Block Optimism for Nonstationary Bandits with Latent Linear Dynamics

Taehyun Hwang **Hyunjun Choi** **Heesang Ann** **Min-hwan Oh**

Seoul National University
Seoul, Republic of Korea
{th.hwang, nschj1, sang3798, minoh}@snu.ac.kr

Abstract

We study an endogenous nonstationary stochastic bandit problem with latent linear dynamics, where actions affect both immediate rewards and the future evolution of an unobserved latent state. Rewards are bilinear in the current action and latent state, inducing history-dependent rewards and a nontrivial long-horizon planning problem. The existing explore-then-commit approach achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret by uniformly exploring to estimate the latent dynamics and then committing to an optimized open-loop action sequence. We show that this rate can be improved via adaptive block-level optimism. Our key step is a cyclic approximation: under stable dynamics, the infinite-memory reward process can be truncated, and the open-loop benchmark can be approximated by optimizing a finite-memory block-level proxy. Building on this reduction, we propose a UCB-based block algorithm that maintains confidence sets for the truncated dynamics parameters and selects blocks optimistically. We prove a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{T})$, significantly improving over the previous $\widetilde{\mathcal{O}}(T^{2/3})$ guarantee for the same model. To the best of our knowledge, this is the first $\widetilde{\mathcal{O}}(\sqrt{T})$ regret guarantee for latent linear-dynamics bandits with bilinear reward observations and an open-loop action-sequence benchmark.

1 Introduction

Many sequential decision-making problems exhibit *endogenous nonstationarity*: the learner's actions affect not only immediate rewards but also the future state of the environment. This phenomenon naturally arises in recommender systems, online advertising, and personalized interventions, where repeated recommendations or interventions can change a user's latent preferences and thereby alter future rewards [18, 22, 21, 33]. While classical stochastic bandits assume fixed reward distributions, much of the nonstationary bandit literature models nonstationarity as exogenous, for example through variation budgets, change points, or drifting parameters [7, 10, 31, 19]. Other action-induced models, such as rotting, recharging, blocking, rebounding, and delay-dependent bandits [27, 34, 20, 6, 26, 8], capture temporal effects at the arm level, but are less suited for structured action spaces where one action can affect the future rewards of many related actions.

A recent work introduces a latent linear-dynamics bandit model in which actions drive an unobserved latent state with linear dynamics and rewards are bilinear in the current action and latent state [11]. This model captures cross-action endogenous effects through a latent state representation, but creates a coupled learning-and-planning problem: the learner must infer the latent input-output dynamics from scalar, temporally correlated rewards while choosing actions whose effects persist over the horizon. The prior explore-then-commit (ETC) algorithm separates these tasks by first estimating the latent dynamics from random exploration and then committing to an optimized open-loop action

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

sequence, yielding a $\widetilde{\mathcal{O}}(T^{2/3})$ regret guarantee [11]. This phase separation leads to a suboptimal exploration–exploitation trade-off, since data collected during exploration are not used for reward maximization, while observations obtained after commitment are not used to revise the plan. This raises a natural question:

Can we design an adaptive exploration strategy for latent linear-dynamics bandits that interleaves learning and planning and achieves sharper regret?

We answer this question affirmatively by exploiting the stability of the latent dynamics. In latent linear-dynamics bandits, past actions can continue to influence future rewards through the evolving latent state, so the induced reward process generally depends on the full action history. Under stable dynamics, however, the influence of far-past actions decays geometrically, allowing this infinite-memory reward process to be approximated by a finite-memory model. This observation enables a cyclic approximation that reduces the original long-horizon planning problem to a stationary block-level learning problem. In this block-level problem, the proxy reward is approximately linear in a truncated representation of the latent dynamics. Building on this reduction, we develop an adaptive optimistic learning strategy that updates its estimate at the end of each block and selects subsequent blocks optimistically. This interleaves learning and planning throughout the horizon and yields a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{T})$, significantly improving over the $\widetilde{\mathcal{O}}(T^{2/3})$ ETC guarantee for the same model. Our main contributions are summarized as follows:

- We first show that, under stable latent dynamics, the influence of far-past actions decays geometrically, allowing the infinite-memory reward process to be approximated by truncating its dependence to a finite window of recent actions. Building on this approximation, we extend the cyclic approximation principle for finite-memory bandits [12, 5] to our latent linear-dynamics setting, showing that the optimal open-loop benchmark can still be approximated by a cyclic block sequence despite the persistent influence of past actions through the latent state.
- Building on this block-level reduction, we propose a block-level UCB algorithm that adaptively explores the latent dynamics while exploiting the current estimate. Unlike the previous ETC method [11], our algorithm does not require a long initial phase of uniform exploration. Instead, it maintains confidence sets for the truncated dynamics parameters, selects each block optimistically, and updates its estimate after every block, so that every block contributes simultaneously to exploration and exploitation.
- We prove that the proposed algorithm achieves a $\widetilde{\mathcal{O}}(\sqrt{T})$-order regret bound. We decompose each block-level reward observation into a linear signal and residual terms. The main challenge is that one residual component contains pre-block process noises and can therefore be correlated with the adaptively selected block, so it cannot be treated as standard martingale noise. We address this dependence by establishing a uniform bound on the corresponding vector-valued tail, which yields valid confidence sets under adaptive block selection and enables the optimism-based regret analysis. This significantly improves over the previous $\widetilde{\mathcal{O}}(T^{2/3})$ ETC guarantee for the same latent linear-dynamics bandit model [11]. To the best of our knowledge, this is the first $\widetilde{\mathcal{O}}(\sqrt{T})$-order regret guarantee for latent linear-dynamics bandits with bilinear reward observations.
- We provide numerical experiments demonstrating that the proposed block-level optimistic algorithm performs favorably against baselines. The results illustrate the empirical benefit of adaptive block-level exploration.

1.1 Related Work

Nonstationary bandits. One common approach models nonstationarity as *exogenous*, for instance through variation budgets [7], switching or drifting reward parameters [10, 31], or nonstationary stochastic linear bandit models [19]. These models capture environments whose reward functions change over time, but the changes are typically not caused by the learner's own actions.

Several lines of work study action-induced endogenous nonstationarity. Rotting bandits [27, 34], recharging bandits [20], blocking bandits [6], rebounding bandits [26], and bandits with delay-dependent payoffs [8] model temporal effects such as satiation, recovery, blocking, or delay-dependent

rewards. These models typically describe how the reward of an individual arm changes as a function of its own past usage or delay. Another line of work studies bandit models with finite memory, where the reward at each round depends on only a bounded window of past actions [12, 5]. In these models, the memory length is specified as part of the problem, so the influence of sufficiently distant actions vanishes by construction.

A closer structural analogue is dynamical linear bandits [29], where actions drive an unobserved linear dynamical system. Their reward is a linear combination of the latent state and the current action, and under stability the relevant benchmark reduces to the best constant action. By contrast, the bilinear reward structure in our setting induces interactions between current and past actions, leading to a finite-horizon open-loop action-sequence benchmark. The closest setting to ours is the latent linear-dynamics bandit model [11], which adopts this bilinear reward structure. Unlike bandit models with finite memory, the resulting reward process has infinite memory because past actions can continue to influence future rewards through the latent state. Compared with the previous ETC approach [11], we show that a UCB-based adaptive algorithm achieves improved statistical efficiency, yielding a sharper regret guarantee.

Online control and nonstochastic control. Our problem is also related to online control of linear dynamical systems. A large literature studies regret minimization in *Linear Quadratic Regulator* (LQR), *Linear Quadratic Gaussian* (LQG), and related control problems, where the learner controls a linear system and incurs state-action costs [1, 14, 28, 13, 24]. More recent work on nonstochastic control allows adversarial perturbations and adversarially chosen convex losses [3, 4, 16, 35]. These works share with ours the use of stability and input-output representations of linear dynamical systems. However, their objective is control with exogenous loss functions, typically convex or quadratic costs in the state, observation, and control, and the benchmark is a feedback controller class. By contrast, our model is a bandit reward-maximization problem in which the only observation is a scalar bilinear reward, and the benchmark is an optimal finite-horizon open-loop action sequence. This bilinear reward structure creates pairwise interactions between current and past actions after unrolling the dynamics, which do not arise in standard LQR/LQG or nonstochastic control formulations.

2 Preliminaries

Notations. For a vector $\mathbf{x} \in \mathbb{R}^d$, we denote its ℓ_2-norm by $\|\mathbf{x}\|_2$. For a positive semidefinite matrix $\mathbf{A}$, we define the $\mathbf{A}$-weighted norm by $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. For a matrix $\mathbf{M}$, we use $\|\mathbf{M}\|_2$ to denote its spectral norm, $\rho(\mathbf{M})$ its spectral radius, and $\lambda_{\max}(\mathbf{M})$ its largest eigenvalue when $\mathbf{M}$ is symmetric. We denote the Frobenius norm of a matrix $\mathbf{M}$ by $\|\mathbf{M}\|_F$. We write $\mathrm{vec}(\mathbf{M})$ for the vectorization of $\mathbf{M}$, and $\otimes$ for the Kronecker product. The vector $\mathbf{0}_d$ denotes the zero vector in $\mathbb{R}^d$. We use $\mathcal{O}(\cdot)$ to hide universal constant factors, and $\widetilde{\mathcal{O}}(\cdot)$ to additionally hide logarithmic factors.

Problem setup. We study an endogenous nonstationary stochastic bandit problem in which rewards are generated through controlled latent dynamics. At each round $t = 1, \dots, T$, the learner selects an action $\mathbf{u}_t \in \mathcal{U} \subset \mathbb{R}^p$ and observes a scalar reward r_t, while the latent state $\mathbf{x}_t \in \mathbb{R}^n$ remains unobserved. The reward depends bilinearly on the current action and the latent state, and the action also affects the future evolution of the latent state. Specifically, the system evolves as

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{w}_t, \qquad r_t = \mathbf{u}_t^\top \mathbf{C}\mathbf{x}_t + z_t, \tag{1}$$

where $\mathbf{x}_1 = \mathbf{0}_n$, and $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, and $\mathbf{C} \in \mathbb{R}^{p\times n}$ are *unknown* matrices. The process noise $\{\mathbf{w}_t\}_{t\geq 1}$ and reward noise $\{z_t\}_{t\geq 1}$ are unobserved zero-mean noise processes.

For a deterministic open-loop action sequence $\mathbf{u}_{1:T} \in \mathcal{U}^T$, unrolling Eq. (1) gives $r_t = \sum_{s=1}^{t-1} \mathbf{u}_t^\top \mathbf{C}\mathbf{A}^{s-1}\mathbf{B}\mathbf{u}_{t-s} + \sum_{s=1}^{t-1} \mathbf{u}_t^\top \mathbf{C}\mathbf{A}^{s-1}\mathbf{w}_{t-s} + z_t$. Therefore, for deterministic open-loop sequences, the expected reward at round t is

$$\mu_t(\mathbf{u}_{1:t}) = \sum_{s=1}^{\infty} \mathbf{u}_t^\top \mathbf{G}_{s-1}\mathbf{u}_{t-s}. \qquad \mathbf{G}_h := \mathbf{C}\mathbf{A}^h\mathbf{B}\,, \tag{2}$$

where we set $\mathbf{u}_s = \mathbf{0}_p$ for all $s \leq 0$. For adaptive policies, we use $\mu_t(\mathbf{u}_{1:t})$ as the action-driven mean-reward functional obtained by removing the process-noise and reward-noise terms from the unrolled

reward. The open-loop benchmark is $\mathrm{OPT}_T := \sup_{\mathbf{u}_{1:T} \in \mathcal{U}^T} \sum_{t=1}^{T} \mu_t(\mathbf{u}_{1:t})$. For an adaptive policy π, the realized action sequence $\mathbf{u}_{1:T}^{\pi}$ is random. We define its *pseudo-regret* with respect to the open-loop benchmark as

$$\mathcal{R}_T^{\mathrm{pseudo}}(\pi) := \mathrm{OPT}_T - \sum_{t=1}^{T} \mu_t(\mathbf{u}_{1:t}^{\pi}).$$

Thus, $\mathcal{R}_T^{\mathrm{pseudo}}(\pi)$ is a random variable through the action sequence generated by the policy. Now, we define the actual expected regret in terms of the realized rewards collected by policy π:

$$\mathrm{Regret}_T(\pi) := \mathrm{OPT}_T - \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} r_t^{\pi}\right].$$

Here the expectation is over the policy randomization and the process and reward noises.

Remark 1. *Recent work on latent linear dynamics bandits [11] analyzes expected regret defined in terms of realized rewards. In this paper, we first establish a high-probability bound on the pseudo-regret $\mathcal{R}_T^{\mathrm{pseudo}}(\pi)$, which is defined with respect to the action-driven mean-reward functional. For adaptive policies, this pseudo-regret need not coincide with actual expected regret, because actions may be correlated with past process noise through previously observed rewards. We then translate the pseudo-regret guarantee into an actual expected regret guarantee for our block-committed algorithm by bounding the resulting adaptation gap (Lemma 8).*

Assumptions. We next state some standard modeling assumptions, commonly used in stochastic bandits and linear dynamical systems, under which we design and analyze our algorithm.

Assumption 1 (Action set)**.** *The action set is compact, i.e., $\mathcal{U}$ is closed and for all $\mathbf{u} \in \mathcal{U}$, $\|\mathbf{u}\|_2 \le 1$.*

Assumption 2 (Stability)**.** *For some $\rho \in (0,1)$ and $M_\rho \ge 1$, $\|\mathbf{A}^h\|_2 \le M_\rho \rho^h$ for all $h \ge 0$.*

Assumption 3 (Sub-Gaussian noise)**.** *Let $\{\mathcal{F}_t\}_{t\ge 0}$ be a filtration containing the interaction history and the latent process. At each round t, the action $\mathbf{u}_t$ is $\mathcal{F}_{t-1}$-measurable, and the noises $(z_t, \mathbf{w}_t)$ are realized after $\mathbf{u}_t$ is chosen. Conditional on $\mathcal{F}_{t-1}$, z_t and $\mathbf{w}_t$ are independent. The reward noise z_t is conditionally σ_z^2-sub-Gaussian, and the process noise $\mathbf{w}_t$ is conditionally sub-Gaussian with variance proxy $\Sigma_w \succeq \mathbf{0}_{n\times n}$: for every $\mathbf{v} \in \mathbb{R}^n$ and every $\alpha \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\alpha z_t) \mid \mathcal{F}_{t-1}] \le \exp\left(\frac{\alpha^2\sigma_z^2}{2}\right) \quad \textit{and} \quad \mathbb{E}\big[\exp(\alpha \mathbf{v}^\top \mathbf{w}_t) \mid \mathcal{F}_{t-1}\big] \le \exp\left(\frac{\alpha^2 \mathbf{v}^\top \mathbf{\Sigma}_w \mathbf{v}}{2}\right).$$

These assumptions are standard in stochastic bandit and linear dynamical system models. Assumption 1 rules out unbounded rewards and ensures that the open-loop benchmark is well-defined over a compact action set. Assumption 2 imposes exponential stability of the latent dynamics, which is equivalent to Schur stability in finite dimensions. This type of stability assumption is common in learning and control of linear dynamical systems, since it keeps the state and rewards well behaved under persistent inputs and process noise, and ensures geometric decay of past actions and noises [35, 30, 25, 36, 32, 11]. Assumption 3 is a standard light-tail condition used for high-probability least-squares and self-normalized martingale concentration arguments.

3 Main Results

3.1 Cyclic Approximation

A central difficulty in the latent-dynamics bandit model is that learning and planning are inherently coupled. The learner must estimate the latent input-output dynamics from temporally correlated rewards, while also choosing actions whose effects propagate through the latent state over the horizon. The ETC approach [11] addresses this challenge by separating learning and planning: it first estimates the latent dynamics from exploratory data and then commits to an optimized open-loop sequence. This phase separation yields a $\widetilde{\mathcal{O}}(T^{2/3})$ regret rate, reflecting a suboptimal exploration–exploitation trade-off. Our goal is to avoid this separation by updating the dynamics estimate and revising the plan adaptively throughout the horizon.

To avoid this rigid phase separation, we exploit the stability of the latent dynamics. Stability makes the effect of far-past actions decay geometrically, allowing the infinite-memory mean-reward functional

to be approximated by a finite-memory proxy. We then use a block construction of length $m + L$: the first m actions serve as a warm-up segment, and the last L positions yield informative rewards whose m-step histories remain inside the block. This turns the long-horizon open-loop problem into a stationary block-level learning problem, in which each block is a composite action and the proxy reward is linear in the truncated Markov parameters.

Finite-memory approximation. We first formalize the finite-memory approximation of the mean reward defined in Eq. (2). For a memory cutoff $m \geq 1$, define the truncated mean reward as

$$\bar{\mu}_t(\mathbf{u}_{1:t}) := \sum_{s=1}^{m} \mathbf{u}_t^\top \mathbf{G}_{s-1} \mathbf{u}_{t-s} \,,$$

where we use the convention $\mathbf{u}_j = \mathbf{0}_p$ for all $j \leq 0$. Under the stability condition in Assumption 2, the truncation error decays geometrically. In particular, since $\|\mathbf{u}\|_2 \leq 1$ for every $\mathbf{u} \in \mathcal{U}$, we have

$$|\mu_t(\mathbf{u}_{1:t}) - \bar{\mu}_t(\mathbf{u}_{1:t})| \leq \Delta_m := \sum_{h=m}^{\infty} \|\mathbf{G}_h\|_2 \leq \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho} \rho^m \,. \tag{3}$$

We refer to Δ_m as the finite-memory truncation error. The derivation of Eq. (3) is provided in Lemma 4 in Appendix A.

Block-level proxy reward. The finite-memory representation allows us to localize the reward within short blocks. Fix an informative suffix length $L \geq 1$, and let the block length be $m + L$. For simplicity, we assume that $T/(m+L)$ is an integer; a nonzero remainder contributes only an additive $\mathcal{O}(m + L)$ term.

For a block $\mathbf{U} = (\mathbf{v}_1, \ldots, \mathbf{v}_{m+L}) \in \mathcal{U}^{m+L}$, we define its proxy reward as the truncated reward accumulated over the last L positions of the block as $\bar{r}_m(\mathbf{U}) := \sum_{j=m+1}^{m+L} \sum_{s=1}^{m} \mathbf{v}_j^\top \mathbf{G}_{s-1} \mathbf{v}_{j-s}$. The first m positions serve as a warm-up segment so that every informative position $j \in \{m+1, \ldots, m+L\}$ has its full m-step action history contained inside the same block. This block construction reveals a *linear reward structure*: $\bar{r}_m(\mathbf{U})$ can be written as a linear function of the truncated Markov parameters. Define

$$\boldsymbol{\theta}_m^* := \begin{bmatrix} \mathrm{vec}(\mathbf{G}_0) \\ \mathrm{vec}(\mathbf{G}_1) \\ \vdots \\ \mathrm{vec}(\mathbf{G}_{m-1}) \end{bmatrix} \in \mathbb{R}^{p^2 m}, \qquad \boldsymbol{\phi}_j(\mathbf{U}) := \begin{bmatrix} \mathbf{v}_{j-1} \otimes \mathbf{v}_j \\ \mathbf{v}_{j-2} \otimes \mathbf{v}_j \\ \vdots \\ \mathbf{v}_{j-m} \otimes \mathbf{v}_j \end{bmatrix} \in \mathbb{R}^{p^2 m},$$

for each informative position $j \in \{m+1, \ldots, m+L\}$. Then we have

$$\bar{r}_m(\mathbf{U}) = \sum_{j=m+1}^{m+L} \sum_{s=1}^{m} \mathbf{v}_j^\top \mathbf{G}_{s-1} \mathbf{v}_{j-s} = \sum_{j=m+1}^{m+L} \sum_{s=1}^{m} \langle \mathbf{v}_{j-s} \otimes \mathbf{v}_j, \mathrm{vec}(\mathbf{G}_{s-1}) \rangle = \sum_{j=m+1}^{m+L} \left\langle \boldsymbol{\phi}_j(\mathbf{U}), \boldsymbol{\theta}_m^* \right\rangle . \tag{4}$$

Thus, under the finite-memory block construction, each block can be viewed as a composite action whose expected proxy reward is linear in the unknown parameter $\boldsymbol{\theta}_m^*$. This linear representation is the key reduction that enables us to apply optimism-based linear bandit methods at the block level.

Cyclic approximation. We next show that the original open-loop benchmark can be approximated by optimizing the block-level proxy reward. Let $\mathbf{U}^* \in \mathrm{argmax}_{\mathbf{U} \in \mathcal{U}^{m+L}} \bar{r}_m(\mathbf{U})$ be a *proxy-optimal block*, and let $\mathbf{U}_{\mathrm{cyc}}^* \in \mathcal{U}^T$ denote its cyclic repetition over the horizon T. The following proposition shows that this cyclic sequence approximates the optimal benchmark up to a controlled additive error.

Proposition 1 (Cyclic approximation)**.** *Suppose that Assumptions 1 and 2 hold. Then, we have*

$$\sup_{\mathbf{u}_{1:T} \in \mathcal{U}^T} \sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t}) - \sum_{t=1}^{T} \bar{\mu}_t\big((\mathbf{U}_{\mathrm{cyc}}^*)_{1:t}\big) \leq \frac{2Tm}{m+L} R_m \,,$$

where $R_m := \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho}$. *Consequently,*

$$\mathrm{OPT}_T - \sum_{t=1}^{T} \mu_t\big((\mathbf{U}_{\mathrm{cyc}}^*)_{1:t}\big) \leq 2T\Delta_m + \frac{2Tm}{m+L} R_m \,,$$

where Δ_m *is the finite-memory truncation error defined in Eq.* (3)*.*

Algorithm 1 `Block-OFUL`

Inputs: horizon $T \in \mathbb{N}$, memory cutoff $m \in \mathbb{N}$, informative suffix length $L \in \mathbb{N}$, regularization parameter $\lambda > 0$, confidence radii $\{\gamma_\tau\}_{\tau \geq 0}$

Set $d = p^2 m$, $N = \lceil T/(m+L) \rceil$, and initialize $\mathbf{V}_0 = \lambda \mathbf{I}_d$, $\hat{\boldsymbol{\theta}}_0 = \mathbf{0}_d$.

for $\tau = 1, 2, \ldots, N$ **do**

Choose $\mathbf{U}_\tau \in \underset{\mathbf{U} \in \mathcal{U}^{m+L}}{\operatorname{argmax}} \sum_{j=m+1}^{m+L} \left(\langle \boldsymbol{\phi}_j(\mathbf{U}), \hat{\boldsymbol{\theta}}_{\tau-1} \rangle + \gamma_{\tau-1} \left\| \boldsymbol{\phi}_j(\mathbf{U}) \right\|_{\mathbf{V}_{\tau-1}^{-1}} \right)$

Play $\mathbf{U}_\tau = (\mathbf{u}_{\tau,1}, \ldots, \mathbf{u}_{\tau,m+L})$ sequentially and observe rewards $r_{\tau,1}, \ldots, r_{\tau,m+L}$

Update $\mathbf{V}_\tau = \mathbf{V}_{\tau-1} + \sum_{j=m+1}^{m+L} \boldsymbol{\phi}_j(\mathbf{U}_\tau) \boldsymbol{\phi}_j(\mathbf{U}_\tau)^\top$, $\hat{\boldsymbol{\theta}}_\tau = \mathbf{V}_\tau^{-1} \left(\sum_{\tau'=1}^{\tau} \sum_{j=m+1}^{m+L} \boldsymbol{\phi}_j(\mathbf{U}_{\tau'}) r_{\tau',j} \right)$

end for

Discussion of Proposition 1. Proposition 1 reduces the original nonstationary planning problem to a stationary block-level optimization problem: it suffices to learn and optimize the proxy reward $\bar{r}_m(\mathbf{U})$ over blocks, while paying the approximation error $2T\Delta_m + 2TmR_m/(m+L)$.

This reduction is related to the block-cyclic construction for linear bandits with finite memory [12, 5], but our setting differs in two important ways. First, in finite-memory bandits, the memory length is part of the model, so after ignoring the first m positions of each block, the remaining rewards are exactly localized within the block. In our latent-dynamics model, the reward process has *infinite memory*: every past action can affect the current reward through the latent state, and the cutoff m is only an approximation parameter. Thus, our block reduction must control both the truncation error Δ_m and the warm-up loss induced at block boundaries, of order $mR_m/(m+L)$. Second, the reward in finite-memory linear bandits [12, 5] is linear in observed action features, whereas our reward depends bilinearly on the current action and an unobserved latent state. After unrolling the dynamics, this induces a quadratic dependence on pairs of past and current actions through unknown Markov parameters. Therefore, Proposition 1 is not merely a reuse of a finite-memory blocking argument; it shows that an endogenous latent-dynamics bandit with infinite memory can be approximated by a stationary linear bandit problem over blocks.

3.2 Algorithm: `Block-OFUL`

Proposition 1 shows that the original nonstationary planning problem can be reduced, up to a controlled approximation error, to learning and optimizing the block-level proxy reward $\bar{r}_m(\mathbf{U})$. Since Eq. (4) represents $\bar{r}_m(\mathbf{U})$ as a linear function of the unknown truncated Markov parameter $\boldsymbol{\theta}_m^*$, we can treat each block $\mathbf{U} \in \mathcal{U}^{m+L}$ as a composite action in a linear bandit problem. Based on this reduction, we propose `Block-OFUL`, an optimistic block-level algorithm described in Algorithm 1.

The algorithm proceeds over blocks of length $m+L$. At the beginning of each block τ, it constructs a confidence set around the ridge estimator $\hat{\boldsymbol{\theta}}_{\tau-1}$ using the Gram matrix $\mathbf{V}_{\tau-1}$, and selects a block $\mathbf{U}_\tau$ that maximizes a UCB on the proxy reward. More specifically, the selected block maximizes the sum of the estimated reward and the uncertainty bonus over the informative suffix $j = m+1, \ldots, m+L$:

$$\mathrm{UCB}_{\tau-1}(\mathbf{U}) := \sum_{j=m+1}^{m+L} \left(\langle \boldsymbol{\phi}_j(\mathbf{U}), \hat{\boldsymbol{\theta}}_{\tau-1} \rangle + \gamma_{\tau-1} \| \boldsymbol{\phi}_j(\mathbf{U}) \|_{\mathbf{V}_{\tau-1}^{-1}} \right).$$

After playing the chosen block sequentially, the learner observes all rewards in the block, but uses only the last L rewards for the regression update. The first m actions are used only to build the within-block history required by the finite-memory proxy, while the informative suffix provides noisy linear observations of $\boldsymbol{\theta}_m^*$.

This design differs from the previous ETC approach [11]. Rather than spending an initial exploration phase to estimate the Markov parameters and then committing to a fixed action sequence, our proposed algorithm interleaves exploration and exploitation at the block level. After each block, it updates its estimate of $\boldsymbol{\theta}_m^*$ and immediately uses the refined confidence set to choose the next block optimistically. This block-level adaptivity is the key mechanism that avoids the sub-optimal exploration–exploitation trade-off inherent in the previous ETC approach.

Remark 2 (Maximization oracle)**.** *The block selection step in Algorithm 1 is formulated as an offline UCB maximization problem over $\mathcal{U}^{m+L}$. Our theoretical analysis assumes access to an exact maximization oracle; similar oracle-based formulations have been widely used in the previous literature [9, 23, 12]. This allows us to focus on the statistical learning aspect of the problem, separately from the implementation of the offline optimizer. In practice, the UCB objective can be optimized approximately, for example by gradient ascent when $\mathcal{U}$ is continuous. Designing computationally efficient oracles tailored to the block-UCB objective, together with corresponding regret guarantees, is an interesting direction for future work.*

3.3 Regret Bound

We now state the main regret guarantee of Algorithm 1. Before stating the theorem, we introduce some notation. For a fixed memory cutoff $m \in \mathbb{N}$, let S_m be an upper bound on the norm of the vectorized truncated Markov parameter $\boldsymbol{\theta}^*_m$, i.e., $\|\boldsymbol{\theta}^*_m\|_2 \le S_m$. Also, let L_ϕ be a uniform upper bound on the lifted features, i.e., $\|\boldsymbol{\phi}_j(\mathbf{U})\|_2 \le L_\phi$ for all $\mathbf{U} \in \mathcal{U}^{m+L}$ and $j \in \{m+1, \dots, m+L\}$.

Lemma 1 (Pseudo-regret bound for Algorithm 1)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let $d = p^2 m$, and let $N = T/(m+L)$ be an integer. For any $\delta \in (0,1)$, set the parameters in Algorithm 1 as $\lambda \ge L_\phi^2$ and $\gamma_\tau = \alpha_\tau + \beta_\tau$, where*

$$\alpha_\tau := \sigma_{\mathcal{B}} \sqrt{d \log\left(1 + \frac{\tau L L_\phi^2}{d\lambda}\right) + 2\log\frac{1}{\delta}} + \sqrt{\lambda} S_m,$$

$$\beta_\tau := \bar{\Delta}_m(\delta) \sqrt{2\tau L d \log\left(1 + \frac{\tau L L_\phi^2}{d\lambda}\right)}, \quad \sigma_{\mathcal{B}}^2 := \sigma_z^2 + \frac{\lambda_{\max}(\boldsymbol{\Sigma}_w)\|\mathbf{C}\|_2^2 M_\rho^2}{(1-\rho)^2},$$

$$\bar{\Delta}_m(\delta) := \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho}\rho^m + 2\rho^m \sqrt{\frac{2\lambda_{\max}(\boldsymbol{\Sigma}_w)\|\mathbf{C}\|_2^2 M_\rho^2}{1-\rho^2} \log\frac{2\cdot 5^p T}{\delta}}.$$

Then, with probability at least $1-\delta$, the pseudo-regret of the policy π of Algorithm 1 satisfies

$$\mathcal{R}_T^{\mathrm{pseudo}}(\pi) \le 2T\Delta_m + \frac{2Tm}{m+L} R_m + 2\gamma_N \left[\frac{2LL_\phi d}{\sqrt{\lambda}} + \sqrt{2Td\log\left(1 + \frac{TL_\phi^2}{\lambda d}\right)}\right].$$

In particular, choosing $m = -\frac{1}{2}\log_\rho T$, $L = \frac{\sqrt{mT}}{d^{3/4}} - m$, and $\lambda = L_\phi^2$ gives

$$\mathcal{R}_T^{\mathrm{pseudo}}(\pi) = \widetilde{\mathcal{O}}\left(d^{3/4}\sqrt{mT} + d\sqrt{T}\right).$$

Now, we provide the actual expected regret guarantee of Algorithm 1.

Theorem 1 (Actual expected regret of Algorithm 1)**.** *Suppose the assumptions of Lemma 1 hold. For a fixed $\delta \in (0,1)$, let $B_T(\delta)$ be the pseudo-regret bound established in Lemma 1. Define $\Delta_0 := \sum_{h=0}^{\infty} \|\mathbf{G}_h\|_2$. Then, we have*

$$\mathrm{Regret}_T(\pi) \le B_T(\delta) + 2\delta T \Delta_0 + \frac{\|\mathbf{C}\|_2 M_\rho}{1-\rho}\sqrt{\frac{p\lambda_{\max}(\boldsymbol{\Sigma}_w)}{1-\rho^2}}\frac{T}{m+L}.$$

Consequently, choosing $\delta = T^{-2}$ and the parameters $m = -\frac{1}{2}\log_\rho T$, $L = \frac{\sqrt{mT}}{d^{3/4}} - m$, $\lambda = L_\phi^2$, up to integer rounding and whenever $L \ge 1$, gives

$$\mathrm{Regret}_T(\pi) = \widetilde{O}\left(d^{3/4}\sqrt{mT} + d\sqrt{T}\right) = \widetilde{O}\left(\left[\frac{p^{3/2}}{\log(1/\rho)^{5/4}} + \frac{p^2}{\log(1/\rho)}\right]\sqrt{T}\right).$$

Discussion of Theorem 1. Theorem 1 establishes an actual expected regret guarantee for Algorithm 1. Compared with the recent ETC approach for the same latent-dynamics bandit model [11], our result improves the regret rate from $\widetilde{\mathcal{O}}(T^{2/3})$ to $\widetilde{\mathcal{O}}(\sqrt{T})$ under the open-loop action-sequence benchmark. The improvement comes from replacing the phase-separated explore-then-commit strategy with

adaptive block-level optimism. In ETC, the learner spends a long initial uniform-exploration phase to estimate the latent dynamics and then commits to a fixed open-loop sequence, which induces the suboptimal exploration–exploitation trade-off leading to the $\widetilde{\mathcal{O}}(T^{2/3})$ rate. In contrast, Algorithm 1 updates its confidence set after each block and selects the next block optimistically, so every block contributes simultaneously to learning and reward maximization. Thus, adaptive exploration at the block level removes the need for a long pure-exploration phase and yields the improved regret.

3.4 Regret Analysis

In this section, we present the main technical ingredients of the regret analysis for Lemma 1. First, we relate the pseudo-regret to the proxy regret for the block-level proxy reward. Second, we establish a high-probability confidence set for the truncated Markov parameter $\boldsymbol{\theta}_m^*$. Finally, we use optimism to bound the proxy regret by a cumulative uncertainty term and control this term through a blockwise elliptical-potential argument.

Pseudo-regret decomposition. Let $N := T/(m+L)$, and let $\mathbf{U}_1, \dots, \mathbf{U}_N$ be the sequence of blocks selected by Algorithm 1. Define the *proxy regret* $\mathcal{R}_{\text{proxy}} := \sum_{\tau=1}^{N} \left(\bar{r}_m(\mathbf{U}^*) - \bar{r}_m(\mathbf{U}_\tau)\right)$, where $\mathbf{U}^* \in \operatorname{argmax}_{\mathbf{U}\in\mathcal{U}^{m+L}} \bar{r}_m(\mathbf{U})$ is the proxy-optimal block. By Proposition 1, the pseudo-regret is bounded by the approximation error plus the proxy regret:

$$\mathcal{R}_T^{\text{pseudo}}(\pi) \le 2T\Delta_m + \frac{2Tm}{m+L} R_m + \mathcal{R}_{\text{proxy}} .$$

Thus, it remains to control $\mathcal{R}_{\text{proxy}}$.

Confidence set for truncated Markov parameters. We establish a confidence set for the truncated Markov parameter $\boldsymbol{\theta}_m^*$. The key step is to express the informative rewards in each block as noisy linear observations of $\boldsymbol{\theta}_m^*$. The following decomposition separates the observed reward into the linear signal, the current-block noise, the action-tail bias due to finite-memory truncation, and the process-noise tail coming from process noises outside the current finite-memory window.

Lemma 2 (Reward decomposition)**.** *Let $N = T/(m+L)$, and assume that $N \in \mathbb{N}$. For each block $\tau \in [N]$, let $\mathbf{U}_\tau = (\mathbf{u}_{\tau,1}, \dots, \mathbf{u}_{\tau,m+L})$, $\mathbf{u}_{\tau,j} = \mathbf{u}_{t_{\tau,j}}$, and $t_{\tau,j} := (\tau-1)(m+L)+j$. For every $j \in \{m+1, \dots, m+L\}$, define $r_{\tau,j} := r_{t_{\tau,j}}$. Then,*

$$r_{\tau,j} = \left\langle \boldsymbol{\phi}_j(\mathbf{U}_\tau), \boldsymbol{\theta}_m^* \right\rangle + \epsilon_{\tau,j} + \eta_{\tau,j} + \zeta_{\tau,j},$$

where

$$\epsilon_{\tau,j} := z_{t_{\tau,j}} + \sum_{h=0}^{j-2} \mathbf{u}_{\tau,j}^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h} , \qquad \text{(current-block noise)}$$

$$\eta_{\tau,j} := \sum_{s=m+1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1} \mathbf{u}_{t_{\tau,j}-s} , \qquad \text{(action-tail bias)}$$

$$\zeta_{\tau,j} := \sum_{h=j-1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h} . \qquad \text{(process-noise tail)}$$

Using the decomposition in Lemma 2, we construct a confidence set for the truncated Markov parameter $\boldsymbol{\theta}_m^*$. The current-block noise $\epsilon_{\tau,j}$ is controlled through a self-normalized martingale concentration argument, whereas the action-tail bias $\eta_{\tau,j}$ and the process-noise tail $\zeta_{\tau,j}$ are controlled by the stability of the latent dynamics.

Lemma 3 (Confidence set)**.** *Let $N = T/(m+L)$, and assume that $N \in \mathbb{N}$. Suppose Assumptions 1–3 hold. Fix $\delta \in (0,1)$, and suppose that $\lambda \ge L_\phi^2$. Then, with probability at least $1-\delta$, for all $\tau \in [N]$, $\|\hat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_m^*\|_{\mathbf{V}_\tau} \le \gamma_\tau$, where γ_τ is the confidence radius specified in Lemma 1.*

Lemma 3 is the key ingredient that enables optimism in our block-level formulation. At a high level, it plays a role analogous to the confidence set used in block-based analyses for linear bandits with finite memory [12]. However, there is an important distinction. In finite-memory linear bandits, the

memory length is part of the model specification, and each informative reward is exactly linear in a finite-memory feature with martingale noise.

In our latent-dynamics model, however, the finite-memory representation is only an approximation of an infinite-memory reward process induced by the latent state dynamics. This distinction is reflected in Lemma 2. Each informative reward contains not only the current-block noise $\epsilon_{\tau,j}$, but also the action-tail bias $\eta_{\tau,j}$ and the process-noise tail $\zeta_{\tau,j}$. The term $\eta_{\tau,j}$ appears because actions more than m steps in the past can still affect the current reward through the latent state, while $\zeta_{\tau,j}$ captures the contribution of process noises generated before the current block. Importantly, these pre-block noises can be correlated with the adaptively selected block through past reward observations, so $\zeta_{\tau,j}$ is not a martingale-difference noise term with respect to the pre-block history. Thus, unlike in the finite-memory setting, blocking does not yield a linear observation model with martingale noise.

We overcome these difficulties by exploiting the exponential stability of $\mathbf{A}$. The action-tail bias is bounded directly through the geometric decay of the Markov parameters. For the process-noise tail, we separate the adaptively chosen action from the underlying vector-valued noise tail and establish a uniform high-probability bound over all informative indices. This bound remains valid for the adaptively selected blocks. The current-block noise is then controlled using a blockwise self-normalized martingale concentration argument. Combining these bounds yields the confidence set in Lemma 3, extending the finite-memory block analysis to our latent-dynamics setting, where effective finite memory arises from stability rather than being built into the model.

Proxy regret. We control the proxy regret on the confidence event from Lemma 3. By optimism, the proxy-optimal block is upper bounded by its UCB score, while Algorithm 1 selects a block maximizing this score. Hence, for every τ,

$$\bar{r}_m(\mathbf{U}^*) - \bar{r}_m(\mathbf{U}_\tau) \le 2\gamma_{\tau-1} \sum_{j=m+1}^{m+L} \|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_{\mathbf{V}_{\tau-1}^{-1}}.$$

Summing over $\tau = 1, \dots, N$ and using the monotonicity of γ_τ, we obtain

$$\mathcal{R}_{\text{proxy}} \le 2\gamma_N \sum_{\tau=1}^{N} \sum_{j=m+1}^{m+L} \|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_{\mathbf{V}_{\tau-1}^{-1}}.$$

It remains to control the cumulative uncertainty term. The main technical issue is that the UCB score uses the pre-block Gram matrix $\mathbf{V}_{\tau-1}$, whereas the elliptical-potential argument naturally applies to sequentially updated within-block Gram matrices. In the previous block-cyclic analysis [12], this mismatch is absorbed by enlarging the confidence radius, which makes the block length appear multiplicatively in the leading estimation term. Instead, we keep the confidence radius tied to $\mathbf{V}_{\tau-1}$ and control the mismatch directly in the uncertainty sum.

Let $\mathbf{V}_{\tau,j} := \mathbf{V}_{\tau-1} + \sum_{i=m+1}^{j} \boldsymbol{\phi}_i(\mathbf{U}_\tau)\boldsymbol{\phi}_i(\mathbf{U}_\tau)^\top$, with $\mathbf{V}_{\tau,m} := \mathbf{V}_{\tau-1}$. We decompose each uncertainty term into

$$\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_{\mathbf{V}_{\tau-1}^{-1}} = \|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_{\mathbf{V}_{\tau,j-1}^{-1}} + \Big(\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_{\mathbf{V}_{\tau-1}^{-1}} - \|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_{\mathbf{V}_{\tau,j-1}^{-1}}\Big).$$

The first term is controlled by the usual elliptical-potential argument, and the second by a Gram-matrix telescoping argument. This yields

$$\sum_{\tau=1}^{N} \sum_{j=m+1}^{m+L} \|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_{\mathbf{V}_{\tau-1}^{-1}} \le \frac{2LL_\phi d}{\sqrt{\lambda}} + \sqrt{2Td \log\left(1 + \frac{TL_\phi^2}{\lambda d}\right)}.$$

Thus, L contributes only through an additive block-update mismatch term, rather than multiplying the leading $\sqrt{T}$-type uncertainty term. Combining this bound with the proxy-regret decomposition proves Lemma 1; the full derivation is deferred to Appendix B.

4 Experiments

We evaluate the proposed algorithm, `Block-OFUL`, on two synthetic instances of the latent linear-dynamics model in Eq. (1). In both instances, the latent-state dimension is $n = 4$, the action

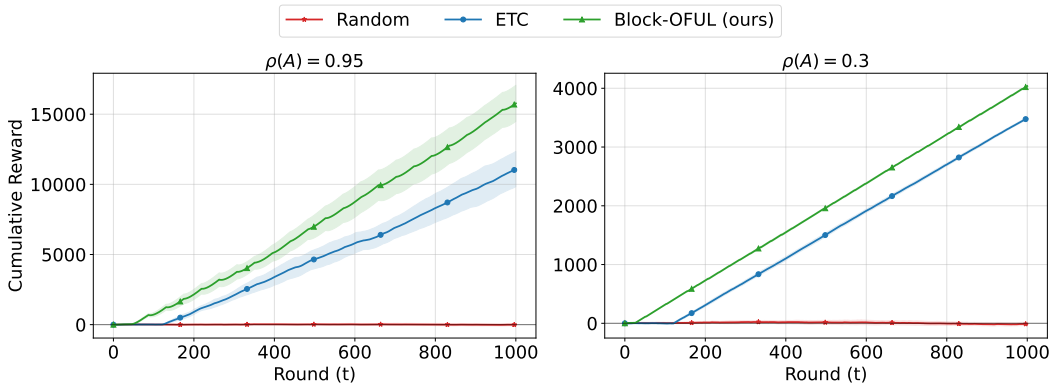

Figure 1: Cumulative reward on latent linear-dynamics bandit instances with different stability levels.

dimension is $p = 3$, and the action set is $\mathcal{U} = \{-1, +1\}^p$. We use the same input and reward matrices

$$\mathbf{B} = \begin{bmatrix} 0.75 & 0.35 & 0.54 \\ 0.01 & 0.58 & 0.46 \\ 0.68 & -0.09 & -0.61 \\ -0.15 & 0.83 & -0.47 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 0.67 & 0.04 & 0.72 & -0.19 \\ 0.20 & 0.65 & 0.00 & 0.76 \\ 0.54 & 0.46 & -0.61 & -0.47 \end{bmatrix},$$

and consider two transition matrices with different stability levels:

$$\mathbf{A}_{0.95} = \begin{bmatrix} 0.52 & 0.14 & 0.27 & -0.18 \\ 0.14 & 0.50 & -0.08 & -0.06 \\ 0.27 & -0.08 & 0.75 & 0.03 \\ -0.18 & -0.06 & 0.03 & 0.52 \end{bmatrix}, \qquad \mathbf{A}_{0.3} = \begin{bmatrix} 0.16 & 0.04 & 0.09 & -0.06 \\ 0.04 & 0.13 & -0.02 & -0.03 \\ 0.09 & -0.02 & 0.23 & 0.00 \\ -0.06 & -0.03 & 0.00 & 0.14 \end{bmatrix}.$$

Here, $\rho(\mathbf{A}_{0.95}) = 0.95$ and $\rho(\mathbf{A}_{0.3}) = 0.3$. These two instances allow us to examine the effect of the persistence of the latent dynamics. We compare `Block-OFUL` with the ETC baseline [11]. For ETC, we use an exploration length of $H = 120$ in both instances. For `Block-OFUL`, we use $(m, L) = (8, 36)$ when $\rho(\mathbf{A}) = 0.95$ and $(m, L) = (2, 24)$ when $\rho(\mathbf{A}) = 0.3$. The reward and process noises are Gaussian with variance proxy 0.01.

Figure 1 shows the cumulative reward over the interaction horizon. Across both stability regimes, `Block-OFUL` consistently accumulates reward faster than ETC, while the random baseline remains close to zero. For the instance with $\rho(\mathbf{A}) = 0.95$, the performance gap between `Block-OFUL` and ETC becomes increasingly pronounced over time, indicating that continued adaptation is particularly beneficial when the effects of past actions decay slowly. For the faster-decaying instance with $\rho(\mathbf{A}) = 0.3$, both methods exhibit more stable reward accumulation, but `Block-OFUL` maintains a consistently higher cumulative-reward trajectory throughout most of the horizon. Overall, these results demonstrate the benefit of adaptive block-level learning over the phase-separated ETC strategy across different levels of dynamical persistence.

5 Conclusion

We studied an endogenous nonstationary stochastic bandit problem with latent linear dynamics, where actions influence both immediate rewards and the future latent state. This structure induces history-dependent rewards and makes long-horizon planning substantially more challenging than in classical stochastic bandits. We showed that, under stable latent dynamics, the original infinite-memory problem can be reduced to a stationary block-level linear bandit problem through cyclic approximation. Based on this reduction, we proposed an optimistic algorithm that learns the truncated Markov parameters while adaptively selecting blocks. Our main result establishes a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{T})$, significantly improving over the previous $\widetilde{\mathcal{O}}(T^{2/3})$ explore-then-commit guarantee for the same model. To the best of our knowledge, this is the first regret guarantee of order $\widetilde{\mathcal{O}}(\sqrt{T})$ for latent linear-dynamics bandits with bilinear rewards and action-controlled latent states. Future work includes developing computationally efficient oracles for block-UCB maximization, extending the analysis to adaptive or feedback benchmarks, and designing parameter-free variants that adapt to unknown stability and memory scales.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant both funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023- 00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008) and the Ascending SNU Future Leader Fellowship through Seoul National University.

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th annual conference on learning theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[3] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

[4] Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Heesang Ann, Hyunjun Choi, Taehyun Hwang, Younghoon Shin, Haeju Cheong, and Min hwan Oh. Generalized linear bandits with memory. In *Forty-third International Conference on Machine Learning*, 2026.

[6] Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. Blocking bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27, 2014.

[8] Leonardo Cella and Nicolo Cesa-Bianchi. Stochastic bandits with delay-dependent payoffs. In *International Conference on Artificial Intelligence and Statistics*, pages 1168–1177. PMLR, 2020.

[9] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

[10] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR, 2019.

[11] Sunmook Choi, Yahya Sattar, Yassir Jedra, Maryam Fazel, and Sarah Dean. Explore-then-commit for nonstationary linear bandits with latent dynamics. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[12] Giulia Clerici, Pierre Laforgue, and Nicolò Cesa-Bianchi. Linear bandits with memory. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

[13] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{t}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

[14] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

[16] Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Algorithmic Learning Theory*, pages 408–421. PMLR, 2020.

[17] Rong Jin, David Simchi-Levi, Li Wang, Xinshang Wang, and Sen Yang. Shrinking the upper confidence bound: A dynamic product selection problem for urban warehouses. *Management Science*, 67(8):4756–4771, 2021.

[18] Komal Kapoor, Karthik Subbian, Jaideep Srivastava, and Paul Schrater. Just in time recommendations: Modeling the dynamics of boredom in activity streams. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 233–242, 2015.

[19] Baekjin Kim and Ambuj Tewari. Randomized exploration for non-stationary stochastic linear bandits. In *Conference on Uncertainty in Artificial Intelligence*, pages 71–80. PMLR, 2020.

[20] Robert Kleinberg and Nicole Immorlica. Recharging bandits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 309–319. IEEE, 2018.

[21] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. Rotating online behavior change interventions increases effectiveness but also increases attrition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[22] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems–a survey. *Knowledge-based systems*, 123:154–162, 2017.

[23] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

[24] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Adaptive control and regret minimization in linear quadratic gaussian (lqg) setting. In *2021 American Control Conference (ACC)*, pages 2517–2522. IEEE, 2021.

[25] Holden Lee. Improved rates for prediction and identification of partially observed linear dynamical systems. In *International Conference on Algorithmic Learning Theory*, pages 668–698. PMLR, 2022.

[26] Liu Leqi, Fatma Kilinc Karzan, Zachary Lipton, and Alan Montgomery. Rebounding bandits for modeling satiation effects. *Advances in Neural Information Processing Systems*, 34:4003–4014, 2021.

[27] Nir Levine, Koby Crammer, and Shie Mannor. Rotting bandits. *Advances in neural information processing systems*, 30, 2017.

[28] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in neural information processing systems*, 32, 2019.

[29] Marco Mussi, Alberto Maria Metelli, and Marcello Restelli. Dynamical linear bandits. In *International Conference on Machine Learning*, pages 25563–25587. PMLR, 2023.

[30] Samet Oymak and Necmiye Ozay. Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 2021.

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

[32] Yahya Sattar, Yassir Jedra, and Sarah Dean. Learning linear dynamics from bilinear observations. In *2025 American Control Conference (ACC)*, pages 3109–3115. IEEE, 2025.

[33] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, 2018.

[34] Julien Seznec, Andrea Locatelli, Alexandra Carpentier, Alessandro Lazaric, and Michal Valko. Rotting bandits are no harder than stochastic ones. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2564–2572. PMLR, 2019.

[35] Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

[36] Y Jennifer Sun, Stephen Newman, and Elad Hazan. Optimal rates for bandit nonstochastic control. *Advances in Neural Information Processing Systems*, 36:21908–21919, 2023.

Contents of Appendix

A Proof of Proposition 1 **14**

B Proof of Theorem 1 **16**

B.1 Proof of Lemma 2 16

B.2 Proof of Lemma 3 17

B.3 Proof of Lemma 1 20

B.4 Proof of Theorem 1 23

B.5 Proof of Technical Lemmas 24

B.5.1 Proof of Lemma 5 24

B.5.2 Proof of Lemma 6 25

B.5.3 Proof of Lemma 7 27

B.5.4 Proof of Lemma 8 29

C Auxiliary Lemmas **30**

D Experimental Details **31**

D.1 Implementation of `Block-OFUL` 31

D.2 Additional Experiments 31

A Proof of Proposition 1

In this section, we provide the technical details for our cyclic approximation. First, we show that under the stability condition, the truncation error decays geometrically.

Lemma 4 (Finite-memory truncation error)**.** *Let $\bar{\mu}_t(\mathbf{u}_{1:t})$ denote the m-truncated version of the mean reward $\mu_t(\mathbf{u}_{1:t})$, defined by*

$$\mu_t(\mathbf{u}_{1:t}) = \sum_{s=1}^{\infty} \mathbf{u}_t^\top \mathbf{G}_{s-1} \mathbf{u}_{t-s}, \qquad \bar{\mu}_t(\mathbf{u}_{1:t}) := \sum_{s=1}^{m} \mathbf{u}_t^\top \mathbf{G}_{s-1} \mathbf{u}_{t-s}.$$

Suppose Assumptions 1 and 2 hold. Then, for every $t \geq 1$, we have

$$|\mu_t(\mathbf{u}_{1:t}) - \bar{\mu}_t(\mathbf{u}_{1:t})| \leq \Delta_m := \sum_{h=m}^{\infty} \|\mathbf{G}_h\|_2 \leq \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho} \rho^m,$$

and

$$|\bar{\mu}_t(\mathbf{u}_{1:t})| \leq R_m := \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho}.$$

Proof of Lemma 4. Recall that for every $t \geq 1$, we have

$$\mu_t(\mathbf{u}_{1:t}) := \sum_{s=1}^{\infty} \mathbf{u}_t^\top \mathbf{G}_{s-1} \mathbf{u}_{t-s}\,.$$

For a fixed memory cutoff $m \geq 1$, the truncated mean reward is defined as

$$\bar{\mu}_t(\mathbf{u}_{1:t}) := \sum_{s=1}^{m} \mathbf{u}_t^\top \mathbf{G}_{s-1} \mathbf{u}_{t-s}\,.$$

Then,

$$\mu_t(\mathbf{u}_{1:t}) - \bar{\mu}_t(\mathbf{u}_{1:t}) = \sum_{s=m+1}^{\infty} \mathbf{u}_t^\top \mathbf{G}_{s-1} \mathbf{u}_{t-s} \,.$$

Since $\|\mathbf{u}\|_2 \le 1$ for every $\mathbf{u} \in \mathcal{U}$, we have

$$\left|\mathbf{u}_t^\top \mathbf{G}_{s-1} \mathbf{u}_{t-s}\right| \le \|\mathbf{u}_t\|_2 \|\mathbf{G}_{s-1}\|_2 \|\mathbf{u}_{t-s}\|_2 \le \|\mathbf{G}_{s-1}\|_2 \,.$$

Therefore,

$$|\mu_t(\mathbf{u}_{1:t}) - \bar{\mu}_t(\mathbf{u}_{1:t})| \le \sum_{s=m+1}^{\infty} \|\mathbf{G}_{s-1}\|_2 = \sum_{h=m}^{\infty} \|\mathbf{G}_h\|_2 \,.$$

Finally, by Assumption 2,

$$\|\mathbf{G}_h\|_2 = \|\mathbf{C}\mathbf{A}^h\mathbf{B}\|_2 \le \|\mathbf{C}\|_2 \|\mathbf{A}^h\|_2 \|\mathbf{B}\|_2 \le \|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho \rho^h.$$

Hence, we have

$$|\mu_t(\mathbf{u}_{1:t}) - \bar{\mu}_t(\mathbf{u}_{1:t})| \le \|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho \sum_{h=m}^{\infty} \rho^h = \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho} \rho^m \,.$$

□

Now, we provide the proof of Proposition 1.

Proposition 1 (Cyclic approximation)**.** *Suppose that Assumptions 1 and 2 hold. Then, we have*

$$\sup_{\mathbf{u}_{1:T} \in \mathcal{U}^T} \sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t}) - \sum_{t=1}^{T} \bar{\mu}_t\big((\mathbf{U}^*_{\mathrm{cyc}})_{1:t}\big) \le \frac{2Tm}{m+L} R_m \,,$$

where $R_m := \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho}$*. Consequently,*

$$\mathrm{OPT}_T - \sum_{t=1}^{T} \mu_t\big((\mathbf{U}^*_{\mathrm{cyc}})_{1:t}\big) \le 2T\Delta_m + \frac{2Tm}{m+L} R_m \,,$$

where Δ_m *is the finite-memory truncation error defined in Eq.* (3)*.*

Proof of Proposition 1. Fix an arbitrary action sequence $\mathbf{u}_{1:T} \in \mathcal{U}^T$, and partition it into consecutive blocks of length $m+L$. By definition,

$$\begin{aligned}
\sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t}) &= \sum_{\tau=1}^{T/(m+L)} \sum_{j=1}^{m+L} \bar{\mu}_{(\tau-1)(m+L)+j}\big(\mathbf{u}_{1:(\tau-1)(m+L)+j}\big) \\
&= \sum_{\tau=1}^{T/(m+L)} \sum_{j=1}^{m} \bar{\mu}_{(\tau-1)(m+L)+j}\big(\mathbf{u}_{1:(\tau-1)(m+L)+j}\big) \\
&\quad + \sum_{\tau=1}^{T/(m+L)} \sum_{j=m+1}^{m+L} \bar{\mu}_{(\tau-1)(m+L)+j}\big(\mathbf{u}_{1:(\tau-1)(m+L)+j}\big) \,.
\end{aligned}$$

By Lemma 4, each prefix term satisfies $|\bar{\mu}_t| \le R_m$. Hence,

$$\sum_{\tau=1}^{T/(m+L)} \sum_{j=1}^{m} \bar{\mu}_{(\tau-1)(m+L)+j}\big(\mathbf{u}_{1:(\tau-1)(m+L)+j}\big) \le \frac{Tm}{m+L} R_m \,. \tag{5}$$

For the suffix term, since $j \ge m+1$, the truncated reward depends only on the previous m actions within the same block. Therefore, for each block τ,

$$\begin{aligned}
\sum_{j=m+1}^{m+L} \bar{\mu}_{(\tau-1)(m+L)+j}\big(\mathbf{u}_{1:(\tau-1)(m+L)+j}\big) &= \bar{r}_m\big(\mathbf{u}_{(\tau-1)(m+L)+1:\tau(m+L)}\big) \\
&\le \bar{r}_m(\mathbf{U}^*),
\end{aligned}$$

where the last inequality follows from the optimality of $\mathbf{U}^* \in \operatorname{argmax}_{\mathbf{U}\in\mathcal{U}^{m+L}} \bar{r}_m(\mathbf{U})$. Combining this with Eq. (5), we obtain

$$\sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t}) \le \frac{T}{m+L}\bar{r}_m(\mathbf{U}^*) + \frac{Tm}{m+L}R_m. \tag{6}$$

We now apply the same decomposition to the cyclic sequence $\mathbf{U}^*_{\mathrm{cyc}}$. For each block, the suffix contribution is exactly $\bar{r}_m(\mathbf{U}^*)$, while the prefix contribution is bounded from below by $-mR_m$. Thus,

$$\sum_{t=1}^{T} \bar{\mu}_t\big((\mathbf{U}^*_{\mathrm{cyc}})_{1:t}\big) \ge -\frac{Tm}{m+L}R_m + \frac{T}{m+L}\bar{r}_m(\mathbf{U}^*). \tag{7}$$

Combining Eqs. (6) and (7), we get

$$\sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t}) - \sum_{t=1}^{T} \bar{\mu}_t\big((\mathbf{U}^*_{\mathrm{cyc}})_{1:t}\big) \le \frac{2Tm}{m+L}R_m\,.$$

Since this holds for every $\mathbf{u}_{1:T} \in \mathcal{U}^T$, we have

$$\sup_{\mathbf{u}_{1:T}\in\mathcal{U}^T} \sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t}) - \sum_{t=1}^{T} \bar{\mu}_t\big((\mathbf{U}^*_{\mathrm{cyc}})_{1:t}\big) \le \frac{2Tm}{m+L}R_m. \tag{8}$$

It remains to transfer this bound from the truncated rewards to the original rewards. By Lemma 4, for every action sequence $\mathbf{u}_{1:T}$,

$$\left|\sum_{t=1}^{T} \mu_t(\mathbf{u}_{1:t}) - \sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t})\right| \le \sum_{t=1}^{T} |\mu_t(\mathbf{u}_{1:t}) - \bar{\mu}_t(\mathbf{u}_{1:t})| \le T\Delta_m\,.$$

Therefore,

$$\begin{aligned}\mathrm{OPT}_T - \sum_{t=1}^{T} \mu_t\big((\mathbf{U}^*_{\mathrm{cyc}})_{1:t}\big) &\le \sup_{\mathbf{u}_{1:T}\in\mathcal{U}^T} \sum_{t=1}^{T} \bar{\mu}_t(\mathbf{u}_{1:t}) - \sum_{t=1}^{T} \bar{\mu}_t\big((\mathbf{U}^*_{\mathrm{cyc}})_{1:t}\big) + 2T\Delta_m \\ &\le \frac{2Tm}{m+L}R_m + 2T\Delta_m,\end{aligned}$$

where the last inequality follows from Eq. (8). □

B Proof of Theorem 1

B.1 Proof of Lemma 2

Lemma 2 (Reward decomposition)**.** *Let $N = T/(m+L)$, and assume that $N \in \mathbb{N}$. For each block $\tau \in [N]$, let $\mathbf{U}_\tau = (\mathbf{u}_{\tau,1}, \dots, \mathbf{u}_{\tau,m+L})$, $\mathbf{u}_{\tau,j} = \mathbf{u}_{t_{\tau,j}}$, and $t_{\tau,j} := (\tau-1)(m+L)+j$. For every $j \in \{m+1, \dots, m+L\}$, define $r_{\tau,j} := r_{t_{\tau,j}}$. Then,*

$$r_{\tau,j} = \big\langle \boldsymbol{\phi}_j(\mathbf{U}_\tau), \boldsymbol{\theta}^*_m \big\rangle + \epsilon_{\tau,j} + \eta_{\tau,j} + \zeta_{\tau,j},$$

where

$$\epsilon_{\tau,j} := z_{t_{\tau,j}} + \sum_{h=0}^{j-2} \mathbf{u}_{\tau,j}^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}\,, \qquad \text{(current-block noise)}$$

$$\eta_{\tau,j} := \sum_{s=m+1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1} \mathbf{u}_{t_{\tau,j}-s}\,, \qquad \text{(action-tail bias)}$$

$$\zeta_{\tau,j} := \sum_{h=j-1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}\,. \qquad \text{(process-noise tail)}$$

Proof of Lemma 2. We first unroll the latent dynamics. Since $\mathbf{x}_1 = \mathbf{0}_n$, for any $t \geq 1$,

$$\begin{aligned}\mathbf{x}_t &= \sum_{i=1}^{t-1} \mathbf{A}^{t-1-i} (\mathbf{B}\mathbf{u}_i + \mathbf{w}_i) \\ &= \sum_{i=1}^{t-1} \mathbf{A}^{t-1-i}\mathbf{B}\mathbf{u}_i + \sum_{i=1}^{t-1} \mathbf{A}^{t-1-i}\mathbf{w}_i.\end{aligned}$$

Therefore,

$$\begin{aligned}r_t &= \mathbf{u}_t^\top \mathbf{C}\mathbf{x}_t + z_t \\ &= \sum_{i=1}^{t-1} \mathbf{u}_t^\top \mathbf{C}\mathbf{A}^{t-1-i}\mathbf{B}\mathbf{u}_i + \sum_{i=1}^{t-1} \mathbf{u}_t^\top \mathbf{C}\mathbf{A}^{t-1-i}\mathbf{w}_i + z_t.\end{aligned}$$

Reindexing the first sum by $s = t - i$, and the second sum by $h = t - 1 - i$, gives

$$r_t = \sum_{s=1}^{\infty} \mathbf{u}_t^\top \mathbf{G}_{s-1}\mathbf{u}_{t-s} + \sum_{h=0}^{\infty} \mathbf{u}_t^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t-1-h} + z_t\,,$$

where $\mathbf{G}_h := \mathbf{C}\mathbf{A}^h\mathbf{B}$, and we use the convention that $\mathbf{u}_s = \mathbf{0}_p$ and $\mathbf{w}_s = \mathbf{0}_n$ for $s \leq 0$.

Now fix a block $\tau \in [N]$ and an informative position $j \in \{m+1, \ldots, m+L\}$, and set $t = t_{\tau,j}$. Since $\mathbf{u}_{t_{\tau,j}} = \mathbf{u}_{\tau,j}$, the above identity yields

$$\begin{aligned}r_{\tau,j} &= \sum_{s=1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1}\mathbf{u}_{t_{\tau,j}-s} + z_{t_{\tau,j}} + \sum_{h=0}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h} \\ &= \sum_{s=1}^{m} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1}\mathbf{u}_{t_{\tau,j}-s} + \sum_{s=m+1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1}\mathbf{u}_{t_{\tau,j}-s} \\ &\quad + z_{t_{\tau,j}} + \sum_{h=0}^{j-2} \mathbf{u}_{\tau,j}^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h} + \sum_{h=j-1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}\,.\end{aligned}$$

Since $j \geq m+1$, for every $s = 1, \ldots, m$, the action $\mathbf{u}_{t_{\tau,j}-s}$ lies in the same block and equals $\mathbf{u}_{\tau,j-s}$. Hence,

$$\begin{aligned}\sum_{s=1}^{m} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1}\mathbf{u}_{t_{\tau,j}-s} &= \sum_{s=1}^{m} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1}\mathbf{u}_{\tau,j-s} \\ &= \left\langle \boldsymbol{\phi}_j(\mathbf{U}_\tau), \boldsymbol{\theta}_m^* \right\rangle,\end{aligned}$$

by the definitions of $\boldsymbol{\phi}_j(\mathbf{U}_\tau)$ and $\boldsymbol{\theta}_m^*$. Collecting the remaining terms according to the definitions of $\epsilon_{\tau,j}$, $\eta_{\tau,j}$, and $\zeta_{\tau,j}$ gives

$$r_{\tau,j} = \left\langle \boldsymbol{\phi}_j(\mathbf{U}_\tau), \boldsymbol{\theta}_m^* \right\rangle + \epsilon_{\tau,j} + \eta_{\tau,j} + \zeta_{\tau,j}\,.$$

□

B.2 Proof of Lemma 3

Lemma 3 (Confidence set)**.** *Let $N = T/(m+L)$, and assume that $N \in \mathbb{N}$. Suppose Assumptions 1–3 hold. Fix $\delta \in (0,1)$, and suppose that $\lambda \geq L_\phi^2$. Then, with probability at least $1-\delta$, for all $\tau \in [N]$, $\|\hat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_m^*\|_{\mathbf{V}_\tau} \leq \gamma_\tau$, where γ_τ is the confidence radius specified in Lemma 1.*

Proof of Lemma 3. For each block $s \in [N]$ and informative position $j \in \{m+1, \ldots, m+L\}$, let

$$\boldsymbol{\phi}_{s,j} := \boldsymbol{\phi}_j(\mathbf{U}_s).$$

Define the block feature matrix and reward vector by

$$\boldsymbol{\Phi}_s := \left[\boldsymbol{\phi}_{s,m+1}, \ldots, \boldsymbol{\phi}_{s,m+L}\right] \in \mathbb{R}^{d \times L}, \qquad \mathbf{r}_s := (r_{s,m+1}, \ldots, r_{s,m+L})^\top \in \mathbb{R}^L.$$

Similarly, define

$$\boldsymbol{\epsilon}_s := (\epsilon_{s,m+1}, \ldots, \epsilon_{s,m+L})^\top, \quad \boldsymbol{\eta}_s := (\eta_{s,m+1}, \ldots, \eta_{s,m+L})^\top, \quad \boldsymbol{\zeta}_s := (\zeta_{s,m+1}, \ldots, \zeta_{s,m+L})^\top.$$

Then Lemma 2 gives the vector form

$$\mathbf{r}_s = \boldsymbol{\Phi}_s^\top \boldsymbol{\theta}_m^* + \boldsymbol{\epsilon}_s + \boldsymbol{\eta}_s + \boldsymbol{\zeta}_s.$$

Recall that

$$\mathbf{V}_\tau = \lambda \mathbf{I}_d + \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\Phi}_s^\top, \qquad \widehat{\boldsymbol{\theta}}_\tau = \mathbf{V}_\tau^{-1} \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \mathbf{r}_s.$$

Using the reward decomposition, we obtain

$$\mathbf{V}_\tau \big(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_m^*\big) = -\lambda \boldsymbol{\theta}_m^* + \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\epsilon}_s + \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s (\boldsymbol{\eta}_s + \boldsymbol{\zeta}_s).$$

Therefore,

$$\left\| \widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_m^* \right\|_{\mathbf{V}_\tau} \le \sqrt{\lambda}\, \|\boldsymbol{\theta}_m^*\|_2 + \underbrace{\left\| \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\epsilon}_s \right\|_{\mathbf{V}_\tau^{-1}}}_{\text{block noise}} + \underbrace{\left\| \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s (\boldsymbol{\eta}_s + \boldsymbol{\zeta}_s) \right\|_{\mathbf{V}_\tau^{-1}}}_{\text{bias}}. \tag{9}$$

The first term is bounded by $\sqrt{\lambda} S_m$. It remains to control the current-block noise term and the bias term. For the current-block noise term, we introduce the following lemmas.

Lemma 5 (Blockwise sub-Gaussianity of the current-block noise)**.** *Let $\mathcal{H}_{\tau-1}$ denote the sigma-field just before the process and reward noises in block τ are revealed; in particular, $\mathbf{U}_\tau$ and $\boldsymbol{\Phi}_\tau$ are $\mathcal{H}_{\tau-1}$-measurable. Under Assumptions 1 and 3, for every $\tau \in [N]$, every $\xi \in \mathbb{R}^L$, and every $\eta \in \mathbb{R}$,*

$$\mathbb{E}\big[\exp\big(\eta \xi^\top \boldsymbol{\epsilon}_\tau\big) \,\big|\, \mathcal{H}_{\tau-1}\big] \le \exp\left(\frac{\eta^2 \sigma_{\mathcal{B}}^2}{2} \|\xi\|_2^2 \right),$$

where

$$\sigma_{\mathcal{B}}^2 := \sigma_z^2 + \lambda_{\max}(\boldsymbol{\Sigma}_w) \frac{\|\mathbf{C}\|_2^2 M_\rho^2}{(1-\rho)^2}.$$

Equivalently, $\boldsymbol{\epsilon}_\tau$ is conditionally $\sigma_{\mathcal{B}}$-sub-Gaussian as a vector in $\mathbb{R}^L$.

Lemma 6 (Blockwise self-normalized bound)**.** *Let $(\mathcal{H}_\tau)_{\tau \ge 0}$ be a filtration. Let $\boldsymbol{\Phi}_\tau \in \mathbb{R}^{d \times L}$ be $\mathcal{H}_{\tau-1}$-measurable, and let $\boldsymbol{\epsilon}_\tau \in \mathbb{R}^L$ be $\mathcal{H}_\tau$-measurable. Assume that $\boldsymbol{\epsilon}_\tau$ is conditionally σ-sub-Gaussian as a vector: for every $\boldsymbol{\xi} \in \mathbb{R}^L$ and every $\alpha \in \mathbb{R}$,*

$$\mathbb{E}\Big[\exp\Big(\alpha \boldsymbol{\xi}^\top \boldsymbol{\epsilon}_\tau\Big) \,\Big|\, \mathcal{H}_{\tau-1}\Big] \le \exp\left(\frac{\alpha^2 \sigma^2}{2} \|\boldsymbol{\xi}\|_2^2 \right).$$

Let $\lambda > 0$, $\mathbf{V}_0 = \lambda \mathbf{I}_d$, and define

$$\mathbf{V}_\tau := \mathbf{V}_0 + \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\Phi}_s^\top, \qquad \mathbf{S}_\tau := \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\epsilon}_s.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds simultaneously for all $\tau \ge 0$:

$$\|\mathbf{S}_\tau\|_{\mathbf{V}_\tau^{-1}} \le \sigma \sqrt{2 \log\left(\frac{\det(\mathbf{V}_\tau)^{1/2}}{\det(\mathbf{V}_0)^{1/2} \delta} \right)}.$$

Lemma 5 shows that the vector of current-block noises $\boldsymbol{\epsilon}_\tau$ is conditionally sub-Gaussian at the block level. This is needed because the L informative rewards within a block are not treated as independent scalar observations; rather, they enter the regression update jointly through the matrix $\boldsymbol{\Phi}_\tau$. Lemma 6 then converts this blockwise sub-Gaussianity into a self-normalized concentration bound for the martingale sum $\sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\epsilon}_s$. It is the block/vector-valued analogue of the standard self-normalized

inequality for scalar martingale noise used in linear bandits [2]. Combining these two lemmas, the event

$$\Omega_{\mathrm{sn}}(\delta) := \left\{ \forall \tau \in [N] : \left\| \sum_{s=1}^{\tau} \mathbf{\Phi}_s \boldsymbol{\epsilon}_s \right\|_{\mathbf{V}_\tau^{-1}} \le \sigma_{\mathcal{B}} \sqrt{2 \log \left(\frac{\det(\mathbf{V}_\tau)^{1/2}}{\det(\lambda \mathbf{I}_d)^{1/2} \delta} \right)} \right\} \tag{10}$$

holds with probability at least $1-\delta$. Now it remains to control the bias term. We use the following high-probability bound for the action-tail bias and the process-noise tail.

Lemma 7 (High-probability bias bound)**.** *Fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the event*

$$\Omega_{\mathrm{bias}}(\delta) := \left\{ \forall \tau \in [N],\ \forall j \in \{m+1, \dots, m+L\},\ |\eta_{\tau,j} + \zeta_{\tau,j}| \le \bar{\Delta}_m(\delta) \right\}$$

holds, where

$$\bar{\Delta}_m(\delta) := \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho} \rho^m + 2\rho^m \sqrt{2\lambda_{\max}(\mathbf{\Sigma}_w) \frac{\|\mathbf{C}\|_2^2 M_\rho^2}{1-\rho^2} \log \frac{2 \cdot 5^p T}{\delta}}\,.$$

By Lemma 7, the event $\Omega_{\mathrm{bias}}(\delta)$ holds with probability at least $1-\delta$. On $\Omega_{\mathrm{bias}}(\delta)$, for every $\tau \in [N]$,

$$\begin{aligned}
\left\| \sum_{s=1}^{\tau} \mathbf{\Phi}_s (\boldsymbol{\eta}_s + \boldsymbol{\zeta}_s) \right\|_{\mathbf{V}_\tau^{-1}} &= \left\| \sum_{s=1}^{\tau} \sum_{j=m+1}^{m+L} \boldsymbol{\phi}_{s,j} (\eta_{s,j} + \zeta_{s,j}) \right\|_{\mathbf{V}_\tau^{-1}} \\
&\le \sum_{s=1}^{\tau} \sum_{j=m+1}^{m+L} |\eta_{s,j} + \zeta_{s,j}| \|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_\tau^{-1}} \\
&\le \bar{\Delta}_m(\delta) \sum_{s=1}^{\tau} \sum_{j=m+1}^{m+L} \|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_\tau^{-1}} \\
&\le \bar{\Delta}_m(\delta) \sqrt{\tau L \sum_{s=1}^{\tau} \sum_{j=m+1}^{m+L} \|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_\tau^{-1}}^2},
\end{aligned}$$

where the last inequality follows from Cauchy–Schwarz.

It remains to bound the sum of squared norms. Let $\mathbf{V}_{s,j-1}$ denote the sequential Gram matrix immediately before adding the feature $\boldsymbol{\phi}_{s,j}$; that is,

$$\mathbf{V}_{s,j-1} := \lambda \mathbf{I}_d + \sum_{q=1}^{s-1} \sum_{i=m+1}^{m+L} \boldsymbol{\phi}_{q,i} \boldsymbol{\phi}_{q,i}^\top + \sum_{i=m+1}^{j-1} \boldsymbol{\phi}_{s,i} \boldsymbol{\phi}_{s,i}^\top.$$

Since $\mathbf{V}_\tau \succeq \mathbf{V}_{s,j-1}$ for all $s \le \tau$, we have

$$\|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_\tau^{-1}} \le \|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_{s,j-1}^{-1}}.$$

Moreover, $\lambda \ge L_\phi^2$ and $\|\boldsymbol{\phi}_{s,j}\|_2 \le L_\phi$ imply

$$\|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_{s,j-1}^{-1}}^2 \le \frac{\|\boldsymbol{\phi}_{s,j}\|_2^2}{\lambda} \le 1.$$

Therefore, the standard elliptical-potential (Lemma 10) bound gives

$$\sum_{s=1}^{\tau} \sum_{j=m+1}^{m+L} \|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_\tau^{-1}}^2 \le \sum_{s=1}^{\tau} \sum_{j=m+1}^{m+L} \|\boldsymbol{\phi}_{s,j}\|_{\mathbf{V}_{s,j-1}^{-1}}^2 \le 2 \log \left(\frac{\det(\mathbf{V}_\tau)}{\det(\lambda \mathbf{I}_d)} \right).$$

Consequently,

$$\left\| \sum_{s=1}^{\tau} \mathbf{\Phi}_s (\boldsymbol{\eta}_s + \boldsymbol{\zeta}_s) \right\|_{\mathbf{V}_\tau^{-1}} \le \bar{\Delta}_m(\delta) \sqrt{2\tau L \log \left(\frac{\det(\mathbf{V}_\tau)}{\det(\lambda \mathbf{I}_d)} \right)}. \tag{11}$$

Now define the good event

$$\Omega_{\text{good}}(\delta) := \Omega_{\text{sn}}(\delta/2) \cap \Omega_{\text{bias}}(\delta/2).$$

By the bounds above and a union bound,

$$\mathbb{P}(\Omega_{\text{good}}(\delta)) \geq 1 - \delta.$$

On $\Omega_{\text{good}}(\delta)$, combining Eqs. (9), (10), and (11), and using $\|\boldsymbol{\theta}_m^*\|_2 \leq S_m$, yields for all $\tau \in [N]$,

$$\left\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_m^*\right\|_{\mathbf{V}_\tau} \leq \sqrt{\lambda} S_m + \sigma_{\mathcal{B}} \sqrt{2 \log\left(\frac{2 \det(\mathbf{V}_\tau)^{1/2}}{\det(\lambda \mathbf{I}_d)^{1/2} \delta}\right)} + \bar{\Delta}_m(\delta/2) \sqrt{2\tau L \log\left(\frac{\det(\mathbf{V}_\tau)}{\det(\lambda \mathbf{I}_d)}\right)}.$$

The right-hand side is exactly the confidence radius γ_τ chosen in Lemma 1. Hence,

$$\left\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}_m^*\right\|_{\mathbf{V}_\tau} \leq \gamma_\tau \qquad \forall \tau \in [N],$$

with probability at least $1 - \delta$. □

B.3 Proof of Lemma 1

Lemma 1 (Pseudo-regret bound for Algorithm 1)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let $d = p^2 m$, and let $N = T/(m + L)$ be an integer. For any $\delta \in (0, 1)$, set the parameters in Algorithm 1 as $\lambda \geq L_\phi^2$ and $\gamma_\tau = \alpha_\tau + \beta_\tau$, where*

$$\alpha_\tau := \sigma_{\mathcal{B}} \sqrt{d \log\left(1 + \frac{\tau L L_\phi^2}{d\lambda}\right) + 2 \log \frac{1}{\delta}} + \sqrt{\lambda} S_m,$$

$$\beta_\tau := \bar{\Delta}_m(\delta) \sqrt{2\tau L d \log\left(1 + \frac{\tau L L_\phi^2}{d\lambda}\right)}, \quad \sigma_{\mathcal{B}}^2 := \sigma_z^2 + \frac{\lambda_{\max}(\boldsymbol{\Sigma}_w) \|\mathbf{C}\|_2^2 M_\rho^2}{(1-\rho)^2},$$

$$\bar{\Delta}_m(\delta) := \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho} \rho^m + 2\rho^m \sqrt{\frac{2\lambda_{\max}(\boldsymbol{\Sigma}_w) \|\mathbf{C}\|_2^2 M_\rho^2}{1-\rho^2} \log \frac{2 \cdot 5^p T}{\delta}}.$$

Then, with probability at least $1 - \delta$, the pseudo-regret of the policy π of Algorithm 1 satisfies

$$\mathcal{R}_T^{\text{pseudo}}(\pi) \leq 2T\Delta_m + \frac{2Tm}{m+L} R_m + 2\gamma_N \left[\frac{2LL_\phi d}{\sqrt{\lambda}} + \sqrt{2Td \log\left(1 + \frac{TL_\phi^2}{\lambda d}\right)}\right].$$

In particular, choosing $m = -\frac{1}{2} \log_\rho T$, $L = \frac{\sqrt{mT}}{d^{3/4}} - m$, and $\lambda = L_\phi^2$ gives

$$\mathcal{R}_T^{\text{pseudo}}(\pi) = \widetilde{\mathcal{O}}\left(d^{3/4}\sqrt{mT} + d\sqrt{T}\right).$$

Proof of Lemma 1. Let $N := T/(m + L)$, and assume $N \in \mathbb{N}$. Define the proxy regret

$$\mathcal{R}_{\text{proxy}} := \sum_{\tau=1}^{N} \left(\bar{r}_m(\mathbf{U}^*) - \bar{r}_m(\mathbf{U}_\tau)\right).$$

On the confidence event, for any block $\mathbf{U} \in \mathcal{U}^{m+L}$,

$$\bar{r}_m(\mathbf{U}) \leq \text{UCB}_{\tau-1}(\mathbf{U}),$$

where

$$\text{UCB}_{\tau-1}(\mathbf{U}) := \sum_{j=m+1}^{m+L} \left(\left\langle \boldsymbol{\phi}_j(\mathbf{U}), \widehat{\boldsymbol{\theta}}_{\tau-1} \right\rangle + \gamma_{\tau-1} \left\|\boldsymbol{\phi}_j(\mathbf{U})\right\|_{\mathbf{V}_{\tau-1}^{-1}}\right).$$

Since $\mathbf{U}_\tau$ maximizes this UCB score,

$$\bar{r}_m(\widetilde{\mathbf{b}}) \le \mathrm{UCB}_{\tau-1}(\widetilde{\mathbf{b}}) \le \mathrm{UCB}_{\tau-1}(\mathbf{U}_\tau).$$

Moreover, the confidence event also implies

$$\bar{r}_m(\mathbf{U}_\tau) \ge \mathrm{UCB}_{\tau-1}(\mathbf{U}_\tau) - 2\gamma_{\tau-1} \sum_{j=m+1}^{m+L} \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}}.$$

Combining the two inequalities gives

$$\bar{r}_m(\widetilde{\mathbf{b}}) - \bar{r}_m(\mathbf{U}_\tau) \le 2\gamma_{\tau-1} \sum_{j=m+1}^{m+L} \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}}.$$

Summing over $\tau = 1, \dots, N$ and using the monotonicity of γ_τ, we obtain

$$\mathcal{R}_{\mathrm{proxy}} \le 2\gamma_N \sum_{\tau=1}^{N} \sum_{j=m+1}^{m+L} \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}}. \tag{12}$$

It remains to control the cumulative uncertainty term. For each block τ, define the within-block sequential Gram matrices

$$\mathbf{V}_{\tau,j} := \mathbf{V}_{\tau-1} + \sum_{i=m+1}^{j} \boldsymbol{\phi}_i(\mathbf{U}_\tau)\boldsymbol{\phi}_i(\mathbf{U}_\tau)^\top, \qquad j = m+1, \dots, m+L,$$

with the convention $\mathbf{V}_{\tau,m} := \mathbf{V}_{\tau-1}$. Then

$$\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}} = \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau,j-1}^{-1}} + \left(\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}} - \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau,j-1}^{-1}} \right).$$

We bound the two terms separately.

First, by Cauchy–Schwarz and the standard elliptical-potential lemma (Lemma 10),

$$\begin{aligned}
\sum_{\tau=1}^{N} \sum_{j=m+1}^{m+L} \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau,j-1}^{-1}} &\le \sqrt{NL \sum_{\tau=1}^{N} \sum_{j=m+1}^{m+L} \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|^2_{\mathbf{V}_{\tau,j-1}^{-1}}} \\
&\le \sqrt{2NL \log\left(\frac{\det(\mathbf{V}_N)}{\det(\lambda \mathbf{I}_d)}\right)} \\
&\le \sqrt{2Td \log\left(1 + \frac{TL_\phi^2}{\lambda d}\right)},
\end{aligned}$$

where the last inequality uses $NL \le T$, $\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\|_2 \le L_\phi$, and the determinant-trace inequality (Lemma 11).

Next, since $\mathbf{V}_{\tau,j-1} \preceq \mathbf{V}_\tau$, we have

$$\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau,j-1}^{-1}} \ge \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_\tau^{-1}}.$$

Hence,

$$\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}} - \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau,j-1}^{-1}} \le \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}} - \left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_\tau^{-1}}.$$

Let $\lambda_{\tau,1}, \dots, \lambda_{\tau,d}$ be the eigenvalues of $\mathbf{V}_\tau$. To control the discrepancy between the pre-block and within-block Gram matrices, we use the Gram-matrix telescoping inequality for block updates

(Lemma 12):

$$\begin{aligned}
&\sum_{\tau=1}^{N}\sum_{j=m+1}^{m+L}\left(\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}}-\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau,j-1}^{-1}}\right)\\
&\quad\le\sum_{\tau=1}^{N}\sum_{j=m+1}^{m+L}\left(\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}}-\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau}^{-1}}\right)\\
&\quad\le\sum_{\tau=1}^{N}\sum_{j=m+1}^{m+L}2\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_2\sum_{k=1}^{d}\left(\lambda_{\tau-1,k}^{-1/2}-\lambda_{\tau,k}^{-1/2}\right)\\
&\quad\le 2LL_\phi\sum_{\tau=1}^{N}\sum_{k=1}^{d}\left(\lambda_{\tau-1,k}^{-1/2}-\lambda_{\tau,k}^{-1/2}\right)\\
&\quad\le 2LL_\phi\sum_{k=1}^{d}\left(\lambda_{0,k}^{-1/2}-\lambda_{N,k}^{-1/2}\right)\\
&\quad\le\frac{2LL_\phi d}{\sqrt{\lambda}},
\end{aligned}$$

where the last step uses $\mathbf{V}_0=\lambda\mathbf{I}_d$.

Combining the two bounds gives

$$\sum_{\tau=1}^{N}\sum_{j=m+1}^{m+L}\left\|\boldsymbol{\phi}_j(\mathbf{U}_\tau)\right\|_{\mathbf{V}_{\tau-1}^{-1}}\le\frac{2LL_\phi d}{\sqrt{\lambda}}+\sqrt{2Td\log\left(1+\frac{TL_\phi^2}{\lambda d}\right)}. \tag{13}$$

Substituting Eq. (13) into Eq. (12) yields

$$\mathcal{R}_{\text{proxy}}\le 2\gamma_N\left[\frac{2LL_\phi d}{\sqrt{\lambda}}+\sqrt{2Td\log\left(1+\frac{TL_\phi^2}{\lambda d}\right)}\right].$$

Finally, we relate the proxy regret to the cumulative regret. Using the finite-memory truncation and block decomposition argument from Proposition 1, for the action sequence generated by Algorithm 1,

$$\mathcal{R}_T^{\text{pseudo}}(\pi)\le 2T\Delta_m+\frac{2Tm}{m+L}R_m+\mathcal{R}_{\text{proxy}}\,.$$

This concludes the first part of theorem.

For the second part, we simplify the regret bound under a specific choice of m and L. Under Assumption 1, the lifted feature satisfies

$$\|\boldsymbol{\phi}_j(\mathbf{U})\|_2^2=\sum_{s=1}^{m}\|\mathbf{u}_{j-s}\otimes\mathbf{u}_j\|_2^2\le\sum_{s=1}^{m}\|\mathbf{u}_{j-s}\|_2^2\|\mathbf{u}_j\|_2^2\le m.$$

Hence, we may take $L_\phi\le\sqrt{m}$. Moreover, by Assumption 2,

$$S_m^2=\|\boldsymbol{\theta}_m^*\|_2^2=\sum_{h=0}^{m-1}\|\mathbf{G}_h\|_F^2=\sum_{h=0}^{m-1}\|\mathbf{C}\mathbf{A}^h\mathbf{B}\|_F^2\le\|\mathbf{C}\|_2^2\|\mathbf{B}\|_F^2M_\rho^2\sum_{h=0}^{m-1}\rho^{2h}.$$

Thus,

$$S_m\le\frac{\|\mathbf{C}\|_2\|\mathbf{B}\|_FM_\rho}{\sqrt{1-\rho^2}},$$

so S_m is uniformly bounded in m by a problem-dependent constant.

Choose

$$m=-\frac{1}{2}\log_\rho T,\qquad L=\frac{\sqrt{mT}}{d^{3/4}}-m,\qquad d=p^2m,$$

up to integer rounding, and set $\lambda = L_\phi^2$. Then $\rho^m = T^{-1/2}$. Since

$$\Delta_m \le \frac{\|\mathbf{C}\|_2 \|\mathbf{B}\|_2 M_\rho}{1-\rho} \rho^m,$$

we have

$$T\Delta_m = \widetilde{\mathcal{O}}(\sqrt{T}).$$

Similarly, from the definition of $\bar{\Delta}_m(\delta)$, we have

$$\bar{\Delta}_m(\delta) = \widetilde{\mathcal{O}}(T^{-1/2}).$$

We next bound the confidence radius. Using $NL \le T$, $L_\phi \le \sqrt{m}$, $\lambda = L_\phi^2$, and the uniform bound on S_m, we obtain

$$\gamma_N = \widetilde{\mathcal{O}}\left(\sqrt{d} + \sqrt{\lambda} S_m + \bar{\Delta}_m(\delta)\sqrt{Td}\right) = \widetilde{\mathcal{O}}(\sqrt{d}).$$

Indeed, $\sqrt{\lambda} S_m \le \sqrt{m}\, S_m$, which is dominated by $\sqrt{d} = p\sqrt{m}$ up to problem-dependent constants, and

$$\bar{\Delta}_m(\delta)\sqrt{Td} = \widetilde{\mathcal{O}}(\sqrt{d}).$$

The block-boundary term satisfies

$$\frac{Tm}{m+L} R_m = \widetilde{\mathcal{O}}\left(\frac{Tm}{L}\right) = \widetilde{\mathcal{O}}\left(d^{3/4}\sqrt{mT}\right),$$

where we used $R_m = \mathcal{O}(1)$ and $L \asymp \sqrt{mT}/d^{3/4}$.

Finally, the statistical term in Lemma 1 is bounded as

$$\begin{aligned}\gamma_N \left[\frac{LL_\phi d}{\sqrt{\lambda}} + \sqrt{Td \log\left(1 + \frac{TL_\phi^2}{\lambda d}\right)}\right] &= \widetilde{\mathcal{O}}\left(\sqrt{d}\left[Ld + \sqrt{Td}\right]\right) \\ &= \widetilde{\mathcal{O}}\left(d^{3/4}\sqrt{mT} + d\sqrt{T}\right).\end{aligned}$$

Combining the truncation error, the block-boundary error, and the statistical term yields

$$\mathcal{R}_T^{\mathrm{pseudo}}(\pi) = \widetilde{\mathcal{O}}\left(d^{3/4}\sqrt{mT} + d\sqrt{T}\right).$$

Since $d = p^2 m$ and $m = \frac{\log T}{2\log(1/\rho)}$, the above bound can also be written, up to logarithmic factors in T, as

$$\mathcal{R}_T^{\mathrm{pseudo}}(\pi) = \widetilde{\mathcal{O}}\left(\left[\frac{p^{3/2}}{\{\log(1/\rho)\}^{5/4}} + \frac{p^2}{\log(1/\rho)}\right]\sqrt{T}\right),$$

where $\widetilde{\mathcal{O}}$ hides logarithmic factors in T and δ^{-1}, as well as problem-dependent constants. In particular, for fixed p and fixed stability parameter ρ, this gives

$$\mathcal{R}_T^{\mathrm{pseudo}}(\pi) = \widetilde{\mathcal{O}}(\sqrt{T}).$$

□

B.4 Proof of Theorem 1

Theorem 1 (Actual expected regret of Algorithm 1)**.** *Suppose the assumptions of Lemma 1 hold. For a fixed $\delta \in (0,1)$, let $B_T(\delta)$ be the pseudo-regret bound established in Lemma 1. Define $\Delta_0 := \sum_{h=0}^{\infty} \|\mathbf{G}_h\|_2$. Then, we have*

$$\mathrm{Regret}_T(\pi) \le B_T(\delta) + 2\delta T\Delta_0 + \frac{\|\mathbf{C}\|_2 M_\rho}{1-\rho}\sqrt{\frac{p\lambda_{\max}(\mathbf{\Sigma}_w)}{1-\rho^2}}\frac{T}{m+L}.$$

Consequently, choosing $\delta = T^{-2}$ and the parameters $m = -\frac{1}{2}\log_\rho T$, $L = \frac{\sqrt{mT}}{d^{3/4}} - m$, $\lambda = L_\phi^2$, up to integer rounding and whenever $L \ge 1$, gives

$$\mathrm{Regret}_T(\pi) = \widetilde{O}\left(d^{3/4}\sqrt{mT} + d\sqrt{T}\right) = \widetilde{O}\left(\left[\frac{p^{3/2}}{\log(1/\rho)^{5/4}} + \frac{p^2}{\log(1/\rho)}\right]\sqrt{T}\right).$$

Proof of Theorem 1. First, we introduce a lemma that shows, for block-committed policies, pseudo-regret and actual expected regret differ only by a boundary-noise term of order $T/(m+L)$.

Lemma 8 (Blockwise adaptation gap)**.** *Suppose Assumptions 1–3 hold, and assume for simplicity that $N := T/(m+L) \in \mathbb{N}$. Consider any $(m+L)$-block-committed policy π, i.e., a policy that chooses the whole block*

$$\mathbf{U}_\tau = (\mathbf{u}_{\tau,1}, \dots, \mathbf{u}_{\tau,m+L})$$

measurably with respect to $\mathcal{H}_{\tau-1} := \mathcal{F}_{(\tau-1)(m+L)}$ before any noise in block τ is revealed, and then plays the block without within-block adaptation. Then

$$\left|\mathrm{Regret}_T(\pi) - \mathbb{E}_\pi\left[\mathcal{R}_T^{\mathrm{pseudo}}(\pi)\right]\right| \le \frac{\|\mathbf{C}\|_2 M_\rho}{1-\rho}\sqrt{\frac{p\,\lambda_{\max}(\mathbf{\Sigma}_w)}{1-\rho^2}}\,\frac{T}{m+L}\,.$$

Let $\mathcal{E}_\delta^\mu$ be the high-probability event from Lemma 1. On $\mathcal{E}_\delta^\mu$, we have

$$\mathcal{R}_T^{\mathrm{pseudo}}(\pi) \le B_T(\delta).$$

Moreover, since $|\mu_t(\mathbf{u}_{1:t})| \le \Delta_0$ for every action sequence,

$$0 \le \mathcal{R}_T^{\mathrm{pseudo}}(\pi) \le 2T\Delta_0$$

deterministically. Therefore,

$$\mathbb{E}_\pi[\mathcal{R}_T^{\mathrm{pseudo}}(\pi)] \le B_T(\delta) + 2\delta T\Delta_0.$$

Combining this with Lemma 8 yields

$$\mathrm{Regret}_T(\pi) \le B_T(\delta) + 2\delta T\Delta_0 + \frac{\|\mathbf{C}\|_2 M_\rho}{1-\rho}\sqrt{\frac{p\,\lambda_{\max}(\mathbf{\Sigma}_w)}{1-\rho^2}}\,\frac{T}{m+L}.$$

For the second claim, observe that

$$\Delta_0 := \sum_{h=0}^{\infty}\|\mathbf{G}_h\|_2 \le \frac{\|\mathbf{C}\|_2\|\mathbf{B}\|_2 M_\rho}{1-\rho}.$$

The stated rate follows by taking $\delta = T^{-2}$ and substituting the choices of m, L, and λ. In particular,

$$\frac{T}{m+L} \asymp \frac{d^{3/4}\sqrt{T}}{\sqrt{m}},$$

so the additional adaptation-gap term is dominated by the leading block-boundary and statistical terms up to logarithmic and problem-dependent factors. □

B.5 Proof of Technical Lemmas

B.5.1 Proof of Lemma 5

Proof of Lemma 5. Fix $\tau \in [N]$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_L)^\top \in \mathbb{R}^L$, and $\alpha \in \mathbb{R}$. For notational simplicity, write the informative position as $j = m + \ell$, where $\ell \in [L]$. By definition,

$$\epsilon_{\tau,m+\ell} = z_{t_{\tau,m+\ell}} + \sum_{h=0}^{m+\ell-2}\mathbf{u}_{\tau,m+\ell}^\top \mathbf{C}\mathbf{A}^h\mathbf{w}_{t_{\tau,m+\ell}-1-h}.$$

Hence,

$$\boldsymbol{\xi}^\top\boldsymbol{\epsilon}_\tau = \sum_{\ell=1}^{L}\xi_\ell z_{t_{\tau,m+\ell}} + \sum_{\ell=1}^{L}\sum_{h=0}^{m+\ell-2}\xi_\ell\mathbf{u}_{\tau,m+\ell}^\top\mathbf{C}\mathbf{A}^h\mathbf{w}_{t_{\tau,m+\ell}-1-h}.$$

We reindex the process-noise terms by the within-block noise index $q \in \{1, \dots, m+L-1\}$, where the corresponding global index is $(\tau-1)(m+L)+q$. Then

$$\boldsymbol{\xi}^\top\boldsymbol{\epsilon}_\tau = \sum_{\ell=1}^{L}\xi_\ell z_{t_{\tau,m+\ell}} + \sum_{q=1}^{m+L-1}\left\langle \mathbf{g}_{\tau,q}(\boldsymbol{\xi}), \mathbf{w}_{(\tau-1)(m+L)+q}\right\rangle,$$

where

$$\mathbf{g}_{\tau,q}(\boldsymbol{\xi}) := \sum_{\ell=1}^{L} \mathbf{1}\{q \le m+\ell-1\}\xi_\ell (\mathbf{A}^{m+\ell-1-q})^\top \mathbf{C}^\top \mathbf{u}_{\tau,m+\ell} \in \mathbb{R}^n .$$

Conditioned on $\mathcal{H}_{\tau-1}$, the actions in $\mathbf{U}_\tau$ are fixed, and the noise variables in the current block are conditionally sub-Gaussian. The reward-noise part satisfies

$$\mathbb{E}\left[\exp\left(\alpha \sum_{\ell=1}^{L} \xi_\ell z_{t_{\tau,m+\ell}}\right)\middle| \mathcal{H}_{\tau-1}\right] \le \exp\left(\frac{\alpha^2\sigma_z^2}{2}\|\boldsymbol{\xi}\|_2^2\right) .$$

For the process-noise part, Assumption 3 gives

$$\mathbb{E}\big[\exp\big(\alpha \left\langle \mathbf{g}_{\tau,q}(\boldsymbol{\xi}), \mathbf{w}_{(\tau-1)(m+L)+q}\right\rangle\big)\big|\mathcal{H}_{\tau-1}\big] \le \exp\left(\frac{\alpha^2}{2}\mathbf{g}_{\tau,q}(\boldsymbol{\xi})^\top \boldsymbol{\Sigma}_w \mathbf{g}_{\tau,q}(\boldsymbol{\xi})\right).$$

Therefore, the process-noise contribution has variance proxy at most

$$\lambda_{\max}(\boldsymbol{\Sigma}_w) \sum_{q=1}^{m+L-1} \|\mathbf{g}_{\tau,q}(\boldsymbol{\xi})\|_2^2.$$

It remains to bound $\sum_{q=1}^{m+L-1} \|\mathbf{g}_{\tau,q}(\boldsymbol{\xi})\|_2^2$. Using $\|\mathbf{u}_{\tau,m+\ell}\|_2 \le 1$ and $\|\mathbf{A}^h\|_2 \le M_\rho \rho^h$, we obtain

$$\|\mathbf{g}_{\tau,q}(\boldsymbol{\xi})\|_2 \le \sum_{\ell=1}^{L} \mathbf{1}\{q \le m+\ell-1\}|\xi_\ell|\|\mathbf{C}\|_2 M_\rho \rho^{m+\ell-1-q} .$$

This is bounded by the convolution of $(|\xi_\ell|)_{\ell=1}^L$ with the nonnegative sequence

$$c_a := \|\mathbf{C}\|_2 M_\rho \rho^a, \qquad a \ge 0.$$

By Young's convolution inequality,

$$\sum_{q=1}^{m+L-1} \|\mathbf{g}_{\tau,q}(\boldsymbol{\xi})\|_2^2 \le \left(\sum_{a=0}^{\infty} c_a\right)^2 \|\boldsymbol{\xi}\|_2^2 = \frac{\|\mathbf{C}\|_2^2 M_\rho^2}{(1-\rho)^2}\|\boldsymbol{\xi}\|_2^2.$$

Combining the reward-noise and process-noise bounds yields

$$\mathbb{E}\Big[\exp\Big(\alpha\boldsymbol{\xi}^\top \boldsymbol{\epsilon}_\tau\Big)\,\Big|\,\mathcal{H}_{\tau-1}\Big] \le \exp\left(\frac{\alpha^2}{2}\left[\sigma_z^2 + \lambda_{\max}(\boldsymbol{\Sigma}_w)\frac{\|\mathbf{C}\|_2^2 M_\rho^2}{(1-\rho)^2}\right]\|\boldsymbol{\xi}\|_2^2\right) .$$

□

B.5.2 Proof of Lemma 6

Proof of Lemma 6. Fix an arbitrary vector $\mathbf{v} \in \mathbb{R}^d$. Since $\boldsymbol{\Phi}_\tau$ is $\mathcal{H}_{\tau-1}$-measurable, the vector $\boldsymbol{\Phi}_\tau^\top \mathbf{v} \in \mathbb{R}^L$ is also $\mathcal{H}_{\tau-1}$-measurable. By the conditional vector sub-Gaussianity assumption, applied with $\boldsymbol{\xi} = \boldsymbol{\Phi}_\tau^\top \mathbf{v}$, we have

$$\mathbb{E}\big[\exp\big(\mathbf{v}^\top \boldsymbol{\Phi}_\tau \boldsymbol{\epsilon}_\tau\big)\,\big|\,\mathcal{H}_{\tau-1}\big] \le \exp\left(\frac{\sigma^2}{2}\|\boldsymbol{\Phi}_\tau^\top \mathbf{v}\|_2^2\right).$$

Equivalently,

$$\mathbb{E}\left[\exp\left(\mathbf{v}^\top \boldsymbol{\Phi}_\tau \boldsymbol{\epsilon}_\tau - \frac{\sigma^2}{2}\mathbf{v}^\top \boldsymbol{\Phi}_\tau \boldsymbol{\Phi}_\tau^\top \mathbf{v}\right)\,\middle|\,\mathcal{H}_{\tau-1}\right] \le 1.$$

Define

$$\mathbf{S}_\tau := \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\epsilon}_s, \qquad \mathbf{V}_\tau := \mathbf{V}_0 + \sum_{s=1}^{\tau} \boldsymbol{\Phi}_s \boldsymbol{\Phi}_s^\top .$$

For each fixed $\mathbf{v} \in \mathbb{R}^d$, define

$$M_\tau(\mathbf{v}) := \exp\left(\mathbf{v}^\top \mathbf{S}_\tau - \frac{\sigma^2}{2}\mathbf{v}^\top(\mathbf{V}_\tau - \mathbf{V}_0)\mathbf{v}\right).$$

Then $(M_\tau(\mathbf{v}))_{\tau \geq 0}$ is a nonnegative supermartingale. Indeed,

$$\begin{aligned}\mathbb{E}[M_\tau(\mathbf{v}) \,|\, \mathcal{H}_{\tau-1}] &= M_{\tau-1}(\mathbf{v})\, \mathbb{E}\left[\exp\left(\mathbf{v}^\top \mathbf{\Phi}_\tau \boldsymbol{\epsilon}_\tau - \frac{\sigma^2}{2}\mathbf{v}^\top \mathbf{\Phi}_\tau \mathbf{\Phi}_\tau^\top \mathbf{v}\right) \,\middle|\, \mathcal{H}_{\tau-1}\right] \\ &\leq M_{\tau-1}(\mathbf{v}).\end{aligned}$$

Moreover, $M_0(\mathbf{v}) = 1$.

We now use the standard mixture argument of Abbasi-Yadkori et al. [2]. Let $\mathbf{\Lambda}$ be a Gaussian random vector, independent of all other randomness, with distribution

$$\mathbf{\Lambda} \sim \mathcal{N}\big(\mathbf{0}_d,\, \sigma^{-2}\mathbf{V}_0^{-1}\big)\,,$$

and let μ denote its distribution. Define the mixture process

$$M_\tau := \mathbb{E}_{\mathbf{\Lambda}}\left[M_\tau(\mathbf{\Lambda})\right] = \int_{\mathbb{R}^d} M_\tau(\mathbf{v})\, \mu(d\mathbf{v}).$$

Since $M_\tau(\mathbf{v}) \geq 0$, Fubini's theorem allows us to interchange the mixture integral and conditional expectation. Hence,

$$\begin{aligned}\mathbb{E}[M_\tau \mid \mathcal{H}_{\tau-1}] &= \mathbb{E}\left[\int_{\mathbb{R}^d} M_\tau(\mathbf{v})\, \mu(d\mathbf{v}) \,\middle|\, \mathcal{H}_{\tau-1}\right] \\ &= \int_{\mathbb{R}^d} \mathbb{E}[M_\tau(\mathbf{v}) \mid \mathcal{H}_{\tau-1}]\, \mu(d\mathbf{v}) \\ &\leq \int_{\mathbb{R}^d} M_{\tau-1}(\mathbf{v})\, \mu(d\mathbf{v}) = M_{\tau-1}.\end{aligned}$$

Therefore, $(M_\tau)_{\tau \geq 0}$ is a nonnegative supermartingale. Also, since $M_0(\mathbf{v}) = 1$ for every $\mathbf{v}$, we have $M_0 = 1$.

We compute M_τ explicitly. The density of $\mathbf{\Lambda}$ is

$$f(\mathbf{v}) = \frac{\det(\sigma^2\mathbf{V}_0)^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{\sigma^2}{2}\mathbf{v}^\top \mathbf{V}_0 \mathbf{v}\right).$$

Hence,

$$\begin{aligned}M_\tau &= \int_{\mathbb{R}^d} \exp\left(\mathbf{v}^\top \mathbf{S}_\tau - \frac{\sigma^2}{2}\mathbf{v}^\top(\mathbf{V}_\tau - \mathbf{V}_0)\mathbf{v}\right) f(\mathbf{v})\, d\mathbf{v} \\ &= \frac{\det(\sigma^2\mathbf{V}_0)^{1/2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp\left(\mathbf{v}^\top \mathbf{S}_\tau - \frac{\sigma^2}{2}\mathbf{v}^\top \mathbf{V}_\tau \mathbf{v}\right) d\mathbf{v}.\end{aligned}$$

Completing the square gives

$$\mathbf{v}^\top \mathbf{S}_\tau - \frac{\sigma^2}{2}\mathbf{v}^\top \mathbf{V}_\tau \mathbf{v} = -\frac{\sigma^2}{2}\left\|\mathbf{v} - \sigma^{-2}\mathbf{V}_\tau^{-1}\mathbf{S}_\tau\right\|_{\mathbf{V}_\tau}^2 + \frac{1}{2\sigma^2}\|\mathbf{S}_\tau\|_{\mathbf{V}_\tau^{-1}}^2.$$

Therefore,

$$\begin{aligned}M_\tau &= \frac{\det(\sigma^2\mathbf{V}_0)^{1/2}}{\det(\sigma^2\mathbf{V}_\tau)^{1/2}} \exp\left(\frac{1}{2\sigma^2}\|\mathbf{S}_\tau\|_{\mathbf{V}_\tau^{-1}}^2\right) \\ &= \frac{\det(\mathbf{V}_0)^{1/2}}{\det(\mathbf{V}_\tau)^{1/2}} \exp\left(\frac{1}{2\sigma^2}\|\mathbf{S}_\tau\|_{\mathbf{V}_\tau^{-1}}^2\right).\end{aligned}$$

By Ville's inequality for nonnegative supermartingales,

$$\mathbb{P}\left(\exists \tau \geq 0 : M_\tau \geq \frac{1}{\delta}\right) \leq \delta.$$

Thus, with probability at least $1-\delta$, for all $\tau \geq 0$,

$$M_\tau < \frac{1}{\delta}.$$

Using the explicit expression for M_τ, this implies

$$\frac{\det(\mathbf{V}_0)^{1/2}}{\det(\mathbf{V}_\tau)^{1/2}} \exp\left(\frac{1}{2\sigma^2}\|\mathbf{S}_\tau\|^2_{\mathbf{V}_\tau^{-1}}\right) < \frac{1}{\delta}.$$

Equivalently,

$$\|\mathbf{S}_\tau\|^2_{\mathbf{V}_\tau^{-1}} \leq 2\sigma^2 \log\left(\frac{\det(\mathbf{V}_\tau)^{1/2}}{\det(\mathbf{V}_0)^{1/2}\delta}\right).$$

Taking square roots gives

$$e\|\mathbf{S}_\tau\|_{\mathbf{V}_\tau^{-1}} \leq \sigma\sqrt{2\log\left(\frac{\det(\mathbf{V}_\tau)^{1/2}}{\det(\mathbf{V}_0)^{1/2}\delta}\right)}.$$

This holds simultaneously for all $\tau \geq 0$, which completes the proof. □

B.5.3 Proof of Lemma 7

Proof of Lemma 7. We first bound the action-tail bias $\eta_{\tau,j}$. Since $\|\mathbf{u}_{\tau,j}\|_2 \leq 1$ and $\|\mathbf{u}_{t_{\tau,j}-s}\|_2 \leq 1$, we have

$$\begin{aligned}
|\eta_{\tau,j}| &= \left|\sum_{s=m+1}^{\infty} \mathbf{u}_{\tau,j}^\top \mathbf{G}_{s-1}\mathbf{u}_{t_{\tau,j}-s}\right| \\
&\leq \sum_{s=m+1}^{\infty} \|\mathbf{u}_{\tau,j}\|_2\|\mathbf{G}_{s-1}\|_2\|\mathbf{u}_{t_{\tau,j}-s}\|_2 \\
&\leq \sum_{s=m+1}^{\infty} \|\mathbf{G}_{s-1}\|_2 \\
&= \sum_{h=m}^{\infty} \|\mathbf{G}_h\|_2 \\
&\leq \sum_{h=m}^{\infty} \|\mathbf{C}\|_2\|\mathbf{A}^h\|_2\|\mathbf{B}\|_2 \\
&\leq \frac{\|\mathbf{C}\|_2\|\mathbf{B}\|_2 M_\rho}{1-\rho}\rho^m =: \Delta_m\,.
\end{aligned}$$

It remains to control the process-noise tail $\zeta_{\tau,j}$. Because $\mathbf{u}_{\tau,j}$ may be chosen adaptively based on past rewards, we do not directly claim that $\zeta_{\tau,j}$ is centered sub-Gaussian after conditioning on the selected action. Instead, define the vector-valued tail

$$\mathbf{y}_{\tau,j} := \sum_{h=j-1}^{\infty} \mathbf{C}\mathbf{A}^h\mathbf{w}_{t_{\tau,j}-1-h} \in \mathbb{R}^p.$$

Then, by $\|\mathbf{u}_{\tau,j}\|_2 \leq 1$,

$$|\zeta_{\tau,j}| = |\mathbf{u}_{\tau,j}^\top \mathbf{y}_{\tau,j}| \leq \|\mathbf{y}_{\tau,j}\|_2.$$

Hence, it suffices to control $\|\mathbf{y}_{\tau,j}\|_2$.

Fix (τ, j). For any unit vector $\mathbf{v} \in \mathbb{S}^{p-1}$, we have

$$\mathbf{v}^\top \mathbf{y}_{\tau,j} = \sum_{h=j-1}^{\infty} \mathbf{v}^\top \mathbf{C}\mathbf{A}^h\mathbf{w}_{t_{\tau,j}-1-h}.$$

This is a centered sub-Gaussian random variable with variance proxy bounded by

$$\sum_{h=j-1}^{\infty} \lambda_{\max}(\mathbf{\Sigma}_w) \left\|(\mathbf{A}^h)^\top \mathbf{C}^\top \mathbf{v}\right\|_2^2 \le \lambda_{\max}(\mathbf{\Sigma}_w) \sum_{h=j-1}^{\infty} \|\mathbf{v}\|_2^2 \|\mathbf{C}\|_2^2 \|\mathbf{A}^h\|_2^2$$
$$\le \lambda_{\max}(\mathbf{\Sigma}_w) \|\mathbf{C}\|_2^2 M_\rho^2 \sum_{h=j-1}^{\infty} \rho^{2h}.$$

Since $j \ge m+1$, we have $j-1 \ge m$, and therefore

$$\sum_{h=j-1}^{\infty} \rho^{2h} \le \sum_{h=m}^{\infty} \rho^{2h} = \frac{\rho^{2m}}{1-\rho^2}.$$

Thus, every one-dimensional projection of $\mathbf{y}_{\tau,j}$ is sub-Gaussian with variance proxy at most

$$\sigma_{\text{tail}}^2 := \lambda_{\max}(\mathbf{\Sigma}_w) \frac{\|\mathbf{C}\|_2^2 M_\rho^2}{1-\rho^2} \rho^{2m}.$$

In particular, for each fixed $\mathbf{v} \in \mathbb{S}^{p-1}$ and every $x > 0$,

$$\mathbb{P}\left(|\mathbf{v}^\top \mathbf{y}_{\tau,j}| > x\right) \le 2\exp\left(-\frac{x^2}{2\sigma_{\text{tail}}^2}\right).$$

Let $\mathcal{N}_{1/2}$ be a $1/2$-net of $\mathbb{S}^{p-1}$ with $|\mathcal{N}_{1/2}| \le 5^p$. Then,

$$\|\mathbf{y}_{\tau,j}\|_2 \le 2 \max_{\mathbf{v} \in \mathcal{N}_{1/2}} |\mathbf{v}^\top \mathbf{y}_{\tau,j}|.$$

Indeed, if $\mathbf{y}_{\tau,j} = 0$, the claim is trivial. Otherwise, set

$$\mathbf{u}^* := \frac{\mathbf{y}_{\tau,j}}{\|\mathbf{y}_{\tau,j}\|_2}.$$

By the definition of a $1/2$-net, there exists $\mathbf{v} \in \mathcal{N}_{1/2}$ such that $\|\mathbf{u}^* - \mathbf{v}\|_2 \le 1/2$. Hence,

$$\begin{aligned}
\|\mathbf{y}_{\tau,j}\|_2 &= \langle \mathbf{u}^*, \mathbf{y}_{\tau,j} \rangle \\
&= \langle \mathbf{v}, \mathbf{y}_{\tau,j} \rangle + \langle \mathbf{u}^* - \mathbf{v}, \mathbf{y}_{\tau,j} \rangle \\
&\le |\langle \mathbf{v}, \mathbf{y}_{\tau,j} \rangle| + \|\mathbf{u}^* - \mathbf{v}\|_2 \|\mathbf{y}_{\tau,j}\|_2 \\
&\le \max_{\mathbf{v}' \in \mathcal{N}_{1/2}} |\langle \mathbf{v}', \mathbf{y}_{\tau,j} \rangle| + \frac{1}{2} \|\mathbf{y}_{\tau,j}\|_2.
\end{aligned}$$

Rearranging gives the claim.

Now, using the projection tail bound and a union bound over $\mathcal{N}_{1/2}$,

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{y}_{\tau,j}\|_2 > \Delta'_m(\delta)\right) &\le \mathbb{P}\left(\max_{\mathbf{v} \in \mathcal{N}_{1/2}} |\mathbf{v}^\top \mathbf{y}_{\tau,j}| > \frac{\Delta'_m(\delta)}{2}\right) \\
&\le \sum_{\mathbf{v} \in \mathcal{N}_{1/2}} \mathbb{P}\left(|\mathbf{v}^\top \mathbf{y}_{\tau,j}| > \frac{\Delta'_m(\delta)}{2}\right) \\
&\le 2|\mathcal{N}_{1/2}| \exp\left(-\frac{(\Delta'_m(\delta))^2}{8\sigma_{\text{tail}}^2}\right).
\end{aligned}$$

Therefore, choosing

$$\Delta'_m(\delta) := 2\sigma_{\text{tail}} \sqrt{2 \log \frac{2 \cdot 5^p T}{\delta}}$$

gives

$$\mathbb{P}\left(\|\mathbf{y}_{\tau,j}\|_2 > \Delta'_m(\delta)\right) \le \frac{\delta}{T}.$$

Since $|\zeta_{\tau,j}| \le \|\mathbf{y}_{\tau,j}\|_2$, we also have

$$\mathbb{P}\left(|\zeta_{\tau,j}| > \Delta'_m(\delta)\right) \le \frac{\delta}{T}.$$

Since the number of informative indices is

$$NL = \frac{T}{m+L} L \le T,$$

a union bound gives

$$\mathbb{P}\left(\exists \tau \in [N],\ \exists j \in \{m+1, \dots, m+L\} : |\zeta_{\tau,j}| > \Delta'_m(\delta)\right) \le \delta.$$

On the complement of this event, for all $\tau \in [N]$ and $j \in \{m+1, \dots, m+L\}$,

$$|\eta_{\tau,j} + \zeta_{\tau,j}| \le |\eta_{\tau,j}| + |\zeta_{\tau,j}| \le \Delta_m + \Delta'_m(\delta) = \bar{\Delta}_m(\delta).$$

□

B.5.4 Proof of Lemma 8

Proof of Lemma 8. For any time t, unrolling the latent dynamics gives

$$r_t = \mu_t(\mathbf{u}_{1:t}) + z_t + \mathbf{u}_t^\top \mathbf{C} \sum_{h=0}^{\infty} \mathbf{A}^h \mathbf{w}_{t-1-h},$$

where we use the convention $\mathbf{w}_s = \mathbf{0}_n$ for $s \le 0$. Since z_t is zero-mean and is revealed after $\mathbf{u}_t$ is chosen, $\mathbb{E}[z_t] = 0$. Therefore,

$$\mathrm{Regret}_T(\pi) - \mathbb{E}_\pi[\mathcal{R}_T^{\mathrm{pseudo}}(\pi)] = -\mathbb{E}_\pi\left[\sum_{t=1}^{T} \mathbf{u}_t^\top \mathbf{C} \sum_{h=0}^{\infty} \mathbf{A}^h \mathbf{w}_{t-1-h}\right].$$

It remains to bound the absolute value of the right-hand side.

Let $t_{\tau,j} := (\tau-1)(m+L) + j$, where $\tau \in [N]$ and $j \in [m+L]$. For the process-noise contribution at time $t_{\tau,j}$, decompose

$$\mathbf{u}_{\tau,j}^\top \mathbf{C} \sum_{h=0}^{\infty} \mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h} = \underbrace{\mathbf{u}_{\tau,j}^\top \mathbf{C} \sum_{h=0}^{j-2} \mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}}_{=:a_{\tau,j}} + \underbrace{\mathbf{u}_{\tau,j}^\top \mathbf{C} \sum_{h=j-1}^{\infty} \mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}}_{=:b_{\tau,j}}.$$

Here, $a_{\tau,j}$ is the contribution of process noises generated within the current block, whereas $b_{\tau,j}$ is the contribution of process noises generated before the current block. Because $\mathbf{U}_\tau$ is $\mathcal{H}_{\tau-1}$-measurable, all actions $\mathbf{u}_{\tau,j}$ are fixed before the noises in block τ are revealed. By the martingale-difference property of the process noise and the tower property,

$$\mathbb{E}[a_{\tau,j} \mid \mathcal{H}_{\tau-1}] = 0, \qquad \text{and hence} \qquad \mathbb{E}[a_{\tau,j}] = 0.$$

It remains to control $b_{\tau,j}$. Let

$$\mathbf{y}_{\tau,j} := \mathbf{C} \sum_{h=j-1}^{\infty} \mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}.$$

Then $b_{\tau,j} = \mathbf{u}_{\tau,j}^\top \mathbf{y}_{\tau,j}$. Using $\|\mathbf{u}_{\tau,j}\|_2 \le 1$,

$$|\mathbb{E}[b_{\tau,j}]| \le \mathbb{E}[\|\mathbf{y}_{\tau,j}\|_2].$$

Since $\mathbf{w}_t$ is conditionally sub-Gaussian with variance proxy $\mathbf{\Sigma}_w$, we have $\mathbb{E}[\mathbf{w}_t \mathbf{w}_t^\top \mid \mathcal{F}_{t-1}] \preceq \mathbf{\Sigma}_w$. Moreover, cross terms between different time indices vanish by the martingale-difference property. Hence,

$$\mathbb{E}\left[\|\mathbf{y}_{\tau,j}\|_2^2\right] \le \sum_{h=j-1}^{\infty} \mathbb{E}\left[\|\mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}\|_2^2\right].$$

For each h, by the covariance bound and Assumption 2,

$$\mathbb{E}\big[\|\mathbf{C}\mathbf{A}^h \mathbf{w}_{t_{\tau,j}-1-h}\|_2^2\big] \le \lambda_{\max}(\mathbf{\Sigma}_w)\|\mathbf{C}\mathbf{A}^h\|_F^2 \le p\,\lambda_{\max}(\mathbf{\Sigma}_w)\|\mathbf{C}\|_2^2 M_\rho^2 \rho^{2h}.$$

Therefore,

$$\mathbb{E}\big[\|\mathbf{y}_{\tau,j}\|_2^2\big] \le p\,\lambda_{\max}(\mathbf{\Sigma}_w)\|\mathbf{C}\|_2^2 M_\rho^2 \frac{\rho^{2(j-1)}}{1-\rho^2}.$$

By Jensen's inequality,

$$\mathbb{E}[\|\mathbf{y}_{\tau,j}\|_2] \le \|\mathbf{C}\|_2 M_\rho \sqrt{\frac{p\,\lambda_{\max}(\mathbf{\Sigma}_w)}{1-\rho^2}}\rho^{j-1}.$$

Summing over all positions in a block gives

$$\sum_{j=1}^{m+L} |\mathbb{E}[b_{\tau,j}]| \le \|\mathbf{C}\|_2 M_\rho \sqrt{\frac{p\,\lambda_{\max}(\mathbf{\Sigma}_w)}{1-\rho^2}} \sum_{j=1}^{m+L} \rho^{j-1} \le \frac{\|\mathbf{C}\|_2 M_\rho}{1-\rho}\sqrt{\frac{p\,\lambda_{\max}(\mathbf{\Sigma}_w)}{1-\rho^2}}.$$

Finally, summing over the $N = T/(m+L)$ blocks yields

$$\Big|\mathrm{Regret}_T(\pi) - \mathbb{E}_\pi[\mathcal{R}_T^{\mathrm{pseudo}}(\pi)]\Big| \le \frac{\|\mathbf{C}\|_2 M_\rho}{1-\rho}\sqrt{\frac{p\,\lambda_{\max}(\mathbf{\Sigma}_w)}{1-\rho^2}}\frac{T}{m+L}.$$

□

C Auxiliary Lemmas

Lemma 9 (Theorem 1 in Abbasi-Yadkori et al. [2]). *Let $\{F_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R \ge 0$. Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is F_{t-1}-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \ge 0$, define*

$$\overline{V}_t = V + \sum_{s=1}^t X_s X_s^\top, \qquad S_t = \sum_{s=1}^t \eta_s X_s.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \ge 0$,

$$\|S_t\|_{\overline{V}_t^{-1}}^2 \le 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2}\det(V)^{-1/2}}{\delta}\right).$$

Lemma 10 (Lemma 15 in Faury et al. [15]). *Let $\{x_s\}_{s=1}^\infty$ be a sequence in $\mathbb{R}^d$ such that $\|x_s\|_2 \le X$ for all $s \in \mathbb{N}$, and let λ be a non-negative scalar. For $t \ge 1$, define $V_t := \sum_{s=1}^{t-1} x_s x_s^\top + \lambda I_d$. The following inequality holds:*

$$\sum_{t=1}^T \|x_t\|_{V_t^{-1}}^2 \le 2\max\left(1, \frac{X^2}{\lambda}\right)\log\frac{\det(V_{T+1})}{\lambda^d}.$$

Lemma 11 (Lemma 10 in Abbasi-Yadkori et al. [2]). *Suppose $X_1, \dots, X_t \in \mathbb{R}^d$ and for any $1 \le s \le t$, $\|X_s\|_2 \le L$, Let $\mathbf{V}_t = \lambda \mathbf{I}_d + \sum_{s=1}^t X_s X_s^\top$ for some $\lambda > 0$. Then,*

$$\det(\mathbf{V}_t) \le (\lambda + tL^2/d)^d.$$

Lemma 12 (Lemma 3 in Jin et al. [17]). *Let $A \in \mathbb{R}^{d\times d}$ be any symmetric positive definite matrix, and let $u_1, u_2, \dots, u_L \in \mathbb{R}^d$ be any vectors. Let $\lambda_1, \lambda_2, \dots, \lambda_d$ be the eigenvalues of A, and let $\nu_1, \nu_2, \dots, \nu_d$ be the eigenvalues of $A + \sum_{k=1}^L u_k u_k^\top$. We have, for any $x \in \mathbb{R}^d$ such that $\|x\|_2 = 1$,*

$$\sqrt{x^\top A^{-1} x} - \sqrt{x^\top \left(A + \sum_{k=1}^L u_k u_k^\top\right)^{-1} x} \le \sum_{i=1}^d \frac{2}{\sqrt{\lambda_i}} - \sum_{i=1}^d \frac{2}{\sqrt{\nu_i}}.$$

D Experimental Details

In this appendix, we provide details of the experiments reported in Section 4.

D.1 Implementation of `Block-OFUL`

For the continuous hypercube $\mathcal{U} = [-1, 1]^p$, there exists an optimal open-loop action sequence whose actions all lie at vertices of the hypercube [11]. Motivated by this property and to enable a direct comparison with the ETC baseline, we restrict the experimental action set to $\mathcal{U} = \{-1, +1\}^p$.

Even with this restriction, exhaustive maximization of the block-UCB objective over all $2^{p(m+L)}$ candidate blocks is computationally prohibitive. We therefore implement the oracle using a multi-start local-search procedure. For each restart, we initialize a block by sampling its actions uniformly from $\mathcal{U}$, and repeatedly apply single-coordinate sign flips whenever they improve the block-UCB objective. The local search terminates when no further improving flip is found. We repeat this procedure from multiple random initializations and select the locally optimized block with the largest UCB objective among all restarts.

D.2 Additional Experiments

Environment. We further evaluate `Block-OFUL` on a structured delayed latent-dynamics environment designed to isolate the effects of delayed action influence and persistent latent memory. The latent state is partitioned into D blocks of dimension p, with $n = pD$, and the action set is $\mathcal{U} = \{-1, +1\}^p$. The system matrices are constructed so that each action is injected into the first latent block, propagated sequentially through the latent state with decay factor ρ, and affects the reward only after reaching the final block:

$$\mathbf{A} = \rho \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I}_p & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{I}_p & \mathbf{0} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{I}_p \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \qquad \mathbf{C} = [\mathbf{0} \quad \mathbf{0} \quad \cdots \quad \mathbf{C}_D],$$

where

$$\mathbf{C}_D = c_1 \mathbf{I}_p + \frac{c_2}{p-1} \left(\mathbf{1}\mathbf{1}^\top - \mathbf{I}_p \right),$$

with $c_1 = 1$ and $c_2 = 0.35$. This construction creates an explicit delayed input–output dependence: increasing D delays the reward-relevant effect of an action, while increasing ρ makes this effect decay more slowly. We use $T = 1000$, $p = 2$, $D \in \{3, 6\}$, and $\rho \in \{0.3, 0.95\}$, with Gaussian reward and process noise.

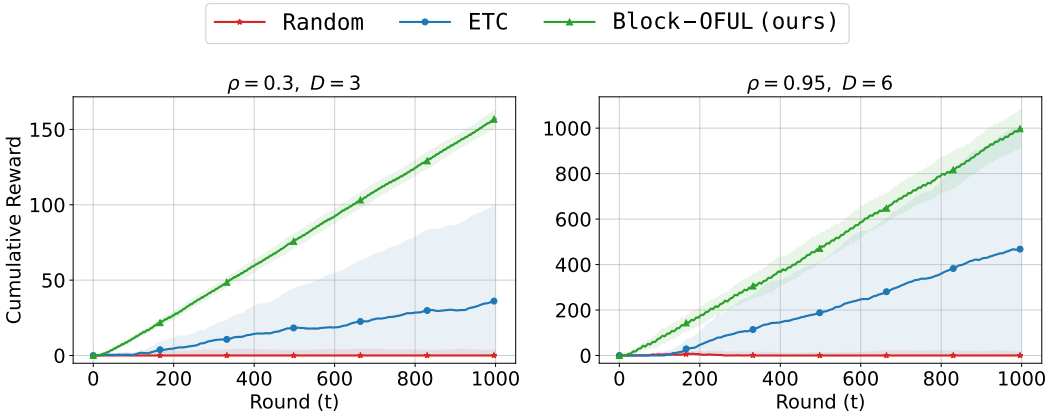

Figure 2: Cumulative reward on delayed latent linear-dynamics bandit instances with different levels of stability and delay.

Figure 2 shows that Block-OFUL consistently achieves higher cumulative reward than ETC and the random baseline in the delayed latent-dynamics environments. The advantage becomes more pronounced in the setting with larger D and ρ close to one, where useful reward information arrives only after a longer delay and the effects of past actions persist for longer. This regime is particularly challenging for ETC because the quality of the committed action sequence depends entirely on the dynamics estimate obtained during its initial exploration phase. In contrast, Block-OFUL continues to incorporate newly observed rewards and revise its block-level decisions throughout the horizon. These results complement the main experiments by showing that the benefit of adaptive block-level optimism persists in environments with explicitly delayed and long-lasting action effects.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the problem setting, the proposed cyclic approximation and block-level optimistic algorithm, and the resulting regret guarantee. They also clearly compare the result with the previous explore-then-commit guarantee for the same model and describe the scope of the contribution as an open-loop benchmark for latent linear-dynamics bandits with bilinear rewards.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper notes that the theoretical analysis assumes access to an exact oracle for block-UCB maximization, an oracle-based formulation also used in prior combinatorial and block-based bandit literature. It discusses computationally efficient oracle design with corresponding regret guarantees as a future direction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theorem, proposition, and lemma states the required assumptions explicitly, and the appendix provides complete proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the synthetic latent-dynamics environment, including the construction of the system matrices, noise distributions, parameter choices, baselines, algorithmic hyperparameters, evaluation metric, and number of independent runs used for averaging the reported cumulative rewards.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use synthetic data generated from the specified latent linear-dynamics model. The supplemental material provides the code and instructions for reproducing the main experimental results, including the environment construction, algorithmic inputs, and plotting procedure.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the synthetic environment, parameter settings, noise levels, baselines, horizon, number of runs, and algorithmic hyperparameters used in the experiments. Since the experiments are synthetic bandit simulations, there are no train/test data splits; the relevant evaluation details are provided through the environment and algorithm configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The experimental results are averaged over 10 independent runs, and the plots report variability across runs using shaded error bands. The paper also specifies how the averages are computed and what randomness is varied across runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The experiments are lightweight synthetic bandit simulations and do not require specialized hardware such as GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is a theoretical and synthetic experimental study of bandit algorithms. It does not involve human subjects, personal data, sensitive data, or deployment in high-stakes settings, and we are not aware of any ethical concerns raised by the work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper focuses on theoretical algorithmic development and synthetic experiments for latent linear-dynamics bandits. It does not make application-specific deployment claims, and the work is not expected to have direct societal impacts beyond general advances in sequential decision-making methodology.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets. The experiments use synthetic data generated from the specified latent linear-dynamics model, and no pretrained generative models, scraped datasets, or sensitive data are involved.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not rely on external datasets, pretrained models, or third-party assets. The experiments use synthetic data generated from the specified latent linear-dynamics model, and all referenced prior works are cited appropriately.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces code for synthetic bandit experiments, and the supplemental material documents how to reproduce the experiments, including the environment construction, algorithm inputs, hyperparameters, and plotting procedure.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.