Dynamical Linear Bandits for Long-Lasting Vanishing Rewards

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

Abstract

In many real-world sequential decision-making problems, an action might not immediately reflect on the feedback and spread its effects over a long time horizon. For instance, in online advertising, investing in a platform produces an increase in *awareness*, but the actual reward (e.g., a *conversion*) might occur far in the future. Furthermore, whether a conversion takes place depends on several factors: how fast the awareness grows; possible awareness vanishing effects; 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 into the future, disregarding possible hidden dynamical effects. In this paper, we introduce a novel setting, Dynamical Linear Bandits (DLB), an extension of 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 the action, and the hidden state evolves according to a linear dynamics. In this way, the effects of each action are delayed by the system evolution, persist over time, and the interplay between the action components is taken into account. After introducing the setting and discussing the notion of optimal policy, we provide an any-time optimistic regret minimization algorithm Dynamical Linear Upper Confidence Bound (`DynLin-UCB`) that suffers regret of order $\widetilde{\mathcal{O}}(cd\sqrt{T})$, where c is a constant dependent on the properties of the linear dynamical evolution. Finally, we conduct a numerical validation on a synthetic environment to show the effectiveness of `DynLin-UCB` in comparison with bandit baselines.

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

Complex Feedback in Online Learning Workshop at the International Conference on Machine Learning, Baltimore, Maryland, USA, 2022.

1. Introduction

In a large variety of sequential decision-making problems, a learner is required choosing 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 might observe a reward (a.k.a. feedback) that is the combined effect of multiple actions played in the past and its realization might span a large time horizon. For instance, in online advertising campaigns, the process that leads to a *conversion*, also known as the marketing funnel (Court et al., 2009), is characterized by complex dynamics and comprises several phases. When multiple campaigns and heterogeneous platforms are involved, a profitable budget investment policy has to account for relationships between campaigns/platforms and how their effects are combined. In this scenario, the conversion attribution problem (Berman, 2018) consists in assigning the credit of a conversion (e.g., a user's purchase of a promoted product) not only to the latest ad the user was exposed to, but also to the previous ones which contributed to generate such a conversion. 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 the brand awareness, while the latter aims at creating the actual conversion. If we evaluate the performances in terms of conversions only, we will observe that impression ads are not instantaneously effective in creating conversions, so we will be tempted to reduce the amount of budget invested in such a campaign. However, this approach is clearly sub-optimal because, as demonstrated in several works (e.g., Hoban & Bucklin, 2015), impression ads increase the chance to convert when a conversion ad is shown after the impression. 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 delayed consequences that simple models cannot capture. It is worth noting that this kind of interplay comprises more general scenarios than simple delay (which are anyway present), including cases where the interaction is governed by dynamics

hidden to the observer.

While this problem can be indubitably modeled as a partially-observable Markov Decision Process (POMDP, Åström, 1965), the complexity of the framework and its generality 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 (a.k.a. feedback) 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 for 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 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 an hidden internal state x_t (e.g., awareness) which evolves via a linear dynamics fed by the action u_t (e.g., the amount invested on each platform) and is 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 x_t and the action 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, reduce to linear bandits when no dynamics is involved. Coming back to the application scenario, the state allows encoding the awareness that accumulates and/or vanishes over time. Furthermore, thanks to the dynamics, the model allows for representing interference and synergy phenomena between platforms. Due to the dynamical nature of the system, the effect of each action persists over time indefinitely but, under stability conditions, it vanishes asymptotically.

Contributions The contributions of this paper are theoretical, algorithmic, and experimental and can be summarized as follows. 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 and the learner observes a noisy reward obtained as a linear combination of the current state and action played. We formally define the hidden-state linear dynamics, the computation of the reward, the notion of policy, and the regret definition. In particular, under stability conditions, we show that the optimal policy corresponds to playing the *constant action* that leads the system to the most profitable steady-state. In Section 3, we propose a novel anytime optimistic regret minimization algorithm, *Dynamical Linear Upper Confidence Bound* (`DynLin-UCB`) for the DLB setting. `DynLin-UCB` takes inspiration from `Lin-UCB` and subdivides the optimization horizon T into increasing-length epochs. In each epoch, an action is selected and kept constant (persisted) so that the system approximately reaches the steady-state. Now, a reliable observation of the reward is collected to update the estimates and optimistically select the next action to play. We provide a regret analysis for `DynLin-UCB` showing that, under certain assumptions, it enjoys $\widetilde{\mathcal{O}}(cd\sqrt{T})$ expected regret, where c is a constant that accounts for the "speed" at which the system reaches the steady-state and d is the dimensionality of the action u. In Section 5, we provide a numerical validation to highlight the properties of our approach in comparison with bandit baselines. The proof of all the results are reported in Appendix A.

Notation Let $a, b \in \mathsf{N}$ with $a \leqslant b$, we denote with $\mathsf{J}a, b\mathsf{K} \coloneqq \{a, \dots, b\}$, with $\mathsf{J}b\mathsf{K} \coloneqq \mathsf{J}1, b\mathsf{K}$, and with $\mathsf{J}a, \infty\mathbb{M} = \{a, a+1, \dots\}$. Let $\mathsf{x}, \mathsf{y} \in \mathsf{R}^n$ be real-valued vectors, we denote with $\langle \mathsf{x}, \mathsf{y} \rangle = \mathsf{x}^\mathsf{T}\mathsf{y} = \sum_{j\ 1}^{n} x_i y_i$ the inner product. For a positive semidefinite matrix $\mathsf{A} \in \mathsf{R}^{n\ \ n}$, we denote with $\|\mathsf{x}\|_\mathsf{A}^2 = \mathsf{x}^\mathsf{T}\mathsf{A}\mathsf{x}$ the weighted 2-norm. The *spectral radius* $\rho(\mathsf{A})$ is the largest absolute value of the eigenvalues of A, the *spectral norm* $\|\mathsf{A}\|_2$ is the square root of the maximum eigenvalue of $\mathsf{A}^\mathsf{T}\mathsf{A}$, the *Frobenius norm* $\|\mathsf{A}\|_F$ is the trace of $\mathsf{A}^\mathsf{T}\mathsf{A}$. We introduce the maximum spectral norm to spectral radius ratio of the powers of A defined as $(\mathsf{A}) = \mathsf{sup}_{\ \yen 0} \|\mathsf{A}\ \|_2/\rho(\mathsf{A})$ (Oymak & Ozay, 2019). We denote with I_n the identity matrix of order n and with $\mathsf{0}_n$ the vector of all zeros of dimension n. A random vector $\mathsf{x} \in \mathsf{R}^n$ is σ^2-subgaussian, in the sense of (Hsu et al., 2012), if for every vector $\in \mathsf{R}^n$ it holds that $\mathsf{E}[\mathsf{exp}(\langle\ , \mathsf{x}\rangle)] \leqslant \mathsf{exp}(\|\ \|_2^2\sigma^2/2)$.

2. Problem Formulation

In this section, we introduce the *Dynamical Linear Bandits* (DLBs), formulate the learning problem, focusing on the learner-environment interaction, assumptions, policies, and definition of regret (Section 2.1). Then, we derive a closed-form expression for the optimal policy for DLBs (Section 2.2).

2.1. Setting

We consider the sequential interaction between a learner and an environment. In a Dynamical Linear Bandit (DLB), the environment is characterized by a *hidden* state, i.e., a n-dimensional real vector, initialized to $\mathsf{x}_1 \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathsf{R}^n$ is the state space. At each round $t \in \mathsf{N}$, the environment is in the hidden state $\mathsf{x}_t \in \mathcal{X}$, the learner chooses

an action, i.e., a d-dimensional real vector $\mathsf{u}_t \in \mathcal{U}$, where $\mathcal{U} \subseteq \mathsf{R}^d$ is the action space. Then, the learner receives a noisy reward $y_t = \langle \ !, \mathsf{x}_t \rangle + \langle \ , \mathsf{u}_t \rangle + \eta_t \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathsf{R}$ is the reward space, $! \in \mathsf{R}^n$, $\in \mathsf{R}^d$ are unknown parameters, and η_t is a zero-mean σ^2–subgaussian random noise, conditioned to the past. Then, the environment evolves to the new state according to the unknown linear dynamics $\mathsf{x}_{t\ 1} = \mathsf{A}\mathsf{x}_t + \mathsf{B}\mathsf{u}_t + \ _t$, where $\mathsf{A} \in \mathsf{R}^{n\ \ n}$ is the dynamic matrix, $\mathsf{B} \in \mathsf{R}^{n\ \ d}$ is the action-state matrix, and $\ _t$ is a zero-mean σ^2–subgaussian random noise, conditioned to the past, independent of η_t.[1]

Remark 2.1. *The setting proposed above is a particular case of a POMDP (Littman et al., 1995), in which the state x_t is non-observable, while the learner has access to the noisy observation y_t that, in our setting, 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 (Lattimore & Szepesvári, 2020) when the hidden state does not affect the reward, i.e., when $! = 0$.*

Markov Parameters We revise a useful representation, widely employed in the LTI literature, that allows expressing y_t in terms of the sequence of the most recent $H + 1 \in \mathsf{N}$ actions $(\mathsf{u}_s)_{s\mathsf{PJ}t\ \ H:t\mathsf{K}}$, reward noise η_t, H state noises $(\ _s)_{s\mathsf{PJ}t\ \ H:t\ \ 1\mathsf{K}}$, and starting state $\mathsf{x}_{t\ \ H}$ (Ho & Kálmán, 1966; Oymak & Ozay, 2019; Tsiamis & Pappas, 2019; Sarkar et al., 2021):

$$y_t = \sum_{s\ \ 0}^{H} \langle \mathsf{h}^{t\mathsf{s}u}, \mathsf{u}_{t\ \ s} \rangle + \underbrace{!^{\mathsf{T}} \mathsf{A}^{H} \mathsf{x}_{t\ \ H}}_{\text{starting state}} + \underbrace{+\eta_t + \sum_{s\ \ 1}^{H} !^{\mathsf{T}} \mathsf{A}^{s\ \ 1}\ _{t\ \ s}}_{\text{noise}}, \tag{1}$$

where the sequence of vectors $\mathsf{h}^{t\mathsf{s}u} \in \mathsf{R}^d$ for every $s \in \mathsf{N} \cup \{0\}$ are called *Markov parameters* and are defined as: $\mathsf{h}^{t0u} =$ and $\mathsf{h}^{t\mathsf{s}u} = \mathsf{B}^{\mathsf{T}}(\mathsf{A}^{s\ \ 1})^{\mathsf{T}} !$ if $s \geqslant 1$. Furthermore, we introduce the *cumulative Markov parameters*, defined for every $s, s^1 \in \mathsf{N} \cup \{0\}$ with $s \leqslant s^1$ as $\mathsf{h}^{\mathsf{J}s;s^1\mathsf{K}} = \sum_{l\ \ s}^{s^1} \mathsf{h}^{t/u}$ and the corresponding limit as $s^1 \to +\infty$, i.e., $\mathsf{h}^{\mathsf{J}s;\ \ 8\mathsf{M}} = \sum_{l\ \ s}^{8} \mathsf{h}^{t/u}$. Finally, we use the abbreviation $\mathsf{h} = \mathsf{h}^{\mathsf{J}0;\ \ 8\mathsf{M}} = \ + \mathsf{B}^{\mathsf{T}}(\mathsf{I} - \mathsf{A})^{\ \mathsf{T}} !$. We will make use of the following standard assumptions.

Assumption 2.1 (Boundedness)**.** *The following inequalities hold: $\| \ \|_2 \leqslant \ $, $\| ! \|_2 \leqslant \ $, $\|\mathsf{B}\|_2 \leqslant B$, $\|\mathsf{u}\|_2 \leqslant U$ with $\mathsf{u} \in \mathcal{U}$, and $\|\mathsf{x}\|_2 \leqslant X$ with $\mathsf{x} \in \mathcal{X}$, and $\sup_{\mathsf{u};\mathsf{u}^1\mathsf{P}U} \langle \ , \mathsf{u} - \mathsf{u}^1 \rangle \leqslant 1$.*[2]

Assumption 2.2 (Stability)**.** *The spectral radius of A is strictly smaller than 1, i.e., $\rho(\mathsf{A}) < 1$, and the maximum spectral norm to spectral radius ratio of the powers of A is bounded, i.e., $\ (\mathsf{A}) < +\infty$.*[3]

The former assumption requires the boundedness of the norms of the relevant vectors, matrices, as well as states and actions; whereas the latter is related to the *stability* of the dynamic matrix A, which is widely employed in discrete–time LTI literature (Oymak & Ozay, 2019; Lale et al., 2020b;a).

Policies and Performance The learner's behavior is modeled by a deterministic *policy* $\_ = (\ _t)_{t\mathsf{PN}}$ defined, for every round $t \in \mathsf{N}$, as $\ _t : \mathcal{H}_{t\ \ 1} \to \mathcal{U}$, mapping the history of observations $H_{t\ \ 1} = (\mathsf{u}_1, y_1, \ldots, \mathsf{u}_{t\ \ 1}, y_{t\ \ 1}) \in \mathcal{H}_{t\ \ 1}$ to an action $\mathsf{u}_t = \ _t(H_{t\ \ 1}) \in \mathcal{U}$, 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 $\underline{\pi}$ is evaluated in terms of the *infinite-horizon expected average reward*:[4]

$$J(\_) := \liminf_{H\ \mathsf{N}\ \ 8} \mathsf{E}\left[\frac{1}{H}\sum_{t\ \ 1}^{H} y_t\right] \tag{2}$$

$$\text{where} \quad \begin{cases} \mathsf{x}_{t\ \ 1} = \mathsf{A}\mathsf{x}_t + \mathsf{B}\mathsf{u}_t + \ _t \\ y_t = \langle \ !, \mathsf{x}_t \rangle + \langle \ , \mathsf{u}_t \rangle + \eta_t \quad , \quad t \in \mathsf{N}, \\ \mathsf{u}_t = \ _t(H_{t\ \ 1}) \end{cases}$$

where the expectation is taken w.r.t. the randomness of the state noise $\ _t$ and reward noise η_t. If a policy $\_$ is *constant*, i.e., $\ _t(H_{t\ \ 1}) = \mathsf{u}$ for every $t \in \mathsf{N}$, we abbreviate $J(\mathsf{u}) = J(\_)$. A policy $_\ $ is an *optimal infinite-horizon policy* if it maximizes the infinite-horizon expected average reward, i.e., $\_\ \in \arg\max_{\underline{\pi}} J(\_)$, whose performance is denoted as $J\ := J(\_\ )$.

Regret The goal of the learner is to minimize the *online expected (policy) regret* by playing a policy $\_$, competing against the optimal infinite-horizon policy $_\ $ over a *learning horizon* $T \in \mathsf{N}$: $\mathsf{E}\, R(\_, T) := \mathsf{E}[\sum_{t\ \ 1}^{T} J\ - y_t]$, where y_t is the sequence of rewards collected by playing $\_$ as in Equation (2). Furthermore, we introduce a different notion of regret, that will turn useful for analysis purposes, that we name *offline expected (policy) regret* that compares $J\ $ with the infinite-horizon performance of the

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

[2] The assumption of the bounded state norm $\}\mathsf{x}\}_2 \ \mathsf{p}\ X$ can be replaced with the assumption of bounded state noise $\ $. As shown in (Agarwal et al., 2019), this assumption can be relaxed 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.

[3] The latter is a mild assumption: if A is diagonalizable as $\mathsf{A}\ \ \mathsf{Q}\ \mathsf{Q}^{\ 1}$, than $\ \mathsf{p}\mathsf{A}\mathsf{q}\ \mathsf{p}\ \}\mathsf{Q}\}_2\}\mathsf{Q}^{\ 1}\}_2$ and it is finite. In particular, if A is symmetric than $\ \mathsf{p}\mathsf{A}\mathsf{q}\ \ 1$.

[4] In Appendix B, we show that, under Assumption 2.2, the infinite-horizon setting is not dissimilar from the finite-horizon setting, provided that the horizon is sufficiently large.

action u_t $\pi_{t-1}(H_{t-1})$ played at each round $t \in [T]$ by the agent: $\mathrm{ER}^{\mathrm{off}}(\pi, T) := \mathbb{E}\left[\sum_{t=1}^{T} J(u^\star) - J(u_t)\right]$. Clearly, the two regret definition coincide when the system does not have a dynamics, i.e., $\{A\}_2 = 0$.

2.2. Optimal Policy

In this section, we derive a closed-form expression for the optimal policy $\pi^\star$ 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 $\pi^\star$ maximizing the infinite-horizon expected average reward $J(\pi)$, as in Equation (2), is given by:

$$\forall (H_{t-1}) \in (\mathbb{R}^d)^{t-1} \times \mathbb{N}, \forall t \in \mathbb{N}: \quad \pi_{t-1}(H_{t-1}) = u^\star \quad (3)$$
$$\text{where } u^\star \in \arg\max_{u \in \mathcal{U}} J(u) = x(u; u).$$

Some remarks are in order. The optimal policy performs the constant action $u^\star \in \mathcal{U}$ which brings the system in the "most profitable" steady-state.[5] Indeed, the expression $x(u; u)$ can be rewritten expanding the cumulative Markov parameter as $\sum_{i=0}^{T} A^i B u \to \sum_{i=0}^{T} (I - A)^{-1} B u$ and is the expression of the steady state $\bar{x} = A\bar{x} + Bu$, when applying action u. It is worth noting the role of Assumption 2.2 which guarantees the existence of the inverse $(I - A)^{-1}$. In this sense, our problem shares the constant nature of the optimal policy with the linear bandit setting (Lattimore & Szepesvári, 2020), 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 focus on open-loop stationary policies only. The reason why this problem does not benefit from closed-loop policies, differently from other classical problems, such as the LQG (Abbasi-Yadkori & Szepesvári, 2011), lies in the linearity of the reward and in the additive nature of the noise components w_t and η_t, which makes its presence irrelevant 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 $u \in \mathcal{U}$, we need to let the system evolve to the steady-state and, then, observe the reward.

3. Algorithm

In this section, we present an anytime optimistic regret minimization algorithm for the DLB setting introduced in Section 2. Dynamical Linear Upper Confidence Bound (DynLin-UCB), whose pseudocode is reported in Algorithm 1, requires the knowledge of an upper-bound $\bar{\rho} < 1$ on the spectral radius of the dynamic matrix A (i.e., $\rho(A) \le \bar{\rho}$) and on the maximum spectral norm to spectral radius ratio $\bar{\gamma} \ge 8$ (i.e., $\|A\| \le \bar{\gamma}\bar{\rho}$), as well as the bounds on the relevant quantities of Assumption 2.1.[6] DynLin-UCB is based on the following simple observation. To assess the quality of an action $u \in \mathcal{U}$, 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 $J(u) = x(u; u)$. We shall show that, under Assumption 2.2, 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 $\bar{\gamma}$. After initializing the Gram matrix $V_0 = I_d$ and the vectors b_0 and $\hat{\theta}_0$ both to 0_d (line 1), DynLin-UCB subdivides the learning horizon T in M epochs. Each epoch $m \in [M]$ is composed of $H_m + 1$ rounds, where $H_m = t_{\log}(m) - t_{\log}(m-1)$ 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):

$$\forall u \in \mathcal{U}: \quad \mathrm{UCB}_t(u) = x(\hat{\theta}_{t-1}; u) + \beta_{t-1} \|y_{t-1} - u\|_{V_{t-1}^{-1}}, \quad (4)$$

where $\hat{\theta}_{t-1} = V_{t-1}^{-1} b_{t-1}$ is the regression estimator of the cumulative Markov parameter as employed in Equation (3) to define the optimal action and $\beta_{t-1} \ge 0$ is an exploration coefficient that will be defined later. Similar to the LinUCB algorithm (Lattimore & Szepesvári, 2020), the index $\mathrm{UCB}_t(u)$ is designed to be optimistic, i.e., $J(u) \le \mathrm{UCB}_t(u)$ in high-probability for all $u \in \mathcal{U}$. Then, the optimistic action $u_t \in \arg\max_{u \in \mathcal{U}} \mathrm{UCB}_t(u)$ is executed (line 5) and persisted for the next H_m rounds (lines 7-10). The length of the epoch H_m is selected such that, under Assumption 2.2, 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 $J(u_t)$. This sample is employed to update the Gram matrix estimate V_t and the vector b_t (line 12), while the samples collected in the previous H_m rounds are discarded (line 8). It is worth noting that by setting $H_m = 0$ for all $m \in [M]$, DynLin-UCB reduces to LinUCB. The following sections provide the concentration of the estimator $\hat{\theta}_{t-1}$ of 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 estimator $\hat{\theta}_t$ of the cumulative Markov param-

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

[6] As an alternative, one can consider a more demanding requirement of the knowledge of a bound on the spectral norm $\|A\|_2$ of A. Similar assumptions regarding the knowledge of analogous quantities are considered in the literature, including decay of Markov operator norms (Simchowitz et al., 2020) and strong stability (Plevrakis & Hazan, 2020).

Algorithm 1 Dynamical Linear Upper Confidence Bound (DynLin-UCB).

Input: Regularization parameter $\lambda > 0$, exploration coefficients $(\beta_{t-1})_{t \in [\![T]\!]}$, spectral radius upper bound $\bar{\rho}$

Initialize $t \leftarrow 1$, $V_0 = \lambda I_d$, $b_0 = 0_d$, $\hat{h}_0 = 0_d$,

Define

$$M = \min\left\{M \in \mathbb{N} : \sum_{m=1}^{M-1} 1 + \lceil \log m / \log(1/\bar{\rho}) \rceil > T\right\} - 1$$

for $m \in \{1, \dots, M\}$ **do**

  Compute $u_t \in \arg\max_{u \in \mathcal{U}} \mathrm{UCB}_t(u)$: $\langle \hat{h}_{t-1}, u \rangle + \beta_{t-1} \|u\|_{V_{t-1}^{-1}}$

  Play arm u_t and observe y_t

  Define $H_m = \lceil \log m / \log(1/\bar{\rho}) \rceil$

  for $j \in \{1, \dots, H_m\}$ **do**

    Update $V_t = V_{t-1}$, $b_t = b_{t-1}$

    $t \leftarrow t + 1$

    Play arm $u_t = u_{t-1}$ and observe y_t

  end

  Update $V_t = V_{t-1} + u_t u_t^T$, $b_t = b_{t-1} + u_t y_t$

  Compute $\hat{h}_t = V_t^{-1} b_t$

  $t \leftarrow t + 1$

end

eter h, as performed in Algorithm 1. For every $m \in [\![M]\!]$, we denote with t_m the last round of epoch m: $t_0 = 0$ and $t_m = t_{m-1} + 1 + H_m$. At the end of each epoch $m \in [\![M]\!]$, we compute the Ridge–regularized linear regression problem:

$$\forall t \in [\![T]\!]: \quad \hat{h}_t = \arg\min_{h \in \mathbb{R}^d} \sum_{l \in [\![M]\!]: t_l \le t_m} (y_{t_l} - \langle h, u_{t_l} \rangle)^2 + \lambda \|h\|_2^2 = V_t^{-1} b_t.$$

We start by presenting 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 $(\hat{h}_t)_{t \in \mathbb{N}}$ be the sequence of solutions of the regression problems computed by Algorithm 1. Then, under Assumption 2.1 and 2.2, for every $\lambda \ge 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that:*

$$\forall t \in \mathbb{N}: \quad \|\hat{h}_t - h\|_{V_t} \le \frac{c_1}{\sqrt{\lambda}} \log(e(t+1)) + c_2\sqrt{\lambda} + \sqrt{2\tilde{\sigma}^2\left(\log\frac{1}{\delta} + \frac{1}{2}\log\frac{\det(V_t)}{\lambda^d}\right)},$$

where:

$$c_1 = U\rho(A)\frac{UB}{1-\rho(A)} + X,$$

$$c_2 = \Theta + \frac{B\rho(A)}{1-\rho(A)},$$

$$\tilde{\sigma}^2 = \sigma^2\left(1 + \frac{\Theta^2\rho(A)^2}{(1-\rho(A)^2)}\right).$$

First of all, it is worth noting that, when $\Theta = 0$ (i.e., $\omega = 0_n$) and there is no effect of the state on the reward, the bound perfectly reduces to the self-normalized concentration for the regression estimation used in linear bandits (Abbasi-Yadkori et al., 2011, Theorem 1). In particular, we recognize the second addendum which is the result of the regularization parameter $\lambda > 0$ and the third one which involves the subgaussianity constant $\tilde{\sigma}^2$, which is related to the joint contribution of the state and reward noises. Furthermore, the first addendum is an additional bias term and derives from the epochs of length $H_m + 1$. We observe that this term conveniently grows logarithmically with the round index t.

Finally, it is worth looking at our result from the more general 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. First note that, although we focus on a MISO system (y_t is a scalar, being our reward), extending our regression estimator to the multiple-output (MIMO) setting is straightforward. Second, the approach proposed in (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 a pretty convenient analysis for recovering the Markov parameters of the predictive form, but introduces significant challenges for recovering the parameters of the original system. Indeed, this requires an application of the Ho-Kalman method (Ho & Kálmán, 1966) which, however, does not preserve the concentration properties in the general case, but only for *persistently exciting* action signals. Our method, instead, forces to play an open-loop policy within a single epoch $m \in [\![M]\!]$ (with logarithmic duration), while the overall behavior is closed-loop.[7] 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 as the exploration coefficient β_t, based on the knowledge of the upper bounds $\bar{\rho} < 1$,

[7] Indeed, the next action to be played depends on the previous-epoch estimates.

$\bar{\ } \geq 8$, and those specified in Assumption 2.1:

$$\forall t \in [\![T]\!]: \quad \beta_t = \frac{\bar{c}_1}{\sqrt{d}} \log(e(t+1)) + \bar{c}_2 \sqrt{2\bar{\sigma}^2\left(\log\frac{1}{\delta} + \frac{d}{2}\log\left(1+\frac{tU^2}{d\lambda}\right)\right)}, \quad (5)$$

where:

$$\bar{c}_1 = U + \frac{UB}{1-\bar{\rho}}X, \qquad \bar{c}_2 = \frac{B}{1-\bar{\rho}}, \qquad \bar{\sigma}^2 = \sigma^2\left(1+\frac{2\bar{\rho}^2}{(1-\bar{\rho}^2)}\right).$$

The analysis poses additional challenges than that of Lin-UCB. Indeed, by a straightforward application of the proof strategy for Lin-UCB, it is possible, under Assumptions 2.1 and 2.2, comfortably obtain an expected offline regret bound, involving just the steady-state performances (details in Appendix A.2):[8]

$$\mathbb{E}[R^{\text{off}}(\text{DynLin-UCB}, T)] = \mathbb{E}\left[\sum_{t=1}^{T} J^* - J(u_t)\right] \leq \widetilde{O}\left(\frac{d\sqrt{T}}{(1-\bar{\rho})^{3/2}}\right). \quad (6)$$

However, when applying action u_t, the DLB does not immediately reach $J(u_t)$, as the system needs to converge to the steady-state according to its dynamics (Equation 2). Consequently, the expected online reward $\mathbb{E}[y_t]$ experiences a transitional phase. In order to obtain a sublinear expected online regret, we need to guarantee that during the transitional phase, when moving from one epoch $m \in [\![M]\!]$ to the next one $m+1$, performance does not degrade too much. This property can be guaranteed by the following assumption.

Assumption 3.1. For every action $u_1, u_2 \in \mathcal{U}$ it holds that:

$$\langle \theta, u_1\rangle + \langle B^T(I-A)^{-T}\omega, u_2\rangle \geq \min_{u \in \{u_1, u_2\}} J(u) = \langle \theta, u\rangle + \langle B^T(I-A)^{-T}\omega, u\rangle.$$

The rationale behind the assumption is the following. We need to guarantee that when we converge to the steady-state associated with action u_2, i.e., $\omega^T(I-A)^{-1}Bu_2$, and we apply action u_1, the instantaneous expected reward $\mathbb{E}[y_t]$ cannot take values below the minimum between the steady-state performances of actions u_1 and u_2, i.e., $J(u_1)$ and $J(u_2)$ respectively. In this way, the transitional phase does not have a relevant impact on the regret rate. It is worth noting that the assumption is surely fulfilled for linear bandits (i.e., $\omega = 0_n$) and strictly proper LTI systems (i.e., $\theta = 0_d$). Under this assumption, we are able to provide a bound on the expected online regret.

Theorem 3.2. Under Assumptions 2.1, 2.2, and 3.1, Algorithm 1 suffers an expected online regret bounded by:

$$\mathbb{E}[R(\text{DynLin-UCB}, T)] = \mathbb{E}\left[\sum_{t=1}^{T} J^* - y_t\right] \leq \widetilde{O}\left(\frac{(1+\|A\|_F)d\sqrt{T}}{(1-\bar{\rho})^{3/2}}\right).$$

Some remarks are in order. First of all, compared to the expected offline regret bound, we obtain a multiplicative factor that corresponds to the Frobenius norm $\|A\|_F$ of the dynamic matrix A. It is worth noting that the Frobenius norm is related to the spectral gap, since $\|A\|_F \leq \sqrt{n}\|A\|_2 \leq \sqrt{n}\,\sigma(A)\,\rho(A)$. Second, if the underlying problem does not have a dynamics, i.e., $\|A\|_F = 0$ and we choose $\bar{\rho} = 0$, we obtain a regret bound of order $\widetilde{O}(d\sqrt{T})$, which corresponds to the regret bound of LinUCB. Clearly, the dependence on $\bar{\rho}$ is relevant and with too large a value of $\bar{\rho}$ compared to the optimization horizon T (e.g., $\bar{\rho} = 1 - 1/T^{1/3}$) could make the regret degenerate to linear. This is a case in which the underlying system is as slow that the whole horizon T is insufficient to approximately reach steady state.

4. Related Works

Our formulation is placed at the intersection of three research areas: (i) bandits with delayed, aggregated, and composite feedback (Joulani et al., 2013), (ii) Partially-Observable Markov Decision Processes (POMDPs, Åström, 1965), and online control for Linear Time-Invariant (LTI) dynamical systems (Hespanha, 2018). In this section, we survey the related approaches in these areas in comparison with our model and algorithm.

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 in (Joulani et al., 2013) and widely exploited in the 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.,

[8] For interpretability reasons, we highlight the dependencies on T, $\bar{\rho}$, and d only and disregard other constant and the logarithmic terms.

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 $\mathcal{O}(\sqrt{T})$ 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 cases, $\mathcal{O}(\sqrt{T})$ regret for strongly convex functions and $\mathcal{O}(T^{2/3})$ for convex functions. Recently, $\mathcal{O}(\sqrt{T})$ regret rate has been obtained for convex functions too, by means of geometric exploration methods (Plevrakis & Hazan, 2020). Furthermore, (Lale et al., 2020a) reach $\mathcal{O}(\log T)$ regret in the case of strongly convex cost functions competing against the best *persistently exciting* controller (i.e., a controller that implicitly maintains a non-null exploration). Some approaches are designed to deal with adversarial noise (Simchowitz et al., 2020). All of these approaches, 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, especially in advertising domains. Our DynLin-UCB works with an arbitrary action space and, thanks to the linearity of the reward in the hidden state, does not require considering complex closed-loop controllers.

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 (Undurti & How, 2010; Kim et al., 2011; Isom et al., 2008) without exploiting the linearity and with no regret guarantees.

5. Numerical Simulations

In this section, we provide a numerical validation of DynLin-UCB in a synthetically generated domain. The goal of this simulation is to highlight the behavior of DynLin-UCB in comparison with bandit baselines, describing advantages and disadvantages. We start by introducing the DLB setting considered and the baselines for comparison, and then discussing the obtained results. The complete experimental results are reported in Appendix E.

Setting We consider a DLB defined by means of the following matrices $A = \mathrm{diag}(0.2, 0, 0.1)$, $B = \mathrm{diag}(0.25, 0, 0.1)$, $\omega = (0, 0.5, 0.1)^T$, $h = (1, 0, 0.1)^T$ and a Gaussian noise with $\sigma = 0.03$ (diagonal covariance matrix for the state noise[9]). This way the spectral gap of the dynamical matrix is $\rho(A) = 0.2$ and $\rho(A) = 1$. Moreover, the cumulative Markov parameter is given by $h = (0.5, 0.5, 0.1)^T$. We consider the action space $U = \{(u_1, u_2, u_3)^T \in [0, 1]^3$ 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 h, will invest the budget in the first component of the action, that is weighted by 0.5. Instead, a "far-sighted" agent, aware of the system evolution, will look at the cumulative Markov parameter h, realizing that the most convenient action is investing in the first component, weighted by 0.5. Therefore, the optimal action is $u^* = (1, 0.5, 0)^T$ leading to $J^* = 0.81$.

Baselines We consider as baselines Lin-UCB (Abbasi-Yadkori et al., 2011) designed for linear bandits, and Exp3 (Auer et al., 2002) usually employed in adversarial settings. Concerning the hyperparameter α for both DynLin-UCB and Lin-UCB, we tested $\alpha \in \{1, \log T\}$.[10] Since Exp3 requires finite actions, we consider a subset of the action space that surely contain the optimal action. Given that U is a polytope and since the objective is linear, we enumerate all the vertices of U, that, in this setting, can-

[9] 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 to the cumulative Markov parameter h and we could have obtained the same results with an equivalent (non-diagonal) representation, by applying an inevitable transformation T as $A' = TAT^{-1}$ and $\omega' = T\omega$, $h' = T^{-T}h$, and $B' = TB$.

[10] For DynLin-UCB, $\log T$ is a nearly optimal choice for α as it can be seen by looking at the first two addenda of the exploration factor in Equation (5).

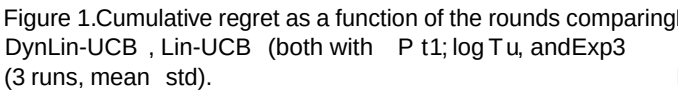


Figure 1. Cumulative regret as a function of the rounds comparing DynLin-UCB, Lin-UCB (both with $P \in \{1, \log T\}$), and Exp3 (3 runs, mean $\pm$ std).

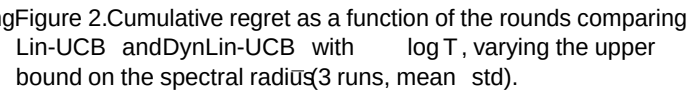


Figure 2. Cumulative regret as a function of the rounds comparing Lin-UCB and DynLin-UCB with $\bar{\rho} = \log T$, varying the upper bound on the spectral radius $\bar{\rho}$ (3 runs, mean $\pm$ std).

not be more than $d! = 6$. The learning rate for Exp3 is set as prescribed in the original paper (Auer et al., 2002).

Comparison with Lin-UCB and Exp3 Figure 1 shows the performance in terms of cumulative regret of DynLin-UCB, Lin-UCB and Exp3. The experiments are conducted over a time horizon of 500k rounds. For DynLin-UCB, we employed, for the sake of this experiment, the true value of the spectral gap, i.e., $\rho(A) = 0.2$. First of all, we observe that Exp3 suffers a significantly large cumulative regret. Moreover, both versions of Lin-UCB suffer linear regret. Indeed, even for a quite fast system ($\rho(A) = 0.2$), ignoring the system dynamics, and the presence of the hidden state, has made Lin-UCB committing to the sub-optimal (myopic) action $a = (0.5, 1, 0)^T$ with performance $J = 0.78 \pm$, with also a relevant variance. On the other hand, DynLin-UCB is able to maintain a smaller and stable (variance is negligible) sublinear regret in both its versions, with a notable advantage when using $\bar{\rho} = \log T$.

Sensitivity to the Choice of $\bar{\rho}$ The upper bound $\bar{\rho}$ of the spectral radius $\rho(A) = 0.2$ represents a crucial parameter of DynLin-UCB. While an overestimation $\bar{\rho} > \rho(A)$ does not compromise the regret rate, but tends to slow down the convergence process, a severe underestimation $\bar{\rho} < \rho(A)$ might prevent learning at all. In Figure 2, we test DynLin-UCB against a misspecification of $\bar{\rho}$, when $\bar{\rho} = \log T$. We can see that by considering $\bar{\rho} > 2\,\rho(A)$, DynLin-UCB experiences a larger regret but still sublinear and smaller w.r.t. Lin-UCB with $\bar{\rho} = \log T$. Even by reducing $\bar{\rho} \in \{0.1, 0.05\}$, DynLin-UCB is able to keep the regret sublinear, showing a remarkable robustness to misspecification. Clearly, when setting $\bar{\rho}$ to 0 makes the regret almost degenerates to linear.

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 a linear dynamics and by an observable noisy reward that linearly combines the hidden state and the played action. This model accounts for scenarios that cannot be easily represented by existing bandit models that consider delayed and aggregated feedback. On top of it, we have proposed a novel any-time optimistic regret minimization approach, DynLin-UCB, that, under suitable assumptions, is able to achieve sub-linear regret. The numerical simulation in a synthetic domains, succeeded in showing that, in a system where the baselines suffer linear regret, our algorithm enjoys sublinear regret. Furthermore, DynLin-UCB proved to be robust to misspecification of its most relevant hyper-parameter $\bar{\rho}$. To the best of our knowledge, this is the first work addressing this family of problems, characterized by a hidden linear dynamics, with a bandit-like approach. Short-term future directions include the study of the complexity of the problem by deriving regret lower bounds and understanding whether the considered assumptions (especially Assumption 3.1) and the knowledge of the upper bound $\bar{\rho}$ are actually unavoidable. Long-term future directions might focus on extending the present approach to non-linear system dynamics (e.g., to model saturation phenomena in the awareness) and non-stationary dynamics (e.g., to account for the natural market evolution).

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011, Budapest, Hungary*, pp. 1–26, 2011.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 2312–2320, 2011.

Agarwal, N., Hazan, E., and Singh, K. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 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. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

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, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, 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, KDD '14, pp. 1097–1105, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569.*

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

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

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

Ho, B. L. and Kálmán, 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 Fox, D. and Gomes, C. P. (eds.), *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 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, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 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 Walsh, T. (ed.), *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pp. 1968–1974. IJCAI/AAAI, 2011.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. Logarithmic regret bound in partially observable linear dynamical systems. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 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, ICML 2021, 18-24 July 2021, Virtual Event*, 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, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, 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.

Littman, M. L., Cassandra, A. R., and Kaelbling, L. P. Learning policies for partially observable environments: Scaling up. In *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pp. 362–370, 1995.

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, ICML 2020, 13-18 July 2020, Virtual Event*, pp. 3348–3356, 2020.

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10154–10164, 2019.

Oymak, S. and Ozay, N. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American Control Conference, ACC 2019, Philadelphia, PA, USA, July 10-12, 2019*, 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, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 4102–4110, 2018.

Plevrakis, O. and Hazan, E. Geometric exploration for online control. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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 Abernethy, J. D. and Agarwal, S. (eds.), *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, 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 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 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, CDC 2019, Nice, France, December 11-13, 2019*, pp. 3648–3654, 2019.

Undurti, A. and How, J. P. An online algorithm for constrained pomdps. In *IEEE International Conference on Robotics and Automation, ICRA 2010, Anchorage, Alaska, USA, 3-7 May 2010*, 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, UAI 2017, Sydney, Australia, August 11-15, 2017*, 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, ICML 2020, 13-18 July 2020, Virtual Event*, 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, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 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. ISSN 0022-247X.

A. Proofs and Derivations

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

A.1. Proofs of Section 2

Theorem 2.1 (Optimal Policy) Under Assumptions 2.1 and 2.2, an optimal policy maximizing the in nite-horizon expected average reward $J(\pi)$, as in Equation (2), is given by:

$$\forall t \in \mathbb{N},\ \forall H_{t-1} \in \mathcal{H}_{t-1}:\quad \pi_t(H_{t-1}) = u^* \tag{3}$$
$$\text{where}\quad u^* \in \arg\max_{u \in \mathcal{U}} \langle u, \theta^* x(h_{t-1}, u_{t-1}) \rangle.$$

Proof. Referring to the notation of Appendix B, we first observe that for every policy π we have $J(\pi) = \liminf_{H\to\infty} J_H(\pi)$, where $J_H(\pi) = \frac{1}{H}\mathbb{E}\left[\sum_{t=1}^{H} y_t\right]$ is the H-horizon expected average reward. Let us start with Equation (20), a xed nite $H \in \mathbb{N}$, and considering the sequence of actions $(u_1, u_2, \dots)$ generated by policy π:

$$\begin{aligned} J_H(\pi) &= \frac{1}{H}\sum_{s=1}^{H} \langle h_{0:H-s}, \theta^* \rangle + \frac{1}{H}\sum_{t=1}^{H} \langle A^{t-1} x_1, \theta^* \rangle \\ &= \frac{1}{H}\sum_{s=1}^{H} \langle h_{1:H-s}, \theta^* \rangle + \frac{1}{H}\sum_{s=1}^{H} \langle h_{H-s+1:H-1}, \mathbb{E}[u_s] \rangle + \frac{1}{H}\sum_{t=1}^{H} \langle A^{t-1} x_1, \theta^* \rangle. \end{aligned}$$

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

$$\begin{aligned} J_H(\pi) &\le \frac{1}{H}\sum_{s=1}^{H} \langle h_{1:H-s}, \theta^* \rangle + \frac{1}{H}\sum_{s=1}^{H} \|h_{H-s+1:H-1}\|_2 \, \|\mathbb{E}[u_s]\|_2 \\ &\quad + \frac{1}{H}\sum_{t=1}^{H} \langle A^{t-1} x_1, \theta^* \rangle =: J_H^{\uparrow}(\pi), \\ J_H(\pi) &\ge \frac{1}{H}\sum_{s=1}^{H} \langle h_{1:H-s}, \theta^* \rangle - \frac{1}{H}\sum_{s=1}^{H} \|h_{H-s+1:H-1}\|_2 \, \|\mathbb{E}[u_s]\|_2 \\ &\quad + \frac{1}{H}\sum_{t=1}^{H} \langle A^{t-1} x_1, \theta^* \rangle =: J_H^{\downarrow}(\pi). \end{aligned}$$

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

$$\begin{aligned} \|h_{H-s+1:H-1}\|_2 &\le \sqrt{\sum_{l=1}^{H-s-1} \left\| B^T (A^l)^T \theta^* \right\|_2^2} \\ &\le B \sqrt{\sum_{l=1}^{H-s-1} \rho(A)^l \rho(A)^l} \\ &\le B\, \rho(A)\, \frac{\rho(A)^{H-s}}{1-\rho(A)}. \end{aligned} \tag{7}$$

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

$$\frac{1}{H}\,\frac{B\,\rho(A)}{1-\rho(A)}\sum_{s=1}^{H} \rho(A)^{H-s} \le \frac{B\,\rho(A)\,(1-\rho(A)^H)}{H(1-\rho(A))^2}.$$

It is simple to observe that the last term approaches zero as $H \to \infty$. Moreover, with analogous argument, it can be proved that $\frac{1}{H}\sum_{t=1}^{H} \|A^{t-1} x_1\|_2 \to 0$ as $H \to \infty$. Thus, we have that $\liminf_{H\to\infty} J_H^{\downarrow}(\pi) \le \liminf_{H\to\infty} J_H^{\uparrow}(\pi)$.

Consequently, by the squeezing theorem of limits, we have:

$$J(\lambda) \le \liminf_{H\to\infty} J_H(\lambda) \le \limsup_{H\to\infty} J_H(\lambda)$$
$$\limsup_{H\to\infty} \frac{1}{H}\sum_{s=1}^{H} \langle x(h), E[u_s]\rangle \le h^T \limsup_{H\to\infty} \frac{1}{H}\sum_{s=1}^{H} E[u_s] .$$

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

A.2. Proofs of Section 3

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

$$\forall t \in \mathbb{N}: \quad \|\theta_t - h\|_{V_t} \le \frac{c_1}{2}\sqrt{\log\left(\frac{\det(V_t)}{\lambda^d}\right) - 1} + c_2 \sqrt{2R^2 \log\left(\frac{1}{\delta}\right) + \frac{1}{2}\log\left(\frac{\det(V_t)}{\lambda^d}\right)} + \sqrt{\lambda}\,;$$

where:

$$c_1 = \frac{UB}{1-\rho(A)}\,,$$
$$c_2 = \frac{B\,\rho(A)}{1-\rho(A)}\,,$$
$$R^2 = \sigma^2\left(1 + \frac{2\rho(A)^2}{(1-\rho(A)^2)}\right) .$$

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 and b_t):[11]

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

We now proceed at defining suitable filtrations. Let $(\mathcal{F}_t)_{t\in[T]}$ such that for every $t \ge 1$, the random variables $(u_1, y_1, \dots, u_{t-1}, y_{t-1}, u_t)$ are $\mathcal{F}_{t-1}$-measurable, i.e., $\mathcal{F}_{t-1} = \sigma(u_1, y_1, \dots, u_{t-1}, y_{t-1}, u_t)$. Let us also consider the filtration indexed by m, denoted with $\mathcal{G} = (\mathcal{G}_m)_{m\in[M]}$ and defined for all $m \in [M]$ as $\mathcal{G}_m = \mathcal{F}_{t_{m-1}}$. Thus, the random variables $\mathcal{G}_{m-1}$-measurable are those realized until the end of epoch $m-1$ except for y_{t_m}.

Since the estimates θ_t do not change within an epoch, we need to guarantee the statement for all rounds $t = t_{m-1}+1, \dots, t_m$ only. For these rounds, we define the following quantities:

$$y_m = y_{t_m},$$
$$u_m = u_{t_m}, \quad \text{(or any } u_l \text{ with } l \in [t_{m-1}+1, t_m] \text{ since they are all equal)}$$
$$r_m = \sum_{s=1}^{H_m} \left(\prod_{s=1}^{T} A^{s-1}\right) t_m - s,$$
$$x_{m-1} = x_{t_{m-1}},$$
$$\theta_m = \theta_{t_m},$$
$$V_m = V_{t_m},$$
$$b_m = b_{t_m} .$$

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

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

$$\begin{aligned}
\mathbb{E}\left[r_m \,|\, \mathcal{F}_{m-1}\right] &= \mathbb{E}\left[\sum_{s=1}^{H_m - 1} A^{s-1} x_{t_m + s} \,\middle|\, \mathcal{F}_{t_m - 1}\right] \\
&= \mathbb{E}\left[r_{t_m} \,|\, \mathcal{F}_{t_m - 1}\right] + \mathbb{E}\left[\sum_{s=1}^{H_m - 1} A^{s-1} \mathbb{E}\left[r_{t_m + s} \,|\, \mathcal{F}_{t_m + s - 1}\right] \,\middle|\, \mathcal{F}_{t_m - 1}\right] = 0,
\end{aligned}$$

since the system is operating by persisting the action after having decided it at the beginning of the epoch. Thus, by exploiting the decomposition in Equation (1), we can write:

$$\begin{aligned}
y_m &= y_{t_m} = x_{[0:H_m - 1]}^\top \theta_m + A^{H_m - 1} x_{t_{m-1}} + \sum_{s=1}^{H_m - 1} A^{s-1} x_{t_m + s} \\
&= x_{[0:H_m - 1]}^\top \theta_m + A^{H_m - 1} x_{m-1} + r_m \\
&= x_{[0:H_m - 1]}^\top \theta_m + x_{[H_m:2H_m - 1]}^\top \theta_m + A^{H_m - 1} x_{m-1} + r_m,
\end{aligned} \tag{8}$$

where we simply exploit the identity $x_{[0:H_m - 1]} = x_{[H_m:2H_m - 1]}$. We now introduce the following vectors and matrices:

$$\Phi_m = \begin{bmatrix} \phi_1^\top \\ \vdots \\ \phi_m^\top \end{bmatrix} \in \mathbb{R}^{m \times d}, \qquad y_m = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m,$$

$$r_m = \begin{bmatrix} r_1 \\ \vdots \\ r_m \end{bmatrix} \in \mathbb{R}^m, \qquad r_m = \begin{bmatrix} A^{H_1 - 2} x_0 \\ \vdots \\ A^{H_m - 2} x_{m-1} \end{bmatrix} \in \mathbb{R}^m,$$

$$g_m = \begin{bmatrix} x_{[H_1:2H_1 - 1]}^\top \theta_1 \\ \vdots \\ x_{[H_m:2H_m - 1]}^\top \theta_m \end{bmatrix} \in \mathbb{R}^m.$$

Using the vectors and matrices above, we observe that $V_m = I + \Phi_m^\top \Phi_m$ and $b_m = \Phi_m^\top y_m$. Furthermore, by exploiting Equation (8), we can write:

$$y_m = \Phi_m h + g_m + r_m + r_m.$$

Let us consider the estimate at $P_{[M]}$:

$$\begin{aligned}
\hat{h}_m = V_m^{-1} b_m &= \left(I + \Phi_m^\top \Phi_m\right)^{-1} \Phi_m^\top y_m \\
&= \left(I + \Phi_m^\top \Phi_m\right)^{-1} \Phi_m^\top \Phi_m h + g_m + r_m + r_m \\
&= h - \left(I + \Phi_m^\top \Phi_m\right)^{-1} h + \left(\Phi_m^\top \Phi_m\right)^{-1} g_m + \left(\Phi_m^\top \Phi_m\right)^{-1} r_m + \left(\Phi_m^\top \Phi_m\right)^{-1} r_m.
\end{aligned}$$

We now proceed at bounding the $\|\cdot\|_{V_m}$-norm, and exploit the triangle inequality:

$$\begin{aligned}
\|\hat{h}_m - h\|_{V_m} &\leq \underbrace{\left\|V_m^{-1} h\right\|_{V_m}}_{(a)} + \underbrace{\left\|V_m^{-1} \Phi_m^\top g_m\right\|_{V_m}}_{(b)} + \underbrace{\left\|V_m^{-1} \Phi_m^\top r_m\right\|_{V_m}}_{(c)} + \underbrace{\left\|V_m^{-1} \Phi_m^\top r_m\right\|_{V_m}}_{(d)}; \\
&\leq \|h\|_{V_m^{-1}} + \left\|\Phi_m^\top g_m\right\|_{V_m^{-1}} + \left\|\Phi_m^\top r_m\right\|_{V_m^{-1}} + \left\|\Phi_m^\top r_m\right\|_{V_m^{-1}};
\end{aligned}$$

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

$$\begin{aligned}(a)^2 &= \lambda^2 \|h\|^2_{\tilde{V}_{m-1}^{-1}} = \lambda^2 h^T \tilde{V}_{m-1}^{-1} h \\ &\le \lambda^2 \|\tilde{V}_{m-1}^{-1}\|_2 \|h\|_2^2 \\ &\le \lambda^2 \|h\|_2^2 \\ &\le \lambda^2 \frac{(B + \sqrt{A}q)^2}{(1 - \sqrt{A}q)^2},\end{aligned}$$

where we observed that $\|\tilde{V}_{m-1}^{-1}\|_2 \le \|\tilde{V}_m^{-1}\|_2 \le \lambda^{-1}$. Finally, we have bounded the norm of h:

$$\begin{aligned}\|h\|_2 &= \Big\| \sum_{s=0}^{\infty} A^{t-s} \Big\|_2^2 \\ &\le \sum_{s=0}^{\infty} \|A^{t-s}\|_2^2 \\ &\le \|A\|_2 \|A\|_2 \sum_{s=1}^{\infty} \|A^{s-1}\|_2^2 \\ &\le \frac{(B + \sqrt{A}q)^2}{(1 - \sqrt{A}q)^2},\end{aligned}$$

where we have exploited Assumptions 2.1 and 2.2.

We now move to term (b):

$$\begin{aligned}(b)^2 &= \|\tilde{U}_m^T g_m\|^2_{\tilde{V}_{m-1}} = g_m^T \tilde{U}_m \tilde{V}_{m-1}^{-1} \tilde{U}_m^T g_m \\ &\le \frac{1}{\lambda} \|g_m^T \tilde{U}_m\|_2^2 \\ &\le \frac{1}{\lambda} \Big\| \sum_{l=1}^{m} x(u_l; h^{H_l}, 2^{-8M}) y(u_l) \Big\|_2^2 \\ &\le \frac{1}{\lambda} \Big(\sum_{l=1}^{m} \|u_l\|_2 \, \|h^{H_l}, 2^{-8M}\|_2 \Big)^2 \\ &\le \frac{U^4 \, 2B^2 \, (1 + \sqrt{A}q)^2}{\lambda (1 - \sqrt{A}q)^2} \Big(\sum_{l=1}^{m} \sqrt{A}q^{H_l - 1} \Big)^2,\end{aligned}$$

where we have employed the following inequality:

$$\begin{aligned}\|h^{H_l}, 2^{-8M}\|_2 &\le \Big\| \sum_{j=H_l}^{T} A^{j-1} B \Big\|_2 \\ &\le \|B\|_2 \sum_{j=H_l}^{\infty} \|A^{j-1}\|_2 \\ &\le \|B\| \sqrt{A}q \frac{\sqrt{A}q^{H_l - 1}}{1 - \sqrt{A}q}.\end{aligned}$$

Let us now consider term (c):

$$(c)^2 \quad \left\| U_m^T r_m \right\|^2_{V_{m-1}} \quad r_m^T U_m V_m^{-1} U_m^T r_m$$

$$\leq \frac{1}{\lambda} \left\| U_m^T r_m \right\|_2^2$$

$$\frac{1}{\lambda} \left\| \sum_{s=1}^{m} \left(x_{l-1}^T A^{H_l - 1} x_{l-1} \right) u_l \right\|_2^2$$

$$\leq \frac{1}{\lambda} \left(\sum_{s=1}^{m} \| x_{l-1} \|_2 \, \| A^{H_l - 1} \|_2 \, \| x_{l-1} \|_2 \, \| u_l \|_2 \right)^2$$

$$\leq \frac{X^2 Z^2 U^2 \, p(A)^2}{\lambda} \left(\sum_{l=1}^{m} p(A)^{H_l - 1} \right)^2 .$$

We now bound the summations, exploiting the inequality $p(A) \leq \gamma$, holding by assumption:

$$\sum_{l=1}^{m} p(A)^{H_l - 1} \leq \sum_{l=1}^{m} p(A)^{\frac{\log l}{\log \frac{1}{\gamma}} - 1}$$

$$\leq \sum_{l=1}^{m} p(A)^{\frac{\log l}{\log \frac{1}{\gamma}}}$$

$$\leq \sum_{l=1}^{m} \exp\left(\frac{\log \frac{1}{p(A)}}{\log \frac{1}{\gamma}} \log l \right)$$

$$\leq \sum_{l=1}^{m} \frac{1}{l} \leq \log(m) + 1 \leq \log(t) + 1 \leq 1 + \log(e^t) + 1 \leq 2 \log(e^t).$$

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 η. For every $l \in [m]$ and R, and properly using the tower law of expectation:

$$\mathbb{E}\left[\exp\left(\eta_l \right) \middle| \mathcal{F}_{t_l - 1} \right] \quad \mathbb{E}\left[\exp\left(\sum_{s=1}^{H_m - 1} \left(x_{t_l - s - 1}^T A^s \right) \eta_{t_l - s} \right) \middle| \mathcal{F}_{t_l - 1} \right]$$

$$\mathbb{E}\left[\exp\left(\lambda \eta_{t_l} \right) \middle| \mathcal{F}_{t_l - 1} \right] \sum_{s=1}^{H_m - 1} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\left(x_{t_l - s - 1}^T A^s \right) \eta_{t_l - s} \right) \middle| \mathcal{F}_{t_l - 1 - s} \right] \middle| \mathcal{F}_{t_l - 1} \right]$$

$$\leq \exp\left(\frac{\lambda^2 R^2}{2} \right) \sum_{s=1}^{H_m - 1} \mathbb{E}\left[\exp\left(\frac{\lambda^2 \left\| x_{t_l - s - 1}^T A^s \right\|_2^2 R^2}{2} \right) \middle| \mathcal{F}_{t_l - 1} \right]$$

$$\leq \exp\left(\frac{\lambda^2 R^2}{2} \right) \sum_{s=1}^{H_m - 1} \exp\left(\frac{\lambda^2 R^2 \, p(A)^2 \, p(A)^{2(s-1)}}{2} \right)$$

$$\leq \exp\left(\frac{\lambda^2 R^2}{2} \right) \left(1 + \lambda^2 R^2 \, p(A)^2 \sum_{s=1}^{\infty} p(A)^{2(s-1)} \right)$$

$$\exp\left(\frac{\lambda^2 R^2}{2} \right) \left(1 + \frac{\lambda^2 R^2 \, p(A)^2}{(1 - p(A)^2)} \right) .$$

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

$$(d)^2 \quad \|U_m^T r_m\|^2_{V_{m-1}} \le 2\sigma^2 \left(1 - \frac{2\sigma^2}{(1-\alpha)^2\sigma^2 + \alpha^2}\right) \log\frac{1}{\delta} + \frac{1}{2}\log\frac{\det V_m}{\det V_d} :$$

□

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

Theorem A.1. Under Assumptions 2.1 and 2.2, having selected λ as in Equation (5), for every $\delta \in (0,1)$, with probability at least $1-\delta$ DynLin-UCB suffers an offline regret R^{off} bounded as:

$$R^{off}(\text{DynLin-UCB}, T) \le \sqrt{8dT\left(\frac{2}{T}+1\right)\left(1 + \frac{\log T}{\log\frac{1}{\delta}}\right)\log\left(1+\frac{TU^2}{d}\right)} :$$

Moreover, the expected offline regret ER^{off} is bounded as:

$$E\,R^{off}(\text{DynLin-UCB}, T) \le \mathcal{O}\left(\frac{d\sqrt{T}}{(1-\alpha)^{3/2}}\right)$$

Proof. For every epoch $m \in [M]$, let us define $r_{m-1} = t_{m-1}$ and define the confidence set $C_{m-1} = \{h \in \mathbb{R}^d : \|h - \hat{h}_{m-1}\|_{V_{m-1}} \le r_{m-1}\}$. Let us start by considering the instantaneous offline regret at epoch $m \in [M]$. Let $u \in \arg\max_{u \in U} x(h, u)$ and let $\tilde{h}_{m-1} \in C_{m-1}$ such that $UCB_{t_{m-1}+1}(u_m) = x(\tilde{h}_{m-1}, u_m)$. Thus, with probability at least $1-\delta$, we have:

$$\begin{aligned}
r_m &= J - J(u_m) = x(h, u) - x(h, u_m) \le x(\tilde{h}_{m-1}, u_m) - x(h, u_m) \\
&\le x(\tilde{h}_{m-1} - h, u_m) && (9)\\
&\le \|\tilde{h}_{m-1} - h\|_{V_{m-1}} \|u_m\|_{V_{m-1}^{-1}} \\
&\le \left(\|\tilde{h}_{m-1} - \hat{h}_{m-1}\|_{V_{m-1}} + \|\hat{h}_{m-1} - h\|_{V_{m-1}}\right)\|u_m\|_{V_{m-1}^{-1}} && (10)\\
&\le 2r_{m-1}\|u_m\|_{V_{m-1}^{-1}} : && (11)
\end{aligned}$$

where line (9) follows from the optimism, line (10) derives from triangle inequality, line (11) is obtained by observing that $h \in C_{m-1}$ with probability at least $1-\delta$, simultaneously for all $m \in [M]$, thanks to Theorem 3.1, having observed that r_{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:

$$R^{off}(\text{DynLin-UCB}, T) \le \sum_{m=1}^{M} (H_m - 1) r_m \le \sqrt{\sum_{m=1}^{M} (H_m - 1)} \sqrt{\sum_{m=1}^{M} r_m^2} :$$

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

$$\sum_{m=1}^{M} (H_m - 1) \le T(H_M - 1) \le T\left(1 + \frac{\log T}{\log\frac{1}{\delta}}\right) :$$

For the second summation, we following the usual derivation for linear bandits, recalling that $r_{m-1} \leq r_{M-1}$ for all $m \in [M]$:

$$\sum_{m=1}^{M} r_m^2 \leq 4 r_{M-1}^2 \sum_{m=1}^{M} \|u_m\|^2_{V_{m-1}^{-1}} \leq 8d r_{M-1}^2 \log\left(1 + \frac{MU^2}{d}\right) \leq 8d \tau_{T-1}^2 \log\left(1 + \frac{TU^2}{d}\right),$$

where the last passage follows from the elliptic potential lemma (Lattimore & Szepesvári, 2020, Lemma 19.4). Putting all together, we obtain:

$$R^{\text{off}}(\text{DynLin-UCB}; T) \leq \sqrt{8dT^{\tau_{T-1}^2}\left(1 + \frac{\log T}{\log \frac{1}{\lambda}}\right)\log\left(1 + \frac{TU^2}{d}\right)}.$$

We can also arrive to a problem-dependent regret bound, by setting $\sup_{u \in \mathcal{U}} \langle h(y_{x,h}), u \rangle = \langle y_{x,h}, u \rangle$. Since the instantaneous regret is either 0 or at least Δ, we have:

$$\begin{aligned} R^{\text{off}}(\text{DynLin-UCB}; T) &\leq \sum_{m=1}^{M} (H_m - 1)\frac{r_m^2}{\Delta} \\ &\leq \frac{H_M - 1}{\Delta} 8d r_{M-1}^2 \log\left(1 + \frac{MU^2}{d}\right) \\ &\leq \frac{8d}{\Delta}\left(1 + \frac{\log T}{\log \frac{1}{\lambda}}\right)\tau_{T-1}^2 \log\left(1 + \frac{TU^2}{d}\right). \end{aligned}$$

By setting $\lambda = 1 - 1/T$, we obtain the of ine regret in expectation, highlighting the dependence on T and d only:

$$\mathbb{E}\, R^{\text{off}}(\text{DynLin-UCB}; T) \leq \mathcal{O}\left(\frac{d\sqrt{T}}{(1-\rho(A))^{3/2}}\right),$$

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

The following lemma relates the expected of ine regret with the expected online regret.

Lemma A.2. Let $T \in \mathbb{N}$ be the optimization horizon. Then, under Assumptions 2.1, 2.2, 3.1, for Algorithm 1, it holds that:

$$\mathbb{E}\, R(\text{DynLin-UCB}; T) \leq \mathcal{O}\left(\frac{1}{(1-\rho(A))^2}\|A\|_F\right) \mathbb{E}\, R^{\text{off}}(\text{DynLin-UCB}; T).$$

Proof. First of all, we observe that for any policy, the cumulative effect of the noise components is zero-mean. Thus, it suf ces to consider the deterministic evolution of the system. Let $(u_1, \ldots, u_M)$ be the sequence of actions played by Algorithm 1 over the M epochs. We consider one epoch $m \in [2, M]$ at a time, starting from the second one. Let us consider the state that is reached at the end of epoch $m-1$, i.e., $x_{t_{m-1}}$ that can be expressed as:

$$x_{t_{m-1}} = A^{H_{m-1}-1} x_{t_{m-2}} + \sum_{l=1}^{H_{m-1}-1} A^{l-1} B u_{m-1}.$$

We compare this with the steady state that would have been reached by applying the same input, i.e., $\bar{x}_{t_{m-1}} = (I - A)^{-1} u_{m-1} + \sum_{l=1}^{\infty} A^{l-1} B u_{m-1}$:

$$\begin{aligned} x_{t_{m-1}} - \bar{x}_{t_{m-1}} &= A^{H_{m-1}-1} x_{t_{m-2}} - \sum_{l=H_{m-2}}^{\infty} A^{l-1} B u_{m-1} \\ &= X \rho(A)^{H_{m-1}-1} + B U \frac{\rho(A)^{H_{m-1}-1}}{1 - \rho(A)} \\ &= \frac{\rho(A)}{m-1} X + \frac{BU}{1-\rho(A)}, \end{aligned} \tag{12}$$

where we used the fact that $\|A^{H_{m-1}-1}\| \leq 1$ (proof, Lemma A.1). We now consider the evolution of the state of the system, starting from $\bar{x}_{t_{m-1}}$ and applying action u_m. To this end, we consider the steady state $\bar{x}_{t_m} = (I - A)^{-1} B u_m$ that we would obtain by constantly applying action u_m. Let $l \in [H_m - 1]$, we denote with $p_{t_{m-1}+l}$ the state reached after l rounds staring from state $\bar{x}_{t_{m-1}}$:

$$\begin{aligned} p_{t_{m-1}+l} &= A^l \bar{x}_{t_{m-1}} + \sum_{s=1}^{l} A^{s-1} B u_m \\ &= A^l \bar{x}_{t_{m-1}} + (I - A^l)(I - A)^{-1} B u_m \\ &= A^l \bar{x}_{t_{m-1}} + (I - A^l)\bar{x}_{t_m}, \end{aligned}$$

where we exploited Lemma A.3. Now, we sum over l, limited to epoch m:

$$\begin{aligned} \sum_{l=1}^{H_m-1} p_{t_{m-1}+l} &= \sum_{l=1}^{H_m-1} A^l \bar{x}_{t_{m-1}} + \sum_{l=1}^{H_m-1} A^l (I - A^l)\bar{x}_{t_m} \\ &= A(I - A^{H_m-1})(I - A)^{-1}(\bar{x}_{t_{m-1}} - \bar{x}_{t_m}) + (H_m - 1)\bar{x}_{t_m}. \end{aligned} \tag{13}$$

Thus, passing to the reward and averaging over the rounds of epoch m, we introduce the following quantity, representing the average reward in epoch m:

$$\bar{W}(u_m) := \frac{1}{H_m - 1} \sum_{l=1}^{H_m-1} \langle p_{t_{m-1}+l}, x \rangle \langle u_m, y \rangle + \frac{1}{H_m - 1} \sum_{l=1}^{H_m-1} \langle x', \bar{x}_{t_m} \rangle \langle u_m, y \rangle.$$

Let us also consider the sequence of states generated starting from state $x_{t_{m-1}}$ and applying action u_m:

$$x_{t_{m-1}+l} = A^l x_{t_{m-1}} + \sum_{s=1}^{l} A^{s-1} B u_m,$$

and the corresponding outputs, averaged over the rounds of epoch m:

$$W(u_m) := \frac{1}{H_m - 1} \sum_{l=1}^{H_m-1} \langle y_{t_{m-1}+l}, x \rangle \langle u_m, y \rangle + \frac{1}{H_m - 1} \sum_{l=1}^{H_m-1} \langle x', x_{t_{m-1}+l} \rangle \langle u_m, y \rangle.$$

Putting all together, and limiting to epoch m, we obtain:

$$J(u_m) = W(u_m) - \underbrace{\big(\bar{W}(u_m) - W(u_m)\big)}_{\text{(a)}} - \underbrace{\big(\bar{W}(u_m) - W(u_m)\big)}_{\text{(b)}}.$$

We now consider one term at a time and start with term (b):

$$\begin{aligned} \text{(b)} \quad \bar{W}(u_m) - W(u_m) &= \langle x, u_m \rangle \langle y, \frac{1}{H_m - 1} \sum_{l=1}^{H_m-1} \langle x', p_{t_{m-1}+l} \rangle \rangle - \langle x, u_m \rangle \langle y, \frac{1}{H_m - 1} \sum_{l=1}^{H_m-1} \langle x', x_{t_{m-1}+l} \rangle \rangle \\ &= \langle y, \frac{1}{H_m - 1} \sum_{l=1}^{H_m-1} \langle x', p_{t_{m-1}+l} - x_{t_{m-1}+l} \rangle \rangle \\ &= \langle y, \frac{1}{H_m - 1} \Big\langle x', \sum_{l=1}^{H_m-1} A^l (\bar{x}_{t_{m-1}} - x_{t_{m-1}}) \Big\rangle \rangle \\ &\leq \frac{\|A\|^2 - \|A\|}{H_m - 1} \cdot \frac{\|B\| U}{1 - \|A\|}, \end{aligned}$$

having bounded $|\omega^T A^l (\bar{x}_{t_{m-1}} - x_{t_{m-1}})| \le \rho(A)^l \rho(A) \|\bar{x}_{t_{m-1}} - x_{t_{m-1}}\|_2$ and, further, by setting $l = 1$ and exploiting Equation (12) for bounding the norm. We now move to term (a), and exploit Equation (13):

$$(a) = J(u_m) - \widetilde{W}(u_m) = \langle \theta, u_m\rangle + \langle \omega, \bar{x}_{t_m}\rangle - \langle \theta, u_m\rangle - \frac{1}{H_m - 1}\sum_{l=1}^{H_m-1} \langle \omega, x_{t_{m-1}+l}\rangle$$

$$= \frac{1}{H_m - 1}\omega^T A (I - A)^{-1}(I - A^{H_m-1})(\bar{x}_{t_{m-1}} - \bar{x}_{t_m}).$$

Let us rename $R = A(I - A)^{-1}(I - A^{H_m-1})$ and recall that the involved quantity is a scalar:

$$\omega^T R (\bar{x}_{t_{m-1}} - \bar{x}_{t_m}) = \operatorname{tr}\left(\omega^T R (\bar{x}_{t_{m-1}} - \bar{x}_{t_m})\right)$$
$$= \operatorname{tr}\left(R (\bar{x}_{t_{m-1}} - \bar{x}_{t_m})\omega^T\right)$$
$$\le \|R\|_F \left\|(\bar{x}_{t_{m-1}} - \bar{x}_{t_m})\omega^T\right\|_F,$$

having exploited the Cauchy-Schwarz inequality over the inner product space of matrices $\operatorname{tr}(XY^T) \le \|X\|_F \|Y\|_F$. For the first norm, we obtain:

$$\|R\|_F \le \|A\|_F \left\|(I - A)^{-1}(I - A^{H_m-1})\right\|_2 = \|A\|_F \left\|\sum_{l=0}^{H_m-2} A^l\right\|_2 \le (H_m - 1)\rho(A)\|A\|_F,$$

using the inequality between matrix norms $\|XY\|_F \le \|X\|_F \|Y\|_2$ and bounding $\left\|\sum_{l=0}^{H_m-2} A^l\right\|_2 \le \rho(A) \sum_{l=0}^{H_m-2} \rho(A)^l \le (H_m - 1)\rho(A)$. For the second term, recalling that $(\bar{x}_{t_{m-1}} - \bar{x}_{t_m})\omega^T$ has rank 1, we obtain:

$$\left\|(\bar{x}_{t_{m-1}} - \bar{x}_{t_m})\omega^T\right\|_F = |\omega^T (\bar{x}_{t_{m-1}} - \bar{x}_{t_m})|$$
$$= |\omega^T (\bar{x}_{t_{m-1}} - \bar{x}_{t_m}) + \theta^T u_m - \mu^* + \mu^* - \theta^T u_m|$$
$$\le |\mu^* - \theta^T u_m - \omega^T \bar{x}_{t_{m-1}} - (\mu^* - (\theta^T u_m + \omega^T \bar{x}_{t_m}))|$$
$$\le \max\{r_m, r_{m-1}\} + r_m$$
$$\le 2(r_{m-1} + r_m),$$

where we used Assumption 3.1 for the first inequality. Putting all together, we have:

$$\mathbb{E} R(\text{DynLin-UCB}, T) = \sum_{t=1}^{T} \mu^* - y_t \tag{14}$$
$$= (\mu^* - y_1) + \sum_{m=2}^{M} (H_m - 1)(\mu^* - W(u_m))$$
$$= \mu^* + U + X + \sum_{m=2}^{M} (H_m - 1)(\mu^* - J(u_m)) \tag{15}$$
$$+ \sum_{m=2}^{M} (H_m - 1)(J(u_m) - \widetilde{W}(u_m)) + \sum_{m=2}^{M} (H_m - 1)(\widetilde{W}(u_m) - W(u_m))$$
$$\le \mu^* + U + X + \sum_{m=2}^{M} (H_m - 1) r_m$$
$$+ 2\rho(A)\|A\|_F \sum_{m=2}^{M} (r_{m-1} + r_m) + \rho(A)^2 \rho(A) \sum_{m=2}^{M} \frac{H_m - 1}{m - 1}; \tag{16}$$
$$\le \mu^* + U + X + \rho(A)^2 \rho(A)(H_M - 1)\log(eM)$$
$$+ (1 + 2\rho(A)\|A\|_F)\, \mathbb{E} R^{\text{off}}(T, \text{DynLin-UCB}), \tag{17}$$

where in line (15) we exploited the inequality $\mu^* - y_1 \le \mu^* + U + X$, we used the bounds on (a) and (b) in line (16), we used the definition of offline regret and bounded the harmonic summation in line (17). □

Theorem 3.2. Under Assumptions 2.1, 2.2, and 3.1, Algorithm 1 suffers an expected online regret bounded by:

$$\mathrm{E}\,R_T(\text{DynLin-UCB}) = \mathrm{E}\sum_{t=1}^{T} J^\star - y_t \leq \mathcal{O}\left(\frac{(1+\|A\|_F)\sqrt{d}\,T}{(1-\bar{\rho})^{3/2}}\right).$$

Proof. We simply combine Theorem A.1 and Lemma A.2. □

A.3. Technical Lemmas

Lemma A.3. Let $A \in \mathbb{R}^{n\times n}$. Then, for every $m \in \mathbb{N}$, it holds that:

$$\sum_{i=0}^{m-1} A^i = (I - A)^{-1}(I - A^m) = (I - A^m)(I - A)^{-1}.$$

Proof. Let us simply write explicitly the expressions:

$$(I - A)^{-1}(I - A^m) = \sum_{i=0}^{\infty} A^i - (I - A^m)\sum_{i=0}^{\infty} A^i = \sum_{i=0}^{\infty} A^i - \sum_{i=m}^{\infty} A^i = \sum_{i=0}^{m-1} A^i.$$

An analogous derivation holds for the second expression. □

B. 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 2.2, 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:

$$J_H(\underline{u}) := \mathrm{E}\,\frac{1}{H}\sum_{t=1}^{H} y_t \quad \text{where} \quad \begin{cases} x_{t+1} = Ax_t + Bu_t + \eta_t \\ y_t = x_t^\top x_t + u_t^\top u_t + \nu_t \end{cases}, \quad t \in [H], \tag{18}$$

where the expectation is taken w.r.t. the randomness of the state noise η_t and the reward noise ν_t. We now show that the optimal policy for the finite-horizon setting is a non-stationary open-loop policy.

Theorem B.1 (Optimal Policy for the H–Horizon Setting). If $H \in \mathbb{N}$, an optimal policy $\underline{u}_H = (u_{H,t})_{t\in[H]}$ maximizing the H-horizon expected average reward $J_H(\underline{u})$ as in Equation (18) is given by:

$$u_{H,t} = -K_{H-t-1}x_{H-t-1}, \quad K_{H-t-1} \in \arg\min_{K\in\mathbb{R}^{d\times n}} \left\|A - BK\right\|_2, \quad x_{H-t-1} = \left(A - BK_{H-t-2}\right)x_{H-t-2} + \eta_{H-t-2}.$$

Proof. We start by expressing for every $t \in [H]$ the reward y_t as a function of the sequence of actions $\underline{u} = (u_1, \ldots, u_H)$ produced by a generic policy $\underline{u}$. By exploiting Equation (3) instanced with $H = t - 1$, we have:

$$y_t = \sum_{s=0}^{t-1} x(t-s, u_{t-s}) + \eta_t^\top A^{t-1}x_1 + \sum_{s=1}^{t-1} \eta_{t-s}^\top A^{s-1}\eta_{t-s}.$$

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

$$\mathrm{E}\,y_t = \sum_{s=0}^{t-1} x(t-s, \mathrm{E}\,u_{t-s}) + \eta_t^\top A^{t-1}\,\mathrm{E}\,x_1.$$

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

$$
\begin{aligned}
J_H(\underline{\theta}) &= \frac{1}{H}\sum_{t=1}^{H} \mathbb{E}[r_t] \\
&= \frac{1}{H}\sum_{t=1}^{H}\sum_{s=0}^{t-1} x(h^{t-s-1}, \mathbb{E}[u_{t-s}])^\top \theta + \frac{1}{H}\sum_{t=1}^{H} (\mathbf{1}^\top A^{t-1})\, \mathbb{E}[x_1] \\
&= \frac{1}{H}\sum_{s=1}^{H}\sum_{t=s}^{H} h^{t-s} \mathbb{E}[u_s]^\top \theta + \frac{1}{H}\sum_{t=1}^{H} (\mathbf{1}^\top A^{t-1})\, \mathbb{E}[x_1] \qquad (19) \\
&= \frac{1}{H}\sum_{s=1}^{H} x(h^{[0:H-s]}, \mathbb{E}[u_s])^\top \theta + \frac{1}{H}\sum_{t=1}^{H} (\mathbf{1}^\top A^{t-1})\, \mathbb{E}[x_1]. \qquad (20)
\end{aligned}
$$

where line (19) is obtained by renaming the indexes of the summations, and (20) comes from the definition of cumulative Markov parameter $h^{[0:H-s]}$. It is now simple to see, as no noise is present in the expression, that the performance $J_H(\underline{\theta})$ is maximized by taking at each round $s \in [H]$ an action $u_s = \pi_s(H-s+1)$ such that whose expectation satisfies $\mathbb{E}[u_s] \in \arg\max_{\mathbb{E}[u_s]} x(h^{[0:H-s]}, \mathbb{E}[u_s])^\top \theta$. Clearly, we can take the deterministic action such that $u_s = \mathbb{E}[u_s]$.

□

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 B.2. Let $H \in \mathbb{N}$. Then, for every policy $\underline{\theta}$ it holds that:

$$
|J_H(\underline{\theta}) - J(\underline{\theta})| \le \frac{B\,U\,\rho(A)\,(1-\rho(A)^H)}{H(1-\rho(A))}.
$$

Proof. Consider two horizons H and $H-1 \in \mathbb{N}$, and let $(u_1, u_2, \dots)$ be the sequence of actions played by policy $\underline{\theta}$. Using Equation (20), we have:

$$
\begin{aligned}
J_H(\underline{\theta}) - J_{H-1}(\underline{\theta}) &= \frac{1}{H}\sum_{s=1}^{H} x(h^{[0:H-s]}, \mathbb{E}[u_s])^\top \theta - \frac{1}{H-1}\sum_{s=1}^{H-1} x(h^{[0:H-1-s]}, \mathbb{E}[u_s])^\top \theta \qquad (21) \\
&= \frac{1}{H}\sum_{s=1}^{H} x(h^{[0:H-s]}, \mathbb{E}[u_s])^\top h - \frac{1}{H-1}\sum_{s=1}^{H-1} x(h^{[0:H-1-s]}, \mathbb{E}[u_s])^\top h \qquad (22) \\
&= \frac{1}{H}\sum_{s=1}^{H} x(h^{[H-s-1:8]}, \mathbb{E}[u_s])^\top \theta - \frac{1}{H-1}\sum_{s=1}^{H-1} x(h^{[H-1-s-1:8]}, \mathbb{E}[u_s])^\top \theta. \qquad (23)
\end{aligned}
$$

As shown in Appendix A.1, we have that the second addendum vanishes as $H-1$ approaches ∞:

$$
\frac{1}{H-1}\sum_{s=1}^{H-1} x(h^{[H-1-s-1:8]}, \mathbb{E}[u_s])^\top \theta \to 0 \qquad \text{when} \qquad H-1 \to \infty.
$$

Concerning the first addendum, we have:

$$
\begin{aligned}
\frac{1}{H}\sum_{s=1}^{H} x(h^{[H-s-1:8]}, \mathbb{E}[u_s])^\top \theta &\le \frac{U}{H}\sum_{s=1}^{H} \|h^{[H-s-1:8]}\|_2 \\
&\le \frac{B\,U\,\rho(A)}{H}\sum_{s=1}^{H} \rho(A)^{H-s} \\
&\le \frac{B\,U\,\rho(A)\,(1-\rho(A)^H)}{H(1-\rho(A))}.
\end{aligned}
$$

□

C. System Identification

This section presents a solution to identify matrices A, B, C, and D characterizing a Linear Dynamical System starting from a single trajectory. We adopt a variant of the Ho-Kalman (Ho & Kálmán, 1966) algorithm. We start from the identification method proposed by Lale et al. (Lale et al., 2020a, Section 3), where authors consider a system of the type:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + \eta_t; \\ y_t &= Cx_t + z_t: \end{aligned} \tag{24}$$

Our setting can be seen as:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + \eta_t; \\ y_t &= Cx_t + Du_t + z_t; \end{aligned} \tag{25}$$

with $x_t; \eta_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^p$, and $y_t; z_t \in \mathbb{R}^m$. The noise over state transition model η_t and output z_t are σ^2-subgaussian random variables.

We consider in this part the standard control problem notation adopted for Linear Dynamical Systems. The mapping to our problem is straightforward by considering $C = v_m^\top$ and $D = v_p^\top$ ($v_m = 1_m$ and $v_p = 1_p$).

In predictive form, the system described in Equation (24) is:

$$\begin{aligned} x_{t+1} &= \bar{A}x_t + Bu_t + Fy_t; \\ y_t &= Cx_t + e_t; \end{aligned}$$

where:

$$\begin{aligned} \bar{A} &= A + FC; \\ F &= -\left(APC^\top + \sigma^2 I\right)\left(CPC^\top + \sigma^2 I\right)^{-1}; \end{aligned}$$

and P is the solution to the following DARE (Discrete Algebraic Riccati Equation):

$$APA^\top - APC^\top\left(CPC^\top + \sigma^2 I\right)^{-1}CPA^\top + \sigma^2 I = P:$$

In order to identify this Linear Dynamical System, we want to detect a matrix G_y:

$$G_y = \begin{bmatrix} CF & CAF & \cdots & CA^{H-1}F & CB & CAB & \cdots & CA^{H-1}B \end{bmatrix}: \tag{26}$$

To identify through least squares method matrix G_y, we construct for each t, a vector r_t:

$$r_t = \begin{bmatrix} y_{t-1}^\top & \cdots & y_{t-H}^\top & u_{t-1}^\top & \cdots & u_{t-H}^\top \end{bmatrix}^\top \in \mathbb{R}^{pm+pq+H}: \tag{27}$$

The system output y_t can be rewritten as:

$$y_t = G_y r_t + e_t + CA^H x_{t-H}:$$

The output of the system under analysis (Equation 25) is:

$$\begin{aligned} y(t+q) &= y_t + Du_t \\ &= G_y r_t + Du_t + e_t + CA^H x_{t-H} \end{aligned}$$

We can incorporate the contribution of Du_t in G_y obtaining G_y:

$$G_y = \begin{bmatrix} CF & CAF & \cdots & CA^{H-1}F & D & CB & CAB & \cdots & CA^{H-1}B \end{bmatrix}:$$

The related vector $_t$ is:

$$_t = \begin{bmatrix} y_{t-1}^\top & \cdots & y_{t-H}^\top & u_t^\top & u_{t-1}^\top & \cdots & u_{t-H}^\top \end{bmatrix}^\top \in \mathbb{R}^{pm+pq+H+p}: \tag{28}$$

The best value of G_y can be found through regularized least squares as in (Lale et al., 2020a, Equation 10):

$$\hat{\Theta}_y \leftarrow \arg\min_X \|X\|_F^2 + \sum_t^t \|y - X\|_2^2; \quad (29)$$

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

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

E. Experimental Results

In this section, we present two more contributions of the experimental evaluation provided in the main paper. First, we construct a simulator by generalizing real-world data using the variant of the Ho-Kalman method presented in Appendix C, and we test DynLin-UCB compared with Lin-UCB, Exp3 and the Expert Policy (i.e., the policy applied by human experts). Second, we study the behavior of DynLin-UCB, Lin-UCB and Exp3 for different magnitude of noises in both the state transition model and the output for the setting presented in the main paper.

DynLin-UCB is considered in two version, depending on the choice of the hyper-parameter α. We consider the case with $\alpha = 1$ (referred as DynLin-UCB (1)), and $\alpha = \log(T)$ (referred as DynLin-UCB (log T)), where T is the total number of steps performed in the experiment.

Baselines In this section, the quality of DynLin-UCB is compared with several baselines. First, we compare our solution with Lin-UCB algorithm, which is a particular case of our solution, that, given the nature of the problem (output is a linear combination of the actions plus stochastic noise), is a straightforward baseline. As already done in Section 5, we consider two version of Lin-UCB, one with hyper-parameter $\alpha = 1$ (referred as Lin-UCB (1)), and the other with $\alpha = \log(T)$ (referred as Lin-UCB (log T)), Second, we compare our algorithm with Exp3 (Auer et al., 2002), an algorithm designed for adversarial bandits.[12] Lastly, only in the case of real-world data, we compare our solution with the human-expert policy (referred as Expert Policy) which does not consider the interactions and delays explicitly. This policy is static, directly generalized from the original dataset by learning the average budget allocation over all platforms from the available data. If such an allocation does not belong to the action space, a projection into U is performed.

E.1. Real-world Data

In this section, we present an experimental evaluation based on real-world data coming 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 & Kálmán, 1966).[13] We used the matrices estimated with Ho-Kalman to build up a simulator and compare the performances of DynLin-UCB and the baselines in a controlled environment. The resulting system has $\rho(A) = 0.67$. To assert the quality of the solution, we evaluate DynLin-UCB in comparison with the baselines presented above for 10^6 steps over 10 runs.

Results Figure 3 gives an overview of the results of this simulation. More in details, Figure 3a shows the performance of each algorithm in terms of cumulative regret, while Figure 3b shows the instantaneous reward. It is worth noting how neither Lin-UCB nor Exp3 are able to converge to the optimal choice. Indeed, they immediately converge to a sub-optimal solution and persist in it. DynLin-UCB, instead, shows a convergence trend towards the optimal policy over time in both the setting ($\alpha = 1$ and $\alpha = \log(T)$), even if the best solution is the one which employs $\alpha = \log(T)$. The Expert Policy which tends to consider instantaneous effect only and does not take into account correlations between platforms is also sub-optimal.

[12] It is worth noting that in the adversarial setting the regret is defined by comparing the learner's behavior with the best fixed action. However, the regret guarantees of Exp3 are fully meaningful for the non-adaptive adversaries only to which our setting cannot be reduced to, since the hidden state evolution (and so the reward) depends on the actual played actions.

[13] The complete method used to retrieve the best model generalizing data is described in Appendix C.

(a) Cumulative Regret (b) Immediate Reward

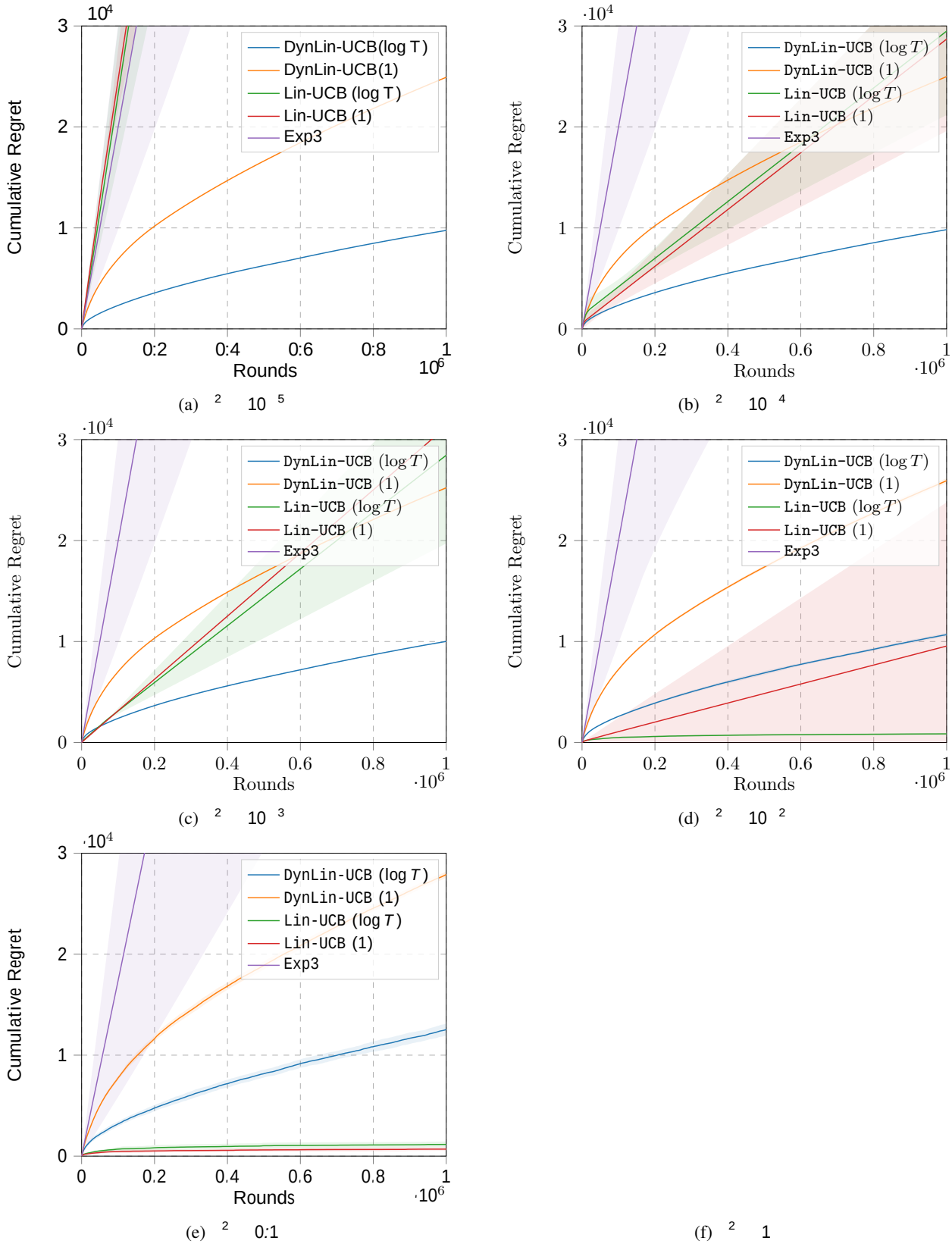
Figure 3. Performance of DynLin-UCB , Lin-UCB and Exp3 , and the Expert Policy in the system generalized from real-world data. (10 runs, mean std)

E.2. Noise Effect

In this experiment, we evaluate the performance of DynLin-UCB and the other bandit baselines described above at different values of noise. The evaluation is performed over 10 runs considering the setting presented in Section 5 and lasts for $T = 10^6$ time steps. The analysis considers noisy state transition model and output subject to different magnitude of noise. The noise in this simulation is a zero-mean Gaussian noise with variance $^2 \in \{10^{-5}; 10^{-4}; 10^{-3}; 10^{-2}; 10^{-1}; 1\}$.

Results Figure 4 shows the results of the experiment for the different values of ^{2}. It is clearly visible how DynLin-UCB performs in almost the same way no matter the noise at which the system is subject to, leading always to sub-linear regret. On the other hand, Lin-UCB regret is different in every simulation we perform. Indeed, with the a low level of noise reaches linear regret and does not converge (showing linear regret), while for large values of noise it converges very quickly. This is due to the nature of the con dence bound of Linear bandits that is not able to take into account such a complex scenario and lead to no guarantees in this setting. Exp3 is not able to reach the optimum in this scenario, independently from the value of the noise^2.

Infrastructure and Computational Time The code used for the results provided in this section has been run on a Intel(R) I5(R) 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. The experiment performed over real-world data takes overall 90 minutes to to run all the algorithms and perform 10 runs over 10^6 rounds each. The experiment to evaluate the performance of DynLin-UCB and the other bandit baselines takes overall 8 hours to perform the 6 experiments of 10 runs each over 10^6 rounds. It is worth noting that the time complexity of DynLin-UCB is upper-bounded by the one of Lin-UCB .



(a) ^{2} 10 ^{5} (b) ^{2} 10 ^{4}

(c) ^{2} 10 ^{3} (d) ^{2} 10 ^{2}

(e) ^{2} 0:1 (f) ^{2} 1

Figure 4. Performance of DynLin-UCB, Lin-UCB and Exp3 at different values of ^{2}. (10 runs, mean std)