# Dynamical Linear Bandits

**Marco Mussi** [1] **Alberto Maria Metelli** [1] **Marcello Restelli** [1]

## Abstract

In many real-world sequential decision-making problems, an action does not immediately reflect on the feedback and spreads its effects over a long time frame. For instance, in online advertising, investing in a platform produces an instantaneous increase of *awareness*, but the actual reward, i.e., a *conversion*, might occur far in the future. Furthermore, whether a conversion takes place depends on: how fast the awareness grows, its vanishing effects, and the synergy or interference with other advertising platforms. Previous work has investigated the Multi-Armed Bandit framework with the possibility of delayed and aggregated feedback, without a particular structure on how an action propagates in the future, disregarding possible dynamical effects. In this paper, we introduce a novel setting, the Dynamical Linear Bandits (DLB), an extension of the linear bandits characterized by a hidden state. When an action is performed, the learner observes a noisy reward whose mean is a linear function of the hidden state and of the action. Then, the hidden state evolves according to linear dynamics, affected by the performed action too. We start by introducing the setting, discussing the notion of optimal policy, and deriving an expected regret lower bound. Then, we provide an optimistic regret minimization algorithm, Dynamical Linear Upper Confidence Bound (`DynLin-UCB`), that suffers an expected regret of order $\widetilde{\mathcal{O}}\Big(\frac{d\sqrt{T}}{(1-\overline{\rho})^{3/2}}\Big)$, where $\overline{\rho}$ is a measure of the stability of the system, and $d$ is the dimension of the action vector. Finally, we conduct a numerical validation on a synthetic environment and on real-world data to show the effectiveness of `DynLin-UCB` in comparison with several baselines.

[1]Politecnico di Milano, Milan, Italy.
Correspondence to: Marco Mussi <marco.mussi@polimi.it>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

## 1. Introduction

In a large variety of sequential decision-making problems, a learner must choose an action that, when executed, determines an evolution of the underlying system state that is hidden to the learner. In these partially observable problems, the learner observes a reward (i.e., feedback) representing the combined effect of multiple actions played in the past. For instance, in online advertising campaigns, the process that leads to a *conversion*, i.e., *marketing funnel* (Court et al., 2009), is characterized by complex dynamics and comprises several phases. When heterogeneous campaigns/platforms are involved, a profitable budget investment policy has to account for the interplay between campaigns/platforms. In this scenario, a conversion (e.g., a user's purchase of a promoted product) should be attributed not only to the latest ad the user was exposed to, but also to previous ones (Berman, 2018).

The *joint* consideration of each funnel phase is a fundamental step towards an optimal investment solution while considering the advertising campaigns/platforms *independently* leads to sub-optimal solutions. Consider, for instance, a simplified version of the funnel with two types of campaigns: *awareness* (i.e., impression) ads and *conversion* ads. The first kind of ad aims at improving brand awareness, while the latter aims at creating the actual conversion. If we evaluate the performances in terms of conversions only, we will discover that impression ads are not instantaneously effective in creating conversions, so we will be tempted to reduce the budget invested in such a campaign. However, this approach is sub-optimal because impression ads increase the chance to convert when a conversion ad is shown after the impression (e.g., Hoban & Bucklin, 2015). In addition, the effect of some ads, especially impression ads delivered via television, may be delayed. It has been demonstrated (Chapelle, 2014) that users remember advertising over time in a vanishing way, leading to consequences that non-dynamical models cannot capture. This kind of interplay comprises more general scenarios than the simple reward delay, including the case where the interaction is governed by a dynamics *hidden* to the observer.

While this scenario can be indubitably modeled as a Partially Observable Markov Decision Process (POMDP, Åström, 1965), the complexity of the framework and its general-

ity are often not required to capture the main features of the problem. Indeed, for specific classes of problems, the Multi-Armed Bandit (MAB, Lattimore & Szepesvári, 2020) literature has explored the possibility of experiencing delayed reward either assuming that the actual reward will be observed, individually, in the future (e.g., Joulani et al., 2013) or with the more realistic assumption that an aggregated feedback is available (e.g., Pike-Burke et al., 2018), with also specific applications to online advertising (Vernade et al., 2017). Although effective in dealing with delay effects and the possibility of a reward spread in the future (Cesa-Bianchi et al., 2018), they do not account for the additional, more complex, dynamical effects, which can be regarded as the evolution of a hidden state.

In this work, we take a different perspective. We propose to model the non-observable dynamical effects underlying the phenomena as a Linear Time-Invariant (LTI) system (Hespanha, 2018). In particular, the system is characterized by a hidden internal state $\mathbf{x}_t$ (e.g., awareness) which evolves via linear dynamics fed by the action $\mathbf{u}_t$ (e.g., amount invested) and affected by noise. At each round, the learner experiences a reward $y_t$ (e.g., conversions), which is a noisy observation that linearly combines the state $\mathbf{x}_t$ and the action $\mathbf{u}_t$. Our goal consists in learning an optimal policy so as to maximize the expected cumulative reward. We call this setting *Dynamical Linear Bandits* (DLBs) that, as we shall see, reduces to linear bandits (Abbasi-Yadkori et al., 2011) when no dynamics are involved. Because of the dynamics, the effect of each action persists over time indefinitely but, under stability conditions, it vanishes asymptotically. This allows representing interference and synergy between platforms, thanks to the dynamic nature of the system.

**Contributions** In Section 2, we introduce the Dynamical Linear Bandit (DLB) setting to represent sequential decision-making problems characterized by a hidden state that evolves linearly according to an *unknown* dynamics. We show that, under stability conditions, the optimal policy corresponds to playing the *constant action* that leads the system to the most profitable steady state. Then, we derive an expected regret lower bound of order $\left(\frac{d\sqrt{T}}{(1-\overline{\rho})^{1/2}}\right)$, being $d$ the dimensionality of the action space and $\overline{\rho}$ the spectral radius of the dynamical matrix of the system evolution law.[1] In Section 3, we propose a novel optimistic regret minimization algorithm, *Dynamical Linear Upper Confidence Bound* (`DynLin-UCB`), for the DLB setting. `DynLin-UCB` takes inspiration from `Lin-UCB` but subdivides the optimization horizon $T$ into increasing-length epochs. In each epoch, an action is selected optimistically and kept constant (i.e., persisted) so that the system approximately reaches the steady state. We provide a regret analysis for `DynLin-UCB` showing that, under certain assumptions, it enjoys $\widetilde{\mathcal{O}}\left(\frac{d\sqrt{T}}{(1-\overline{\rho})^{3/2}}\right)$ expected regret. In Section 5, we provide a numerical validation, with both synthetic and real-world data, compared with bandit baselines. The proofs of all the results are reported in Appendix B.

**Notation** Let $a, b \in \mathbb{N}$ with $a \leq b$, we introduce the symbols: $[a, b] := \{a, \dots, b\}$, $[b] := [1, b]$, and $[a, \infty) := \{a, a+1, \dots\}$. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we denote with $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{j=1}^n x_i y_i$ the inner product. For a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we denote with $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ the weighted 2-norm. The *spectral radius* $\rho(\mathbf{A})$ is the largest absolute value of the eigenvalues of $\mathbf{A}$, the *spectral norm* $\|\mathbf{A}\|_2$ is the square root of the maximum eigenvalue of $\mathbf{A}^\top \mathbf{A}$. We introduce the maximum spectral norm to spectral radius ratio of the powers of $\mathbf{A}$ defined as $\overline{\rho}(\mathbf{A}) = \sup_{\tau \geqslant 0} \|\mathbf{A}^\tau\|_2 \{\rho(\mathbf{A})^\tau$ (Oymak & Ozay, 2019). We denote with $\mathbf{I}_n$ the identity matrix of order $n$ and with $\mathbf{0}_n$ the vector of all zeros of dimension $n$. A random vector $\mathbf{x} \in \mathbb{R}^n$ is $\sigma^2$-subgaussian, in the sense of Hsu et al. (2012), if for every vector $\boldsymbol{\zeta} \in \mathbb{R}^n$ it holds that $\mathbb{E} \exp(\langle \mathbf{x}, \boldsymbol{\zeta} \rangle - \langle \mathbb{E}[\mathbf{x}], \boldsymbol{\zeta} \rangle) \leqslant \exp(\|\boldsymbol{\zeta}\|_2^2 \sigma^2 / 2)$.

## 2. Setting

In this section, we introduce the *Dynamical Linear Bandits* (DLBs), the learner-environment interaction, assumptions, and regret (Section 2.1). Then, we derive a closed-form expression for the optimal policy for DLBs (Section 2.2). Finally, we derive a lower bound to the regret, highlighting the intrinsic complexities of the DLB setting (Section 2.3).

### 2.1. Problem Formulation

In a Dynamical Linear Bandit (DLB), the environment is characterized by a *hidden* state, i.e., a $n$-dimensional real vector, initialized to $\mathbf{x}_1 \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ is the state space. At each round $t \in \mathbb{N}$, the environment is in the hidden state $\mathbf{x}_t \in \mathcal{X}$, the learner chooses an action, i.e., a $d$-dimensional real vector $\mathbf{u}_t \in \mathcal{U}$, where $\mathcal{U} \subseteq \mathbb{R}^d$ is the action space. Then, the learner receives a noisy reward $y_t = \langle \boldsymbol{\omega}, \mathbf{x}_t \rangle + \langle \boldsymbol{\theta}, \mathbf{u}_t \rangle + \eta_t \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}$ is the reward space, $\boldsymbol{\omega} \in \mathbb{R}^n$, $\boldsymbol{\theta} \in \mathbb{R}^d$ are unknown parameters, and $\eta_t$ is a zero-mean $\sigma^2$–subgaussian random noise, conditioned to the past. Then, the environment evolves to the new state according to the unknown linear dynamics $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t$, where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the dynamic matrix, $\mathbf{B} \in \mathbb{R}^{n \times d}$ is the action-state matrix, and $\boldsymbol{\epsilon}_t$ is a zero-mean $\sigma^2$–subgaussian random noise, conditioned to the past, independent of $\eta_t$.[2]

**Remark 2.1.** *The setting proposed above is a particular case of a POMDP (Åström, 1965), in which the state $\mathbf{x}_t$ is non-observable, while the learner accesses the noisy*

[1] The smaller $\overline{\rho}$, the faster the system reaches its steady state.

[2] $n$ is the *order* of the LTI system (Kalman, 1963). We make no assumption on the value of $n$ and on its knowledge.

*observation $y_t$ that corresponds to the noisy reward too. Furthermore, the setting can be viewed as a MISO (Multiple Input Single Output) discrete-time LTI system (Kalman, 1963). Finally, the DLB reduces to (non-contextual) linear bandit (Abbasi-Yadkori et al., 2011) when the hidden state does not affect the reward, i.e., when $\boldsymbol{\omega} = \mathbf{0}$.*

**Markov Parameters** We revise a useful representation, that for every $H \in \mathbb{N}$ allows expressing $y_t$ in terms of the sequence of the most recent $H$ actions $(\mathbf{u}_s)_{s\in[t-H,t)}$, reward noise $\eta_t$, $H$ state noises $(\boldsymbol{\epsilon}_s)_{s\in[t-H,t-1)}$, and starting state $\mathbf{x}_{t-H}$ (Ho & Kalman, 1966; Oymak & Ozay, 2019; Tsiamis & Pappas, 2019; Sarkar et al., 2021):

$$y_t = \underbrace{\sum_{s=0}^{H} \langle \mathbf{h}^{\{s\}}, \mathbf{u}_{t-s} \rangle}_{\text{action effect}} + \underbrace{\boldsymbol{\omega}^\top \mathbf{A}^H \mathbf{x}_{t-H}}_{\text{starting state}} + \underbrace{\eta_t + \sum_{s=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{s-1} \boldsymbol{\epsilon}_{t-s}}_{\text{noise}}, \quad (1)$$

where the sequence of vectors $\mathbf{h}^{\{s\}} \in \mathbb{R}^d$ for every $s \in \mathbb{N}$ are called *Markov parameters* and are defined as: $\mathbf{h}^{\{0\}} = \boldsymbol{\theta}$ and $\mathbf{h}^{\{s\}} = \mathbf{B}^\top \mathbf{A}^{s-1} \boldsymbol{\omega}$ if $s \geq 1$. Furthermore, we introduce the *cumulative Markov parameters*, defined for every $s, s' \in \mathbb{N}$ with $s \leq s'$ as $\mathbf{h}^{[s,s']} = \sum_{l=s}^{s'} \mathbf{h}^{\{l\}}$ and the corresponding limit as $s' \to \infty$, i.e., $\mathbf{h}^{[s,+\infty)} = \sum_{l=s}^{+\infty} \mathbf{h}^{\{l\}}$. Finally, we use the abbreviation $\mathbf{h} = \mathbf{h}^{[0,+\infty)} = \boldsymbol{\theta}^\top \mathbf{B} \otimes \mathbf{A}^\top \boldsymbol{\omega}$.

We will make use of the following standard assumption related to the *stability* of the dynamic matrix $\mathbf{A}$, widely employed in discrete–time LTI literature (Oymak & Ozay, 2019; Lale et al., 2020a;b).

**Assumption 2.1** (Stability)**.** *The spectral radius of $\mathbf{A}$ is strictly smaller than 1, i.e., $\rho(\mathbf{A}) < 1$, and the maximum spectral norm to spectral radius ratio of the powers of $\mathbf{A}$ is bounded, i.e., $\kappa(\mathbf{A}) < \infty$.*[3]

**Policies and Performance** The learner's behavior is modeled via a deterministic *policy* $\underline{\boldsymbol{\pi}} = (\boldsymbol{\pi}_t)_{t\in\mathbb{N}}$ defined, for every round $t \in \mathbb{N}$, as $\boldsymbol{\pi}_t : \mathcal{H}_{t-1} \to \mathcal{U}$, mapping the history of observations $H_{t-1} = (\mathbf{u}_1, y_1, \ldots, \mathbf{u}_{t-1}, y_{t-1}) \in \mathcal{H}_{t-1}$ to an action $\mathbf{u}_t = \boldsymbol{\pi}_t(H_{t-1}) \in \mathcal{U}$, where $\mathcal{H}_{t-1} = \mathcal{U} \times \mathcal{Y})^{t-1}$ is the set of histories of length $t - 1$. The performance of a policy $\underline{\boldsymbol{\pi}}$ is evaluated in terms of the *(infinite-horizon) expected average reward*:

$$J(\underline{\boldsymbol{\pi}}) : = \liminf_{H\to+\infty} \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{H} y_t\right], \quad (2)$$

$$\text{where} \quad \begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t \\ y_t = \langle \boldsymbol{\omega}, \mathbf{x}_t \rangle + \langle \boldsymbol{\theta}, \mathbf{u}_t \rangle + \eta_t, \quad \forall t \in \mathbb{N}, \\ \mathbf{u}_t = \boldsymbol{\pi}_t(H_{t-1}) \end{cases}$$

where the expectation is taken w.r.t. the randomness of the state noise $\boldsymbol{\epsilon}_t$ and reward noise $\eta_t$. If a policy $\underline{\boldsymbol{\pi}}$ is *constant*, i.e., $\boldsymbol{\pi}_t(H_{t-1}) = \mathbf{u}$ for every $t \in \mathbb{N}$, we abbreviate $J(\mathbf{u}) = J(\underline{\boldsymbol{\pi}})$. A policy $\underline{\boldsymbol{\pi}}^*$ is an *optimal policy* if it maximizes the expected average reward, i.e., $\underline{\boldsymbol{\pi}}^* \in \arg\max_{\underline{\boldsymbol{\pi}}} J(\underline{\boldsymbol{\pi}})$, and its performance is denoted by $J^* := J(\underline{\boldsymbol{\pi}}^*)$.

We further introduce the following assumption that requires the boundedness of the norms of the relevant quantities.

**Assumption 2.2** (Boundedness)**.** *There exist $B, U, X > 0$ s.t.: $\|\boldsymbol{\theta}\|_2 \leq B$, $\|\boldsymbol{\omega}\|_2 \leq B$, $\|\mathbf{B}\|_2 \leq B$, $\sup_{\mathbf{u}\in U} \|\mathbf{u}\|_2 \leq U$, and $\sup_{\mathbf{x}\in X} \|\mathbf{x}\|_2 \leq X$, $\sup_{\mathbf{u}\in U} |J(\mathbf{u})| \leq 1$.*[4]

**Regret** The *regret* suffered by playing a policy $\underline{\boldsymbol{\pi}}$, competing against the optimal infinite-horizon policy $\underline{\boldsymbol{\pi}}^*$ over a *learning horizon* $T \in \mathbb{N}$ is given by:

$$R(\underline{\boldsymbol{\pi}}, T) : = TJ^* - \sum_{t=1}^{T} y_t, \quad (3)$$

where $y_t$ is the sequence of rewards collected by playing $\underline{\boldsymbol{\pi}}$ as in Equation (2). The goal of the learner consists in minimizing the *expected regret* $\mathbb{E}\, R(\underline{\boldsymbol{\pi}}, T)$, where the expectation is taken w.r.t. the randomness of the reward.

### 2.2. Optimal Policy

In this section, we derive a closed-form expression for the optimal policy $\underline{\boldsymbol{\pi}}^*$ for the infinite–horizon objective function, as introduced in Equation (2).

**Theorem 2.1** (Optimal Policy)**.** *Under Assumptions 2.1 and 2.2, an optimal policy $\underline{\boldsymbol{\pi}}^*$ maximizing the (infinite-horizon) expected average reward $J(\underline{\boldsymbol{\pi}})$ (Equation 2), for every round $t \in \mathbb{N}$ and history $H_{t-1} \in \mathcal{H}_{t-1}$ is given by:*

$$\boldsymbol{\pi}_t^*(H_{t-1}) = \mathbf{u}^* \quad \text{where} \quad \mathbf{u}^* \in \operatorname*{Argmax}_{\mathbf{u}\in U} J(\mathbf{u}) \quad \langle \mathbf{h}, \mathbf{u} \rangle. \quad (4)$$

Some remarks are in order. The optimal policy plays the *constant* action $\mathbf{u}^* \in \mathcal{U}$ which brings the system in the "most profitable" steady-state.[5] Indeed, the expression $\langle \mathbf{h}, \mathbf{u} \rangle$ can be rewritten expanding the cumulative Markov parameter as $\boldsymbol{\theta}^\top + \boldsymbol{\omega}^\top \mathbf{B} + \boldsymbol{\omega}^\top \mathbf{A}\mathbf{B}\mathbf{u}^* + \cdots$ and $\bar{\mathbf{x}}^* = \mathbf{B} + \mathbf{A}\mathbf{B}\mathbf{u}^* + \cdots$ is the expression of the steady state $\bar{\mathbf{x}}^* = \mathbf{A}\bar{\mathbf{x}}^* + \mathbf{B}\mathbf{u}^*$, when applying action $\mathbf{u}^*$. It is worth noting the role of Assumption 2.1 which guarantees the existence of the inverse $(\mathbf{I} - \mathbf{A})^{-1}$. In this sense, our problem shares the constant nature of the optimal policy with the linear bandit setting (Abbasi-Yadkori et al., 2011), although ours is characterized by an evolving state, which introduces a new trade-off in the action selection. From the LTI system perspective, this implies that we can restrict to *open-loop stationary* policies. The reason why DLBs do not benefit from *closed-loop* policies, differently from other classical problems, such as

[3] The latter is a mild assumption: if $\mathbf{A}$ is diagonalizable as $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$, then $\kappa(\mathbf{A}) \leq \|\mathbf{Q}\|_2 \|\mathbf{Q}^{-1}\|_2$ and it is finite. In particular, if $\mathbf{A}$ is symmetric then $\kappa(\mathbf{A}) = 1$.

[4] The assumption of the bounded state norm $\|\mathbf{x}\|_2 \leq X$ holds whenever the state noise $\boldsymbol{\epsilon}$ is bounded. As shown by Agarwal et al. (2019), this assumption can be relaxed, for unbounded subgaussian noise, by conditioning to the event that none of the noise vectors are ever large at the cost of an additional $\log T$ factor in the regret.

[5] In Appendix C, we show that the optimal policy is non–stationary for the finite–horizon case.

the LQG (Abbasi-Yadkori & Szepesvári, 2011), lies in the linearity of the reward $y_t$ and in the additive noise $\eta_t$ and $\epsilon_t$, making their presence irrelevant (in expectation) for control purposes. Nonetheless, as we shall see, our problem poses additional challenges compared to linear bandits since, in order to assess the quality of an action $\mathbf{u} \in \mathcal{U}$, instantaneous rewards are not reliable, and we need to let the system evolve to the steady state and, only then, observe the reward.

### 2.3. Regret Lower Bound

In this section, we provide a lower bound to the expected regret that any learning algorithm suffers when addressing the learning problem in a DLB.

**Theorem 2.2** (Lower Bound)**.** *For any policy $\underline{\boldsymbol{\pi}}$ (even stochastic), there exists a DLB fulfilling Assumptions 2.1 and 2.2, such that for sufficiently large $T \geq \mathcal{O}\left(\frac{d^2}{1-\rho(\mathbf{A})}\right)$, policy $\underline{\boldsymbol{\pi}}$ suffers an expected regret lower bounded by:*

$$\mathbb{E}R(\underline{\boldsymbol{\pi}}, T) \geq \left(\frac{d\sqrt{T}}{(1-\rho(\mathbf{A}))^{\frac{1}{2}}}\right).$$

The lower bound highlights the main challenges of the DLB learning problem. First of all, we observe a dependence on $1/(1-\rho(\mathbf{A}))$, being $\rho(\mathbf{A})$ the spectral radius of the matrix $\mathbf{A}$. This is in line with the intuition that, as $\rho(\mathbf{A})$ approaches 1, the problem becomes more challenging. Furthermore, we note that when $\rho(\mathbf{A}) = 0$, i.e., the problem has no dynamical effects, the lower bound matches the one of linear bandits (Lattimore & Szepesvári, 2020). It is worth noting that, for technical reasons, the result of Theorem 2.2 is derived under the assumption that, at every round $t \in [T]$, the agent observes *both* the state $\mathbf{x}_t$ and the reward $y_t$ (see Appendix B). Clearly, this represents a simpler setting w.r.t. DLBs (in which $\mathbf{x}_t$ is hidden) and, consequently, Theorem 2.2 is a viable lower bound for DLBs too.

## 3. Algorithm

In this section, we present an *optimistic* regret minimization algorithm for the DLB setting. *Dynamical Linear Upper Confidence Bound* (`DynLin-UCB`), whose pseudocode is reported in Algorithm 1, requires the knowledge of an upper-bound $\overline{\rho} > 1$ on the spectral radius of the dynamic matrix $\mathbf{A}$ (i.e., $\rho(\mathbf{A}) < \overline{\rho}$) and on the maximum spectral norm to spectral radius ratio $\overline{\gamma} > 8$ (i.e., $\|\mathbf{A}\| < \overline{\gamma}$), as well as the bounds on the relevant quantities of Assumption 2.2.[6]

`DynLin-UCB` is based on the following simple observation. To assess the quality of action $\mathbf{u} \in \mathcal{U}$, we need to *persist* in applying it so that the system approximately reaches the corresponding steady state and, then, observe the reward $y_t$, representing a reliable estimate of $J(\mathbf{u}) = \mathbf{x}\mathbf{h}, \mathbf{u}\mathbf{y}$. We shall show that, under Assumption 2.1, the number of rounds needed to approximately reach such a steady state is logarithmic in the learning horizon $T$ and depends on the upper bound of the spectral norm $\overline{\rho}$. After initializing the Gram matrix $\mathbf{V}_0 = \lambda \mathbf{I}_d$ and the vectors $\mathbf{b}_0$ and $\widehat{\mathbf{h}}_0$ both to $\mathbf{0}_d$ (line 1), `DynLin-UCB` subdivides the learning horizon $T$ in $M \leq T$ *epochs*. Each epoch $m \in [M]$ is composed of $H_m = 1$ rounds, where $H_m = \lceil \log m / \log(1/\overline{\rho}) \rceil$ is logarithmic in the epoch index $m$. At the beginning of each epoch, $m \in [M]$, `DynLin-UCB` computes the upper confidence bound (UCB) index (line 4) defined for every $\mathbf{u} \in \mathcal{U}$ as:

$$\mathrm{UCB}_t(\mathbf{u}) : \langle \mathbf{x}\widehat{\mathbf{h}}_{t-1}, \mathbf{u}\mathbf{y} \rangle + \beta_{t-1} \|\mathbf{u}\|_{\mathbf{V}_{t-1}^{-1}}, \qquad (5)$$

where $\widehat{\mathbf{h}}_{t-1} = \mathbf{V}_{t-1}^{-1}\mathbf{b}_{t-1}$ is the Ridge regression estimator of the cumulative Markov parameter $\mathbf{h}$, as in Equation (4) and $\beta_{t-1} \geq 0$ is an exploration coefficient to be defined later. Similar to `Lin-UCB` (Abbasi-Yadkori et al., 2011), the index $\mathrm{UCB}_t(\mathbf{u})$ is designed to be optimistic, i.e., $J(\mathbf{u}) \leq \mathrm{UCB}_t(\mathbf{u})$ in high-probability for all $\mathbf{u} \in \mathcal{U}$. Then, the optimistic action $\mathbf{u}_t \in \arg\max_{\mathbf{u} \in U} \mathrm{UCB}_t(\mathbf{u})$ is executed (line 6) and persisted for the next $H_m$ rounds (lines 8-11). The length of the epoch $H_m$ is selected such that, under Assumption 2.1, the system has approximately reached the steady state after $H_m - 1$ rounds. In this way, at the end of epoch $m$, the reward $y_t$ is an almost-unbiased sample of the steady-state performance $J(\mathbf{u}_t)$. This sample is employed to update the Gram matrix estimate $\mathbf{V}_t$ and the vector $\mathbf{b}_t$ (line 13), while the samples collected in the previous $H_m$ rounds are discarded (line 9). It is worth noting that by setting $H_m = 0$ for all $m \in [M]$, `DynLin-UCB` reduces to `Lin-UCB`. The following sections provide the concentration of the estimator $\widehat{\mathbf{h}}_{t-1}$ of $\mathbf{h}$ (Section 3.1) and the regret analysis of `DynLin-UCB` (Section 3.2).

### 3.1. Self-Normalized Concentration Inequality for the Cumulative Markov Parameter

In this section, we provide a self-normalized concentration result for the estimate $\widehat{\mathbf{h}}_t$ of the cumulative Markov parameter $\mathbf{h}$. For every epoch $m \in [M]$, we denote with $t_m$ the last round of epoch $m$: $t_0 = 0$ and $t_m = t_{m-1} + H_m$. At the end of each epoch $m$, we solve the Ridge regression problem, defined for every round $t \in [T]$ as:

$$\widehat{\mathbf{h}}_t \in \arg\min_{\widetilde{\mathbf{h}} \in \mathbb{R}^d} \sum_{l \in [M]: t_l \leq t_m} (y_{t_l} - \langle \mathbf{x}\widetilde{\mathbf{h}}, \mathbf{u}_{t_l} \rangle)^2 + \lambda \|\widetilde{\mathbf{h}}\|_2^2 = \mathbf{V}_t^{-1}\mathbf{b}_t.$$

[6] As an alternative, one can consider a more demanding requirement of the knowledge of a bound on the spectral norm $\|\mathbf{A}\|_2$ of $\mathbf{A}$. Similar assumptions regarding the knowledge of analogous quantities are considered in the literature, e.g., *decay of Markov operator norms* (Simchowitz et al., 2020) and *strong stability* (Plevrakis & Hazan, 2020), spectral norm bound (Lale et al., 2020a). As a side note, the knowledge of $\overline{\rho} \geq \rho(\mathbf{A})$ (or an equivalent quantity) is proved to be unavoidable by Theorem 2.2. Indeed, if no restriction on $\rho(\mathbf{A})$ is enforced (i.e., just $\rho(\mathbf{A}) < 1$), one can always consider the DLB in which $\rho(\mathbf{A}) = 1 - 1/T$ making the regret lower bound degenerate to linear.

**Algorithm 1:** DynLin-UCB.

**Input :** Regularization parameter $\lambda > 0$, exploration coefficients $(\beta_t)_{t \in [T]}$, spectral radius upper bound $0 < \overline{\rho} < 1$

Initialize $t = 1$, $\mathbf{V}_0 = \lambda \mathbf{I}_d$, $\mathbf{b}_0 = \mathbf{0}_d$, $\widehat{\mathbf{h}}_0 = \mathbf{0}_d$,
Define $M = \min\{M' \in \mathbb{N} : \sum_{m=1}^{M'} \lceil \frac{\log m}{\log(1/\overline{\rho})} \rceil > T\} - 1$
**for** $m \in [\![M]\!]$ **do**
  Compute $\mathbf{u}_t \in \arg\max_{\mathbf{u} \in \mathcal{U}} \mathrm{UCB}_t(\mathbf{u})$
    where $\mathrm{UCB}_t(\mathbf{u}) = \langle \widehat{\mathbf{h}}_{t-1}, \mathbf{u} \rangle + \beta_{t-1} \|\mathbf{u}\|_{\mathbf{V}_{t-1}^{-1}}$
  Play arm $\mathbf{u}_t$ and observe $y_t$
  Define $H_m = \lceil \frac{\log m}{\log(1/\overline{\rho})} \rceil$
  **for** $j \in [\![H_m]\!]$ **do**
    Update $\mathbf{V}_t = \mathbf{V}_{t-1}$, $\mathbf{b}_t = \mathbf{b}_{t-1}$
    $t \leftarrow t + 1$
    Play arm $\mathbf{u}_t = \mathbf{u}_{t-1}$ and observe $y_t$
  **end**
  Update $\mathbf{V}_t = \mathbf{V}_{t-1} + \mathbf{u}_t \mathbf{u}_t^\top$, $\mathbf{b}_t = \mathbf{b}_{t-1} + \mathbf{u}_t y_t$
  Compute $\widehat{\mathbf{h}}_t = \mathbf{V}_t^{-1} \mathbf{b}_t$
  $t \leftarrow t + 1$
**end**

We now present the following self-normalized maximal concentration inequality and, then, we compare it with the existing results in the literature.

**Theorem 3.1** (Self-Normalized Concentration)**.** *Let $(\widehat{\mathbf{h}}_t)_{t \in \mathbb{N}}$ be the sequence of solutions of the Ridge regression problems of Algorithm 1. Then, under Assumption 2.1 and 2.2, for every $\lambda \geq 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, simultaneously for all rounds $t \in \mathbb{N}$, it holds that:*

$$\left\| \widehat{\mathbf{h}}_t - \mathbf{h} \right\|_{\mathbf{V}_t} < \frac{\zeta_1}{\sqrt{\lambda}} \log(e/t)^{1+q} + c_2 \sqrt{\lambda} \sqrt{2\widetilde{\sigma}^2 \left( \log\left(\frac{1}{\delta}\right) + \frac{1}{2} \log\left( \frac{\det \mathbf{V}_t}{\lambda^d} \right) \right)},$$

*where $c_1 = U + (1+q)\left(\frac{UB}{1-\rho(\mathbf{A})} + X\right)$, $c_2 = \frac{\Omega B \Phi(\mathbf{A})}{1-\rho(\mathbf{A})}$, and $\widetilde{\sigma}^2 = \sigma^2 \left(1 + \frac{\Omega^2 \Phi(\mathbf{A})^2}{1-\rho(\mathbf{A})^2}\right)$.*

First, we note that when $\mathbf{u}_t = \mathbf{0}$ ($\omega = \mathbf{0}_n$), i.e., the state does not affect the reward, the bound perfectly reduces to the self-normalized concentration used in linear bandits (Abbasi-Yadkori et al., 2011, Theorem 1). In particular, we recognize the second term due to the regularization parameter $\lambda > 0$ and the third one, which involves the subgaussianity parameter $\widetilde{\sigma}^2$, related to the joint contribution of the state and reward noises. Furthermore, the first term is an additional bias that derives from the epochs of length $H_m > 1$. The choice of the value $H_m$ represents one of the main technical novelties that, on the one hand, leads to a bias that conveniently grows logarithmically with $t$ and, on the other hand, can be computed without the knowledge of $T$.

It is worth looking at our result from the perspective of learning the LTI system parameters. We can compare our Theorem 3.1 with the concentration presented in (Lale et al., 2020a, Appendix C), which represents, to the best of our knowledge, the only result for the closed-loop identification of LTI systems with non-observable states. First, note that, although we focus on a MISO system ($y_t$ is a scalar, being our reward), extending our estimator to multiple-outputs (MIMO) is straightforward. Second, the approach of (Lale et al., 2020a) employs the *predictive form* of the LTI system to cope with the correlation introduced by closed-loop control. This choice allows for convenient analysis of the estimated Markov parameters of the predictive form. However, recovering the parameters of the original system requires an application of the Ho-Kalman method (Ho & Kalman, 1966) which, unfortunately, does not preserve the concentration properties in general, but only for *persistently exciting* actions. Our method, instead, forces to play an open-loop policy within a single epoch (each with logarithmic duration), while the overall behavior is closed-loop, as the next action depends on the previous-epoch estimates. In this way, we are able to provide a concentration guarantee on the parameters of the original system without assuming additional properties on the action signal.

### 3.2. Regret Analysis

In this section, we provide the analysis of the regret of DynLin-UCB, when we select the exploration coefficient $\beta_t$ based on the knowledge of the upper bounds $\overline{\rho} > 1$, $\overline{\sigma} > 0$, and those specified in Assumption 2.2, defined for every round $t \in [\![T]\!]$ as:

$$\beta_t : \frac{\overline{\zeta}_1}{\sqrt{\lambda}} \log(e/t)^{1+q} + \overline{c}_2 \sqrt{\lambda} \sqrt{2\overline{\sigma}^2 \left( \log\left(\frac{1}{\delta}\right) + \frac{d}{2} \log\left(1 + \frac{tU^2}{d\lambda}\right) \right)},$$

where $\overline{c}_1 = U + (1+q)\left(\frac{UB}{1-\overline{\rho}} + X\right)$, $\overline{c}_2 = \frac{\Omega B \overline{\Phi}}{1-\overline{\rho}}$, and $\overline{\sigma}^2 = \sigma^2 \left(1 + \frac{\Omega^2 \overline{\Phi}^2}{1-\overline{\rho}^2}\right)$. The following result provides the bound on the expected regret of DynLin-UCB.

**Theorem 3.2** (Upper Bound)**.** *Under Assumptions 2.1 and 2.2, selecting $\beta_t$ as in Equation (6) and $\delta \in (0, 1/T)$, DynLin-UCB suffers an expected regret bounded as (highlighting the dependencies on $T$, $\overline{\rho}$, $d$, and $\sigma$ only):*

$$\mathbb{E}\, R(\pi^{\texttt{DynLin-UCB}}, T) < \mathcal{O}\left( \frac{d\sigma \sqrt{\overline{T}} \log T^{\frac{3}{2}}}{1-\overline{\rho}} + \frac{\sqrt{d\overline{T}} \log T^2}{(1-\overline{\rho})^{\frac{3}{2}}} + \frac{1}{(1-\rho(\mathbf{A}))^2} \right).$$

*Proof Sketch.* The analysis of DynLin-UCB poses additional challenges compared to that of Lin-UCB (Abbasi-Yadkori et al., 2011) because of the dynamic effects of the hidden state. The idea behind the proof is to first derive

a bound on a different notion of regret, i.e., the *offline regret*: $R^{\text{off}}(\underline{\boldsymbol{\pi}}, T) = TJ^* - \sum_{t=1}^{T} J(\mathbf{u}_t)$, that compares $J^*$ with the steady-state performance $J(\mathbf{u}_t)$ of the action $\mathbf{u}_t = \boldsymbol{\pi}_t(H_{t-1})$ (Theorem B.2). This analysis of $R^{\text{off}}(\underline{\boldsymbol{\pi}}, T)$ can be comfortably carried out, by adopting a proof strategy similar to that of `Lin-UCB`. However, when applying action $\mathbf{u}_t$, the DLB does not immediately reach the performance $J(\mathbf{u}_t)$ as the expected reward $\mathbb{E}[y_t]$ experiences a transitional phase before converging to the steady state. Under stability (Assumption 2.1), it is possible to show that the expected offline regret and the expected regret differ by a constant: $|\mathbb{E}\, R(\underline{\boldsymbol{\pi}}, T) - \mathbb{E}\, R^{\text{off}}(\underline{\boldsymbol{\pi}}, T)| = \mathcal{O}(1/(1-\rho(\mathbf{A}))^2)$ (Lemma B.1). □

Some observations are in order. We first note a dependence on the term $1/(1-\overline{\rho})$, which, in turn, depends on the upper bound $\overline{\rho}$ of the spectral gap $\rho(\mathbf{A})$. If the system does not display a dynamics, i.e., we can set $\overline{\rho} = 0$, we obtain a regret bound that, apart from logarithmic terms, coincides with that of `Lin-UCB`, i.e., $\widetilde{\mathcal{O}}(d\sigma\sqrt{T})$. Instead, for slow-converging systems, i.e., $\overline{\rho} \to 1$, the regret bound enlarges, as expected. Clearly, a value of $\overline{\rho}$ too large compared to the optimization horizon $T$ (e.g., $\overline{\rho} = 1 - 1/\sqrt{T^{1/3}}$) makes the regret bound degenerate to linear. This is a case in which the underlying system is so slow that the whole horizon $T$ is insufficient to approximately reach the steady state. Third, the regret bound is the sum of three components: the first one depends on the subgaussian proxy $\sigma$ and is due to the noisy estimation of the relevant quantities; the second one is a bias due to the epoch-based structure of `DynLin-UCB`; finally, the third one is constant (does not depend on $T$) accounts for the time needed to reach the steady state.

**Remark 3.1** (Regret upper bound (Theorem 3.2) and lower bound (Theorem 2.2) Comparison)**.** *Apart from logarithmic terms, we notice a tight dependence on $d$ and on $T$. Instead, concerning the spectral properties of $\mathbf{A}$, in the upper bound, we experience a dependence on $1/(1-\overline{\rho})$ raised to a higher power (either 1 for the $\sqrt{T}$ term multiplied by $d$ and 3/2 for the term multiplied by $\sqrt{d}$) w.r.t. the exponent appearing in the lower bound (i.e., 1/2). It is currently an open question whether the lower bound is not tight (which is obtained for a simpler setting in which the state is observable $\mathbf{x}_t$) or whether more efficient algorithms for DLBs can be designed. Furthermore, Theorem 3.2 highlights the impact of the upper bound $\overline{\rho}$ compared with the true $\rho(\mathbf{A})$.*

## 4. Related Works

In this section, we survey and compare the literature with a particular focus on bandits with delayed, aggregated, and composite feedback (Joulani et al., 2013) and online control for Linear Time-Invariant (LTI) systems (Hespanha, 2018). Additional related works are reported in Appendix A.

**Bandits with Delayed/Aggregated/Composite Feedback** The Multi-Armed Bandit setting has been widely employed as a principled approach to address sequential decision-making problems (Lattimore & Szepesvári, 2020). The possibility of experiencing delayed rewards has been introduced by Joulani et al. (2013) and widely exploited in advertising applications (Chapelle, 2014; Vernade et al., 2017). A large number of approaches have extended this setting either considering stochastic delays (Vernade et al., 2020), unknown delays (Li et al., 2019; Lancewicki et al., 2021), arm-dependent delays (Manegueu et al., 2020), non-stochastic delays (Ito et al., 2020; Thune et al., 2019; Jin et al., 2022). Some methods relaxed the assumption that the individual reward is revealed after the delay expires, admitting the possibility of receiving anonymous feedback, which can be aggregated (Pike-Burke et al., 2018; Zhang et al., 2021) or composite (Cesa-Bianchi et al., 2018; Garg & Akash, 2019; Wang et al., 2021). Most of these approaches are able to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, plus additional terms depending on the extent of the delay. In our DLBs, the reward is generated over time as a combined effect of past and present actions through a *hidden state*, while these approaches generate the reward instantaneously and reveal it (individually or in aggregate) to the learner in the future and no underlying state dynamics is present.

**Online Control of Linear Time-Invariant Systems** The particular structure imposed by linear dynamics makes our approach comparable to LTI online control for partially observable systems (e.g., Lale et al., 2020b; Simchowitz et al., 2020; Plevrakis & Hazan, 2020). While the dynamical model is similar, in online control of LTI systems, the perspective is quite different. Most of the works either consider the Linear Quadratic Regulator (Mania et al., 2019; Lale et al., 2020b) or (strongly) convex objective functions (Mania et al., 2019; Simchowitz et al., 2020; Lale et al., 2020a), achieving, in most of the cases $\widetilde{\mathcal{O}}(\sqrt{T})$ regret for strongly convex functions and $\widetilde{\mathcal{O}}(T^{2/3})$ for convex functions. Recently, $\widetilde{\mathcal{O}}(\sqrt{T})$ regret rate has been obtained for convex function too, by means of geometric exploration methods (Plevrakis & Hazan, 2020). Compared to `DynLin-UCB`, the algorithm of Plevrakis & Hazan (2020) considers general convex costs but assumes the observability of the state and limits to the class of disturbance response controllers (Li & Bosch, 1993) that do not include the constant policy. Moreover, the regret bound of Plevrakis & Hazan (2020) differs from Theorem 3.2, as it shows a cubic dependence on the system order[7] and an implicit non-trivial dependence on the dynamic matrix $\mathbf{A}$. Instead, our Theorem 3.2 is remarkably independent of the system order $n$. Furthermore, Lale et al. (2020a) reach $\mathcal{O}(\log(T))$ regret in the case of strongly convex cost functions compet-

[7] This holds for *known* cost functions. Instead, for *unknown* costs, the exponent becomes 24 (Plevrakis & Hazan, 2020).

ing against the best *persistently exciting* controller (i.e., a controller implicitly maintaining a non-null exploration). Some approaches are designed to deal with adversarial noise (Simchowitz et al., 2020). All of these solutions, however, look for the best closed-loop controller within a specific class, e.g., disturbance response control (Li & Bosch, 1993). These controllers, however, do not allow us to easily incorporate constraints on the action space, which could be of crucial importance in practice, e.g., in advertising domains. DynLin-UCB works with an arbitrary action space and, thanks to the linearity of the reward, does not require complex closed-loop controllers.

## 5. Numerical Simulations

In this section, we provide numerical validations of DynLin-UCB in both a synthetic scenario and a domain obtained from real-world data. The goal of these simulations is to highlight the behavior of DynLin-UCB in comparison with bandit baselines, describing advantages and disadvantages. The first experiment is a synthetic setting in which we can evaluate the performances of all the solutions and the sensitivity of DynLin-UCB w.r.t. the $\overline{\rho}$ parameter (Section 5.1). Then, we show a comparison in a DLB scenario retrieved from real-world data (Section 5.2). The code of the experiments can be found at https://github.com/marcomussi/DLB. Details and additional experiments can be found in Appendix E.

**Baselines** We consider as main baseline Lin-UCB (Abbasi-Yadkori et al., 2011), designed for linear bandits. We include Exp3 (Auer et al., 1995) usually employed in (non-adaptive) adversarial settings, and its extension to $k$-length memory (adaptive) adversaries Exp3-k by Dekel et al. (2012).[8] Additionally, we perform a comparison with algorithms for regret minimization in non-stationary environments: D-Lin-UCB (Russac et al., 2019), an extension of Lin-UCB for non-stationary settings, and AR2 (Chen et al., 2021), a bandit algorithm for processes presenting temporal structure. Lastly, in the case of real-world data, we compare our solution with a human-expert policy (Expert). This policy is directly generalized from the original dataset by learning via regression the average budget allocation over all platforms from the available data.

For the baselines which do not support vectorial actions, we perform a discretization of the action space $\mathcal{U}$ that surely contains optimal action. Concerning the hyperparameters of the baselines, whenever possible, they are selected as in the respective original papers. The experiments are presented with a regularization parameter $\lambda \in \{1, \log T\}$ for the algorithms which require it (i.e., DynLin-UCB, Lin-UCB, and D-Lin-UCB).[9] Further information about the hyperparameters of the baselines and the adopted optimistic exploration bounds are presented in Appendix E.1.

[8] $k$ is proportional to $\log M\{\log(1\{\overline{\rho}\})$. In Appendix A.3 we elaborate on the use of adversarial bandit algorithms for DLBs.

### 5.1. Synthetic Data

**Setting** We consider a DLB defined by the following matrices $\mathbf{A} = \mathrm{diag}(0.2, 0, 0.1)$, $\mathbf{B} = \mathrm{diag}(0.25, 0, 0.1)$, $\boldsymbol{\theta} = (0, 0.5, 0.1)^\intercal$, $\boldsymbol{\omega} = (1, 0, 0.1)^\intercal$ and a Gaussian noise with $\sigma = 0.01$ (diagonal covariance matrix for the state noise).[10] This way, the spectral gap of the dynamical matrix is $\rho(\mathbf{A}) = 0.2$ and $\|\mathbf{A}\| = 1$. Moreover, the cumulative Markov parameter is given by $\mathbf{h} = (0.56, 0.5, 0.11)^\intercal$. We consider the action space $\mathcal{U} = \{(u_1, u_2, u_3)^\intercal \in [0, 1.5]^3$ with $u_1 + u_2 + u_3 \leq 1.5\}$ that simulates a total budget of 1.5 to be allocated to the three platforms. Thus, a "myopic" agent would simply look at how the action immediately propagates to the reward through $\boldsymbol{\theta}$, and will invest the budget in the second component of the action, which is weighted by 0.5. Instead, a "far-sighted" agent, aware of the system evolution, will look at the cumulative Markov parameter $\mathbf{h}$, realizing that the most convenient action is investing in the first component, weighted by 0.56. Therefore, the optimal action is $\mathbf{u}^* = (1, 0.5, 0)^\intercal$ leading to $J^* = 0.81$.

**Comparison with the bandit baselines** Figure 1 shows the performance in terms of cumulative regret of DynLin-UCB, Lin-UCB, D-Lin-UCB, AR2, Exp3, and Exp3-k. The experiments are conducted over a time horizon of 1 million rounds. For DynLin-UCB, we employed, for the sake of this experiment, the true value of the spectral gap, i.e., $\overline{\rho} = \rho(\mathbf{A}) = 0.2$. First of all, we observe that both Exp3 and Exp3-k suffers a significantly large cumulative regret. Similar behavior is displayed by AR2. Moreover, all the versions of Lin-UCB and D-Lin-UCB suffer linear regret. The best performance of D-Lin-UCB is obtained when the discount factor $\gamma$ is close to 1 (the weights take the form $w_t = \gamma^{-t}$), and the behavior is comparable with the one of Lin-UCB. Even for a quite fast system ($\rho(\mathbf{A}) = 0.2$), ignoring the system dynamics, and the presence of the hidden state, has made both Lin-UCB and D-Lin-UCB commit (in their best version, with $\lambda = \log T$) to the sub-optimal (myopic) action $\mathbf{u}^\circ = (0.5, 1, 0)^\intercal$ with performance $J^\circ = 0.78 < J^*$, with also a relevant variance. On the other hand, DynLin-UCB is able to maintain

[9] For DynLin-UCB, $\log T$ is a nearly optimal choice for $\lambda$ as it can be seen by looking at the first two addenda of the exploration factor in Equation (6).

[10] It is worth noting that the decision of using diagonal matrices is just for explanation purposes and w.l.o.g. (at least in the class of diagonalizable dynamic matrices). Indeed, we are just interested in the cumulative Markov parameter $\mathbf{h}$ and we could have obtained the same results with an equivalent (non-diagonal) representation, by applying an inevitable transformation $\mathbf{T}$ as $\mathbf{A}^1 = \mathbf{TAT}^{-1}$, $\boldsymbol{\omega}^1 = \mathbf{T}^\intercal\boldsymbol{\omega}$, and $\mathbf{B}^1 = \mathbf{TB}$.

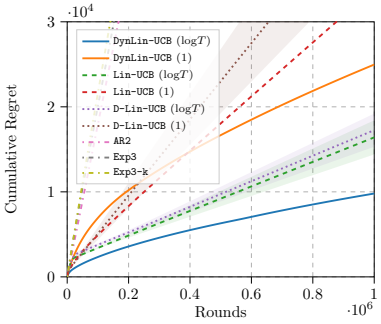



*Figure 1.* Cumulative regret as a function of the rounds comparing `DynLin-UCB` and the other bandit baselines (50 runs, mean std).

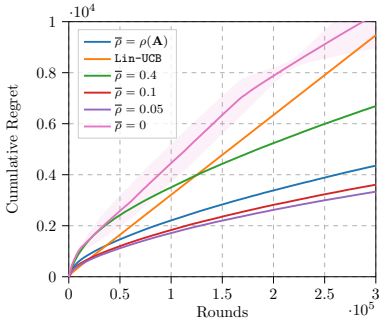



*Figure 2.* Cumulative regret as a function of the rounds comparing `Lin-UCB`, and `DynLin-UCB` with $\lambda$ $\log T$, varying the upper bound on the spectral radius $\overline{\rho}$ (50 runs, mean std).

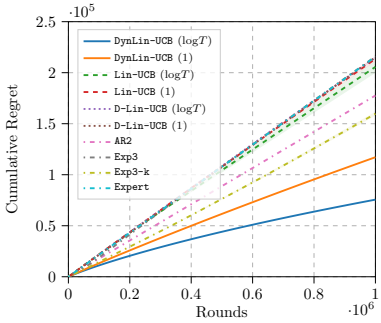



*Figure 3.* Cumulative regret for `DynLin-UCB`, the other bandit baselines and the `Expert` in the system generalized from real-world data (50 runs, mean std).

a smaller and stable (variance is negligible) sublinear regret in both its versions, with a notable advantage when using $\lambda$ $\log T$.

**Sensitivity to the Choice of $\overline{\rho}$** The upper bound $\overline{\rho}$ of the spectral radius $\rho(\mathbf{A})$ 0.2 represents a crucial parameter of `DynLin-UCB`. While an overestimation $\overline{\rho}$ " $\rho(\mathbf{A})$ does not compromise the regret rate but tends to slow down the convergence process, a severe underestimation $\overline{\rho}$ ! $\rho(\mathbf{A})$ might prevent learning at all. In Figure 2, we test `DynLin-UCB` against a misspecification of $\overline{\rho}$, when $\lambda$ $\log T$. We can see that by considering $\overline{\rho}$ $2\rho(\mathbf{A})$, `DynLin-UCB` experiences a larger regret but still sublinear and smaller w.r.t. `Lin-UCB` with $\lambda$ $\log T$. Even by reducing $\overline{\rho}$ P t0.1, 0.05u, `DynLin-UCB` is able to keep the regret sublinear, showing remarkable robustness to misspecification. Clearly, setting $\overline{\rho}$ 0 makes the regret almost degenerate to linear.

### 5.2. Real-world Data

We present an experimental evaluation based on real-world data coming from three web advertising platforms (`Facebook`, `Google`, and `Bing`), related to several campaigns for an invested budget of 5 Million EUR over 2 years. Starting from such data, we learn the best DLB model by means of a specifically designed variant of the Ho-Kalman algorithm (Ho & Kalman, 1966).[11] We used the learned model to build up a simulator. The resulting system has $\rho(\mathbf{A})$ 0.67. We evaluate `DynLin-UCB` against the baselines for $T$ $10^6$ steps over 50 runs.

**Results** Figure 3 shows the results in terms of cumulative regret. It is worth noting that no algorithm, except for `DynLin-UCB`, is able to converge to the optimal choice. Indeed, they immediately commit to a sub-optimal solution. `DynLin-UCB`, instead, shows a convergence trend towards the optimal policy over time for both $\lambda$ 1 and $\lambda$ $\log T$, even if the best-performing version is the one which employs $\lambda$ $\log T$. The `Expert`, which has a preference towards maximizing the instantaneous effect of the actions only and does not take into account correlations between platforms, displays a sub-optimal performance.

[11] See Appendix D.

## 6. Discussion and Conclusions

In this paper, we have introduced the Dynamical Linear Bandits (DLBs), a novel model to represent sequential decision-making problems in which the system is characterized by a non-observable hidden state that evolves according to linear dynamics and by an observable noisy reward that linearly combines the hidden state and the action played. This model accounts for scenarios that cannot be easily represented by existing bandit models that consider delayed and aggregated feedback. We have derived a regret lower bound that highlights the main complexities of the DLB problem. Then, we have proposed a novel optimistic regret minimization approach, `DynLin-UCB`, that, under stability assumption, is able to achieve sub-linear regret. The numerical simulation in both synthetic and real-world domains succeeded in showing that, in a setting where the baselines mostly suffer linear regret, our algorithm consistently enjoys sublinear regret. Furthermore, `DynLin-UCB` proved to be robust to misspecification of its most relevant hyper-parameter $\overline{\rho}$. To the best of our knowledge, this is the first work addressing this family of problems, characterized by hidden linear dynamics, with a simple, yet effective, bandit-like approach. Short-term future directions include efforts in closing the gap between the regret lower and upper bounds. Long-term future directions should focus on extending the present approach to non-linear system dynamics and embedding in the algorithm additional budget constraints enforced over the optimization horizon.

## Acknowledgements

This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

## References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *The 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

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

Agarwal, N., Hazan, E., and Singh, K. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pp. 10175–10184, 2019.

Åström, K. J. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pp. 322–331. IEEE, 1995.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Bacchiocchi, F., Genalti, G., Maran, D., Mussi, M., Restelli, M., Gatti, N., and Metelli, A. M. Autoregressive bandits. *CoRR*, abs/2212.06251, 2022.

Berman, R. Beyond the last touch: Attribution in online advertising. *Marketing Science*, 37(5):771–792, 2018.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pp. 750–773, 2018.

Chapelle, O. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1097–1105. Association for Computing Machinery, 2014.

Chen, Q., Golrezaei, N., and Bouneffouf, D. Dynamic bandits with temporal structure. *Available at SSRN 3887608*, 2021.

Court, D., Elzinga, D., Mulder, S., and Vetvik, O. J. The consumer decision journey. *McKinsey Quarterly*, 3:96–107, 2009.

Dekel, O., Tewari, A., and Arora, R. Online bandit learning against an adaptive adversary: from regret to policy regret. In *International Conference on Machine Learning*, 2012.

Garg, S. and Akash, A. K. Stochastic bandits with delayed composite anonymous feedback. *CoRR*, abs/1910.01161, 2019.

Gur, Y., Zeevi, A., and Besbes, O. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, pp. 199–207, 2014.

Hespanha, J. P. *Linear Systems Theory: Second Edition*. Princeton University Press, 2018.

Ho, B. L. and Kalman, R. E. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Hoban, P. R. and Bucklin, R. E. Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research*, 52(3):375–393, 2015.

Hsu, D., Kakade, S., and Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

Isom, J. D., Meyn, S. P., and Braatz, R. D. Piecewise linear dynamic programming for constrained pomdps. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 291–296. AAAI Press, 2008.

Ito, S., Hatano, D., Sumita, H., Takemura, K., Fukunaga, T., Kakimura, N., and Kawarabayashi, K. Delay and cooperation in nonstochastic linear bandits. In *Advances in Neural Information Processing Systems*, 2020.

Jin, T., Lancewicki, T., Luo, H., Mansour, Y., and Rosenberg, A. Near-optimal regret for adversarial MDP with delayed bandit feedback. *CoRR*, abs/2201.13172, 2022.

Joulani, P., György, A., and Szepesvári, C. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1453–1461, 2013.

Kalman, R. E. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

Kim, D., Lee, J., Kim, K., and Poupart, P. Point-based value iteration for constrained pomdps. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 1968–1974, 2011.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Logarithmic regret bound in partially observable linear dynamical systems. In *Advances in Neural Information Processing Systems*, 2020a.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Regret minimization in partially observable linear quadratic control. *CoRR*, abs/2002.00082, 2020b.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. Stochastic multi-armed bandits with unrestricted delay distributions. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5969–5978, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, B., Chen, T., and Giannakis, G. B. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 993–1002, 2019.

Li, H. X. and Bosch, P. P. J. V. D. A robust disturbance-based control and its application. *International Journal of Control*, 58(3):537–554, 1993.

Manegueu, A. G., Vernade, C., Carpentier, A., and Valko, M. Stochastic bandits with arm-dependent delays. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3348–3356, 2020.

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pp. 10154–10164, 2019.

Nobari, S. DBA: dynamic multi-armed bandit algorithm. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pp. 9869–9870, 2019.

Oymak, S. and Ozay, N. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American Control Conference*, pp. 5655–5661, 2019.

Pike-Burke, C., Agrawal, S., Szepesvári, C., and Grünewälder, S. Bandits with delayed, aggregated anonymous feedback. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4102–4110, 2018.

Plevrakis, O. and Hazan, E. Geometric exploration for online control. In *Advances in Neural Information Processing Systems*, 2020.

Russac, Y., Vernade, C., and Cappé, O. Weighted linear bandits for non-stationary environments. In *Advances in Neural Information Processing Systems*, pp. 12017–12026, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite time LTI system identification. *J. Mach. Learn. Res.*, 22:26:1–26:61, 2021.

Simchowitz, M., Singh, K., and Hazan, E. Improper learning for non-stochastic control. In *Conference on Learning Theory*, volume 125, pp. 3320–3436. PMLR, 2020.

Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pp. 6538–6547, 2019.

Tsiamis, A. and Pappas, G. J. Finite sample analysis of stochastic system identification. In *58th IEEE Conference on Decision and Control*, pp. 3648–3654, 2019.

Undurti, A. and How, J. P. An online algorithm for constrained pomdps. In *IEEE International Conference on Robotics and Automation*, pp. 3966–3973. IEEE, 2010.

Vernade, C., Cappé, O., and Perchet, V. Stochastic bandit models for delayed conversions. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brückner, M. Linear bandits with stochastic delayed feedback. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9712–9721, 2020.

Wang, S., Wang, H., and Huang, L. Adaptive algorithms for multi-armed bandit with composite and anonymous feedback. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 10210–10217, 2021.

Zhang, M., Tsuchida, R., and Ong, C. S. Gaussian process bandits with aggregated feedback. *CoRR*, abs/2112.13029, 2021.

Åström, K. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

# A. Additional Related Works

In this appendix, we report additional details about the related works.

## A.1. Delayed/Aggregated Feedback with DLBs

In this appendix, we show how we can model *delayed* and *composite* feedback with DLBs. For the delayed feedback, we focus on the case in which either the delay is fixed to the value $\tau \in 1$, i.e., the reward of the pull performed at round $t$ is experienced at round $t \quad \tau$. For the composite feedback, we assume that the reward of the pull performed at round $t$ is spread over the next $\tau \in 1$ rounds with fixed weights $\p w_1, \dots, w_\tau \q$. Denoting with $R_t$ the full reward (not observed) due to the pull performed at round $t$, the agent at round $t$ observes the weighted sum of the rewards reported below:[12]

$$\sum_{l=1}^{\tau} w_l R_{t-l}. \tag{6}$$

These two cases can be modeled as DLBs with a suitable encoding of the arms and choice of matrices. In particular, assuming to have $K$ arms, we take the arm set $\mathcal{U}$ to be the canonical basis of $\R^K$, and we denote with $\boldsymbol{\mu}$ the vector of expected rewards. We define $\boldsymbol{\theta} \quad \mathbf{0}$ and:

$$\mathbf{A} \quad \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \in \R^{\tau \times \tau}, \qquad \mathbf{B} \quad \begin{pmatrix} \boldsymbol{\mu}_K^T \\ \mathbf{0}_K^T \\ \mathbf{0}_K^T \\ \vdots \\ \mathbf{0}_K^T \end{pmatrix} \in \R^{\tau \times K},$$

$$\boldsymbol{\omega}_{\text{delay}} \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \in \R^{\tau}, \qquad \boldsymbol{\omega}_{\text{composite}} \quad \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_\tau \end{pmatrix} \in \R^{\tau}.$$

However, DLBs cannot model random or adversarial delays. Nevertheless, DLBs can capture scenarios of composite feedback in which the reward is spread over an infinite number of rounds. Keeping the $K$-armed case introduced above, we can consider the simplest example of a reward that spreads as an autoregressive process AR(1) with parameter $\gamma \in \p 0, 1 \q$, that cannot be represented using the standard composite feedback. In such a case, we simply need a system with order $n \quad 1$ with matrices (actually scalars):

$$\mathbf{A} \quad \gamma, \qquad \mathbf{B} \quad \mathbf{u}^T, \qquad \boldsymbol{\omega} \quad 1.$$

Clearly, one can consider AR($m$) processes (Bacchiocchi et al., 2022) by employing systems of order $n \quad m \ \i \ 1$.

## A.2. Partially Observable Markov Decision Processes

As already noted, looking at DLBs in their generality, we realize that our model is a particular subclass of the Partially Observable Markov Decision Processes (POMDP, Åström, 1965). However, in the POMDP literature, no particular structure of the hidden state dynamics is assumed. The specific linear dynamics are rarely considered, as well as the possibility of a reward that is a linear combination of the hidden state and the action. Nevertheless, several works accounted for the presence of constraints (Isom et al., 2008; Undurti & How, 2010; Kim et al., 2011) without exploiting the linearity and without regret guarantees.

## A.3. Adversarial Bandits

It is worth elaborating on the adaptation of adversarial MAB algorithms to this setting. First, since the reward distribution in DLBs depends at every round $t$ on the sequence of actions played by the agent prior to $t$, we can reduce the DLB setting to an adversarial bandit with an *adaptive* (or non-oblivious) adversary. Second, such an adversary must have *infinite memory* in principle. Third, our regret definition of Section 2 is a *policy regret* (Dekel et al., 2012) that compares the algorithm performance against playing the optimal policy in hindsight from the beginning, as opposed to the *external regret* often

[12] It is worth noting that the fixed-delay case is a particular case of composite feedback, where $w_1 \quad w_\tau \ 1 \quad 0$ and $w_\tau \quad 1$.

employed for non-adaptive adversaries. It is well known that for infinite-memory adaptive adversaries, no algorithm can achieve sublinear policy regret. Nevertheless, for DLB setting, we know that the effect of the past is always vanishing (given Assumption 2.1 enforcing $\rho(\mathbf{A}) < 1$), so we can approximate our setting as a *finite-memory* setting, by considering memory length $k = \lceil \frac{\log M}{\log 1/\rho} \rceil$, where $M$ is the one defined in Algorithm 1 (line 2), with an additional regret term only logarithmic in the optimization horizon $T$. Then, given this approximation, we can make use of an adversarial bandit algorithm (designed for non-adaptive adversaries) in the framework proposed by Dekel et al. (2012) to make it effective for the finite-memory adaptive adversary setting. In the case of an optimal algorithm, such as `Exp3` (Auer et al., 2002), suffering an external regret of order $\widetilde{\mathcal{O}}(\sqrt{MT})$, being $M$ the number of arms, the version to address this finite-memory adaptive adversary setting suffers a regret bounded by $\widetilde{\mathcal{O}}(k^2 M^{1/3} T^{2/3})$, as shown in Theorem 2 of Dekel et al. (2012).

### A.4. Other Approaches

Non-stationary bandits (Gur et al., 2014) can be regarded as bandits with a hidden state that evolves through a (possibly non-linear) dynamics. The main difference compared with our DLBs is that the hidden state evolves in an *uncontrollable* way, i.e., it does not depend on the sequence of actions performed so far. Russac et al. (2019) extend the linear bandit setting by considering a non-stationary evolution of the parameter $\boldsymbol{\theta}_t^*$. The notion of *dynamic* bandit is further studied by Chen et al. (2021), where an auto-regressive process is considered for the evolution of the reward through time and by Nobari (2019) that propose a practical approach to cope with this setting.

## B. Proofs and Derivations

In this section, we provide the proofs we have omitted in the main paper.

### B.1. Proofs of Section 2

Before we proceed, we introduce a different notion of regret useful for analysis purposes, that we name *offline regret*. This notion of regret compares $J^*$ with the steady-state performance of the action $\mathbf{u}_t = \boldsymbol{\pi}_t(H_{t-1})$ played at each round $t \in [T]$ by the agent:

$$R^{\text{off}}(\underline{\boldsymbol{\pi}}, T) := TJ^* - \sum_{t=1}^{T} J(\mathbf{u}_t). \tag{7}$$

We denote with $\mathbb{E}R^{\text{off}}(\underline{\boldsymbol{\pi}}, T)$ the *expected offline regret*, where the expectation is taken w.r.t. the randomness of the reward. Clearly, the two notions of regret coincide when the system has no dynamics.

The following result relates the offline and the (online) expected regret.

**Lemma B.1.** *Under Assumptions 2.1 and 2.2, for any policy $\underline{\boldsymbol{\pi}}$, it holds that:*

$$\left|\mathbb{E}\, R^{off}(\underline{\boldsymbol{\pi}}, T) - \mathbb{E}\, R(\underline{\boldsymbol{\pi}}, T)\right| \leq \frac{\rho(\mathbf{A})BU}{(1-\rho(\mathbf{A}))^2} + \frac{\rho(\mathbf{A})X}{1-\rho(\mathbf{A})}.$$

*Proof.* First of all, we observe that for any policy, the cumulative effect of the noise components is zero-mean. Thus, it suffices to consider the deterministic evolution of the system. For every $t \in [T]$, let us denote with $\mathbb{E}[y_t]$ the expected reward at time $t$ and with $J(\mathbf{u}_t)$ as the steady-state performance when executing action $\mathbf{u}_t$:

$$\mathbb{E}[y_t] = \sum_{s=0}^{t-1} \langle \mathbf{h}^{\{s\}}, \mathbf{u}_{t-s} \rangle = \boldsymbol{\omega}^\top \mathbf{A}^{t-1}\mathbf{x}_1 + \boldsymbol{\theta}^\top \mathbf{u}_t + \boldsymbol{\omega}^\top \sum_{s=1}^{t-1} \mathbf{A}^{s-1}\mathbf{B}\mathbf{u}_{t-s} = \boldsymbol{\omega}^\top \mathbf{A}^{t-1}\mathbf{x}_1,$$

$$J(\mathbf{u}_t) = \boldsymbol{\theta}^\top \mathbf{u}_t + \boldsymbol{\omega}^\top (\mathbf{I}_d - \mathbf{A})^{-1} \mathbf{u}_t = \boldsymbol{\theta}^\top \mathbf{u}_t + \boldsymbol{\omega}^\top \sum_{s=0}^{+\infty} \mathbf{A}^s \mathbf{u}_t.$$

We now proceed by summing over $t \in [T]$. First of all, we consider the following preliminary result involving $y_t$, which is obtained by rearranging the summations:

$$\sum_{t=1}^{T} \mathbb{E}[y_t] = \boldsymbol{\theta}^\top \sum_{t=1}^{T} \mathbf{u}_t + \boldsymbol{\omega}^\top \sum_{t=1}^{T} \sum_{s=1}^{t-1} \mathbf{A}^{s-1}\mathbf{B}\mathbf{u}_{t-s} + \boldsymbol{\omega}^\top \sum_{t=1}^{T} \mathbf{A}^{t-1}\mathbf{x}_1$$

$$\boldsymbol{\theta}^\top \sum_{t=1}^{T} \mathbf{u}_t \quad \boldsymbol{\omega}^\top \sum_{t=1}^{T-1} \left( \sum_{s=0}^{T-t-1} \mathbf{A}^s \right) \mathbf{B}\mathbf{u}_t \quad \boldsymbol{\omega}^\top \sum_{t=1}^{T} \mathbf{A}^{t-1}\mathbf{x}_1.$$

Thus, we have:

$$\begin{aligned}
\left| \sum_{t=1}^{T} \mathsf{p} J \mathsf{p} \mathbf{u}_t \mathsf{q} \quad \mathsf{Er} y_t \mathsf{sq} \right| \quad & \left| \boldsymbol{\omega}^\top \sum_{t=1}^{T} \left( \sum_{s=0}^{+\infty} \mathbf{A}^s \quad \sum_{s=0}^{T-t-1} \mathbf{A}^s \right) \mathbf{B}\mathbf{u}_t \quad \boldsymbol{\omega}^\top \sum_{t=1}^{T} \mathbf{A}^{t-1}\mathbf{x}_1 \right| \\
& \left| \boldsymbol{\omega}^\top \sum_{t=1}^{T} \left( \sum_{s=T-t}^{+\infty} \mathbf{A}^s \right) \mathbf{B}\mathbf{u}_t \quad \boldsymbol{\omega}^\top \sum_{t=1}^{T} \mathbf{A}^{t-1}\mathbf{x}_1 \right| \\
\text{¤} \quad & \mathsf{p} \mathbf{A} \mathsf{q} BU \sum_{t=1}^{T} \sum_{s=T-t}^{+\infty} \rho \mathsf{p} \mathbf{A} \mathsf{q}^s \qquad \mathsf{p} \mathbf{A} \mathsf{q} X \sum_{t=1}^{T} \rho \mathsf{p} \mathbf{A} \mathsf{q}^{t-1} \qquad (8) \\
\text{¤} \quad & \frac{\mathsf{p} \mathbf{A} \mathsf{q} BU}{1 \quad \rho \mathsf{p} \mathbf{A} \mathsf{q}} \sum_{t=1}^{T} \rho \mathsf{p} \mathbf{A} \mathsf{q}^{T-t} \quad \frac{\mathsf{p} \mathbf{A} \mathsf{q} X}{1 \quad \rho \mathsf{p} \mathbf{A} \mathsf{q}} \qquad (9) \\
\text{¤} \quad & \frac{\mathsf{p} \mathbf{A} \mathsf{q} BU}{\mathsf{p} 1 \quad \rho \mathsf{p} \mathbf{A} \mathsf{q} \mathsf{q}^2} \quad \frac{\mathsf{p} \mathbf{A} \mathsf{q} X}{1 \quad \rho \mathsf{p} \mathbf{A} \mathsf{q}}, \qquad (10)
\end{aligned}$$

where line (8) follows from Assumptions 2.1 and 2.2, lines (9) and (10) follow from bounding the summations with the series. The result follows by observing that:

$$\mathsf{E}\, R^{\text{off}} \mathsf{p} \underline{\boldsymbol{\pi}}, T \mathsf{q} \quad \mathsf{E}\, R \mathsf{p} \underline{\boldsymbol{\pi}}, T \mathsf{q} \quad \sum_{t=1}^{T} \mathsf{p} J \mathsf{p} \mathbf{u}_t \mathsf{q} \quad \mathsf{Er} y_t \mathsf{sq}.$$

□

**Theorem 2.2** (Lower Bound)**.** *For any policy $\underline{\boldsymbol{\pi}}$ (even stochastic), there exists a DLB fulfilling Assumptions 2.1 and 2.2, such that for sufficiently large $T$ ¥ $\mathcal{O}\left(\frac{d^2}{1-\rho(\mathbf{A})}\right)$, policy $\underline{\boldsymbol{\pi}}$ suffers an expected regret lower bounded by:*

$$\mathsf{E} R \mathsf{p} \underline{\boldsymbol{\pi}}, T \mathsf{q} \text{ ¥ } \left( \frac{d \sqrt{T}}{\mathsf{p} 1 \quad \rho \mathsf{p} \mathbf{A} \mathsf{q} \mathsf{q}^{\frac{1}{2}}} \right).$$

*Proof.* To derive the lower bound, we take inspiration from the construction of (Lattimore & Szepesvári, 2020) for linear bandits (Theorem 24.1). We consider a class of DLBs defined in terms of fixed $0 \text{ ¤ } \rho \quad 1$ and $0 \text{ ¤ } \epsilon \text{ ¤ } \rho$ with $\boldsymbol{\omega} \quad \mathbf{1}_d$, $\boldsymbol{\theta} \quad \frac{2(1-\rho)+\epsilon}{2(1-(\rho-\epsilon))} \mathbf{1}_d$, $\mathbf{B} \quad \mathsf{p} 1 \quad \rho \mathsf{q} \mathbf{I}_d$ and with a diagonal dynamical matrix $\mathbf{A} \quad \mathsf{diag} \mathsf{p} \mathbf{a} \mathsf{q}$, defined in terms of the vector $\mathbf{a}$ belonging to the set $\mathcal{A} \quad \mathsf{t} \rho, \rho \quad \epsilon \mathsf{u}^d$. The available actions are $\mathcal{U} \quad \mathsf{t} \quad 1, 1 \mathsf{u}^d$. Let us note that $|\mathcal{A}| \quad |\mathcal{U}| \quad 2^d$. Thus, in our set of DLBs, the vector $\mathbf{a}$ fully characterizes the problem. Moreover, we observe that, given the diagonal $\mathbf{a} \quad \mathsf{diag} \mathsf{p} \mathbf{A} \mathsf{q}$, we can compute the cumulative Markov parameter $\mathbf{h}_\mathbf{a} \quad \mathsf{sign} \mathsf{p} \mathbf{a} \mathsf{q} \frac{\epsilon}{2(1-(\rho-\epsilon))}$.[13] As a consequence the optimal action can be defined as $\mathbf{u}_\mathbf{a}^* \quad \mathsf{sign} \mathsf{p} \mathbf{a} \mathsf{q}$, whose performance is given by $J_\mathbf{a}^* \quad \mathsf{x} \mathbf{h}_\mathbf{a}, \mathbf{u}_\mathbf{a}^* \mathsf{y} \quad \frac{\epsilon d}{2(1-(\rho-\epsilon))}$.

Let us consider the probability distribution over the canonical bandit model induced by executing a policy $\underline{\boldsymbol{\pi}}$ in a DLB characterized by the diagonal of the dynamical matrix $\mathbf{a}$ P $\mathcal{A}$ and with Gaussian diagonal noise:

$$\mathsf{P}_\mathbf{a} \quad \prod_{t=1}^{T} \mathcal{N} \mathsf{p} \mathbf{x}_{t+1} | \mathbf{A}\mathbf{x}_t \quad \mathbf{B}\mathbf{u}_t, \sigma^2 \mathbf{I}_d \mathsf{q} \mathcal{N} \mathsf{p} y_t | \mathsf{x} \boldsymbol{\theta}, \mathbf{u}_t \mathsf{y} \quad \mathsf{x} \boldsymbol{\omega}, \mathbf{x}_t \mathsf{y}, \sigma^2 \mathsf{q} \pi_t \mathsf{p} \mathbf{u}_t | H_{t-1} \mathsf{q},$$

where $H_{t-1}$ is the history of observations up to time $t \quad 1$. We denote with $\mathsf{E}_\mathbf{a}$ the expectation induced by the distribution $\mathsf{P}_\mathbf{a}$. For every $i$ P J$d$K, let us now consider an alternative DLB instance that differs on the dynamical matrix only. Specifically:

$$\mathbf{a}'_j \quad \begin{cases} \mathbf{a}_j & \text{if } j \quad i \\ \rho & \text{if } j \quad i \text{ and } \mathbf{a}_j \quad \rho \quad \epsilon, \\ \rho \quad \epsilon & \text{if } j \quad i \text{ and } \mathbf{a}_j \quad \rho \end{cases} \qquad @ j \text{ P J} d \text{K}.$$

[13] For a vector $\mathbf{v}$ P $\mathsf{R}^d$, we denote with $\mathsf{sign} \mathsf{p} \mathbf{v} \mathsf{q}$ P t $1, 1 \mathsf{u}^d$ the vector of the signs of the components of $\mathbf{v}$. It is irrelevant how we convene to define the sign of 0.

By relative entropy identities (Lattimore & Szepesvári, 2020), let $\mathbf{A}$ diag$(\mathbf{a})$ and $\mathbf{A}'$ diag$(\mathbf{a}')$, we have:

$$D_{\mathrm{KL}}(\mathsf{P}_{\mathbf{a}}, \mathsf{P}_{\mathbf{a}'}) = \mathsf{E}_{\mathbf{a}}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(\mathcal{N}(\mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t, \sigma^2\mathbf{I}_d), \mathcal{N}(\mathbf{A}'\mathbf{x}_t + \mathbf{B}\mathbf{u}_t, \sigma^2\mathbf{I}_d)\right)\right]$$
$$= \frac{1}{2\sigma^2}\sum_{t=1}^{T}\mathsf{E}_{\mathbf{a}}\left[\left\|(\mathbf{A} - \mathbf{A}')\,\mathbf{x}_t\right\|_2^2\right] \leq \epsilon^2\mathsf{E}_{\mathbf{a}}\left[\mathbf{x}_{t,i}^2\right].$$

We proceed at properly bounding the KL-divergence, letting $\mathbf{e}_i$ be the $i$-th vector of the canonical basis of $\mathsf{R}^d$ and convening that $\mathbf{x}_0 = \mathbf{0}_d$:

$$\mathsf{E}_{\mathbf{a}}\left[\mathbf{x}_{t,i}^2\right] \leq \mathsf{E}_{\mathbf{a}}\left[\left(\sum_{s=1}^{t-1}\mathbf{e}_i^{\mathrm{T}}\mathbf{A}^s\mathbf{B}\mathbf{u}_{t-s} + \sum_{s=1}^{t-1}\mathbf{e}_i^{\mathrm{T}}\mathbf{A}^s\boldsymbol{\epsilon}_{t-s}\right)^2\right]$$
$$\leq \mathsf{E}_{\mathbf{a}}\left[\left(1 + \rho\sum_{s=1}^{t-1}\mathbf{a}_i^s\mathbf{u}_{t-s,i} + \sum_{s=1}^{t-1}\mathbf{a}_i^s\boldsymbol{\epsilon}_{t-s,i}\right)^2\right]$$
$$\leq \mathsf{E}_{\mathbf{a}}\Bigg[\underbrace{1 + \rho^2\sum_{s=1}^{t-1}\sum_{l=1}^{t-1}\mathbf{a}_i^{s+l}\mathbf{u}_{t-s,i}\mathbf{u}_{t-l,i}}_{\text{(a)}} \underbrace{+ 2 + 2\rho\sum_{s=1}^{t-1}\sum_{l=1}^{t-1}\mathbf{a}_i^{s+l}\mathbf{u}_{t-s,i}\boldsymbol{\epsilon}_{t-l,i}}_{\text{(b)}} \underbrace{+ \sum_{s=1}^{t-1}\sum_{l=1}^{t-1}\mathbf{a}_i^{s+l}\boldsymbol{\epsilon}_{t-s,i}\boldsymbol{\epsilon}_{t-l,i}}_{\text{(c)}}\Bigg]$$

Let us start with (a):

$$1 + \rho^2\mathsf{E}_{\mathbf{a}}\left[\sum_{s=1}^{t-1}\sum_{l=1}^{t-1}\mathbf{a}_i^{s+l}\mathbf{u}_{t-s,i}\mathbf{u}_{t-l,i}\right] \leq 1 + \rho^2\sum_{s=1}^{t-1}\sum_{l=1}^{t-1}\rho^{s+l} \leq 1,$$

having observed that $|\mathbf{u}_{t-s,i}|, |\mathbf{u}_{t-l,i}| \leq 1$, that $|\mathbf{a}_i| \leq \rho$, and bounding the summations with the series. Let us move to (b):

$$2\rho\mathsf{E}_{\mathbf{a}}\left[\sum_{s=1}^{t-1}\sum_{l=1}^{t-1}\mathbf{a}_i^{s+l}\mathbf{u}_{t-s,i}\boldsymbol{\epsilon}_{t-l,i}\right] = 2\rho\mathsf{E}_{\mathbf{a}}\left[\sum_{s=1}^{t-1}\sum_{l=s+1}^{t-1}\mathbf{a}_i^{s+l}\mathbf{u}_{t-s,i}\boldsymbol{\epsilon}_{t-l,i}\right]$$
$$= 2\rho\,\mathsf{E}_{\mathbf{a}}\left[\sum_{l=1}^{t-1}\sum_{s=l}^{t-1}\mathbf{a}_i^{s+l}\mathbf{u}_{t-s,i}\boldsymbol{\epsilon}_{t-l,i}\right] \to 0$$
$$\leq 2\rho\sum_{s=1}^{t-1}\sum_{l=s+1}^{t-1}\rho^{s+l}\mathsf{E}_{\mathbf{a}}\left[|\boldsymbol{\epsilon}_{t-l,i}|\right]$$
$$\leq \frac{\sigma}{1-\rho}\sqrt{\frac{2}{\pi}},$$

having observed that $\mathbf{u}_{t-s,i}$ and $\boldsymbol{\epsilon}_{t-l,i}$ are independent when $s \neq l$ and $\boldsymbol{\epsilon}_{t-l,i}$ has zero mean, that $|\mathbf{u}_{t-s,i}| \leq 1$, that $\mathbf{a}_i^{s+l} \leq \rho^{s+l}$, and that the expectation of the absolute value of random variable normally distributed is given by $\mathsf{E}\left[|\boldsymbol{\epsilon}_{t-l,i}|\right] = \sigma\sqrt{\frac{2}{\pi}}$. Finally, let us consider (c):

$$\mathsf{E}_{\mathbf{a}}\left[\sum_{s=1}^{t-1}\sum_{l=1}^{t-1}\mathbf{a}_i^{s+l}\boldsymbol{\epsilon}_{t-s,i}\boldsymbol{\epsilon}_{t-l,i}\right] = \mathsf{E}_{\mathbf{a}}\left[\sum_{s=1}^{t-1}\mathbf{a}_i^{2s}\boldsymbol{\epsilon}_{t-s,i}\boldsymbol{\epsilon}_{t-s,i}\right] + 2\,\mathsf{E}_{\mathbf{a}}\left[\sum_{s=1}^{t-2}\sum_{l=s+1}^{t-1}\mathbf{a}_i^{s+l}\boldsymbol{\epsilon}_{t-s,i}\boldsymbol{\epsilon}_{t-l,i}\right] \to 0$$
$$\leq \sigma^2\sum_{s=1}^{t-1}\rho^{2s} \leq \frac{\sigma^2}{1-\rho^2} \leq \frac{\sigma^2}{1-\rho},$$

having observed that the noise vectors $\boldsymbol{\epsilon}_{t-l,i}$ and $\boldsymbol{\epsilon}_{t-s,i}$ are independent whenever $s \neq l$, that $\mathsf{E}_{\mathbf{a}}\left[\boldsymbol{\epsilon}_{t-s,i}^2\right] = \sigma^2$, and having bounded the sum with the series. Coming back to the original bound, we have:

$$\mathsf{E}_{\mathbf{a}}\left[\mathbf{x}_{t,i}^2\right] \leq 1 + \frac{1}{1-\rho}\left(\sigma^2 + 2\sigma\sqrt{\frac{2}{\pi}}\right).$$

For $i \in [d]$ and $\mathbf{a} \in \mathcal{A}$, we introduce the symbol:

$$p_{\mathbf{a},i} \quad \mathbb{P}_{\mathbf{a}}\left(\sum_{t=1}^{T} \mathbf{1}\{\operatorname{sign}(\mathbf{u}_{t,i}) \neq \operatorname{sign}(\mathbf{h}_{\mathbf{a},i}[u])\} \geq \frac{T}{2}\right).$$

Thus, for $\mathbf{a}$ and $\mathbf{a}'$ defined as above, by the Bretagnolle-Huber inequality (Lattimore & Szepesvári, 2020, Theorem 14.2), we have:

$$\begin{aligned} p_{\mathbf{a},i} \quad p_{\mathbf{a}',i} &\geq \frac{1}{2}\exp(-) \quad D_{\mathrm{KL}}(\mathbb{P}_{\mathbf{a}}, \mathbb{P}_{\mathbf{a}'}) \quad \frac{1}{2}\exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\mathbb{E}_{\mathbb{P}}\left[\left\|\begin{pmatrix}\mathbf{A} & \mathbf{A}'\end{pmatrix}\mathbf{x}_t\right\|_2^2\right]\right) \\ &\geq \frac{1}{2}\exp\left(-\frac{T\epsilon^2}{2}\left(\frac{1}{\sigma^2} \quad \frac{1}{1-\rho}\left(1-\frac{2}{\sigma}\sqrt{\frac{2}{\pi}}\right)\right)\right) \\ &\geq \frac{1}{2}\exp\left(-\frac{2T\epsilon^2}{1-\rho}\right), \end{aligned}$$

having selected $\sigma^2 \quad 1$. We use the notation $\sum_{\mathbf{a}_{-i}}$ to denote the multiple summation $\sum_{\mathbf{a}_1,\ldots,\mathbf{a}_{i-1},\mathbf{a}_{i+1},\ldots,\mathbf{a}_d \in \{\rho,\rho-\epsilon\}^{d-1}}$:

$$\begin{aligned} \sum_{\mathbf{a}\in A} 2^{-d}\sum_{i=1}^{d} p_{\mathbf{a},i} \quad \sum_{i=1}^{d}\sum_{\mathbf{a}_{-i}} 2^{-d} \sum_{\mathbf{a}_i \in \{\rho,\rho-\epsilon\}} p_{\mathbf{a},i} \\ \geq \sum_{i=1}^{d}\sum_{\mathbf{a}_{-i}} 2^{-d} \quad \frac{1}{2}\exp\left(-\frac{2T\epsilon^2}{1-\rho}\right) \\ \frac{d}{4}\exp\left(-\frac{2T\epsilon^2}{1-\rho}\right). \end{aligned}$$

Therefore, with this averaging argument, we can conclude that there exists $\mathbf{a}^* \in \mathcal{A}$ such that $\sum_{i=1}^{d} p_{\mathbf{a}^*,i} \geq \frac{d}{4}\exp\left(-\frac{2T\epsilon^2}{1-\rho}\right)$. For this choice $\mathbf{a}^*$, we consider $\mathbf{u}^*_{\mathbf{a}^*} \quad \operatorname{sign}(\mathbf{a}^*) \in \mathcal{U}$, we can proceed to the lower bound on the expected offline regret:

$$\begin{aligned} \mathbb{E}R^{\mathrm{off}}(\underline{\boldsymbol{\pi}}, T) \quad & \sum_{t=1}^{T} \mathbb{E}_{\mathbf{a}^*}\langle \mathbf{h}_{\mathbf{a}^*}, \mathbf{u}^*_{\mathbf{a}^*} \quad \mathbf{u}_t\rangle \\ & \sum_{t=1}^{T}\mathbb{E}_{\mathbf{a}^*}\left[\sum_{i=1}^{d}\mathbf{1}\{\operatorname{sign}(\mathbf{u}_{t,i}) \quad \operatorname{sign}(\mathbf{h}_{\mathbf{a}^*,i})[u]\}\frac{\epsilon}{1 \quad (\rho \quad \epsilon)}\right] \\ & \frac{\epsilon}{1 \quad (\rho \quad \epsilon)}\sum_{t=1}^{T}\sum_{i=1}^{d}\mathbb{P}_{\mathbf{a}^*}\left(\operatorname{sign}(\mathbf{u}_{t,i}) \quad \operatorname{sign}(\mathbf{h}_{\mathbf{a}^*,i})\right) \\ \geq & \frac{T\epsilon}{2(1 \quad (\rho \quad \epsilon))}\sum_{i=1}^{d}\mathbb{P}_{\mathbf{a}^*}\left(\sum_{t=1}^{T}\mathbf{1}\{\operatorname{sign}(\mathbf{u}_{t,i}) \quad \operatorname{sign}(\mathbf{h}_{\mathbf{a}^*,i})[u] \geq \frac{T}{2}\right) \\ & \frac{T\epsilon}{2(1 \quad (\rho \quad \epsilon))}\sum_{i=1}^{d} p_{\mathbf{a}^*,i} \geq \frac{Td\epsilon}{8(1 \quad (\rho \quad \epsilon))}\exp\left(-\frac{2T\epsilon^2}{1-\rho}\right). \end{aligned}$$

We now maximize over $0 < \epsilon < \rho$. To this end, we perform the substitution $\epsilon \quad \frac{(1-\rho)\widetilde{\epsilon}}{1-\widetilde{\epsilon}}$, with $0 < \widetilde{\epsilon} < \rho$:

$$\frac{Td\epsilon}{8(1 \quad (\rho \quad \epsilon))}\exp\left(-\frac{2T\epsilon^2}{1-\rho}\right) \quad \frac{Td\widetilde{\epsilon}}{8}\exp\left(-\frac{2\widetilde{\epsilon}^2 T(1 \quad \rho)}{(1 \quad \widetilde{\epsilon})^2}\right) \geq \frac{Td\widetilde{\epsilon}}{8}\exp\left(-8\widetilde{\epsilon}^2 T(1 \quad \rho)\right),$$

where the last inequality holds for $\widetilde{\epsilon} < \frac{1}{2}$. We not take $\widetilde{\epsilon} \quad \frac{1}{\sqrt{8T(1-\rho)}}$ which is smaller than $\frac{1}{2}$ if $T \geq \frac{1}{2(1-\rho)}$, to get:

$$\mathbb{E}R^{\mathrm{off}}(\underline{\boldsymbol{\pi}}, T) \geq \frac{d\sqrt{T}}{\sqrt{512e(1 \quad \rho)}}.$$

Notice that with this choice of $\widetilde{\epsilon}$ (and, consequently, of $\epsilon$), for sufficiently large $T$, we fulfill Assumption 2.2. Indeed:

$$\boldsymbol{\theta} \quad 1 \quad \frac{1}{\sqrt{32T(1 \quad \rho)}}, \quad J^*_{\mathbf{a}} \quad \frac{d}{\sqrt{32T(1 \quad \rho)}}.$$

Thus, we require $T \geq \mathcal{O}\left(\frac{d^2}{1-\overline{\rho}}\right)$. Finally, to convert this result to the expected regret, we employ Lemma B.1:

$$\mathbb{E}R^{\text{off}}(\underline{\boldsymbol{\pi}}, T) \geq \mathbb{E}R^{\text{off}}(\underline{\boldsymbol{\pi}}, T) \quad \frac{d}{1 \quad \rho}.$$

Under the constraint $T \geq \mathcal{O}\left(\frac{d^2}{1-\overline{\rho}}\right)$, we observe that:

$$\mathbb{E}R^{\text{off}}(\underline{\boldsymbol{\pi}}, T) \geq \quad \left(\frac{d \quad ? \quad \overline{T}}{(1 \quad \rho)^{\frac{1}{2}}}\right).$$

□

**Theorem 2.1** (Optimal Policy)**.** *Under Assumptions 2.1 and 2.2, an optimal policy $\underline{\boldsymbol{\pi}}^*$ maximizing the (infinite-horizon) expected average reward $J(\underline{\boldsymbol{\pi}})$ (Equation 2), for every round $t \in \mathbb{N}$ and history $H_{t-1} \in \mathcal{H}_{t-1}$ is given by:*

$$\boldsymbol{\pi}_t^*(H_{t-1}) \quad \mathbf{u}^* \quad \textit{where} \quad \mathbf{u}^* \in \operatorname*{argmax}_{\mathbf{u} \in U} J(\mathbf{u}) \quad \langle \mathbf{h}, \mathbf{u} \rangle. \tag{4}$$

*Proof.* Referring to the notation of Appendix C, we first observe that for every policy $\underline{\boldsymbol{\pi}}$, we have $J(\underline{\boldsymbol{\pi}}) \quad \liminf_{H \to +\infty} J_H(\underline{\boldsymbol{\pi}})$, where $J_H(\underline{\boldsymbol{\pi}}) \quad \frac{1}{H}\mathbb{E}[\sum_{t=1}^{H} y_t]$, is the $H$-horizon expected average reward. Let us start with Equation (18), a fixed finite $H \in \mathbb{N}$, and considering the sequence of actions $(\mathbf{u}_1, \mathbf{u}_2, \dots)$ generated by policy $\underline{\boldsymbol{\pi}}$:

$$\begin{aligned} J_H(\underline{\boldsymbol{\pi}}) \quad & \frac{1}{H}\sum_{s=1}^{H} \langle \mathbf{h}^{(0, H-s)}, \mathbb{E}[\mathbf{u}_s] \rangle \quad \frac{1}{H}\sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1}\,\mathbb{E}[\mathbf{x}_1] \\ & \frac{1}{H}\sum_{s=1}^{H} \langle \mathbf{h}, \mathbb{E}[\mathbf{u}_s] \rangle \quad \frac{1}{H}\sum_{s=1}^{H} \langle \mathbf{h}^{(H-s+1, +\infty)}, \mathbb{E}[\mathbf{u}_s] \rangle \quad \frac{1}{H}\sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1}\,\mathbb{E}[\mathbf{x}_1]. \end{aligned}$$

Now, we consider two bounds on $J_H(\underline{\boldsymbol{\pi}})$, obtained by an application of Cauchy-Schwarz inequality on the second addendum:

$$\begin{aligned} J_H(\underline{\boldsymbol{\pi}}) \leq \quad & \frac{1}{H}\sum_{s=1}^{H} \langle \mathbf{h}, \mathbb{E}[\mathbf{u}_s] \rangle \quad \frac{1}{H}\sum_{s=1}^{H} \left\| \mathbf{h}^{(H-s+1, +\infty)} \right\|_2 \|\mathbb{E}[\mathbf{u}_s]\|_2 \\ & \frac{1}{H}\sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1}\,\mathbb{E}[\mathbf{x}_1] \quad : J_H^{\uparrow}(\underline{\boldsymbol{\pi}}), \\ J_H(\underline{\boldsymbol{\pi}}) \geq \quad & \frac{1}{H}\sum_{s=1}^{H} \langle \mathbf{h}, \mathbb{E}[\mathbf{u}_s] \rangle \quad \frac{1}{H}\sum_{s=1}^{H} \left\| \mathbf{h}^{(H-s+1, +\infty)} \right\|_2 \|\mathbb{E}[\mathbf{u}_s]\|_2 \\ & \frac{1}{H}\sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1}\,\mathbb{E}[\mathbf{x}_1] \quad : J_H^{\downarrow}(\underline{\boldsymbol{\pi}}). \end{aligned}$$

Concerning the term $\|\mathbb{E}[\mathbf{u}_s]\|_2$, we have that $\|\mathbb{E}[\mathbf{u}_s]\|_2 \leq \mathbb{E}[\|\mathbf{u}_s\|_2] \leq U$, having used Jensen's inequality and under Assumption 2.2. Regarding the second term, using Assumptions 2.1 and 2.2, we obtain:

$$\begin{aligned} \left\| \mathbf{h}^{(H-s+1, +\infty)} \right\|_2 \quad & \left\| \sum_{l=H-s+1}^{+\infty} \mathbf{B}^\top (\mathbf{A}^{l-1})^\top \boldsymbol{\omega} \right\|_2 \\ & \leq B \quad \sum_{l=H-s+1}^{+\infty} \quad \|\mathbf{A}\|\rho(\mathbf{A})^{l-1} \\ & B \quad \|\mathbf{A}\|\frac{\rho(\mathbf{A})^{H-s}}{1 \quad \rho(\mathbf{A})}. \end{aligned} \tag{11}$$

Plugging this result into the summation over $s$, we obtain:

$$\frac{1}{H} \quad \frac{B \quad \|\mathbf{A}\|}{1 \quad \rho(\mathbf{A})} \sum_{s=1}^{H} \rho(\mathbf{A})^{H-s} \quad \frac{B \quad \|\mathbf{A}\|(1 \quad \rho(\mathbf{A})^H)}{H(1 \quad \rho(\mathbf{A}))^2}.$$

It is simple to observe that the last term approaches zero as $H \to \infty$. Moreover, with an analogous argument, it can be proved that $\left\| \frac{1}{H}\sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1}\,\mathbb{E}[\mathbf{x}_1] \right\|_2 \to 0$ as $H \to \infty$. Thus, we have that $\liminf_{H \to +\infty} J_H^{\downarrow}(\underline{\boldsymbol{\pi}})$

$\liminf_{H\to+\infty} J_H^{\uparrow}(\underline{\pi})$. Consequently, by the squeezing theorem of limits, we have:

$$J(\underline{\pi}) \quad \liminf_{H\to+\infty} J_H^{\uparrow}(\underline{\pi}) \quad \liminf_{H\to+\infty} J_H^{\downarrow}(\underline{\pi})$$
$$\liminf_{H\to+\infty} \frac{1}{H}\sum_{s=1}^{H}\langle \mathbf{h}, \mathbb{E}[\mathbf{u}_s]\rangle \quad \mathbf{h}^{\top}\left(\liminf_{H\to+\infty}\frac{1}{H}\sum_{s=1}^{H}\mathbb{E}[\mathbf{u}_s]\right).$$

It follows that an optimal policy is a policy that plays the constant action $\mathbf{u}^* = \arg\max_{\mathbf{u}\in U}\langle \mathbf{h}, \mathbf{u}\rangle$. □

### B.2. Proofs of Section 3

**Theorem 3.1** (Self-Normalized Concentration)**.** *Let $(\widehat{\mathbf{h}}_t)_{t\in\mathbb{N}}$ be the sequence of solutions of the Ridge regression problems of Algorithm 1. Then, under Assumption 2.1 and 2.2, for every $\lambda > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all rounds $t \in \mathbb{N}$, it holds that:*

$$\left\|\widehat{\mathbf{h}}_t - \mathbf{h}\right\|_{\mathbf{V}_t} \le \frac{c_1}{\sqrt{\lambda}}\log(et) + c_2\sqrt{\lambda}\sqrt{2\widetilde{\sigma}^2\left(\log\left(\frac{1}{\delta}\right) + \frac{1}{2}\log\left(\frac{\det(\mathbf{V}_t)}{\lambda^d}\right)\right)},$$

*where $c_1 = U = (\mathbf{A})\left(\frac{UB}{1-\rho(\mathbf{A})} + X\right)$, $c_2 = \frac{\Omega B\Phi(\mathbf{A})}{1-\rho(\mathbf{A})}$, and $\widetilde{\sigma}^2 = \sigma^2\left(1 + \frac{\Omega^2\Phi(\mathbf{A})^2}{1-\rho(\mathbf{A})^2}\right)$.*

*Proof.* First of all, let us properly relate the round $t \in [T]$ and the index of the epoch $m \in [M]$. For every epoch $m \in [M]$, we denote with $t_m$ the last round of epoch $m$ (i.e., the one in which we update the relevant matrices $\mathbf{V}_t$ and $\mathbf{b}_t$):[14]

$$t_0 = 0, \qquad t_m = t_{m-1} + H_m.$$

We now proceed to define suitable filtrations. Let $\mathcal{F} = (\mathcal{F}_t)_{t\in[T]}$ such that for every $t \ge 1$, the random variables $(\mathbf{u}_1, y_1, \ldots, \mathbf{u}_{t-1}, y_{t-1}, \mathbf{u}_t)$ are $\mathcal{F}_{t-1}$-measurable, i.e., $\mathcal{F}_{t-1} = \sigma(\mathbf{u}_1, y_1, \ldots, \mathbf{u}_{t-1}, y_{t-1}, \mathbf{u}_t)$. Let us also consider the filtration indexed by $m$, denoted with $\widetilde{\mathcal{F}} = (\widetilde{\mathcal{F}}_m)_{m\in[M]}$ and defined for all $m \in [M]$ as $\widetilde{\mathcal{F}}_m = \mathcal{F}_{t_{m+1}-1}$. Thus, the random variables $\widetilde{\mathcal{F}}_{m-1}$-measurable are those realized until the end of epoch $m$ except for $y_{t_m}$.

Since the estimates $\widehat{\mathbf{h}}_t$ do not change within an epoch, we need to guarantee the statement for all rounds $(t_m)_{m\in[M]}$ only. For these rounds, we define the following quantities:

$$\widetilde{y}_m = y_{t_m},$$
$$\widetilde{\mathbf{u}}_m = \mathbf{u}_{t_m}, \quad \text{(or any } \mathbf{u}_l \text{ with } l \in [t_{m-1}+1, t_m] \text{ since they are all equal)}$$
$$\widetilde{\xi}_m = \eta_{t_m} + \sum_{s=1}^{H_m+1}\boldsymbol{\omega}^{\top}\mathbf{A}^{s-1}\boldsymbol{\epsilon}_{t_m-s},$$
$$\widetilde{\mathbf{x}}_{m-1} = \mathbf{x}_{t_{m-1}},$$
$$\widetilde{\mathbf{h}}_m = \widehat{\mathbf{h}}_{t_m},$$
$$\widetilde{\mathbf{V}}_m = \mathbf{V}_{t_m},$$
$$\widetilde{\mathbf{b}}_m = \mathbf{b}_{t_m}.$$

We prove that $(\widetilde{\xi}_m)_{m\in[M]}$ is a martingale difference process adapted to the filtration $\widetilde{\mathcal{F}}$. To this end, we recall that, by construction, $(\eta_t)_{t\in[T]}$ and $(\boldsymbol{\epsilon}_t)_{t\in[T]}$ are martingale difference processes adapted to the filtration $\mathcal{F}$. It is clear that $\widetilde{\xi}_m$ is $\mathcal{F}_m$-measurable and, being $\sigma^2$-subgaussian it is absolutely integrable. Furthermore, using the tower law of expectation:

$$\mathbb{E}\left[\widetilde{\xi}_m|\widetilde{\mathcal{F}}_{m-1}\right] = \mathbb{E}\left[\eta_{t_m} + \sum_{s=1}^{H_m+1}\boldsymbol{\omega}^{\top}\mathbf{A}^{s-1}\boldsymbol{\epsilon}_{t_m-s}|\mathcal{F}_{t_m-1}\right]$$
$$= \mathbb{E}[\eta_{t_m}|\mathcal{F}_{t_m-1}] + \mathbb{E}\left[\sum_{s=1}^{H_m+1}\boldsymbol{\omega}^{\top}\mathbf{A}^{s-1}\,\mathbb{E}[\boldsymbol{\epsilon}_{t_m-s}|\mathcal{F}_{t_m-s-1}]|\mathcal{F}_{t_m-1}\right] = 0,$$

since the system is operating by persisting the action after having decided it at the beginning of the epoch. Thus, by

[14] It is worth noting that the variables $t_m$ are deterministic.

exploiting the decomposition in Equation (1), we can write:

$$\begin{aligned}\widetilde{y}_m \quad y_{t_m} \quad & \mathsf{x}\mathbf{h}^{\mathsf{J}0,H_m+1\mathsf{K}}, \widetilde{\mathbf{u}}_m\mathsf{y} \quad \boldsymbol{\omega}^\top \mathbf{A}^{H_m+1}\mathbf{x}_{t_{m-1}} \quad \eta_{t_m} \quad \sum_{s=1}^{H_m+1} \boldsymbol{\omega}^\top \mathbf{A}^{s-1}\boldsymbol{\epsilon}_{t_m-s} \\ & \mathsf{x}\mathbf{h}^{\mathsf{J}0,H_m+1\mathsf{K}}, \widetilde{\mathbf{u}}_m\mathsf{y} \quad \boldsymbol{\omega}^\top \mathbf{A}^{H_m+1}\widetilde{\mathbf{x}}_{m-1} \quad \widetilde{\xi}_m \\ & \mathsf{x}\mathbf{h}, \widetilde{\mathbf{u}}_m\mathsf{y} \quad \mathsf{x}\mathbf{h}^{\mathsf{J}H_m+2,\infty\mathsf{M}}, \widetilde{\mathbf{u}}_m\mathsf{y} \quad \boldsymbol{\omega}^\top \mathbf{A}^{H_m+1}\widetilde{\mathbf{x}}_{m-1} \quad \widetilde{\xi}_m, \end{aligned} \tag{12}$$

where we simply exploit the identity $\mathbf{h} \quad \mathbf{h}^{\mathsf{J}0,H_m+1\mathsf{K}} \quad \mathbf{h}^{\mathsf{J}H_m+2,\infty\mathsf{M}}$. We now introduce the following vectors and matrices:

$$\widetilde{\mathbf{U}}_m \quad \begin{pmatrix} \widetilde{\mathbf{u}}_1^\top \\ \vdots \\ \widetilde{\mathbf{u}}_m^\top \end{pmatrix} \mathsf{P}\, \mathsf{R}^{m\times d}, \qquad \widetilde{\mathbf{y}}_m \quad \begin{pmatrix} \widetilde{y}_1 \\ \vdots \\ \widetilde{y}_m \end{pmatrix} \mathsf{P}\, \mathsf{R}^m,$$

$$\widetilde{\boldsymbol{\xi}}_m \quad \begin{pmatrix} \widetilde{\xi}_1 \\ \vdots \\ \widetilde{\xi}_m \end{pmatrix} \mathsf{P}\, \mathsf{R}^m, \qquad \widetilde{\boldsymbol{\nu}}_m \quad \begin{pmatrix} \boldsymbol{\omega}^\top \mathbf{A}^{H_1+2}\widetilde{\mathbf{x}}_0 \\ \vdots \\ \boldsymbol{\omega}^\top \mathbf{A}^{H_m+2}\widetilde{\mathbf{x}}_{m-1} \end{pmatrix} \mathsf{P}\, \mathsf{R}^m,$$

$$\widetilde{\mathbf{g}}_m \quad \begin{pmatrix} \mathsf{x}\mathbf{h}^{\mathsf{J}H_1+1,\infty\mathsf{M}}, \widetilde{\mathbf{u}}_1\mathsf{y} \\ \vdots \\ \mathsf{x}\mathbf{h}^{\mathsf{J}H_m+1,\infty\mathsf{M}}, \widetilde{\mathbf{u}}_m\mathsf{y} \end{pmatrix} \mathsf{P}\, \mathsf{R}^m.$$

Using the vectors and matrices above, we observe that $\widetilde{\mathbf{V}}_m \quad \lambda\mathbf{I} \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{U}}_m$ and $\widetilde{\mathbf{b}}_m \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{y}}_m$. Furthermore, by exploiting Equation (12), we can write:

$$\widetilde{\mathbf{y}}_m \quad \widetilde{\mathbf{U}}_m\mathbf{h} \quad \widetilde{\mathbf{g}}_m \quad \widetilde{\boldsymbol{\nu}}_m \quad \widetilde{\boldsymbol{\xi}}_m.$$

Let us consider the estimate at $m \mathsf{P} \mathsf{J}MK$:

$$\begin{aligned}\widetilde{\mathbf{h}}_m \quad \widetilde{\mathbf{V}}_m^{-1}\widetilde{\mathbf{b}}_m \quad & \left(\lambda\mathbf{I} \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{U}}_m\right)^{-1}\widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{y}}_m \\ & \left(\lambda\mathbf{I} \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{U}}_m\right)^{-1}\widetilde{\mathbf{U}}_m^\top\left(\widetilde{\mathbf{U}}_m\mathbf{h} \quad \widetilde{\mathbf{g}}_m \quad \widetilde{\boldsymbol{\nu}}_m \quad \widetilde{\boldsymbol{\xi}}_m\right) \\ & \mathbf{h} \quad \left(\lambda\mathbf{I} \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{U}}_m\right)^{-1}\left( \quad \lambda\mathbf{h} \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{g}}_m \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\boldsymbol{\nu}}_m \quad \widetilde{\mathbf{U}}_m^\top\widetilde{\boldsymbol{\xi}}_m\right).\end{aligned}$$

We now proceed at bounding the $\}\ \}_{\widetilde{\mathbf{V}}_m}$-norm, and exploit the triangle inequality:

$$\begin{aligned}\left\|\widetilde{\mathbf{h}}_m \quad \mathbf{h}\right\|_{\widetilde{\mathbf{V}}_m} \ \mathsf{¤}\ & \lambda\left\|\widetilde{\mathbf{V}}_m^{-1}\mathbf{h}\right\|_{\widetilde{\mathbf{V}}_m} \quad \left\|\widetilde{\mathbf{V}}_m^{-1}\widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{g}}_m\right\|_{\widetilde{\mathbf{V}}_m} \quad \left\|\widetilde{\mathbf{V}}_m^{-1}\widetilde{\mathbf{U}}_m^\top\widetilde{\boldsymbol{\nu}}_m\right\|_{\widetilde{\mathbf{V}}_m} \quad \left\|\widetilde{\mathbf{V}}_m^{-1}\widetilde{\mathbf{U}}_m^\top\widetilde{\boldsymbol{\xi}}_m\right\|_{\widetilde{\mathbf{V}}_m} \\ & \underbrace{\lambda\,\}\mathbf{h}\}_{\widetilde{\mathbf{V}}_m^{-1}}}_{\text{(a)}} \quad \underbrace{\left\|\widetilde{\mathbf{U}}_m^\top\widetilde{\mathbf{g}}_m\right\|_{\widetilde{\mathbf{V}}_m^{-1}}}_{\text{(b)}} \quad \underbrace{\left\|\widetilde{\mathbf{U}}_m^\top\widetilde{\boldsymbol{\nu}}_m\right\|_{\widetilde{\mathbf{V}}_m^{-1}}}_{\text{(c)}} \quad \underbrace{\left\|\widetilde{\mathbf{U}}_m^\top\widetilde{\boldsymbol{\xi}}_m\right\|_{\widetilde{\mathbf{V}}_m^{-1}}}_{\text{(d)}},\end{aligned}$$

where we simply exploited the identity $\}\mathbf{V}^{-1}\mathbf{x}\}_{\mathbf{V}}^2 \quad \mathbf{x}^\top\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{x} \quad \mathbf{x}^\top\mathbf{V}^{-1}\mathbf{x} \quad \}\mathbf{x}\}_{\mathbf{V}^{-1}}^2$. We now bound one term at a time. Let us start with (a):

$$\begin{aligned}\text{(a)}^2 \quad \lambda^2\,\}\mathbf{h}\}_{\widetilde{\mathbf{V}}_m^{-1}}^2 \quad & \lambda^2\mathbf{h}^\top\widetilde{\mathbf{V}}_m^{-1}\mathbf{h} \\ & \mathsf{¤}\ \lambda^2\left\|\widetilde{\mathbf{V}}_m^{-1}\right\|_2\}\mathbf{h}\}_2^2 \\ & \mathsf{¤}\ \lambda\,\}\mathbf{h}\}_2^2 \\ & \mathsf{¤}\ \lambda\left( \quad \frac{B \quad \mathsf{p}\mathbf{A}\mathsf{q}}{1 \quad \rho\mathsf{p}\mathbf{A}\mathsf{q}}\right)^2,\end{aligned}$$

where we observed that $\left\|\widetilde{\mathbf{V}}_m^{-1}\right\|_2 \ \mathsf{¤}\ \left\|\widetilde{\mathbf{V}}_m\right\|_2^{-1} \ \mathsf{¤}\ \lambda^{-1}$. Finally, we have bounded the norm of $\mathbf{h}$:

$$\}\mathbf{h}\}_2 \quad \left\|\sum_{s=0}^{+\infty}\mathbf{h}^{\{s\}}\right\|_2$$

$$\begin{aligned}
&\leq \sum_{s=0}^{+\infty} \left\| \mathbf{h}^{\{s\}} \right\|_2 \\
&\leq \|\theta\|_2 \quad \|\omega\|_2 \|\mathbf{B}\|_2 \sum_{s=1}^{+\infty} \|\mathbf{A}\|^{s-1} \\
&\leq \frac{B \quad \|\mathbf{A}\|}{1 \quad \rho\|\mathbf{A}\|},
\end{aligned}$$

where we have exploited Assumptions 2.1 and 2.2.

We now move to term (b):

$$\begin{aligned}
(\mathrm{b})^2 \quad \left\| \widetilde{\mathbf{U}}_m^\top \widetilde{\mathbf{g}}_m \right\|^2_{\widetilde{\mathbf{V}}_m^{-1}} \quad & \widetilde{\mathbf{g}}_m^\top \widetilde{\mathbf{U}}_m \widetilde{\mathbf{V}}_m^{-1} \widetilde{\mathbf{U}}_m^\top \widetilde{\mathbf{g}}_m \\
&\leq \frac{1}{\lambda} \left\| \widetilde{\mathbf{g}}_m^\top \widetilde{\mathbf{U}}_m \right\|_2^2 \\
&\quad \frac{1}{\lambda} \left\| \sum_{l=1}^{m} \mathbf{x}\widetilde{\mathbf{u}}_l, \mathbf{h}^{\mathbf{J}H_l+2,\infty\mathbf{M}} \mathbf{y}\widetilde{\mathbf{u}}_l \right\|_2^2 \\
&\leq \frac{1}{\lambda} \left( \sum_{l=1}^{m} \|\widetilde{\mathbf{u}}_l\|_2^2 \left\| \mathbf{h}^{\mathbf{J}H_l+2,\infty\mathbf{M}} \right\|_2 \right)^2 \\
&\leq \frac{U^4 \quad {}^2B^2 \quad \|\mathbf{A}\|^2}{\lambda\|\mathbf{1} \quad \rho\|\mathbf{A}\|\|^2} \quad \left( \sum_{l=1}^{m} \rho\|\mathbf{A}\|^{H_l+1} \right)^2,
\end{aligned}$$

where we have employed the following inequality:

$$\begin{aligned}
\left\| \mathbf{h}^{\mathbf{J}H_l+2,\infty\mathbf{M}} \right\|_2 \quad & \left\| \omega^\top \sum_{j=H_l+2}^{+\infty} \mathbf{A}^{j-1}\mathbf{B} \right\|_2 \\
&\leq \|\omega\|_2 \|\mathbf{B}\|_2 \sum_{j=H_l+2}^{+\infty} \left\|\mathbf{A}^{j-1}\right\|_2 \\
&\leq B \quad \|\mathbf{A}\| \frac{\rho\|\mathbf{A}\|^{H_l+1}}{1 \quad \rho\|\mathbf{A}\|}.
\end{aligned}$$

Let us now consider term (c):

$$\begin{aligned}
(\mathrm{c})^2 \quad \left\| \widetilde{\mathbf{U}}_m^\top \widetilde{\boldsymbol{\nu}}_m \right\|^2_{\widetilde{\mathbf{V}}_m^{-1}} \quad & \widetilde{\boldsymbol{\nu}}_m^\top \widetilde{\mathbf{U}}_m \widetilde{\mathbf{V}}_m^{-1} \widetilde{\mathbf{U}}_m^\top \widetilde{\boldsymbol{\nu}}_m \\
&\leq \frac{1}{\lambda} \left\| \widetilde{\mathbf{U}}_m^\top \widetilde{\boldsymbol{\nu}}_m \right\|_2^2 \\
&\quad \frac{1}{\lambda} \left\| \sum_{s=1}^{m} \omega^\top \mathbf{A}^{H_l+1} \widetilde{\mathbf{x}}_{l-1} \widetilde{\mathbf{u}}_l \right\|_2^2 \\
&\leq \frac{1}{\lambda} \left( \sum_{s=1}^{m} \|\omega\|_2 \left\|\mathbf{A}^{H_l+1}\right\|_2 \|\widetilde{\mathbf{x}}_{l-1}\|_2 \|\widetilde{\mathbf{u}}_l\|_2 \right)^2 \\
&\leq \frac{X^2 \quad {}^2U^2 \quad \|\mathbf{A}\|^2}{\lambda} \quad \left( \sum_{l=1}^{m} \rho\|\mathbf{A}\|^{H_l+1} \right)^2.
\end{aligned}$$

We now bound the summations, exploiting the inequality $\rho\|\mathbf{A}\| \leq \overline{\rho}$, holding by assumption:

$$\begin{aligned}
\sum_{l=1}^{m} \rho\|\mathbf{A}\|^{H_l+1} \quad & \sum_{l=1}^{m} \rho\|\mathbf{A}\|^{\left\lfloor \frac{\log l}{\log \frac{1}{\overline{\rho}}} \right\rfloor + 1} \\
&\leq \sum_{l=1}^{m} \rho\|\mathbf{A}\|^{\frac{\log l}{\log \frac{1}{\overline{\rho}}}}
\end{aligned}$$

$$\sum_{l=1}^{m} \exp\left( \frac{\log \frac{1}{\rho(\mathbf{A})}}{\log \frac{1}{\rho}} \log l \right)$$
$$\sum_{l=1}^{m} \frac{1}{l} \leq \log(m) + 1 \leq 1 + \log(t) + 1 \leq \log(et) + 1,$$

having exploited the fact that $m \leq t$ and the bound with the integral to the harmonic sum.

Finally, we consider term (d). In this case, we apply Theorem 1 of (Abbasi-Yadkori et al., 2011), observing that the conditions are satisfied. To this end, we first need to determine the subgaussianity constant for the noise process $\widetilde{\xi}_l$. For every $l \in [m]$ and $\zeta \in \mathbb{R}$, and properly using the tower law of expectation:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\zeta\widetilde{\xi}_l\right)|\widetilde{\mathcal{F}}_{l-1}\right] &\leq \mathbb{E}\left[\exp\left(\zeta\eta_{t_l} + \zeta\sum_{s=1}^{H_m+1}\boldsymbol{\omega}^\top\mathbf{A}^{s-1}\boldsymbol{\epsilon}_{t_l-s}\right)|\mathcal{F}_{t_l-1}\right] \\
&= \mathbb{E}\exp(\zeta\eta_{t_l}|\mathcal{F}_{t_l-1}) \prod_{s=1}^{H_m+1}\mathbb{E}\left[\mathbb{E}\left[\exp\left(\zeta\boldsymbol{\omega}^\top\mathbf{A}^{s-1}\boldsymbol{\epsilon}_{t_l-s}\right)|\mathcal{F}_{t_l-1-s}\right]|\mathcal{F}_{t_l-1}\right] \\
&\leq \exp\left(\frac{\zeta^2\sigma^2}{2}\right)\prod_{s=1}^{H_m+1}\mathbb{E}\left[\exp\left(\frac{\zeta^2\{\boldsymbol{\omega}^\top\mathbf{A}^{s-1}\}_2^2\sigma^2}{2}\right)|\mathcal{F}_{t_l-1}\right] \\
&\leq \exp\left(\frac{\zeta^2\sigma^2}{2}\right)\prod_{s=1}^{H_m+1}\exp\left(\frac{\zeta^2\|\mathbf{A}\|^2\rho(\mathbf{A})^{2(s-1)}\sigma^2}{2}\right) \\
&\leq \exp\left(\frac{\sigma^2\zeta^2}{2}\left(1 + \|\mathbf{A}\|^2\sum_{s=1}^{+\infty}\rho(\mathbf{A})^{2(s-1)}\right)\right) \\
&\leq \exp\left(\frac{\sigma^2\zeta^2}{2}\left(1 + \frac{\|\mathbf{A}\|^2}{1-\rho(\mathbf{A})^2}\right)\right).
\end{aligned}$$

Thus, simultaneously for all $m \in [M]$, with probability at least $1-\delta$, it holds that:

$$\text{(d)}^2 \quad \left\|\widetilde{\mathbf{U}}_m^\top\widetilde{\boldsymbol{\xi}}_m\right\|^2_{\widetilde{\mathbf{V}}_m^{-1}} \leq 2\sigma^2\left(1 + \frac{\|\mathbf{A}\|^2}{1-\rho(\mathbf{A})^2}\right)\left(\log\left(\frac{1}{\delta}\right) + \frac{1}{2}\log\left(\frac{\det\left(\widetilde{\mathbf{V}}_m\right)}{\lambda^d}\right)\right).$$

□

We now proceed at bounding the offline regret $R^{\text{off}}$ and, then, relating the offline regret $R^{\text{off}}$ with the online regret $R$, as defined in the main paper.

**Theorem B.2** (Offline Regret Upper Bound)**.** *Under Assumptions 2.1 and 2.2, having selected $\beta_t$ as in Equation (6), for every $\delta \in (0, 1]$, with probability at least $1-\delta$, `DynLin-UCB` suffers an offline regret $R^{off}$ bounded as:*

$$R^{off}(\underline{\boldsymbol{\pi}}^{\texttt{DynLin-UCB}}, T) \leq \sqrt{8dT\beta_{T-1}^2\left(1 + \frac{\log T}{\log\frac{1}{\overline{\rho}}}\right)\log\left(1 + \frac{TU^2}{d\lambda}\right)}.$$

*Moreover, by setting $\delta = 1/T$, highlighting the dependencies on $T$, $\overline{\rho}$, $d$, and $\sigma$ only, the expected offline regret $\mathbb{E}\,R^{off}$ is bounded as:*

$$\mathbb{E}\,R^{off}(\underline{\boldsymbol{\pi}}^{\texttt{DynLin-UCB}}, T) \leq \mathcal{O}\left(d\sigma\sqrt{\overline{T}}\log T^{\frac{3}{2}} + \sqrt{dT}\log T^2\left(1 - \overline{\rho}^{\frac{3}{2}}\right)\right).$$

*Proof.* For every epoch $m \in [M]$, let us define $\widetilde{\beta}_{m-1} = \beta_{t_{m-1}}$ and define the confidence set $\mathcal{C}_{m-1} = \{\widetilde{\mathbf{h}} \in \mathbb{R}^d : \|\widetilde{\mathbf{h}} - \widetilde{\mathbf{h}}_{m-1}\|_{\widetilde{\mathbf{V}}_{m-1}} \leq \widetilde{\beta}_{m-1}\}$. Let us start by considering the instantaneous offline regret $\widetilde{r}_m$ at epoch $m \in [M]$. Let $\mathbf{u}^* \in \arg\max_{\mathbf{u}\in U}\langle\mathbf{h}, \mathbf{u}\rangle$ and let $\widetilde{\mathbf{h}}_{m-1}^{\uparrow} \in \mathcal{C}_{m-1}$ such that $\text{UCB}_{t_{m-1}+1}(\widetilde{\mathbf{u}}_m) = \langle\widetilde{\mathbf{h}}_{m-1}^{\uparrow}, \widetilde{\mathbf{u}}_m\rangle$. Thus, with probability at least $1-\delta$, we have:

$$\widetilde{r}_m = J^* - J(\widetilde{\mathbf{u}}_m) = \langle\mathbf{h}, \mathbf{u}^*\rangle - \langle\mathbf{h}, \widetilde{\mathbf{u}}_m\rangle \leq \langle\widetilde{\mathbf{h}}_{m-1}^{\uparrow}, \widetilde{\mathbf{u}}_m\rangle$$

$$\begin{aligned}
&\leq \langle \widetilde{\mathbf{h}}_{m-1}^{\uparrow}, \widetilde{\mathbf{u}}_m \rangle && (13)\\
&\leq \left\| \widetilde{\mathbf{h}}_{m-1}^{\uparrow} \right\|_{\widetilde{\mathbf{V}}_{m-1}} \|\widetilde{\mathbf{u}}_m\|_{\widetilde{\mathbf{V}}_{m-1}^{-1}}\\
&\leq \left( \left\| \widetilde{\mathbf{h}}_{m-1}^{\uparrow} \right\|_{\widetilde{\mathbf{V}}_{m-1}} - \left\| \widetilde{\mathbf{h}}_{m-1} \right\|_{\widetilde{\mathbf{V}}_{m-1}} \right) \|\widetilde{\mathbf{u}}_m\|_{\widetilde{\mathbf{V}}_{m-1}^{-1}} && (14)\\
&\leq 2\widetilde{\beta}_{m-1} \|\widetilde{\mathbf{u}}_m\|_{\widetilde{\mathbf{V}}_{m-1}^{-1}} . && (15)
\end{aligned}$$

where line (13) follows from the optimism, line (14) derives from triangle inequality, line (15) is obtained by observing that $\mathbf{h} \in \mathcal{C}_{m-1}$ with probability at least $1 - \delta$, simultaneously for all $m \in [M]$, thanks to Theorem 3.1, having observed that $\widetilde{\beta}_{m-1}$ is larger than the right hand side of Theorem 3.1.

We now move to the cumulative offline regret over the whole horizon $T$, by decomposing w.r.t. the epochs and recalling that we pay the same instantaneous regret within each epoch:

$$R^{\text{off}}(\texttt{DynLin-UCB}, T) = \sum_{m=1}^{M} H_m (1 - \widetilde{r}_m) \leq \sqrt{\sum_{m=1}^{M} H_m (1 - \widetilde{r}_m^2)} \sqrt{\sum_{m=1}^{M} \widetilde{r}_m^2}.$$

Concerning the first summation, we proceed as follows, recalling that $M \leq T$ and $H_m \leq H_M$ for all $m \in [M]$:

$$\sum_{m=1}^{M} H_m (1 - \widetilde{r}_m^2) \leq T H_M (1 - \widetilde{r}_M^2) \leq T \left( 1 - \frac{\log T}{\log \frac{1}{\overline{\rho}}} \right).$$

For the second summation, we follow the usual derivation for linear bandits, recalling that $\widetilde{\beta}_{M-1} \geq \max(1, \widetilde{\beta}_{m-1})$ for all $m \in [M]$ and that under Assumption 2.2 we have that $\widetilde{r}_m^2 \leq 2$. In particular:

$$\widetilde{r}_m^2 \leq \min\left\{2, 2\widetilde{\beta}_{M-1} \|\widetilde{\mathbf{u}}_m\|_{\widetilde{\mathbf{V}}_{m-1}^{-1}}\right\} \leq 2\widetilde{\beta}_{M-1} \min\left\{1, \|\widetilde{\mathbf{u}}_m\|_{\widetilde{\mathbf{V}}_{m-1}^{-1}}\right\}.$$

Plugging this inequality into the second summation, we obtain:

$$\begin{aligned}
\sum_{m=1}^{M} \widetilde{r}_m^2 &\leq 4\widetilde{\beta}_{M-1}^2 \sum_{m=1}^{M} \min\left\{1, \|\widetilde{\mathbf{u}}_m\|_{\widetilde{\mathbf{V}}_{m-1}^{-1}}^2\right\}\\
&\leq 8d\widetilde{\beta}_{M-1}^2 \log\left(1 + \frac{MU^2}{d\lambda}\right) \leq 8d\beta_{T-1}^2 \log\left(1 + \frac{TU^2}{d\lambda}\right),
\end{aligned}$$

where the last passage follows from the elliptic potential lemma (Lattimore & Szepesvári, 2020, Lemma 19.4). Putting all together, we obtain the inequality holding with probability at least $1 - \delta$:

$$R^{\text{off}}(\texttt{DynLin-UCB}, T) \leq \sqrt{8dT\beta_{T-1}^2 \left(1 - \frac{\log T}{\log \frac{1}{\overline{\rho}}}\right) \log\left(1 + \frac{TU^2}{d\lambda}\right)},$$

having observed that $\widetilde{\beta}_{M-1} \leq \beta_{T-1}$ We can also arrive at a problem-dependent regret bound, by setting : $\inf_{\mathbf{u} \in U : \langle \mathbf{h}, \mathbf{u} \rangle < \langle \mathbf{h}, \mathbf{u}^* \rangle} \langle \mathbf{h}, \mathbf{u}^* - \mathbf{u} \rangle$ (if it exists $> 0$). Since the instantaneous regret is either 0 or at least , we have:

$$\begin{aligned}
R^{\text{off}}(\texttt{DynLin-UCB}, T) &\leq \sum_{m=1}^{M} H_m (1 - \widetilde{r}_m^2)\\
&\leq \frac{H_M}{1} 8d\widetilde{\beta}_{M-1}^2 \log\left(1 + \frac{MU^2}{d\lambda}\right)\\
&\leq \frac{8d}{} \left(1 - \frac{\log T}{\log \frac{1}{\overline{\rho}}}\right) \beta_{T-1}^2 \log\left(1 + \frac{TU^2}{d\lambda}\right).
\end{aligned}$$

By setting $\delta = 1/T$, replacing the value of $\beta_{T-1}$, we obtain the offline regret in expectation, highlighting the dependence on $T$, $\overline{\rho}$, $d$, and $\sigma$ only:

$$\mathbb{E}\, R^{\text{off}}(\texttt{DynLin-UCB}, T) \leq \mathcal{O}\left( \frac{d\sigma \sqrt{T} \log T^{\frac{3}{2}}}{1 - \overline{\rho}} \quad \sqrt{\frac{dT}{}} \log T^2 \right),$$

where we used the fact that $\frac{1}{\log \frac{1}{\overline{\rho}}} \leq \frac{1}{1-\overline{\rho}}$ and $\rho(\mathbf{A}) \leq \overline{\rho}$. □

The following lemma relates the expected offline regret with the expected online regret.

**Theorem 3.2** (Upper Bound)**.** *Under Assumptions 2.1 and 2.2, selecting $\beta_t$ as in Equation (6) and $\delta$ $1\{T$, `DynLin-UCB` suffers an expected regret bounded as (highlighting the dependencies on $T$, $\overline{\rho}$, $d$, and $\sigma$ only):*

$$\mathbb{E}\, R(\underline{\pi}^{\texttt{DynLin-UCB}}, T) \in \mathcal{O}\left( \frac{d\sigma \sqrt{\overline{T}} \log T^{\frac{3}{2}}}{1 - \overline{\rho}} \quad \frac{\sqrt{d\overline{T}} \log T^2}{(1-\overline{\rho})^{\frac{3}{2}}} \quad \frac{1}{(1-\rho(\mathbf{A}))^2} \right).$$

*Proof.* The result is simply obtained by exploiting the offline regret bound of Theorem B.2 and by upper bounding the expected regret using Lemma B.1. □

## C. Finite-Horizon Setting

In this section, we compare the finite-horizon setting with the infinite-horizon one presented in the main paper. We shall show that under Assumption 2.1, the two settings tend to coincide when the horizon is sufficiently large. Let us start by introducing the *$H$–horizon expected average reward*, with $H \in \mathbb{N}$ being the optimization horizon:

$$J_H(\underline{\pi}) := \mathbb{E}\left[ \frac{1}{H} \sum_{t=1}^{H} y_t \right] \quad \text{where} \quad \begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \boldsymbol{\epsilon}_t \\ y_t = \langle \boldsymbol{\omega}, \mathbf{x}_t \rangle + \langle \boldsymbol{\theta}, \mathbf{u}_t \rangle + \eta_t \\ \mathbf{u}_t = \pi_t(H_{t-1}) \end{cases}, \quad t \in [H], \tag{16}$$

where the expectation is taken w.r.t. the randomness of the state noise $\boldsymbol{\epsilon}_t$ and reward noise $\eta_t$. We now show that the optimal policy for the finite-horizon setting is a non-stationary open-loop policy.

**Theorem C.1** (Optimal Policy for the $H$–Horizon Setting)**.** *If $H \in \mathbb{N}$, an optimal policy $\underline{\pi}^*_H = (\pi^*_{H,t})_{t \in [H]}$ maximizing the $H$-horizon expected average reward $J(\underline{\pi})$ as in Equation (16) is given by:*

$$\forall t \in [H], \ \forall H_{t-1} \in \mathcal{H}_{t-1} : \quad \pi^*_{H,t}(H_{t-1}) = \mathbf{u}^*_{H,t} \quad \textit{where} \quad \mathbf{u}^*_{H,t} \in \operatorname*{arg\,max}_{\mathbf{u} \in U} \langle \mathbf{h}^{[0, H-t]}, \mathbf{u} \rangle.$$

*Proof.* We start by expressing for every $t \in [H]$ the reward $y_t$ as a function of the sequence of actions $\underline{\mathbf{u}} = (\mathbf{u}_1, \dots, \mathbf{u}_H)$ produced by a generic policy $\underline{\pi}$. By exploiting Equation (4) instanced with $H = t - 1$, we have:

$$y_t = \sum_{s=0}^{t-1} \langle \mathbf{h}^{\{s\}}, \mathbf{u}_{t-s} \rangle + \boldsymbol{\omega}^\top \mathbf{A}^{t-1} \mathbf{x}_1 + \eta_t = \sum_{s=1}^{t-1} \boldsymbol{\omega}^\top \mathbf{A}^{s-1} \boldsymbol{\epsilon}_{t-s}.$$

By computing the expectation, using linearity, and recalling that the noises are zero-mean, we obtain:

$$\mathbb{E}[y_t] = \sum_{s=0}^{t-1} \langle \mathbf{h}^{\{s\}}, \mathbb{E}[\mathbf{u}_{t-s}] \rangle + \boldsymbol{\omega}^\top \mathbf{A}^{t-1} \mathbb{E}[\mathbf{x}_1].$$

By averaging over $t \in [H]$, we obtain the $H$-horizon expected average reward:

$$\begin{aligned} J_H(\underline{\pi}) &= \frac{1}{H} \sum_{t=1}^{H} \mathbb{E}[y_t] \\ &= \frac{1}{H} \sum_{t=1}^{H} \sum_{s=0}^{t-1} \langle \mathbf{h}^{\{s\}}, \mathbb{E}[\mathbf{u}_{t-s}] \rangle + \frac{1}{H} \sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1} \mathbb{E}[\mathbf{x}_1] \\ &= \frac{1}{H} \sum_{s=1}^{H} \left( \sum_{t=s}^{H} \mathbf{h}^{\{t-s\}} \right)^\top \mathbb{E}[\mathbf{u}_s] + \frac{1}{H} \sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1} \mathbb{E}[\mathbf{x}_1] \qquad (17) \\ &= \frac{1}{H} \sum_{s=1}^{H} \langle \mathbf{h}^{[0, H-s]}, \mathbb{E}[\mathbf{u}_s] \rangle + \frac{1}{H} \sum_{t=1}^{H} \boldsymbol{\omega}^\top \mathbf{A}^{t-1} \mathbb{E}[\mathbf{x}_1]. \qquad (18) \end{aligned}$$

where line (17) is obtained by renaming the indexes of the summations, and line (18) comes from the definition of cumulative Markov parameter $\mathbf{h}^{[0,H-s]}$. It is now simple to see, as no noise is present in the expression, that the performance $J_H(\underline{\pi})$ is maximized by taking at each round $s \in \mathbb{N}$ an action $\mathbf{u}^*_s = \pi^*_s(H_{s-1})$ such that whose expectation satisfies $\mathbb{E}[\mathbf{u}^*_s] \in \arg\max_{\mathbb{E}[\mathbf{u}_s]} \langle \mathbf{h}^{[0,H-s]}, \mathbb{E}[\mathbf{u}_s] \rangle$. Clearly, we can take the deterministic action such that $\mathbf{u}^*_s = \mathbb{E}[\mathbf{u}^*_s]$. □

We now show that for sufficiently large $H$, the $H$-horizon expected average reward $J_H$ tends to coincide with the infinite-horizon expected average reward.

**Proposition C.2.** *Let $H$ P N. Then, for every policy $\underline{\pi}$ it holds that:*

$$|J_H p\underline{\pi}q \quad Jp\underline{\pi}q| ¤ \frac{BU \quad pAqp1 \quad \rho pAq^H q}{Hp1 \quad \rho pAqq}.$$

*Proof.* Consider two horizons $H \quad H'$ P N, and let $pu_1, u_2, \dots q$ be the sequence of actions played by policy $\underline{\pi}$. Using Equation (18), we have:

$$J_H p\underline{\pi}q \quad J_{H'} p\underline{\pi}q \quad \frac{1}{H}\sum_{s=1}^{H} xh^{J0,H-sK}, Eru_s sy \quad \frac{1}{H'}\sum_{s=1}^{H'} xh^{J0,H'-sK}, Eru_s sy \tag{19}$$

$$\frac{1}{H}\sum_{s=1}^{H} xh^{J0,H-sK} \quad h, Eru_s sy \quad \frac{1}{H'}\sum_{s=1}^{H'} xh^{J0,H'-sK} \quad h, Eru_s sy \tag{20}$$

$$\frac{1}{H}\sum_{s=1}^{H} xh^{JH-s+1,+\infty M}, Eru_s sy \quad \frac{1}{H'}\sum_{s=1}^{H'} xh^{JH'-s+1,+\infty M}, Eru_s sy. \tag{21}$$

As shown in Appendix B.1, we have that the second addendum vanishes as $H'$ approaches 8:

$$\frac{1}{H'}\left|\sum_{s=1}^{H'} xh^{JH'-s+1,+\infty M}, Eru_s sy\right| Ñ 0 \qquad \text{when} \qquad H' Ñ 8.$$

Concerning the first addendum, we have:

$$\begin{aligned}\frac{1}{H}\left|\sum_{s=1}^{H} xh^{JH-s+1,+\infty M}, Eru_s sy\right| &¤ \frac{U}{H}\sum_{s=1}^{H}\left\|h^{JH-s+1,+\infty M}\right\|_2 \\ &¤ \frac{BU \quad pAq}{H}\sum_{s=1}^{H}\rho pAq^{H-s} \\ &\quad \frac{BU \quad pAqp1 \quad \rho pAq^H q}{Hp1 \quad \rho pAqq}.\end{aligned}$$

□

## D. System Identification

This section presents a solution to identify matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ characterizing an LTI system starting from a single trajectory. We adopt a variant of the Ho-Kalman (Ho & Kalman, 1966) algorithm. We start from the identification method proposed by Lale et al. (2020a, Section 3), where authors consider a system of the type (strictly proper):

$$\begin{aligned}\mathbf{x}_{t+1} \quad \mathbf{A}\mathbf{x}_t \quad \mathbf{B}\mathbf{u}_t \quad \boldsymbol{\epsilon}_t,\\ \tilde{\mathbf{y}}_t \quad \mathbf{C}\mathbf{x}_t \quad \mathbf{z}_t.\end{aligned} \tag{22}$$

Our setting can be seen as (not strictly proper):

$$\begin{aligned}\mathbf{x}_{t+1} \quad \mathbf{A}\mathbf{x}_t \quad \mathbf{B}\mathbf{u}_t \quad \boldsymbol{\epsilon}_t,\\ \mathbf{y}_t \quad \mathbf{C}\mathbf{x}_t \quad \mathbf{D}\mathbf{u}_t \quad \mathbf{z}_t,\end{aligned} \tag{23}$$

with $\mathbf{x}_t, \boldsymbol{\epsilon}_t$ P $\mathsf{R}^n$, $\mathbf{u}_t$ P $\mathsf{R}^p$, and $\mathbf{y}_t, \mathbf{z}_t$ P $\mathsf{R}^m$. The noise over state transition model $\boldsymbol{\epsilon}_t$ and output $\mathbf{z}_t$ are $\sigma^2$-subgaussian random variables. We consider in this part the standard control problem notation adopted for LTI systems. The mapping to our problem is straightforward by considering $\mathbf{C} \quad \boldsymbol{\omega}^{\mathrm{T}}$ and $\mathbf{D} \quad \boldsymbol{\theta}^{\mathrm{T}}$. In predictive form, the system described in Equation (22) is:

$$\begin{aligned}\hat{\mathbf{x}}_{t+1} \quad \mathbf{A}\hat{\mathbf{x}}_t \quad \mathbf{B}\mathbf{u}_t \quad \mathbf{F}\tilde{\mathbf{y}}_t,\\ \tilde{\mathbf{y}}_t \quad \mathbf{C}\hat{\mathbf{x}}_t \quad \mathbf{e}_t,\end{aligned}$$

where:

$$\mathbf{A} \quad \mathbf{A} \quad \mathbf{FC},$$

$$\mathbf{F} \quad \mathbf{A\Sigma C}^{\mathsf{T}}\mathsf{p}\mathbf{C\Sigma C}^{\mathsf{T}} \quad \sigma^2\mathbf{I}\mathsf{q}^{-1},$$

and $\mathbf{\Sigma}$ is the solution to the following DARE (Discrete Algebraic Riccati Equation):

$$\mathbf{\Sigma} \quad \mathbf{A\Sigma A}^{\mathsf{T}} \quad \mathbf{A\Sigma C}^{\mathsf{T}}\mathsf{p}\mathbf{C\Sigma C}^{\mathsf{T}} \quad \sigma^2\mathbf{I}\mathsf{q}^{-1}\mathbf{C\Sigma A}^{\mathsf{T}} \quad \sigma^2\mathbf{I}.$$

In order to identify this LTI system, we want to detect a matrix $\widetilde{\mathcal{G}}_y$:

$$\widetilde{\mathcal{G}}_y \quad \left[\mathbf{CF} \quad \mathbf{CAF} \quad \dots \quad \mathbf{CA}^{H-1}\mathbf{F} \quad \mathbf{CB} \quad \mathbf{CAB} \quad \dots \mathbf{CA}^{H-1}\mathbf{B}\right]. \tag{24}$$

To identify through least squares method matrix $\widetilde{\mathcal{G}}_y$, we construct for each $t$, a vector $\widetilde{\phi}_t$:

$$\widetilde{\phi}_t \quad \left[\mathbf{y}_{t-1}^{\mathsf{T}} \quad \dots \quad \mathbf{y}_{t-H}^{\mathsf{T}} \quad \mathbf{u}_{t-1}^{\mathsf{T}} \quad \dots \quad \mathbf{u}_{t-H}^{\mathsf{T}}\right]^{\mathsf{T}} \mathsf{P}\,\mathsf{R}^{(m+p)H}. \tag{25}$$

The system output $\widetilde{\mathbf{y}}_t$ can be rewritten as:

$$\widetilde{\mathbf{y}}_t \quad \widetilde{\mathcal{G}}_y\widetilde{\phi}_t \quad \mathbf{e}_t \quad \mathbf{CA}^H\mathbf{x}_{t-H}.$$

The output of the system under analysis (Equation 23) is:

$$\mathbf{y}\mathsf{p}t\mathsf{q} \quad \widetilde{\mathbf{y}}_t \quad \mathbf{Du}_t \quad \widetilde{\mathcal{G}}_y\widetilde{\phi}_t \quad \mathbf{Du}_t \quad \mathbf{e}_t \quad \mathbf{CA}^H\mathbf{x}_{t-H}$$

We can incorporate the contribution of $\mathbf{Du}_t$ in $\widetilde{\mathcal{G}}_y$ obtaining $\mathcal{G}_y$:

$$\mathcal{G}_y \quad \left[\mathbf{CF} \quad \mathbf{CAF} \quad \dots \quad \mathbf{CA}^{H-1}\mathbf{F} \quad \mathbf{D} \quad \mathbf{CB} \quad \mathbf{CAB} \quad \dots \mathbf{CA}^{H-1}\mathbf{B}\right].$$

The related vector $\phi_t$ is:

$$\phi_t \quad \left[\mathbf{y}_{t-1}^{\mathsf{T}} \quad \dots \quad \mathbf{y}_{t-H}^{\mathsf{T}} \quad \mathbf{u}_t^{\mathsf{T}} \quad \mathbf{u}_{t-1}^{\mathsf{T}} \quad \dots \quad \mathbf{u}_{t-H}^{\mathsf{T}}\right]^{\mathsf{T}} \mathsf{P}\,\mathsf{R}^{(m+p)H+p}. \tag{26}$$

The best value of $\mathcal{G}_y$ can be found through regularized least squares as in Lale et al. (2020a, Equation 10):

$$\widehat{\mathcal{G}}_y \quad \arg\min_X \lambda\}\mathbf{X}\}_F^2 \quad \sum_{\tau=t-H}^{t} \}\mathbf{y}_\tau \quad \mathbf{X}\phi_\tau\}_2^2, \tag{27}$$

where $\}\ \}_F$ represents the Frobenius norm.

The matrix $\mathbf{D}$ can be directly retrieved from $\widehat{\mathcal{G}}_y$. In order to get matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, we remove the values related to $\mathbf{D}$ from $\widehat{\mathcal{G}}_y$ and we retrieve $\widetilde{\mathcal{G}}_y$. From now on, we refer to the algorithm proposed in Lale et al. (2020a, Appendix B).

## E. Integration on Numerical Simulations

This section is divided in three parts. First, in Section E.1, we provide additional information about the baselines, their hyperparameters and the optimistic bounds. Second, in Section E.2, we provide all the matrices and vectors generalized to run the real-world experiment. Third, in Section E.3, we provide further results for the simulations presented in Section 5.1.

### E.1. Additional Notes on the Baselines

As mentioned in Section 5, the chosen baselines are `Lin-UCB` (Abbasi-Yadkori et al., 2011), `D-Lin-UCB` (Russac et al., 2019), `AR2` (Chen et al., 2021), `Exp3` (Auer et al., 1995), `Exp3-k` (Dekel et al., 2012; Auer et al., 1995) and the `Expert` (the latter available only in the case of real-world data). All the hyperparameters, whenever possible, are set as prescribed in the original papers. The bounds used for the exploration are adjusted in order to be able to fairly compete in this setting, and are considered as follows:

$$\beta_t^{\texttt{Lin-UCB}} : \quad \bar{c}_2 \overset{?}{} \bar{\lambda} \quad \sqrt{2\bar{\sigma}^2\left(\log\left(\frac{1}{\delta}\right) \quad \frac{d}{2}\log\left(1 \quad \frac{tU^2}{d\lambda}\right)\right)},$$

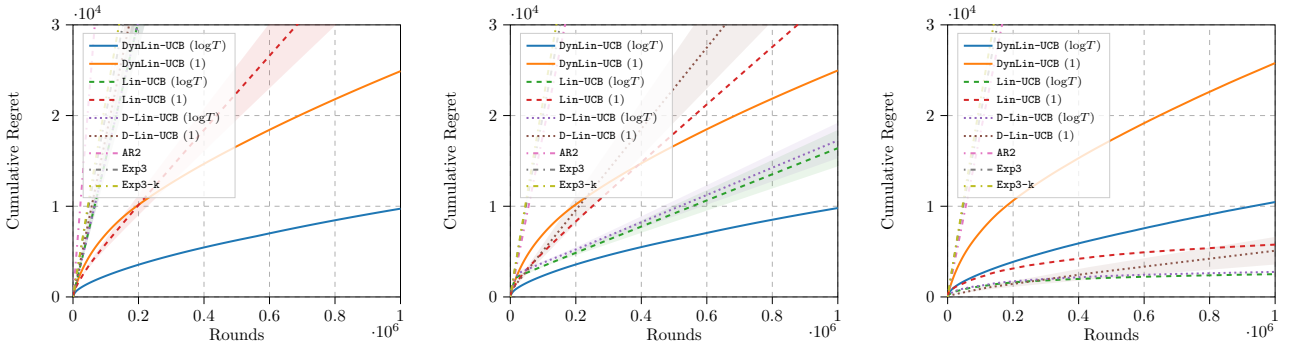
$$\beta_t^{\texttt{D-Lin-UCB}} : \quad \bar{c}_2 \overset{?}{} \bar{\lambda} \quad \sqrt{2\bar{\sigma}^2\left(\log\left(\frac{1}{\delta}\right) \quad \frac{d}{2}\log\left(1 \quad \frac{tU^2}{d\lambda}\left(\frac{1 \quad \gamma^{2t}}{1 \quad \gamma^2}\right)\right)\right)},$$

where $\bar{c}_2$ and $\bar{\sigma}^2$ are as prescribed in Section 3.2, and the hyperparameter $\gamma$ of `D-Lin-UCB` is tuned.

For `AR2`, the hyperparameter $\alpha$, describing the correlation over time is considered equal to $\rho\mathsf{p}\mathbf{A}\mathsf{q}$.

In the case of `Exp3`, the rewards are rescaled in order to make them range in $\mathsf{r}0, 1\mathsf{s}$ with high probability, as follows:

$$\bar{r}_t \quad \frac{r_t \quad 2\xi}{4\xi}, \qquad \text{where} \qquad \xi \quad \left(\quad \frac{B}{1 \quad \rho\mathsf{p}\mathbf{A}\mathsf{q}}\right)U.$$




(a) $\sigma$ 0.001 (b) $\sigma$ 0.01 (c) $\sigma$ 0.1

*Figure 4.* Performance of DynLin-UCB, Lin-UCB, D-Lin-UCB, AR2, Exp3 and Exp3-k at different values of $\sigma$. (50 runs, mean std)

Furthermore, in the case of Exp3-k, the batch dimension $k$ is considered as:

$$k \quad \left\lceil \frac{\log M}{\log 1\{\overline{\rho}} \right\rceil,$$

where $M$ is the one defined in Algorithm 1 (line 2). This batch size $k$ ensures that, at each time $t$, the contribution of actions $\mathbf{u}_s$ is negligible, with $s$ P J$t$ $k$ 1K. The rewards collected in the same batch are averaged and transformed as in Exp3.

### E.2. Further Information on the Real-world Setting

The real-world setting is generalized through a dataset containing real data related to the budgets invested in each advertising platform (i.e., the $\mathbf{u}_t$) and the overall generated conversions (i.e., the $y_t$) collected from three of the most important advertising platforms of the web (Facebook, Google, and Bing), related to a large number of campaigns for a value of more than 5 Million USD over 2 years. Starting from such data, we generalized the best model by means of a specifically designed variant of the Ho-Kalman algorithm (Ho & Kalman, 1966), as described in Appendix D. We used the matrices estimated with Ho-Kalman to build up a simulator. The resulting system has $\rho$pAq 0.67, and is characterized as follows:

$$\mathbf{A} \begin{pmatrix} 0.38 & 0.33 & 0.6 \\ 0.07 & 0.76 & 0.54 \\ 0.18 & 0.34 & 0.05 \end{pmatrix}, \quad \mathbf{B} \begin{pmatrix} 0.14 & 0.34 & 0.05 \\ 0.17 & 0.03 & 0.01 \\ 0.04 & 0.09 & 0.17 \end{pmatrix}, \quad \boldsymbol{\omega} \begin{pmatrix} 0.61 \\ 0.04 \\ 0.13 \end{pmatrix}, \quad \boldsymbol{\theta} \begin{pmatrix} 0.13 \\ 0.41 \\ 0.02 \end{pmatrix}.$$

### E.3. Additional Numerical Simulations

These additional results are obtained in the setting presented in Section 5.1. However, here, we want to analyze the behavior of DynLin-UCB and the other bandit baselines at different magnitudes of noise in both the state transition model and the output. The noise in this simulation is a zero-mean Gaussian noise with $\sigma$ P t0.001, 0.01, 0.1u.

**Results** Figure 4 shows the results of the experiment for the different values of $\sigma$. It is clearly visible how DynLin-UCB performs in almost the same way no matter the noise to which the system is subject, always leading to sub-linear regret. On the other hand, the cumulative regret of both Lin-UCB and D-Lin-UCB is different in every simulation we perform. Indeed, with a low level of noise (Figure 4a) reaches linear regret and does not converge, while for large values of noise, it converges very quickly (Figure 4c). This is due to the nature of the confidence bound of linear bandits, which is not able to take into account such a complex scenario and leads to no guarantees in this setting. Exp3, Exp3-k, and AR2 are not able to reach the optimum in this scenario, independently from the noise magnitude $\sigma$, and provide large values of (linear) regret.

### E.4. Computational Time

The code used for the results provided in this section has been run on an Intel(R) I5 8259U @ 2.30GHz CPU with 8 GB of LPDDR3 system memory. The operating system was macOS 12.2.1, and the experiments have been run on *Python* 3.9.7. A single run of DynLin-UCB takes 110 seconds to run. It is worth noting that the time complexity of DynLin-UCB is upper-bounded by the one of Lin-UCB.