Linear Bandits with Non-i.i.d. Noise

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the linear stochastic bandit problem, relaxing the standard i.i.d. assumption on the observation noise. As an alternative to this restrictive assumption, we allow the noise terms across rounds to be sub-Gaussian but interdependent, with dependencies that decay over time. To address this setting, we develop new confidence sequences using a recently introduced reduction scheme to sequential probability assignment, and use these to derive a bandit algorithm based on the principle of optimism in the face of uncertainty. We provide regret bounds for the resulting algorithm, expressed in terms of the decay rate of the strength of dependence between observations. Among other results, we show that our bounds recover the standard rates up to a factor of the mixing time for geometrically mixing observation noise.

1 Introduction

The linear bandit problem (Abe and Long, 1999; Auer, 2003) is an instance of a multi-armed bandit framework, where the expected reward is linear in the feature vector representing the chosen arm. More concretely, it is a sequential decision-making problem, where an agent each round picks an arm X_t, and receives a reward $Y_t = \langle \theta^\star, X_t \rangle + \varepsilon_t$, with $\theta^\star$ a fixed parameter unknown to the agent, and ε_t zero-mean random noise. This framework has gained significant attention in the literature as it yields analytic tools that can be applied to several concrete applications, such as online advertising (Abe et al., 2003), recommendation systems (Li et al., 2010; Korkut and Li, 2021), and dynamic pricing (Cohen et al., 2020).

A popular strategy to tackle linear bandits leverages the principle of *optimism in the face of uncertainty*, via upper confidence bound (UCB) algorithms. The idea of optimism can be traced back to Lai and Robbins (1985), and its application to linear bandits was already advanced by Auer (2003). Since then, this approach has been improved and analysed by several works (Abbasi-Yadkori et al., 2011; Lattimore and Szepesvári, 2020; Flynn et al., 2023). This class of methods requires constructing an adaptive sequence of confidence sets that, with high probability, contain the true parameter $\theta^\star$. Each round, the agent selects the arm maximising the expected reward under the most optimistic parameter (in terms of reward) in the current confidence set. UCB-based algorithms have become popular as they are often easy to implement and come with tight worst-case regret guarantees.

For a UCB algorithm to perform well, it is necessary that the confidence sets are tight, which can be ensured by taking advantage of the structure of the problem. In this paper, our focus is on studying various assumptions on the observation noise. A commonly studied situation is when $(\varepsilon_t)_{t\geq 0}$ consists of a sequence of i.i.d. realisations of some bounded or sub-Gaussian random variable (see Lattimore and Szepesvári, 2020, Chapter 20). Often, the standard analysis can be extended to the case in which the realisation are not independent, but conditionally centred and sub-Gaussian (Abbasi-Yadkori et al., 2011). Yet, in real-world settings, this assumption is often unrealistic, as one can expect the presence of interdependencies among the noise at different rounds. For instance, in the context of advertisement selection, the noise models the ensemble of external factors that influence the

Submitted to the 18th European Workshop on Reinforcement Learning (EWRL 2025). Do not distribute.

user's choice on whether to click or not an ad. The i.i.d. assumption implies that across different rounds these external factors are completely independent. In practice, the user choice will be affected by temporally correlated events, such as recent browsing history or exposure to similar content. Therefore, a more realistic assumption is to allow the dependencies to decay with time, rather than being completely absent. This way to model dependencies, often referred to as *mixing*, is common to study concentration for sums of non-i.i.d. random variables, with applications to machine learning (Bradley, 2005; Mohri and Rostamizadeh, 2008; Abélès et al., 2025).

In the present paper we relax the assumption that the noise is conditionally zero-mean in the bandit problem, and we allow for the presence of dependencies. Concretely, we replace the standard conditionally sub-Gaussian setting with a more general formulation that accounts for conditional dependence of the noise on the past, by introducing a natural notion of *mixing sub-Gaussianity*. Within this context, we introduce a UCB algorithm for which we rigorously establish regret guarantees. There are two key challenges for our approach: constructing a valid confidence sequence under dependent noise, and deriving a regret upper bound for the UCB algorithm that we propose.

We derive the confidence sequence by adapting the *online-to-confidence-sets* technique to accommodate temporal dependencies in the noise. This approach, originally introduced by Abbasi-Yadkori et al. (2011) and recently extended and improved (Jun et al., 2017; Lee et al., 2024; Clerico et al., 2025), involves constructing an abstract online learning game whose regret guarantees can be turned into a confidence sequence. To deal with the dependencies in the noise, we modify the standard online-to-confidence-sets framework by introducing delays in the feedback received within the abstract online game. This approach is inspired by the recent work of Abélès et al. (2025) on extending online-to-PAC conversions to non-i.i.d. mixing data sets in the context of deriving generalisation bounds for statistical learning. There, a delayed-feedback trick similar to ours is employed to derive statistical guarantees (generalisation bounds) from an abstract online learning game.

For the regret analysis of the bandit algorithm, we also need to face some challenges due to the correlated observation noise. We address these by introducing delays into the decision-making policy as well. This makes our approach superficially similar to algorithms used in the rich literature on bandits with delayed feedback (see, e.g., Vernade et al., 2020a; Howson et al., 2023). These works consider delay as part of the problem statement and not part of the solution concept, and are thus orthogonal to our work. In particular, a simple adaptation of results from this literature would not suffice for dealing with dependent observations, which we tackle by developing new concentration inequalities. Another line of work that is conceptually related to ours is that of non-stationary bandits (Garivier and Moulines, 2008; Russac et al., 2019). In that setting, the parameter vector $\theta_t^\star$ evolves in time according to a nonstationary stochastic process, and the observation noise remains i.i.d., once again making for a rather different problem with its own challenges. Namely, the main obstacle to overcome is that comparing with the optimal sequence of actions becomes impossible unless strong assumptions are made about the sequence of parameter vectors. A typical trick to deal with these nonstationarities is to discard old observations (which may have been generated by a very different reward function), and use only recent rewards for decision-making. This is the polar opposite of our approach that is explicitly *disallowed* to use recent rewards, which clearly highlights how different these problems are. That said, there exists an intersection between the worlds of delayed and nontationary bandits (Vernade et al., 2020b), and thus we would not discard the possibility of eventually building a bridge between bandits with nonstationary reward functions and bandits with nonstationary observation noise. For simplicity, we focus on the second of these two components in this paper.

Notation. Throughout the paper, we will often use the following notations. For u and v in $\mathbb{R}^p$, we let $\langle u, v\rangle$ denote their dot product. $\|u\|_2 = \sqrt{\langle u, u\rangle}$ is the Euclidean norm, while for a non-negative definite $(p \times p)$-matrix A, $\|u\|_A = \sqrt{\langle u, Au\rangle}$ is a semi-norm (a norm if the matrix is strictly positive definite). For $r > 0$, $\mathcal{B}(r)$ denotes the closed centred Euclidean ball in $\mathbb{R}^p$ with radius r. Given a non-empty set $U \subseteq \mathbb{R}^p$, we let Δ_U denote the space of (Borel) probability measures on $\mathbb{R}^p$ whose support in U. Finally, $(u_t)_{t \geq t_0}$ denotes a sequence indexed on the integers, with t_0 its smallest index.

2 Preliminaries on linear bandits

We consider a version of the classic problem of regret minimisation in stochastic linear bandits, where an agent needs to make a sequence of decisions (or pick an *arm*) from a given contextual decision set that may change over the sequence of rounds. We assume that the environment is oblivious to the actions of the agent, in the sense that the decision sets are determined in advance, and do not depend neither on the realisations of the noise nor on the agent's arm-selection strategy.

Concretely, we define the problem as follows. Let $\theta^\star \in \mathbb{R}^p$ be a parameter vector that is unknown to the learning agent. We assume as known an upper bound $B > 0$ on its euclidean norm (namely, $\theta^\star \in \mathcal{B}(B)$). Fix a sequence of decision sets $(\mathcal{X}_t)_{t\geq 1}$ in $\mathbb{R}^p$. We assume that for all t we have $\mathcal{X}_t \subseteq \mathcal{B}(1)$. At each round t, the agent is required to pick an arm $X_t \in \mathcal{X}_t$, and receives the reward $Y_t = \langle \theta^\star, X_t\rangle + \varepsilon_t$. The sequence $(\varepsilon_t)_{t\geq 1}$ represents the random feedback noise. The noise across different rounds is typically assumed to be conditionally centred and to have well behaved tails. For instance, a common assumption is to ask that $\mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}]$ is centred and sub-Gaussian, where $\mathcal{F}_t = \sigma(\varepsilon_1, \dots, \varepsilon_t)$ is the σ-field generated by the noise. This is the assumption this work relaxes. We also remark that, more generally, one can consider the case where the X_t as well are randomised, namely contain additional randomness that is not included in the noise. To take this into account, one can add this other source or randomness in the filtration. However, since in our case we will only consider a non-randomised bandit algorithm, we omit this to simplify our analysis.

The agent aims to find a good strategy to pick arms X_t that lead to a high expected T-round reward $\sum_{t=1}^T \langle X_t, \theta^\star\rangle$. To compare their performance to that of an agent playing each round the best available arm (in expectation), we define the *regret* after T rounds as

$$\mathrm{Reg}(T) = \sum_{t=1}^{t} \sup_{x\in\mathcal{X}_t} \left(\langle x, \theta^\star\rangle - \langle X_t, \theta^\star\rangle\right).$$

A common approach to tackle the linear bandit problem is to follow an *upper confidence bound* (UCB) strategy. This involves the following protocol. At each round t, we first derive a confidence set $\mathcal{C}_{t-1}$, based on the arm-reward pairs $(X_s, Y_s)_{s\leq t-1}$. This is a random set (as it depends on the past noise realisations), which must be constructed ensuring that $\theta^\star \in \mathcal{C}_{t-1}$ with high probability. More precisely, the regret can be effectively controlled if one can ensure that $\theta^\star$ uniformly belongs to every set $(\mathcal{C}_t)_{t\geq 1}$, with high probability (a property often referred to as *anytime validity*). Then, for every available arm x, we let

$$\mathrm{UCB}_{\mathcal{C}_{t-1}}(x) = \max_{\theta\in\mathcal{C}_{t-1}} \langle x, \theta\rangle .$$

By definition, this is a high-probability upper bound on $\langle x, \theta^\star\rangle$, which justifies the name "upper confidence bound". The idea is then to *optimistically* pick as $X_t \in \mathcal{X}_t$ the arm maximising $\mathrm{UCB}_{\mathcal{C}_{t-1}}$.

A key technical challenge in designing a UCB algorithm is to construct the anytime valid confidence sequence $(\mathcal{C}_t)_{t\geq 1}$. Typically, under sub-Gaussian assumptions on the noise, these sets take the form of an ellipsoid, centred on a (regularised) maximum likelihood estimator. Explicitly, we often have

$$\mathcal{C}_t = \left\{\theta \in \Theta \,:\, \|\theta - \widehat{\theta}_t\|_{V_t}^2 \leq \beta_t^2\right\},$$

where $\widehat{\theta}_t$ is the least-squares estimator of $\theta^\star$, V_t is the *feature-covariance* matrix and β_t is a radius carefully chosen so that the high-probability coverage requirement is satisfied. In this work, to construct the csonfidence sets we will leverage an *online-to-confidence-set-conversion* approach, a method that reduces the problem of proving statistical concentration bounds to proving existence of well-performing algorithms for an associated game of *sequential probability assignment*. We refer to Section 4 for more details on our technique to construct the confidence sequence.

3 Linear bandits with non-i.i.d. observation noise

We study a variant of the standard linear stochastic bandit problem where the observation-noise variables feature dependencies across different rounds. We focus on the case of weakly stationary noise, meaning we assume all the ε_t to have the same marginal distribution. However, the core assumption we make is what we call *mixing sub-Gaussianity*. This provides a way to control how dependencies decay as the time between two observations increases. It is defined in terms of a sequence of mixing coefficients ϕ_d, which quantify this decay.

Assumption 1 (Mixing sub-Gaussianity). *Fix $\sigma > 0$ and let $\phi = (\phi_d)_{d\geq 0}$ be a non-negative and non-increasing sequence. We say that the random sequence $(\epsilon_t)_{t\geq 1}$ is (σ, ϕ)-*mixing sub-Gaussian *if ε_t is centred and σ-sub-Gaussian for every t, and, for all $d \geq 0$ and all $t > d$, we have*

$$\left|\mathbb{E}\left[\epsilon_t \,|\mathcal{F}_{t-d}\right]\right| \leq \phi_d \tag{1}$$

and
$$\mathbb{E}\left[\exp\lambda(\epsilon_t - \mathbb{E}\left[\epsilon_t \,|\mathcal{F}_{t-d}\right])\,|\mathcal{F}_{t-d}\right] \le e^{\frac{\lambda^2\sigma^2}{2}}\,, \qquad \forall\lambda > 0\,. \tag{2}$$

Clearly, the above assumption generalises the standard conditionally sub-Gaussian assumption (that can be recovered by setting $\phi_d = 0$ for all t), sometimes considered in the bandit literature. Although this might look like an unusual mixing assumption, it is very natural for our problem at hand, and can be weaker than standard mixing hypotheses. For instance, if the noise sequence is φ-mixing (see Bradley, 2005) and each ε_t is centred and bounded in $[-a, b]$, it is straightforward to check that $|\mathbb{E}[\varepsilon_t|\mathcal{F}_{t-d}]| \le (a+b)\phi_d$, and so Assumption 1 is satisfied since the boundedness automatically implies sub-Gaussianity. In the rest of the paper we assume $\sigma = 1$ for simplicity.

Under Assumption 1, we can build the confidence sequence needed for our UCB algorithm. We state this result below, but defer the explicit derivation to Section 4 (see Corollary 1 there).

Proposition 1. *For some given ϕ, let the noise satisfy Assumption 1 with $\sigma = 1$. Fix $\delta \in (0,1)$, $\lambda > 0$, and $d \ge 1$. For $t \ge 1$ let*

$$\mathcal{C}_t = \left\{\theta \in \mathcal{B}(B) \,:\, \tfrac{1}{2}\|\theta - \widehat{\theta}_t\|^2_{V_t} \le \tfrac{dp}{2}\log\tfrac{(B+1)^2 e \max(dp, t+d)}{dp} + 2\lambda B^2 + t\phi_d(B+1) + d\log\tfrac{d}{\delta}\right\}\,,$$

where $V_t = \sum_{s=1}^t X_t X_t^\top + \lambda\mathrm{Id}$, and $\widehat{\theta}_t =$ arg min$_{\theta\in\mathcal{B}(B)}\sum_{s=1}^t(\langle\theta, X_t\rangle - Y_t)^2$. Then, $(\mathcal{C}_t)_{t\ge1}$ is an anytime valid confidence sequence, in the sense that

$$\mathbb{P}\big(\theta^\star \in \mathcal{C}_t\,,\ \forall t \ge 1\big) \ge 1 - \delta\,.$$

Leveraging the confidence sequence above, we can define a UCB approach for our problem (Algorithm 1). At a high level, the algorithm operates by taking the confidence sets defined in Proposition 1, and selecting the arm optimistically, as in the standard UCB. A key point is that a delay d is introduced, which at round t restricts the agent to use only the information available from the first $t-d$ rounds. Although the actual technical reason behind this restriction will become fully clear only with the analysis of the coming sections, one can intuitively think of it as a way to prevent overfitting to recent noise, which might be highly correlated. If d is sufficiently large, the noise observed in each round t will be sufficiently decorrelated from the previous observations, which allows accurate estimation and uncertainty quantification of the true parameter $\theta^\star$ and the associated rewards.

Algorithm 1 Mixing-LinUCB

```
set d > 0
for i ∈ {1, 2, ... d} do
    play an arbitrary X_i and observe Y_i
end for
for t ∈ {d + 1, ... } do
    X_t = arg max_{x∈X_t} UCB_{C_{t−d}}(x), where C_{t−d} is as in Proposition 1
    play X_t and observe reward Y_t
end for
```

In Section 5 we provide a detailed analysis of the regret of the algorithm that we proposed. For instance, assuming that the mixing coefficients decay exponentially as $\phi_d = Ce^{-d/\tau}$ (*geometric mixing*), we show that the regret can be upper bounded in high probability as

$$\mathrm{Reg}(T) \le \mathcal{O}\left(\tau p\sqrt{T}\log(T)^2 + \tau\log T\sqrt{pT\log T}\right).$$

We refer to Theorem 2 and Corollary 2 in Section 5 for more details.

4 Constructing the confidence sequence

In this section we derive a confidence sequence for linear models with non-i.i.d. noise. First, we briefly describe the online-to-confidence-set conversion scheme from Clerico et al. (2025), which serves as our starting point. We then extend this technique to handle mixing noise.

4.1 Online-to-confidence set conversion for i.i.d. data

Before proceeding for the analysis of mixing sub-Gaussian noise, which is the focus of this work, we start by describing how to derive a confidence sequence when the noise is independent (or conditionally) centred and sub-Gaussian across different rounds, as in Clerico et al. (2025). The online-to-confidence sets framework that we consider instantiates an abstract game played between an *online learner* and an *environment*. We define the squared loss $\ell_s(\theta) = \frac{1}{2}(\langle \theta, X_s \rangle - Y_s)^2$. For each round $s = 1, \ldots, t$, the following steps are repeated:

1. the environment reveals X_s to the learner;
2. the learner plays a distribution $Q_s \in \Delta_{\mathbb{R}^p}$;
3. the environment reveals Y_s to the learner;
4. the learner suffers the log loss $\mathcal{L}_s(Q_s) = -\log \int_{\mathbb{R}^p} \exp(-\ell_s(\theta)) \mathrm{d}Q_s(\theta)$.

This game is a special case of a well-studied problem called *sequential probability assignment* (Cesa-Bianchi and Lugosi, 2006). The learner can use any strategy to choose $Q_1, \ldots, Q_t$, as long as each Q_s depends only on $X_1, Y_1, \ldots, X_{s-1}, Y_{s-1}, X_s$. We define the *regret* of the learner against a (possibly data-dependent) comparator $\bar{\theta} \in \mathbb{R}^p$ as

$$\mathrm{Regret}_t(\bar{\theta}) = \sum_{s=1}^{t} \mathcal{L}_s(Q_s) - \sum_{s=1}^{t} \ell_s(\bar{\theta})\,.$$

Clerico et al. (2025) provide a regret bound upper bound (Proposition 3.1 there) for when the learner's strategy is from an *exponential weighted average* (EWA) forecaster with a centred Gaussian prior Q_1. However, to account for the presence of dependencies in our analysis, we will need the prior's support to be bounded. We hence state here a regret bound (whose proof is deferred to Appendix A.1) for the regret of an EWA forecaster with a uniform prior.

Proposition 2. *Fix $B > 0$ and consider the EWA forecaster with as prior the uniform distribution on $\mathcal{B}(B+1)$. Then, for all $\bar{\theta} \in \mathcal{B}(B)$ and any $t \geq 1$,*

$$\mathrm{Regret}_t(\bar{\theta}) \leq \frac{p}{2} \log \frac{(B+1)^2 e \max(p, t)}{p}\,.$$

We remark that, by adding and subtracting the total log loss of the learner, the excess loss of $\theta^\star$ (relative to $\bar{\theta}$) can be rewritten as

$$\sum_{s=1}^{t} \ell_s(\theta^\star) - \sum_{s=1}^{t} \ell_s(\bar{\theta}) = \mathrm{Regret}_t(\bar{\theta}) + \sum_{s=1}^{t} \ell_s(\theta^\star) - \sum_{s=1}^{t} \mathcal{L}_s(Q_s)\,. \tag{3}$$

This simple decomposition is the key idea in the online-to-confidence sets scheme.

Since the noise is conditionally sub-Gaussian and the distributions played by the online learner are predictable (Q_s cannot depend on Y_s), $\sum_{s=1}^t \ell_s(\theta^\star) - \sum_{s=1}^t \mathcal{L}_s(Q_s)$ is the logarithm of a non-negative super-martingale (cf. the no-hypercompression inequality in Grünwald, 2007 or Proposition 2.1 in Clerico et al., 2025) with respect to the noise filtration $(\mathcal{F}_t)_{t\geq 1}$. For simplicity, as already mentioned in Section 2 and since this will be the case for our bandit strategy, we assume throughout the paper that X_t is fully determined given the past noise . Henceforth, from Ville's inequality (a classical anytime valid Markov-like inequality that holds for non-negative super-martingales) one can easily derive that $\theta^\star \in \mathcal{C}_t$ (uniformly for all t) with probability at least $1 - \delta$, where

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^p : \sum_{s=1}^{t} \ell_s(\theta) - \sum_{s=1}^{t} \ell_s(\bar{\theta}) \leq \mathrm{Regret}_t(\bar{\theta}) + \log \frac{1}{\delta} \right\}\,.$$

This result can be relaxed by replacing $\mathrm{Regret}_t(\bar{\theta})$ by any known regret upper bound for the online algorithm used in the abstract game (*e.g.*, the bound of Proposition 2 for the EWA forecaster).

4.2 Confidence sequence under mixing sub-Gaussian noise

The standard online-to-confidence sets scheme relies on the fact that $\sum_{s=1}^t \ell_s(\theta^\star) - \sum_{s=1}^t \mathcal{L}_s(Q_s)$ is the logarithm of a non-negative super-martingale, whose fluctuations can be controlled uniformly in time thanks to Ville's inequality. However, this property hinges on the fact that the noise is assumed to be conditionally centred and sub-Gaussian, which now is not anymore the case. Yet, thanks to our mixing assumption, if we restrict our focus on rounds that are sufficiently far apart, the mutual dependencies get weaker, and the exponential of the sum behaves *almost* like a martingale. This insight suggests to partition the rounds into blocks, whose elements are mutually far apart, then apply concentration results to each block, and finally use a union bound to recover the desired confidence sequence spanning all rounds. We remark that this is a classical approach to derive concentration results for mixing processes, often referred to as the *blocking* technique (Yu, 1994).

In order for the online-to-confidence sets scheme to leverage the blocking strategy outlined above, the abstract online game used for the analysis must be designed in a way that is compatible with the block structure. To address this point, we adopt an approach inspired by Abélès et al. (2025), who introduced delays in the feedbacks received by the online learner in order to address a similar challenge. More precisely, we will now consider the following *delayed-feedback* version of the online game. Fix a delay $d > 0$. For each round $s = 1, \dots, t$, the following steps are repeated:

1. the environment reveals to the learner X_s, which is assumed to be $\mathcal{F}_{s-d}$-measurable;
2. the learner plays a distribution $Q_s \in \Delta_{\mathbb{R}^p}$;
3. if $s > d$, the environment reveals Y_{s-d+1} to the learner;
4. the learner suffers the log loss $\mathcal{L}_s(Q_s) = -\log \int_{\mathbb{R}^p} \exp(-\ell_s(\theta)) \mathrm{d}Q_s(\theta)$.

Note that the delay d only applies for the rewards, while Q_s can still depend on X_s. Indeed, the choice of X_s in our mixing UCB algorithm is already "delayed", as it depends on $\mathcal{C}_{t-d}$ (see Algorithm 1).

Of course, in this setting the decomposition of (3) is still valid. We now want to deal with the concentration of $\sum_{s=1}^t \ell_s(\theta^\star) - \sum_{s=1}^t \mathcal{L}_s(Q_s)$ via the blocking technique. For convenience, let us write $D_t = \ell_t(\theta^\star) - \mathcal{L}_t(Q_t)$. We denote as $S^{(i)} = (S_k^{(i)})_{k \geq 1}$ the subsequence defined as $S_k^{(i)} = \sum_{j=1}^k D_{i+(j-1)d}$. The key idea is now that each of these $S^{(i)}$ behaves as the log of a martingale, up to a cumulative remainder that accounts for the conditional mean shift in the mixing sub-Gaussianity assumption. In particular, Ville's inequality and a union bound yield the following.

Lemma 1. *Fix a delay $d > 0$ and $\delta \in (0,1)$. We have that*

$$\mathbb{P}\left(\sum_{s=1}^t \big(\ell_s(\theta^\star) - \mathcal{L}_s(Q_s)\big) \leq t\phi_d B + d \log \frac{d}{\delta}\,,\ \forall t \geq 1 \right) \geq 1 - \delta\,.$$

Now that we have a concentration result to control S_t, we only need to be able to upper bound the regret of an algorithm for the "delayed" online game that we are considering. To this purpose, we propose the following approach. We run d independent EWA forecaster (with uniform prior), each one only making prediction and receiving the feedback once every d rounds. More explicitly, the first forecaster acts at rounds 1, $d+1$, $2d+1$..., the second at round 2, $d+2$, $2d+2$..., and so on. As a direct consequence of Proposition 2, by summing the individual regret upper bounds we get a regret bound for the joint forecaster, which at each round returns the distribution predicted by the currently active forecaster. This technique of partitioning rounds into blocks for the regret analysis of online learning is common in the literature (*e.g.*, see Weinberger and Ordentlich, 2002).

Lemma 2. *Fix $B > 0$, $d > 0$, and consider a strategy with d independent EWA forecasters outlined above, all initialised with the uniform distribution on $\mathcal{B}(B+1)$ as prior. For all $\bar\theta \in \mathcal{B}(B)$ and $t \geq 1$,*

$$\mathrm{Regret}_t(\bar\theta) \leq \frac{dp}{2} \log \frac{(B+1)^2 e \max(dp, t+d)}{dp}\,.$$

Putting together what we have, we get a confidence sequence suitable for our mixing UCB algorithm.

Theorem 1. *Consider the setting introduced above. Fix $\delta \in (0,1)$ and a delay $d > 0$. Assume as known that $\theta^\star \in \mathcal{B}(B)$. Let $\widehat\theta_t =$ arg min$_{\theta \in \mathcal{B}(B)} \{\sum_{s=1}^t \ell_s(\theta)\}$ and $\Lambda_t = \sum_{s=1}^t X_s X_s^\top$. Define*

$$\mathcal{C}_t = \left\{ \theta \in \mathcal{B}(B) \,:\, \tfrac{1}{2}\|\theta - \widehat\theta_t\|_{\Lambda_t}^2 \leq \tfrac{dp}{2} \log \tfrac{(B+1)^2 e \max(dp, t+d)}{dp} + t\phi_d (B+1) + d \log \tfrac{d}{\delta} \right\}\,.$$

Then, $(\mathcal{C}_t)_{t\geq 1}$ *is an anytime valid confidence sequence for* $\theta^\star$*, namely*

$$\mathbb{P}\big(\theta^\star \in \mathcal{C}_t\,,\ \forall t \geq 1\big) \leq 1 - \delta\,.$$

Proof. The optimality of $\widehat{\theta}_t$ implies $\sum_{s=1}^t \langle \theta - \widehat{\theta}_t, \nabla \ell_s(\widehat{\theta}_t)\rangle \geq 0$, for all $\theta \in \mathcal{B}(B)$. As $\sum_{s=1}^t \ell_s$ is quadratic, it equals its second order Taylor expansion around $\widehat{\theta}_t$ and its Hessian is everywhere Λ_t. So,

$$\frac{1}{2}\|\theta - \widehat{\theta}_t\|^2_{\Lambda_t} \leq \frac{1}{2}\|\theta - \widehat{\theta}_t\|^2_{\Lambda_t} + \sum_{s=1}^t \big\langle \theta - \widehat{\theta}_t, \nabla \ell_s(\widehat{\theta}_t)\big\rangle = \sum_{s=1}^t \big(\ell_s(\theta) - \ell_s(\widehat{\theta}_t)\big)\,,$$

for any $\theta \in \mathcal{B}(B)$. This, together with (3), Lemma 1, and Lemma 2, yields the conclusion. ☐

We remark that the confidence sets of Theorem 1 take the form of the intersection between the ball $\mathcal{B}(B)$ and the "ellipsoid" $\{\theta : \|\theta - \widehat{\theta}_t\|_{\Lambda_t} \leq \beta_t\}$, for a suitable radius β_t. In order to implement and analyse the bandit algorithm, it will be more convenient to work with a relaxation of these sets, a pure ellipsoid not intersected with $\mathcal{B}(B)$. We make this explicit in the following corollary.

Corollary 1. *Fix* $\lambda > 0$*,* $d > 0$*, and* $\delta \in (0, 1)$*. For* $t \geq 1$*, let* $V_t = \Lambda_t + \lambda \mathrm{Id}$*. Assuming that* $\theta^\star \in \mathcal{B}(B)$*, the following compact ellipsoids define an anytime valid confidence sequence for* $\theta^\star$*:*

$$\mathcal{C}_t = \Big\{\theta \in \mathcal{B}(B)\,:\, \tfrac{1}{2}\|\theta - \widehat{\theta}_t\|^2_{V_t} \leq \tfrac{dp}{2}\log\tfrac{(B+1)^2 e \max(dp, t+d)}{dp} + 2\lambda B^2 + t\phi_d(B+1) + d\log\tfrac{d}{\delta}\Big\}\,.$$

Proof. Let $\beta_t^2 = dp\log\frac{(B+1)^2 e \max(dp,t+d)}{dp} + 2t\phi_d(B+1) + 2d\log\frac{d}{\delta}$. From Theorem 1, with probability at least $1-\delta$, uniformly for every t, $\|\theta^\star - \widehat{\theta}_t\|^2_{\Lambda_t} \leq \beta_t^2$. Adding to both sides of this inequality $\frac{\lambda}{2}\|\theta^\star - \widehat{\theta}_t\|_2^2$, and relaxing the RHS using that $\|\theta^\star - \widehat{\theta}_t\|_2^2 \leq 4B^2$, we conclude. ☐

5 Regret bounds for Mixing-LinUCB

In this section, we establish worst-case and gap-dependent cumulative regret bounds for mixing UCB algorithm (Mixing Lin-UCB). However, to account for the fact that Mixing-LinUCB selects actions with delays, the standard elliptical potential arguments must be modified. Throughout this section, we let $R_t = \langle \theta^\star, X_t^\star - X_t\rangle$ (where $X_t^\star = \arg\max_{x\in\mathcal{X}_t}\langle\theta^\star, x\rangle$) denote the regret in round t, and $\beta_t^2 = dp\log\frac{(B+1)^2 e\max(dp,t+d)}{dp} + 4\lambda B^2 + 2t\phi_d(B+1) + 2d\log\frac{d}{\delta}$ denote the squared radius of the ellipsoid $\mathcal{C}_t$ in Corollary 1.

5.1 Worst-case regret bounds

First, following the regret analysis in Abbasi-Yadkori et al. (2011) (see also Section 19.3 in Lattimore and Szepesvári, 2020), we upper bound the instantaneous regret. From our boundedness assumptions ($\theta^\star \in \mathcal{B}(B)$ and $\mathcal{X}_t \subseteq \mathcal{B}(1)$), we easily deduce that $R_t \leq 2B$. Under the event that our confidence sequence contains $\theta^\star$ at every step t, we have another bound on R_t. If we define $\widetilde{\theta}_{t-d} \in \mathcal{C}_{t-d}$ to be the point at which $\langle\widetilde{\theta}_{t-d}, X_t\rangle = \mathrm{UCB}_{\mathcal{C}_{t-d}}(X_t)$, then from the definition of X_t we have

$$\langle\theta^\star, X_t^\star\rangle \leq \max_{x\in\mathcal{X}_t}\max_{\theta\in\mathcal{C}_{t-d}}\langle\theta, x\rangle = \max_{x\in\mathcal{X}_t}\mathrm{UCB}_{\mathcal{C}_{t-d}}(x) = \mathrm{UCB}_{\mathcal{C}_{t-d}}(X_t) = \langle\widetilde{\theta}_{t-d}, X_t\rangle\,.$$

Recall that, for all s, $V_s = \Lambda_s + \lambda\mathrm{Id}$, which is invertible as $\lambda > 0$. Thus, by Cauchy-Schwarz,

$$R_t \leq \langle\widetilde{\theta}_{t-d} - \theta^\star, X_t\rangle \leq \|\widetilde{\theta}_{t-d} - \theta^\star\|_{V_{t-d}}\|X_t\|_{V_{t-d}^{-1}} \leq 2\beta_{t-d}\|X_t\|_{V_{t-d}^{-1}}\,.$$

This means that the instantaneous regret satisfies the bound

$$R_t \leq 2\max(B, \beta_{t-d})\min(1, \|X_t\|_{V_{t-d}^{-1}})\,. \tag{4}$$

Next, we separate the regret suffered in the first d rounds and the remaining $T-d$ rounds. We then use Cauchy-Schwarz once more, and the fact that β_t is increasing in t, to obtain

$$\begin{aligned}\mathrm{Reg}(T) &\leq 2dB + \sqrt{(T-d)\textstyle\sum_{t=d+1}^{T} R_t^2} \\ &\leq 2dB + \sqrt{4(T-d)\max(B^2, \beta_{T-d}^2)\textstyle\sum_{t=d+1}^{T}\min(1, \|X_t\|^2_{V_{t-d}^{-1}})}\,.\end{aligned}$$

At this point, we must depart from the standard linear UCB analysis (Abbasi-Yadkori et al., 2011; Lattimore and Szepesvári, 2020). We bound the sum of the *elliptical potentials* $\sum_{t=d+1}^{T}\min(1, \|X_t\|^2_{V_{t-d}^{-1}})$ using the following variant of the well-known "elliptical potential lemma" (see Appendix), which accounts for the fact that the feature covariance matrix V_{t-d} is updated with a delay of d steps.

Lemma 3. *For all $T \geq d+1$,*

$$\sum_{t=d+1}^{T}\min(1, \|X_t\|^2_{V_{t-d}^{-1}}) \leq 2dp\log(1 + \tfrac{T}{\lambda dp})\,.$$

We can now state a worst-case regret upper bound for Mixing-LinUCB.

Theorem 2. *Fix $\lambda = 1/B^2$, $d > 0$ and $\delta \in (0,1)$. With probability at least $1-\delta$, for all $T > d$, the regret of Mixing-LinUCB satisfies*

$$\mathrm{Reg}(T) \leq 2dB + \sqrt{8dpT\max(B^2, \beta_T^2)\log(1 + \tfrac{B^2 T}{dp})}\,.$$

From the definition of β_T, we see that this regret bound is of the order

$$\mathrm{Reg}(T) = \mathcal{O}\left(dB + dp\sqrt{T}\log\tfrac{TB}{dp} + T\sqrt{Bdp\phi_d\log\tfrac{TB}{dp}} + d\sqrt{pT\log\tfrac{TB}{p\delta}}\right)\,.$$

For any fixed (*i.e.*, not depending on T) delay d, this regret bound is linear in T. To obtain meaningful regret bounds, it is therefore crucial to set d as a function of T and the rate at which the mixing coefficients decay to zero. We point out that if T is unknown, one could probably use a more general framework where the delay is time dependent which might lead to non-trivial results, but we do not pursue this here. Under the assumption that the noise variables are either geometrically or algebraically mixing, we obtain the following worst-case regret bounds.

Corollary 2. *Suppose that the noise satisfies Assumption 1 with $\phi_d = Ce^{-\frac{d}{\tau}}$ for some $C, \tau > 0$ (*geometric mixing*), and set $d = \lceil \tau\log\frac{BCT}{p}\rceil$. Then, the regret of Mixing-LinUCB satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}\left(\tau p\sqrt{T}\left(\log\tfrac{TB\max(1,C)}{p}\right)^2 + p\sqrt{T}\tau\log\tfrac{TB\max(1,C)}{p} + \tau\log\tfrac{BCT}{p}\sqrt{pT\log\tfrac{TB}{p\delta}}\right)\,.$$

Corollary 3. *Suppose that the noise satisfies Assumption 1 with $\phi_d = Cd^{-r}$ for some $C > 0$ and $r > 0$ (*algebraic mixing*), and set $d = \lceil CT^{1/(1+r)}\rceil$. Then, the regret of Mixing-LinUCB satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}\left(CBT^{1/(1+r)} + CT^{\frac{3+r}{2(1+r)}}\left(p\log\tfrac{TB}{p} + \sqrt{Bp\log\tfrac{T^{r/(1+r)}B}{Cp}} + \sqrt{p\log\tfrac{TB}{p\delta}}\right)\right)\,.$$

Up to a factor of $\tau\log T$, the bound for geometrically mixing noise matches the regret bound for linear UCB with i.i.d. noise. This bound is trivial for $r \leq 1$, however for $r > 1$ we get sublinear regret, and in particular we recover standard rates up to logarithmic factors in the limit where $r \to \infty$.

5.2 Gap-dependent regret bounds

Under the assumption that, each round, the gap between the expected reward of the optimal arm and the expected reward of any other arm is at least $\Delta > 0$, we get regret bounds with better dependence on T. More precisely, define the *minimum gap* $\Delta = \min_{t \in [T]} \min_{x \in \mathcal{X}_t : x \neq X_t^\star} \langle X_t^\star - x, \theta^\star \rangle$, and assume that $\Delta > 0$. Since we either have $R_t = 0$ or $R_t \geq \Delta > 0$, it follows that

$$R_t \leq R_t^2 / \Delta \,.$$

In our worst-case analysis, we showed that

$$\sum_{t=d+1}^{T} R_t^2 \leq 8dp \max(B^2, \beta_T^2) \log(1 + \tfrac{T}{\lambda dp}) \,.$$

Combined with the previous inequality, we obtain the following gap-dependent regret bound.

Theorem 3. *Fix $\lambda = 1/B^2$, $d > 0$, and $\delta \in (0, 1)$. With probability at least $1 - \delta$, for all $T > d$, the regret of Mixing-LinUCB satisfies*

$$\mathrm{Reg}(T) \leq 2dB + \frac{8dp}{\Delta} \max(B^2, \beta_T^2) \log\left(1 + \frac{B^2 T}{dp}\right) \,.$$

Similarly to the worst-case bound in Theorem 2, for any fixed $d > 0$, this regret bound is linear in T. By setting d as a suitable function of T, we obtain the following gap-dependent regret bounds under geometrically or algebraically mixing noise.

Corollary 4. *Suppose that the noise variables are geometrically mixing and set $d = \lceil \tau \log \frac{BCT}{p} \rceil$. Then the regret of Mixing-LinUCB satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}\left(\frac{8\tau p}{\Delta} \left(\log \frac{BCT}{p} \right)^2 \log\left(1 + \frac{B^2 T}{p\tau \log \frac{BCT}{p}}\right) \left(\frac{p}{2} \log \frac{T}{p\tau} + \log \frac{\tau \log \frac{BCT}{p}}{\delta} \right) \right) \,.$$

Corollary 5. *Suppose that the noise variables are algebraically mixing and set $d = \lceil CT^{1/(1+r)} \rceil$. Then the regret of Mixing-LinUCB satisfies*

$$\mathrm{Reg}(T) = \mathcal{O}\left(\frac{8Cp}{\Delta} T^{\frac{2}{1+r}} \log\left(1 + \frac{B^2 T}{pCT^{1/(1+r)}}\right) \left(\frac{p}{2} \log \frac{(B+1)^2 eT}{p} + \log \frac{CT^{1/(1+r)}}{\delta} \right) \right) \,.$$

6 Conclusion

We leave several interesting questions open for future research. Some of these are listed below.

An important limitation of our algorithm is that it requires the knowledge of the mixing coefficients (or at least an upper-bound on them). It would be interesting to explore the possibility of relaxing this assumption and to design an algorithm which infers the mixing coefficients while minimizing the regret. We note that the problem of estimating mixing coefficients is already a hard problem on its own right, with tight sample-complexity results only available in special cases such as Markov chains (Hsu et al., 2019; Wolfer, 2020). We also note that in order to recover the standard rate for the regret bound, the delay d introduced in our algorithm need to be chosen as a function of the horizon T. We believe that this could be fixed at little conceptual expense by using time-varying delay in the analysis, but we did not attempt to work out the (potentially non-trivial) details here.

Another limitation is that our analysis assumed throughout that the adversary picking the decision sets $\mathcal{X}_t$ is oblivious, which is typically not required in linear bandit problems. For us, this was necessary to avoid potential statistical dependence between decision sets and the nonstationary observations. We believe that this issue can be handled at least for some classes of adversaries. For instance, it is easy to see that our analysis would carry through under the assumption that the decision sets be selected based on delayed information only. We leave the investigation of this question under more realistic assumptions open for future work.

References

Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3: 397–422, 2003.

Naoki Abe, Alan W. Biermann, and Philip M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Melda Korkut and Andrew Li. Disposable linear bandits for online recommendations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5), 2021.

Maxime C Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

T.L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Hamish Flynn, David Reeb, Melih Kandemir, and Jan R Peters. Improved algorithms for stochastic linear bandits using tail bounds for martingale mixtures. *Advances in Neural Information Processing Systems*, 36:45102–45136, 2023.

Richard C. Bradley. Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. *NeurIPS*, 2008.

Baptiste Abélès, Eugenio Clerico, and Gergely Neu. Generalization bounds for mixing processes via delayed online-to-PAC conversions. In *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, 2025.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 4474–4482, 2024.

Eugenio Clerico, Hamish Flynn, Wojciech Kotłowski, and Gergely Neu. Confidence sequences for generalized linear models via regret analysis, 2025. URL `https://arxiv.org/abs/2504.16555`.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR, 2020a.

Benjamin Howson, Ciara Pike-Burke, and Sarah Filippi. Delayed feedback in generalised linear bandits revisited. In *International Conference on Artificial Intelligence and Statistics*, pages 6095–6119. PMLR, 2023.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

Claire Vernade, Andras Gyorgy, and Timothy Mann. Non-stationary delayed bandits with intermediate observations. In *International Conference on Machine Learning*, pages 9722–9732. PMLR, 2020b.

Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006.

Peter D. Grünwald. *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

M.J. Weinberger and E. Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7), 2002.

Daniel Hsu, Aryeh Kontorovich, David A Levin, Yuval Peres, Csaba Szepesvári, and Geoffrey Wolfer. Mixing time estimation in reversible markov chains from a single sample path. *The Annals of Applied Probability*, 29(4):2439–2480, 2019.

Geoffrey Wolfer. Mixing time estimation in ergodic markov chains from a single trajectory with contraction methods. In *Algorithmic Learning Theory*, pages 890–905, 2020.

A Technical Appendices and Supplementary Material

A.1 Proof of Proposition 2

For the EWA forecaster with prior Q_1, we can rewrite the regret via a standard telescoping argument (see Lemma B.1 in Clerico et al., 2025) as

$$\mathrm{Regret}_t(\bar\theta) = -\log \int \exp\left(-\sum_{s=1}^t \ell_s(\theta) + \sum_{s=1}^t \ell_s(\bar\theta)\right) \mathrm{d}Q_1(\theta)\,.$$

Using the variational representation of the KL divergence, this can be upper bounded as

$$\begin{aligned}\mathrm{Regret}_t(\bar\theta) &= \inf_Q \left\{\int \sum_{s=1}^t \ell_s(\theta)\mathrm{d}Q(\theta) - \sum_{s=1}^t \ell_s(\bar\theta) + D_{\mathrm{KL}}(Q||Q_1)\right\}\\ &\le \inf_{c\in(0,1]} \left\{\int \sum_{s=1}^t \ell_s(\theta)\mathrm{d}P_c(\theta) - \sum_{s=1}^t \ell_s(\bar\theta) + D_{\mathrm{KL}}(P_c||Q_1)\right\}\,,\end{aligned}$$

where P_c is the uniform measure on the closed Euclidean ball of radius c in $\mathbb{R}^p$, centred at $\bar\theta$. We remark that for all $c \in (0, 1]$, $P_c \ll Q_1$. Therefore, for all $c \in (0, 1]$,

$$D_{\mathrm{KL}}(P_c||Q_1) = \int p \log \frac{B+1}{c} \mathrm{d}Q_1(\theta) = p \log \frac{B+1}{c}\,.$$

Taking a second-order Taylor expansion of the total squared loss around $\bar\theta$, and using the fact that the mean of P_c is $\bar\theta$, we obtain

$$\sum_{s=1}^t \int_{\mathbb{R}^p} \big(\ell_s(\theta) - \ell_s(\bar\theta)\big)\mathrm{d}P_c(\theta) = \sum_{s=1}^t \int_{\mathbb{R}^p} \Big(\big\langle \theta-\bar\theta, \nabla\ell_s(\bar\theta)\big\rangle + \tfrac{1}{2}\big\langle \theta-\bar\theta, X_s\big\rangle^2\Big)\mathrm{d}P_c(\theta) \le \frac{tc^2}{2}\,,$$

where we used that $\|X_s\|_2 \le 1$ for all s in the last inequality. Combining everything so far, we obtain

$$\mathrm{Regret}_t(\bar\theta) \le \inf_{c\in(0,1]} \left\{p \log \frac{B+1}{c} + \frac{tc^2}{2}\right\} \le \frac{p}{2} \log \frac{(B+1)^2 e \max(p,t)}{p}\,,$$

where the last term is obtained taking $c = \min(1, \sqrt{p/t})$.

A.2 Proof of Lemma 1

Let $D_t = \ell_t(\theta^\star) - \mathcal{L}_t(Q_t)$ and $\lambda_t(\theta) = \langle \theta - \theta^\star, X_t\rangle$. It is easy to check that

$$D_t = \log \int e^{\lambda_t(\theta)\varepsilon_t - \lambda_t(\theta)^2/2} \mathrm{d}Q_t(\theta)\,.$$

Fix $i \in \{1, \dots, d\}$. We denote as $S^{(i)} = (S_k^{(i)})_{k\ge1}$ the subsequence defined as $S_k^{(i)} = \sum_{j=1}^k D_{i+(j-1)d}$. We also define $\mathcal{F}_k^{(i)} = \mathcal{F}_{i+(k-1)d}$. It is easy to check that $(S_k^{(i)})_{k\ge1}$ is adapted with respect to $(\mathcal{F}_k^{(i)})_{k\ge1}$. Now, let $M_k^{(i)} = \exp(S_k^{(i)} - (k-1)(2B+1)\phi_d)$. We will show that $(M_k^{(i)})_{k\ge1}$ is a super-martingale with respect to $(\mathcal{F}_k^{(i)})_{k\ge1}$, with initial expectation bounded by 1. For this it is enough to show that for any $k \ge 1$ we have $\mathbb{E}[e^{D_{i+(k-1)d}-(2B+1)\phi_d}|\mathcal{F}_{k-1}^{(i)}] \le 1$. This is true for $k = 1$ (where we let $\mathcal{F}_0^{(i)}$ be the trivial σ-field, or more generally a σ-field independent of the noise). Indeed, as $i \le d$, X_i is $\mathcal{F}_0$ measurable and hence independent of ε_i. From Assumption 1, we know that ε_i is sub-Gaussian, and so $\mathbb{E}[e^{D_i}] \le 1$.

Let us now check the case $k \ge 2$. For convenience, we define $t_k^{(i)} = i + (k-1)d$. We note that $\mathcal{F}_{t_k^{(i)}} = \mathcal{F}_k^{(i)}$. We have

$$\begin{aligned}&\mathbb{E}[e^{D_{i+(k-1)d}-(2B+1)\phi_d}|\mathcal{F}_{k-1}^{(i)}]\\ &\qquad= \mathbb{E}\left[\int \exp\big(\lambda_{t_k^{(i)}}(\theta)\varepsilon_{t_k^{(i)}} - \lambda_{t_k^{(i)}}(\theta)^2/2 - (2B+1)\phi_d\big)\mathrm{d}Q_{t_k^{(i)}}(\theta)\middle|\mathcal{F}_{k-1}^{(i)}\right]\,.\end{aligned}$$

Now, $Q_{t_k^{(i)}}$ only depends on the noise up to $\varepsilon_{t_k^{(i)}-d} = \varepsilon_{t_{k-1}^{(i)}}$, thanks to the delayed bandit framework. Henceforth, we can swap the conditional expectation and the integral. In a similar way, we can bring $\exp\big(-\lambda_{t_k^{(i)}}(\theta)^2/2 - (2B+1)\phi_d\big)$ outside of the conditional expectation, as it is $\mathcal{F}_{k-1}^{(i)}$ measurable. We get

$$\begin{aligned}
&\mathbb{E}[e^{D_{i+(k-1)d}-(2B+1)\phi_d}|\mathcal{F}_{k-1}^{(i)}] \\
&\quad= \int \mathbb{E}\left[\exp\big(\lambda_{t_k^{(i)}}(\theta)\varepsilon_{t_k^{(i)}}\big)\Big|\mathcal{F}_{k-1}^{(i)}\right]\exp\big(-\lambda_{t_k^{(i)}}(\theta)^2/2 - (2B+1)\phi_d\big)\mathrm{d}Q_{t_k^{(i)}}(\theta) \\
&\quad\le \int \exp\big(\lambda_{t_k^{(i)}}(\theta)^2/2 + \lambda_{t_k^{(i)}}\mathbb{E}[\varepsilon_{t_k^{(i)}}|\mathcal{F}_{k-1}^{(i)}]\big)\exp\big(-\lambda_{t_k^{(i)}}(\theta)^2/2 - (2B+1)\phi_d\big)\mathrm{d}Q_{t_k^{(i)}}(\theta) \\
&\quad\le \int \exp\big(\big|\lambda_{t_k^{(i)}}(\theta)\big|\phi_d - (2B+1)\phi_d\big)\mathrm{d}Q_{t_k^{(i)}}(\theta)\,,
\end{aligned}$$

where the two inequalities use the sub-Gaussianity and mixing properties of Assumption 1. Now, by construction $Q_{t_k^{(i)}}$ has support on $\mathcal{B}(B+1)$, and for every $\theta \in \mathcal{B}(B+1)$

$$\big|\lambda_{t_k^{(i)}}(\theta)\big| \le \|\theta - \theta^\star\|_2\|X_{t_k^{(i)}}\|_2 \le 2B+1\,,$$

where we also used that $\|X_{t_k^{(i)}}\|_2 \le 1$, as for all t we are assuming that $\mathcal{X}_t \subseteq \mathcal{B}(1)$. We thus conclude that $(M_k^{(i)})_{k\ge1}$ is indeed a super-martingale, non-negative and with initial value bounded by 1. By Ville's inequality it follows that

$$\mathbb{P}\big(S_k^{(i)} \le k(2B+1)\phi_d + \log\tfrac{d}{\delta}\,,\ \forall k\ge1\big) \ge 1 - \tfrac{\delta}{d}\,.$$

Now that we have proven that we have a super-martingale for each block, the desired anytime valid concentration result follows directly from a simple union bound.

A.3 Proof of Lemma 2

Fix $t \ge 1$, and let $i \in \{1,\dots,d\}$ and $k \ge 1$ be such that $t = i + (k-1)d$. Let $I_j = \{j + d\mathbb{N}\} \cap \{1,\dots,t\}$, for $j \in \{1,\dots d\}$. We consider d independent EWA forecaster (all initialised with the uniform prior on $\mathcal{B}(B+1)$). The j^{th} forecaster only acts and receive feedback from the rounds in I_j. We note that the j^{th} forecaster acts for t_j rounds, where $t_j = k$ if $j \ge i$, and $t_j = k-1$ otherwise. We denote as $R^{(j)}$ the regret of the j^{th} forecaster (which only takes into account the losses at the rounds in I_j, with comparator $\bar\theta$. By Proposition 2 we get

$$\mathrm{Regret}_t(\bar\theta) = \sum_{j=1}^d R^{(j)} \le \sum_{j=1}^d \frac{p}{2}\log\frac{(B+1)^2 e\max(p,t_j)}{p}\,.$$

We conclude by noticing that, for all j, $t_j \le (t+d)/d$.

A.4 Proof of Lemma 3

We recall the standard Elliptical Potential Lemma (see e.g. Lemma 11 in Abbasi-Yadkori et al., 2011), which we will use in our proof of Lemma 3.

Lemma 4 (Elliptical Potential Lemma)**.** *Let $(X_t)_t$ be any sequence of vectors in $\mathbb{R}^p$ satisfying $\max_{t\in[T]}\|X_t\|_2 \le L$ and let $V_T = \sum_{t=1}^T X_tX_t^\top + \lambda I$, for some $\lambda > 0$. Then*

$$\sum_{t=1}^T \min(1, \|X_t\|^2_{V_{t-1}^{-1}}) \le 2p\log(1 + \tfrac{TL^2}{\lambda p})\,.$$

Next, we introduce some notation. For $t > d$, define $(i(t),k(t)) \in [d]\times[K]$ such that $t = i(t)+k(t)d$ and let

$$V_{k(t)-1}^{i(t)} = \sum_{k=0}^{k(t)-1} X_k^{i(t)}(X_k^{i(t)})^\top + \lambda\mathrm{Id}\,,$$

where $X_k^{i(t)} = X_{i(t)+kd}$. With this notation, we can state the following lemma.

Lemma 5. *For any $t > d$, we have*

$$V_{t-d} \succcurlyeq V^{i(t)}_{k(t)-1}\,,$$

which implies that $\|X_t\|^2_{V^{-1}_{t-d}} \leq \|X_t\|^2_{\left(V^{i(t)}_{k(t)-1}\right)^{-1}}$ for any $t > d$.

Proof. Notice that we can write $V_{t-d} = \sum_{s=1}^{t-d} X_s X_s^\top + \lambda \mathrm{Id} = V^{i(t)}_{k(t)} + \sum_{s=1, s\notin S_t}^{t-d} X_s X_s^\top$ where $S_t := \{s = i(t) + (k-1)d, k \in [k(t)]\}$ is the set of indices $(i(t), i(t)+d, \ldots, i(t) + (k(t)-1)d)$. The statement now follows from the fact that $\sum_{s=1, s\notin S_t}^{t-d} X_s X_s^\top \succcurlyeq 0$. □

We are now ready to prove Lemma 3. For now, let us assume that $T = Kd$, for some $K > 1$. Using Lemma 5 and then Lemma 4, we have

$$\begin{aligned}
\sum_{t=d+1}^{T} \min(1, \|X_t\|^2_{V^{-1}_{t-d}}) &\leq \sum_{t=d+1}^{T} \min(1, \|X^{i(t)}_{k(t)}\|^2_{(V^{i(t)}_{k(t)-1})^{-1}}) \\
&= \sum_{i=1}^{d} \sum_{k=1}^{K-1} \min(1, \|X^i_k\|^2_{(V^i_{k-1})^{-1}}) \\
&\leq 2dp \log(1 + \tfrac{(K-1)L^2}{\lambda p})\,.
\end{aligned}$$

One can verify that if T is not divisible by d, the above inequality still holds if we replace K by $\lceil \frac{T}{d} \rceil$. Therefore, regardless of whether T is divisible by d, we have

$$\sum_{t=d+1}^{T} \min(1, \|X_t\|^2_{V^{-1}_{t-d}}) \leq 2dp \log(1 + \tfrac{TL^2}{\lambda dp})\,.$$

This concludes the proof of Lemma 3.

A.5 Proof of Corollary 2 and Corollary 3

We start by recalling the general result

$$\mathrm{Reg}(T) = \mathcal{O}\left(\underbrace{dB}_{(1)} + \underbrace{dp\sqrt{T}\log \tfrac{TB}{dp}}_{(2)} + \underbrace{T\sqrt{Bdp\phi_d \log \tfrac{TB}{dp}}}_{(3)} + \underbrace{d\sqrt{pT\log \tfrac{TB}{p\delta}}}_{(4)} \right)\,. \tag{5}$$

To simplify the following calculations, we do not force d to be a positive integer. One can always round d without changing the rates of the regret bounds.

Geometric Mixing:

Assume $d = \tau \log \frac{BCT}{p}$. We notice that the term (1) is logarithmic in T and thus negligible. From the definition of geometric mixing, it holds that $\phi_d = Ce^{-\frac{d}{\tau}} = \frac{p}{BT}$. Therefore,

$$(3) \leq p\sqrt{\tau T} \log \frac{TB}{p}\,.$$

Substituting the value of d yields the desired bounds for terms (2) and (4) in Equation 5, and hence the desired statement.

Algebraic mixing:

Assume $d = CT^{\frac{1}{1+r}}$, we notice that in this case since $\phi_d = Cd^{-r}$, we have $d\phi_d = Cd^{1-r}$. In particular this implies that $T\sqrt{d\phi_d} = T^{\frac{3+r}{2(1+r)}}$ and thus

$$(3) \leq C\sqrt{Bp\log\frac{TB}{p}}T^{\frac{3+r}{2(1+r)}}$$

The same way (2) and (4) are of order $d\sqrt{T} = T^{\frac{3+r}{2(1+r)}}$ and replacing in Equation 5 yields the desired statement.

A.6 Proof of Corollary 4 and Corollary 5

We start by recalling the general result

$$\mathrm{Reg}(T) \leq 2dB + \frac{8dp}{\Delta}\max(B^2, \beta_T^2)\log\left(1 + \frac{B^2 T}{dp}\right),$$

where $\beta_T^2 = \underbrace{dp\log\frac{(B+1)^2 e\max(dp, T+d)}{dp}}_{(1)} + \underbrace{2T\phi_d(B+1)}_{(2)} + \underbrace{2d\log\frac{d}{\delta}}_{(3)}$.

Geometric Mixing:

Assume $d = \tau\log\frac{BCT}{p}$, then $(2) = \frac{2p(B+1)}{BC}$ is a constant. Hence we have

$$\mathrm{Reg}(T) \leq 2dB + \frac{8d^2p}{\Delta}\left(p\log\frac{(B+1)^2 e\max(dp, T+d)}{dp} + 2\log\frac{d}{\delta} + \frac{2p(B+1)}{dBC}\right)\log\left(1 + \frac{B^2T}{dp}\right),$$

which under the assumption that $\beta_T \geq B$ and replacing d by its definition yields

$$\begin{aligned}\mathrm{Reg}(T) &\leq 2B\tau\log\frac{BCT}{p} \\ &+ \frac{8\tau^2 p}{\Delta}\log\left(1 + \frac{B^2T}{p\tau\log\frac{BCT}{p}}\right)\left(\left(\log\frac{BCT}{p}\right)^2\left(p\log\frac{(B+1)^2eT}{p} + 2\tau\frac{\log\frac{BCT}{p}}{\delta}\right) + \frac{2p(B+1)}{BC}\right).\end{aligned}$$

If Δ is constant, then for large T, the first term and the constant part coming from (2) become negligible. Therefore,

$$\mathrm{Reg}(T) = \mathcal{O}\left(\frac{8\tau^2 p}{\Delta}\log\left(1 + \frac{B^2T}{p\tau\log\frac{BCT}{p}}\right)\left(\log\frac{BCT}{p}\right)^2\left(\frac{p}{2}\log\frac{(B+1)^2eT}{p} + \tau\frac{\log\frac{BCT}{p}}{\delta}\right)\right)$$

Algebraic mixing:

Assume $d = CT^{\frac{1}{1+r}}$, then we have

$$\beta_T^2 \leq CT^{\frac{1}{1+r}}\log\frac{(B+1)^2eT}{p} + 2C(B+1)T^{\frac{2}{1+r}} + 2CT^{\frac{1}{1+r}}\log\frac{CT^{\frac{1}{1+r}}}{\delta}.$$

Under the regime where $2dB \leq \frac{8dp}{\Delta}\max(B^2, \beta_T^2)\log\left(1 + \frac{B^2T}{dp}\right)$ and $B \leq \beta_T$ this leads to

$$\mathrm{Reg}(T) = \mathcal{O}\left(\frac{8Cp}{\Delta}T^{\frac{2}{1+r}}\log\left(1 + \frac{B^2T}{pCT^{1/(1+r)}}\right)\left(\frac{p}{2}\log\frac{(B+1)^2eT}{p} + \log\frac{CT^{1/(1+r)}}{\delta}\right)\right).$$