Explore-then-Commit for Nonstationary Linear Bandits with Latent Dynamics

Sunmook Choi Cornell
Yahya Sattar Cornell
Yassir Jedra Imperial College London
Maryam Fazel U Washington
Sarah Dean Cornell

Abstract

We study a nonstationary bandit problem where rewards depend on both actions and latent states, the latter governed by unknown linear dynamics. Crucially, the state dynamics also depend on the actions, resulting in tension between short-term and long-term rewards. We propose an explore-then-commit algorithm for a finite horizon T. During the exploration phase, random Rademacher actions enable estimation of the Markov parameters of the linear dynamics, which characterize the action-reward relationship. In the commit phase, the algorithm uses the estimated parameters to design an optimized action sequence for long-term reward. Our proposed algorithm achieves $\tilde{\mathcal{O}}(pT^{2/3})$ regret where p is the action dimension. Our analysis handles two key challenges: learning from temporally correlated rewards, and designing action sequences with optimal long-term reward. We address the first challenge by providing near-optimal sample complexity and error bounds for system identification using bilinear rewards. We address the second challenge by proving an equivalence with indefinite quadratic optimization over a hypercube, a known NP-hard problem. We provide a sub-optimality guarantee for this problem, enabling our regret upper bound. Lastly, we propose a semidefinite relaxation with Goemans-Williamson rounding as a practical approach.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Many application domains, like personalized recommendations or online advertising, require sequential decision-making under uncertainty. Classical bandit algorithms address the trade-off between reducing uncertainty and optimizing performance in environments where rewards do not depend on the algorithm's past decisions. Many real world systems exhibit temporal dependencies – actions influence not only immediate reward, but also the future state of the environment (Schedl et al., 2018). This paper studies such a setting where decisions propagate through latent dynamics, leading to correlations that fundamentally change both how to learn and how to act optimally.

We study a nonstationary bandit problem in which reward depends *bilinearly* on the *current action* and an *unobserved latent state* that evolves according to a stable linear dynamical system. Formally,

$$r_t = \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{x}_t + z_t, \quad \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{w}_t, \quad (1)$$

where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are unknown matrices, $\boldsymbol{u}_t$ is a bounded action chosen by the learner, $\boldsymbol{w}_t, z_t$ are random noise processes, and $\boldsymbol{x}_t$ is the latent state, at time $t \geq 0$. Unlike classical multi-armed bandits with i.i.d. rewards, here actions not only determine immediate payoffs but also propagate through the state dynamics, creating temporal correlations in rewards, and making parameter estimation and optimal action sequence search significantly more challenging.

We observe that the problem can be resolved in the following ways. The issue of temporal correlation to estimate unknown parameters is addressed by improving upon recent results from system identification (Sattar et al., 2025b). While the state-space parameters may not be identifiable from the observed rewards (without additional assumptions on $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$), it is possible to obtain high-probability bounds on the estimation error for the *Markov parameters*, which characterize the action-reward relationship. In order to obtain an optimal long-term reward, we show that it is sufficient to use only this action-reward representation. In particu-

lar, selecting an optimal action sequence is equivalent to solving an indefinite quadratic problem over a hypercube, where the quadratic function is defined in terms of the Markov parameters.

Combining these two insights, we propose an *explore–then–commit* (ETC) algorithm. In the exploration phase, the learner applies random Rademacher actions to estimate the system's Markov parameters. In the commit phase, the algorithm plays an action sequence which is a solution to an indefinite quadratic optimization problem. We show that, with high probability, this algorithm achieves sublinear regret, scaling as $\tilde{\mathcal{O}}(pT^{2/3})$, where p is the action dimension.

Our setting bears relation to several other nonstationary bandits, including restless bandits (Whittle, 1988), rebounding bandits (Leqi et al., 2021), and bandits with underlying state (Khosravi et al., 2024), to name a few (see §7 for further discussion). Key novelties of our setting include that states can be of arbitrary dimension, the evolution can depend on interactions between dimensions (i.e., the matrix $\boldsymbol{A}$ need not be diagonal), and actions directly affect states in a potentially correlated manner (through the matrix $\boldsymbol{B}$). Furthermore, unlike many multi-armed bandit settings, we consider a continuous action space. Our problem also bears resemblance to model-based reinforcement learning for linear dynamics (Dean et al., 2018; Simchowitz, 2020; Mania et al., 2019; Lale et al., 2020b,a). However, in our setting, the reward is bilinear rather than quadratic, and there are no observations other than reward, which provides only partial information about the underlying state.

Contributions: In this paper, we make the following contributions:

- **Framework:** We propose a novel nonstationary bandit problem in which the action affects both the current reward (through a bilinear interaction) as well as the future rewards (through a latent state).
- **Algorithm:** We propose an explore-then-commit algorithm for our bandit problem. After exploration with random Rademacher actions, we estimate the Markov parameters which characterize the action-reward relationship. We then commit to an optimized action sequence obtained by solving a semidefinite relaxation of indefinite quadratic optimization over a hypercube — a known NP-hard problem.
- **Estimation:** We provide near-optimal sample complexity and estimation error bounds for learning Markov parameters. Specifically, in terms of the exploration length H, our error rate scales as $\tilde{\mathcal{O}}(1/\sqrt{H})$, and we require $H \gtrsim \tilde{\mathcal{O}}(d_M)$, where d_M is the dimension of unknown Markov parameters.

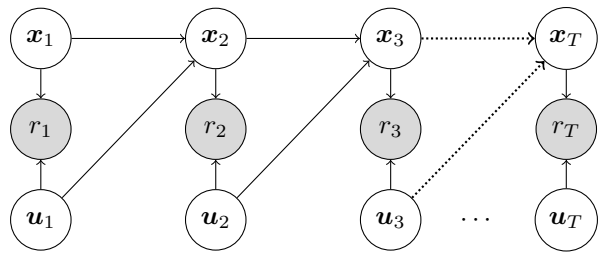


Figure 1: Graphical Model for Non-Stationary Bandits with controlled Latent Dynamics.

- **Regret:** We prove an upper bound on regret scaling as $\tilde{\mathcal{O}}(pT^{2/3})$ assuming perfect commit phase optimization, and additionally provide a suboptimality guarantee for the semidefinite relaxation with the Goemans-Williamson rounding approach.

Organization: §2 formalizes the model and the regret benchmark, defines the action set, and discusses the computational difficulty of the problem. §3 introduces the ETC algorithm and its regret guarantee. §4 presents our parameter estimation results with sample complexity and error bounds. §5 discusses our regret analysis. §6 presents numerical experiments. §7 shows related work to this paper, and §8 concludes the paper.

Notations: We use boldface lowercase/uppercase letters to denote vectors/matrices. The ℓ_2-norm and ℓ_∞-norm of a vector $\boldsymbol{x}$ are denoted by $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$, respectively. The spectral radius, the spectral norm, and the Frobenius norm of a matrix $\boldsymbol{X}$ are denoted by $\rho(\boldsymbol{X}), \|\boldsymbol{X}\|$, and $\|\boldsymbol{X}\|_F$, respectively. The largest and smallest eigenvalue of a square matrix $\boldsymbol{X}$ are denoted by $\lambda_{\max}(\boldsymbol{X})$ and $\lambda_{\min}(\boldsymbol{X})$. The operation $\otimes$ denotes the Kronecker product. We use $\gtrsim$ and $\lesssim$ for inequalities that hold up to an absolute constant factor. The notation $\tilde{\mathcal{O}}$ hides constants and logarithmic terms. Lastly, we use $\mathbf{0}_d$ and $\mathbf{0}_{m\times n}$ to denote the zero vector in $\mathbb{R}^d$ and the zero matrix in $\mathbb{R}^{m\times n}$, respectively.

2 PROBLEM FORMULATION

We consider a nonstationary stochastic bandit problem with controlled[1] latent dynamics and a bilinear reward model. At each round $t=0,1,\ldots,T$, the learner selects an action $\boldsymbol{u}_t$ from a bounded action set $\mathcal{U}\subseteq\mathbb{R}^p$, and receives a reward $r_t\in\mathbb{R}$. More specifically, the reward is bilinear in the latent state $\boldsymbol{x}_t\in\mathbb{R}^n$ and the action $\boldsymbol{u}_t\in\mathcal{U}$, whereas the latent state evolves according to a linear dynamical system, as defined in (1), where $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}$ are unknown matrices of appropriate dimensions, and $\boldsymbol{w}_t$, z_t are random zero-mean noise pro-

[1] i.e., the state is nontrivially influenced by the actions

cesses. Without loss of generality, we assume $\boldsymbol{x}_0 = \boldsymbol{0}_n$. Throughout, we assume $\boldsymbol{A}$ is Schur-stable, that is, $\rho(\boldsymbol{A}) < 1$, and the noise processes are i.i.d. zero-mean sub-Gaussian, with $\boldsymbol{w}_t$ having variance proxy $\boldsymbol{\Sigma}_w$ and z_t having variance proxy σ_z^2.

2.1 Objective, Action Set, and Regret

Our objective is to maximize the expected cumulative reward $\mathbb{E}\left[\sum_{t=0}^{T} r_t\right]$ over the horizon T, by choosing actions from a bounded set. For simplicity, we consider the action set to be a centered hypercube in $\mathbb{R}^p$. In other words, we choose

$$\mathcal{U} = \{\boldsymbol{u} \in \mathbb{R}^p \colon \|\boldsymbol{u}\|_\infty \leq 1\} = [-1, 1]^p. \tag{2}$$

Let $\{r_t^\star\}_{t=0}^T$ denote the rewards collected under the *optimal open-loop action sequence*[2], and $\{r_t^\pi\}_{t=0}^T$ denote the rewards collected by policy π. Then, the regret up to round T of a policy π is defined as:

$$R_T(\pi) := \mathbb{E}\left[\sum_{t=0}^{T} r_t^\star - \sum_{t=0}^{T} r_t^\pi\right]. \tag{3}$$

2.2 Optimal Open-Loop Actions

In this section, we suppose that $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are known, and show that the optimal action sequence is the solution to an indefinite quadratic optimization problem with ℓ_∞-norm constraint. By unrolling the latent state through (1), we have $r_0 = z_0$, and for $t \geq 1$,

$$r_t = \boldsymbol{u}_t^\top \boldsymbol{C} \sum_{i=0}^{t-1} \boldsymbol{A}^{t-i-1}(\boldsymbol{B}\boldsymbol{u}_i + \boldsymbol{w}_i) + z_t. \tag{4}$$

Since $\boldsymbol{w}_t$ and z_t have zero-mean, the expected cumulative reward is given by

$$\mathbb{E}\left[\sum_{t=0}^{T} r_t\right] = \boldsymbol{u}_{0:T}^\top \boldsymbol{M}_T \boldsymbol{u}_{0:T}, \tag{5}$$

where $\boldsymbol{u}_{0:T} := \begin{bmatrix}\boldsymbol{u}_T^\top & \boldsymbol{u}_{T-1}^\top & \cdots & \boldsymbol{u}_0^\top\end{bmatrix}^\top \in \mathbb{R}^{p(T+1)}$, and $\boldsymbol{M}_T \in \mathbb{R}^{p(T+1)\times p(T+1)}$ is a block Toeplitz matrix with (i,j)-th $p \times p$ block given by

$$(\boldsymbol{M}_T)_{ij} := \begin{cases} \boldsymbol{C}\boldsymbol{A}^{j-i-1}\boldsymbol{B} & (i < j) \\ \boldsymbol{0}_{p\times p} & (i \geq j). \end{cases} \tag{6}$$

With these definitions, the optimal action sequence and the corresponding cumulative reward are given as follows. This cumulative reward is our regret baseline.

[2]An open-loop policy specifies the action sequence in advance as a function of time only, independently of the realized reward observations.

Proposition 1. *Let $\boldsymbol{S}_T := \boldsymbol{M}_T + \boldsymbol{M}_T^\top$. Then, the optimal open-loop action sequence $\boldsymbol{u}_{0:T}^\star$ is the solution to the problem:*

$$\begin{aligned} \max_{\boldsymbol{u}_{0:T}} \quad & \frac{1}{2}\boldsymbol{u}_{0:T}^\top \boldsymbol{S}_T \boldsymbol{u}_{0:T} \\ \textit{subject to} \quad & \boldsymbol{u}_t \in [-1,1]^p, \quad \forall\, t = 0, 1, \ldots, T \end{aligned} \tag{7}$$

Note that the maximum value of (7) (i.e. the optimal expected cumulative reward) is at least $\frac{1}{2}\lambda_{\max}(\boldsymbol{S}_T)$ because $\|\boldsymbol{u}_{0:T}\|_2 \leq 1$ implies $\|\boldsymbol{u}_{0:T}\|_\infty \leq 1$.

This optimal action sequence represents the optimal *open-loop* strategy for accruing reward, in contrast to a *feedback* or *closed-loop* policy[3] (Bar-Shalom and Tse, 2003). It captures the problem of selecting optimal *sequences* of actions to maximize the long-term reward, a consideration which is not present in the classical bandit settings.

2.3 Combinatorial Actions

The problem (7) is the maximization of an indefinite quadratic function over the hypercube. We show that it suffices to consider a discrete set of actions corresponding to the vertices of the hypercube

$$\mathcal{U} = \{-1, +1\}^p \tag{8}$$

so that each coordinate of every action is restricted to ± 1.

Proposition 2. *Let $\boldsymbol{S}$ be an arbitrary $d \times d$ symmetric real matrix with nonnegative diagonal entries. Then, a maximizer of $\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}$ over $[-1,1]^d$ exists at a vertex $\boldsymbol{v} \in \{-1,+1\}^d$, that is,*

$$\max_{\|\boldsymbol{x}\|_\infty \leq 1} \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} = \max_{\boldsymbol{v}\in\{-1,+1\}^d} \boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v}. \tag{9}$$

We defer the proof of Proposition 2 to Appendix A.1. We therefore take the action set to be $\mathcal{U} = \{-1,+1\}^p$ instead of the hypercube $[-1,1]^p$ in the remainder of this paper. Under this choice of action set, the optimization (7) reduces to an instance of *quadratic unconstrained binary optimization* (QUBO), which is known to be NP-hard, as it generalizes classical problems such as MaxCut. Consequently, solving for the optimal actions exactly is computationally intractable. This motivates the use of semidefinite relaxations and randomized rounding schemes in §6.

3 MAIN RESULT: EXPLORE-THEN-COMMIT

We propose an explore–then–commit (ETC) algorithm for minimizing regret tailored to bandits with latent

[3]Both feedback and closed-loop policies select actions adaptively based on the available history.

linear dynamics and bilinear rewards. The ETC algorithm runs in two phases, *an exploration phase* and *a commit phase*. In §3.1, we describe the *exploration phase*, which consists in estimating the so-called *Markov parameters*[4] of the system. In §3.2, we describe the *commit phase*, where the learner uses the estimated Markov parameters to formulate an open-loop optimization problem and design a sequence of actions for the remaining horizon. Finally, in §3.3, we state our main theoretical result: with appropriate choices of exploration and truncation lengths, the ETC algorithm achieves a regret of order $\tilde{\mathcal{O}}(T^{2/3})$. We summarize the overall procedure in Algorithm 1.

c

Algorithm 1 Explore-then-Commit

Require: Horizon T, exploration length H, truncation length L

I. Exploration Phase

1: For $t = 0, \dots, H$, play $\boldsymbol{u}_t \sim \mathrm{Unif}(\{-1,+1\}^p)$ and observe reward r_t.

2: Find an estimate $\hat{\boldsymbol{G}}$ of the first L Markov parameters via least squares using $\{(\boldsymbol{u}_t, r_t)\}_{t=0}^{H}$ as in (15).

3: Construct $\hat{\boldsymbol{S}}_{T-H-1} = \frac{1}{2}(\hat{\boldsymbol{M}}_{T-H-1} + \hat{\boldsymbol{M}}_{T-H-1}^\top)$ where $\boldsymbol{M}_{T-H-1}$ is defined via $\hat{\boldsymbol{G}}$ as in (6).

II. Commit Phase

4: Find a sequence $(\boldsymbol{u}_{H+1}^\pi, \dots, \boldsymbol{u}_T^\pi)$ that solves,

$$\max_{\boldsymbol{u}_{H+1:T}} \tfrac{1}{2}\boldsymbol{u}_{H+1:T}^\top \hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{u}_{H+1:T}$$
$$\text{s.t.} \quad \boldsymbol{u}_t \in \{-1,+1\}^p, \quad \forall t = H+1, \dots, T.$$

by **(a)** SDP relaxation with Goemans-Williamson rounding, or **(b)** sign-iteration method.

5: For $t{=}H{+}1, \dots, T$, play $\boldsymbol{u}_t^\pi$ and observe reward r_t.

3.1 Exploration Phase

In the exploration phase, the learner selects actions independently from the Rademacher distribution $\mathrm{Unif}\{-1,+1\}^p$. This distribution satisfies the action constraint (see §2.3) and provides sufficient excitation of the system, ensuring that the Markov parameters can be consistently estimated (Sattar et al., 2025b). Further discussion of persistence of excitation is deferred to Appendix B.1.

The trajectory $\{(\boldsymbol{u}_t, r_t)\}_{t=0}^{H}$ collected in this phase is used to estimate the system's Markov parameters. Concretely, we form nonlinear regressors/features using the previous L actions $(\boldsymbol{u}_{t-1}, \dots, \boldsymbol{u}_{t-L})$ together with the current action $\boldsymbol{u}_t$, and use linear regression to predict the current reward r_t. The least-squares solution gives an estimate of the first L Markov parameters $\{\boldsymbol{C}\boldsymbol{A}^k\boldsymbol{B}\}_{k=0}^{L-1}$.

Since the matrix $\boldsymbol{A}$ is Schur-stable, the influence of terms beyond lag L decays geometrically. Hence, it suffices to estimate only the first L parameters. The details of the regression procedure and error analysis are deferred to §4.

3.2 Commit Phase

After exploration, the learner uses the estimated Markov parameters to construct an open-loop optimization problem aimed at maximizing the expected cumulative reward from $t = H{+}1$ to $t = T$. Because the Rademacher actions and the noise processes are zero-mean, the theoretical objective will be

$$\mathbb{E}\left[\sum_{t=H+1}^{T} r_t\right] = \frac{1}{2}\boldsymbol{u}_{H+1:T}^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}_{H+1:T} \tag{10}$$

with $\boldsymbol{S}_{T-H-1} = \boldsymbol{M}_{T-H-1} + \boldsymbol{M}_{T-H-1}^\top$ where $\boldsymbol{M}_{T-H-1}$ is a block Toeplitz matrix with similar structure to Equation (6) (see Appendix A.2).

Since the true parameters are unknown, we replace them with estimates from the exploration phase. Specifically, we form $\hat{\boldsymbol{S}}_{T-H-1}$ by substituting the estimated blocks for $\boldsymbol{C}\boldsymbol{A}^k\boldsymbol{B}$ whenever $k < L$ and setting the blocks to $\boldsymbol{0}_{p\times p}$ for $k \geq L$, consistent with the Schur-stability of $\boldsymbol{A}$. This yields the commit-phase optimization problem:

$$\begin{aligned} \max_{\boldsymbol{u}_{H+1:T}} \quad & \frac{1}{2}\boldsymbol{u}_{H+1:T}^\top \hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{u}_{H+1:T} \\ \text{subject to} \quad & \boldsymbol{u}_{H+1:T} \in \{-1,+1\}^{p(T-H)} \end{aligned} \tag{11}$$

The structure of this problem is identical to (7), but with estimated rather than true parameters. As noted in §2.3, the problem (11) is NP-hard. We therefore rely on practical approaches to obtain tractable approximations, which will be discussed in §6.1 and §6.2.

3.3 Regret Guarantee

Our main result establishes that the ETC algorithm achieves sublinear regret with high probability.

Theorem 3. *With exploration length $H = \tilde{\mathcal{O}}(T^{2/3})$ and truncation length $L = \Theta(\log T)$, the explore-then-commit algorithm π (Alg. 1) achieves regret*

$$R_T(\pi) = \tilde{\mathcal{O}}(pT^{2/3})$$

with high probability, where p is the action dimension.

[4] The Markov parameters corresponding to the system (1) are typically defined, for example in Sattar et al. (2025b), as the sequence of matrices $\{\boldsymbol{C}\boldsymbol{A}^k\boldsymbol{B}\}_{k\geq 0}$.

The proof of Theorem 3 is discussed in §5. Note that the regret inevitably has a high-probability bound because the Markov parameter estimation error, which affects the regret, has a high-probability bound.

This rate arises from a simple trade-off. The algorithm incurs linear loss during exploration ($\sim H$) and suffers estimation error of order $\tilde{\mathcal{O}}(\sqrt{1/H})$ across the remaining horizon ($\sim T$). Optimizing the bound over H yields the $\tilde{\mathcal{O}}(pT^{2/3})$ rate with $H = \tilde{\mathcal{O}}(T^{2/3})$. Lastly, taking $L = \Theta(\log T)$ makes the truncation error negligible, as already noted in §3.1 where Schur stability implies geometric decay of higher-order terms.

4 MAIN RESULT: PARAMETER ESTIMATION

In this section, we use the action-reward samples $\{(\boldsymbol{u}_t, r_t)\}_{t=0}^{H}$ from the exploration phase to estimate a map between the actions and rewards. Similar to Sattar et al. (2025b), we can estimate Markov parameters by regressing the rewards r_t to an expression defined by the history of actions $\{\boldsymbol{u}_\tau\}_{\tau\le t}$. For $t \ge L$, the reward r_t depends on the past L actions as follows:

$$r_t = \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L} + \boldsymbol{u}_t^\top \boldsymbol{C} \sum_{i=0}^{L-1} \boldsymbol{A}^i (\boldsymbol{B}\boldsymbol{u}_{t-i-1} + \boldsymbol{w}_{t-i-1}) + z_t. \quad (12)$$

To ease notation, for any sequence of vectors $\{\boldsymbol{q}_t\}_{t=0}^{T}$, let $\bar{\boldsymbol{q}}_t := \begin{bmatrix}\boldsymbol{q}_t^\top & \boldsymbol{q}_{t-1}^\top & \cdots & \boldsymbol{q}_{t-L+1}^\top\end{bmatrix}^\top$ denote a concatenation of past L vectors starting from $t \ge L$. If we let

$$\boldsymbol{G} := \begin{bmatrix}\boldsymbol{C}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}\boldsymbol{B} & \cdots & \boldsymbol{C}\boldsymbol{A}^{L-1}\boldsymbol{B}\end{bmatrix} \in \mathbb{R}^{p\times pL} \quad (13)$$

then (12) can be compactly written as

$$\begin{aligned} r_t &= \boldsymbol{u}_t^\top \boldsymbol{G}\bar{\boldsymbol{u}}_{t-1} + \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L} + \boldsymbol{u}_t^\top \boldsymbol{F}\bar{\boldsymbol{w}}_{t-1} + z_t \\ &= \mathbf{vec}(\boldsymbol{G})^\top \tilde{\boldsymbol{u}}_t + \zeta_t \end{aligned} \quad (14)$$

where we define the covariates $\tilde{\boldsymbol{u}}_t := \bar{\boldsymbol{u}}_{t-1} \otimes \boldsymbol{u}_t \in \mathbb{R}^{p^2 L}$, the effective noise $\zeta_t := \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L} + \boldsymbol{u}_t^\top \boldsymbol{F}\bar{\boldsymbol{w}}_{t-1} + z_t$, and $\boldsymbol{F} := [\boldsymbol{C} \;\; \boldsymbol{C}\boldsymbol{A} \;\; \cdots \;\; \boldsymbol{C}\boldsymbol{A}^{L-1}]$. Hence, we can formulate the following least-squares problem to estimate the unknown parameter $\boldsymbol{G}$ from $\{(\boldsymbol{u}_t, r_t)\}_{t=0}^{H}$:

$$\hat{\boldsymbol{G}} = \underset{\boldsymbol{G}\in\mathbb{R}^{p\times pL}}{\arg\min} \sum_{t=L+1}^{H} \left(r_t - \mathbf{vec}(\boldsymbol{G})^\top \tilde{\boldsymbol{u}}_t\right)^2. \quad (15)$$

Let $\tilde{\boldsymbol{U}} := \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\top$ for brevity. When $\tilde{\boldsymbol{U}}$ is full rank, the solution to the least-squares problem above is $\mathbf{vec}(\hat{\boldsymbol{G}}) = \tilde{\boldsymbol{U}}^{-1} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t r_t$, and the corresponding estimation error is given by

$$\mathbf{vec}(\hat{\boldsymbol{G}}) - \mathbf{vec}(\boldsymbol{G}) = \tilde{\boldsymbol{U}}^{-1} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \zeta_t. \quad (16)$$

In this paper, we provide improved sample complexity and error bounds (compared with Sattar et al. (2025b)), under the following assumption.

Assumption 1. **(a)** *$\boldsymbol{A}$ is Schur-stable;* **(b)** *$\boldsymbol{w}_t$ and z_t are i.i.d. centered sub-Gaussian with variance proxy $\boldsymbol{\Sigma}_w$ and σ_z, respectively;* **(c)** $\boldsymbol{u}_t \overset{i.i.d.}{\sim} \mathrm{Unif}\left(\{-1,+1\}^p\right)$.

Under Assumption 1, we derive a bound on sample complexity and an estimation error of (16) in Theorem 4. We note the following fact regarding the stability condition. According to Gelfand's formula, for all $\rho > \rho(\boldsymbol{A})$, the quantity $\phi(\boldsymbol{A}, \rho) := \sup_{k\in\mathbb{Z}_+}(\|\boldsymbol{A}^k\|/\rho^k)$ is finite. Hence, if $\rho(\boldsymbol{A}) < 1$, for all $\rho \in (\rho(\boldsymbol{A}), 1)$, we have $\|\boldsymbol{A}^k\| \le \phi(\boldsymbol{A}, \rho)\rho^k$ for all $k \in \mathbb{Z}_+$. The proof is shown in Appendix B.7.

Theorem 4. *Under Assumption 1, let $\{(\boldsymbol{u}_t, r_t)\}_{t=0}^{H}$ be a single trajectory of action-reward pairs collected from the system (1). For given $\delta \in (0,1)$, suppose*

$$H - L \gtrsim (L+1)\left(p^2 L \log(p^2 L) + \log\left(\frac{L+1}{\delta}\right)\right).$$

With probability at least $1-\delta$, we have $\|\hat{\boldsymbol{G}} - \boldsymbol{G}\|_F \lesssim$

$$\frac{\phi(\boldsymbol{A},\rho)\rho^L \|\boldsymbol{C}\|}{1-\rho}\sqrt{\frac{p^3 L \lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}_w)\log(1/\delta)}{H-L}} + (\sigma_z + \sqrt{p}\,\Lambda_w^{(L)})\sqrt{\frac{p^2 L \log(p^2 L(H-L)) + \log(L/\delta)}{H-L}}$$

for $\Lambda_w^{(L)} := \sum_{k=0}^{L-1}\sqrt{\lambda_{\max}(\boldsymbol{C}\boldsymbol{A}^k \boldsymbol{\Sigma}_w (\boldsymbol{A}^k)^\top \boldsymbol{C}^\top)}$.

The proof of Theorem 4 is presented in Appendix B.2. The recent work (Sattar et al., 2025b) shows that such a randomized design yields an estimation error bound of order $\tilde{\mathcal{O}}(1/\sqrt{\delta H})$ with probability at least $1-\delta$. In this work, we sharpen this guarantee by exploiting the sub-Gaussian structure of the noise processes, obtaining an improved bound of order $\tilde{\mathcal{O}}(\sqrt{\log(1/\delta)/H})$ with probability at least $1-\delta$.

The first term in our error bound corresponds to the approximation error because we use only L Markov parameters to reconstruct the reward function. It decays exponentially with L when the latent dynamics is strictly stable because the norm of Markov parameters decreases exponentially. The second term corresponds to the error due to noisy reward model and noisy latent dynamics update. This term depends linearly on the noise variance proxies.

The choice of actions $\boldsymbol{u}_t \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\left(\{-1,+1\}^p\right)$ leads to better persistence of excitation than Sattar et al. (2025b). In particular, the sample complexity in Theorem 4 scales as $\tilde{\mathcal{O}}(p^2 L^2)$ compared to $\tilde{\mathcal{O}}(p^4 L^3)$, and it is optimal in the dimension of Markov parameters

p^2. Secondly, the error bound in Theorem 4 depends on the failure probability δ through $\log(1/\delta)$. This is significantly better than the $1/\delta$ dependence in Sattar et al. (2025b) which considers a setting where noise processes can be heavy-tailed. The sub-Gaussian noise assumption in this paper allows using better concentration arguments, such as self-normalized bounds for martingales and Freedman's inequality, to obtain optimal dependence on the failure probability.

5 REGRET ANALYSIS

In this section we prove Theorem 3. Additional details and proofs of intermediate results are deferred to Appendix C. Recall the definition of regret from (3). From Proposition 1 and Equation (10), we obtain

$$R_T(\pi)=\frac{1}{2}(\boldsymbol{u}^\star_{0:T})^\top \boldsymbol{S}_T \boldsymbol{u}^\star_{0:T}-\frac{1}{2}(\boldsymbol{u}^\pi_{H+1:T})^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}^\pi_{H+1:T}$$

where $\boldsymbol{u}^\pi_{H+1:T}$ is the optimized action sequence which maximizes (11). For the purpose of analysis, let $\tilde{\boldsymbol{u}}_{H+1:T}$ be an action sequence that maximizes Equation (10) with true parameters under the constraint $\tilde{\boldsymbol{u}}_t \in \{-1,+1\}^p$ for $t = H+1,\dots,T$. Then, we decompose the regret as follows:

$$R_T(\pi) = \frac{1}{2}\left(R_{1,T} + R_{2,T} + R_{3,T}\right) \tag{17}$$

where

$$\begin{aligned}
R_{1,T} &:= (\boldsymbol{u}^\star_{0:T})^\top \boldsymbol{S}_T \boldsymbol{u}^\star_{0:T} - (\tilde{\boldsymbol{u}}_{H+1:T})^\top \boldsymbol{S}_{T-H-1}\tilde{\boldsymbol{u}}_{H+1:T}\\
R_{2,T} &:= (\tilde{\boldsymbol{u}}_{H+1:T})^\top \boldsymbol{S}_{T-H-1}\tilde{\boldsymbol{u}}_{H+1:T}\\
&\qquad - (\boldsymbol{u}^\pi_{H+1:T})^\top \hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{u}^\pi_{H+1:T}\\
R_{3,T} &:= (\boldsymbol{u}^\pi_{H+1:T})^\top (\hat{\boldsymbol{S}}_{T-H-1} - \boldsymbol{S}_{T-H-1})\boldsymbol{u}^\pi_{H+1:T}.
\end{aligned}$$

Intuitively, $R_{1,T}$ captures the sub-optimality between the full horizon and the commit-phase horizon, $R_{2,T}$ captures the sub-optimality between the true and estimated dynamics, and finally $R_{3,T}$ captures the error from parameter estimation. We analyze these three terms in detail in §5.1 and §5.2.

5.1 Upper Bounding $R_{1,T}$

The term $R_{1,T}$ quantifies the loss from truncating the horizon at H, capturing both short-term dependencies within the exploration phase and long-term dependencies between exploration and commit phases through the dynamics.

Proposition 5. *Let $\rho \in (\rho(\boldsymbol{A}), 1)$ be given. Then*

$$R_{1,T} \le 2p\kappa^2\left(\alpha H + \beta\right)$$

where $\alpha = 1 + \frac{\phi(\boldsymbol{A},\rho)\rho}{1-\rho}$, $\beta = \frac{\phi(\boldsymbol{A},\rho)\rho}{(1-\rho)^2} + 1$, and $\kappa = \max\{\|\boldsymbol{B}\|, \|\boldsymbol{C}\|\}$.

The proof of Proposition 5 is deferred to Appendix C.1. Note that the upper bound in 5 depends linearly on H, implying that the available reward grows linearly with the horizon. There is also a second term, a constant term independent of H. The coefficients α and β depend on the stability of the state dynamics $\boldsymbol{A}$, and they can be understood as effective memory capacity (Kumar et al., 2024). The first term arises from short-term dependencies within the exploration-phase. The second term is related to long-term dependencies which, due to stability, do not depend on the horizon.

5.2 Upper Bounding $R_{2,T}$ and $R_{3,T}$

The terms $R_{2,T}$ and $R_{3,T}$ both depend on the estimation error ϵ and the truncation error ρ^L, and their bounds are given in the following proposition.

Proposition 6. *Let $\rho \in (\rho(\boldsymbol{A}), 1)$ be given and $\epsilon > 0$ be the high-probability parameter estimation error, i.e., $\|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F \le \epsilon$. Then, with high probability,*

$$\max\{R_{2,T}, R_{3,T}\} \le 2p(T-H)\left(\epsilon + \kappa^2\gamma_L\right)$$

where $\kappa = \max\{\|\boldsymbol{B}\|, \|\boldsymbol{C}\|\}$ and $\gamma_L = \frac{\phi(\boldsymbol{A},\rho)\rho^L}{1-\rho}$.

The proof of Proposition 6 is deferred to Appendix C.2. We can observe that the bound depends on the length of commit-phase $T-H$, the estimation error ϵ, and the truncation error ρ^L.

5.3 Final Regret Analysis

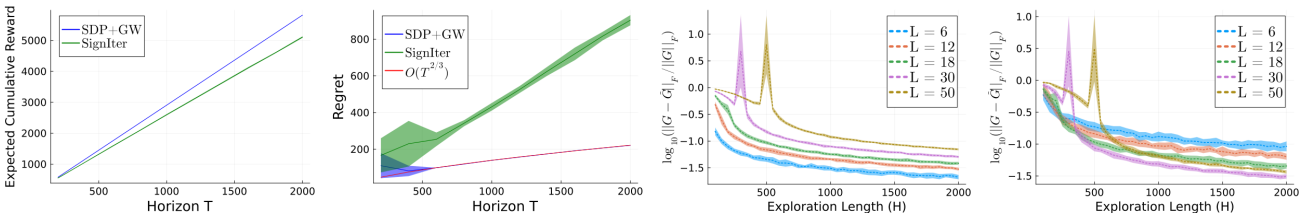
Finally, combining Propositions 5 and 6 yields the following high-probability bound:

$$R_T(\pi) \le 2p\kappa^2(\alpha H + \beta) + 4pT(\epsilon + \kappa^2\gamma_L). \tag{18}$$

From Theorem 4, we can take $\epsilon \lesssim \sqrt{\frac{\log(1/\delta)}{H-L}}$ with probability at least $1-\delta$. Optimizing over H gives $H = \tilde{\mathcal{O}}(T^{2/3})$, and with $L = \Theta(\log T)$ we obtain the high-probability regret bound $R_T(\pi) = \tilde{\mathcal{O}}(pT^{2/3})$. Note that we define regret to be the expected cumulative rewards, where expectation is taken both over actions (in the exploration phase) and noise processes. However, we still have a high-probability bound on the regret because the estimation error bound holds with high probability.

6 NUMERICAL EXPERIMENTS

In this section, we present numerical experiments with synthetic data. To obtain tractable solutions to (7) and (11) and analyze sub-optimality, we investigate two methods: (i) a semidefinite relaxation combined with Goemans-Williamson random hyperplane



(a) Oracle (b) Regret (c) $\rho(\boldsymbol{A}) = 0.1$ (d) $\rho(\boldsymbol{A}) = 0.9$

Figure 2: Each curve shows a mean over 20 experiments, with shaded regions indicating ± 1 standard deviation. (a) Expected cumulative reward under the oracle benchmark, approximated by semidefinite relaxation with Goemans-Williamson rounding (`SDP+GW`) and by the sign-iteration method (`SignIter`). (b) The regret of the explore-then-commit algorithm measured against the `SDP+GW` oracle benchmark, compared with the theoretical $\tilde{\mathcal{O}}(T^{2/3})$ rate. (c)–(d) Relative error of Markov parameter estimation for different truncation lengths L, under systems with spectral radii $\rho(\boldsymbol{A}) = 0.1$ and $\rho(\boldsymbol{A}) = 0.9$.

rounding, and (ii) a heuristic sign-iteration method. The code for experiments can be found in https://github.com/sdean-group/EtC-latent-bandits.

6.1 SDP Relaxation and Rounding

Our problem is to maximize a quadratic form $\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}$ for some symmetric $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ over $\boldsymbol{x} \in \{-1, 1\}^n$, which is equivalent to

$$\begin{aligned} \text{maximize} \quad & \operatorname{tr}(\boldsymbol{W}\boldsymbol{X}) \\ \text{subject to} \quad & \boldsymbol{X} \succeq 0, \ \operatorname{rank}(\boldsymbol{X}) = 1 \\ & \boldsymbol{X}_{ii} = 1, \quad i = 1, \dots, n. \end{aligned} \tag{19}$$

Dropping the rank constraint yields the semidefinite relaxation, which we solve using the Mosek solver (ApS, 2025).

To obtain a feasible binary matrix from a solution $\boldsymbol{X}$ to the relaxed problem, we apply Goemans-Williamson (GW) random hyperplane rounding algorithm (Goemans and Williamson, 1995): factor $\boldsymbol{X} = \boldsymbol{V}^\top \boldsymbol{V}$, sample $\boldsymbol{r} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, and set a vector

$$\boldsymbol{x} = \begin{bmatrix} \operatorname{sign}(\boldsymbol{r}^\top \boldsymbol{v}_1) & \cdots & \operatorname{sign}(\boldsymbol{r}^\top \boldsymbol{v}_n) \end{bmatrix}^\top$$

where $\boldsymbol{v}_i$ is the ith column vector of $\boldsymbol{V}$. Then the matrix $\boldsymbol{x}\boldsymbol{x}^\top$ is rank-one and feasible for (19). In practice, we repeat the rounding multiple times and keep the best value of $\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}$. We will hereafter refer to this method as `SDP+GW`.

For MaxCut problems, `SDP+GW` achieves an α-approximation algorithm with $\alpha \approx 0.87856$. While our objective is not exactly MaxCut, we derive a similar lower bound that depends on $\boldsymbol{W}$ (see Appendix D.1).

6.2 Sign-Iteration Method

As an alternative to relaxation and rounding, we also consider a heuristic method called *sign-iteration* (henceforth `SignIter`) to maximize $\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}$ over $\boldsymbol{x} \in \{-1, +1\}^n$. Starting from a random $\boldsymbol{x}^{(0)}$, the update is

$$x_i^{(k+1)} = \begin{cases} \operatorname{sign}\big((\boldsymbol{W}\boldsymbol{x}^{(k)})_i\big), & (\boldsymbol{W}\boldsymbol{x}^{(k)})_i \neq 0, \\ x_i^{(k)}, & \text{otherwise}, \end{cases}$$

and the procedure repeats until convergence or a maximum number of iterations. To mitigate dependence on initialization, the method is run multiple times and the best objective value is returned.

6.3 Experimental Evaluation of Regret

We consider a simple latent dynamics and reward function specified by

$$\boldsymbol{A} = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.15 & 0 \\ 0 & 0 & 0.12 \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0.5 & 0.4 \end{bmatrix}, \ \boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.3 \end{bmatrix}.$$

We set the noise processes to be Gaussian with variances $\boldsymbol{\Sigma}_w = (0.01)^2 \boldsymbol{I}_3$ and $\sigma_z = 0.01$. To select the exploration and truncation lengths, we perform a grid search to determine constants c_1, c_2 in $H = c_1 T^{2/3}$ and $L = c_2 \log T$, calibrated at $T = 1500$. These constants were then fixed and applied across all regret experiments.

We ran the ETC algorithm under this setup, repeating each experiment 20 times with different noise seeds. Regret was computed using the two benchmarks described in §6.1 and §6.2. As shown in Figure 2a, `SDP+GW` consistently attains higher benchmark values than `SignIter`, indicating that the latter rarely approaches the true optimum. We therefore adopt the `SDP+GW` benchmark as the reference for regret.

Figure 2b reports regrets measured against this benchmark. The `SDP+GW` regret grows slowly and closely follows the $\tilde{\mathcal{O}}(T^{2/3})$ rate, consistent with our theoretical

analysis. In contrast, the `SignIter` regret increases much more rapidly, nearly linearly in T, demonstrating that the heuristic commit phase is substantially suboptimal relative to the `SDP+GW` benchmark.

6.4 Parameter Estimation

In this section, we show that the parameters are estimated effectively. We generate random instances with $n = 5$, $p = 3$, and Schur-stable $\boldsymbol{A}$ (scaled from i.i.d. $\mathcal{N}(0, 1/n)$ entries), with $\boldsymbol{B}, \boldsymbol{C}$ similarly and noise levels $\boldsymbol{\Sigma}_w = (0.05)^2 \boldsymbol{I}_n$, $\sigma_z = 0.05$. We study the estimation error of Markov parameters for two spectral radii, $\rho(\boldsymbol{A}) = 0.1$ and $\rho(\boldsymbol{A}) = 0.9$, by repeating each experiment 20 times with different noise seeds.

Figures 2c and 2d show that the relative estimation error decreases with exploration length H. For small $\rho(\boldsymbol{A})$, shorter truncation lengths L yield smaller errors, while for large $\rho(\boldsymbol{A})$ the trend reverses, reflecting a trade-off between memory of states and number of Markov parameters. We also observe a characteristic double-descent effect (Nakkiran et al., 2020) as the regression problem transitions from under- to over-determined. For further discussion and derivations, see Sattar et al. (2025b) and Appendix D.2.

6.5 Comparison of Practical Methods

To evaluate the performance of the two approaches, we compare them against the true optimum of the regret-benchmark problem (7). Since this optimization is NP-hard in general, we restrict to small instances (n=3, p=2, T=5, . . . ,16) where brute force is feasible. Both `SDP+GW` and `SignIter` are randomized, so we vary the number of rounding trials (r=1, 10, 30) and repeat each experiment 20 times with different seeds.

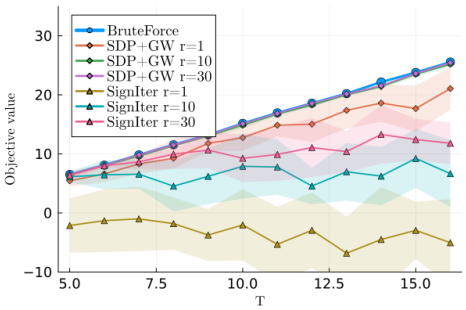


Figure 3: Comparison between `SDP+GW` and `SignIter` to the true optimum from the brute force method.

Figure 3 shows `SignIter` becomes increasingly suboptimal as T grows, even with more rounding trials. In contrast, `SDP+GW` consistently returns values very close to the brute-force optimum; even a single trial (r=1) mostly outperforms all `SignIter` cases. These results indicate that `SDP+GW` is an effective and reliable method for the commit-phase optimization.

7 RELATED WORK

Our formulation is connected to, but fundamentally different from, several classical bandit settings. Unlike classic multi-armed bandits (MAB), we consider continuous (or combinatorially many) actions and temporally correlated rewards. Many settings which generalize MAB to nonstationary rewards have been investigated, including restless bandits (Whittle, 1988), rotting bandits (Levine et al., 2017; Seznec et al., 2019), recharging bandits (Kleinberg and Immorlica, 2018), blocking bandits (Basu et al., 2019), decaying/rebounding bandits (Heidari et al., 2016; Leqi et al., 2021), delay-dependent rewards (Cella and Cesa-Bianchi, 2020), and bandits with underlying state (Khosravi et al., 2024). Most of these works consider finitely many discrete actions that affect rewards in a simple and uncorrelated manner, e.g. repeatedly playing the same action results in decayed or improved reward.

The linear bandit setting (Abbasi-Yadkori et al., 2011) naturally models correlations between actions, but classically considers a static parameter, equivalent in our setting to a fixed but unknown state. The nonstationary linear bandit setting (Russac et al., 2019) allows for drift, but requires bounded variation to provide meaningful regret guarantees. Recently, Trella et al. (2024) consider a nonstationary linear bandit with auto-regressive structure, but do not model any impact of past actions on the rewards. Closely related to our work in motivation, Clerici et al. (2024) study linear bandits with memory. Our setting differs in two regards: the state depends on all past actions, not just finitely many, and the actions affect the state linearly, leading to a quadratic dependence in the reward, rather than through sublinear powers of a covariance-like 'memory matrix'.

The structure of our reward function bears resemblance to both linear contextual bandits (Lattimore and Szepesvári, 2020) and bilinear bandits (Jun et al., 2019). Unlike these settings, our latent state is unobserved. Furthermore, unlike the contextual setting, the state is affected by past actions; unlike for bilinear bandits, the state is influenced through the dynamics and cannot be chosen directly.

Our setting is also related to partially observable Markov decision processes (POMDPs) (Aberdeen, 2003; Cayci et al., 2024; Efroni et al., 2022). Unlike much of the POMDP literature, which studies belief-based closed loop policies, often in multi-episode or

infinite-horizon settings, we compete with the optimal open-loop action sequence and study within a single finite-horizon trajectory. Moreover, by exploiting the linear latent dynamics and bilinear reward structure, we obtain polynomial sample complexity for system identification and a sublinear regret guarantee. Such structural exploitation is generally not available in algorithms for arbitrary POMDPs.

Finally, our setting relates to model-based reinforcement learning for linear dynamical systems (Dean et al., 2018; Simchowitz, 2020; Mania et al., 2019; Lale et al., 2020b,a). Compared to classical linear control, we point out two key differences. First, classical models assume action-independent observations, whereas our rewards depend bilinearly on actions. Second, quadratic costs in prior work yield linear optimal policies, whereas our selection is NP-hard.

8 CONCLUSION

We study a nonstationary bandit problem where rewards depend bilinearly on actions and latent states evolving under unknown linear dynamics. We propose an explore-then-commit algorithm that combines the parameter estimation with open-loop action optimization. Our analysis shows that the algorithm achieves $\tilde{\mathcal{O}}(pT^{2/3})$ regret with high probability. To address the NP-hard commit-phase optimization, we employed semidefinite relaxation and a rounding scheme, demonstrating its effectiveness empirically.

A natural direction for future work is to move beyond open-loop policies toward adaptive strategies. The recent work (Sattar et al., 2025a) considers a related structure in control theory, studying an optimal control problem of minimizing quadratic cost under linear dynamics with bilinear observations. It shows that the optimal feedback policy is nonlinear in the estimated state, suggesting that extending open-loop analysis to feedback or closed-loop designs is both challenging and an important direction for future research.

Acknowledgements

The authors are grateful to the anonymous reviewers for their time and effort in reviewing this manuscript and providing thoughtful feedback. S.D. was partly supported by NSF CCF 2312774, NSF OAC-2311521, NSF IIS-2442137, a PCCW Affinito-Stewart Award, a gift to the LinkedIn-Cornell Bowers CIS Strategic Partnership, and an AI2050 Early Career Fellowship program at Schmidt Sciences. M.F. was supported in part by awards NSF TRIPODS II 2023166, NSF CCF 2007036, NSF CCF 2212261, NSF CCF 2312775, and by an Amazon-UW Hub gift award. Y.J. was partially supported by the Knut and Alice Wallenberg Foundation Postdoctoral Scholarship Program at MIT-KAW 2022.0366.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Aberdeen, D. (2003). A (revised) survey of approximate methods for solving partially observable markov decision processes. Technical report, Technical report, National ICT Australia.

ApS, M. (2025). *MOSEK Optimizer API for Julia 11.0.29*.

Bar-Shalom, Y. and Tse, E. (2003). Dual effect, certainty equivalence, and separation in stochastic control. *IEEE Transactions on Automatic Control*, 19(5):494–500.

Basu, S., Sen, R., Sanghavi, S., and Shakkottai, S. (2019). Blocking bandits. *Advances in Neural Information Processing Systems*, 32.

Cayci, S., He, N., and Srikant, R. (2024). Finite-time analysis of natural actor-critic for pomdps. *SIAM Journal on Mathematics of Data Science*, 6(4):869–896.

Cella, L. and Cesa-Bianchi, N. (2020). Stochastic bandits with delay-dependent payoffs. In *International Conference on Artificial Intelligence and Statistics*, pages 1168–1177. PMLR.

Clerici, G., Laforgue, P., Cesa Bianchi, N., et al. (2024). Linear bandits with memory. *Transactions on Machine Learning Research*, (5):1–26.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197.

Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S. (2022). Provable reinforcement learning with a short-term memory. In *International Conference on Machine Learning*, pages 5832–5850. PMLR.

Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145.

Heidari, H., Kearns, M. J., and Roth, A. (2016). Tight policy regret bounds for improving and decaying bandits. In *IJCAI*, pages 1562–1570.

Jun, K.-S., Willett, R., Wright, S., and Nowak, R. (2019). Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pages 3163–3172. PMLR.

Khosravi, K., Leme, R. P., Podimata, C., and Tsorvantzis, A. (2024). Preferences evolve and so should your bandits: Bandits with evolving states for online platforms. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 101–101.

Kleinberg, R. and Immorlica, N. (2018). Recharging bandits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 309–319. IEEE.

Kumar, P. R. and Varaiya, P. (2015). *Stochastic systems: Estimation, identification, and adaptive control*. SIAM.

Kumar, R., Dean, S., and Kleinberg, R. (2024). Online convex optimization with unbounded memory. *Advances in Neural Information Processing Systems*, 36.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. (2020a). Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. (2020b). Regret minimization in partially observable linear quadratic control. *arXiv preprint arXiv:2002.00082*.

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

Leqi, L., Kilinc Karzan, F., Lipton, Z., and Montgomery, A. (2021). Rebounding bandits for modeling satiation effects. *Advances in Neural Information Processing Systems*, 34:4003–4014.

Levine, N., Crammer, K., and Mannor, S. (2017). Rotting bandits. *Advances in neural information processing systems*, 30.

Mania, H., Tu, S., and Recht, B. (2019). Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32.

Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. (2020). Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*.

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32.

Sattar, Y., Choi, S., Jedra, Y., Fazel, M., and Dean, S. (2025a). Sub-optimality of the separation principle for quadratic control from bilinear observations. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pages 3862–3867. IEEE.

Sattar, Y., Jedra, Y., and Dean, S. (2024). Learning linear dynamics from bilinear observations. *arXiv e-prints*, pages arXiv–2409.

Sattar, Y., Jedra, Y., and Dean, S. (2025b). Learning linear dynamics from bilinear observations. In *2025 American Control Conference (ACC)*, pages 3109–3115. IEEE.

Sattar, Y., Jedra, Y., Fazel, M., and Dean, S. (2025c). Finite sample identification of partially observed bilinear dynamical systems. *arXiv preprint arXiv:2501.07652*.

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., and Elahi, M. (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116.

Seznec, J., Locatelli, A., Carpentier, A., Lazaric, A., and Valko, M. (2019). Rotting bandits are no harder than stochastic ones. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2564–2572. PMLR.

Simchowitz, M. (2020). Making non-stochastic control (almost) as easy as stochastic. *Advances in Neural Information Processing Systems*, 33:18318–18329.

Trella, A. L., Dempsey, W., Gazi, A. H., Xu, Z., Doshi-Velez, F., and Murphy, S. A. (2024). Nonstationary latent auto-regressive bandits. *arXiv preprint arXiv:2402.03110*.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix: Explore-then-Commit for Nonstationary Linear Bandits with Latent Dynamics

A PROBLEM FORMULATION

In Section 2, we formulated the problem concretely, and we stated that it suffices to optimize the objective over the vertices of the hypercube in Section 2.3. We give a proof of Proposition 2 in Section A.1. Moreover, in Section 3.2, we introduced a theoretical objective to the commit phase. We derive the theoretical commit-phase objective in detail in Section A.2.

A.1 Proof of Proposition 2

Proposition 2. *Let $\boldsymbol{S}$ be an arbitrary $d \times d$ symmetric real matrix with nonnegative diagonal entries. Then, a maximizer of $\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}$ over $[-1,1]^d$ exists at a vertex $\boldsymbol{v} \in \{-1,+1\}^d$, that is,*

$$\max_{\|\boldsymbol{x}\|_\infty \le 1} \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} \;=\; \max_{\boldsymbol{v}\in\{-1,+1\}^d} \boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v}. \tag{20}$$

Proof. Let s_{ij} be the (i,j)-th entry of $\boldsymbol{S}$. Fix $\boldsymbol{x} = (x_1,\dots,x_d) \in \mathbb{R}^d$ with $\|\boldsymbol{x}\|_\infty \le 1$, and choose an arbitrary index k. Define $\phi_k(x_k) = \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}$ as a function of $x_k \in [-1,1]$, that is,

$$\phi_k(x_k) = \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} = s_{kk}x_k^2 + 2b_k x_k + C(x_{-k})$$

where we denote $b_k = \sum_{i\neq k} s_{ik}x_i$ and $C(x_{-k}) = 2\sum_{\substack{i<j\\ i,j\neq k}} s_{ij}x_i x_j + \sum_{i\neq k} s_{ii}x_i^2$. Since $s_{kk} \ge 0$, $\phi_k(x_k)$ is convex, so its maximum over $[-1,1]$ is attained at an endpoint. Thus, replacing x_k by whichever of ± 1 gives the larger value (monotonically) increases $\boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}$. Repeat this coordinatewise for all indices. After d steps, we arrive at a vertex $\boldsymbol{v} \in \{-1,+1\}^d$ with $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v} \ge \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x}$. Since $\boldsymbol{x}$ was arbitrarily feasible, we have

$$\max_{\|\boldsymbol{x}\|_\infty \le 1} \boldsymbol{x}^\top \boldsymbol{S}\boldsymbol{x} = \max_{\boldsymbol{v}\in\{-1,+1\}^d} \boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v}$$

which completes the proof. □

A.2 Derivation of Theoretical Commit-Phase Objective

In the commit phase of the explore-then-commit algorithm, the goal is to maximize the expected cumulative reward from $t = H+1$ to $t = T$. In Section 3.2, we derived that the theoretical objective of the commit phase is

$$\mathbb{E}\left[\sum_{t=H+1}^{T} r_t\right] = \frac{1}{2}\boldsymbol{u}_{H+1:T}^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}_{H+1:T}$$

subject to $\boldsymbol{u}_{H+1:T} \in \{-1,+1\}^{p(T-H)}$. In this section, we formally derive that the objective is valid.

After the exploration phase, the commit phase starts with the state $\boldsymbol{x}_{H+1}$, which is different from the initial state $\boldsymbol{x}_0 = \boldsymbol{0}_n$. Hence, the conditional expectation of cumulative reward will be

$$\mathbb{E}\left[\sum_{t=H+1}^{T} r_t \,\middle|\, \boldsymbol{x}_{H+1}\right] = \sum_{t=H+1}^{T} \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^{t-H-1}\boldsymbol{x}_{H+1}$$

$$+ \begin{bmatrix} \boldsymbol{u}_T \\ \boldsymbol{u}_{T-1} \\ \vdots \\ \vdots \\ \boldsymbol{u}_{H+2} \\ \boldsymbol{u}_{H+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{0}_{p\times p} & \boldsymbol{CB} & \boldsymbol{CAB} & \cdots & \cdots & \boldsymbol{CA}^{T-H-2}\boldsymbol{B} \\ \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & \boldsymbol{CB} & \cdots & \cdots & \boldsymbol{CA}^{T-H-3}\boldsymbol{B} \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \mathbf{0}_{p\times p} & \boldsymbol{CB} & \boldsymbol{CAB} \\ \vdots & & & \ddots & \mathbf{0}_{p\times p} & \boldsymbol{CB} \\ \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & \cdots & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_T \\ \boldsymbol{u}_{T-1} \\ \vdots \\ \vdots \\ \boldsymbol{u}_{H+2} \\ \boldsymbol{u}_{H+1} \end{bmatrix}$$

$$= \sum_{t=H+1}^{T} \boldsymbol{u}_t^\top \boldsymbol{CA}^{t-H-1}\boldsymbol{x}_{H+1} + \frac{1}{2}\boldsymbol{u}_{H+1:T}^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}_{H+1:T}$$

On the other hand, the state $\boldsymbol{x}_{H+1}$ is $\mathbf{0}_n$ in expectation because we assumed that $\boldsymbol{x}_0 = \mathbf{0}_n$, $\mathbb{E}[\boldsymbol{u}_t] = \mathbf{0}_p$, and $\mathbb{E}[\boldsymbol{w}_t] = \mathbf{0}_n$ for $t = 0, \ldots, H$. Therefore, applying the law of total expectation gives the desired objective (10).

B PARAMETER ESTIMATION

In this section, we provide details for parameter estimation. In Section B.1, we explain what persistence of excitation is and how it is related to parameter estimation. Across the following sections, we derive the estimation error of Markov parameters.

B.1 Persistence of Excitation

Definition 1 (Persistence of Excitation (Kumar and Varaiya, 2015))**.** *The action sequence $\{\boldsymbol{u}_t\}$ is said to be* persistently exciting *if there is a positive definite matrix $\boldsymbol{U}$ such that for all large n,*

$$\frac{1}{n}\sum_{k=0}^{n-1} \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top \succeq \boldsymbol{U}$$

where $\boldsymbol{\phi}_k$ is a regressor vector constructed from recent inputs $\boldsymbol{u}_t$.

Intuitively, persistent excitation in control theory means that the action signal provides sufficient richness of the system's dynamics so that its parameters can be uniquely identified from input-output data. More importantly, in quantitative aspect, the minimum eigenvalue of $\sum_{k=0}^{n-1} \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top$ grows linearly with n, meaning the regressors remain uniformly informative over time. This plays a crucial role in finite sample statistical guarantees, which we will talk about later.

Now we recall the least squares problem that we discussed in Section 4:

$$\hat{\boldsymbol{G}} = \underset{\boldsymbol{G}\in\mathbb{R}^{p\times pL}}{\arg\min} \sum_{t=L+1}^{H} \left(r_t - \mathbf{vec}(\boldsymbol{G})^\top \tilde{\boldsymbol{u}}_t\right)^2$$

where $\tilde{\boldsymbol{u}}_t = \bar{\boldsymbol{u}}_{t-1} \otimes \boldsymbol{u}_t$ and $\bar{\boldsymbol{u}}_{t-1} = \begin{bmatrix} \boldsymbol{u}_{t-1}^\top & \boldsymbol{u}_{t-2}^\top & \cdots & \boldsymbol{u}_{t-L}^\top \end{bmatrix}^\top$. The solution can be found by solving the normal equation which is

$$\mathbf{vec}(\hat{\boldsymbol{G}}) = \left(\sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\top\right)^\dagger \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t r_t.$$

If our action sequence is persistently exciting, then the matrix $\tilde{\boldsymbol{U}} = \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \tilde{\boldsymbol{u}}_t^\top$ is invertible (since positive definite), implying that the parameter $\boldsymbol{G}$ is uniquely determined. Now we show that random Rademacher action sequence is indeed persistently exciting using a result in the recent work (Sattar et al., 2025b). We state the result for completeness, and the notations are modified to be consistent to our paper.

Assumption 2 (Action Properties (Sattar et al., 2025b))**.** *$\boldsymbol{u}_0, \ldots, \boldsymbol{u}_H$ are i.i.d. stochastic isotropic vectors with zero mean and satisfy at least one of the following:*

(a) There exists a scalar $\beta > 0$ such that $\|\boldsymbol{u}_t\|_2 \le \beta$ for all $t = 0, \ldots, H$.

(b) There exists a scalar $m_4 > 0$ such that $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}[(\boldsymbol{v}^\top \tilde{\boldsymbol{u}}_t)^4] \le m_4$ for all $t = 0, \dots, H$ where $\tilde{\boldsymbol{u}}_t$ is as defined in Section 4.

Theorem 3 (Persistence of Excitation (Sattar et al., 2025b)). *Consider an i.i.d. sequence of random actions $\boldsymbol{u}_0, \dots, \boldsymbol{u}_H$ with number of samples satisfying*

$$H - L \gtrsim \gamma_1 (L+1) \left(\log \left(\frac{2(L+1)}{\delta} \right) + \gamma_2 p^2 L \right).$$

Suppose either of the following two conditions hold: **(1)** *$\{\boldsymbol{u}_t\}_{t=0}^H$ satisfies Assumption 2(a), $\gamma_1 = \beta^4 L$, and $\gamma_2 = 1$;* **(2)** *$\{\boldsymbol{u}_t\}_{t=0}^H$ satisfies Assumption 2(b), $\gamma_1 = m_4$, and $\gamma_2 = \log(1 + 16p^2 L/\delta)$. Then, with probability at least $1 - \delta$, we have*

$$\lambda_{\min}(\tilde{\boldsymbol{U}}) \ge (H - L)/4.$$

Theorem 3 shows that action distributions that satisfy at least one of the properties in Assumption 2 are persistently exciting with high probability. Moreover, the theorem implies that action distribution satisfying Assumption 2(b) gives a better sample complexity condition. Observe that the Rademacher distribution $\mathrm{Unif}\{-1, +1\}^p$ is centered and isotropic. It is clear that $\boldsymbol{u}_t \sim \mathrm{Unif}\{-1, +1\}p$ is bounded with $\beta = \sqrt{p}$. Moreover, in Section B.6, we show that the covariates $\tilde{\boldsymbol{u}}_t$ constructed from Rademacher random vectors have bounded fourth moments $m_4 = 9$. Hence, Theorem 3 implies that the number of samples required for persistence of excitation is

$$H - L \gtrsim (L+1) \left(\log \left(\frac{2(L+1)}{\delta} \right) + p^2 L \log \left(1 + \frac{16 p^2 L}{\delta} \right) \right)$$

given $\delta \in (0, 1)$.

On the other hand, we improve the result by exploiting that our covariates are bounded almost surely. The improvement comes from modifying the proof of Proposition 6.5 in Sattar et al. (2024) as follows: (1) since $\tilde{\boldsymbol{u}}_t$ constructed from Rademacher random vectors are bounded almost surely, we have $\|\tilde{\boldsymbol{U}}\| \le (H - L)p^2 L$ almost surely. Hence, the event $\mathcal{E}$ in the proof is modified as $\mathcal{E} := \{\|\tilde{\boldsymbol{U}}\| \le 2p^2 L(H - L)\}$; (2) because of the modified event $\mathcal{E}$, we also change the value of ϵ in the covering argument to $\epsilon = 1/(8p^2 L)$. The rest of the proof follows similarly.

Following the proof with the argument above, we can obtain an improved sample complexity for persistent excitation as

$$H - L \ge 32(L+1) m_4 \left(\log \left(\frac{2(L+1)}{\delta} \right) + p^2 L \log(1 + 16 p^2 L) \right) =: N(p, L, \delta) \tag{21}$$

where $m_4 = 9$, an upperbound of the fourth moment of $\tilde{\boldsymbol{u}}_t$. For future reference, define $N(p, L, \delta)$ to be the right-hand side of (21).

B.2 Proof of Theorem 4

Theorem 4. *Under Assumption 1, let $\{(\boldsymbol{u}_t, r_t)\}_{t=0}^H$ be a single trajectory of action-reward pairs collected from the system (1). For given $\delta \in (0, 1)$, suppose*

$$H - L \gtrsim (L+1) \left(p^2 L \log(p^2 L) + \log \left(\frac{L+1}{\delta} \right) \right).$$

With probability at least $1 - \delta$, we have

$$\|\hat{\boldsymbol{G}} - \boldsymbol{G}\|_F \lesssim \frac{\phi(\boldsymbol{A}, \rho) \rho^L \|\boldsymbol{C}\|}{1 - \rho} \sqrt{\frac{p \lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}_w) \log(1/\delta)}{H - L}} + (\sigma_z + \sqrt{p}\, \Lambda_w^{(L)}) \sqrt{\frac{p^2 L \log(p^2 L(H - L)) + \log(L/\delta)}{H - L}}$$

for $\Lambda_w^{(L)} := \sum_{k=0}^{L-1} \sqrt{\lambda_{\max}(\boldsymbol{C}\boldsymbol{A}^k \boldsymbol{\Sigma}_w (\boldsymbol{A}^k)^\top \boldsymbol{C}^\top)}$.

We use the same notations used in Section 4. From (16), we analyze the error by decomposing it as follows:

$$\|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F = \left\| \tilde{\boldsymbol{U}}^{-1} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \zeta_t \right\|_2 \leq \epsilon_1 + \epsilon_2$$

where

$$\epsilon_1 = \left\| \tilde{\boldsymbol{U}}^{-1} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L}) \right\|_2 \quad \text{and} \quad \epsilon_2 = \left\| \tilde{\boldsymbol{U}}^{-1} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{F}\bar{\boldsymbol{w}}_{t-1} + z_t) \right\|_2 .$$

The error ϵ_1 and ϵ_2 are about truncation bias and multiplier process, respectively. In Section B.3, we derive an upper bound of ϵ_1, and in Section B.4, we derive an upper bound of ϵ_2. In Section B.5, we give a final error bound for parameter estimation.

B.3 Analysis of Truncation Bias

In this section, we analyze the error related to the truncation bias:

$$\epsilon_1 = \left\| \tilde{\boldsymbol{U}}^{-1} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L}) \right\|_2 .$$

Throughout the analysis of ϵ_1, we assume the sample complexity (21) which leads the persistence of excitation with high probability. More specifically, given $\delta \in (0,1)$, we assume $H - L \geq N(p, L, \delta/4)$ so that

$$\lambda_{\min}(\tilde{\boldsymbol{U}}) \geq \frac{H-L}{4}$$

with probability $1 - \frac{\delta}{4}$. The error ϵ_1 has an upperbound

$$\epsilon_1 \leq \frac{1}{\lambda_{\min}(\tilde{\boldsymbol{U}})} \left\| \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L}) \right\|_2$$

by the consistency of norms. Given $t \geq 1$, $\boldsymbol{x}_t$ can be expressed in terms of previous actions and noises as

$$\boldsymbol{x}_t = \sum_{i=1}^{t} \boldsymbol{A}^{t-i} \left(\boldsymbol{B}\boldsymbol{u}_{i-1} + \boldsymbol{w}_{i-1} \right)$$

when $\boldsymbol{x}_0 = \boldsymbol{0}_n$. Observing that $\tilde{\boldsymbol{u}}_t \in \{-1, +1\}^{p^2 L}$, we have

$$\begin{aligned}
\left\| \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L}) \right\|_2 &= \sup_{\boldsymbol{\theta} \in S^{p^2 L - 1}} \boldsymbol{\theta}^\top \left(\sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L \boldsymbol{x}_{t-L}) \right) \\
&= \sum_{t=L+1}^{H} \sum_{i=1}^{t-L} \left(\sup_{\boldsymbol{\theta} \in S^{p^2 L - 1}} \boldsymbol{\theta}^\top \tilde{\boldsymbol{u}}_t \right) \left(\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^{h-i} (\boldsymbol{B}\boldsymbol{u}_{i-1} + \boldsymbol{w}_{i-1}) \right) \\
&= \sum_{i=1}^{H-L} \sum_{t=L+i}^{H} \left(\sup_{\boldsymbol{\theta} \in S^{p^2 L - 1}} \boldsymbol{\theta}^\top \tilde{\boldsymbol{u}}_t \right) \left(\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^{t-i} (\boldsymbol{B}\boldsymbol{u}_{i-1} + \boldsymbol{w}_{i-1}) \right) \\
&\leq \sqrt{p^2 L} \left| \sum_{i=1}^{H-L} \left(\sum_{t=L+i}^{H} \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^{t-i} \right) (\boldsymbol{B}\boldsymbol{u}_{i-1} + \boldsymbol{w}_{i-1}) \right|
\end{aligned}$$

Now we apply the following lemma which is a version of Freedman's inequality (adapted from Sattar et al. (2025c)).

Lemma 1. *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Let $(\boldsymbol{\eta}_t)_{t\geq 1}$ is a sequence of zero-mean, σ^2-sub-Gaussian random vectors taking values in $\mathbb{R}^d$, such that $\boldsymbol{\eta}_t$ is $\mathcal{F}_t$-measurable for all $t \geq 1$. Let $(\boldsymbol{x}_t)_{t\geq 1}$ be a sequence of random vectors taking values in $\mathbb{R}^d$ such that for all $t \geq 1$, $\boldsymbol{x}_t$ is $\mathcal{F}_{t-1}$-measurable and $\|\boldsymbol{x}_t\|_2 \leq K$ almost surely for some $K > 0$. Then for all $\delta \in (0,1)$, $T \geq 1$,*

$$\Pr\left(\left|\sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{\eta}_t\right| \leq \sigma K\sqrt{2T\log(2/\delta)}\right) \geq 1-\delta.$$

First, define a filtration where $\mathcal{F}_0 = \sigma(\boldsymbol{u}_{0:H})$ and $\mathcal{F}_i = \sigma(\boldsymbol{u}_{0:H}, \boldsymbol{w}_{0:i-1})$ for $i \geq 1$. With the notations in Lemma 1, take $\boldsymbol{x}_i = \left(\sum_{t=L+i}^{H} \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^{t-i}\right)^\top$ and $\boldsymbol{\eta}_i = \boldsymbol{B}\boldsymbol{u}_{i-1} + \boldsymbol{w}_{i-1}$.

Claim 1. *$\boldsymbol{\eta}_i$ is σ^2-sub-Gaussian where $\sigma^2 = \lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}_w)$.*

Proof. Let $\boldsymbol{v} \in \mathbb{R}^n$ be a unit vector, then $\mathbb{E}[\boldsymbol{v}^\top \boldsymbol{\eta}_i] = 0$. Given $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda(\boldsymbol{v}^\top \boldsymbol{\eta}_i)\right)\right] = \mathbb{E}\left[\exp\left(\lambda \boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{u}_{i-1}\right)\right]\mathbb{E}\left[\exp\left(\lambda \boldsymbol{v}^\top \boldsymbol{w}_{i-1}\right)\right]$$

because $\boldsymbol{u}_{i-1}$ and $\boldsymbol{w}_{i-1}$ are independent. First, since $\boldsymbol{w}_{i-1}$ is sub-Gaussian with variance proxy $\boldsymbol{\Sigma}_w$, we have $\mathbb{E}\left[\exp\left(\lambda \boldsymbol{v}^\top \boldsymbol{w}_{i-1}\right)\right] \leq \exp\left(\frac{1}{2}\lambda^2 \boldsymbol{v}^\top \boldsymbol{\Sigma}_w \boldsymbol{v}\right)$. Now we compute $\mathbb{E}\left[\exp\left(\lambda \boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{u}_{i-1}\right)\right]$. Let $\boldsymbol{b} = \boldsymbol{B}^\top \boldsymbol{v}$, and let b_k and $(\boldsymbol{u}_{t-1})_k$ be kth coordinate entry of $\boldsymbol{b}$ and $\boldsymbol{u}_{t-1}$, respectively. Then

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{u}_{i-1}\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{k=1}^{p} b_k (\boldsymbol{u}_{i-1})_k\right)\right] \\
&= \prod_{k=1}^{p} \mathbb{E}\left[\exp\left(\lambda b_k (\boldsymbol{u}_{i-1})_k\right)\right] \qquad (\because \text{ each coordinate of } \boldsymbol{u}_{i-1} \text{ is independent}) \\
&= \prod_{k=1}^{p} \frac{e^{\lambda b_k} + e^{-\lambda b_k}}{2} \\
&\leq \prod_{k=1}^{p} \exp\left(\frac{\lambda^2 b_k^2}{2}\right) \\
&= \exp\left(\frac{\lambda^2}{2}\sum_{k=1}^{p} b_k^2\right) \\
&= \exp\left(\frac{1}{2}\lambda^2 \boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{v}\right)
\end{aligned}$$

Hence, since $\boldsymbol{v} \in \mathbb{R}^n$ is a unit vector, we have

$$\mathbb{E}\left[\exp\left(\lambda(\boldsymbol{v}^\top \boldsymbol{\eta}_i)\right)\right] \leq \exp\left(\frac{\lambda^2}{2}\boldsymbol{v}^\top\left(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}_w\right)\boldsymbol{v}\right) \leq \exp\left(\frac{\lambda^2}{2}\lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}_w)\right)$$

which completes the proof. □

Claim 2. *Given $\rho \in (\rho(\boldsymbol{A}), 1)$, $\|\boldsymbol{x}_i\|_2 \leq K$ with $K = \sqrt{p}\|\boldsymbol{C}\|\frac{\phi(\boldsymbol{A},\rho)\rho^L}{1-\rho}$.*

Proof. We use Corollary 1 in order to bound $\|\boldsymbol{x}_i\|_2$ as follows:

$$\|\boldsymbol{x}_i\|_2 \leq \sum_{t=L+i}^{H} \|\boldsymbol{A}^{t-i}\|\|\boldsymbol{C}\|\,\|\boldsymbol{u}_h\|_2 \leq \sqrt{p}\|\boldsymbol{C}\| \sum_{t=L+i}^{H} \phi(\boldsymbol{A},\rho)\rho^{t-i} \leq \sqrt{p}\|\boldsymbol{C}\|\frac{\phi(\boldsymbol{A},\rho)\rho^L}{1-\rho}$$

□

According to Lemma 1, given $\delta \in (0,1)$, with probability at least $1 - \frac{\delta}{4}$,

$$\left|\sum_{i=1}^{H-L}\left(\sum_{t=L+i}^{H} \boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^{t-i}\right)(\boldsymbol{B}\boldsymbol{u}_{i-1} + \boldsymbol{w}_{i-1})\right| = \left|\sum_{i=1}^{H-L} \boldsymbol{x}_i^\top \boldsymbol{\eta}_i\right| \leq \frac{\phi(\boldsymbol{A},\rho)\rho^L\|\boldsymbol{C}\|}{1-\rho}\sqrt{2p(H-L)\lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{\Sigma}_w)\log(8/\delta)}.$$

Therefore, combining with persistence of excitation, the truncation bias term will be bounded by

$$\begin{aligned}\left\|\tilde{\boldsymbol{U}}^{-1}\sum_{t=L+1}^{H}\tilde{\boldsymbol{u}}_t(\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L\boldsymbol{x}_{t-L})\right\|_2 &\leq \frac{1}{\lambda_{\min}(\tilde{\boldsymbol{U}})}\left\|\sum_{t=L+1}^{H}\tilde{\boldsymbol{u}}_t(\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^L\boldsymbol{x}_{t-L})\right\|_2 \\ &\leq \frac{\sqrt{p^2L}}{\lambda_{\min}(\tilde{\boldsymbol{U}})}\left|\sum_{i=1}^{H-L}\left(\sum_{t=L+i}^{H}\boldsymbol{u}_t^\top \boldsymbol{C}\boldsymbol{A}^{t-i}\right)(\boldsymbol{B}\boldsymbol{u}_{i-1}+\boldsymbol{w}_{i-1})\right| \\ &\leq \frac{4p^{3/2}\phi(\boldsymbol{A},\rho)\rho^L\|\boldsymbol{C}\|}{1-\rho}\sqrt{\frac{2L\lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top+\boldsymbol{\Sigma}_w)\log(8/\delta)}{H-L}}\end{aligned}$$

with probability at least $1 - \frac{\delta}{2}$.

B.4 Analysis of Multiplier Process

In this section, we analyze the error ϵ_2, which is related to multiplier process. We again first assume that $H - L \geq N(p, L, \delta/4)$ for given $\delta \in (0,1)$ to get the result of persistent excitation, that is,

$$\lambda_{\min}(\tilde{\boldsymbol{U}}) \geq \frac{H-L}{4}$$

with probability $1 - \frac{\delta}{4}$.

The term related to multiplier process has an upper bound:

$$\epsilon_2 = \left\|\tilde{\boldsymbol{U}}^{-1}\sum_{t=L+1}^{H}\tilde{\boldsymbol{u}}_t(\boldsymbol{u}_t^\top \boldsymbol{F}\bar{\boldsymbol{w}}_{t-1} + z_t)\right\|_2 \leq \frac{\left\|\tilde{\boldsymbol{U}}^{-1/2}\sum_{t=L+1}^{H}\tilde{\boldsymbol{u}}_t(\boldsymbol{u}_t^\top \boldsymbol{F}\bar{\boldsymbol{w}}_{t-1}+z_t)\right\|_2}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \tag{22}$$

$$\leq \underbrace{\frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}}\left\|\tilde{\boldsymbol{U}}^{-1/2}\sum_{t=L+1}^{H} z_t\tilde{\boldsymbol{u}}_t\right\|_2}_{\epsilon_{2,1}} \tag{23}$$

$$+ \underbrace{\frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}}\left\|\tilde{\boldsymbol{U}}^{-1/2}\sum_{t=L+1}^{H}\tilde{\boldsymbol{u}}_t(\boldsymbol{u}_t^\top \boldsymbol{F}\bar{\boldsymbol{w}}_{t-1})\right\|_2}_{\epsilon_{2,2}} \tag{24}$$

To analyze the errors $\epsilon_{2,1}$ and $\epsilon_{2,2}$, we use Theorem 6 in Section B.7.

B.4.1 Analysis of $\epsilon_{2,1}$

Recall that $(z_t)_{t\geq 0}$ is i.i.d. zero-mean σ^2-sub-Gaussian. Define $\mathcal{F}_t = \sigma(\boldsymbol{u}_{0:H}, z_{0:t})$. Then $\tilde{\boldsymbol{u}}_t$ is $\mathcal{F}_{t-1}$-measurable and z_t is $\mathcal{F}_t$-measurable. Observing that z_t is independent of $\mathcal{F}_{t-1}$, z_t is conditionally σ_z^2-sub-Gaussian, i.e.,

$$\mathbb{E}\left[z_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}[z_t] = 0 \quad \text{and} \quad \mathbb{E}\left[e^{\lambda z_t} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[e^{\lambda z_t}\right] \leq \exp\left(\frac{\lambda^2\sigma_z^2}{2}\right)$$

Let $d = p^2L$. We apply the theorem with $\boldsymbol{V} = \boldsymbol{I}_d$. Define

$$\boldsymbol{V}_H = \boldsymbol{I}_d + \sum_{t=L+1}^{H}\tilde{\boldsymbol{u}}_t\tilde{\boldsymbol{u}}_t^\top = \boldsymbol{I}_d + \tilde{\boldsymbol{U}}, \quad \text{and} \quad \boldsymbol{s}_H = \sum_{t=L+1}^{H} z_t\tilde{\boldsymbol{u}}_t.$$

Then, for any $\delta > 0$, with probability at least $1 - \frac{\delta}{4}$,

$$\left\| \boldsymbol{V}_H^{-1/2} \boldsymbol{s}_H \right\|_2 = \left\| \left(\boldsymbol{I}_d + \tilde{\boldsymbol{U}} \right)^{-1/2} \sum_{t=L+1}^{H} z_t \tilde{\boldsymbol{u}}_t \right\|_2 \leq \sigma_z \sqrt{2 \log \left(\frac{4 \det(\boldsymbol{V}_H)^{1/2}}{\delta} \right)}. \tag{25}$$

According to Corollary 2 with $\lambda = 1$, we have

$$\left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} z_t \tilde{\boldsymbol{u}}_t \right\|_2 \leq \sqrt{1 + \frac{1}{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \left(\boldsymbol{I}_d + \tilde{\boldsymbol{U}} \right)^{-1/2} \sum_{t=L+1}^{H} z_t \tilde{\boldsymbol{u}}_t \right\|_2 .$$

Combining the inequality with (25), we get

$$\left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} z_t \tilde{\boldsymbol{u}}_t \right\|_2 \leq \sigma_z \sqrt{2 \left(1 + \frac{1}{\lambda_{\min}(\tilde{\boldsymbol{U}})} \right) \log \left(\frac{4 \det(\boldsymbol{V}_H)^{1/2}}{\delta} \right)}. \tag{26}$$

Claim 3. $\lambda_{\max}(\tilde{\boldsymbol{U}}) \leq d(H - L)$

Proof. If we show that $\tilde{\boldsymbol{U}} \preceq d(H-L)\boldsymbol{I}_d$, then the claim is proved. We can show that $d(H-L)\boldsymbol{I}_d - \tilde{\boldsymbol{U}}$ is positive semidefinite by Cauchy-Schwarz inequality:

$$\boldsymbol{x}^\top \tilde{\boldsymbol{U}} \boldsymbol{x} = \sum_{t=L+1}^{H} (\tilde{\boldsymbol{u}}_t^\top \boldsymbol{x})^2 \leq \sum_{t=L+1}^{H} \|\tilde{\boldsymbol{u}}_t\|_2^2 \, \|\boldsymbol{x}\|_2^2 = d(H-L) \, \|\boldsymbol{x}\|_2^2$$

Since the inequality above holds for all $\boldsymbol{x} \in \mathbb{R}^d$, the proof is done. □

By Claim 3,

$$\det(\boldsymbol{V}_H) = \det(\boldsymbol{I}_d + \tilde{\boldsymbol{U}}) \leq \left(1 + \lambda_{\max}(\tilde{\boldsymbol{U}}) \right)^d \leq \left(1 + d(H-L) \right)^d.$$

From (26), we now have

$$\left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} z_t \tilde{\boldsymbol{u}}_t \right\|_2 \leq \sigma_z \sqrt{2 \left(1 + \frac{1}{\lambda_{\min}(\tilde{\boldsymbol{U}})} \right) \log \frac{4 \left(1 + d(H-L)\right)^{d/2}}{\delta}}. \tag{27}$$

so that

$$\frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} z_t \tilde{\boldsymbol{u}}_t \right\|_2 \leq \frac{\sigma_z \sqrt{1 + \lambda_{\min}(\tilde{\boldsymbol{U}})}}{\lambda_{\min}(\tilde{\boldsymbol{U}})} \sqrt{2 \log \frac{4 \left(1 + d(H-L)\right)^{d/2}}{\delta}}.$$

Now we combine with persistence of excitation. Under the sample complexity requirement, since we have $\lambda_{\min}(\tilde{\boldsymbol{U}}) \geq (H-L)/4$ with probability at least $1 - \frac{\delta}{4}$, we have

$$1 + \lambda_{\min}(\tilde{\boldsymbol{U}}) \leq \left(\frac{4}{H-L} + 1 \right) \lambda_{\min}(\tilde{\boldsymbol{U}}).$$

If we choose H and L with $H - L \geq 4$, then $2(H-L) \geq H - L + 4$, implying

$$\frac{\sqrt{1 + \lambda_{\min}(\tilde{\boldsymbol{U}})}}{\lambda_{\min}(\tilde{\boldsymbol{U}})} \leq \sqrt{\frac{H - L + 4}{\lambda_{\min}(\tilde{\boldsymbol{U}})(H-L)}} \leq \sqrt{\frac{2}{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \leq \sqrt{\frac{8}{H-L}}.$$

Moreover, under $H - L \geq 4$, we have $d(H-L) \geq 2$ so that

$$\log(1 + d(H-L))^{d/2} = d \log \sqrt{1 + d(H-L)} \leq d \log(d(H-L))$$

for $d(H-L) \geq 2$.

By combining all, with probability at least $1 - \frac{\delta}{2}$,

$$\frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} z_t \tilde{\boldsymbol{u}}_t \right\|_2 \leq 4\sigma_z \sqrt{\frac{d\log(d(H-L)) + \log(4/\delta)}{H-L}}.$$

provided that $H - L \geq N(p, L, \delta/4)$.

B.4.2 Analysis of $\epsilon_{2,2}$

We analyze the error $\epsilon_{2,2}$ by decomposing in the following way:

$$\begin{aligned}
\frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{F} \bar{\boldsymbol{w}}_{t-1}) \right\|_2 &= \frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \left(\sum_{k=0}^{L-1} \boldsymbol{u}_t^\top \boldsymbol{C} \boldsymbol{A}^k \boldsymbol{w}_{t-k-1} \right) \right\| \\
&\leq \sum_{k=0}^{L-1} \frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \left(\boldsymbol{u}_t^\top \boldsymbol{C} \boldsymbol{A}^k \boldsymbol{w}_{t-k-1} \right) \right\|
\end{aligned}$$

For each $k = 0, 1, \ldots, L-1$, define a filtration $\mathcal{F}_t^{(k)} = \sigma(\boldsymbol{u}_{0:H}, \boldsymbol{w}_{0:t-k-1})$ for $t \geq L+1$. Let $\eta_t^{(k)} = \boldsymbol{u}_t^\top \boldsymbol{C} \boldsymbol{A}^k \boldsymbol{w}_{t-k-1}$. Then $\eta_t^{(k)}$ is $\mathcal{F}_t^{(k)}$-measurable. Observe that $\eta_t^{(k)} \mid \mathcal{F}_{t-1}^{(k)}$ is mean zero sub-Gaussian because

$$\mathbb{E}\left[\eta_t^{(k)} \mid \mathcal{F}_{t-1}^{(k)} \right] = \boldsymbol{u}_t^\top \boldsymbol{C} \boldsymbol{A}^k \mathbb{E}[\boldsymbol{w}_{t-k-1}] = 0$$

and, for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \eta_t^{(k)} \right) \middle| \mathcal{F}_{t-1}^{(k)} \right] &= \mathbb{E}\left[\exp\left(\lambda \boldsymbol{u}_t^\top \boldsymbol{C} \boldsymbol{A}^k \boldsymbol{w}_{t-k-1} \right) \middle| \mathcal{F}_{t-1}^{(k)} \right] \\
&\leq \exp\left(\frac{\lambda^2}{2} \boldsymbol{u}_t^\top \boldsymbol{C} \boldsymbol{A}^k \boldsymbol{\Sigma}_w (\boldsymbol{A}^k)^\top \boldsymbol{C}^\top \boldsymbol{u}_h \right) \\
&\leq \exp\left(\frac{\lambda^2}{2} \lambda_{\max}(\boldsymbol{C} \boldsymbol{A}^k \boldsymbol{\Sigma}_w (\boldsymbol{A}^k)^\top \boldsymbol{C}^\top) \, \|\boldsymbol{u}_t\|_2^2 \right) \\
&= p \lambda_{\max}\left(\boldsymbol{C} \boldsymbol{A}^k \boldsymbol{\Sigma}_w (\boldsymbol{A}^k)^\top \boldsymbol{C}^\top \right)
\end{aligned}$$

Let $\boldsymbol{S}_k = \boldsymbol{C} \boldsymbol{A}^k \boldsymbol{\Sigma}_w (\boldsymbol{A}^k)^\top \boldsymbol{C}^\top$ for brevity. We repeat a similar argument in Section B.4.1. By the persistence of excitation result, the event $\lambda_{\min}(\tilde{\boldsymbol{U}}) \geq (H-L)/4$ holds with probability $1 - \delta/4$, provided that $H - L \geq N(p, L, \delta/4)$. Conditioning on this event, for each fixed $k \in \{0, 1, \ldots, L-1\}$, with probability at least $1 - \left(\frac{\delta}{4L} \times L + \frac{\delta}{4} \right)$, we have

$$\frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} \eta_t^{(k)} \tilde{\boldsymbol{u}}_t \right\| \leq 4\sqrt{p \lambda_{\max}(\boldsymbol{S}_k)} \sqrt{\frac{d\log(d(H-L)) + \log(4L/\delta)}{H-L}}.$$

Applying a union bound over $k = 0, 1, \ldots, L-1$, we obtain, with probability at least $1 - \delta/2$,

$$\begin{aligned}
\frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t (\boldsymbol{u}_t^\top \boldsymbol{F} \bar{\boldsymbol{w}}_{t-1}) \right\|_2 &\leq \sum_{k=0}^{L-1} \frac{1}{\sqrt{\lambda_{\min}(\tilde{\boldsymbol{U}})}} \left\| \tilde{\boldsymbol{U}}^{-1/2} \sum_{t=L+1}^{H} \tilde{\boldsymbol{u}}_t \left(\boldsymbol{u}_t^\top \boldsymbol{C} \boldsymbol{A}^k \boldsymbol{w}_{t-k-1} \right) \right\| \\
&\leq 4\sqrt{\frac{d\log(d(H-L)) + \log(4L/\delta)}{H-L}} \sum_{k=0}^{L-1} \sqrt{p \, \lambda_{\max}(\boldsymbol{S}_k)}
\end{aligned}$$

provided that $H - L \geq N(p, L, \delta/4)$.

B.5 Final Parameter Estimation Error Bound

We now combine the errors ϵ_1 and ϵ_2 (and $\epsilon_{2,1}$, $\epsilon_{2,2}$). Under the assumption of persistence of excitation, we have

$$\epsilon_1 \leq \frac{4p^{3/2}\phi(\boldsymbol{A},\rho)\rho^L\|\boldsymbol{C}\|}{1-\rho}\sqrt{\frac{2L\lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top+\boldsymbol{\Sigma}_w)\log(8/\delta)}{H-L}}$$

with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned}
\epsilon_2 &\leq \epsilon_{2,1}+\epsilon_{2,2} \\
&\leq 4\sigma_z\sqrt{\frac{p^2L\log(p^2L(H-L))+\log(4/\delta)}{H-L}} \\
&\qquad + 4\sqrt{\frac{p^2L\log(p^2L(H-L))+\log(4L/\delta)}{H-L}}\sum_{k=0}^{L-1}\sqrt{p\,\lambda_{\max}(\boldsymbol{C}\boldsymbol{A}^k\boldsymbol{\Sigma}_w(\boldsymbol{A}^k)^\top\boldsymbol{C}^\top)} \\
&\leq 4\left(\sigma_z+\sqrt{p}\sum_{k=0}^{L-1}\sqrt{\lambda_{\max}(\boldsymbol{C}\boldsymbol{A}^k\boldsymbol{\Sigma}_w(\boldsymbol{A}^k)^\top\boldsymbol{C}^\top)}\right)\sqrt{\frac{p^2L\log(p^2L(H-L))+\log(4L/\delta)}{H-L}}
\end{aligned}$$

with probability at least $1-\frac{3\delta}{4}$, where $d=p^2L$ and $\boldsymbol{S}_k=\boldsymbol{C}\boldsymbol{A}^k\boldsymbol{\Sigma}_w(\boldsymbol{A}^k)^\top\boldsymbol{C}^\top$. As a result, we obtain a final bound for the multiplier process. With probability at least $1-\delta$,

$$\begin{aligned}
\|\boldsymbol{G}-\hat{\boldsymbol{G}}\|_F &\leq \epsilon_1+\epsilon_2 \\
&\leq \frac{4p^{3/2}\phi(\boldsymbol{A},\rho)\rho^L\|\boldsymbol{C}\|}{1-\rho}\sqrt{\frac{2L\lambda_{\max}(\boldsymbol{B}\boldsymbol{B}^\top+\boldsymbol{\Sigma}_w)\log(8/\delta)}{H-L}} \\
&\qquad + 4\left(\sigma_z+\sqrt{p}\sum_{k=0}^{L-1}\sqrt{\lambda_{\max}(\boldsymbol{C}\boldsymbol{A}^k\boldsymbol{\Sigma}_w(\boldsymbol{A}^k)^\top\boldsymbol{C}^\top)}\right)\sqrt{\frac{p^2L\log(p^2L(H-L))+\log(4L/\delta)}{H-L}}
\end{aligned}$$

provided that $H-L\geq N(p,L,\delta/4)$.

B.6 Proof of Bounded Fourth Moment of Covariates

In this section, we prove that our covariate $\tilde{\boldsymbol{u}}_t$ has bounded fourth moment. Observe that the fourth moment of $\tilde{\boldsymbol{u}}_t\in\mathbb{R}^{p^2L}$ is defined to be

$$\sup_{\substack{\|\boldsymbol{v}\|_2=1\\ \boldsymbol{v}\in\mathbb{R}^{p^2L}}}\mathbb{E}\left[(\boldsymbol{v}^\top\tilde{\boldsymbol{u}}_t)^4\right]$$

since $\tilde{\boldsymbol{u}}_t$ is centered and isotropic.

Given $\boldsymbol{v}\in\mathbb{R}^{p^2L}$ with $\|\boldsymbol{v}\|_2=1$, let $\boldsymbol{V}\in\mathbb{R}^{p\times pL}$ such that $\mathbf{vec}(\boldsymbol{V})=\boldsymbol{v}$ (so that $\|\boldsymbol{V}\|_F=1$). By the property of Kronecker product, we have

$$\mathbb{E}\left[(\boldsymbol{v}^\top\tilde{\boldsymbol{u}}_t)^4\right]=\mathbb{E}\left[((\bar{\boldsymbol{u}}_{t-1}\otimes\boldsymbol{u}_t)^\top\boldsymbol{v})^4\right]=\mathbb{E}\left[(\boldsymbol{u}_t^\top\boldsymbol{V}\bar{\boldsymbol{u}}_{t-1})^4\right].$$

Let $\boldsymbol{V}_1,\ldots,\boldsymbol{V}_L$ be the $p\times p$ matrices such that

$$\boldsymbol{V}=\begin{bmatrix}\boldsymbol{V}_1 & \boldsymbol{V}_2 & \cdots & \boldsymbol{V}_L\end{bmatrix}$$

then $\boldsymbol{u}_t^\top\boldsymbol{V}\bar{\boldsymbol{u}}_{t-1}=\sum_{k=1}^L\boldsymbol{u}_t^\top\boldsymbol{V}_k\boldsymbol{u}_{t-k}$. For simplicity, let $z_k=\boldsymbol{u}_t^\top\boldsymbol{V}_k\boldsymbol{u}_{t-k}$. Then, we have

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{u}_t^\top\boldsymbol{V}\bar{\boldsymbol{u}}_{t-1})^4\right] &= \mathbb{E}\left[\left(\sum_{k=1}^L z_k\right)^4\right] \\
&= \sum_{k=1}^L\mathbb{E}[z_k^4]+4\sum_{k\neq l}\mathbb{E}[z_k^3z_l]+6\sum_{k<l}\mathbb{E}[z_k^2z_l^2]+12\sum_{\substack{k\neq l,k\neq m\\ l\neq m}}\mathbb{E}[z_k^2z_lz_m]+24\sum_{k<l<m<n}\mathbb{E}[z_kz_lz_mz_n]
\end{aligned}$$

Since $\boldsymbol{u}_t, \boldsymbol{u}_{t-1}, \ldots, \boldsymbol{u}_{t-L}$ are independent, we know $\mathbb{E}[z_k] = 0$, and moreover, $\mathbb{E}\left[z_k \mid \boldsymbol{u}_t\right] = 0$. Then, for $k \neq l$,

$$\begin{aligned}
\mathbb{E}[z_k z_l] &= \mathbb{E}\left[\mathbb{E}\left[z_k z_l \mid \boldsymbol{u}_t\right]\right] && \text{(law of total expectation)} \\
&= \mathbb{E}\left[\mathbb{E}\left[z_k \mid \boldsymbol{u}_t\right] \mathbb{E}\left[z_l \mid \boldsymbol{u}_t\right]\right] && (z_k \text{ and } z_l \text{ are conditionally independent.}) \\
&= 0
\end{aligned}$$

Likewise, we can show that $\mathbb{E}[z_k^3 z_l] = 0$ for $k \neq l$, $\mathbb{E}[z_k^2 z_l z_m] = 0$ for $k \neq l$, $k \neq m$, and $l \neq m$, and $\mathbb{E}[z_k z_l z_m z_n] = 0$ for $k < l < m < n$, showing that

$$\mathbb{E}\left[(\boldsymbol{v}^\top \tilde{\boldsymbol{u}}_t)^4\right] = \mathbb{E}\left[(\boldsymbol{u}_t^\top \boldsymbol{V} \bar{\boldsymbol{u}}_{t-1})^4\right] = \sum_{k=1}^{L} \mathbb{E}[z_k^4] + 6 \sum_{k<l} \mathbb{E}[z_k^2 z_l^2]. \tag{28}$$

Now we upper bound $\mathbb{E}[z_k^4]$ for $k = 1, \ldots, L$, as follows: Let $\boldsymbol{v}_{k,1}, \ldots, \boldsymbol{v}_{k,p}$ denote the columns of $\boldsymbol{V}_k$, and $(\boldsymbol{u}_{t-k})_i$ denote the i-th entry of the Rademacher random vector $\boldsymbol{u}_{t-k}$. Then, we have

$$\begin{aligned}
\mathbb{E}[z_k^4] = \mathbb{E}\left[\left(\boldsymbol{u}_t^\top \boldsymbol{V}_k \boldsymbol{u}_{t-k}\right)^4\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{p} \boldsymbol{u}_t^\top \boldsymbol{v}_{k,i} (\boldsymbol{u}_{t-k})_i\right)^4\right], \\
&\overset{\text{(a)}}{=} \sum_{i=1}^{p} \mathbb{E}\left[\left(\boldsymbol{u}_t^\top \boldsymbol{v}_{k,i} (\boldsymbol{u}_{t-k})_i\right)^4\right] + 6 \sum_{i<j} \mathbb{E}\left[\left(\boldsymbol{u}_t^\top \boldsymbol{v}_{k,i} (\boldsymbol{u}_{t-k})_i\right)^2 \left(\boldsymbol{u}_t^\top \boldsymbol{v}_{k,j} (\boldsymbol{u}_{t-k})_j\right)^2\right], \\
&\overset{\text{(b)}}{=} \sum_{i=1}^{p} \mathbb{E}\left[\left(\boldsymbol{u}_t^\top \boldsymbol{v}_{k,i}\right)^4\right] + 6 \sum_{i<j} \mathbb{E}\left[\left(\boldsymbol{u}_t^\top \boldsymbol{v}_{k,i}\right)^2 \left(\boldsymbol{u}_t^\top \boldsymbol{v}_{k,j}\right)^2\right], \\
&\overset{\text{(c)}}{\leq} 3 \sum_{i=1}^{p} \|\boldsymbol{v}_{k,i}\|_2^4 + 18 \sum_{i<j} \|\boldsymbol{v}_{k,i}\|_2^2 \|\boldsymbol{v}_{k,j}\|_2^2, \\
&\leq 9 \left(\sum_{i=1}^{p} \|\boldsymbol{v}_{k,i}\|_2^2\right)^2 = 9 \|\boldsymbol{V}_k\|_F^4,
\end{aligned} \tag{29}$$

where we obtained (a) from using the same arguments as used for simplifying $\mathbb{E}\left[(\boldsymbol{u}_t^\top \boldsymbol{V} \bar{\boldsymbol{u}}_{t-1})^4\right]$ above, (b) follows from the independence of $(\boldsymbol{u}_{t-k})_i$, $(\boldsymbol{u}_{t-k})_i$ and $\mathbb{E}[(\boldsymbol{u}_{t-k})_i^4] = E[(\boldsymbol{u}_{t-k})_i^2 (\boldsymbol{u}_{t-k})_j^2] = 1$, and lastly (c) is obtained from the fact that, given fixed vetors $\boldsymbol{q}, \boldsymbol{q}_1, \boldsymbol{q}_2$, for any Rademacher random vector $\boldsymbol{u}$, we have $\mathbb{E}[(\boldsymbol{q}^\top \boldsymbol{u})^4] \leq 3 \|\boldsymbol{q}\|_2^4$ and $\mathbb{E}[(\boldsymbol{q}_1^\top \boldsymbol{u})^2 (\boldsymbol{q}_2^\top \boldsymbol{u})^2] \leq 3 \|\boldsymbol{q}_1\|_2^2 \|\boldsymbol{q}_2\|_2^2$.

Next, we upper bound $\mathbb{E}[z_k^2 z_l^2]$ for $k < l$ as follows: Let $\boldsymbol{v}_{k,1}, \ldots, \boldsymbol{v}_{k,p}$ denote the columns of $\boldsymbol{V}_k$, $\boldsymbol{v}_{l,1}, \ldots, \boldsymbol{v}_{l,p}$ denote the columns of $\boldsymbol{V}_l$, and $(\boldsymbol{u}_{t-k})_i$, $(\boldsymbol{u}_{t-l})_i$ denote the i-th entry of the Rademacher random vectors $\boldsymbol{u}_{t-k}$, $\boldsymbol{u}_{t-l}$ respectively. Then, we have

$$\begin{aligned}
\mathbb{E}[z_k^2 z_l^2] &= \mathbb{E}\left[\left(\boldsymbol{u}_t^\top \boldsymbol{V}_k \boldsymbol{u}_{t-k}\right)^2 \left(\boldsymbol{u}_t^\top \boldsymbol{V}_l \boldsymbol{u}_{t-l}\right)^2\right], \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{p} \boldsymbol{u}_t^\top \boldsymbol{v}_{k,i} (\boldsymbol{u}_{t-k})_i\right)^2 \left(\sum_{j=1}^{p} \boldsymbol{u}_t^\top \boldsymbol{v}_{l,j} (\boldsymbol{u}_{t-l})_j\right)^2\right], \\
&\overset{\text{(i)}}{=} \sum_{i=1}^{p} \sum_{j=1}^{p} \mathbb{E}\left[\left(\boldsymbol{u}_t^\top \boldsymbol{v}_{k,i}\right)^2 \left(\boldsymbol{u}_t^\top \boldsymbol{v}_{l,j}\right)^2\right], \\
&\overset{\text{(ii)}}{\leq} 3 \sum_{i=1}^{p} \sum_{j=1}^{p} \|\boldsymbol{v}_{k,i}\|_2^2 \|\boldsymbol{v}_{l,j}\|_2^2 = 3 \|\boldsymbol{V}_k\|_F^2 \|\boldsymbol{V}_l\|_F^2,
\end{aligned} \tag{30}$$

where we get (i) from the independence of $(\boldsymbol{u}_{t-k})_i, (\boldsymbol{u}_{t-l})_j$ for all $k < l$, and $i, j = 1, \ldots, p$, and (ii) is obtained from using the same argument as used to get (c) in (29). Finally, combining (29), and (30) into (28), we obtain the following upper bound,

$$\mathbb{E}\left[(\boldsymbol{v}^\top \tilde{\boldsymbol{u}}_t)^4\right] = \sum_{k=1}^{L} \mathbb{E}[z_k^4] + 6 \sum_{k<l} \mathbb{E}[z_k^2 z_l^2],$$

$$\leq 9\sum_{k=1}^{L}\|\boldsymbol{V}_k\|_F^4 + 18\sum_{k<l}\|\boldsymbol{V}_k\|_F^2\|\boldsymbol{V}_l\|_F^2 .$$
$$= 9\left(\sum_{k=1}^{L}\|\boldsymbol{V}_k\|_F^2\right)^2 = 9\|\boldsymbol{V}\|_F^2 = 9. \tag{31}$$

B.7 Miscellaneous

Lemma 2. *If the matrix norm $\|\cdot\|$ is consistent, the quantity*

$$\phi(\boldsymbol{A},\rho) = \sup_{k\in\mathbb{Z}_+}\frac{\|\boldsymbol{A}^k\|}{\rho^k}$$

is finite for all $\rho > \rho(\boldsymbol{A})$.

Proof. Recall Gelfand's formula: for any consistent matrix norm $\|\cdot\|$, we have

$$\rho(A) = \lim_{k\to\infty}\|\boldsymbol{A}^k\|^{1/k} \tag{32}$$

and the limit approaches from above. By Gelfand's formula, for all $\rho > \rho(\boldsymbol{A})$, there exists N such that

$$\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}^k\|^{1/k} < \rho$$

whenever $k \geq N$. Then, we have

$$\phi(\boldsymbol{A},\rho) = \max\left\{\max_{k=1,\ldots,N-1}\frac{\|\boldsymbol{A}^k\|}{\rho^k}, \sup_{k\geq N}\frac{\|\boldsymbol{A}^k\|}{\rho^k}\right\}$$

implying that $\phi(\boldsymbol{A},\rho)$ is finite because $\|\boldsymbol{A}^k\|/\rho^k < 1$ for $k \geq N$. □

Corollary 1. *Given $\rho > \rho(\boldsymbol{A})$, for all $k \geq 1$, we have $\|\boldsymbol{A}^k\| \leq \phi(\boldsymbol{A},\rho)\rho^k$.*

Proposition 5. *The following holds for $\lambda > 0$:*

$$\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right)^{-1} \preceq \tilde{\boldsymbol{U}}^{-1} \preceq \left(1 + \frac{\lambda}{\lambda_{\min}(\tilde{\boldsymbol{U}})}\right)\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right)^{-1}$$

Proof. Since $\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}} \succeq \tilde{\boldsymbol{U}}$ for $\lambda > 0$, we have $\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right)^{-1} \preceq \tilde{\boldsymbol{U}}^{-1}$. In order to show the other Loewner order, it is equivalent to show that

$$\tilde{\boldsymbol{U}} \succeq \frac{\lambda_{\min}(\tilde{\boldsymbol{U}})}{\lambda_{\min}(\tilde{\boldsymbol{U}}) + \lambda}\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right).$$

We want to find $\gamma > 0$ such that $\tilde{\boldsymbol{U}} - \gamma\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right) \succeq 0$, that is, $\lambda_{\min}\left(\tilde{\boldsymbol{U}} - \gamma\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right)\right) \geq 0$.

$$\begin{aligned}\lambda_{\min}\left(\tilde{\boldsymbol{U}} - \gamma\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right)\right) &= \lambda_{\min}\left((1-\gamma)\tilde{\boldsymbol{U}} - \gamma\lambda\boldsymbol{I}_d\right)\\ &= (1-\gamma)\lambda_{\min}(\tilde{\boldsymbol{U}}) - \gamma\lambda\\ &= \lambda_{\min}(\tilde{\boldsymbol{U}}) - \gamma\left(\lambda_{\min}(\tilde{\boldsymbol{U}}) + \lambda\right) \geq 0\end{aligned}$$

Hence, for $0 < \gamma \leq \frac{\lambda_{\min}(\tilde{\boldsymbol{U}})}{\lambda_{\min}(\tilde{\boldsymbol{U}})+\lambda}$, or $\frac{1}{\gamma} \geq 1 + \frac{\lambda}{\lambda_{\min}(\tilde{\boldsymbol{U}})}$, we have

$$\tilde{\boldsymbol{U}}^{-1} \preceq \frac{1}{\gamma}\left(\lambda\boldsymbol{I}_d + \tilde{\boldsymbol{U}}\right)^{-1}.$$

This completes the proof. □

Corollary 2. *For any $\boldsymbol{x} \in \mathbb{R}^d$, we have*

$$\left\|\tilde{\boldsymbol{U}}^{-1/2}\boldsymbol{x}\right\|_2 \leq \sqrt{1+\frac{\lambda}{\lambda_{\min}(\tilde{\boldsymbol{U}})}}\left\|\left(\lambda \boldsymbol{I}_d+\tilde{\boldsymbol{U}}\right)^{-1/2}\boldsymbol{x}\right\|_2.$$

Proof. Since $\left(1+\frac{\lambda}{\lambda_{\min}(\tilde{\boldsymbol{U}})}\right)\left(\lambda \boldsymbol{I}_d+\tilde{\boldsymbol{U}}\right)^{-1}-\tilde{\boldsymbol{U}}^{-1}$ is positive semi-definite by Proposition 5, for any $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$\boldsymbol{x}^\top \tilde{\boldsymbol{U}}^{-1}\boldsymbol{x} \leq \left(1+\frac{\lambda}{\lambda_{\min}(\tilde{\boldsymbol{U}})}\right)\boldsymbol{x}^\top\left(\lambda \boldsymbol{I}_d+\tilde{\boldsymbol{U}}\right)^{-1}\boldsymbol{x}.$$

which completes the proof. □

Theorem 6 (Self-Normalized Bound for Vector-Valued Martingales, (Abbasi-Yadkori et al., 2011))**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}\left[e^{\lambda \eta_t} \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{x}_t$ is $\mathcal{F}_{t-1}$-measurable. Assume that $\boldsymbol{V}$ is a $d \times d$ positive definite matrix. For any $t \geq 0$, define

$$\bar{\boldsymbol{V}}_t = \boldsymbol{V}+\sum_{s=1}^{t}\boldsymbol{x}_s\boldsymbol{x}_s^\top \qquad \boldsymbol{s}_t = \sum_{s=1}^{t}\eta_s\boldsymbol{x}_s.$$

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

$$\|\bar{\boldsymbol{V}}^{-1/2}\boldsymbol{s}_t\| \leq R\sqrt{2\log\left(\frac{\det(\bar{\boldsymbol{V}}_t)^{1/2}\det(\boldsymbol{V})^{-1/2}}{\delta}\right)}.$$

C REGRET ANALYSIS

In this section, we derive the regret in detail and prove Proposition 5 and 6. Recall that

$$\begin{aligned}
&\boldsymbol{u}_{0:T}^\star = \arg\max_{\boldsymbol{u}_{0:T}} \mathbb{E}\left[\sum_{t=0}^{T} r_t\right] = \frac{1}{2}\boldsymbol{u}_{0:T}^\top \boldsymbol{S}_T \boldsymbol{u}_{0:T} \quad \text{subject to } \boldsymbol{u}_{0:T} \in \{-1,+1\}^{p(T+1)},\\
&\boldsymbol{u}_{0:H}^\pi \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{-1,+1\}^p), \quad \text{and}\\
&\boldsymbol{u}_{H+1:T}^\pi = \arg\max_{\boldsymbol{u}_{H+1:T}} \frac{1}{2}\boldsymbol{u}_{H+1:T}^\top \hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{u}_{H+1:T}^\top \quad \text{subject to } \boldsymbol{u}_{H+1:T} \in \{-1,+1\}^{p(T-H)}.
\end{aligned}$$

where π denotes the ETC algorithm policy. Let $\{r_t^\star\}_{t=0}^T$ denote the random rewards collected under $\boldsymbol{u}_{0:T}^\star$, and $\{r_t^\pi\}_{t=0}^T$ denote the random rewards collected by policy π. Then the regret of π is

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} r_t^\star - \sum_{t=0}^{T} r_t^\pi\right] \quad \text{where} \quad \mathbb{E}\left[\sum_{t=0}^{T} r_t^\star\right] = \frac{1}{2}(\boldsymbol{u}_{0:T}^\star)^\top \boldsymbol{S}_T \boldsymbol{u}_{0:T}.$$

For the expected cumulative reward for π, we split into the two phases. For $t = 0, \ldots, H$, we have $\{\boldsymbol{u}_t^\pi\}_{t=0}^\pi \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{-1,+1\}^p)$, and note that the distribution has zero-mean and identity covariance. Since the conditional expectation is

$$\mathbb{E}\left[\sum_{t=1}^{H} r_t^\pi \,\middle|\, \{\boldsymbol{u}_t^\pi\}_{t=0}^H\right] = (\boldsymbol{u}_{0:H}^\pi)^\top \boldsymbol{M}_H \boldsymbol{u}_{0:H}^\pi = \mathbf{tr}(\boldsymbol{M}_H \boldsymbol{u}_{0:H}^\pi (\boldsymbol{u}_{0:H}^\pi)^\top),$$

the law of total expectation implies that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{H} r_t^\pi\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^{H} y_t^\pi \,\middle|\, \{\boldsymbol{u}_t^\pi\}_{t=0}^{H}\right]\right] \\ &= \mathbb{E}\left[\mathbf{tr}(\boldsymbol{M}_H \boldsymbol{u}_{0:H}^\pi (\boldsymbol{u}_{0:H}^\pi)^\top)\right] \\ &= \mathbf{tr}\left(\boldsymbol{M}_H \mathbb{E}\left[\boldsymbol{u}_{0:H}^\pi (\boldsymbol{u}_{0:H}^\pi)^\pi\right]\right) \\ &= \mathbf{tr}(\boldsymbol{M}_H \boldsymbol{I}_{p(H+1)}) \\ &= 0\end{aligned}$$

Now we consider the remaining steps. Let $\boldsymbol{x}_{H+1}^\pi$ be the random state after the exploration, then we know that $\mathbb{E}\left[\boldsymbol{x}_{H+1}^\pi\right] = \mathbf{0}_n$ because $\boldsymbol{x}_0 = \mathbf{0}_n$ and the expectation of noise processes and Rademacher actions are zero. Since we have

$$\mathbb{E}\left[\sum_{t=H+1}^{T} r_t^\pi \,\middle|\, \boldsymbol{x}_{H+1}^\pi\right] = \sum_{t=H+1}^{T} (\boldsymbol{u}_t^\pi)^\top \boldsymbol{C}\boldsymbol{A}^{t-H-1}\boldsymbol{x}_{H+1}^\pi + \frac{1}{2}(\boldsymbol{u}_{H+1:T}^\pi)^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}_{H+1:T}^\pi$$

the law of total expectation again implies that

$$\mathbb{E}\left[\sum_{t=H+1}^{T} r_t^\pi\right] = \frac{1}{2}(\boldsymbol{u}_{H+1:T}^\pi)^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}_{H+1:T}^\pi.$$

Therefore, the regret is

$$R_T(\pi) = \frac{1}{2}(\boldsymbol{u}_{0:T}^\star)^\top \boldsymbol{S}_T \boldsymbol{u}_{0:T} - \frac{1}{2}(\boldsymbol{u}_{H+1:T}^\pi)^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}_{H+1:T}^\pi.$$

For the regret analysis, let

$$\tilde{\boldsymbol{u}}_{H+1:T} = \underset{\boldsymbol{u}_{H+1:T}}{\arg\max}\, \frac{1}{2}\boldsymbol{u}_{H+1:T}^\top \boldsymbol{S}_{T-H-1}\boldsymbol{u}_{H+1:T} \quad \text{subject to} \quad \boldsymbol{u}_{H+1:T} \in \{-1,+1\}^{p(T-H)}.$$

Then, we decompose the regret as follows:

$$R_T(\pi) = \frac{1}{2}\left(R_{1,T} + R_{2,T} + R_{3,T}\right) \tag{33}$$

where

$$\begin{aligned}R_{1,T} &:= (\boldsymbol{u}_{0:T}^\star)^\top \boldsymbol{S}_T \boldsymbol{u}_{0:T}^\star - (\tilde{\boldsymbol{u}}_{H+1:T})^\top \boldsymbol{S}_{T-H-1}\tilde{\boldsymbol{u}}_{H+1:T} \\ R_{2,T} &:= (\tilde{\boldsymbol{u}}_{H+1:T})^\top \boldsymbol{S}_{T-H-1}\tilde{\boldsymbol{u}}_{H+1:T} - (\boldsymbol{u}_{H+1:T}^\pi)^\top \hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{u}_{H+1:T}^\pi \\ R_{3,T} &:= (\boldsymbol{u}_{H+1:T}^\pi)^\top (\hat{\boldsymbol{S}}_{T-H-1} - \boldsymbol{S}_{T-H-1})\boldsymbol{u}_{H+1:T}^\pi.\end{aligned}$$

C.1 Proof of Proposition 5

Proposition 5. *Let $\rho \in (\rho(\boldsymbol{A}), 1)$ be given. Then*

$$R_{1,T} \le 2p\kappa^2\left(\alpha H + \beta\right),$$

for $\alpha = 1 + \frac{\phi(\boldsymbol{A},\rho)\rho}{1-\rho}$, $\beta = \frac{\phi(\boldsymbol{A},\rho)\rho}{(1-\rho)^2} + 1$, and $\kappa = \max\{\|\boldsymbol{B}\|, \|\boldsymbol{C}\|\}$.

Proof. Observe that $\boldsymbol{S}_{T-H-1}$ is a submatrix of $\boldsymbol{S}_T$. Define $\tilde{\boldsymbol{S}}_T$ by zero-padding the matrix $\boldsymbol{S}_{T-H-1}$ so that its size is the same as $\boldsymbol{S}_T$, that is,

$$\tilde{\boldsymbol{S}}_T = \begin{bmatrix} \boldsymbol{S}_{T-H-1} & \mathbf{0}_{p(T-H)\times p(H+1)} \\ \mathbf{0}_{p(H+1)\times p(T-H)} & \mathbf{0}_{p(H+1)\times p(H+1)} \end{bmatrix}.$$

Let $\boldsymbol{\Delta}_T = \boldsymbol{S}_T - \tilde{\boldsymbol{S}}_T$, then

$$\begin{aligned}
R_{1,T} &= (\boldsymbol{u}^\star_{0:T})^\top \boldsymbol{S}_T \boldsymbol{u}^\star_{0:T} - (\tilde{\boldsymbol{u}}_{H+1:T})^\top \boldsymbol{S}_{T-H-1} \tilde{\boldsymbol{u}}_{H+1:T} \\
&= \max_{\substack{\|\boldsymbol{u}\|_\infty = 1 \\ \boldsymbol{u} \in \mathbb{R}^{p(T+1)}}} \boldsymbol{u}^\top \boldsymbol{S}_T \boldsymbol{u} - \max_{\substack{\|\boldsymbol{v}\|_\infty = 1 \\ \boldsymbol{v} \in \mathbb{R}^{p(T-H)}}} \boldsymbol{v}^\top \boldsymbol{S}_{T-H-1} \boldsymbol{v} \\
&= \max_{\substack{\|\boldsymbol{u}\|_\infty = 1 \\ \boldsymbol{u} \in \mathbb{R}^{p(T+1)}}} \boldsymbol{u}^\top \boldsymbol{S}_T \boldsymbol{u} - \max_{\substack{\|\boldsymbol{v}\|_\infty = 1 \\ \boldsymbol{v} \in \mathbb{R}^{p(T+1)}}} \boldsymbol{v}^\top \tilde{\boldsymbol{S}}_T \boldsymbol{v} \\
&= \max_{\substack{\|\boldsymbol{u}\|_\infty = 1 \\ \boldsymbol{u} \in \mathbb{R}^{p(T+1)}}} \min_{\substack{\|\boldsymbol{v}\|_\infty = 1 \\ \boldsymbol{v} \in \mathbb{R}^{p(T+1)}}} \left[\boldsymbol{u}^\top (\tilde{\boldsymbol{S}}_T + \boldsymbol{\Delta}_T) \boldsymbol{u} - \boldsymbol{v}^\top \tilde{\boldsymbol{S}}_T \boldsymbol{v} \right] \\
&\leq \max_{\substack{\|\boldsymbol{u}\|_\infty = 1 \\ \boldsymbol{u} \in \mathbb{R}^{p(T+1)}}} \boldsymbol{u}^\top \boldsymbol{\Delta}_T \boldsymbol{u} && \text{(take } \boldsymbol{v} = \boldsymbol{u}\text{)}
\end{aligned}$$

To be concrete for $\boldsymbol{\Delta}_T$, observe that

$$\boldsymbol{\Delta}_T = \begin{bmatrix} \boldsymbol{0}_{p(T-H)\times p(T-H)} & \boldsymbol{R} \\ \boldsymbol{R}^\top & \boldsymbol{S}_{\text{tail}} \end{bmatrix}$$

where

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{C}\boldsymbol{A}^{T-H-1}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}^{T-H}\boldsymbol{B} & \cdots & \boldsymbol{C}\boldsymbol{A}^{T-1}\boldsymbol{B} \\ \boldsymbol{C}\boldsymbol{A}^{T-H-2}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}^{T-H-1}\boldsymbol{B} & \cdots & \boldsymbol{C}\boldsymbol{A}^{T-2}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}\boldsymbol{B} & \cdots & \boldsymbol{C}\boldsymbol{A}^{H}\boldsymbol{B} \end{bmatrix} \in \mathbb{R}^{p(T-H)\times p(H+1)}$$

and

$$\boldsymbol{S}_{\text{tail}} = \begin{bmatrix} \boldsymbol{0}_{p\times p} & \boldsymbol{C}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}\boldsymbol{B} & \cdots & \boldsymbol{C}\boldsymbol{A}^{H-2}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}^{H-1}\boldsymbol{B} \\ (\boldsymbol{C}\boldsymbol{B})^\top & \boldsymbol{0}_{p\times p} & \boldsymbol{C}\boldsymbol{B} & \cdots & \boldsymbol{C}\boldsymbol{A}^{H-3}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}^{H-2}\boldsymbol{B} \\ (\boldsymbol{C}\boldsymbol{A}\boldsymbol{B})^\top & (\boldsymbol{C}\boldsymbol{B})^\top & \boldsymbol{0}_{p\times p} & \cdots & \boldsymbol{C}\boldsymbol{A}^{H-4}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{A}^{H-3}\boldsymbol{B} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ (\boldsymbol{C}\boldsymbol{A}^{H-2}\boldsymbol{B})^\top & (\boldsymbol{C}\boldsymbol{A}^{H-3}\boldsymbol{B})^\top & (\boldsymbol{C}\boldsymbol{A}^{H-4}\boldsymbol{B})^\top & \cdots & \boldsymbol{0}_{p\times p} & \boldsymbol{C}\boldsymbol{B} \\ (\boldsymbol{C}\boldsymbol{A}^{H-1}\boldsymbol{B})^\top & (\boldsymbol{C}\boldsymbol{A}^{H-2}\boldsymbol{B})^\top & (\boldsymbol{C}\boldsymbol{A}^{H-3}\boldsymbol{B})^\top & \cdots & (\boldsymbol{C}\boldsymbol{B})^\top & \boldsymbol{0}_{p\times p} \end{bmatrix} \in \mathbb{R}^{p(H+1)\times p(H+1)}$$

By Proposition 7, we can find an upper bound in terms of the spectral norms of the blocks as follows:

$$\max_{\substack{\|\boldsymbol{u}\|_\infty = 1 \\ \boldsymbol{u} \in \mathbb{R}^{p(T+1)}}} \boldsymbol{u}^\top \boldsymbol{\Delta}_T \boldsymbol{u} \leq \underbrace{2p \sum_{i=0}^{H} \sum_{j=0}^{T-H-1} \|\boldsymbol{C}\boldsymbol{A}^{i+j}\boldsymbol{B}\|}_{\text{blocks in } \boldsymbol{R}, \boldsymbol{R}^\top} + \underbrace{2p \sum_{l=1}^{H} \sum_{k=1}^{l} \|\boldsymbol{C}\boldsymbol{A}^{k-1}\boldsymbol{B}\|}_{\text{blocks in } \boldsymbol{S}_{\text{tail}}}$$

We use Corollary 1 in order to bound the upper bound above. Let $\rho \in (\rho(\boldsymbol{A}), 1)$ be given. Then the term related to the blocks in $\boldsymbol{R}$ and $\boldsymbol{R}^\top$ is bounded by,

$$\begin{aligned}
2p \sum_{i=0}^{H} \sum_{j=0}^{T-H-1} \|\boldsymbol{C}\boldsymbol{A}^{i+j}\boldsymbol{B}\| &= 2p\|\boldsymbol{C}\boldsymbol{B}\| + 2p \sum_{j=1}^{T-H-1} \|\boldsymbol{C}\boldsymbol{A}^{j}\boldsymbol{B}\| + 2p \sum_{i=1}^{H} \sum_{j=0}^{T-H-1} \|\boldsymbol{C}\boldsymbol{A}^{i+j}\boldsymbol{B}\| \\
&\leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\| \left(1 + \sum_{j=1}^{T-H-1} \|\boldsymbol{A}^j\| + \sum_{i=1}^{H} \sum_{j=0}^{T-H-1} \|\boldsymbol{A}^{i+j}\| \right) \\
&\leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\| \left(1 + \phi(\boldsymbol{A}, \rho) \sum_{j=1}^{T-H-1} \rho^j + \phi(\boldsymbol{A}, \rho) \sum_{i=1}^{H} \sum_{j=0}^{T-H-1} \rho^{i+j} \right) \\
&\leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\| \left(1 + \phi(\boldsymbol{A}, \rho) \sum_{j=1}^{\infty} \rho^j + \phi(\boldsymbol{A}, \rho) \sum_{i=1}^{\infty} \rho^i \sum_{j=0}^{\infty} \rho^j \right)
\end{aligned}$$

$$\leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\|\left(1+\frac{\phi(\boldsymbol{A},\rho)\rho}{1-\rho}+\frac{\phi(\boldsymbol{A},\rho)\rho}{(1-\rho)^2}\right)$$

Similarly, the term related to the blocks in $\boldsymbol{S}_{\text{tail}}$ is bounded by,

$$\begin{aligned}
2p\sum_{l=1}^{H}\sum_{k=1}^{l}\|\boldsymbol{C}\boldsymbol{A}^{k-1}\boldsymbol{B}\| &= 2pH\|\boldsymbol{C}\boldsymbol{B}\| + 2p\sum_{l=1}^{H-1}\sum_{k=1}^{l}\|\boldsymbol{C}\boldsymbol{A}^{k}\boldsymbol{B}\| \\
&\leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\|\left(H+\phi(\boldsymbol{A},\rho)\sum_{l=1}^{H-1}\sum_{k=1}^{l}\rho^k\right) \\
&\leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\|\left(H+\phi(\boldsymbol{A},\rho)\sum_{l=1}^{H-1}\sum_{k=1}^{\infty}\rho^k\right) \\
&\leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\|\left(H+\phi(\boldsymbol{A},\rho)\sum_{l=1}^{H-1}\frac{\rho}{1-\rho}\right) \\
&= 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\|\left(H+\frac{(H-1)\phi(\boldsymbol{A},\rho)\rho}{1-\rho}\right)
\end{aligned}$$

Combining two bounds, we get

$$R_{1,T} \leq \max_{\substack{\|\boldsymbol{u}\|_\infty=1 \\ \boldsymbol{u}\in\mathbb{R}^{p(T+1)}}} \boldsymbol{u}^\top\boldsymbol{\Delta}_T\boldsymbol{u} \leq 2p\|\boldsymbol{B}\|\|\boldsymbol{C}\|\left(\left(1+\frac{\phi(\boldsymbol{A},\rho)}{1-\rho}\right)H+\frac{\phi(\boldsymbol{A},\rho)\rho}{(1-\rho)^2}+1\right)$$

which completes the proof. □

C.2 Proof of Proposition 6

Proposition 6. *Let $\rho \in (\rho(\boldsymbol{A}),1)$ be given and $\epsilon > 0$ be the high-probability parameter estimation error, i.e., $\|\boldsymbol{G}-\hat{\boldsymbol{G}}\|_F \leq \epsilon$. Then, with high probability,*

$$\max\{R_{2,T}, R_{3,T}\} \leq 2p(T-H)\left(\epsilon+\kappa^2\gamma_L\right)$$

where $\kappa = \max\{\|\boldsymbol{B}\|,\|\boldsymbol{C}\|\}$ and $\gamma_L = \frac{\phi(\boldsymbol{A},\rho)\rho^L}{1-\rho}$.

Proof. Observe the following bounds for $R_{2,T}$ and $R_{3,T}$.

$$\begin{aligned}
R_{2,T} &= (\tilde{\boldsymbol{u}}_{H+1:T})^\top\boldsymbol{S}_{T-H-1}\tilde{\boldsymbol{u}}_{H+1:T} - (\boldsymbol{u}^\pi_{H+1:T})^\top\hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{u}^\pi_{H+1:T} \\
&= \max_{\|\boldsymbol{u}\|_\infty=1}\boldsymbol{u}^\top\boldsymbol{S}_{T-H-1}\boldsymbol{u} - \max_{\|\boldsymbol{v}\|_\infty=1}\boldsymbol{v}^\top\hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{v} \\
&= \max_{\|\boldsymbol{u}\|_\infty=1}\min_{\|\boldsymbol{v}\|_\infty=1}\left[\boldsymbol{u}^\top\boldsymbol{S}_{T-H-1}\boldsymbol{u} - \boldsymbol{v}^\top\hat{\boldsymbol{S}}_{T-H-1}\boldsymbol{v}\right] \\
&\leq \max_{\|\boldsymbol{u}\|_\infty=1}\boldsymbol{u}^\top(\boldsymbol{S}_{T-H-1}-\hat{\boldsymbol{S}}_{T-H-1})\boldsymbol{u} \qquad \text{(take } \boldsymbol{v}=\boldsymbol{u}\text{)}
\end{aligned}$$

$$\begin{aligned}
R_{3,T} &= (\boldsymbol{u}^\pi_{H+1:T})^\top(\hat{\boldsymbol{S}}_{T-H-1}-\boldsymbol{S}_{T-H-1})\boldsymbol{u}^\pi_{H+1:T} \\
&\leq \max_{\|\boldsymbol{u}\|_\infty=1}\boldsymbol{u}^\top(\hat{\boldsymbol{S}}_{T-H-1}-\boldsymbol{S}_{T-H-1})\boldsymbol{u}
\end{aligned}$$

According to Proposition 7, we now see that $R_{2,T}$ and $R_{3,T}$ can be bounded by a same quantity, which will be specified in this proof.

Let $\hat{\boldsymbol{M}}_{T-H-1}$ be the upper diagonal part of $\hat{\boldsymbol{S}}_{T-H-1}$, that is, $\hat{\boldsymbol{S}}_{T-H-1} = \hat{\boldsymbol{M}}_{T-H-1} + \hat{\boldsymbol{M}}^\top_{T-H-1}$. Recall that

$\boldsymbol{S}_{T-H-1} = \boldsymbol{M}_{T-H-1} + \boldsymbol{M}_{T-H-1}^\top$ where

$$\boldsymbol{M}_{T-H-1} = \begin{bmatrix} \boldsymbol{0}_{p\times p} & \boldsymbol{CB} & \boldsymbol{CAB} & \cdots & \cdots & \boldsymbol{CA}^{T-H-2}\boldsymbol{B} \\ \boldsymbol{0}_{p\times p} & \boldsymbol{0}_{p\times p} & \boldsymbol{CB} & \cdots & \cdots & \boldsymbol{CA}^{T-H-3}\boldsymbol{B} \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \boldsymbol{0}_{p\times p} & \boldsymbol{CB} & \boldsymbol{CAB} \\ \vdots & & & \ddots & \boldsymbol{0}_{p\times p} & \boldsymbol{CB} \\ \boldsymbol{0}_{p\times p} & \boldsymbol{0}_{p\times p} & \boldsymbol{0}_{p\times p} & \cdots & \boldsymbol{0}_{p\times p} & \boldsymbol{0}_{p\times p} \end{bmatrix} \in \mathbb{R}^{p(T-H)\times p(T-H)}.$$

Observe that

$$\boldsymbol{u}^\top(\boldsymbol{S}_{T-H-1} - \hat{\boldsymbol{S}}_{T-H-1})\boldsymbol{u} = 2\boldsymbol{u}^\top(\boldsymbol{M}_{T-H-1} - \hat{\boldsymbol{M}}_{T-H-1})\boldsymbol{u}.$$

Let $\boldsymbol{\Delta}_{ij}$ denote the (i,j)-th $p\times p$ block of $\boldsymbol{M}_{T-H-1} - \hat{\boldsymbol{M}}_{T-H-1}$. By Proposition 7,

$$\max_{\|\boldsymbol{u}\|_\infty=1} \boldsymbol{u}^\top(\boldsymbol{M}_{T-H-1} - \hat{\boldsymbol{M}}_{T-H-1})\boldsymbol{u} \le p\left(\sum_{(i,j)\in\mathcal{I}} \|\boldsymbol{\Delta}_{ij}\| + \sum_{(i,j)\in\mathcal{I}^c} \|\boldsymbol{\Delta}_{ij}\|_F\right)$$

where $\mathcal{I} = \{(i,j)\colon L+1 \le i+L < j \le T-H\}$. Define $\mathcal{I}_k = \{(k,j)\colon (k,j)\in\mathcal{I}\}$ and $\mathcal{J}_k = \{(k,j)\colon (k,j)\in\mathcal{I}^c\}$ for $k = 1,\dots,T-H$, then

$$\mathcal{I} = \dot{\cup}_{k=1}^{T-H-L-1}\mathcal{I}_k, \quad \mathcal{J} = \dot{\cup}_{k=1}^{T-H}\mathcal{J}_k,$$

and $\mathcal{I}_k = \varnothing$ for $k = T-H-L,\dots,T-H$. For each $k = 1,\dots,T-H-L-1$,

$$\sum_{(k,j)\in\mathcal{I}_k} \|\boldsymbol{\Delta}_{ij}\| + \sum_{(k,j)\in\mathcal{J}_k} \|\boldsymbol{\Delta}_{ij}\|_F = \sum_{\ell=L+1}^{T-H-k} \|\boldsymbol{CA}^{\ell-1}\boldsymbol{B}\| + \|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F$$

and for $k = T-H-L,\dots,T-H$,

$$\sum_{(k,j)\in\mathcal{I}_k} \|\boldsymbol{\Delta}_{ij}\| + \sum_{(k,j)\in\mathcal{J}_k} \|\boldsymbol{\Delta}_{ij}\|_F = 0 + \sum_{(k,j)\in\mathcal{J}_k} \|\boldsymbol{\Delta}_{ij}\|_F \le \|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F$$

Therefore, we obtain

$$\begin{aligned}
\max_{\|\boldsymbol{u}\|_\infty=1} \boldsymbol{u}^\top(\boldsymbol{S}_{T-H-1} - \hat{\boldsymbol{S}}_{T-H-1})\boldsymbol{u} &= 2\max_{\|\boldsymbol{u}\|_\infty=1} \boldsymbol{u}^\top(\boldsymbol{M}_{T-H-1} - \hat{\boldsymbol{M}}_{T-H-1})\boldsymbol{u} \\
&\le 2p\left(\sum_{(i,j)\in\mathcal{I}} \|\boldsymbol{\Delta}_{ij}\| + \sum_{(i,j)\in\mathcal{I}^c} \|\boldsymbol{\Delta}_{ij}\|_F\right) \\
&= 2p\sum_{k=1}^{T-H}\left(\sum_{(k,j)\in\mathcal{I}_k} \|\boldsymbol{\Delta}_{ij}\| + \sum_{(k,j)\in\mathcal{J}_k} \|\boldsymbol{\Delta}_{ij}\|_F\right) \\
&\le 2p\sum_{k=1}^{T-H-L-1}\left(\sum_{\ell=L+1}^{T-H-k} \|\boldsymbol{CA}^{\ell-1}\boldsymbol{B}\| + \|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F\right) + 2p(L+1)\|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F \\
&\le 2p\sum_{k=1}^{T-H-L-1}\sum_{\ell=L+1}^{T-H-k} \|\boldsymbol{CA}^{\ell-1}\boldsymbol{B}\| + 2p(T-H)\|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F
\end{aligned}$$

We again use Corollary 1. Given $\rho \in (\rho(\boldsymbol{A}), 1)$,

$$\sum_{k=1}^{T-H-L-1}\sum_{\ell=L+1}^{T-H-k} \|\boldsymbol{CA}^{\ell-1}\boldsymbol{B}\| \le \phi(\boldsymbol{A},\rho)\|\boldsymbol{B}\|\|\boldsymbol{C}\| \sum_{k=1}^{T-H-L-1}\sum_{\ell=L+1}^{T-H-k} \rho^{\ell-1}$$

$$
\begin{aligned}
&= \phi(\boldsymbol{A},\rho)\|\boldsymbol{B}\|\|\boldsymbol{C}\| \sum_{k=1}^{T-H-L-1} \frac{\rho^L - \rho^{T-H-k}}{1-\rho} \\
&= \phi(\boldsymbol{A},\rho)\|\boldsymbol{B}\|\|\boldsymbol{C}\| \left(\frac{(T-H-L-1)\rho^L}{1-\rho} - \sum_{k=1}^{T-H-L-1} \frac{\rho^{T-H-k}}{1-\rho} \right) \\
&\leq (T-H)\phi(\boldsymbol{A},\rho)\|\boldsymbol{B}\|\|\boldsymbol{C}\| \frac{\rho^L}{1-\rho}
\end{aligned}
$$

In conclusion, we get a final upper bound as follows:

$$
\max_{\|\boldsymbol{u}\|_\infty=1} \boldsymbol{u}^\top (\boldsymbol{S}_{T-H-1} - \hat{\boldsymbol{S}}_{T-H-1})\boldsymbol{u} \leq 2p(T-H) \left(\|\boldsymbol{B}\|\|\boldsymbol{C}\| \frac{\phi(\boldsymbol{A},\rho)\rho^L}{1-\rho} + \|\boldsymbol{G} - \hat{\boldsymbol{G}}\|_F \right)
$$

This completes the proof. □

C.3 Miscellaneous

Lemma 3. *For a square $n \times n$ matrix $\boldsymbol{M} = (M_{ij})$, we have*

$$
\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} \leq n\|\boldsymbol{M}\|_F \quad \textit{and} \quad \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} \leq n\|\boldsymbol{M}\|_2
$$

whenever $\|\boldsymbol{x}\|_\infty \leq 1$.

Proof. Assuming $\|\boldsymbol{x}\|_\infty \leq 1$, the first result can be proved as follows:

$$
\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} = \sum_{i,j=1}^{n} M_{ij}x_ix_j \leq \sum_{i,j=1}^{n} |M_{ij}| \leq \sqrt{n^2}\sqrt{\sum_{i,j=1}^{n} |M_{ij}|^2} = n\|\boldsymbol{M}\|_F
$$

where the second inequality is obtained using Cauchy-Schwarz inequality. The second result can be readily obtained:

$$
\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} \leq \|\boldsymbol{M}\|_2\|\boldsymbol{x}\|_2^2 \leq n\|\boldsymbol{M}\|_2
$$

because $\|\boldsymbol{x}\|_\infty = 1$ implies $\|\boldsymbol{x}\|_2 \leq \sqrt{n}$. □

Proposition 7. *Let $\boldsymbol{A}$ be a block matrix given by*

$$
\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1n} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{n1} & \boldsymbol{A}_{n2} & \cdots & \boldsymbol{A}_{nn} \end{bmatrix}
$$

where $\boldsymbol{A}_{ij} \in \mathbb{R}^{p\times p}$. Let $\mathcal{I}$ be any subset of $\{(i,j) \colon 1 \leq i,j \leq n\}$. Then we have the following:

$$
\max_{\|\boldsymbol{x}\|_\infty=1} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \leq p \left(\sum_{(i,j)\in\mathcal{I}} \|\boldsymbol{A}_{ij}\|_2 + \sum_{(i,j)\in\mathcal{I}^c} \|\boldsymbol{A}_{ij}\|_F \right) \tag{34}
$$

where $\boldsymbol{x} \in \mathbb{R}^{pn}$.

Proof. Let $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1^\top & \cdots & \boldsymbol{x}_n^\top \end{bmatrix}^\top$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ for each i. First, observe the equality:

$$
\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \sum_{i,j=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{A}_{ij}\boldsymbol{x}_j.
$$

For arbitrarily chosen index pair (i,j), Lemma 3 implies that

$$
\boldsymbol{x}_i^\top \boldsymbol{A}_{ij}\boldsymbol{x}_j \leq p\|\boldsymbol{A}_{ij}\|_F \quad \text{and} \quad \boldsymbol{x}_i^\top \boldsymbol{A}_{ij}\boldsymbol{x}_j \leq p\|\boldsymbol{A}_{ij}\|_2.
$$

This completes the proof. □

D NUMERICAL EXPERIMENTS

In this section, we give further details regarding numerical experiments. In Section D.1, we give a lower bound of the expected output of Goemans-Williamson random hyperplane rounding method, which is similar but different from MaxCut problems. In Section D.2, we describe further discussion about the experimental results in parameter estimation. In Section D.3, we design additional numerical experiments to show that our practical approach works for dense matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, and we discuss the experimental results with different spectral radii of $\boldsymbol{A}$.

D.1 Goemans-Williamson Random Hyperplane Rounding

We recall the problem stated in Section 6.1.

$$\begin{aligned} \text{maximize } & \mathbf{tr}(\boldsymbol{W}\boldsymbol{X}) \\ \text{subject to } & \boldsymbol{X} \succeq 0, \operatorname{rank}(\boldsymbol{X}) = 1 \\ & \boldsymbol{X}_{ii} = 1, \quad i = 1, \dots, n. \end{aligned} \tag{19}$$

Its semidefinite relaxation is

$$\begin{aligned} \text{maximize } & \mathbf{tr}(\boldsymbol{W}\boldsymbol{X}) \\ \text{subject to } & \boldsymbol{X} \succeq 0 \\ & \boldsymbol{X}_{ii} = 1, \quad i = 1, \dots, n. \end{aligned} \tag{35}$$

In Section 6.1, we discussed Goemans-Williamson random hyperplane rounding method to find a feasible point of (19) from its relaxed problem (35). For MaxCut problems, `SDP+GW` achieves an α-approximation algorithm with $\alpha \approx 0.87856$. This means that, for any given instance of MaxCut problem, the algorithm is guaranteed on average to return a solution whose value is at least 87.856% of the true optimum. In the next proposition, we derive a similar result where α depends on the matrix $\boldsymbol{W}$ in (19).

Let ν and ν_{rlx} be the optimal value of the original problem (19) and its relaxed problem (35), respectively. It is clear that $\nu_{\text{rlx}} \geq \nu$. Let $\boldsymbol{X}^\star = \boldsymbol{V}^\top \boldsymbol{V}$ be a solution to (35) where we denote $\boldsymbol{v}_i$ to be the ith column of $\boldsymbol{V}$. Note that the constraint $\boldsymbol{X}_{ii} = 1$ implies that $\boldsymbol{v}_i^\top \boldsymbol{v}_i = 1$. According to Goemans-Williamson random hyperplane rounding method, for $\boldsymbol{r} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, the point $\boldsymbol{x}_f$ where

$$\boldsymbol{x}_f = \begin{bmatrix} \operatorname{sign}(\boldsymbol{v}_1^\top \boldsymbol{r}) & \cdots & \operatorname{sign}(\boldsymbol{v}_n^\top \boldsymbol{r}) \end{bmatrix}^\top$$

is a feasible point of (19). The expected value ν_{GW} from the rounding algorithm is

$$\nu_{\text{GW}} = \mathbb{E}\left[\mathbf{tr}(\boldsymbol{W}\boldsymbol{x}_f\boldsymbol{x}_f^\top)\right] = \mathbf{tr}(\boldsymbol{W}\boldsymbol{X}_f)$$

where $(\boldsymbol{X}_f)_{ij} = \mathbb{E}\left[\operatorname{sign}(\boldsymbol{v}_i^\top \boldsymbol{r})\operatorname{sign}(\boldsymbol{v}_j^\top \boldsymbol{r})\right]$.

Proposition 8. *The expected value ν_{GW} has a lower bound that depends on $\boldsymbol{W}$. Specifically, we have*

$$\nu_{\text{GW}} \geq \alpha \nu_{\text{rlx}} - (1-\alpha) \sum_{i=1}^{n} \sum_{j=1}^{n} |\boldsymbol{W}_{ij}|$$

where

$$\alpha = \min_{x \in (-1,1)} \frac{1 - \frac{\arccos(x)}{\pi}}{\frac{1+x}{2}} = \min_{x \in (-1,1)} \frac{\frac{\arccos(x)}{\pi}}{\frac{1-x}{2}} \approx 0.87856.$$

Proof. Observe that we have

$$\begin{aligned} (\boldsymbol{X}_f)_{ij} = \mathbb{E}\left[\operatorname{sign}(\boldsymbol{v}_i^\top \boldsymbol{r})\operatorname{sign}(\boldsymbol{v}_j^\top \boldsymbol{r})\right] &= \Pr\left(\operatorname{sign}(\boldsymbol{v}_i^\top \boldsymbol{r}) = \operatorname{sign}(\boldsymbol{v}_j^\top \boldsymbol{r})\right) + (-1)\Pr\left(\operatorname{sign}(\boldsymbol{v}_i^\top \boldsymbol{r}) \neq \operatorname{sign}(\boldsymbol{v}_j^\top \boldsymbol{r})\right) \\ &= 1 - 2\Pr\left(\operatorname{sign}(\boldsymbol{v}_i^\top \boldsymbol{r}) \neq \operatorname{sign}(\boldsymbol{v}_j^\top \boldsymbol{r})\right) \\ &= 1 - \frac{2}{\pi}\arccos\left(\boldsymbol{v}_i^\top \boldsymbol{v}_j\right) \end{aligned}$$

where $\theta_{ij} = \arccos\left(\boldsymbol{v}_i^\top \boldsymbol{v}_j\right)$. The last equality holds since each $\boldsymbol{v}_i$ is a unit vector. Therefore, we have

$$\nu_{\mathrm{GW}} = \mathbf{tr}(\boldsymbol{W}\boldsymbol{X}_f) = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{W}_{ij}\left(1 - \frac{2}{\pi}\arccos\left(\boldsymbol{v}_i^\top \boldsymbol{v}_j\right)\right).$$

On the other hand, the optimal value to (35) is

$$\nu_{\mathrm{rlx}} = \mathbf{tr}(\boldsymbol{W}\boldsymbol{X}^\star) = \mathbf{tr}(\boldsymbol{W}\boldsymbol{V}^\top\boldsymbol{V}) = \mathbf{tr}(\boldsymbol{V}\boldsymbol{W}\boldsymbol{V}^\top) = \mathbf{tr}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{W}_{ij}\boldsymbol{v}_i\boldsymbol{v}_j^\top\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{W}_{ij}\boldsymbol{v}_i^\top\boldsymbol{v}_j \tag{36}$$

Observe the two inequalities that is valid for all $t \in (-1, 1)$:

$$1 - \frac{\arccos(t)}{\pi} \geq \alpha\frac{1+t}{2}, \quad \frac{\arccos(t)}{\pi} \geq \alpha\frac{1-t}{2}$$

where $\alpha = 0.87856....$ From each inequality, we can deduce the two inequalities:

$$\begin{aligned} 1 - \frac{2}{\pi}\arccos(t) &\geq \alpha(1+t) - 1 = \alpha t + (\alpha - 1) \\ \frac{2}{\pi}\arccos(t) - 1 &\geq \alpha(1-t) - 1 = -\alpha t + (\alpha - 1). \end{aligned}$$

Combining two inequalities together, we have

$$s\left(1 - \frac{2}{\pi}\arccos(t)\right) \geq s\alpha t + (\alpha - 1)$$

for $s = \pm 1$. Therefore, letting $s_{ij} = \mathrm{sign}(\boldsymbol{W}_{ij})$, we have

$$\begin{aligned} \nu_{\mathrm{GW}} &= \sum_{i=1}^{n}\sum_{j=1}^{n}|\boldsymbol{W}_{ij}|s_{ij}\left(1 - \frac{2}{\pi}\arccos(\boldsymbol{v}_i^\top\boldsymbol{v}_j)\right) \\ &\geq \sum_{i=1}^{n}\sum_{j=1}^{n}\left(|\boldsymbol{W}_{ij}|s_{ij}\alpha(\boldsymbol{v}_i^\top\boldsymbol{v}_j) + (\alpha - 1)\right) \\ &= \alpha\nu_{\mathrm{rlx}} - (1-\alpha)\sum_{i=1}^{n}\sum_{j=1}^{n}|\boldsymbol{W}_{ij}| \end{aligned}$$

by (36). This completes the proof. □

D.2 Parameter Estimation

In Section 6.4, we discussed how the relative estimation error changes with different truncation lengths L under systems with spectral radii $\rho(\boldsymbol{A}) = 0.1$ and $\rho(\boldsymbol{A}) = 0.9$. In this section, we discuss the experimental result in detail.

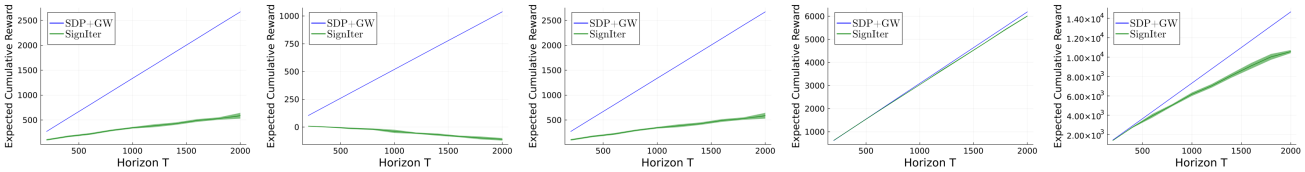
In the experiment, we consider the case when $n = 5$ and $p = 3$. The Schur-stable matrix $\boldsymbol{A}$ is generated with i.i.d. $\mathcal{N}(0, 1/n)$ entries and scaled to a desired spectral radius. The matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ are generated with i.i.d. $\mathcal{N}(0, 1/n)$ and $\mathcal{N}(0, 1/p)$, respectively. The noise processes are chosen to be Gaussian, that is, $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}_n, (0.05)^2\boldsymbol{I}_n\right)$ and $z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, (0.05)^2\right)$. We repeated each experiment 20 times with different noise seeds.

In Figures 2c and 2d, we can see that the relative error of Markov parameters tends to decrease as the exploration length H increases. It is a natural phenomenon which explains that more data samples give better estimation. When $\rho(\boldsymbol{A}) = 0.1$ (consider Figure 2c), we can also observe that larger truncation length L gives larger error. If we have larger L, we have more parameters (i.e., $\boldsymbol{CB}, \boldsymbol{CAB}, \ldots, \boldsymbol{CA}^{L-1}\boldsymbol{B}$) to estimate, but the spectral radius $\rho(\boldsymbol{A})$ too small to capture longer memory of the system, giving larger estimation error. On the other hand, when $\rho(\boldsymbol{A}) = 0.9$ (consider Figure 2d), we can see that it tends to have smaller errors when the truncation length L is

larger. It is because the spectral radius $\rho(\boldsymbol{A})$ is large enough to capture memory of the system and give help to estimate more number of Markov parameters.

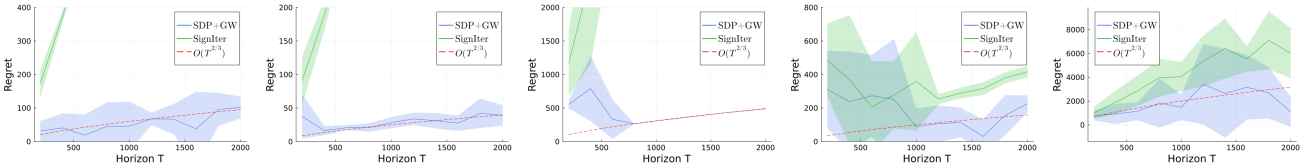
Moreover, we can observe double-descent phenomenon (Nakkiran et al., 2020) in both plots, which happens when a model whose number of parameters is about the same as the number of data used to estimate the model. In our problem, the regression model has p^2L number of parameters, and the number of samples used is $H - L$, implying that the peak would be shown at $H = L + p^2L$. We can see the peaks at $H = 300$ and $H = 500$ for $L = 30$ and $L = 50$, respectively, in Figures 2c and 2d. For $L = 6, 12, 18$, the peak will be obtained at $H = 60, 120, 180$, but it is not shown in the plot because we only plotted the curves at $H \in \{100, 150, \ldots, 2000\}$.

D.3 Additional Experiments



(a) $\rho(\boldsymbol{A}) = 0.1$ (b) $\rho(\boldsymbol{A}) = 0.3$ (c) $\rho(\boldsymbol{A}) = 0.5$ (d) $\rho(\boldsymbol{A}) = 0.7$ (e) $\rho(\boldsymbol{A}) = 0.9$

Figure 4: Expected cumulative reward under the oracle benchmark, approximated by semidefinite relaxation and Goemans-Williamson rounding (`SDP+GW`) and by the sign-iteration method (`SignIter`), in different spectral radii $\rho(\boldsymbol{A})$.



(a) $\rho(\boldsymbol{A}) = 0.1$ (b) $\rho(\boldsymbol{A}) = 0.3$ (c) $\rho(\boldsymbol{A}) = 0.5$ (d) $\rho(\boldsymbol{A}) = 0.7$ (e) $\rho(\boldsymbol{A}) = 0.9$

Figure 5: The regret curves of the explore-then-commit algorithm measured from `SDP+GW` and `SignIter` against the `SDP+GW` oracle benchmark, compared with the theoretical $\tilde{\mathcal{O}}(T^{2/3})$ rate.

In Section 6.3, we used a simple latent dynamics and reward function. Especially, we set the matrix $\boldsymbol{A}$ diagonal with small spectral radius (which is $\rho(\boldsymbol{A}) = 0.3$). This section is devoted to discuss whether our practical approaches, `SDP+GW` and `SignIter`, would work for dense matrices.

In order to generate dense matrices, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are generated in the same way as we did in the experiment for parameter estimation described in Section D.2 except that we use $n = 3$ and $p = 2$ in this experiment. We set the noise processes to be Gaussian, that is, $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}_n, (0.01)^2\boldsymbol{I}_n)$ and $z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (0, 01)^2)$. We conducted our experiments for $\rho(\boldsymbol{A}) \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ by scaling the same matrix $\boldsymbol{A}$, and we performed a grid search to determine constants c_1, c_2 in $H = c_1T^{2/3}$ and $L = c_2 \log T$, calibrated at $T = 1500$. The grid was chosen to be $c_1 \in \left\{T_0^{\epsilon} \colon T_0 = 1500,\ \epsilon \in \{0, \pm 0.02, \pm 0.05, \pm 0.1\}\right\}$ and $c_2 \in \{0.75, 1.0, 1.25\}$. For all cases of $\rho(\boldsymbol{A})$, an optimal value of c_1 is $1500^{-0.1} \approx 0.48$. On the other hand, optimal values of c_2 for $\rho(\boldsymbol{A}) = 0.1, 0.3, 0.5, 0.7$, and 0.9 are 0.75, 0.75, 1.0, 1.25, and 1.25, respectively, which is a consistent result with the experiment in parameter estimation. Each experiment was repeated 10 times with different noise seeds, except for the grid search.

Recall that a feasible solution to our optimization problems (7) and (11) is obtained by random rounding methods discussed in Section 6.1 and Section 6.2. Since the matrices are more complicated than the ones in Section 6.3, we might expect that rounding methods need more trials to get feasible points that are close to optimal. Hence we increased the number of rounding trials for both Goemans-Williamson random hyperplane rounding method and sign-iteration method as $\rho(\boldsymbol{A})$ increases. Moreover, we also increased the maximum number of iterations for sign-iteration method. The number of rounding trials is chosen to be 256, 320, 384, 448, and 512 for $\rho(\boldsymbol{A}) = 0.1, 0.3, 0.5, 0.7$, and 0.9, respectively, and the maximum number of iterations for the sign-iteration method is chosen to be 200, 250, 300, 350, and 400, respectively.

In Figure 4, we can see that `SDP+GW` always gives the larger value of the expected cumulative reward. Therefore, we choose `SDP+GW` as our reward benchmark. In Figure 5, the regret curves are shown which are computed from two practical approaches against the `SDP+GW` oracle benchmark with different spectral radii. We can see that it tends to accrue more rewards when $\rho(\boldsymbol{A})$ is closer to 1 because the states decay slower. In addition, the numerical experiment verifies that, even though the regret curve is more noisy than the simple case in Section 6.3, the curves are close to our theoretical sublinear $\tilde{\mathcal{O}}(T^{2/3})$ rates.