbb{E}(\langle \theta_t, X_t \rangle \mathbb{I}_{U_m^c}) = \mathbb{E}(\mathbb{I}_{U_m^c} \mathbb{E}(\langle \theta_t, X_t \rangle | \bar{\mathcal{G}}_m \vee \mathcal{G}_m)) \tag{40}$$

$$= \mathbb{E}(\mathbb{I}_{U_m^c} \langle X_t, \mathbb{E}(\theta_t | \bar{\mathcal{G}}_m \vee \mathcal{G}_m) \rangle). \tag{41}$$

Hence, for any $m \in 1, \ldots, \lfloor n/k \rfloor$ and $t = mk + \ell$ for some $\ell \in 2, \ldots, k$ we have,

$$\begin{aligned}
&|\mathbb{E}((\langle \theta_t, X_t \rangle - \mathbb{E}\langle \bar{\theta}_t, \bar{X}_t \rangle) \mathbb{I}_{U_m^c})| \\
&\quad \leq \mathbb{E}(\mathbb{I}_{U_m^c} \langle X_t, |\mathbb{E}(\theta_t | \bar{\mathcal{G}}_m \vee \mathcal{G}_m) - \mathbb{E}\, \theta_t| \rangle)) && (42) \\
&\quad = \int_{U_m^c} \langle X_t, |\mathbb{E}(\theta_t | \bar{\mathcal{G}}_m \vee \mathcal{G}_m) - \mathbb{E}\, \theta_t| \rangle dP && (43) \\
&\quad \leq \int_{U_m^c} \|X_t\|_2 \left\| \mathbb{E}(\theta_t | \bar{\mathcal{G}}_m \vee \mathcal{G}_m) - \mathbb{E}\, \theta_t \right\|_2 dP && (44) \\
&\quad \leq P(U_m^c) \left\| \mathbb{E}(\theta_t | \bar{\mathcal{G}}_m \vee \mathcal{G}_m) - \mathbb{E}\, \theta_t \right\|_{\mathcal{L}_\infty} && (45) \\
&\quad \leq 2\varphi_k \|\theta_t\|_{\mathcal{L}_\infty} && (46)
\end{aligned}$$

where (42) follows from (39) and (41); (44) and (45) follow from Cauchy-Schwarz and Hölder inequalities respectively, and (46) follows from (32). We obtain,

$$\begin{aligned}
&\sum_{m=1}^{\lfloor n/k \rfloor} \sum_{\ell=2}^{k} |\mathbb{E}(Y_{mk+\ell} - \bar{Y}_{mk+\ell})| \\
&\quad = \sum_{m=0}^{\lfloor n/k \rfloor} \sum_{\ell=2}^{k} |\mathbb{E}((Y_{mk+\ell} - \bar{Y}_{mk+\ell}) \mathbb{I}_{U_m^c})| \\
&\qquad + |\mathbb{E}((Y_{mk+\ell} - \bar{Y}_{mk+\ell}) \mathbb{I}_{U_m})| && (47) \\
&\quad \leq 2n\varphi_k \|\theta_t\|_{\mathcal{L}_\infty} + 2n^2 \beta_k \|\theta_t\|_{\mathcal{L}_\infty} && (48) \\
&\quad \leq 4n^2 \varphi_k \|\theta_t\|_{\mathcal{L}_\infty} && (49)
\end{aligned}$$

where (47) follows from noting that by design, $\mathbb{E}\bar{Y}_t = \mathbb{E}Y_t$ for all $t \in 1, \ldots, k$ as the algorithm sets $X_t = \mathbf{x}_0$ for some constant $\mathbf{x}_0 \in \mathcal{A}$ independent of the data, (48) follows from (36) and (46), and (49) is due to the fact that in

general $\beta_k \leq \varphi_k$ for all $k \in \mathbb{N}$ [5, Proposition 3.11 - vol. I pp. 76]. Therefore, by (31) and (49) we obtain

$$\begin{aligned}
&\left|\mathcal{R}_{\boldsymbol{\pi}(n)} - \mathcal{R}_{\bar{\boldsymbol{\pi}}}(n)\right| \\
&\leq \sum_{i=0}^{\lfloor n/k \rfloor} \left|\mathbb{E}Y_{ik+1} - \mathbb{E}\bar{Y}_{ik+1}\right| \\
&\qquad + \sum_{m=0}^{\lfloor n/k \rfloor} \sum_{\ell=2}^{k} \left|\mathbb{E}Y_{mk+\ell} - \mathbb{E}\bar{Y}_{mk+\ell}\right| \\
&\leq 6n^2 \varphi_k \left\|\theta_t\right\|_{\mathcal{L}_\infty} . \qquad (50)
\end{aligned}$$

It remains to calculate the regret of $\bar{\boldsymbol{\pi}} = \{\bar{X}_t : t \in 1, \ldots, n\}$. The payoff $\bar{Y}_t = \langle \bar{\theta}_t, \bar{X}_t \rangle$ obtained via the policy $\bar{\boldsymbol{\pi}}$ at time-step t can be decomposed as

$$\bar{Y}_t = \langle \theta^*, \bar{X}_t \rangle + \eta_t \qquad (51)$$

where

$$\eta_t := \langle \bar{\theta}_t - \theta^*, \bar{X}_t \rangle. \qquad (52)$$

For each $m = 0, 1, \ldots, \lfloor n/k \rfloor$, set

$$S_m = \sum_{i=0}^{m} \eta_{ik+1} \bar{X}_{ik+1}$$

define

$$V_m := \sum_{\substack{m'=0 \\ s=m'k+1}}^{m} \bar{X}_s \bar{X}_s^\top$$

and let $I \in \mathbb{R}^{d \times d}$ be the identity matrix. Consider the estimator

$$\widehat{\theta}_m := (\lambda I + V_m)^{-1} \sum_{\substack{m'=0 \\ s=m'k+1}}^{m} \bar{Y}_s \bar{X}_s$$

and observe that

$$\begin{aligned}
\widehat{\theta}_m &= (\lambda I + V_m)^{-1} \left(S_m + \sum_{\substack{m'=0 \\ s=m'k+1}}^{m} \bar{X}_s \bar{X}_s^\top \theta^* \right) \qquad (53) \\
&= (\lambda I + V_m)^{-1} \left(S_m + V_m \theta^* \right). \qquad (54)
\end{aligned}$$

Let $\zeta := 2\left\|\theta_t\right\|_{\mathcal{L}_\infty}$. We can write,

$$\begin{aligned}
&\left\|\widehat{\theta}_m - \theta^*\right\|_{\zeta^2(\lambda I+V_m)} \\
&= \left\|(\lambda I + V_m)^{-1}\left(S_m + V_m\theta^*\right) - \theta^*\right\|_{\zeta^2(\lambda I+V_m)} \\
&\leq \left\|S_m\right\|_{\zeta^2(\lambda I+V_m)^{-1}} \\
&\qquad + \zeta\lambda^{1/2}({\theta^*}^\top(I-(\lambda I+V_m)^{-1}V_m)\theta^*)^{1/2} \\
&= \left\|S_m\right\|_{\zeta^2(\lambda I+V_m)^{-1}} + \zeta\lambda({\theta^*}^\top(\lambda I+V_m)^{-1}\theta^*)^{1/2} \\
&\leq \left\|S_m\right\|_{\zeta^2(\lambda I+V_m)^{-1}} + \zeta\lambda^{1/2}\left\|\theta^*\right\|_2 \qquad (55)
\end{aligned}$$

where (55) follows from noting that V_m is positive semi-definite and Löwner matrix order is reversed through inversion, so that

$${\theta^*}^\top(\lambda I+V_m)^{-1}\theta^* \leq {\theta^*}^\top(\lambda I)^{-1}\theta^* = \lambda^{-1}\left\|\theta^*\right\|_2^2.$$

Observe that $\bar{X}_t$ for $t = mk+1,\ m = 0,1,\ldots,\lfloor n/k\rfloor$ is $\bar{\mathcal{G}}_m$-measurable, and that by construction $\bar{\theta}_{ik+1}$ for $i = 0,1,\ldots,m$ are iid. Thus,

$$\mathbb{E}(\eta_t) = \mathbb{E}(\langle\bar{X}_t, \mathbb{E}(\theta^* - \theta_t|\bar{\mathcal{G}}_m)\rangle) = 0.$$

Furthermore, it is straightforward to verify that by Cauchy-Schwarz inequality and noting that X_t takes values in the unit ball, η_t is ζ-subGaussian, i.e. for all $\alpha\in\mathbb{R}$ and every $m = 0,1,\ldots,\lfloor n/k\rfloor$ and $t = mk+1$ we have, $\mathbb{E}(e^{\alpha\eta_t}|\bar{\mathcal{G}}_m)) \leq e^{\alpha^2\zeta^2/2}$ almost surely. In particular,

$$\begin{aligned}
&\mathbb{E}(\exp\{\langle x,X_t\rangle\eta_t\}|\bar{\mathcal{G}}_m)) \\
&\qquad \leq \exp\{\langle x, X_t\rangle^2\zeta^2/2\} \\
&\qquad = \exp\left\{\frac{\zeta^2\left\|x\right\|^2_{X_tX_t^\top}}{2}\right\} \qquad (56)
\end{aligned}$$

almost surely for all $x\in\mathbb{R}^d$. Define

$$M_m(x) := \exp\{\langle x, S_m\rangle - \frac{\zeta^2\left\|x\right\|^2_{V_m}}{2}\}$$

for $x\in\mathbb{R}^d$ and $m = 0,1,\ldots,\lfloor n/k\rfloor$. We have

$$\begin{aligned}
&\mathbb{E}(M_m(x)|\bar{\mathcal{G}}_m) \\
&= M_{m-1}(x)\exp\left\{-\frac{\zeta^2}{2}\left\|x\right\|^2_{\bar{X}_{mk+1}\bar{X}_{mk+1}^\top}\right\} \\
&\qquad\qquad \times\mathbb{E}\left(\exp\left\{\eta_{mk+1}\langle x,\bar{X}_{mk+1}\rangle\right\}\Big|\bar{\mathcal{G}}_m\right) \qquad (57) \\
&\leq M_{m-1}(x) \qquad (58)
\end{aligned}$$

where (57) follows from the fact that $\bar{X}_{mk+1}$ is $\bar{\mathcal{G}}_m$-measurable, and (58) follows from (56). Moreover, by (56) and $\bar{\mathcal{G}}_0$-measurability of $\bar{X}_1$, for every $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
&\mathbb{E}(M_0(x)) \\
&= \mathbb{E}\left(-\frac{\zeta^2 \left\|x\right\|^2_{\bar{X}_1\bar{X}_1^\top}}{2}\mathbb{E}\left(\exp\left\{\langle x, \bar{X}_1\rangle\eta_1\right\}\Big|\bar{\mathcal{G}}_0\right)\right) \\
&\leq 1. \qquad (59)
\end{aligned}$$

Let $W : \Omega \to \mathbb{R}^d$ be a d-dimensional Gaussian random vector with mean $\mathbf{0} \in \mathbb{R}^d$ and covariance matrix $(\zeta^2\lambda)^{-1}I \in \mathbb{R}^{d\times d}$; denote by P_W its distribution on $\mathbb{R}^d$. Define

$$\widetilde{M}_m := \int_{\mathbb{R}^d} M_m(x) dP_W(x) \qquad (60)$$

for each $m \in 0, 1, \ldots, \lfloor n/k \rfloor$. Observe that by (59) and Fubini's theorem we have

$$\begin{aligned}
\mathbb{E}\widetilde{M}_0 &= \mathbb{E}\left(\int_{\mathbb{R}^d} M_0(x) dP_W(x)\right) \\
&= \int_{\mathbb{R}^d} \mathbb{E}M_0(x) dP_W(x) \\
&\leq 1 \qquad (61)
\end{aligned}$$

Furthermore, by completing the square in the integrand, we can write

$$\begin{aligned}
&\widetilde{M}_m \\
&= \exp\left\{\frac{1}{2}\left\|S_m\right\|^2_{\zeta^2(\lambda I + V_m)^{-1}}\right\}\sqrt{\frac{\lambda^d}{\det(\lambda I + V_m)}} \qquad (62)
\end{aligned}$$

On the other hand, by Fubini's theorem together with (58), we have that $\widetilde{M}_m$ is a non-negative super-martingale, i.e.

$$\begin{aligned}
\mathbb{E}(\widetilde{M}_m|\bar{\mathcal{G}}_m) &= \int_{\mathbb{R}^d} \mathbb{E}(M_m(x)|\bar{\mathcal{G}}_m) dP_W \\
&\leq \int_{\mathbb{R}^d} M_{m-1}(x) \\
&= \widetilde{M}_{m-1}. \qquad (63)
\end{aligned}$$

As a result, by Doob's maximal inequality (see, e.g. [14, Theorem 7.3.1 - pp. 132]) for every $\delta > 0$ it holds that

$$\begin{aligned}
&\Pr\left(\sup_{m\in\mathbb{N}} \log \widetilde{M}_m \geq \log(1/\delta)\right) \\
&\qquad = \Pr\left(\sup_{m\in\mathbb{N}} \widetilde{M}_m \geq \frac{1}{\delta}\right) \\
&\qquad \leq \delta\mathbb{E}\widetilde{M}_0 \qquad (64) \\
&\qquad \leq \delta \qquad (65)
\end{aligned}$$

where the last inequality follows from (61). Consider the event E defined as

$$E := \Big\{ \exists m : \|S_m\|^2_{\zeta^2(\lambda I + V_m)^{-1}} \geq 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\lambda I + V_m)}{\lambda^d}\right) \Big\}.$$

By (62) and (64), we have

$$\Pr(E) \leq \delta. \tag{66}$$

Define b_n^δ such that

$$\sqrt{b_n^\delta} := 2\sqrt{\lambda}\,\|\theta_t\|_{\mathcal{L}_\infty} + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{n}{k\lambda d}\right)}$$

and let

$$C_m := \left\{ \theta \in \Theta : \left\| \theta - \widehat{\theta}_m \right\|^2_{\zeta^2(\lambda I + V_m)} \leq b_n^\delta \right\}. \tag{67}$$

By (55) and (66), with probability at least $1-\delta$ it holds that,

$$\begin{aligned} &\left\| \widehat{\theta}_m - \theta^* \right\|_{\zeta^2(\lambda I + V_m)} \\ &\leq \zeta\sqrt{\lambda} + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\lambda I + V_m)}{\lambda^d}\right)} \\ &\leq b_n^\delta \end{aligned} \tag{68}$$

where the second inequality follows from the definition of ζ as well as from [4, Equation 20.9, pp. 261]. Then, it immediately follows that

$$\Pr(\{\exists m : \theta^* \notin C_m\}) \leq \delta. \tag{69}$$

Consider the instantaneous regret

$$r_{mk+1} := \langle \theta^*, \bar{X}^*_{mk+1} - \bar{X}_{mk+1} \rangle$$

of $\bar{\boldsymbol{\pi}}$ for $m = 1, \ldots, \lfloor n/k \rfloor - 1$, where $\bar{X}^*_m$ is an optimal action at $mk+1$ so that

$$\bar{X}^*_m \in \operatorname*{argmax}_{x \in C_{m-1}} \langle \theta^*, x \rangle$$

almost surely. Let us recall that the algorithm selects

$$\bar{X}_{mk+1} \in \operatorname*{argmax}_{x \in \mathcal{A}} \max_{\theta \in C_{\max\{0, m-1\}}} \langle \theta, x \rangle.$$

With probability at least $1-\delta$ we have,

$$\langle \theta^*, \bar{X}^*_{mk+1} \rangle \leq \max_{\theta \in C_{\max\{0,m-1\}}} \langle \theta, \bar{X}^*_{mk+1} \rangle \tag{70}$$

$$\leq \max_{\theta \in C_{\max\{0,m-1\}}} \langle \theta, \bar{X}_{mk+1} \rangle. \tag{71}$$

Fix some

$$\bar{\theta}_{mk+1} \in \underset{\theta \in C_{\max\{0,m-1\}}}{\operatorname{argmax}} \langle \theta, \bar{X}_{mk+1} \rangle.$$

With probability at least $1-\delta$ it holds that,

$$\begin{aligned}
&r_{mk+1} \\
&\leq \langle \bar{\theta}_{mk+1} - \theta^*, \bar{X}_{mk+1} \rangle \\
&\leq \langle \bar{\theta}_{mk+1} - \theta^*, \bar{\theta}_{mk+1} - \theta^* \rangle^{1/2} \langle \bar{X}_{mk+1}, \bar{X}_{mk+1} \rangle^{1/2} \\
&\leq \left\| \bar{\theta}_{mk+1} - \theta^* \right\|_{\zeta^2(\lambda I + V_{\max\{0,m-1\}})} \\
&\qquad\qquad \times \left\| \bar{X}_{mk+1} \right\|_{\zeta^2(\lambda I + V_{\max\{0,m-1\}})^{-1}} \\
&\leq \sqrt{b_n^\delta} \times \zeta \lambda^{-1/2} \left\| \bar{X}_{mk+1} \right\|_2 \\
&\leq \zeta \sqrt{\frac{b_n^\delta}{\lambda}}
\end{aligned} \tag{72}$$

where in much the same way as with (55), (72) follows from noting that, V_m is positive semi-definite so the matrix order is reversed through inversion, i.e.

$$x^\top (\lambda I + V_m)^{-1} x \leq x^\top (\lambda I)^{-1} x$$

for all $x \in \mathbb{R}^d$. It follows that

$$\begin{aligned}
\sum_{m=1}^{\lfloor n/k \rfloor - 1} r_{mk+1} &\leq \sqrt{\frac{n}{k} \sum_{m=1}^{\lfloor n/k \rfloor - 1} r_{mk+1}^2} \\
&\leq 2 \left\| \theta_t \right\|_{\mathcal{L}_\infty} \sqrt{\frac{n b_n^\delta}{k\lambda}}
\end{aligned} \tag{73}$$

which in turn yields

$$\begin{aligned}
&\mathcal{R}_{\bar{\boldsymbol{\pi}}}(n) \\
&\leq (1-\delta) k \left(\left\| \theta_t \right\|_{\mathcal{L}_\infty} + \sum_{m=1}^{\lfloor n/k \rfloor - 1} \mathbb{E} r_{mk+1} \right) + \delta \left\| \theta_t \right\|_{\mathcal{L}_\infty} \\
&\leq \left\| \theta_t \right\|_{\mathcal{L}_\infty} \left((1-\delta) k \left(1 + 2\sqrt{\frac{n b_n^\delta}{k\lambda}} \right) + \delta \right)
\end{aligned} \tag{74}$$

By (50) and (74), taking $\delta = 1/n$, and noting that $\varphi_k \leq a e^{-\gamma k}$ for some $a, \gamma \in (0, \infty)$, we have,

$$\begin{aligned}
\frac{\mathcal{R}_{\boldsymbol{\pi}}(n)}{\left\| \theta_t \right\|_{\mathcal{L}_\infty}} \leq \frac{1}{n} &+ 6n^2 a e^{-\gamma k} \\
&+ k(1 + 4\sqrt{n} \left\| \theta_t \right\|_{\mathcal{L}_\infty}) \\
&+ k \sqrt{\frac{8dn \log(n(1 + \frac{n}{\lambda d}))}{\lambda}}
\end{aligned} \tag{75}$$

Optimizing (75) for k we obtain $k^\star$ given by

$$\left\lceil \log\left(\frac{6a\gamma n^2}{1+4\sqrt{n}\,\|\theta_t\|_{\mathcal{L}_\infty}+\sqrt{\frac{8dn\log(n(1+\frac{n}{\lambda d}))}{\lambda}}} \right)^{1/\gamma} \right\rceil \tag{76}$$

For $n \geq \left\lceil \frac{3a\gamma\sqrt{\lambda}}{2\sqrt{\lambda}\|\theta_t\|_{\mathcal{L}_\infty}+\sqrt{2}} \right\rceil$ and $k = \max\{1, k^\star\}$ we have,

$$\begin{aligned} &\frac{\mathcal{R}_{\boldsymbol{\pi}}(n)}{\|\theta_t\|_{\mathcal{L}_\infty}} \\ &\leq \frac{1}{n} + \left(\frac{12(\sqrt{2\lambda}+4\sqrt{2\lambda}\,\|\theta_t\|_{\mathcal{L}_\infty}+1)}{\gamma\sqrt{2\lambda}} \right) \\ &\qquad \times \sqrt{2dn\log(n(1+\frac{n}{\lambda d}))}\log(n). \end{aligned}$$

□

Algorithm 2 LinMix-UCB (∞ - horizon)

Input: regularization parameter λ; φ-mixing rate parameters: $a, \gamma \in (0, \infty)$

$n_0 \leftarrow \max\left\{1, \left\lceil \frac{3a\gamma\sqrt{\lambda}}{2\sqrt{\lambda}\|\theta_t\|_{\mathcal{L}_\infty}+\sqrt{2}} \right\rceil \right\}$

for $i = 0, 1, 2, \ldots$ **do**

$n_i \leftarrow 2^i n_0$

Run Algorithm1$(n_i, \lambda, a, \gamma)$ from $t = (2^i - 1)n_0 + 1$ to $t = (2^{i+1} - 1)n_0$

Algorithm 1 can be turned into an infinite-horizon strategy using a classical doubling-trick. The procedure is outlined in Algorithm 2 below. As in the finite-horizon setting, the algorithm aims to minimize the regret with respect to (2), in the case where the φ-mixing coefficients of the process $(\theta_t,\ t \in \mathbb{N})$ satisfy

$$\varphi_m \leq ae^{-\gamma m}$$

for some fixed $a, \gamma \in (0, \infty)$ and all $m \in \mathbb{N}$.

The algorithm works as follows. At every round $i = 0, 1, 2, \ldots$ a horizon is determined as

$$n_i = 2^i n_0$$

with

$$n_0 := \max\left\{1, \left\lceil \frac{3a\gamma\sqrt{\lambda}}{2\sqrt{\lambda}\,\|\theta_t\|_{\mathcal{L}_\infty}+\sqrt{2}} \right\rceil \right\}$$

and the algorithm plays the finite-horizon strategy specified in Algorithm 1 from $t = \sum_{j=0}^{i-1} n_j$ to $t = \sum_{j=0}^{i} n_j$. The regret of this algorithm is given in Theorem 2.

Theorem 2. *Suppose, as in Theorem 1, that there exist $a, \gamma \in (0, \infty)$ such that φ-mixing coefficients of the stationary process $(\theta_t,\ t \in \mathbb{N})$ satisfy $\varphi_m \leq ae^{-\gamma m}$ for all $m \in \mathbb{N}$. Then, the regret (with respect to $\widetilde{\nu}_n$) of LinMix-UCB (infinite horizon) after n rounds of play satisfies*

$$\begin{aligned}&\frac{\mathcal{R}_{\boldsymbol{\pi}}(n)}{2\left\|\theta_t\right\|_{\mathcal{L}_\infty}}\\ &\leq n_0 + C(\log_2(n+1)+1)\log(2(n+1))\\ &\quad\times\sqrt{(n+1)d\times\log\left(2(n+1)(1+\frac{2(n+1)}{\lambda d})\right)}\end{aligned}$$

where $C := \left(\frac{12(\sqrt{2\lambda}+4\sqrt{2\lambda}\|\theta_t\|_{\mathcal{L}_\infty}+1)}{\gamma\sqrt{2\lambda}}\right)$, *and*

$$n_0 := \max\left\{1, \left\lceil \frac{3a\gamma\sqrt{\lambda}}{2\sqrt{\lambda}\left\|\theta_t\right\|_{\mathcal{L}_\infty}+\sqrt{2}}\right\rceil\right\}$$

and $\lambda > 0$ is the regularization parameter.

Proof. For $n \in \mathbb{N}$, let

$$j(n) := \min\{i \in \mathbb{N} : \sum_{i=0}^{i} n_i \geq n\}.$$

Recall that the algorithm plays the finite-horizon strategy of Algorithm 1 during non-overlapping intervals of length $n_i = 2^i n_0,\ i = 0, 1, 2, \ldots$ with

$$n_0 := \max\left\{1, \left\lceil \frac{3a\gamma\sqrt{\lambda}}{2\sqrt{\lambda}\left\|\theta_t\right\|_{\mathcal{L}_\infty}+\sqrt{2}}\right\rceil\right\}.$$

By Theorem 1 and that

$$\sum_{i=0}^{j(n)} \frac{1}{n_i} \leq n_0 \sum_{i=0}^{\infty} 2^{-i} \leq 2n_0,$$

we have the following upper-bound on $\frac{\mathcal{R}_{\boldsymbol{\pi}}(n)}{\|\theta_t\|_{\mathcal{L}_\infty}}$,

$$2n_0 + C\sum_{i=0}^{j(n)} \log(n_i)\sqrt{2dn_i\log(n_i(1+\frac{n_i}{\lambda d}))} \tag{77}$$

with the constant C as given in the theorem statement. The result follows from (77), and the fact that as follows from the definition of $j(n)$ we have,

$$\sum_{i=0}^{j(n)} n_i = n_0(2^{j(n)+1}-1) \geq n$$

so that

$$j(n) = \left\lceil \log_2(\frac{n}{n_0}+1)\right\rceil$$

and $2^{j(n)} \leq 2(n+1)$. □

IV. Outlook

We have formulated a generalization of both the classical linear bandits with iid noise, and the finite-armed restless bandits. In the problem that we have considered, an unknown $\mathbb{R}^d$-valued stationary φ-mixing sequence of parameters $(\theta_t,\ t \in \mathbb{N})$ gives rise to the payoffs. We have provided an approximation of the optimal restless linear bandit strategy via a UCB-type algorithm, in the case where the process $(\theta_t,\ t \in \mathbb{N})$ has an exponential mixing rate. The regret of the proposed algorithm, namely LinMix-UCB, with respect a more relaxed oracle which always plays a multiple of $\mathbb{E}\theta_t$, is shown to be

$$\mathcal{O}\left(\sqrt{dn\,\mathrm{polylog}(n)}\right).$$

Our results differ from that of [8] which tackles the (simpler) finite-armed restless φ-mixing bandit problem, in that they do not require an exponential φ-mixing rate in order to ensure an $\mathcal{O}(\log n)$ regret with respect to the highest stationary mean in their setting. With only a finite number of arms to play, they are able to base their approach on a Hoeffding-type inequality for φ-mixing processes. However, this does not extend to our linear bandit framework. A natural open problem is the derivation of a lower-bound on the regret with respect to ν_n (or with respect to the best switching strategy in the finite-armed setting of [8]).

A. Knowledge of the mixing rates.

Our algorithms require (bounds on) the true φ mixing rate parameters. Although this assumption is standard in time-series analysis, an intriguing objective would be to relax this assumption and infer the mixing parameters while maximizing the expected cumulative payoff. However, for reasons outlined below, this is a challenging endeavor that is beyond the scope of the present paper.

The problem of estimating the mixing coefficients of a process from its sample-paths has only recently garnered attention, and the results so far concern the full-information setting. For example, [15] have proposed strongly consistent estimators for the α and β mixing coefficients of a stationary ergodic process. Their results are necessarily asymptotic, as it is provably impossible to obtain rates of convergence for empirical measures in this general class of processes. However, since finite-time analysis is crucial in a bandit problem, convergence rates are required for the estimators to be effectively deployed as part of a bandit strategy. On the other hand, for the more restrictive class of geometrically ergodic Markov chains, convergence rates for β-mixing estimators can be achieved under density conditions or when the state space is finite [13], [16]. Furthermore, the existing estimators are designed for fully observed sample-paths and cannot readily provide estimates in the bandit context considered in the present paper where the process $(\theta_t, t \in \mathbb{N})$ is only partially observed.

Estimating the mixing coefficients while playing a restless bandit strategy requires more care compared to estimation from a fully observed sample-path. As demonstrated in [8, Example 1], in this framework, a policy can introduce strong couplings between past and future payoffs, resulting in a payoff sequence with a completely different dependency structure than that of the original process.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2025.3533299

Finally, estimating φ-mixing coefficients is significantly more challenging than estimating β-mixing coefficients. A key challenge in estimating φ_m lies in conditioning on potentially rare events with small probabilities. To the best of our knowledge, no estimator for φ_m exists, even in the full-information setting.

B. Towards a relaxation.

In light of the challenges in consistently estimating φ_m, one might consider incorporating a sequential hypothesis test into Algorithm 2 to determine (at least asymptotically) whether φ_m is indeed upper bounded by $ae^{-\gamma m}$, without directly estimating φ_m. Noting that, in general, $\beta_m \leq \varphi_m,\ m \in \mathbb{N}$ (see, e.g. [5]), and under some mild assumptions such as the summability of the φ-mixing coefficients, one may be able to adopt the asymptotically consistent hypothesis test for the β-mixing rate from [15], which is based on estimates of β_m rather than φ_m. In order to use this test as part of a bandit strategy, the Type I and Type II errors of the test would need to be controlled; this can be done if an upper-bound on $\sum_{m\in\mathbb{N}} \varphi_m$ is known and using an appropriate concentration bound, such as [6, Corollary 2.1]. As a result, we conjecture that Algorithm 2 can be modified so that at each $i \in \mathbb{N}$, some $\sqrt{n_i}$ samples are allocated to test the β-mixing rate. Recall that $n_i,\ i \in \mathbb{N}$ is an increasing sequence, with each element representing the length of the i^{th} batch on which the finite-horizon algorithm is executed. The modified algorithm would then continue to apply the bandit strategy on $n_i - \sqrt{n_i}$ samples, operating under the null hypothesis that $\varphi_m \leq ae^{-\gamma m}$. The process would halt once the null hypothesis is rejected. Maintaining the number of samples per test at the order of $\sqrt{n_i}$ ensures that the algorithm's regret remains unaffected. Although this roadmap appears promising, it requires thorough exploration to assess its feasibility. We defer this investigation to future work.

REFERENCES

[1] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[2] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[3] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[4] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[5] R. C. Bradley, *Introduction to Strong Mixing Conditions*. Kendrick Press, 2007, vol. 1,2, 3.

[6] E. Rio, *Asymptotic theory of weakly dependent random processes*. Springer, 2017, vol. 80.

[7] R. Ortner, D. Ryabko, P. Auer, and R. Munos, "Regret bounds for restless markov bandits," *Theoretical Computer Science*, vol. 558, pp. 62–76, 2014.

[8] S. Grünewälder and A. Khaleghi, "Approximations of the restless bandit problem," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 514–550, 2019.

[9] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.

[10] Q. Chen, N. Golrezaei, and D. Bouneffouf, "Non-stationary bandits with auto-regressive temporal dependency," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7895–7929, 2023.

[11] P. Mattila, *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. Cambridge university press, 1999, no. 44.

[12] H. C. Berbee, "Random walks with stationary increments and renewal theory," *Mathematisch Centrum*, 1979.

[13] S. Grünewälder and A. Khaleghi, "Estimating the mixing coefficients of geometrically ergodic markov processes," *arXiv preprint arXiv:2402.07296*, 2024.

[14] A. N. Shiryaev, *Probability-2*. Springer, 2019, vol. 95.

[15] A. Khaleghi and G. Lugosi, "Inferring the mixing properties of a stationary ergodic process from a single sample-path," *IEEE Transactions on Information Theory*, 2023.

[16] G. Wolfer and P. Alquier, "Optimistic estimation of convergence in markov chains with the average-mixing time," *arXiv preprint arXiv:2402.10506*, 2024.