HAL open science

Weighted Linear Bandits for Non-Stationary Environments

Yoan Russac, Claire Vernade, Olivier Cappé

To cite this version:

Yoan Russac, Claire Vernade, Olivier Cappé. Weighted Linear Bandits for Non-Stationary Environments. NeurIPS 2019 - 33rd Conference on Neural Information Processing Systems, Dec 2019, Vancouver, Canada. hal-02291460v1

HAL Id: hal-02291460

https://inria.hal.science/hal-02291460v1

Submitted on 18 Sep 2019 (v1), last revised 19 Mar 2020 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Weighted Linear Bandits for Non-Stationary Environments

Yoan Russac
CNRS, Inria, ENS, Université PSL
yoan.russac@ens.fr

Claire Vernade
Deepmind
vernade@google.com

Olivier Cappé
CNRS, Inria, ENS, Université PSL
olivier.cappe@cnrs.fr

Abstract

We consider a stochastic linear bandit model in which the available actions correspond to arbitrary context vectors whose associated rewards follow a *non-stationary* linear regression model. In this setting, the unknown regression parameter is allowed to vary in time. To address this problem, we propose D-LinUCB, a novel optimistic algorithm based on discounted linear regression, where exponential weights are used to smoothly forget the past. This involves studying the deviations of the sequential weighted least-squares estimator under generic assumptions. As a by-product, we obtain novel deviation results that can be used beyond non-stationary environments. We provide theoretical guarantees on the behavior of D-LinUCB in both slowly-varying and abruptly-changing environments. We obtain an upper bound on the dynamic regret that is of order $d^{2/3}B_T^{1/3}T^{2/3}$, where B_T is a measure of non-stationarity (d and T being, respectively, dimension and horizon). This rate is known to be optimal. We also illustrate the empirical performance of D-LinUCB and compare it with recently proposed alternatives in simulated environments.

1 Introduction

Multi-armed bandits offer a class of models to address sequential learning tasks that involve exploration-exploitation trade-offs. In this work we are interested in structured bandit models, known as stochastic linear bandits, in which linear regression is used to predict rewards [1, 2, 22].

A typical application of bandit algorithms based on the linear model is online recommendation where actions are items to be, for instance, efficiently arranged on personalized web pages to maximize some conversion rate. However, it is unlikely that customers' preferences remain stable and the collected data becomes progressively obsolete as the interest for the items evolve. Hence, it is essential to design adaptive bandit agents rather than restarting the learning from scratch on a regular basis. In this work, we consider the use of weighted least-squares as an efficient method to progressively forget past interactions. Thus, we address sequential learning problems in which the parameter of the linear bandit is evolving with time.

Our first contribution consists in extending existing deviation inequalities to sequential weighted least-squares. Our result applies to a large variety of bandit problems and is of independent interest. In particular, it extends the recent analysis of heteroscedastic environments by [18]. It can also be useful to deal with class imbalance situations, or, as we focus on here, in non-stationary environments.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

As a second major contribution, we apply our results to propose D-LinUCB, an adaptive linear bandit algorithm based on carefully designed exponential weights. D-LinUCB can be implemented fully recursively —without requiring the storage of past actions— with a numerical complexity that is comparable to that of LinUCB. To characterize the performance of the algorithm, we provide a unified regret analysis for abruptly-changing or slowly-varying environments.

The setting and notations are presented below and we state our main deviation result in Section 2. Section 3 is dedicated to non-stationary linear bandits: we describe our algorithms and provide regret upper bounds in abruptly-changing and slowly-varying environments. We complete this theoretical study with a set of experiments in Section 4.

1.1 Model and Notations

The setting we consider in this paper is a non-stationary variant of the stochastic linear bandit problem considered in [1, 22], where, at each round $t \geq 1$, the learner

- receives a finite set of feasible actions $\mathcal{A}_t \subset \mathbb{R}^d$;
- chooses an action $A_t \in \mathcal{A}_t$ and receives a reward X_t such that

$$X_t = \langle A_t, \theta_t^\star \rangle + \eta_t, \tag{1}$$

  where $\theta_t^\star \in \mathbb{R}^d$ is an unknown parameter and η_t is, conditionally on the past, a $\sigma-$subgaussian random noise.

The action set $\mathcal{A}_t$ may be arbitrary but its components are assumed to be bounded, in the sense that $\|a\|_2 \leq L, \forall a \in \mathcal{A}_t$. The time-varying parameter is also assumed to be bounded: $\forall t, \|\theta_t^\star\|_2 \leq S$. We further assume that $|\langle a, \theta_t^\star \rangle| \leq 1, \forall t, \forall a \in \mathcal{A}_t$, (obviously, this could be guaranteed by assuming that $L = S = 1$, but we indicate the dependence in L and S in order to facilitate the interpretation of some results). For a positive definite matrix M and a vector x, we denote by $\|x\|_M$ the norm $\sqrt{x^\top M x}$.

The goal of the learner is to minimize the expected *dynamic regret* defined as

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle a, \theta_t^\star \rangle - X_t\right] = \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \langle a - A_t, \theta_t^\star \rangle . \tag{2}$$

Even in the stationary case —i.e., when $\theta_t^\star = \theta^\star$—, there is, in general, no single fixed best action in this model.

When making stronger structural assumption on $\mathcal{A}_t$, one recovers specific instances that have also been studied in the literature. In particular, the canonical basis of $\mathbb{R}^d$, $\mathcal{A}_t = \{e_1, \ldots, e_d\}$, yields the familiar —non contextual— multi-armed bandit model [20]. Another variant, studied by [15] and others, is obtained when $\mathcal{A}_t = \{e_1 \otimes a_t, \ldots, e_k \otimes a_t\}$, where $\otimes$ denotes the Kronecker product and a_t is a time-varying context vector shared by the k actions.

1.2 Related Work

There is an important literature on online learning in changing environments. For the sake of conciseness, we restrict the discussion to works that consider specifically the stochastic linear bandit model in (1), including its restriction to the simpler (non-stationnary) multi-armed bandit model. Note that there is also a rich line of works that consider possibly non-linear contextual models in the case where one can make probabilistic assumptions on the contexts [10, 23].

Controlling the regret with respect to the non-stationary optimal action defined in (2) depends on the assumptions that are made on the time-variations of $\theta_t^\star$. A generic way of quantifying them is through a *variation bound* $B_T = \sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2$ [4, 6, 11], similar to the penalty used in the group fused Lasso [8]. The main advantage of using the variation budget is that is includes both *slowly-varying* and *abruptly-changing* environments. For the $K-$armed bandits with known B_T, [4–6] achieve the tight dynamic regret bound of $O(K^{1/3} B_T^{1/3} T^{2/3})$. For linear bandits, [11, 12] propose an algorithm based on the use of a sliding-window and provide a $O(d^{2/3} B_T^{1/3} T^{2/3})$ dynamic regret bound; since this contribution is close to ours, we discuss it further in Section 3.2.

A more specific non-stationary setting arises when the number of changes in the parameter is bounded by Γ_T, as in traditional change-point models. The problem is usually referred to as *switching bandits* or *abruptly-changing* environments. It is, for instance, the setting considered in the work by Garivier and Moulines [14], who analyzed the dynamic regret of UCB strategies based on either a sliding-window or exponential discounting. For both policies, they prove upper bounds on the regret in $O(\sqrt{\Gamma_T T})$ when Γ_T is known. They also provide a lower bound in a specific non-stationary setting, showing that $R(T) = \Omega(\sqrt{T})$. The algorithm ideas can be traced back to [19]. [28] shows that an horizon-independent version of the sliding window algorithm can also be analyzed in a slowly-varying setting. [17] analyze windowing and discounting approaches to address dynamic pricing guided by a (time-varying) linear regression model. Discount factors have also been used with Thomson sampling in dynamic environments as in [16, 26].

In abruptly-changing environments, the alternative approach relies on change-point detection [3, 7, 9, 29, 30]. A bound on the regret in $O((\frac{1}{\epsilon^2} + \frac{1}{\Delta})\log(T))$ is proven by [30], where ϵ is the smallest gap that can be detected by the algorithm, which had to be given as prior knowledge. [9] proves a minimax bound in $O(\sqrt{\Gamma_T K T})$ if Γ_T is known. [7] achieves a rate of $O(\sqrt{\Gamma_T K T})$ without any prior knowledge of the gaps or Γ_T. In the contextual case, [29] builds on the same idea: they use a pool of `LinUCB` learners called *slave models* as experts and they add a new model when no existing slave is able to give good prediction, that is, when a change is detected. A limitation however of such an approach is that it can not adapt to some slowly-varying environments, as will be illustrated in Section 4. From a practical viewpoint, the methods based either on sliding window or change-point detection require the storage of past actions whereas those based on discount factors can be implemented fully recursively.

Finally, non-stationarity may also arise in more specific scenarios connected, for instance, to the decaying attention of the users, as investigated in [21, 24, 27]. In the following, we consider the general case where the parameters satisfy the variation bound, i.e., $\sum_{t=1}^{T-1} \|\theta_t^\star - \theta_{t+1}^\star\|_2 \leq B_T$ and we propose an algorithm based on discounted linear regression.

2 Confidence Bounds for Weighted Linear Bandits

In this section, we consider the concentration of the weighted regularized least-squares estimator, when used with general weights and regularization parameters. To the best of our knowledge there is no such results in the literature for sequential learning —i.e., when the current regressor may depend on the random outcomes observed in the past. The particular case considered in Lemma 5 of [18] (heteroscedastic noise with optimal weights) stays very close to the unweighted case and we show below how to extend this result. We believe that this new bound is of interest beyond the specific model considered in this paper. For the sake of clarity, we first focus on the case of regression models with fixed parameter, where $\theta_t^\star = \theta^\star$, for all t.

First consider a deterministic sequence of regularization parameters $(\lambda_t)_{t\geq 1}$. The reason why these should be non-constant for weighted least-squares will appear clearly in Section 3. Next, define by $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ the filtration associated with the random observations. We assume that both the actions A_t and positive weights w_t are predictable, that is, they are $\mathcal{F}_{t-1}$ measurable.

Defining by

$$\hat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\arg\min} \left(\sum_{s=1}^{t} w_s (X_s - \langle A_s, \theta \rangle)^2 + \lambda_t / 2 \|\theta\|_2^2 \right)$$

the regularized weighted least-squares estimator of $\theta^\star$ at time t, one has

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} w_s A_s X_s \quad \text{where} \quad V_t = \sum_{s=1}^{t} w_s A_s A_s^\top + \lambda_t I_d, \tag{3}$$

and I_d denotes the d-dimensional identity matrix. We further consider an arbitrary sequence of positive parameters $(\mu_t)_{t\geq 1}$ and define the matrix

$$\widetilde{V}_t = \sum_{s=1}^{t} w_s^2 A_s A_s^\top + \mu_t I_d. \tag{4}$$

$\widetilde{V}$ is strongly connected to the variance of the estimator $\hat{\theta}_t$, which involves the squares of the weights $(w_s^2)_{s\geq 1}$. For the time being, μ_t is arbitrary and will be set as a function of λ_t in order to optimize the deviation inequality.

We then have the following maximal deviation inequality.

Theorem 1. *For any $\mathcal{F}_t$-predictable sequences of actions $(A_t)_{t\geq 1}$ and positive weights $(w_t)_{t\geq 1}$ and for all $\delta > 0$,*

$$\mathbb{P}\left(\forall t, \|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t} \leq \frac{\lambda_t}{\sqrt{\mu_t}}S + \sigma\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2\sum_{s=1}^t w_s^2}{d\mu_t}\right)}\right) \geq 1-\delta.$$

The proof of this theorem is deferred to the appendix and combines an argument using the method of mixtures and the use of a proper stopping time. The standard result used for least-squares [20, Chapter 20] is recovered by taking $\mu_t = \lambda_t$ and $w_t = 1$ (note that $\widetilde{V}_t$ is then equal to V_t). When the weights are not equal to 1, the appearance of the matrix $\widetilde{V}_t$ is a consequence of the fact that the variance terms are proportional to the squared weights w_t^2, while the least-squares estimator itself is defined with the weights w_t. In the weighted case, the matrix $V_t\widetilde{V}_t^{-1}V_t$ must be used to define the confidence ellipsoid.

An important property of the least-squares estimator is to be scale-invariant, in the sense that multiplying all weights $(w_s)_{1\leq s\leq t}$ and the regularization parameter λ_t by a constant leaves the estimator $\hat{\theta}_t$ unchanged. In Theorem 1, the only choice of sequence $(\mu_t)_{t\geq 1}$ that is compatible with this scale-invariance property is to take μ_t proportional to λ_t^2: then the matrix $V_t\widetilde{V}_t^{-1}V_t$ becomes scale-invariant (*i.e.* unchanged by the transformation $w_s \mapsto \alpha w_s$) and so does the upper bound of $\|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t}$ in Theorem 1. In the following, we will stick to this choice, while particularizing the choice of the weights w_t to allow for non-stationary models.

It is possible to extend this result to heteroscedastic noise, when η_t is σ_t sub-Gaussian and σ_t is $\mathcal{F}_{t-1}$ measurable, by defining $\widetilde{V}_t$ as $\sum_{s=1}^t w_s^2\sigma_s^2 A_s A_s^\top + \mu_t I_d$. In the next section, we will also use an extension of Theorem 1 to the non-stationary model presented in (1) . In this case, Theorem 1 holds with $\theta^\star$ replaced by $V_t^{-1}\big(\sum_{s=1}^t w_s A_s A_s^\top \theta_s^\star + \lambda_t\theta_r^\star\big)$, where r is an arbitrary time index (proposition 3 in Appendix). The fact that r can be chosen freely is a consequence of the assumption that the sequence of L2-norms of the parameters $(\theta_t^\star)_{t\geq 1}$ is bounded by S.

3 Application to Non-stationary Linear Bandits

In this section, we consider the non-stationary model defined in (1) and propose a bandit algorithm in Section 3.1, called Discounted Linear Upper Confidence Bound (`D-LinUCB`), that relies on weighted least-squares to adapt to changes in the parameters $\theta_t^\star$. Analyzing the performance of `D-LinUCB` in Section 3.2, we show that it achieves reliable performance both for abruptly changing or slowly drifting parameters.

3.1 The `D-LinUCB` Algorithm

Being adaptive to parameter changes indeed implies to reduce the influence of observations that are far back in the past, which suggests using weights w_t that increase with time. In doing so, there are two important caveats to consider. First, this can only be effective if the sequence of weights is growing sufficiently fast (see the analysis in the next section). We thus consider exponentially increasing weights of the form $w_t = \gamma^{-t}$, where $0 < \gamma < 1$ is the discount factor.

Next, due to the absence of assumptions on the action sets $\mathcal{A}_t$, the regularization is instrumental in obtaining guarantees of the form given in Theorem 1. In fact, if $w_t = \gamma^{-t}$ while λ_t does not increase sufficiently fast, then the term $\log\big(1 + (L^2\sum_{s=1}^t w_s^2)/(d\mu_t)\big)$ will eventually dominate the radius of the confidence region since we choose μ_t proportional to λ_t^2. This occurs because there is no guarantee that the algorithm will persistently select actions A_t that span the entire space. With this in mind, we consider an increasing regularization factor of the form $\lambda_t = \gamma^{-t}\lambda$, where $\lambda > 0$ is a hyperparameter.

Note that due to the scale-invariance property of the weighted least-square estimator, we can equivalently consider that at time t, we are given *time-dependent* weights $w_{t,s} = \gamma^{t-s}$, for $1 \leq s \leq t$ and that $\hat{\theta}_t$ is defined as

$$\underset{\theta \in \mathbb{R}^d}{\arg\min} \big(\sum_{s=1}^{t} \gamma^{t-s} (X_s - \langle A_s, \theta \rangle)^2 + \lambda/2 \|\theta\|_2^2 \big).$$

For numerical stability reasons, this form is preferable and is used in the statement of Algorithm 1. In the analysis of Section 3.2 however we revert to the standard form of the weights, which is required to apply the concentration result of Section 1. We are now ready to describe `D-LinUCB` in Algorithm 1.

Algorithm 1: `D-LinUCB`

Input: Probability δ, subgaussianity constant σ, dimension d, regularization λ, upper bound for actions L, upper bound for parameters S, discount factor γ.

Initialization: $b = 0_{\mathbb{R}^d}$, $V = \lambda I_d$, $\widetilde{V} = \lambda I_d$, $\hat{\theta} = 0_{\mathbb{R}^d}$

for $t \geq 1$ **do**

  Receive $\mathcal{A}_t$, compute $\beta_{t-1} = \sqrt{\lambda} S + \sigma \sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(1 + \frac{L^2(1-\gamma^{2(t-1)})}{\lambda d (1-\gamma^2)}\right)}$

  for $a \in \mathcal{A}_t$ **do**

    Compute $\mathrm{UCB}(a) = a^\top \hat{\theta} + \beta_{t-1} \sqrt{a^\top V^{-1} \widetilde{V} V^{-1} a}$

  $A_t = \arg\max_a (\mathrm{UCB}(a))$

  Play action A_t and **receive reward** X_t

  Updating phase: $V = \gamma V + A_t A_t^\top + (1-\gamma)\lambda I_d$, $\widetilde{V} = \gamma^2 \widetilde{V} + A_t A_t^\top + (1-\gamma^2)\lambda I_d$

  $b = \gamma b + X_t A_t$, $\hat{\theta} = V^{-1} b$

3.2 Analysis

As discussed previously, we consider weights of the form $w_t = \gamma^{-t}$ (where $0 < \gamma < 1$) in the `D-LinUCB` algorithm. In accordance with the discussion at the end of Section 1, Algorithm 1 uses $\mu_t = \gamma^{-2t}\lambda$ as the parameter to define the confidence ellipsoid around $\hat{\theta}_{t-1}$. The confidence ellipsoid $\mathcal{C}_t$ is defined as $\left\{\theta : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1}\right\}$ where

$$\beta_t = \lambda \sqrt{S} + \sigma \sqrt{2 \log(1/\delta) + d \log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d (1-\gamma^2)}\right)}. \tag{5}$$

Using standard algebraic calculations together with the remark above about scale-invariance it is easily checked that at time t Algorithm 1 selects the action A_t that maximizes $\langle a, \theta \rangle$ for $a \in \mathcal{A}_t$ and $\theta \in \mathcal{C}_t$. The following theorem bounds the regret resulting from Algorithm 1.

Theorem 2. *Assuming that $\sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2 \leq B_T$, the regret of the* `D-LinUCB` *algorithm is bounded for all $\gamma \in (0,1)$ and integer $D \geq 1$, with probability at least $1-\delta$, by*

$$R_T \leq 2LDB_T + \frac{4L^3 S}{\lambda} \frac{\gamma^D}{1-\gamma} T + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{T \log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}. \tag{6}$$

The first two terms of the r.h.s. of (6) are the result of the bias due to the non-stationary environment. The last term is the consequence of the high probability bound established in the previous section and an adaptation of the technique used in [1].

We give the complete proof of this result in appendix. The high-level idea of the proof is to isolate bias and variance terms. However, in contrast with the stationary case, the confidence ellipsoid $\mathcal{C}_t$ does not necessarily contain (with high probability) the actual parameter value $\theta_t^\star$ due to the (unknown) bias arising from the time variations of the parameter. We thus define

$$\bar{\theta}_t = V_{t-1}^{-1} \left(\sum_{s=1}^{t-1} \gamma^{-s} A_s A_s^\top \theta_s^\star + \lambda \gamma^{-(t-1)} \theta_t^\star \right)$$

which is an action-dependent analogue of the parameter value $\theta^\star$ in the stationary setting (although this is a random value). As mentioned in section 2, $\bar{\theta}_t$ does belong to $\mathcal{C}_t$ with probability at least $1-\delta$ (see Proposition 3 in Appendix). The regret may then be split as

$$R_T \leq 2L \sum_{t=1}^{T} \|\theta_t^\star - \bar{\theta}_t\|_2 + \sum_{t=1}^{T} \langle A_t, \theta_t - \bar{\theta}_t \rangle \quad \text{(with probability at least } 1-\delta\text{)},$$

where $(A_t, \theta_t) = \arg\max_{(a \in \mathcal{A}_t, \theta \in \mathcal{C}_t)} \langle a, \theta \rangle$. The rightmost term can be handled by proceeding as in the case of stationary linear bandits, thanks to the deviation inequality obtained in Section 2. The first term in the r.h.s. can be bounded deterministically, from the assumption made on $\sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2$. In doing so, we introduce the analysis parameter D that, roughly speaking, corresponds to the window length equivalent to a particular choice of discount factor γ: the bias resulting from observations that are less than D time steps apart may be bounded in term of D while the remaining ones are bounded globally by the second term of the r.h.s. of (6). This sketch of proof is substantially different from the arguments used by [11] to analyze their sliding window algorithm (called `SW-LinUCB`). We refer to the appendix for a more detailed analysis of these differences. Interestingly, the regret bound of Theorem 2 holds despite the fact that the true parameter $\theta_t^\star$ may not be contained in the confidence ellipsoid $\mathcal{C}_{t-1}$, in contrast to the proof of [14].

It can be checked that, as T tends to infinity, the optimal choice of the analysis parameter D is to take $D = \log(T)/(1-\gamma)$. Further assuming that one may tune γ as a function of the horizon T and the variation upper bound B_T yields the following result.

Corollary 1. *By choosing $\gamma = 1 - (B_T/(dT))^{2/3}$, the regret of the* `D-LinUCB` *algorithm is asymptotically upper bounded with high probability by a term $O(d^{2/3} B_T^{1/3} T^{2/3})$ when $T \to \infty$.*

This result is favorable as it corresponds to the same order as the lower bound established by [4]. More precisely, the case investigated by [4] corresponds to a non-contextual model with a number of changes that grows with the horizon. On the other hand, the guarantee of Corollary 1 requires horizon-dependent tuning of the discount factor γ, which opens interesting research issues (see also [11]).

4 Experiments

This section is devoted to the evaluation of the empirical performance of `D-LinUCB`. We first consider two simulated low-dimensional environments that illustrate the behavior of the algorithms when confronted to either abrupt changes or slow variations of the parameters. The analysis of the previous section, suggests that `D-LinUCB` should behave properly in both situations. We then consider a more realistic scenario in Section 4.2, where the contexts are high-dimensional and extracted from a data set of actual user interactions with a web service.

For benchmarking purposes, we compare `D-LinUCB` to the Dynamic Linear Upper Confidence Bound (`dLinUCB`) algorithm proposed by [29] and with the Sliding Window Linear UCB (`SW-LinUCB`) of [11]. The principle of the `dLinUCB` algorithm is that a master bandit algorithm is in charge of choosing the best `LinUCB` slave bandit for making the recommendation. Each slave model is built to run in each one of the different environments. The choice of the slave model is based on a lower confidence bound for the so-called *badness* of the different models. The badness is defined as the number of times the expected reward was found to be far enough from the actual observed reward on the last τ steps, where τ is a parameter of the algorithm. When a slave is chosen, the action proposed to a user is the result of the `LinUCB` algorithm associated with this slave. When the action is made, all the slave models that were good enough are updated and the models whose badness were too high are deleted from the pool of slaves models. If none of the slaves were found to be sufficiently good, a new slave is added to the pool.

The other algorithm that we use for comparison is `SW-LinUCB`, as presented in [11]. Rather than using exponentially increasing weights, a hard threshold is adopted. Indeed, the actions and rewards included in the l-length sliding window are used to estimate the linear regression coefficients. We expect `D-LinUCB` and `SW-LinUCB` to behave similarly as they both may be shown to have the same sort of regret guarantees (see appendix).

In the case of abrupt changes, we also compare these algorithms to the Oracle Restart LinUCB (`LinUCB-OR`) strategy that would know the change-points and simply restart, after each change, a

new instance of the LinUCB algorithm. The regret of this strategy may be seen as an empirical lower bound on the optimal behavior of an online learning algorithm in abruptly changing environments.

In the following figures, the vertical red dashed lines correspond to the change-points (in abrupt changes scenarios). They are represented to ease the understanding but except for LinUCB-OR, they are of course unknown to the learning algorithms. When applicable, the blue dashed lines correspond to the average detection time of the breakpoints with the dLinUCB algorithm. For D-LinUCB the discount parameter is chosen as $\gamma = 1 - (\frac{B_T}{dT})^{2/3}$. For SW-LinUCB the window's length is set to $l = (\frac{dT}{B_T})^{2/3}$, where $d = 2$ in the experiment. Those values are theoretically supposed to minimize the asymptotic regret. For the Dynamic Linear UCB algorithm, the badness is estimated from $\tau = 200$ steps, as in the experimental section of [29].

4.1 Synthetic data in abruptly-changing or slowly-varying scenarios

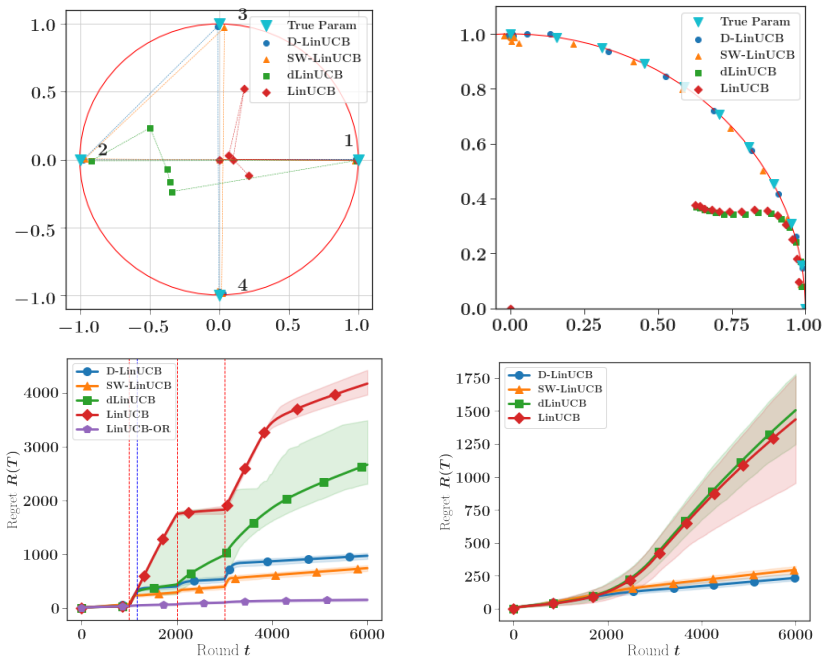

Figure 1: Performances of the algorithms in the abruptly-changing environment (on the left), and, the slowly-varying environment (on the right). The upper plots correspond to the estimated parameter and the lower ones to the accumulated regret, both are averaged on $N = 100$ independent experiments

In this first experiment, we observe the empirical performance of all algorithms in an abruptly changing environment of dimension 2 with 3 breakpoints. The number of rounds is set to $T = 6000$. The light blue triangles correspond to the different positions of the true unknown parameter $\theta_t^\star$: before $t = 1000$, $\theta_t^\star = (1, 0)$; for $t \in [\![1000, 2000]\!]$, $\theta_t^\star = (-1, 0)$; for $t \in [\![2000, 3000]\!]$, $\theta_t^\star = (0, 1)$; and, finally, for $t > 3000$, $\theta_t^\star = (0, -1)$. This corresponds to a hard problem as the sequence of parameters is widely spread in the unit ball. Indeed it forces the algorithm to adapt to big changes, which typically requires a longer adaptation phase. On the other hand, it makes the detection of changes easier, which is an advantage for dLinUCB. In the second half of the experiment (when $t \geq 3000$) there is no change, LinUCB struggles to catch up and suffers linear regret for long periods after the last change-point. The results of our simulations are shown in the left column of Figure 1. On the top row we show a 2-dimensional scatter plot of the estimate of the unknown parameters $\hat{\theta}_t$ every 1000 steps averaged on 100 independent experiment. The bottom row corresponds to the regret averaged over 100 independent experiments with the upper and the lower 5% quantiles. In this environment, with 1-subgaussian random noise, dLinUCB struggles to detect the change-points. Over the 100 experiments, the first change-point was detected in 95% of the runs, the second was never detected and the third only in 6% of the runs, thus limiting the effectiveness of the dLinUCB approach. When decreasing the variance of the noise, the performance of dLinUCB improves and gets closer to

the performance of the oracle restart strategy `LinUCB-OR`. It is worth noting that for both `SW-LinUCB` and `D-LinUCB`, the estimator $\hat{\theta}_t$ adapts itself to non-stationarity and is able to follow $\theta_t^\star$ (with some delay), as shown on the scatter plot. Predictably, `LinUCB-OR` achieves the best performance by restarting exactly whenever a change-point happens.

The second experiment corresponds to a slowly-changing environment. It is easier for `LinUCB` to keep up with the adaptive policies in this scenario. Here, the parameter $\theta_t^\star$ starts at $(1$ and moves continuously counter-clockwise on the unit-circle up to the position $[0, 1]$ in 3000 steps. We then have a steady period of 3000 steps. For this sequence of parameters, $B_T = \sum_{t=1}^{T-1} \|\theta_t^\star - \theta_{t+1}^\star\|_2 = 1.57$. The results are reported in the right column of Figure 1. Unsurprisingly, `dLinUCB` does not detect any change and thus displays the same performance as `LinUCB`. `SW-LinUCB` and `D-LinUCB` behaves similarly and are both robust to such an evolution in the regression parameters. The performance of `LinUCB-OR` is not reported here, as restarting becomes ineffective when the changes are too frequent (here, during the first 3000 time steps, there is a change at every single step). The scatter plot also gives interesting information: $\hat{\theta}_t$ tracks $\theta_t^\star$ quite effectively for both `SW-LinUCB` and `D-LinUCB` but the two others algorithms lag behind. `LinUCB` will eventually catch up if the length of the stationary period becomes larger.

4.2 Simulation based on a real dataset

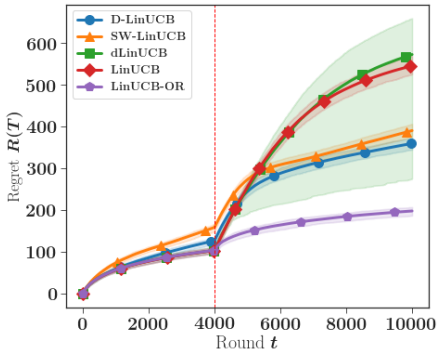

Figure 2: Behavior of the different algorithms on large-dimensional data

`D-LinUCB` also performs well in high-dimensional space ($d = 50$). For this experiment, a dataset providing a sample of 30 days of Criteo live traffic data [13] was used. It contains banners that were displayed to different users and contextual variables, including the information of whether the banner was clicked or not. We kept the categorical variables $cat1$ to $cat9$, together with the variable $campaign$, which is a unique identifier of each campaign. Beforehand, these contexts have been one-hot encoded and 50 of the resulting features have been selected using a Singular Value Decomposition. $\theta^\star$ is obtained by linear regression. The rewards are then simulated using the regression model with an additional Gaussian noise of variance $\sigma^2 = 0.15$. At each time step, the different algorithms have the choice between two 50-dimensional contexts drawn at random from two separate pools of 10000 contexts corresponding, respectively, to clicked or not clicked banners. The non-stationarity is created by switching 60% of $\theta^\star$ coordinates to $-\theta^\star$ at time 4000, corresponding to a partial class inversion. The cumulative dynamic regret is then averaged over 100 independent replications. The results are shown on Figure 2. In the first stationary period, `LinUCB` and `dLinUCB` perform better than the adaptive policies by using all available data, whereas the adaptive policies only use the most recent events. After the breakpoint, `LinUCB` suffers a large regret, as the algorithm fails to adapt to the new environment. In this experiment, `dLinUCB` does not detect the change-point systematically and performs similarly as `LinUCB` on average, it can still outperform adaptive policies from time to time when the breakpoint is detected as can be seen with the 5% quantile. `D-LinUCB` and `SW-LinUCB` adapt more quickly to the change-point and perform significantly better than the non-adaptive policies after the breakpoint. Of course, the oracle policy `LinUCB-OR` is the best performing policy. The take-away message is that there is no free lunch: in a stationary period by using only the most recent events `SW-LinUCB` and `D-LinUCB` do not perform as good as a policy that uses all the available information. Nevertheless, after a breakpoint, the recovery is much faster with the adaptive policies.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[3] P. Auer, P. Gajane, and R. Ortner. Adaptively tracking the best arm with an unknown number of distribution changes. In *European Workshop on Reinforcement Learning 14*, 2018.

[4] O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

[5] O. Besbes, Y. Gur, and A. Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

[6] O. Besbes, Y. Gur, and A. Zeevi. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *Available at SSRN 2436629*, 2018.

[7] L. Besson and E. Kaufmann. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *arXiv preprint arXiv:1902.01575*, 2019.

[8] K. Bleakley and J.-P. Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*, 2011.

[9] Y. Cao, W. Zheng, B. Kveton, and Y. Xie. Nearly optimal adaptive procedure for piecewise-stationary bandit: a change-point detection approach. *arXiv preprint arXiv:1802.03692*, 2018.

[10] Y. Chen, C.-W. Lee, H. Luo, and C.-Y. Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. *arXiv preprint arXiv:1902.00980*, 2019.

[11] W. C. Cheung, D. Simchi-Levi, and R. Zhu. Learning to optimize under non-stationarity. *arXiv preprint arXiv:1810.03024*, 2018.

[12] W. C. Cheung, D. Simchi-Levi, and R. Zhu. Hedging the drift: Learning to optimize under non-stationarity. *arXiv preprint arXiv:1903.01461*, 2019.

[13] Diemert Eustache, Meynet Julien, P. Galland, and D. Lefortier. Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, Halifax, NS, Canada, August, 14, 2017*. ACM, 2017.

[14] A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

[15] A. Goldenshluger and A. Zeevi. A linear response bandit problem. *Stoch. Syst.*, 3(1):230–261, 2013.

[16] N. Gupta, O.-C. Granmo, and A. Agrawala. Thompson sampling for dynamic multi-armed bandits. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 1. IEEE, 2011.

[17] N. B. Keskin and A. Zeevi. Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, 42(2):277–307, 2017.

[18] J. Kirschner and A. Krause. Information directed sampling and bandits with heteroscedastic noise. *arXiv preprint arXiv:1801.09667*, 2018.

[19] L. Kocsis and C. Szepesvári. Discounted ucb. *In: 2nd Pascal Challenge Workshop*, 2006.

[20] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019.

[21] N. Levine, K. Crammer, and S. Mannor. Rotting bandits. In *Advances in Neural Information Processing Systems*, pages 3074–3083, 2017.

[22] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, 2010.

[23] H. Luo, C.-Y. Wei, A. Agarwal, and J. Langford. Efficient contextual bandits in non-stationary worlds. *arXiv preprint arXiv:1708.01799*, 2017.

[24] Y. Mintz, A. Aswani, P. Kaminsky, E. Flowers, and Y. Fukuoka. Non-stationary bandits with habituation and recovery dynamics. *arXiv preprint arXiv:1707.08423*, 2017.

[25] V. H. Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

[26] V. Raj and S. Kalyani. Taming non-stationary bandits: A bayesian approach. *arXiv preprint arXiv:1707.09727*, 2017.

[27] J. Seznec, A. Locatelli, A. Carpentier, A. Lazaric, and M. Valko. Rotting bandits are no harder than stochastic ones. *arXiv preprint arXiv:1811.11043*, 2018.

[28] L. Wei and V. Srivatsva. On abruptly-changing and slowly-varying multiarmed bandit problems. In *2018 Annual American Control Conference (ACC)*, pages 6291–6296. IEEE, 2018.

[29] Q. Wu, N. Iyer, and H. Wang. Learning contextual bandits in a non-stationary environment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 495–504, New York, NY, USA, 2018. ACM.

[30] J. Y. Yu and S. Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1177–1184. ACM, 2009.

Appendix

A Confidence Bounds for Weighted Linear Bandits

A.1 Preliminary results

In this section we give the main results for obtaining Theorem 1. For the sake of conciseness all the results will be stated with σ-subgaussian noises but the proofs will be done with the particular value of $\sigma = 1$. The model we consider is the one defined by equation (1), where we recall that $(\eta_s)_s$ is, conditionally on the past, a sequence of σ-subgaussian random noises. The results of this section are close to the one proposed in [1] but our results are valid with a sequence of predictable weights.

We introduce the quantity $S_t = \sum_{s=1}^t w_s A_s \eta_s$ and $\widetilde{V}_t = \sum_{s=1}^t w_s^2 A_s A_s^\top + \mu_t I_d$. When the regularization term is omitted, let $\widetilde{V}_t(0) = \sum_{s=1}^t w_s^2 A_s A_s^\top$. The filtration associated with the random observations is denoted $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ such that A_t is $\mathcal{F}_{t-1}$-measurable and η_t is $\mathcal{F}_t$-measurable. The weights are also assumed to be predictable. The following lemma is an extension to the weighted case of Lemma 8 of [1].

Lemma 1. *Let $(w_t)_{t\geq 1}$ be a sequence of predictable and positive weights. Let $x \in \mathbb{R}^d$ be arbitrary and consider for any $t \geq 1$*

$$M_t(x) = \exp\left(\frac{1}{\sigma} x^\top S_t - \frac{1}{2} x^\top \widetilde{V}_t(0) x\right).$$

Let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Then $M_\tau(x)$ is almost surely well-defined and

$$\forall x \in \mathbb{R}^d, \mathbb{E}[M_\tau(x)] \leq 1.$$

Proof. First, we prove that $\forall x \in \mathbb{R}^d, (M_t(x))_{t=0}^\infty$ is a super-martingale.

Let $x \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}[M_t(x)|\mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(x^\top S_{t-1} + x^\top w_t A_t \eta_t - 1/2 x^\top (\widetilde{V}_{t-1}(0) + w_t^2 A_t A_t^\top) x\right) |\mathcal{F}_{t-1}\right] \\
&= M_{t-1}(x) \mathbb{E}\left[\exp(x^\top w_t A_t \eta_t - \frac{1}{2} w_t^2 x^\top A_t A_t^\top x) |\mathcal{F}_{t-1}\right] \\
&= M_{t-1}(x) \exp(-\frac{1}{2} w_t^2 x^\top A_t A_t^\top x) \mathbb{E}\left[\exp(x^\top w_t A_t \eta_t) |\mathcal{F}_{t-1}\right] \\
&\leq M_{t-1}(x) \exp(-\frac{1}{2} w_t^2 x^\top A_t A_t^\top x) \exp(1/2 w_t^2 (x^\top A_t)^2) \\
&= M_{t-1}(x).
\end{aligned}$$

The second equality comes from the fact that S_{t-1} and $\widetilde{V}_{t-1}$ are $\mathcal{F}_{t-1}$-measurable. The inequality is the definition of the conditional 1-subgaussianity where we also use the $\mathcal{F}_{t-1}$-measurability of w_t.

Using this supermartingale property, we have $\mathbb{E}[M_t(x)] \leq 1$. The convergence theorem for non-negative supermartingales ensures that $M_\infty(x) = \lim_{t\to\infty} M_t(x)$ is almost surely well defined. By introducing the stopped supermartingale $\mathcal{M}_t(x) = M_{\min(t,\tau)}(x)$, we have $M_\tau(x) = \lim_{t\to\infty} \mathcal{M}_t(x)$. Knowing that $\mathcal{M}_t(x)$ is also a supermartingale, we have

$$\mathbb{E}[\mathcal{M}_t(x)] = \mathbb{E}[M_{\min(t,\tau)}(x)] \leq \mathbb{E}[M_{\min(0,\tau)}(x)] = \mathbb{E}[M_0(x)] = 1.$$

By using Fatou's lemma:

$$\mathbb{E}[M_\tau(x)] = \mathbb{E}[\liminf_{t\to\infty} \mathcal{M}_t(x)] \leq \liminf_{t\to\infty} \mathbb{E}[\mathcal{M}_t(x)] \leq 1.$$

□

In the next lemma, we will integrate $M_t(x)$ with respect to a time-dependent probability measure. This is the key for allowing sequential regularizations in the concentration inequality stated in Theorem 1. This lemma is inspired by the method of mixtures first presented in [25]. The idea of

using time-varying probability measures is inspired from the proof of Theorem 11 in [18]. The two following lemmas are included in the appendix so that the article is self-contained. There are not a mere consequence of the results in [1] because of the time-dependent regularization parameters. As explained in Section 3, this is unavoidable when using exponential weights to avoid the vanishing effect of the regularization.

Lemma 2. *Let $(h_t)_t$ be a sequence of probability measures on $\mathbb{R}^d$. We define $\widetilde{M_t} = \int_{\mathbb{R}^d} M_t(x) dh_t(x)$. Then,*

$$\forall t, \mathbb{E}[\widetilde{M_t}] \leq 1$$

Proof.

$$\begin{aligned}
\mathbb{E}[\widetilde{M_t}] = \int \widetilde{M_t}\, d\mathbb{P} &= \int \left(\int_{\mathbb{R}^d} M_t(x) dh_t(x) \right) d\mathbb{P} \\
&= \int_{\mathbb{R}^d} \left(\int M_t(x) d\mathbb{P} \right) dh_t(x) \quad \text{(Fubini's theorem)} \\
&= \int_{\mathbb{R}^d} \mathbb{E}[M_t(x)] dh_t(x) \\
&\leq \int_{\mathbb{R}^d} dh_t(x) \quad \text{(Lemma 1)} \\
&\leq 1. \quad (h_t \text{ probability measure.})
\end{aligned}$$

□

Lemma 2 is a warm-up for the next lemma and is helpful for understanding why Lemma 3 holds. It is valid for any fixed time t. The next step is to give its equivalent in a stopped version in the specific case of gaussian random vectors.

Lemma 3. *Let $(\mu_t)_t$ be a deterministic sequence of regularization parameters. Let $\mathcal{F}_\infty = \sigma\left(\cup_{t=1}^{\infty} \mathcal{F}_t\right)$ be the tail σ-algebra of the filtration $(\mathcal{F}_t)_t$. Let $X = (X_t)_{t\geq 1}$ be an independent sequence of gaussian random vectors such that $X_t \sim \mathcal{N}(0, \frac{1}{\mu_t} I_d) = h_t$ with X independent of $\mathcal{F}_\infty$. We define*

$$\bar{M}_t(\mu_t) = \mathbb{E}[M_t(X_t)|\mathcal{F}_\infty] = \int_{\mathbb{R}^d} M_t(x) f_{\mu_t}(x) dx,$$

where f_{μ_t} is the probability density function associated with h_t defined as,

$$f_{\mu_t}(x) = \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \exp(-\frac{\mu_t x^\top x}{2}).$$

Let τ be a stopping time with respect to the filtration $(\mathcal{F}_t)_t$ then,

$$\mathbb{E}[\bar{M}_\tau(\mu_\tau)] \leq 1.$$

Proof. We can use the result of Lemma 1 which gives $\forall x \in \mathbb{R}^d$, $\mathbb{E}[M_\tau(x)] \leq 1$.

We have,

$$\begin{aligned}
\mathbb{E}[\bar{M}_\tau(\mu_\tau)] &= \mathbb{E}[\mathbb{E}[M_\tau(X_\tau)|\mathcal{F}_\infty]] = \mathbb{E}[\mathbb{E}[\mathbb{E}[M_\tau(X_\tau)|\mathcal{F}_\infty]|(X_t)_{t\geq 1}]] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[M_\tau(X_\tau)|(X_t)_{t\geq 1}]|\mathcal{F}_\infty]] \leq 1.
\end{aligned}$$

The inequality is a consequence of Lemma 1 as, conditionally to the sequence $(X_t)_t$, $M_\tau(X_\tau)$ is of the form $M_\tau(x)$ with a fixed x. □

We finally state the main result needed to obtain Theorem 1.

Proposition 1. *For $(w_s)_{s\geq 1}$ a sequence of predictable and positive weights, $\forall \delta > 0$, the following deviation inequality holds*

$$\mathbb{P}\left(\exists t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sigma \sqrt{2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\mu_t^d}\right)} \right) \leq \delta.$$

Proof. For a fixed t,

$$\begin{aligned}
\bar{M}_t(\mu_t) &= \int_{\mathbb{R}^d} M_t(x) f_{\mu_t}(x) dx \\
&= \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(x^\top S_t - \frac{1}{2}\|x\|^2_{\mu_t I_d} - \frac{1}{2}\|x\|^2_{\widetilde{V}_t(0)}\right) dx \\
&= \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(x^\top S_t - \frac{1}{2}\|x\|^2_{\widetilde{V}_t}\right) dx \\
&= \frac{1}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}} - \frac{1}{2}\|x - \widetilde{V}_t^{-1} S_t\|^2_{\widetilde{V}_t}\right) dx \\
&= \frac{\exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}}\right)}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|x - \widetilde{V}_t^{-1} S_t\|^2_{\widetilde{V}_t}\right) dx \\
&= \frac{\exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}}\right)}{\sqrt{(2\pi)^d \det(1/\mu_t I_d)}} \sqrt{(2\pi)^d \det\left(\widetilde{V}_t^{-1}\right)} \\
&= \exp\left(\frac{1}{2}\|S_t\|^2_{\widetilde{V}_t^{-1}}\right) \sqrt{\frac{\det(\mu_t I_d)}{\det(\widetilde{V}_t)}}.
\end{aligned}$$

We introduce the particular stopping time,

$$\tau = \min\left\{t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\det(\mu_t I_d)}\right)}\right\}.$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\exists t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\det(\mu_t I_d)}\right)}\right) = \mathbb{P}(\tau < \infty) \\
&= \mathbb{P}\left(\tau < \infty, \|S_\tau\|_{\widetilde{V}_\tau^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_\tau)}{\det(\mu_\tau I_d)}\right)}\right) \\
&\leq \mathbb{P}\left(\|S_\tau\|_{\widetilde{V}_\tau^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_\tau)}{\det(\mu_\tau I_d)}\right)}\right) \\
&= \mathbb{P}\left(\exp\left(\frac{1}{2}\|S_\tau\|^2_{\widetilde{V}_\tau^{-1}}\right) \sqrt{\frac{\det(\mu_\tau I_d)}{\det(\widetilde{V}_\tau)}} \geq \frac{1}{\delta}\right) \\
&\leq \delta \mathbb{E}[\bar{M}_\tau(\mu_\tau)] \text{ (Markov's inequality)} \leq \delta \text{ (Lemma 3).}
\end{aligned}$$

□

A.2 Proof of Theorem 1

We recall that Theorem 1 is established in a stationary environment where $\forall t \geq 1,\ \theta_t^\star = \theta^\star$.

Proof. First note that,

$$\begin{aligned}
\hat{\theta}_t &= V_t^{-1} \sum_{s=1}^{t} w_s A_s X_s \\
&= V_t^{-1} \sum_{s=1}^{t} w_s A_s (A_s^\top \theta^\star + \eta_s) \quad \text{(Equation 1)}
\end{aligned}$$

$$= V_t^{-1}\left(\sum_{s=1}^{t} w_s A_s A_s^\top \theta^\star + \lambda_t \theta^\star - \lambda_t \theta^\star\right) + V_t^{-1} S_t = \theta^\star - \lambda_t V_t^{-1}\theta^\star + V_t^{-1} S_t.$$

Thus,

$$\hat{\theta}_t - \theta^\star = V_t^{-1} S_t - \lambda_t V_t^{-1}\theta^\star. \tag{7}$$

$\forall x \in \mathbb{R}^d, \forall t > 0$, we have

$$\begin{aligned}|x^\top(\hat{\theta}_t - \theta^\star)| &\leq \|x\|_{V_t^{-1}\widetilde{V}_t V_t^{-1}}\left(\|V_t^{-1}S_t\|_{V_t\widetilde{V}_t^{-1}V_t} + \|\lambda_t V_t^{-1}\theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t}\right)\\ &\leq \|x\|_{V_t^{-1}\widetilde{V}_t V_t^{-1}}\left(\|S_t\|_{\widetilde{V}_t^{-1}} + \lambda_t\|\theta^\star\|_{\widetilde{V}_t^{-1}}\right).\end{aligned}$$

By applying the previous inequality with $x = V_t\widetilde{V}_t^{-1}V_t(\hat{\theta}_t - \theta^\star)$, we have

$$\forall t, \|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t} \leq \|S_t\|_{\widetilde{V}_t^{-1}} + \lambda_t\|\theta^\star\|_{\widetilde{V}_t^{-1}}.$$

Knowing that $\widetilde{V}_t \geq \mu_t I_d$ and that $\widetilde{V}_t$ is positive definite, we have $\|\theta^\star\|_{\widetilde{V}_t^{-1}} \leq \frac{1}{\sqrt{\mu_t}}\|\theta^\star\|_2$.

Finally,

$$\forall t, \|\hat{\theta}_t - \theta^\star\|_{V_t\widetilde{V}_t^{-1}V_t} \leq \|S_t\|_{\widetilde{V}_t^{-1}} + \frac{\lambda_t}{\sqrt{\mu_t}}\|\theta^\star\|_2. \tag{8}$$

From Proposition 1, we obtain the following any time high probability lower bound for $\|S_t\|_{\widetilde{V}_t^{-1}}$,

$$\mathbb{P}\left(\forall t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \leq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\mu_t^d}\right)}\right) \geq 1-\delta.$$

Therefore by using inequality 8,

$$\mathbb{P}\left(\forall t \geq 0, \|\hat{\theta}_t - \theta^\star\|_{\widetilde{V}_t^{-1}} \leq \frac{\lambda_t}{\sqrt{\mu_t}}S + \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\widetilde{V}_t)}{\mu_t^d}\right)}\right) \geq 1-\delta.$$

We obtain the exact formula of Theorem 1 by upper bounding $\det(\widetilde{V}_t)$ as proposed in Proposition 2 □

B D-LinUCB Analysis

In this section, the environment is non-stationary, which means that the unknown parameter $\theta^\star$ may evolve over time and is denoted $\theta_t^\star$. The reward generation process in the one presented in Equation (1).

B.1 Preliminary results

In this section, V_t and $\widetilde{V}_t$ are defined by

$$V_t = \sum_{s=1}^{t}\gamma^{-s}A_sA_s^\top + \lambda\gamma^{-t}I_d, \quad \widetilde{V}_t = \sum_{s=1}^{t}\gamma^{-2s}A_sA_s^\top + \lambda\gamma^{-2t}I_d.$$

We recall the definition of β_t:

$$\beta_t = \lambda\sqrt{S} + \sigma\sqrt{2\log(1/\delta) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}.$$

With $\hat{\theta}_t$ defined in equation (3), the confidence ellipsoid we consider is defined by

$$\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1}\right\}. \tag{9}$$

Theorem 1 can be applied with this choice of weights and regularization. We combine it with an upper bound for $\det(\widetilde{V}_t)$ given below.

Proposition 2 (Determinant inequality for the weighted design matrix)**.** *Let $(\lambda_t)_t$ be a deterministic sequence of regularization parameters. Let $V_t = \sum_{s=1}^t w_s A_s A_s^\top + \lambda_t I_d$ be the weighted design matrix. Under the assumption $\forall t, \|A_t\|_2 \leq L$, the following holds*

$$\det(V_t) \leq \left(\lambda_t + \frac{L^2 \sum_{s=1}^t w_s}{d}\right)^d.$$

Proof.

$$\begin{aligned}
\det(V_t) &= \prod_{i=1}^d l_i \quad (l_i \text{ are the eigenvalues}) \\
&\leq \left(\frac{1}{d}\sum_{i=1}^d l_i\right)^d \quad (\text{AM-GM inequality}) \\
&\leq \left(\frac{1}{d}\text{trace}(V_t)\right)^d \leq \left(\frac{1}{d}\sum_{s=1}^t w_s \text{trace}(A_s A_s^\top) + \lambda_t\right)^d \\
&\leq \left(\frac{1}{d}\sum_{s=1}^t w_s \|A_s\|_2^2 + \lambda_t\right)^d \leq \left(\lambda_t + \frac{L^2}{d}\sum_{s=1}^t w_s\right)^d.
\end{aligned}$$

□

Corollary 2. *In the specific case where the weights are given by $w_t = \gamma^{-t}$ with $0 < \gamma < 1$. Proposition 2 can be rewritten*

$$\det(V_t) \leq \left(\lambda_t + \frac{L^2(\gamma^{-t}-1)}{d(1-\gamma)}\right)^d = \left(\lambda\gamma^{-t} + \frac{L^2(\gamma^{-t}-1)}{d(1-\gamma)}\right)^d.$$

We also have,

$$\det(\widetilde{V}_t) \leq \left(\mu_t + \frac{L^2(\gamma^{-2t}-1)}{d(1-\gamma^2)}\right)^d = \left(\lambda\gamma^{-2t} + \frac{L^2(\gamma^{-2t}-1)}{d(1-\gamma^2)}\right)^d.$$

Proof. Apply Proposition 2 and use $\sum_{s=1}^t \gamma^{-s} = \frac{\gamma^{-t}-1}{1-\gamma}$ and $\sum_{s=1}^t \gamma^{-2s} = \frac{\gamma^{-2t}-1}{1-\gamma^2}$. □

Corollary 2 and Proposition 1 yield the following result.

Corollary 3. *$\forall \delta > 0$, with the weights $w_t = \gamma^{-t}$ and $0 < \gamma < 1$, we have*

$$\mathbb{P}\left(\exists t \geq 0, \|S_t\|_{\widetilde{V}_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}\right) \leq \delta.$$

Thanks to this corollary we are now ready to show that $\bar{\theta}_t$ belongs to $\mathcal{C}_{t-1}$ with high probability.

Proposition 3. *Let $\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \beta_{t-1}\right\}$ denote the confidence ellipsoid. Let $\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_s A_s^\top \theta_s^\star + \lambda\gamma^{-(t-1)}\theta_t^\star\right)$. Then, $\forall \delta > 0$,*

$$\mathbb{P}\left(\forall t \geq 1, \bar{\theta}_t \in \mathcal{C}_t\right) \geq 1 - \delta.$$

Proof.

$$\bar{\theta}_t - \hat{\theta}_{t-1} = V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_s A_s^\top \theta_s^\star + \lambda\gamma^{-(t-1)}\theta_t^\star - \sum_{s=1}^{t-1}\gamma^{-s}A_s X_s\right)$$

$$= V_{t-1}^{-1}\left(\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star + \lambda\gamma^{-(t-1)}\theta_t^\star - \sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top\theta_s^\star - \sum_{s=1}^{t-1}\gamma^{-s}A_s\eta_s\right)$$
$$= -V_{t-1}^{-1}S_{t-1} + \lambda\gamma^{-(t-1)}V_{t-1}^{-1}\theta_t^\star.$$

Therefore,

$$\begin{aligned}\|\bar{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} &\leq \|S_{t-1}\|_{\widetilde{V}_{t-1}^{-1}} + \lambda\gamma^{-(t-1)}\|\theta_t^\star\|_{\widetilde{V}_{t-1}^{-1}} \\ &\leq \|S_{t-1}\|_{\widetilde{V}_{t-1}^{-1}} + \sqrt{\lambda}S \quad (\widetilde{V}_{t-1}^{-1} \leq 1/(\gamma^{-2(t-1)}\lambda)I_d \text{ and } \|\theta_t^\star\|_2 \leq S) \\ &\leq \beta_{t-1} \quad \text{(Corollary 3).}\end{aligned}$$

□

B.2 Control of the norm of actions

Lemma 4. *Let* $V_t = \sum_{s=1}^t \gamma^{-s}A_sA_s^\top + \lambda\gamma^{-t}I_d$ *and* $\widetilde{V}_t = \sum_{s=1}^t \gamma^{-2s}A_sA_s^\top + \lambda\gamma^{-2t}I_d$ *and* $0 < \gamma < 1$. *We have*

$$\forall t,\ V_t^{-1}\widetilde{V}_t\,V_t^{-1} \leq \gamma^{-t}\,V_t^{-1}.$$

Proof.

$$\widetilde{V}_t = \sum_{s=1}^t \gamma^{-2s}A_sA_s^\top + \lambda\gamma^{-2t}I_d \leq \gamma^{-t}\sum_{s=1}^t \gamma^{-s}A_sA_s^\top + \lambda\gamma^{-2t}I_d = \gamma^{-t}V_t.$$

Consequently,

$$V_t^{-1}\widetilde{V}_tV_t^{-1} \leq \gamma^{-t}V_t^{-1}V_tV_t^{-1} \leq \gamma^{-t}V_t^{-1}.$$

□

Thanks to Lemma 4 we establish the following proposition,

Proposition 4.

$$\sum_{t=1}^T \min\left(1, \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right) \leq 2\sum_{t=1}^T \log\left(1 + \gamma^{-t}\|A_t\|^2_{V_{t-1}^{-1}}\right) \leq 2\log\left(\frac{\det(V_T)}{\lambda^d}\right).$$

Proof. We first use the fact that: $\forall x \geq 0, \min(1,x) \leq 2\log(1+x)$.

$$\begin{aligned}\min\left(1, \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right) &\leq 2\log\left(1 + \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right) \\ &\leq 2\log\left(1 + \gamma^{-(t-1)}\|A_t\|^2_{V_{t-1}^{-1}}\right) \quad \text{(Lemma 4)} \\ &\leq 2\log\left(1 + \gamma^{-t}\|A_t\|^2_{V_{t-1}^{-1}}\right) \quad (\gamma \leq 1).\end{aligned}$$

Furthermore,

$$V_t \geq \gamma^{-t}A_tA_t^\top + V_{t-1} \geq V_{t-1}^{1/2}(I_d + \gamma^{-t}V_{t-1}^{-1/2}A_tA_t^\top V_{t-1}^{-1/2})V_{t-1}^{1/2}.$$

Given that all those matrices are symmetric positive definite, the previous inequality implies that

$$\begin{aligned}\det(V_t) &\geq \det(V_{t-1})\det(1 + (\gamma^{-t/2}V_{t-1}^{-1/2}A_t)(\gamma^{-t/2}V_{t-1}^{-1/2}A_t)^\top) \\ &\geq \det(V_{t-1})\left(1 + \gamma^{-t}\|A_t\|^2_{V_{t-1}^{-1}}\right) \quad \left(\text{Using } \det(I_d + xx^\top) = 1 + \|x\|_2^2\right).\end{aligned}$$

Therefore,

$$\frac{\det(V_T)}{\det(V_0)} = \prod_{t=1}^T \frac{\det(V_t)}{\det(V_{t-1})} \geq \prod_{t=1}^T (1 + \gamma^{-t}\|A_t\|^2_{V_{t-1}^{-1}}).$$

Finally by applying the log function to the previous inequality,

$$\sum_{t=1}^{T} \min\left(1, \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right) \leq 2\sum_{t=1}^{T} \log\left(1+\gamma^{-t}\|A_t\|^2_{V_{t-1}^{-1}}\right) \leq 2\log\left(\frac{\det(V_T)}{\det(V_0)}\right).$$

□

Corollary 4.

$$\sqrt{\sum_{t=1}^{T} \min\left(1, \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right)} \leq \sqrt{2d}\sqrt{T\log\left(\frac{1}{\gamma}\right) + \log\left(1+\frac{L^2}{d\lambda(1-\gamma)}\right)}.$$

Proof. The proof of this corollary is based on the previous lemma and on Corollary 2. We have

$$\begin{aligned}\log\left(\frac{\det(V_T)}{\det(V_0)}\right) &\leq \log\left(\frac{1}{\lambda^d}\left(\lambda\gamma^{-T} + \frac{L^2(\gamma^{-T}-1)}{d(1-\gamma)}\right)^d\right) \quad \text{(Corollary 2)}\\ &\leq dT\log\left(\frac{1}{\gamma}\right) + d\log\left(1+\frac{L^2}{d\lambda(1-\gamma)}\right).\end{aligned}$$

□

B.3 Proof of Theorem 2

In this subsection we give the proof of Theorem 2 for the high probability upper-bound of the regret for `D-LinUCB`.

Proof.

First step: Upper bound for the instantaneous regret.

Let $A_t^\star = \arg\max_{a\in\mathcal{A}_t}\langle a, \theta_t^\star\rangle$ and $\theta_t = \arg\max_{\theta\in\mathcal{C}_t}\langle A_t, \theta\rangle$. We have,

$$\begin{aligned}r_t &= \max_{a\in\mathcal{A}_t}\langle a, \theta_t^\star\rangle - \langle A_t, \theta_t^\star\rangle = \langle A_t^\star - A_t, \theta_t^\star\rangle\\ &= \langle A_t^\star - A_t, \bar{\theta}_t\rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t\rangle.\end{aligned}$$

Under the event $\{\forall t > 0,\ \bar{\theta}_t \in \mathcal{C}_t\}$, that occurs with probability at least $1-\delta$ thanks to Proposition 3, we have,

$$\langle A_t^\star, \bar{\theta}_t\rangle \leq \arg\max_{\theta\in\mathcal{C}_t}\langle A_t^\star, \theta\rangle = \text{UCB}_t(A_t^\star) \leq \text{UCB}_t(A_t) = \arg\max_{\theta\in\mathcal{C}_t}\langle A_t, \theta\rangle = \langle A_t, \theta_t\rangle. \tag{10}$$

Then, with probability at least $1-\delta$, $\forall t > 0$,

$$\begin{aligned}r_t &\leq \langle A_t, \theta_t - \bar{\theta}_t\rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t\rangle\\ &\leq \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} + \|A_t^\star - A_t\|_2\|\theta_t^\star - \bar{\theta}_t\|_2 \quad \text{(Cauchy-Schwarz)}\\ &\leq \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} + 2L\|\theta_t^\star - \bar{\theta}_t\|_2 \quad (\forall a\in\mathcal{A}_t\|a\|_2 \leq L).\end{aligned}$$

As discussed in Section 3.2, the two terms are upper bounded using different techniques. The first term is handled with the equivalent in a non-stationary environment of the deviation inequality of Theorem 1 and the second term is the equivalent of the bias.

Second step: Upper bound for $\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}}$.

We have,

$$\|\theta_t - \bar{\theta}_t\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq \|\theta_t - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} + \|\bar{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}\widetilde{V}_{t-1}^{-1}V_{t-1}} \leq 2\beta_{t-1},$$

where the last inequality holds because under our assumption $\bar{\theta}_t \in \mathcal{C}_t$ with high probability and by definition $\theta_t \in \mathcal{C}_t$.

Third step: Upper bound for the bias.

Let $D > 0$,

$$\begin{aligned}
\|\theta_t^\star - \bar{\theta}_t\|_2 &= \|V_{t-1}^{-1}\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top(\theta_s^\star - \theta_t^\star)\|_2 \\
&\leq \|\sum_{s=t-D}^{t-1} V_{t-1}^{-1}\gamma^{-s}A_sA_s^\top(\theta_s^\star - \theta_t^\star)\|_2 + \|V_{t-1}^{-1}\sum_{s=1}^{t-D-1}\gamma^{-s}A_sA_s^\top(\theta_s^\star - \theta_t^\star)\|_2 \\
&\leq \|\sum_{s=t-D}^{t-1} V_{t-1}^{-1}\gamma^{-s}A_sA_s^\top\sum_{p=s}^{t-1}(\theta_p^\star - \theta_{p+1}^\star)\|_2 + \|\sum_{s=1}^{t-D-1}\gamma^{-s}A_sA_s^\top(\theta_s^\star - \theta_t^\star)\|_{V_{t-1}^{-2}} \\
&\leq \|\sum_{p=t-D}^{t-1} V_{t-1}^{-1}\gamma^{-s}A_sA_s^\top\sum_{s=t-D}^{p}(\theta_p^\star - \theta_{p+1}^\star)\|_2 + \frac{1}{\lambda}\sum_{s=1}^{t-D-1}\gamma^{t-1-s}\|A_sA_s^\top(\theta_s^\star - \theta_t^\star)\|_2 \\
&\leq \sum_{p=t-D}^{t-1}\|V_{t-1}^{-1}\sum_{s=t-D}^{p}\gamma^{-s}A_sA_s^\top(\theta_p^\star - \theta_{p+1}^\star)\|_2 + \frac{2L^2S}{\lambda}\sum_{s=1}^{t-D-1}\gamma^{t-1-s} \\
&\leq \sum_{p=t-D}^{t-1}\lambda_{\max}\left(V_{t-1}^{-1}\sum_{s=t-D}^{p}\gamma^{-s}A_sA_s^\top\right)\|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{2L^2S}{\lambda}\frac{\gamma^D}{1-\gamma}.
\end{aligned}$$

The first inequality is a consequence of the triangular inequality. The third inequality uses that $V_{t-1}^{-2} \leq (\frac{\gamma^{t-1}}{\lambda})^2 I_d$. In the last inequality, we have used the fact that for a symmetric matrix $M \in \mathcal{M}_d(\mathbb{R})$ and a vector $x \in \mathbb{R}^d$, $\|Mx\|_2 \leq \lambda_{\max}(M)\|x\|_2$.

Furthermore, for x such that $\|x\|_2 \leq 1$, we have that for $t - D \leq p \leq t-1$,

$$\begin{aligned}
x^\top V_{t-1}^{-1}\sum_{s=t-D}^{p}\gamma^{-s}A_sA_s^\top x &\leq x^\top V_{t-1}^{-1}\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top x + \lambda\gamma^{-(t-1)}x^\top V_{t-1}^{-1}x \\
&\leq x^\top V_{t-1}^{-1}(\sum_{s=1}^{t-1}\gamma^{-s}A_sA_s^\top + \lambda\gamma^{-(t-1)}I_d)x = x^\top x \leq 1.
\end{aligned}$$

Therefore, for all p such that $t - D \leq p \leq t-1$, $\lambda_{\max}\left(V_{t-1}^{-1}\sum_{s=t-D}^{p}\gamma^{-s}A_sA_s^\top\right) \leq 1$.

By combining the second and the third step, with probability at least $1-\delta$:

$$r_t \leq 2L\sum_{p=t-D}^{t-1}\|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3S}{\lambda}\frac{\gamma^D}{1-\gamma} + 2\beta_{t-1}\|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}.$$

The assumption $|\langle A_t, \theta_t^\star\rangle| \leq 1$ also implies $r_t \leq 2$. Hence, with probability at least $1-\delta$:

$$r_t \leq 2L\sum_{p=t-D}^{t-1}\|\theta_p^\star - \theta_{p+1}^\star\|_2 + 4L^3S\frac{\gamma^D}{1-\gamma} + 2\beta_{t-1}\min(1, \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}). \tag{11}$$

To conclude the proof we use the results of Subsection B.2.

Final step:

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} r_t \\
&\leq 2L\sum_{t=1}^{T}\sum_{p=t-D}^{t-1}\|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3S}{\lambda}\frac{\gamma^D}{1-\gamma}T + 2\beta_T\sum_{t=1}^{T}\min\left(1, \|A_t\|_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right) \\
&\leq 2L\sum_{t=1}^{T}\sum_{p=t-D}^{t-1}\|\theta_p^\star - \theta_{p+1}^\star\|_2 + \frac{4L^3S}{\lambda}\frac{\gamma^D}{1-\gamma}T + 2\beta_T\sqrt{T}\sqrt{\sum_{t=1}^{T}\min\left(1, \|A_t\|^2_{V_{t-1}^{-1}\widetilde{V}_{t-1}V_{t-1}^{-1}}\right)}
\end{aligned}$$

$$\leq 2LB_T D + \frac{4L^3 S}{\lambda}\frac{\gamma^D}{1-\gamma}T + 2\sqrt{2}\beta_T\sqrt{dT}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}.$$

In the first inequality, we use that $t \mapsto \beta_t$ is increasing. The second inequality is an application of the Cauchy-Schwarz inequality to the third term and the last inequality is an application of Corollary 4. □

B.4 Proof of Corollary 1

Proof. Let γ be defined as $\gamma = 1 - (\frac{B_T}{dT})^{2/3}$ and $D = \frac{\log(T)}{(1-\gamma)}$. With this choice of γ, D is equivalent to $d^{2/3}B_T^{-2/3}T^{2/3}\log(T)$. Thus, DB_T is equivalent to $d^{2/3}B_T^{1/3}T^{2/3}\log(T)$.

In addition,

$$\gamma^D = \exp(D\log(\gamma)) = \exp\left(\frac{\log(\gamma)}{1-\gamma}\log(T)\right) \sim 1/T.$$

Hence, $T\gamma^D\frac{1}{1-\gamma}$ behaves as $d^{2/3}T^{2/3}B_T^{-2/3}$.

Furthermore, $\log(1/\gamma) \sim d^{-2/3}B_T^{2/3}T^{-2/3}$, implying that $T\log(1/\gamma) \sim d^{-2/3}B_T^{2/3}T^{1/3}$.

As a result, it holds that, $\beta_T\sqrt{dT}\sqrt{T\log(1/\gamma) + \log\left(1 + \frac{L^2}{d\lambda(1-\gamma)}\right)}$ is equivalent to $dT^{1/2}\sqrt{\log(T/B_T)}\sqrt{d^{-2/3}B_T^{2/3}T^{1/3}} = d^{2/3}B_T^{1/3}T^{2/3}\sqrt{\log(T/B_T)}$.

By adding those three terms and neglecting the log factors, we obtain the desired result. □

C A new analysis of the `SW-LinUCB` algorithm

In this section we propose a new analysis of the `SW-LinUCB` algorithm. This is useful as the proof provided in [11] has several gaps. First, Lemma 2 of [11] is presented as a specific case of the analysis of [1]. It would hold in the case of a growing window, where the argument developed in [1] could be used, but not with a sliding window, where past actions are removed from the design matrix. Furthermore, Theorem 2 of [11] that bounds $|\langle x, \hat{\theta}_{t-1} - \theta_t^\star\rangle|$ for any fixed direction x with high probability is used in equation (42) with x replaced by A_t, whereas A_t is a random variable strongly related to $\hat{\theta}_{t-1}$.

We only mention this analysis in the Appendix because the deviation inequalities established for the weighted model can not be used. Nevertheless, we believe that this analysis gives new insights on the problem with a sliding window.

C.1 Deviation inequality

Let us introduce some notations to clarify the model. We suppose that there is a sliding window of length l, such that the estimate of the unknown parameter at time t is based on the l last observations. The optimization program solved is

$$\hat{\theta}_t = \underset{\theta\in\mathbb{R}^d}{\arg\min}\left(\sum_{s=\max(1,t-l+1)}^{t}(X_s - \langle A_s, \theta\rangle)^2 + \lambda/2\|\theta\|_2^2)\right).$$

One has

$$\hat{\theta}_t = V_t^{-1}\sum_{s=\max(1,t-l+1)}^{t} A_s X_s, \quad \text{where} \quad V_t = \sum_{s=\max(1,t-l+1)}^{t} A_s A_s^\top + \lambda I_d. \tag{12}$$

The expression linking the matrices V_t and V_{t-1} is the following

$$V_t = V_{t-1} + A_t A_t^\top - A_{t-l}A_{t-l}^\top.$$

The specificity of the sliding window model is that at time t, to update the design matrix, a new action vector A_t is added but the oldest term A_{t-l} is also removed . When considering the equivalent of the quantity $M_t(x)$ defined in the Appendix A, the property of supermartingale does not hold anymore because of this loss of information. For this reason, all the reasoning that was done in [1] can not be applied directly.

The reward generation process we consider is still the one presented in Equation 1. As for the `D-LinUCB` model, the results are stated with σ-subgaussian random noises but the proofs are done with $\sigma = 1$. Let $S_t = \sum_{s=\max(1,t-l+1)}^{t} A_s \eta_s$. We start by giving the proof of the analogue of Lemma 2 presented in [11]. We give an instantaneous deviation inequality.

Proposition 5 (Instantaneous deviation inequality with a sliding window)**.** *Let t be a fixed time instant. For all $\delta > 0$,*

$$\mathbb{P}\left(\|S_t\|_{V_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)}\right) \leq \delta.$$

Proof. We present an interesting trick in this proof for avoiding the loss of information due to the sliding window that is only usable for instantaneous deviation inequalities.

Let t be the time instant of interest. We assume that $t \geq l$. We know that the estimate $\hat{\theta}_t$ is only based on observations between time $t - l + 1$ to t. The trick is to create a fictive regression model starting a time $t - l$ and receiving the exact same information as the true model between the time instants $t - l + 1$ to t.

To ease the understanding of the proof, the notations with dotted symbols refer to the fictive model. Let u be a time instant in $[\![t-l, t]\!]$. Let $\dot{V}_u = \sum_{\max(1,t-l+1)}^{u} A_s A_s^\top + \lambda I_d$, $\dot{S}_u = \sum_{\max(1,t-l+1)}^{u} A_s \eta_s$ and $\dot{M}_u(x) = \exp(x^\top \dot{S}_u - x^\top \dot{V}_u(0)x/2)$. Once again, $\dot{V}_u(0) = \sum_{\max(1,t-l+1)}^{u} A_s A_s^\top$ corresponds to the design matrix without the regularization term. By definition, $\forall x \in \mathbb{R}^d, \dot{M}_{t-l}(x) = 1$.

Using the 1-subgaussianity and following the lines of the proof of Lemma 1,

$$\mathbb{E}[\dot{M}_u(x)|\mathcal{F}_{u-1}] \leq \dot{M}_{u-1}(x).$$

Therefore, $\forall u \in [\![t - l, t]\!], \mathbb{E}[\dot{M}_u(x)] \leq \mathbb{E}[\dot{M}_{t-l}(x)] = 1$. In particular for $u = t$, $\forall x \in \mathbb{R}^d, \mathbb{E}[\dot{M}_t(x)] \leq 1$. By introducing a measure of probability $h = \mathcal{N}(0, \frac{1}{\lambda} I_d)$, we still have $\mathbb{E}\left[\int \dot{M}_t(x) dh(x)\right] \leq 1$ using a similar reasoning than in Lemma 2. We can also give an exact formula for $\int \dot{M}_t(x) dh(x)$ with the chosen h. Let us remark that $\dot{S}_t = S_t$ and $\dot{V}_t = V_t$.

$$\begin{aligned}
\int_{\mathbb{R}^d} \dot{M}_t(x) dh(x) &= \frac{1}{\sqrt{(2\pi)^d \det(1/\lambda I_d)}} \int_{\mathbb{R}^d} \exp\left(x^\top S_t - \frac{1}{2}\|x\|^2_{\lambda I_d} - \frac{1}{2}\|x\|^2_{V_t(0)}\right) dx \\
&= \frac{1}{\sqrt{(2\pi)^d \det(1/\lambda I_d)}} \int_{\mathbb{R}^d} \exp\left(1/2\|S_t\|^2_{V_t^{-1}} - 1/2\|x - V_t^{-1} S_t\|^2_{V_t}\right) dx \\
&= \frac{\exp\left(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right)}{\sqrt{(2\pi)^d \det(1/\lambda I_d)}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|x - V_t^{-1} S_t\|^2_{V_t}\right) dx \\
&= \frac{\exp\left(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right)}{\sqrt{(2\pi)^d \det(1/\lambda I_d)}} \sqrt{(2\pi)^d \det\left(V_t^{-1}\right)} \\
&= \exp\left(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right) \sqrt{\frac{\det(\lambda I_d)}{\det(V_t)}}.
\end{aligned}$$

For this reason,

$$\mathbb{P}\left(\|S_t\|_{V_t^{-1}} \geq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t)}{\det(\lambda I_d)}\right)}\right)$$

$$
\begin{aligned}
&= \mathbb{P}\left(\exp\left(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right)\sqrt{\frac{\det(\lambda I_d)}{\det(V_t)}} \geq \frac{1}{\delta}\right) \\
&\leq \delta\mathbb{E}\left[\int_{\mathbb{R}^d} \dot{M}_t(x)dh(x)\right] \text{ (Markov's inequality)} \\
&\leq \delta.
\end{aligned}
$$

□

The next step is to upper-bound the quantity $\det(V_t)$ similarly as in Proposition 2 for the weighted model.

Proposition 6 (Determinant inequality for the design matrix with a sliding window)**.** *In the specific case where V_t is defined as $V_t = \sum_{s=\max(1,t-l+1)}^{t} A_s A_s^\top + \lambda I_d$. Under the assumption $\forall t, \|A_t\|_2 \leq L$, the following holds,*

$$
\det(V_t) \leq \left(\lambda + \frac{L^2 \min(t,l)}{d}\right)^d.
$$

The proof of this proposition is the same as in Proposition 2. By using the previous inequality, we can obtain the following proposition,

Proposition 7. *When using a sliding window model where the last l terms are considered, for all $\delta > 0$,*

$$
\mathbb{P}\left(\exists t \leq T, \|S_t\|_{V_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{T}{\delta}\right) + d\log\left(1 + \frac{L^2\min(t,l)}{\lambda d}\right)}\right) \leq \delta.
$$

Proof.

$$
\begin{aligned}
&\mathbb{P}\left(\exists t \leq T, \|S_t\|_{V_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{T}{\delta}\right) + d\log\left(1 + \frac{L^2\min(t,l)}{\lambda d}\right)}\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left(\|S_t\|_{V_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{T}{\delta}\right) + d\log\left(1 + \frac{L^2\min(t,l)}{\lambda d}\right)}\right) \\
&\leq \sum_{t=1}^{T} \mathbb{P}\left(\|S_t\|_{V_t^{-1}} \geq \sigma\sqrt{2\log\left(\frac{T}{\delta}\right) + \log\left(\frac{\det(V_t)}{\lambda^d}\right)}\right) \\
&\leq \sum_{t=1}^{T} \frac{\delta}{T} \quad \text{(Proposition 5)} \leq \delta.
\end{aligned}
$$

□

C.2 Regret analysis

The regret analysis of the `SW-LinUCB` algorithm is similar to the one proposed for `D-LinUCB`. We start by defining the confidence ellipsoid used by the algorithm `SW-LinUCB`.

With the `SW-LinUCB` algorithm, the β_t term is defined in the following way,

$$
\beta_t = \sqrt{\lambda}S + \sigma\sqrt{2\log\left(\frac{T}{\delta}\right) + d\log\left(1 + \frac{L^2\min(t,l)}{\lambda d}\right)} \tag{13}
$$

Remark: The cost of loosing some information at each step due to the sliding window when $t > l$ is the term $\log\left(\frac{T}{\delta}\right)$ rather than $\log\left(\frac{1}{\delta}\right)$ in the definition of β_t.

Note that due to the use of a union bound technique the confidence radius is larger than the one suggested in [11]. Nevertheless, this was not taken into account in simulations for `SW-LinUCB`.

Proposition 8. *Let* $\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}^{-1}} \leq \beta_{t-1}\right\}$ *denote the confidence ellipsoid. Let* $\bar{\theta}_t = V_{t-1}^{-1}\left(\sum_{s=\max(1,t-l)}^{t-1} A_s A_s^\top \theta_s^\star + \lambda\theta_t^\star\right)$*. Then,* $\forall\delta > 0$*,*

$$\mathbb{P}\left(\forall t \geq 1, \bar{\theta}_t \in \mathcal{C}_t\right) \geq 1 - \delta.$$

Proof.

$$\begin{aligned}\bar{\theta}_t - \hat{\theta}_{t-1} &= V_{t-1}^{-1}\left(\sum_{s=\max(1,t-l)}^{t-1} A_s A_s^\top \theta_s^\star + \lambda\theta_t^\star - \sum_{s=\max(1,t-l)}^{t-1} A_s A_s^\top \theta_s^\star - \sum_{s=\max(1,t-l)}^{t-1} A_s \eta_s\right)\\ &= -V_{t-1}^{-1} S_{t-1} + \lambda V_{t-1}^{-1}\theta_t^\star.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\bar{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}^{-1}} &\leq \|S_{t-1}\|_{V_{t-1}^{-1}} + \lambda\|\theta_t^\star\|_{V_{t-1}^{-1}}\\ &\leq \|S_{t-1}\|_{V_{t-1}^{-1}} + \sqrt{\lambda}S \quad (V_{t-1}^{-1} \leq \frac{1}{\lambda}I_d)\\ &\leq \beta_{t-1} \quad \text{(with probability} \geq 1-\delta \text{ thanks to Proposition 7).}\end{aligned}$$

□

We need to bound the quantity $\sum_{t=1}^T \min\left(1, \|A_t\|_{V_{t-1}^{-1}}^2\right)$. An analysis of this quantity is already proved in [11]. Nevertheless, we provide a simpler analysis in the following proposition.

Proposition 9. *With the sliding window model, the following upper bound holds,*

$$\sum_{t=1}^T \min\left(1, \|A_t\|_{V_{t-1}^{-1}}^2\right) \leq 2d\lceil T/l\rceil \log\left(1 + \frac{lL^2}{\lambda d}\right).$$

Proof. We start by rewriting the sum as follows.

$$\sum_{t=1}^T \min\left(1, \|A_t\|_{V_{t-1}^{-1}}^2\right) = \sum_{k=0}^{\lceil T/l\rceil - 1}\sum_{t=kl+1}^{(k+1)l} \min\left(1, \|A_t\|_{V_{t-1}^{-1}}^2\right)$$

For the k-th block of length l we define the matrix $W_t^{(k)} = \sum_{s=kl+1}^t A_s A_s^\top + \lambda I_d$. We also have $\forall t \in [\![kl, (k+1)l]\!], V_t \geq W_t^{(k)}$ as every term in $W_t^{(k)}$ is contained in V_t and the extra-terms in V_t correspond to positive definite matrices. The matrices are definite positive, thus $V_t^{-1} \leq (W_t^{(k)})^{-1}$ and consequently,

$$\sum_{k=0}^{\lceil T/l\rceil - 1}\sum_{t=kl+1}^{(k+1)l} \min\left(1, \|A_t\|_{V_{t-1}^{-1}}^2\right) \leq \sum_{k=0}^{\lceil T/l\rceil - 1}\sum_{t=kl+1}^{(k+1)l} \min\left(1, \|A_t\|_{(W_{t-1}^{(k)})^{-1}}^2\right)$$

Furthermore, $\forall t \in [\![kl, (k+1)l]\!]$ we have,

$$\det(W_t^{(k)}) = \det(W_{t-1}^{(k)})\left(1 + \|A_t\|_{(W_{t-1}^{(k)})^{-1}}^2\right).$$

With positive definitive matrices whose determinants are strictly positive, this implies that

$$\frac{\det(W_{(k+1)l}^{(k)})}{\det(W_{kl}^{(k)})} = \prod_{t=kl+1}^{(k+1)l} \frac{\det(W_t^{(k)})}{\det(W_{t-1}^{(k)})} = \prod_{t=kl+1}^{(k+1)l}\left(1 + \|A_t\|_{(W_{t-1}^{(k)})^{-1}}^2\right).$$

By definition we have $W_{kl}^{(k)} = \lambda I_d$ and $\forall x \geq 0, \min(1, x) \leq 2\log(1+x)$. So,

$$\sum_{t=1}^T \min\left(1, \|A_t\|_{V_{t-1}^{-1}}^2\right) \leq 2\sum_{k=0}^{\lceil T/l\rceil - 1}\sum_{t=kl+1}^{(k+1)l} \log\left(1 + \|A_t\|_{(W_{t-1}^{(k)})^{-1}}^2\right)$$

$$\leq 2 \sum_{k=0}^{\lceil T/l \rceil - 1} \log \left(\frac{\det(W_{(k+1)l}^{(k)})}{\lambda^d} \right).$$

Knowing that $W_{(k+1)l}^{(k)}$ contains exactly l terms allows us to give the following bound (by following the proof of Proposition 2),

$$\det(W_{(k+1)l}^{(k)}) \leq \left(\lambda + \frac{L^2 l}{d} \right)^d.$$

Finally,

$$\sum_{t=1}^{T} \min \left(1, \|A_t\|_{V_{t-1}^{-1}}^2 \right) \leq 2d \lceil T/l \rceil \log \left(1 + \frac{L^2 l}{\lambda d} \right).$$

□

With those results we can give a high probability upper bound for the cumulative dynamic regret of the `SW-LinUCB` algorithm.

Theorem 3. *Assuming that $\sum_{s=1}^{T-1} \|\theta_s^\star - \theta_{s+1}^\star\|_2 \leq B_T$, the regret of the* `SW-LinUCB` *algorithm may be bounded for all $l > 0$, with probability at least $1 - \delta$, by*

$$R_T \leq 2LB_T l + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{\lceil T/l \rceil} \sqrt{\log \left(1 + \frac{L^2 l}{\lambda d} \right)},$$

where β_T is defined in Equation (13).

Proof.

1rst step: Upper bound for the instantaneous regret

Defining $A_t^\star = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta_t^\star \rangle$ and $\theta_t = \arg\max_{\theta \in \mathcal{C}_t} \langle A_t, \theta \rangle$. We have,

$$\begin{aligned} r_t &= \max_{a \in \mathcal{A}_t} \langle a, \theta_t^\star \rangle - \langle A_t, \theta_t^\star \rangle = \langle A_t^\star - A_t, \theta_t^\star \rangle \\ &= \langle A_t^\star - A_t, \bar{\theta}_t \rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t \rangle \end{aligned}$$

Under the event $\{\forall t > 0,\ \bar{\theta}_t \in \mathcal{C}_t\}$, that occurs with probability at least $1 - \delta$ thanks to Proposition 8,

$$\langle A_t^\star, \bar{\theta}_t \rangle \leq \arg\max_{\theta \in \mathcal{C}_t} \langle A_t^\star, \theta \rangle = \text{UCB}_t(A_t^\star) \leq \text{UCB}_t(A_t) = \arg\max_{\theta \in \mathcal{C}_t} \langle A_t, \theta \rangle = \langle A_t, \theta_t \rangle \tag{14}$$

Using Inequality (14), with probability larger than $1 - \delta$, $\forall t > 0$,

$$\begin{aligned} r_t &\leq \langle A_t, \theta_t - \bar{\theta}_t \rangle + \langle A_t^\star - A_t, \theta_t^\star - \bar{\theta}_t \rangle \\ &\leq \|A_t\|_{V_{t-1}^{-1}} \|\theta_t - \bar{\theta}_t\|_{V_{t-1}} + \|A_t^\star - A_t\|_2 \|\theta_t^\star - \bar{\theta}_t\|_2 \quad \text{(Cauchy-Schwarz)} \\ &\leq \|A_t\|_{V_{t-1}^{-1}} \|\theta_t - \bar{\theta}_t\|_{V_{t-1}} + 2L \|\theta_t^\star - \bar{\theta}_t\|_2 \quad \text{(Bounded action assumption)}. \end{aligned}$$

As for the analysis of the regret for the `D-LinUCB` algorithm, the two terms are upper bounded using different techniques. The first term is handled with the deviation inequality of Proposition 8.

2nd step: Upper bound for $\|\theta_t - \bar{\theta}_t\|_{V_{t-1}}$

We have,

$$\|\theta_t - \bar{\theta}_t\|_{V_{t-1}} \leq \|\theta_t - \hat{\theta}_{t-1}\|_{V_{t-1}} + \|\bar{\theta}_t - \hat{\theta}_{t-1}\|_{V_{t-1}} \leq 2\beta_{t-1}.$$

Where the last inequality holds because under our assumption $\bar{\theta}_t \in \mathcal{C}_t$ with probability at least $1 - \delta$ and by definition $\theta_t \in \mathcal{C}_t$.

3rd step: Upper bound for the bias.

This step is similar to the proof proposed in [11] for Lemma 1.

$$\begin{aligned}
\|\theta_t^\star - \bar{\theta}_t\|_2 &= \left\| V_{t-1}^{-1} \left(\sum_{s=\max(1,t-l)}^{t-1} A_s A_s^\top (\theta_s^\star - \theta_t^\star) \right) \right\|_2 \\
&\leq \left\| \sum_{s=\max(1,t-l)}^{t-1} V_{t-1}^{-1} A_s A_s^\top \sum_{p=s}^{t-1} (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 \\
&\leq \left\| \sum_{p=\max(1,t-l)}^{t-1} V_{t-1}^{-1} \sum_{s=\max(1,t-l)}^{p} A_s A_s^\top (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 \\
&\leq \sum_{p=\max(1,t-l)}^{t-1} \left\| V_{t-1}^{-1} \sum_{s=\max(1,t-l)}^{p} A_s A_s^\top (\theta_p^\star - \theta_{p+1}^\star) \right\|_2 \\
&\leq \sum_{p=\max(1,t-l)}^{t-1} \lambda_{\max} \left(V_{t-1}^{-1} \sum_{s=\max(1,t-l)}^{p} A_s A_s^\top \right) \|\theta_p^\star - \theta_{p+1}^\star\|_2 .
\end{aligned}$$

Furthermore, for $x \in \mathbb{R}^d$ such that $\|x\|_2 \leq 1$, we have that for $\max(1, t-l) \leq p \leq t-1$,

$$\begin{aligned}
x^\top V_{t-1}^{-1} \sum_{s=\max(1,t-l)}^{p} A_s A_s^\top x &\leq x^\top V_{t-1}^{-1} \sum_{s=\max(1,t-l)}^{t-1} A_s A_s^\top x + \lambda x^\top V_{t-1}^{-1} x \\
&\leq x^\top V_{t-1}^{-1} \left(\sum_{s=\max(1,t-l)}^{t-1} A_s A_s^\top + \lambda I_d \right) x = x^\top x \leq 1.
\end{aligned}$$

By combining the second and the third step,

$$r_t \leq 2L \sum_{p=\max(1,t-l)}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + 2\beta_{t-1} \|A_t\|_{V_{t-1}^{-1}} .$$

By using the assumption $\forall a \in \mathcal{A}_t, |\langle A_t, \theta_t^\star \rangle| \leq 1$, we also have $r_t \leq 2$. So, with probability greater than $1-\delta$,

$$r_t \leq 2L \sum_{p=\max(1,t-l)}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + 2\beta_{t-1} \min\left(1, \|A_t\|_{V_{t-1}^{-1}}\right) . \tag{15}$$

To conclude the proof, we use the results of Proposition 9.

<u>Final step:</u>

$$\begin{aligned}
R_T = \sum_{t=1}^{T} r_t &\leq 2L \sum_{t=1}^{T} \sum_{p=\max(1,t-l)}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + 2\beta_T \sum_{t=1}^{T} \min\left(1, \|A_t\|_{V_{t-1}^{-1}}\right) \\
&\leq 2L \sum_{t=1}^{T} \sum_{p=\max(1,t-l)}^{t-1} \|\theta_p^\star - \theta_{p+1}^\star\|_2 + 2\beta_T \sqrt{T} \sqrt{\sum_{t=1}^{T} \min\left(1, \|A_t\|_{V_{t-1}^{-1}}^2\right)} \\
&\leq 2LB_T l + 2\sqrt{2}\beta_T \sqrt{dT} \sqrt{\lceil T/l \rceil} \sqrt{\log\left(1 + \frac{lL^2}{\lambda d}\right)} .
\end{aligned}$$

In the first inequality, we use the fact that $t \mapsto \beta_t$ is increasing. The second inequality is an application of the Cauchy-Schwarz inequality to the second term. The last inequality is an application of Proposition 9 □

By denoting $\tilde{O}$ the function growth when omitting the logarithmic terms, we have the following Corollary.

Corollary 5 (Asymptotic regret bound for `SW-LinUCB`)**.** *If B_T is known, by choosing $l = (\frac{dT}{B_T})^{2/3}$, the regret of the* `SW-LinUCB` *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}B_T^{1/3}T^{2/3})$ when $T \to \infty$.*

If B_T is unknown, by choosing $l = d^{2/3}T^{2/3}$, the regret of the `SW-LinUCB` *algorithm is asymptotically upper bounded with high probability by a term $\tilde{O}(d^{2/3}B_T T^{2/3})$ when $T \to \infty$.*

Proof. With this particular choice of l, we have:

$$lB_T \sim d^{2/3}T^{2/3}B_T^{1/3}.$$

β_T as defined by equation (13) is equivalent to $\sqrt{d\log(T)}$.

$\sqrt{T}\sqrt{\lceil T/l \rceil}$ has a similar behavior than $d^{-1/3}T^{1-1/3}B_T^{1/3}$, consequently the behavior of $\beta_T\sqrt{dT}\sqrt{\lceil T/l \rceil}\sqrt{\log\left(1+\frac{lL^2}{\lambda d}\right)}$ is similar to $d^{2/3}B_T^{1/3}T^{2/3}\sqrt{\log(T)}\sqrt{\log(T/B_T)}$.

By neglecting the logarithmic term, we have with high probability,

$$R_T = \tilde{O}_{T\to\infty}(d^{2/3}B_T^{1/3}T^{2/3}).$$

□