Episodic Linear Quadratic Regulators with Low-rank Transitions

Tianyu Wang* Lin F. Yang†

Abstract

Linear Quadratic Regulators (LQR) achieve enormous successful real-world applications. Very recently, people have been focusing on efficient learning algorithms for LQRs when their dynamics are unknown. Existing results effectively learn to control the unknown system using number of episodes depending polynomially on the system parameters, including the ambient dimension of the states. These traditional approaches, however, become inefficient in common scenarios, e.g., when the states are high-resolution images. In this paper, we propose an algorithm that utilizes the intrinsic system low-rank structure for efficient learning. For problems of rank-m, our algorithm achieves a K-episode regret bound of order $\widetilde{O}(m^{3/2}K^{1/2})$. Consequently, the sample complexity of our algorithm only depends on the rank, m, rather than the ambient dimension, d, which can be orders-of-magnitude larger.

1 Introduction

The classic problems of control and reinforcement learning have again captured considerable interests, following the recent successes in challenging domains like video games (Mnih et al., 2015) and GO (Silver et al., 2016). A classic and charming control model is the Linear Quadratic Regulator (LQR) model. In an LQR problem, the system state transitions are linear, and the cost to be minimized is quadratic in states and control actions. Simply formulated, LQR models successfully solve many challenging and important real-world problems, e.g., autonomous aerial vehicle (AAV) control (Abbeel et al., 2007), robotic arms (Li and Todorov, 2004; Platt Jr et al., 2010), and humandroid control (Mason et al., 2014).

In an LQR problem, the transition of states, $x \in \mathbb{R}^d$, depends linearly on the current state, the control action, $u \in \mathbb{R}^{d_u}$ played by the agent, and a noise $w \in \mathbb{R}^d$. Symbolically, the system evolves as $x \leftarrow Ax + Bu + w$, where A and B are matrices that describe the system dynamics. Every time the agent executes an action, an immediate quadratic cost (over the state and actions) is incurred, and the system transits to the next state. The goal of the agent is to find a policy that minimizes the expected total cost (in a period of time starting from any state).

If the dynamics $M := [A \;\; B]$ are unknown, the optimal policy is usually not directly attainable. Suppose we allow an agent to interact with the system for a certain amount of time. One common measure of the agent's performance is *regret*: the difference between the total cost that would be incurred by the unknown optimal policy and that by the agent. A good agent would achieve a regret upper bounded by a sub-linear function of the amount of time she is allowed to play. In this case, the *average regret per unit time* tends to zero, and hence the agent's performance becomes closer to an unknown optimal policy when she is allowed to interact with system longer. The amount of time that an agent takes to obtain a small (e.g. constant) average regret is called the *sample complexity of learning*.

There are extensive studies on learning to control an unknown LQR system. As we will discuss in related works, prior works (Abbasi-Yadkori and Szepesvári, 2011; Ibrahimi et al., 2012; Ouyang et al., 2017) achieve (at best) $\widetilde{\mathcal{O}}\left(\text{poly}(d)\sqrt{T}\right)$ regret rate, despite possibly stronger assumptions. The algorithms proposed in these papers all achieve sublinear regret bound. However, their sample complexities are at least proportional to the total number of parameters in the model $M = [A \;\; B]$ (i.e., polynomial in d). This complexity, however, can be big in practice. For instance, in video games, the states – video frames – have a huge number of pixels. Problems as classic as the `Mountain-Car` and `Cart-Pole` (Brockman et al., 2016a) can also suffer from this problem when learned with images – the images representing the states have thousands of pixels. Nevertheless, we notice that **(1)** in all these examples, the intrinsic dimension (i.e., the internal state space is about 3-6 dimensions) is actually low, and **(2)** Suh and Tedrake (2020) recently shows linear models with visual feedbacks achieve surprisingly good control. We therefore propose to use LQR for low-rank problems and ask the following question.
Is there an online algorithm that learns to control an unknown LQR system with number of samples only depends on the intrinsic model complexity?

*tianyu@cs.duke.edu, Duke University.
†linyang@ee.ucla.edu, UCLA.

To make this precise, we consider a basic low intrinsic complexity setting: the states have an underlying *low-rank representation* (please refer to Definition 1 for details). Formally, we consider the episodic online LQR control problem. In each episode of such problems, the agent starts from a random/adversarial initial state, and execute H control actions to finish. After H steps, the agent starts over from another initial state, and the next episode begins. Our goal is to minimize the number of episodes for the agent to achieve a constant average regret (per episode). In this paper, we answer the above question by proposing an algorithm that obtains a K-eposide regret bound of order

$$\widetilde{\mathcal{O}}\left(m^{3/2}\sqrt{K}\right)^{1},$$

where m bounds the rank of the matrix $M = [A \quad B]$. Our algorithm corresponds to a sample complexity of poly(m). This order can be significantly smaller than previous results, which depend polynomially on the ambient dimension d.

Our result is a technically involved combination of the Optimism in the Face of Uncertainty (OFU) principle (Dani et al., 2008; Abbasi-Yadkori and Szepesvári, 2011; Lale et al., 2019; Kveton et al., 2017), and low-rank approximations, e.g., principle component analysis (PCA) (Jolliffe, 1986; Vaswani and Narayanamurthy, 2017). On a high level, our algorithm can be viewed as a model-based algorithm for closed-loop control. We estimate the system dynamics using least-squares combined with a PCA projection. In each episode, with the learned model and its uncertainty estimation, we compute an optimistic control policy that carefully balances exploration and exploitation. We then execute the control policy to obtain a new episode of data and at the same time incur provably small regret.

In our analysis, the core technical difficulty comes from the interleaving of PCA and LQR transitions: for each episode, the new data points can potentially lie outside the subspace identified by the PCA projection of previous episodes. Therefore a plain adoption of previous results fail to give a bound that is independent of d. In order to handle this issue, we project all data points to the subspace identified by the PCA at the very last episode, and carefully handle the difference between the PCA subspaces across episodes. By modeling the process as a rank-deficient self-normalized random process, we show that our algorithm provably achieves a sublinear regret that only depends on the internal rank of the system.

Related works. The history of control theory can date back to the study of governors by Maxwell (Maxwell, 1868), where he linearized differential equations of motion. This work, together with the classic Riccati equation (Riccati, 1720; Bittanti, 1996), builds the root foundation of modern LQR. Similar to many control problems, LQR problems can be classified into open-loop problems (Ljung and Söderström, 1983; Helmicki et al., 1991; Chen and Nett, 1993; Box et al., 2015; Hardt et al., 2018; Tu et al., 2017; Dean et al., 2017) versus closed-loop problems. Compared to open-loop problems, closed-loop problems are closer to a reinforcement-learning setting – feedbacks of the environment are used in an interactive fashion. In this paper, we focus our attention to the closed-loop LQR problem, and use LQR problems to refer to closed-loop LQR problems from now on.

In recent years, as motivated by an increasing amount of real-world data-driven applications, more interests are attracted to the learning to control problems – control problems where the system dynamics are unknown. The learning to control problem is also known as system identification (for observable systems) in classic terms (Kalman, 1960), and learning based model predictive control methods have been developed by the control community (e.g., Aswani et al., 2013; Koller et al., 2018).

Among works on learning to control for LQR problems, some sit in a "bandit" setting, i.e. one observation right after one action, and no rollouts are allowed. Abbasi-Yadkori and Szepesvári (2011) use the optimistic principle, and obtained a regret bound of order $\widetilde{\mathcal{O}}\left(f(d)\sqrt{T}\right)$, where $f(d)$ could be exponential in the ambient dimension d. Ibrahimi et al. (2012) makes a sparsity assumption on the system dynamics $M = [A \quad B]$ and achieves a regret bound of order $\widetilde{\mathcal{O}}\left(d\sqrt{T}\right)$. Yet the dependence on the ambient dimension d is not removed. Simchowitz and Foster (2020) proposes to use ϵ-greedy exploration and achieves optimal rate in terms of the ambient dimension. For observable systems, online control with system identification for Linear Quadratic Gaussian models have also been studied (Lale et al., 2020), and a regret depends polynomially on the ambient dimension is derived. Assuming a correct specification of the prior distribution, Bayesian methods have also been applied to learning to control LQRs (Abeille and Lazaric, 2017; Ouyang et al., 2017; Abeille and Lazaric, 2018). However, all the above mentioned algorithms admit regret (at best) of order $\widetilde{\mathcal{O}}\left(poly(d)\sqrt{T}\right)$.

LQR problems has also been studied in a "non-bandit" setting, i.e. rollouts of trajectories are permitted. In this setting, efforts on extracting the intrinsic dimension have been made. The concepts of "Bellman rank" (Jiang et al., 2017) and "witness rank" (Sun et al., 2019) are proposed as dimension measures in this "non-bandit" setting. These methods, however, are not as sample efficient as ours due to the need of rollouts.

[1] $\tilde{\mathcal{O}}$ omits factors that are *poly-log* in the inputs, as well as factors depending on H. A formal statement can be found in Theorem 1.

2 Preliminaries and Notations

In an episodic LQR problem, the agent starts the k-th episode from a random initial state sampled from μ: $x_{k,1} \sim \rho$. The agent then executes $H-1$ controls to finish this episode. Episode k ends at $h = H$, and the agent starts over from $h = 1$and the $(k+1)$-th episode starts, where a new initial state $x_{k+1,1}$ is sampled from ρ. At each step (k, h), the next state of the system depends linearly on the current observed state $\hat{x}_{k,h} \in \mathcal{X} \subseteq \mathbb{R}^d$ and the action taken $u_{k,h} \in \mathcal{X} \subseteq \mathbb{R}^{d_u}$ plus a noise term. In other words, there are matrices $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times d_u}$, such that the next state can be described as

$$\hat{x}_{k,h+1} = A\hat{x}_{k,h} + Bu_{k,h} + w_{k,h},$$

where $w_{k,h} \in \mathbb{R}^d$ is a mean-zero noise, $\hat{x}_{k,h+1}$ is the observed state. We write $M = [A, B]$ and $\hat{z}_{k,h} = [\hat{x}_{k,h}^\top, u_{k,h}^\top]^\top$. The transition can then be rewritten as $\hat{x}_{k,h+1} = M\hat{z}_{k,h} + w_{k,h}$.

At each time (k, h), the system transits from $\hat{x}_{k,h}$ to $\hat{x}_{k,h+1}$ and receives an immediate cost

$$c_{k,h} = \hat{x}_{k,h}^\top Q_h \hat{x}_{k,h} + u_{k,h}^\top R_h u_{k,h}, \tag{1}$$

where Q_h $(h \in [H])$ and R_h $(h \in [H-1])$ are known positive definite matrices, and $R_H = 0$.

The goal of the agent is to learn a policy $\pi : [H] \times \mathbb{R}^d \to \mathbb{R}^{d'}$, such that the following objective is minimized for all $h \in [H]$

$$J_{k,h}^\pi(M, x) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} c_{k,h} \middle| \hat{x}_{k,h} = x \right], \tag{2}$$

where $c_{k,h}$ is the immediate cost per step and $\mathbb{E}_\pi$ is over the random trajectory generated by policy π starting from x at (k, h). When it is clear from context, we omit k from $J_{k,h}^\pi$ and write $J_{k,h}^\pi$ as J_h^π.

From Bellman optimality (Bellman et al., 1954; Bertsekas, 2004), for a system with dynamics $M = [A \quad B]$, the optimal policy π^* is given by (e.g., Bertsekas, 2004): $\forall x \in \mathcal{X}, h \in [H-1]$,

$$\begin{aligned} \pi_h^*(x) &:= \mathcal{K}_h(M)x, \quad \text{where} \\ \mathcal{K}_h(M) &:= -\left(R_h + B^\top \Psi_{h+1}(M)B\right)^{-1} B^\top \Psi_{h+1}(M)A, \end{aligned} \tag{3}$$

and $\Psi_h(M)$ is defined by the Riccati iteration:

$$\begin{aligned} &\Psi_H(M) := Q_H, \\ &\Psi_{h-1}(M) := Q_h + A^\top \Psi_h(M)A \\ &\quad - A^\top \Psi_h(M)B\left(R_h + B^\top \Psi_h(M)B\right)^{-1} B^\top \Psi_h(M)A, \end{aligned} \tag{4}$$

for $h < H$. When it is clear from context, we simply write Ψ_h for $\Psi_h(M)$. As we will later discuss in Section 3.2, the matrix $K_h(M)$ is always well-defined, since the matrix $R_h + B^\top \Psi_{h+1} B$ is invertible for arbitrary $M = [A, B]$ (e.g., Bertsekas (2004)). More details regarding well-definedness of the control are in Appendix C. This property allows us to estimate the the dynamics M and design a control policy based on the estimation. For a system with dynamics $M = [A, B]$, the cost under an optimal policy can be written as follows (e.g., p. 229, Chapter 5, Vol. I, Bertsekas (2004)),

$$\begin{aligned} &J_h^*(M, x) = x^\top \left[\Psi_{h+1}(M)\right] x + \psi_h, \\ &\psi_h := \psi_{h+1} + \mathbb{E}_{w_{h+1}}\left[w_{h+1}^\top \left[\Psi_{h+1}(M)\right] w_{h+1} \right],\ \psi_H := 0. \end{aligned} \tag{5}$$

Controllable Subspace. In control theory, it is often assumed that the linear quadratic system is controllable, i.e., the matrix

$$\begin{bmatrix} B & AB & A^2B \cdots A^{d-1}B \end{bmatrix}$$

is of full row rank. When the above condition is satisfied, we say that $[A \quad B]$ is a pair of *controllable matrices*. Intuitively, a pair of controllable matrices allows the agent's action (control) to influence all dimensions of the system. In our problem, we assume a low-rank version of controllability.

Definition 1. *We say that a pair of matrices $[A \quad B]$ is rank-m controllable if there exists a matrix L of m orthonormal columns, such that such that $A = LL^\top ALL^\top$, $B = LL^\top B$ and $\begin{bmatrix} L^\top AL & L^\top B \end{bmatrix}$ is a pair of controllable matrices. The projection matrix $P = LL^\top$ is called the true **projection matrix for system** $M = [A \quad B]$.*

Intuitively, a pair of rank-m matrices defines a transition dynamic such that, if there were no noise, the states would always lie on a rank-m subspace, and when restricted to this rank-m subspace, the system is controllable. When m equals the dimension of the state space, rank-m controllability is equivalent to controllability. Throughout the rest of the paper, we assume that the system we are considering is rank-m controllable for a fixed $m \leq d$. We focus on settings where $m \ll d$ and $m = \Theta(d_u)$, where d_u is the dimension of the action (control) space.

In a finite-horizon discrete-time setting, for positive definite R_h and Q_h, the optimal control law can be solved via dynamic programming. (See e.g., p. 150, Chapter 4, Vol. I; p. 229, Chapter 5, Vol. I, Bertsekas, 2004). We assume rank-m controllability so that the controls can influence the entire subspace on which the noiseless states lie. Problems such as the connection between rank-m controllability and convergence of the Riccati iteration might be an interesting future direction.

Performance Measure. We use regret to measure the performance of the algorithm. For LQR problems in this paper, if the true system dynamics are $M_* = [A_* \quad B_*]$, the regret of the first K episodes is defined as:

$$Reg(K) = \sum_{k=1}^{K} J_1^{\pi_k}(M_*, \hat{x}_{k,1}) - J_1^*(M_*, \hat{x}_{k,1}), \tag{6}$$

where $\hat{x}_{k,1}$ is the starting state for episode k, $J_1^{\pi_k}$ is computed from (2), and $J_1^*(M_*, \hat{x}_{k,1})$ can be calculated from (5). As discussed above, $J_1^*(M_*, x_{k,1})$ is the (expected) cost of an optimal policy for episode k, and $J_1^{\pi_k}(M_*, x_{k,1})$ is the (expected) cost of the policy executed during episode k. As common in bandit and online learning setting, we want a sub-linearly growing regret. This ensures that the strategy incurs optimal costs given enough time.

3 Learning with Low-rank Structure

In learning settings, the system dynamics are unknown, and the task is to learn a good policy as we interact with the environment. We will apply the optimism in the face of uncertainty (OFU) principle to learn a good policy. In particular, we maintain an estimation of the system dynamics as well as its uncertainty. To control the system, we will be "optimistic" in the uncertainty ball of our model estimator, i.e., we will use the best possible (in terms of cost) model that satisfies our uncertainty estimation to solve for the next policy. In order to obtain such "optimistic" estimations, we (i) apply a PCA projection to the observed data and use a least-square regression to fit the system dynamics; and (ii) we search a "confidence region" close to this estimated system dynamics. This confidence region uses the uncertainty in both the regression and the PCA projection.

Throughout our analysis, we use the following common stability assumption.

Assumption 1 (Stability). *Let $M_* = [A_* \quad B_*]$ be the true system dynamics. We assume that* [2]
(1) *For all $h \in [H]$, $\|A_* + B_* \mathcal{K}_h(M_*)\|_2 \leq r$ for some $r < 1$ and $\|\mathcal{K}_h(M_*)\|_2 \leq C$ for some constant C, where $\mathcal{K}_h(M_*)$ is defined in (3). We assume that $\|M_*\| \leq C$ for some constant C.*
(2) *The control mapping $\mathcal{K}_h$ is Lipschitz near M_*: there exist constants C and D such that $\|\mathcal{K}_h(M_*) - \mathcal{K}_h(M)\|_2 \leq D\|M_* - M\|_2$ for all M such that $\|M_* - M\|_2 \leq C$.*
(3) *The noises $w_{k,h}$ satisfy: $\forall k \geq 1, h \in [H]$, $\forall h \in [1, H]$, $\|w_{k,h}\|_2 \leq C_w < 1$ and that $2C_w + r \leq 1$.*
(4) *The initial states for each episode is bounded: for any $k \geq 1$, $\|\hat{x}_{k,1}\|_2 \leq 1$.*

In Assumption 1, item (1) in is a type of stability assumption. Stability is standard and usually assumed for control problems (Ibrahimi et al., 2012; Matni et al., 2017; Dean et al., 2019). Item (2) is actually naturally true, since all mappings are continuous, and thus Lipschitz continuous on a compact region. Items (3) and (4) are assumed for simplicity, since we can always rescale the space so that these two are satisfied.

We also assume rank-m controllability of the problem.

Assumption 2. *We assume that the system dynamics matrices $[A_* \quad B_*]$ are rank-m controllable (Definition 1). We use P_* to denote the true projection matrix for this system (Definition 1).*

The above low-rank assumption distinguishes our problem from a general LQR problem. Utilizing the low-rankness can improve the regret dependence on dimension significantly. We also make a noise assumption.

Assumption 3 (Noise). *We assume that at any $h \in [H]$ and $k \geq 1$, the noise $w_{k,h}$ is (1) independent of all other randomness, (2) $\mathbb{E}[w_{k,h}] = 0$ for any $h \in [H]$, and $k \geq 1$. (3*) Without loss of generality, there exists constant σ^2, such that $\mathbb{E}[w_{k,h} w_{k,h}^\top] = \sigma^2 I_d$ for all k and h.*

[2] For simplicity, we use C to denote most boundedness constants. The only exception is the noise bound C_w, and consequently the constant $C_{\max}$ as per defined in Algorithm 1. This does not lose any generality.

Items (3*) can be relaxed by combining Remark 3 in (Abbasi-Yadkori and Szepesvári, 2011) and PCA analysis with general noises (Vaswani and Narayanamurthy, 2017). We focus on this standard noise setting for a cleaner presentation, while our results generalize to different noise settings.

For representation simplicity, we introduce the following assumption.

Assumption 4 (Initial Distribution)**.** *We assume there exists $\lambda_- > 0$, such that the unseen starting state $x_{k,1}$ satisfies*

$$\lambda_m\left(\mathbb{E}_\rho\left[x_{k,1}\left(x_{k,1}\right)^\top\right]\right) \geq \lambda_-,$$

where $\mathbb{E}_\rho$ is the expectation with respect to the initial distribution ρ, and $\lambda_m(\cdot)$ returns the m-th eigenvalue of a matrix.

Note that this is a mild assumption on certain exploratory property of the initial distribution. If we do not have such an initial distribution, we can use a fraction of steps to maximize the top-m eigenvalues. Maximizing the top m eigenvalues can be done by simply keep playing a same control u (because of low-rank controllability). Our analysis techniques still carry through.

Before formulating our algorithm, we define the following notations.

- For state vector $\hat{x}_{k,h}$ and controls vector $u_{k,h}$ at (k,h), we write $\hat{z}_{k,h} := \begin{bmatrix}\hat{x}_{k,h}\\ u_{k,h}\end{bmatrix}$.
- For any $k \geq 1$, we define for following matrices

$$\hat{X}_k := \left[\hat{x}_{h',k'}\right]_{h'\in[H-1],k'\in[k-1]}, \tag{7}$$

$$\hat{X}_k^{\text{next}} := \left[\hat{x}_{h',k'}\right]_{h'\in[2,H],k'\in[k-1]}, \tag{8}$$

$$U_k := \left[u_{h',k'}\right]_{h'\in[H-1],k'\in[k-1]}, \tag{9}$$

$$W_k := \left[w_{h',k'}\right]_{h'\in[H-1],k'\in[k-1]}, \tag{10}$$

$$\hat{Z}_k := \begin{bmatrix}\hat{X}_k^\top & U_k^\top\end{bmatrix}^\top \tag{11}$$

In the above, $\hat{X}_k$ is the collection of observed states (h runs from 1 to $H-1$): Each column in $\hat{X}_k$ is an observed state. Similarly, $\hat{X}_k^{\text{next}}$ is the collection of observed "next" states (h runs from 2 to H). U_k, W_k and $\hat{Z}_k$ are controls, noises (not directly observable), and state-control pairs respectively. Using the above notations, we can write

$$\hat{X}_k^{\text{next}} = M_*\hat{Z}_k + W_k, \quad \text{or}$$
$$\hat{X}_k^{\text{next}} = A_*\hat{X}_k + B_*U_k + W_k.$$

With the above notations introduced, we can proceed to design our control rules.

3.1 Online Control Design

With the above notations (Eq. 8, 11) and true dynamics matrix M_*, we have $\hat{X}_k^{\text{next}} = M_*\hat{Z}_k + W_k$. Thus to approximate M_*, we can consider finding a matrix M for the following problem

$$\min_M \left\|M\hat{Z}_k - \hat{X}_k^{\text{next}}\right\|_F^2 + \frac{1}{2}\|M\|_F^2. \tag{12}$$

Since the solution to this matrix ridge regression problem is $\left(\hat{Z}_k\hat{Z}_k^\top + I_{d+d_u}\right)^{-1}\hat{Z}_k\left(X_k^{\text{next}}\right)^\top$, we combine a PCA with this solution, and design a rank-m estimate. To compute this PCA, we apply a singular value decomposition to $\hat{X}_{k-1}$. Let L_k be the matrix of left singular vectors of $\hat{X}_{k-1}$. Let $\bar{L}_k$ be the columns of L_k that corresponds to the top m singular values. The learned projections at episode k are

$$P_k := \bar{L}_k\bar{L}_k^\top \quad \text{and} \quad P_k^{\text{aug}} := \begin{bmatrix}P_k & 0_{d\times d_u}\\ 0_{d_u\times d} & I_{d_u}\end{bmatrix}. \tag{13}$$

The projection P_k is for states $\hat{x}_{k,h}$, and P_k^{aug} is multiplied to state-action pairs $\hat{z}_{k,h}$. More specifically, P_k applies to $\hat{x}_{k,h}$ and projects the states to a rank-m subspace. P_k^{aug} applies to $\hat{z}_{k,h} = \begin{bmatrix}\hat{x}_{k,h}\\ u_{k,h}\end{bmatrix}$, projects the states $\hat{x}_{k,h}$ to

a rank-m subspace and preserves the controls $u_{k,h}$. With the learned projections P_k and P_k^{aug}, we compute the following quantities:

$$\widetilde{V}_k := \left(\hat{Z}_k \hat{Z}_k^\top + I_{d+d_u}\right), \quad V_k := P_k^{\text{aug}} \widetilde{V}_k P_k^{\text{aug}},$$
$$\bar{Z}_k := P_k^{\text{aug}} \hat{Z}_k, \quad \bar{X}_k^{\text{next}} := P_k \hat{X}_k^{\text{next}}. \tag{14}$$

With the above quantities, we can estimate the system dynamics M_* by

$$M_k^\top = V_k^\dagger \bar{Z}_k \left(\bar{X}_k^{\text{next}}\right)^\top, \tag{15}$$

where † is pseudo-inverse operator. Intuitively, (15) is a low-rank approximation to the solution of the problem in (12).

Now we define confidence region around M_k. Fix a parameter δ that controls the probability that the regret behaves nicely. After a warmup period $K_{\min}$, the confidence region $\mathcal{C}^{(k)}$ ($k > K_{\min}$) is defined as $\mathcal{C}^{(k)} := \mathcal{C}_* \cap \mathcal{C}_1^{(k)} \cap \mathcal{C}_2^{(k)}$, where $\mathcal{C}_*$ is a fixed closed set that always contains M_*, and

$$\mathcal{C}_1^{(k)} := \left\{M : \|(M - M_k)(I_{d+d_u} - P_k^{\text{aug}})\|_2 \lesssim G_{k,\delta}\right\},$$
$$\text{with } G_{k,\delta} = \widetilde{\Theta}\left(\frac{1}{\sqrt{k}}\right), \tag{16}$$
$$\mathcal{C}_2^{(k)} := \left\{M : \left\|(M - M_k)V_k^{1/2}\right\|_2^2 \lesssim \beta_{k,\delta}\right\},$$
$$\text{with } \beta_{k,\delta} = \widetilde{\Theta}(1). \tag{17}$$

For practitioners, the values of $G_{k,\delta}$ and $\beta_{k,\delta}$ can be rescaled by proper constants, and $\mathcal{C}_*$ can be properly chosen to regularize the norm of learned transitions. For theoretical purpose, we use

$$\mathcal{C}_* := \{M = [A \quad B] : \|A + B\mathcal{K}_h(M)\|_2 \le r \text{ and } \|M\| \le C\},$$

and use $G_{k,\delta}$ and $\beta_{k,\delta}$ as detailed in Appendix A. Intuitively, $\mathcal{C}_1^{(k)}$ defines a region (around M_k) that is perpendicular to the PCA projection, and $\mathcal{C}_2^{(k)}$ defines a region around M_k parallel to the PCA projection. As we will show later, with high probability, $M_* \in \mathcal{C}^{(k)} = \mathcal{C}_* \cap \mathcal{C}_1^{(k)} \cap \mathcal{C}_2^{(k)}$. We will search within $\mathcal{C}^{(k)}$ for an optimistic estimate $\widetilde{M}_k$. Specifically, within the confidence region $\mathcal{C}^{(k)}$, we find an optimistic estimation $\widetilde{M}_k \in \mathcal{C}^{(k)}$, such that

$$\widetilde{M}_k \in \arg\min_{M \in \mathcal{C}^{(k)}} J_1^*(M, \hat{x}_{k,1}), \tag{18}$$

where $J_1^*(M, \hat{x}_{k,1})$ is the optimal cost if the system transition were M, and can be computed using (5). With this estimation $\widetilde{M}_k$, we play a policy $\pi^{(k)} = \left\{\pi_h^{(k)}\right\}_{h\in[H]}$:

$$\pi_h^{(k)}(x) := \mathcal{K}_h(\widetilde{M}_k)x. \tag{19}$$

where $\mathcal{K}_h(\widetilde{M}_k)$ is defined in (3). Our strategy is summarized in Algorithm 1.

3.2 Well-definedness of the Control

Before analyzing the algorithm performance, we first need to show that the algorithm is well-defined, i.e., (19) is well-defined. To show this, we need

(i) $\widetilde{M}_k$ is well-defined (with high probability);

(ii) Given any $\widetilde{M}_k$, the matrix $\mathcal{K}_h(\widetilde{M}_k)$ (Eq. 3) is well-defined.

For item (i), in Proposition 1, we show that $\mathcal{C}^{(k)}$ is closed, bounded and non-empty (with high probability), which shows we can find $\widetilde{M}_k \in \arg\min_{M\in\mathcal{C}^{(k)}} J_1^*(M, \hat{x}_{k,1})$ with high probability.

Proposition 1. *The regions enclosed by $\mathcal{C}^{(k)}$ ($k \in (K_{\min}, K]$) are closed and bounded. Also, under event $\mathcal{E}_{K,\delta}$, $\mathcal{C}^{(k)}$ is non-empty.*

Algorithm 1 Low-rank LQR with OFU

Parameters: ***horizon*** $H > 0$, ***probability parameter*** δ, ***dimension*** d, ***true rank*** m, ***constant bounds*** C and $C_{\max} := 4C_w + 2\sqrt{2}C_w^2$, ***minimal eigenvalue*** λ_- (Assumption 4), ***total number of episodes*** K, ***warm-up period*** $K_{\min} = 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$.

▷ *In practice, the above parameters can be chosen more freely.*

Warmup: For the first $K_{\min}$ episodes, randomly play controls from a bounded set.

for $k = K_{\min}+1, K_{\min}+2, \cdots, K$ **do**

  Observe $\hat{x}_{k,1}$.

  Compute $\widetilde{M}_k := \arg\min_{M\in\mathcal{C}^{(k)}} J_1^*(M, \hat{x}_{k,1})$, where $J_1^*(M, \hat{x}_{k,1})$ is computed from (5).

  for $h = 1, 2, \cdots, H-1$ **do**

    Execute the control $u_{k,h} = K_h(\widetilde{M}_k)\hat{x}_{k,h}$, where $K_h(\widetilde{M}_k)$ is defined in (3).

    Observe the next state $\hat{x}_{k,h+1}$.

  end for

  Gather the observations into $\hat{Z}_k$, $\hat{X}_k$, and $\hat{X}_k^{\text{next}}$.

end for

Proposition 1 is essentially a boundedness/stability result. This ensures that the controls $u_{k,h}$ are of reasonable length. A proof of Proposition 1 can be found in Appendix C.

For item (ii), given any $M = [A \quad B]$, which defines a transition system, the matrix $K_h(M)$ (Eq. 3) is well-defined for all $h \in [H]$. Since $\mathcal{K}_h(M) := -\left(R_h + B^\top\Psi_{h+1}B\right)^{-1}B^\top\Psi_{h+1}A$, it is sufficient to show Ψ_h (defined in Eq. 4) is positive semi-definite. This is because positive semi-definiteness of Ψ_{h+1}, together with positive definiteness of R_h, ensures that $R_h + B^\top\Psi_{h+1}B$ is invertible.

The positive semi-definiteness of Ψ_h is below in Lemma 1. Its proof can be found in textbooks covering Linear Quadratic Regulators (e.g., Bertsekas (2004)). We provide a proof in Appendix C for completeness.

Lemma 1. *The matrix Ψ_h is positive semi-definite for any $h \in [H]$ and M, provided that Q_h, R_h are positive definite.*

4 Regret Analysis

The regret can be bounded as in Theorem 1.

Theorem 1. *Under Assumptions 1-4, for any $\delta > 0$, with probability at least $1-(4K+2)\delta$, the regret for the first K ($K > K_{\min}^2$, $K_{\min} := 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$) episodes satisfies*

$$\mathrm{Reg}(K) \le \mathcal{O}\left(\left(H^{5/2} + m^{3/2}H\right)\sqrt{K}\,\mathrm{polylog}\left(\frac{K}{\delta}\right)\right),$$

where $\mathcal{O}$ omits (poly)-logarithmic terms in H and d.

To bound the regret, we first show that the event $\mathcal{E}_{K,\delta} := \left\{M_* \in \mathcal{C}^{(k)}, \quad \forall\, k \in (K_{\min}, K]\right\}$ holds with high probability (Section 4.1). Then we bound the regret under event $\mathcal{E}_{K,\delta}$ (Section 4.2).

4.1 Part I: $M_* \in \mathcal{C}_1^{(k)} \cap \mathcal{C}_2^{(k)}$ for all $k = K_{\min}+1, K_{\min}+2, \cdots, K$ with high probability

Part Ia: $M_* \in \mathcal{C}_1^{(k)}$ with high probability. To show $M_* \in \mathcal{C}_1^{(k)}$ with high probability, we need to show that the projection error is small. In other words, we need to bound the term $\|P_* - P_k\|_2$. The tool we will use is Lemma 2, which extends previous results (Corollary 2.7 by Vaswani and Narayanamurthy (2017), Theorem 1 by Lale et al. (2019)) to our case.

Lemma 2. *Let P_* be the true projection matrix for the rank-m controllable system $M_* = [A_* \quad B_*]$. Let $C^{\max} := 4C_w + 2\sqrt{2}C_w^2$. Suppose Assumptions 1-4 hold. For $k > K_{\min} = 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$, with probability at least $1-3\delta$,*

$$\|P_* - P_k\|_2 \le G_{k,\delta}, \quad \textit{where } G_{k,\delta} := \widetilde{\Theta}\left(\frac{1}{\sqrt{k}}\right).$$

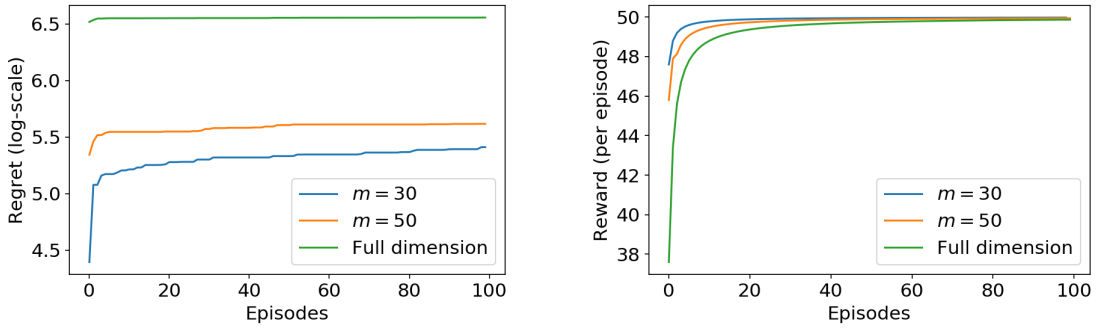

Figure 1: All solid line plots are averaged over 5 runs. In both subfigures, $H = 50$. The $m = 30$ and $m = 50$ plots are Algorithm 1 with $m = 30$ and $m = 50$ respectively. The "Full dimension" curves are OFU algorithms without our PCA projection, which represents previous algorithms (Abbasi-Yadkori and Szepesvári, 2011). *Left:* Cost regret (in log-scale) against episode. Cost regret is defined by our LQR formulation: larger cost corresponds to bad cart/pole positions and large velocities, and regret is larger if cost is larger. This shows that empirical results agree with our theory: regret is smaller when m is smaller. *Right:* Reward per episode against episode. The environment gives a unit reward if the cart/pole are in good positions. In terms of both performance metrics, our algorithm achieves better performance than previous methods that do not utilize low-rankness.

The proof of Lemma 2 uses the Davis-Kahan theorem on principle angle between column spans, as well as concentration results for matrix martingales. More details of this proof can be found in Appendix D.

Part Ib: $M_* \in \mathcal{C}_2^{(k)}$ with high probability. To show $M_* \in \mathcal{C}_2^{(k)}$ with high probability, we need to bound the quantity $\left\|(M_* - M_k)V_k^{1/2}\right\|_2^2$. By definition of V_k (in Eq. 14), the norm of V_k increases with k. At roughly the same rate, the residual $\|M_* - M_k\|_2$ decreases with k because of the learning nature. Using this observation, we can get that, with high probability, $\left\|(M_* - M_k)V_k^{1/2}\right\|_2^2 \leq \widetilde{\Theta}(1)$. The full proof requires a rank-deficit self-normalized process formulation, whose details are in Appendix E.

Combining Part Ia and Part Ib immediately gives Lemma 3.

Lemma 3. *With probability at least* $1 - 4K\delta$, *for* $K > K_{\min} = 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$, *the following event is true:*

$$\mathcal{E}_{K,\delta} := \left\{M_* \in \mathcal{C}^{(k)},\ \forall k = K_{\min}+1, \cdots, K\right\}. \tag{20}$$

Some more details on Lemma 3 can be found in Appendix F. In the next part, we will bound the regret under event $\mathcal{E}_{K,\delta}$.

4.2 Part II: Bound the Regret under $\mathcal{E}_{K,\delta}$

Under event $\mathcal{E}_{K,\delta}$, with the OFU principle, we can decompose the regret as in Proposition 2.

Proposition 2. *Let* $\widetilde{\Psi}_{k,h} := \Psi_h(\widetilde{M}_k)$ *computed by (4). Under event* $\mathcal{E}_{K,\delta}$ *(*$K > K_{\min}^2$*), we have*

$$\begin{aligned}\mathrm{Reg}(K) &\leq \mathcal{O}\left(H\sqrt{K}\right)\\ &+ \sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left(\Delta_{k,h} + \Delta'_{k,h} + \Delta''_{k,h}\right),\end{aligned} \tag{21}$$

where

- $\Delta_{k,h} := \mathbb{E}_{k,h}\left[J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right)\right] - J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right)$, *and* $\mathbb{E}_{k,h}$ *is the expectation conditioning on* $\mathcal{F}_{k,h}$ *– all randomness before time* (k,h).
- $\Delta'_{k,h} := \|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} - \mathbb{E}_{k,h}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}}\right]$,
- $\Delta''_{k,h} := \|M_*\hat{z}_{k,h}\|_{\widetilde{\Psi}_{k,h+1}} - \left\|\widetilde{M}_k\hat{z}_{k,h}\right\|_{\widetilde{\Psi}_{k,h+1}}$.

Proposition 2 is a computational result, and a detailed derivation is in Appendix G.

Next, in Lemmas 4 and 5 below, we bound the the right-hand-side of (21).

Lemma 4. *Under Assumptions 1-4, with probability at least* $1-2\delta$*, we have*

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H-1}\Delta_{k,h}\right| \leq \mathcal{O}\left(\sqrt{KH^3\log\frac{2}{\delta}}\right), \quad \textit{and}$$
$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H-1}\Delta'_{k,h}\right| \leq \mathcal{O}\left(\sqrt{HK\log\frac{2}{\delta}}\right).$$

We can use the Azuma's inequality to derive Lemma 4. The proof of Lemma 4 is in Appendix H. For the regret from $\Delta''_{k,h}$ terms, we use Lemma 5.

Lemma 5. *Let Assumptions 1-4 hold. Under event* $\mathcal{E}_{K,\delta}$ *(*$K > K_{\min}$*), we have*

$$\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\Delta''_{k,h} \leq \widetilde{\mathcal{O}}\left(\left(H^{5/2}+m^{3/2}H\right)\sqrt{K}\right). \tag{22}$$

The proof for Lemma 5 is longer. At a high level, this proof uses the following three observations. (i) M_* are M_k are both rank-m. (ii) The learned projections across episodes are not too far away. Specifically, $\sum_{k=\lceil K\rceil+1}^{K}\|P_*-P_k\|_2 \leq \widetilde{\mathcal{O}}\left(\sqrt{K}\right)$ and $\sum_{k=\lceil K\rceil+1}^{K}\|P_K-P_k\|_2 \leq \widetilde{\mathcal{O}}\left(\sqrt{K}\right)$. (iii) If everything is projected to the subspace identified by P_K, then this term can be carefully handled by extending previous results for the full rank case (Lemma 7 in (Yang and Wang, 2020)). More details on proving this Lemma can be found in Appendix I.

Now we insert Lemmas 4 and (5) into (21) and get, under event $\mathcal{E}_{K,\delta}$, $Reg(K) = \widetilde{\mathcal{O}}\left(\left(H^{5/2}+m^{3/2}H\right)\sqrt{K}\right)$, which proves Theorem 1.

5 Experiments

In the section, we empirically study Algorithm 1 by deploying it to the *Cart-Pole* problem (Brockman et al., 2016b). Our results (Figure 1) show that utilizing low-rankness can significantly improve the regret order. It is worth-noting that controlling *Cart-Pole* from pixels is not easy (Lillicrap et al., 2015). In our study of the *Cart-Pole* problem, the state space is velocities (of the cart and the pole tip) together with pixels (images describing the *Cart-Pole* environment). To deploy our algorithm, we first formulate the *Cart-Pole* control as an LQR problem. Specifically, we assume the state transitions follow a linear model. Also, we let the quadratic cost penalize both bad cart/pole positions (bad pixels values) and large velocities. For the performance measure, we study both (1) costs from our LQR formulation, and (2) rewards from the *Cart-Pole* environment. For (1), the costs are computed from our LQR formulation. For (2), the environment gives a unit reward if the cart and pole are in good positions, and a zero reward otherwise. The results in (1) from LQR formulation empirically verify our theoretical analysis – smaller m values give smaller regrets in the LQR formulation. The results in (2) show that our algorithm solves the *Cart-Pole* problem faster than previous methods (Abbasi-Yadkori and Szepesvári, 2011). We also empirically verify Assumption 4: In common problems, the starting states are on a low-rank space. This study is summarized in Figure 2. More details on experiment setup are in Appendix K.

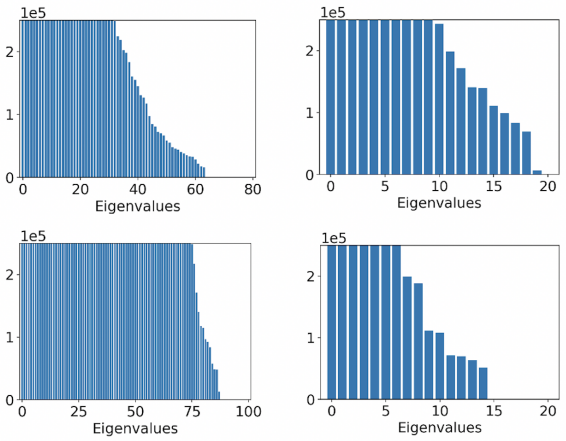

Figure 2: The four bar plots are eigenvalue distributions of starting state covariance (unnormalized) in `OpenAI Gym` problems Brockman et al. (2016b). This shows that Assumption 4 is empirically true. For the four barplots, upper left: Pendulum, upper right: Acrobot, lower left: LunarLander, lower right: Cart-Pole.

6 Conclusion

In this paper we provide a provably efficient reinforcement learning algorithm for controlling LQR systems with unknown dynamics. We show that even if the states of the system are of high dimension, our algorithm learns efficiently as long as the system has some intrinsic low-dimensional representation, i.e., the states transition happens in a low-dimensional subspace. Our algorithm leverages online LQR control and low-rank approximation techniques to achieve balanced exploration and exploitation inside the low-dimensional subspace. Numerical studies demonstrate the efficacy of our approach.

References

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

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

Abbeel, P., Coates, A., Quigley, M., and Ng, A. Y. (2007). An application of reinforcement learning to aerobatic helicopter flight. In *Advances in neural information processing systems*, pages 1–8.

Abeille, M. and Lazaric, A. (2017). Thompson sampling for linear-quadratic control problems. In *AISTATS 2017-20th International Conference on Artificial Intelligence and Statistics*.

Abeille, M. and Lazaric, A. (2018). Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9.

Aswani, A., Gonzalez, H., Sastry, S. S., and Tomlin, C. (2013). Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216–1226.

Bellman, R. et al. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515.

Bertsekas, D. P. (2004). *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 3 edition.

Bittanti, S. (1996). History and prehistory of the riccati equation. In *Proceedings of 35th IEEE Conference on Decision and Control*, volume 2, pages 1599–1604. IEEE.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016a). Openai gym.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016b). Openai gym. *arXiv preprint arXiv:1606.01540*.

Chen, J. and Nett, C. N. (1993). The caratheodory-fejer problem and h/sub/spl infin//identification: a time domain approach. In *Proceedings of 32nd IEEE Conference on Decision and Control*, pages 68–73. IEEE.

Dai, H. and Bai, Z.-Z. (2011). On eigenvalue bounds and iteration methods for discrete algebraic riccati equations. *Journal of Computational Mathematics*, pages 341–366.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *COLT*.

de la Peña, V. H., Lai, T. L., and Shao, Q.-M. (2009). Multivariate self-normalized processes with matrix normalization. *Self-Normalized Processes: Limit Theory and Statistical Applications*, pages 193–203.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2017). On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2019). On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47.

Hardt, M., Ma, T., and Recht, B. (2018). Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44.

Helmicki, A. J., Jacobson, C. A., and Nett, C. N. (1991). Control oriented system identification: a worst-case/deterministic approach in h/sub infinity. *IEEE Transactions on Automatic control*, 36(10):1163–1176.

Ibrahimi, M., Javanmard, A., and Roy, B. V. (2012). Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pages 2636–2644.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713.

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Koller, T., Berkenkamp, F., Turchetta, M., and Krause, A. (2018). Learning-based model predictive control for safe exploration. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6059–6066. IEEE.

Kveton, B., Szepesvári, C., Rao, A., Wen, Z., Abbasi-Yadkori, Y., and Muthukrishnan, S. (2017). Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644*.

Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. (2019). Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*.

Lale, S., Azizzadenesheli, K., Hassibi, B., and Anandkumar, A. (2020). Regret bound of adaptive control in linear quadratic gaussian (lqg) systems. *arXiv preprint arXiv:2003.05999*.

Li, W. and Todorov, E. (2004). Iterative linear quadratic regulator design for nonlinear biological movement systems. In *ICINCO (1)*, pages 222–229.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Ljung, L. and Söderström, T. (1983). *Theory and practice of recursive identification*. MIT press.

Mason, S., Righetti, L., and Schaal, S. (2014). Full dynamics lqr control of a humanoid robot: An experimental study on balancing and squatting. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 374–379. IEEE.

Matni, N., Wang, Y.-S., and Anderson, J. (2017). Scalable system level synthesis for virtually localizable systems. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 3473–3480. IEEE.

Maxwell, J. C. (1868). I. on governors. *Proceedings of the Royal Society of London*, (16):270–283.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Ouyang, Y., Gagrani, M., and Jain, R. (2017). Control of unknown linear systems with thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE.

Platt Jr, R., Tedrake, R., Kaelbling, L., and Lozano-Pérez, T. (2010). Belief space planning assuming maximum likelihood observations. In *Proceedings of the Robotics: Science and Systems Conference, 6th*.

Riccati, J. (1720). Personal communication to giovanni rizzetti.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

Simchowitz, M. and Foster, D. (2020). Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR.

Suh, H. and Tedrake, R. (2020). The surprising effectiveness of linear models for visual foresight in object pile manipulation. *arXiv preprint arXiv:2002.09093*.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933.

Tropp, J. A. (2011). User-friendly tail bounds for matrix martingales. Technical report, CALIFORNIA INST OF TECH PASADENA.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

Tu, S., Boczar, R., Packard, A., and Recht, B. (2017). Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*.

Vaswani, N. and Narayanamurthy, P. (2017). Finite sample guarantees for pca in non-isotropic and data-dependent noise. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 783–789. IEEE.

Yang, L. F. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *ICML 2020*.

A Notations and Algorithm Details

Table 1: List of notations

Symbol	Definition
H, K	horizon (steps per episode), number of episodes
h, k	index of horizon, episode
δ	parameter that controls event probabilities
d, d_u, m	dimension of state, dimension of control, true state rank
$Q_h,\ R_h$	positive definite matrices for state cost and control cost at $h \in [H-1]$
$Q_H,\ R_H$	positive definite matrices for state cost at H, $R_H = 0$.
$M_* = [A_* \quad B_*]$	true transition dynamics
$\hat{x}_{k,h},\ u_{k,h},\ w_{k,h},$	(observed) state, control, noise at (k,h), $\hat{z}_{k,h} = [\hat{x}_{k,h}^\top \quad u_{k,h}^\top]^\top$
$\hat{z}_{k,h}$	(observed) state-control pair at (k,h)
$x_{k,h},\ z_{k,h}$	underlying state (noise removed), underlying state-control at (k,h)
$\Psi_h(M)$	defined in (4)
ψ_h	cost from noise (given transition M) in Eq. 5
$J_h^*(M,x)$	cost under optimal policy for system governed M, at state x and step h.
$K_h(M)$	matrix that defines optimal control law at step h for system M (Eq. 3)
$\hat{X}_k,\ \hat{X}_k^{\text{next}}$	$\hat{X}_k := [\hat{x}_{h',k'}]_{\substack{h'\in[H-1]\\k'\in[k-1]}},\ \hat{X}_k^{\text{next}} := [\hat{x}_{h',k'}]_{\substack{h'\in[2,H]\\k'\in[k-1]}}$
U_k	$U_k := [u_{h',k'}]_{\substack{h'\in[2,H]\\k'\in[k-1]}}$
$\hat{Z}_k$	$\hat{Z}_k := \begin{bmatrix}\hat{X}_k\\ U_k\end{bmatrix}$
$X_k,\ Z_k$	$X_k := [x_{h',k'}]_{\substack{h'\in[H-1]\\k'\in[k-1]}},\ Z_k := \begin{bmatrix}X_k\\ U_k\end{bmatrix}$
P_k	projection matrix at k (from top m left singular values of $\hat{X}_{k-1}$)
P_k^{aug}	$P_k^{\text{aug}} = \begin{bmatrix}P_k & 0_{d\times d_u}\\ 0_{d_u\times d} & I_{d_u}\end{bmatrix}$
$\widetilde{V}_k,\ V_k$	$\widetilde{V}_k := \hat{Z}_k\hat{Z}_k^\top + I_{d+d_u},\quad V_k = P_k^{\text{aug}}\widetilde{V}_k P_k^{\text{aug}}$
$\bar{Z}_k,\ \bar{X}_k^{\text{next}}$	$\bar{Z}_k := P_k^{\text{aug}}\hat{Z}_k,\quad P_k\hat{X}_k^{\text{next}}$
M_k	$M_k^\top = \left(V_k^\dagger \bar{Z}_k \bar{X}_k^{\text{next}}\right)^\top$, the estimation of M_* at k
$\mathcal{C}_1^{(k)},\ \mathcal{C}_2^{(k)}$	confidence set perpendicular to P_k^{aug}, parallel to P_k^{aug} at k (Eq. 24, 25)
$\mathcal{C}^{(k)}$	confidence at k, $\mathcal{C}^{(k)} := \mathcal{C}_* \cap \mathcal{C}_1^{(k)} \cap \mathcal{C}_2^{(k)}$
$G_{k,\delta}, \beta_{k,\delta}$	radius of $\mathcal{C}_1^{(k)}$, $\mathcal{C}_2^{(k)}$, precisely defined in Eq. 26, 27
$\widetilde{M}_k$	optimistic estimation of M_* at episode k
$\widetilde{\Psi}_{k,h}$	$\Psi_h(M)$ quantity computed with $\widetilde{M}_k$
σ^2	noise covariance $\mathbb{E}[w_{k,h}w_{k,h}^\top] = \sigma^2 I_d$ (assumed for readability)
$\lambda_-\ (\lambda_- > 0)$	bound on m-th eigenvalue of start covariance: $\lambda_m\Big(\mathbb{E}[x_{k,1}x_{k,1}^\top]\Big) \geq \lambda_-$
C	bound on $\|Q_h\|_2$ and $\|R_h\|_2$: $\|Q_h\|_2 \leq C$ and $\|R_h\|_2 \leq C$
C_w	bound on $\|w_{k,h}\|_2$, $\|w_{k,h}\|_2 \leq C_w$
$C_{\max}$	constant $C_{\max} := 4C_w + 2\sqrt{2}C_w^2$
$K_{\min}$	warm-up period, $K_{\min} := 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$

A.1 Precise Definition for $\mathcal{C}^{(k)}$

For $k > K_{\min} := 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$, $\mathcal{C}^{(k)} := \mathcal{C}_* \cap \mathcal{C}_1^{(k)} \cap \mathcal{C}_2^{(k)}$, where

$$\mathcal{C}_* := \{M = [A \quad B] : \|M\|_2 \le 1\} \tag{23}$$

$$\mathcal{C}_1^{(k)} := \left\{M : \|(M - M_k)(I_{d+d_u} - P_k^{\text{aug}})\|_2 \lesssim G_{k,\delta}\right\}, \tag{24}$$

$$\mathcal{C}_2^{(k)} := \left\{M : \left\|(M - M_k)V_k^{1/2}\right\|_2^2 \lesssim \beta_{k,\delta}\right\}, \tag{25}$$

$$G_{k,\delta} := \frac{C_{\max}\sqrt{Hk\log\frac{d}{\delta}}}{k\lambda_- - k^{3/4}H\log\frac{m}{\delta} - C_{\max}\sqrt{Hk\log\frac{d}{\delta}}}, \quad C_{\max} := 4C_w + 2\sqrt{2}C_w^2, \tag{26}$$

$$\beta_{k,\delta} := 1 + 4C^2 G_{k,\delta}^2 Hk + G'_{k,\delta}, \quad G'_{k,\delta} := m^2 C_w^2 \log\left(\delta^{-1}\sqrt{1 + \frac{HkC^2}{m}}\right). \tag{27}$$

Note that when $k > K_{\min}$, we have $G_{k,\delta} > 0$, thus $C_1^{(k)}$ is well-defined for $k > K_{\min}$.

A.2 A Note on Constants

Note. Throughout the proof, we will overload notations to use C to denote all constants. For example, we will simultaneously say $\|M_*\| \le C$, $2\|M_*\| \le C$, and $\|M_*^2\| \le C$. The only exception is that we use C_w to denote the bound on noise. Also, this constant C does not depend on K, H, λ_-, δ or C_w, and will be omitted in $\mathcal{O}(\cdot)$ notations.

B Preparation: Computational Propositions

Proposition 3. *Recall* $\widetilde{V}_k = I_{d+d_u} + \hat{Z}_k\hat{Z}_k^\top$ *and* $V_k = P_k^{\text{aug}}\widetilde{V}_k P_k^{\text{aug}}$. *We have*

$$V_k^\dagger = P_k^{\text{aug}}\widetilde{V}_k^{-1}P_k^{\text{aug}} = \left(\bar{Z}_k\bar{Z}_k^\top + P_k^{\text{aug}}\right)^\dagger, \tag{28}$$

where $\bar{Z}_k = P_k^{\text{aug}}\hat{Z}_k$.

Proof. Let L_k^{aug} be the matrix of orthonormal columns such that $L_k^{\text{aug}}\left(L_k^{\text{aug}}\right)^\top = P_k^{\text{aug}}$. Then one has,

$$\begin{aligned}
V_k^\dagger &= \left(P_k^{\text{aug}}\widetilde{V}_k P_k^{\text{aug}}\right)^\dagger \\
&= \left(L_k^{\text{aug}}\left(L_k^{\text{aug}}\right)^\top \widetilde{V}_k L_k^{\text{aug}}\left(L_k^{\text{aug}}\right)^\top\right)^\dagger \\
&= \left(\left(L_k^{\text{aug}}\right)^\top\right)^\dagger \left(\left(L_k^{\text{aug}}\right)^\top \widetilde{V}_k L_k^{\text{aug}}\right)^\dagger \left(L_k^{\text{aug}}\right)^\dagger \\
&= L_k^{\text{aug}}\left(\left(L_k^{\text{aug}}\right)^\top \widetilde{V}_k L_k^{\text{aug}}\right)^\dagger \left(L_k^{\text{aug}}\right)^\top \\
&= L_k^{\text{aug}}\left(L_k^{\text{aug}}\right)^\dagger \widetilde{V}_k^\dagger \left(\left(L_k^{\text{aug}}\right)^\top\right)^\dagger \left(L_k^{\text{aug}}\right)^\top \\
&= L_k^{\text{aug}}\left(L_k^{\text{aug}}\right)^\top \widetilde{V}_k^{-1} L_k^{\text{aug}}\left(L_k^{\text{aug}}\right)^\top \\
&= P_k^{\text{aug}}\widetilde{V}_k^{-1}P_k^{\text{aug}}
\end{aligned}$$

Also we have,

$$\begin{aligned}
V_k = P_k^{\text{aug}}\widetilde{V}_k P_k^{\text{aug}} &= P_k^{\text{aug}}\left(\hat{Z}_k\hat{Z}_k^\top + I_{d+d_u}\right)P_k^{\text{aug}} \\
&= P_k^{\text{aug}}\hat{Z}_k\hat{Z}_k^\top P_k^{\text{aug}} + P_k^{\text{aug}} \\
&= \bar{Z}_k\bar{Z}_k^\top + P_k^{\text{aug}}.
\end{aligned}$$

□

Proposition 4. *Recall*

$$\widetilde{V}_k := I_{d+d_u} + \sum_{h=1}^{H-1} \sum_{k'=1}^{k-1} \hat{z}_{k',h} \left(\hat{z}_{k',h}\right)^\top.$$

Fix K. Let L_K^{aug} be the matrix of orthonormal columns such that $L_K^{\mathrm{aug}} \left(L_K^{\mathrm{aug}}\right)^\top = P_K^{\mathrm{aug}}$. For $k \in [1, K]$, let $V_{K,k} := P_K^{\mathrm{aug}} \widetilde{V}_k P_K^{\mathrm{aug}}$, and let $D_{K,k} := \left(L_K^{\mathrm{aug}}\right)^\top \widetilde{V}_k L_K^{\mathrm{aug}}$. We have

$$V_{K,k-1}^\dagger = L_K^{\mathrm{aug}} D_{K,k-1}^{-1} \left(L_K^{\mathrm{aug}}\right)^\top. \tag{29}$$

Proof. Using the similar argument for Proposition 3, we can prove this proposition. □

Proposition 5. *Let Assumption 2 be true. Let P_* be the true projection matrix for system $M_* = [A_* \quad B_*]$. Let $P_*^{aug} := \begin{bmatrix} P_* & 0_{d \times d_u} \\ 0_{d_u \times d} & I_{d_u} \end{bmatrix}$. Then we have $M_* = P_* M_* = M_* P_*^{\mathrm{aug}} = P_* M_* P_*^{\mathrm{aug}}$.*

Proof. By Assumption 2, the true projection matrix for system $M_* = [A_* \quad B_*]$ satisfies $M_* = [P_* A_* P_* \quad P_* B_*]$. Thus, by using $P_* = P_* P_*$, we have

$$M_* = [A_* \quad B_*] = [P_* A_* P_* \quad P_* B_*] = P_* [P_* A_* P_* \quad P_* B_*] = P_* M_*,$$

$$M_* P_*^{\mathrm{aug}} = [P_* A_* P_* \quad P_* B_*] P_*^{\mathrm{aug}} = [P_* A_* P_* P_* \quad P_* B_* I_{d_u}] = M_*.$$

□

C Well-Definedness of the Algorithm

Proposition 1. *The regions enclosed by $\mathcal{C}^{(k)}$ ($k \in (K_{\min}, K]$) are closed and bounded. Also, under event $\mathcal{E}_{K,\delta}$, $\mathcal{C}^{(k)}$ is non-empty.*

Proof. **$\mathcal{C}^{(k)}$ is closed and bounded.** It is clear that the regions are closed. Also by definition, $\mathcal{C}^{(k)}$ is bounded since $\mathcal{C}_*$ is bounded.

$\mathcal{C}^{(k)}$ is non-empty (with high probability). Under event $\mathcal{E}_{K,\delta}$, by Lemma 3, $M_* \in \mathcal{C}^{(k)}$, which shows $\mathcal{C}^{(k)}$ is non-empty.

□

Lemma 1. *The matrix $\Psi_h(M)$ is positive semi-definite for any $h \in [H]$ and M, provided that Q_h, R_h are positive definite.*

Proof. By Bellman optimality, we know the optimal cost is given by

$$J_h^*(M, x) = \min_a \left\{ x^\top Q_h x + u^\top R_h u + \overline{J_{h+1}^*} \right\}, \qquad \text{where}$$
$$\overline{J_{h+1}^*} := \mathbb{E}_{w_h}[J_{h+1}^*(M, Ax + Bu + w_h)].$$

Solving this dynamic programming problem gives (pp.150, Chapter 4, Vol I; pp. 229, Chapter 5, Vol. I Bertsekas (2004))

$$J_h^*(M, x) = x^\top \Psi_h x + \psi_h,$$

where

$$x^\top \Psi_H(M) x = x^\top Q_H x,$$
$$x^\top \Psi_h(M) x = \min_u \left[x^\top Q_h x + u^\top R_h u + (Ax + Bu)^\top \Psi_{h+1}(M) (Ax + Bu) \right], \quad h < H. \tag{30}$$

First, we know from definition that $\Psi_H(M) = Q_H$ is positive definite. Inductively, given $\Psi_{h+1}(M)$ positive semi-definite, we have from (30) that $x^\top \Psi_{h+1}(M) x \geq 0$, for any x. This is because minimization preserves non-negativity. This shows that $\Psi_h(M)$ is positive semi-definite for all h. □

C.1 Boundedness Results

Based on Assumptions 1-4, the matrices $\widetilde{\Psi}_h(M)$ (defined in Eq. 4) for any $M \in \mathcal{C}^k$ ($k \in (K_{\min}, K]$) are bounded. Also, the states, and controls are bounded.

Proposition 6. *For all $k \in (K_{\min}, K]$ and $h \in [H]$, there exists a constant C, such that, for all $M \in \mathcal{C}^{(k)}$,*

$$\|\Psi_h(M)\|_2 \leq C \quad \text{and} \quad \|\mathcal{K}_h(M)\|_2 \leq C. \tag{31}$$

Proof. Recall for $M = [A \quad B]$, the quantities $\Psi_h(M)$ are recursively computed by

$$\begin{aligned}\Psi_H(M) &:= Q_H,\\ \Psi_{h-1}(M) &:= Q_h + A^\top \Psi_h(M) A - A^\top \Psi_h(M) B \left(R_h + B^\top \Psi_h(M) B\right)^{-1} B^\top \Psi_h(M) A, \text{ for } h < H.\end{aligned} \tag{32}$$

Any $M \in \mathcal{C}^{(k)}$ is of bounded norm. This is because $\mathcal{C}_*$ only includes matrices with norm smaller than C.

By positive definiteness of R_h, Q_h, $M = [A \quad B]$, we know that $\Psi_{h-1}(M)$ is of bounded norm that is independent of H by eigenvalue bounds on Riccati iterations (Theorem 3.1, item (i), Dai and Bai, 2011). Note that we can apply this specific result (Theorem 3.1, item (i), Dai and Bai, 2011) even if our system is heterogeneous (in terms of R_h and Q_h).

Thus, by boundedness of A, B and $\Psi_h(M)$, we have

$$\|\mathcal{K}_h(M)\|_2 = \| \left(R_h + B^\top \Psi_{h+1}(M) B\right)^{-1} B^\top \Psi_{h+1}(M) A\|_2 \leq C \tag{33}$$

for some constant C.

□

Proposition 7. *Under Assumptions 1-4 and event $\mathcal{E}_{K,\delta}$, our algorithm satisfies*

$$\|\hat{x}_{k,h}\|_2 \leq 1 \quad \text{and} \quad \|\hat{z}_{k,h}\|_2 \leq C$$

for all $k \in [\max(K_{\min}, K'_{\min}) + 1, K]$ and $h \in [H]$, where $K'_{\min} := \left\{k : G_{k,\delta}^2 + \beta_{k,\delta}^2 \leq \frac{C_w^2}{C^4}\right\}$.

Note that $K'_{\min}$ is a constant since $G_{k,\delta} = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{k}}\right)$ and $\beta_{k,\delta} = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{k}}\right)$.

Proof. In Eq. 19, our control is defined by

$$u_{k,h} = \pi_h^{(k)}(x) = \mathcal{K}_h(\widetilde{M}_k)\hat{x}_{k,h}, \tag{34}$$

Also, since $M_* = [A_* \quad B_*]$, if $\|x_{k,h}\|_2 \leq 1$,

$$\begin{aligned}\|\hat{x}_{k,h+1}\|_2 &= \left\|A_*\hat{x}_{k,h} + B_*\mathcal{K}_h(\widetilde{M}_k)\hat{x}_{k,h}\right\|_2 + \left\|B_*\mathcal{K}_h(\widetilde{M}_*)\hat{x}_{k,h} - B_*\mathcal{K}_h(\widetilde{M}_k)\hat{x}_{k,h}\right\|_2 + \|w_{k,h}\|_2\\ &\leq r + \left\|B_*\mathcal{K}_h(\widetilde{M}_*)\hat{x}_{k,h} - B_*\mathcal{K}_h(\widetilde{M}_k)\hat{x}_{k,h}\right\|_2 + C_w,\end{aligned} \tag{35}$$

where the last line uses Assumption 1 (item (1)). Also by Assumption 1 (item (2)), we have

$$\left\|B_*\mathcal{K}_h(M_*)\hat{x}_{k,h} - B_*\mathcal{K}_h(\widetilde{M}_k)\hat{x}_{k,h}\right\|_2 \leq \|B_*\|_2\|\widetilde{M}_k - M_*\|_2 \leq C^2\|\widetilde{M}_k - M_*\|_2.$$

As proved in Section 4.1 (and related appendix sections), under event $\mathcal{E}_{K,\delta}$, we use Pythagoras theorem to get

$$\left\|M_* - \widetilde{M}_k\right\|_2 \leq \sqrt{G_{k,\delta}^2 + \beta_{k,\delta}}.$$

This means, for $k > K'_{\min} := \left\{k : G_{k,\delta}^2 + \beta_{k,\delta}^2 \leq \frac{C_w^2}{C^4}\right\}$, under event $\mathcal{E}_{K,\delta}$,

$$\left\|B_*\mathcal{K}_h(M_*)\hat{x}_{k,h} - B_*\mathcal{K}_h(\widetilde{M}_k)\hat{x}_{k,h}\right\|_2 \leq C^2\left\|M_* - \widetilde{M}_k\right\|_2 \leq C^2\sqrt{G_{k,\delta}^2 + \beta_{k,\delta}} \leq C_w.$$

By using Assumption 1 (item (3)) in (35) , we get

$$\|\hat{x}_{k,h+1}\|_2 \leq 1.$$

Inductively, this means $\|\hat{x}_{k,h}\|_2 \leq 1$ for all k, h. Then by Proposition 6, we know $\|\hat{z}_{k,h}\|_2 \leq \left\|\mathcal{K}_h(\widetilde{M}_k)\hat{x}_{k,h}\right\|_2 \leq C$.

□

D Projection Error Analysis (Lemma 2)

In this section, we prove Lemma 2. Our arguments follow the same mechanism as the ones by Vaswani and Narayanamurthy (2017); Lale et al. (2019). We will use the following notation:

- $\mathbb{E}_k$: the expectation conditioned on all randomness up to time $k-1$.

We first need the following Azuma-Hoeffding inequality for positive semi-definite matrices, which appears as Theorem 3.1 in Tropp (2012).

Lemma 6 (Matrix Chernoff Tropp (2011, 2012))**.** *Consider a finite adapted sequence $\{\mathbf{X}_k\}$ of positive-semidefinite matrices with dimension p, and suppose that $\lambda_{\max}(\mathbf{X}_k) \leq R$ almost surely. Define the finite series $\mathbf{Y} := \sum_k \mathbf{X}_k$ and $\mathcal{Y} := \sum_k \mathbb{E}_k \mathbf{X}_k$, where $\mathbb{E}_k$ is the expectation conditioned on all randomness before $\mathbf{X}_k$.*

Then for all $\mu \geq 0$,

$$\mathbb{P}\left\{\lambda_{\min}(\mathbf{Y}) \leq (1-\delta)\mu \;\; \textit{and} \;\; \lambda_{\min}\left(\mathcal{Y} \geq \mu\right)\right\} \leq d\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\frac{\mu}{R}} \tag{36}$$

We also need the following Lemma by Tropp (2012).

Lemma 7 (Matrix Azuma Tropp (2011, 2012))**.** *Consider a matrix martingale $\{\mathbf{Y}_k : k = 0, 1, 2, \cdots\}$ whose values are self-adjoint matrices with dimension d, and let $\{\mathbf{X}_k : k = 1, 2, 3, \cdots\}$ be the difference sequence. Assume that the difference sequence satisfies:*

$$\mathbb{E}_k \mathbf{X}_k = \mathbf{0},$$

and that there exists a deterministic matrix sequence $\{\mathbf{A}_k : k = 1, 2, 3, \cdots\}$ and $\mathbf{X}_k^2 \preceq \mathbf{A}_k^2$ almost surely for $k = 1, 2, 3, \cdots$.

Define the following:

$$\mathbf{Y}_k := \sum_{j=1}^{k} \mathbf{X}_j, \qquad \sigma_M^2 := \left\| \sum_{j=1}^{k} A_k^2 \right\|_2 \tag{37}$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(Y_k\right) \geq t\right\} \leq d \exp\left\{-\frac{t^2}{8\sigma_M^2}\right\}. \tag{38}$$

We also need the following Davis-Kahan $\sin\Theta$ theorem.

Theorem 2 (Davis-Kahan)**.** *Let $S, W \in \mathbb{R}^{d\times d}$ be symmetric matrices, and let $\hat{S}_k = S + W$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$ be the eigenvalues of S and $\hat{S}$ respectively. Define the eigenvalue decompositions of S and $\hat{S}$:*

$$S = \begin{bmatrix} L & L_0 \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_0 \end{bmatrix} \begin{bmatrix} R & R_0 \end{bmatrix}$$

$$\hat{S}_k = \begin{bmatrix} \hat{L} & \hat{L}_0 \end{bmatrix} \begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Lambda}_0 \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{R}_0 \end{bmatrix},$$

where Λ (resp. $\hat{\Lambda}$) is the diagonal matrix of the top m eigenvalues of S (resp. $\hat{S}$), and L (resp. $\hat{L}$) is the matrix of the corresponding eigenvectors of S (resp. $\hat{S}$).

If $\lambda_m > \hat{\lambda}_{m+1}$, then $\sin\Theta_m$, the sine of the largest principal angle between the column spans of L and $\hat{L}$, can be upper bounded by

$$\sin\Theta_m \leq \frac{\left\|\hat{S}_k L - L\Lambda\right\|_2}{\lambda_m - \hat{\lambda}_{m+1}}.$$

In addition,

$$\left\|LL^\top - \hat{L}\hat{L}^\top\right\|_2 = \sin\Theta_m \leq \frac{\left\|\hat{S}_k L - L\Lambda\right\|_2}{\lambda_m - \hat{\lambda}_{m+1}}.$$

Next, we recap Lemma 2 and provide a proof.

Lemma 2. *Assume the the conditions in Assumption 4 hold. Then for $k > K_{\min}$, where*

$$K_{\min} := 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\},$$

with probability at least $1-3\delta$,

$$\|P_* - P_k\|_2 \le G_{k,\delta},$$

where

$$G_{k,\delta} := \frac{C_{\max}\sqrt{Hk\log\frac{d}{\delta}}}{k\lambda_- - k^{3/4}H\log\frac{m}{\delta} - C_{\max}\sqrt{Hk\log\frac{d}{\delta}}}.$$

Proof. Let P_* be the true projection matrix of the system. For indexing simplicity, we consider $\|P_{k+1} - P_*\|_2$. Recall that $\hat{x}_{k',h} = x_{k',h} + w_{k',h}$ where $x_{k',h}$ lies on a low-rank space.

Let L be the rank-m orthonormal matrix such that $P_* = LL^\top$. We set

$$\hat{S}_k := \sum_{\substack{h<H\\k'<k}} \hat{x}_{k',h}\hat{x}_{k',h}^\top = \hat{X}_k\hat{X}_k^\top$$
$$S_k := \sum_{\substack{h<H\\k'<k}} x_{k',h}x_{k',h}^\top + LL^\top W_k W_k^\top LL^\top.$$

Throughout the rest of the proof, for a symmetric matrix A, we use $\lambda_i(A)$ to denote the i-th largest eigenvalue of A.

Step 1: High probability lower bound on $\lambda_m(S_k)$.

High level sketch for this step: *Use the Matrix Chernoff Bound (Lemma 6) and Assumption 4 to show a positive lower bound on $\lambda_m(S_k)$.*

Since $x_{k,h}$ lies in the subspace spanned by L (Recall $P_* = LL^\top$), we let $b_{k,h}$ be proper m-dimensional vector such that $x_{k,h} = Lb_{k,h}$. Thus

$$X_kX_k^\top + LL^\top W_kW_k^\top LL^\top = \sum_{\substack{h<H\\k'<k}} Lb_{k',h}b_{k',h}^\top L^\top + \sum_{\substack{h<H\\k'<k}} LL^\top w_{k',h}w_{k',h}^\top LL^\top,$$

and

$$\begin{aligned}
&\lambda_{\min}\left(\sum_{\substack{h<H\\k'<k}}\left[b_{k,h}b_{k,h}^\top + L^\top w_{k,h}w_{k,h}^\top L\right]\right)\\
=&\lambda_m\left(L\sum_{\substack{h<H\\k'<k}}\left[b_{k,h}b_{k,h}^\top + L^\top w_{k,h}w_{k,h}^\top L\right]L^\top\right)\\
=&\lambda_m\left(\sum_{\substack{h<H\\k'<k}}\left[x_{k,h}x_{k,h}^\top + LL^\top w_{k,h}w_{k,h}LL^\top\right]\right)\\
=&\lambda_m\left(S_k\right). \qquad (39)
\end{aligned}$$

By Assumption 4,

$$\lambda_{\min}(\mathbb{E}_k[b_{k,1}b_{k,1}^\top]) = \lambda_m(\mathbb{E}_k[x_{k,1}x_{k,1}^\top] \ge \lambda_-. \qquad (40)$$

By taking conditional expectations and eigenvalues, we have,

$$
\begin{aligned}
&\lambda_m\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\mathbb{E}_{k'}[Lb_{k',h}b_{k',h}^\top L^\top + LL^\top w_{k,h}w_{k,h}^\top LL^\top]\right)\\
=&\lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\mathbb{E}_{k'}[b_{k',h}b_{k',h}^\top + L^\top w_{k,h}w_{k,h}^\top L]\right) && \text{(since } L \text{ is orthonormal and } P_* = LL^\top\text{)}\\
\geq&\lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\mathbb{E}_{k'}[b_{k',h}b_{k',h}^\top]\right) + \lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\mathbb{E}_{k'}[L^\top w_{k,h}w_{k,h}^\top L]\right) && \text{(by Lidskii inequality)}\\
\geq&k\lambda_- + \lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\mathbb{E}_{k'}[L^\top w_{k,h}w_{k,h}^\top L]\right) && \text{(by Eq. 40)}\\
=&(k-1)\lambda_- + (k-1)(H-1)\sigma^2. && (41)
\end{aligned}
$$

Let $\mathcal{F}_k$ be all randomness by end of episode $k-1$. Then $\left\{\sum_{h=1}^{H-1} b_{k,h}b_{k,h}^\top\right\}_k$ are adapted to the filtration $\{\mathcal{F}_k\}_k$, since $\left\{\sum_{h=1}^{H-1} x_{k,h}x_{k,h}^\top\right\}_k$ are adapted to $\{\mathcal{F}_k\}_k$ and $b_{k,h}$ are determined by $x_{k,h}$ and the constant matrix L.

By Assumption 1, $\lambda_{\max}\left(\sum_{h=1}^{H-1} b_{k',h}b_{k',h}^\top\right) = \lambda_{\max}\left(\sum_{h=1}^{H-1} x_{k',h}x_{k',h}^\top\right) \leq H$. Then we apply Lemma 6 to the matrices $\left\{\sum_{h=1}^{H-1} b_{k',h}b_{k',h}^\top\right\}_{k'}$, and set $\delta = k^{-1/4}$ to get

$$
\begin{aligned}
&\mathbb{P}\Bigg[\lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Lw_{k,h}w_{k,h}^\top L^\top\right)\right)\\
&\qquad\leq\left(1-k^{-1/4}\right)\lambda_{\min}\left(\sum_{k'=1}^{k-1}\mathbb{E}_{k'}\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Lw_{k,h}w_{k,h}^\top L^\top\right)\right)\Bigg]\\
\leq&m\left[\frac{\exp\left\{-k^{-1/4}\right\}}{\left(1-k^{-1/4}\right)^{\left(1-k^{-1/4}\right)}}\right]^{\frac{\lambda_{\min}\left(\sum_{k'=1}^{k-1}\mathbb{E}_{k'}\left[\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Lw_{k,h}w_{k,h}^\top L^\top\right)\right]\right)}{H}}. && (42)
\end{aligned}
$$

Since, from (41),

$$
\lambda_m\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\mathbb{E}_{k'}[Lb_{k',h}b_{k',h}^\top L^\top + LL^\top w_{k,h}w_{k,h}^\top LL^\top]\right) \geq (k-1)\lambda_- + (k-1)(H-1)\sigma^2,
$$

we have

$$
\begin{aligned}
&\mathbb{P}\Bigg[\lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Lw_{k,h}w_{k,h}^\top L^\top\right)\right)\\
&\qquad\leq\left(1-k^{-1/4}\right)\left[(k-1)\lambda_- + (k-1)(H-1)\sigma^2\right]\Bigg]\\
\leq&\mathbb{P}\Bigg[\lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Lw_{k,h}w_{k,h}^\top L^\top\right)\right)\\
&\qquad\leq\left(1-k^{-1/4}\right)\lambda_{\min}\left(\sum_{k'=1}^{k-1}\mathbb{E}_{k'}\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Lw_{k,h}w_{k,h}^\top L^\top\right)\right)\Bigg]\\
\leq&m\left[\frac{\exp\left\{-k^{-1/4}\right\}}{\left(1-k^{-1/4}\right)^{\left(1-k^{-1/4}\right)}}\right]^{\frac{\lambda_{\min}\left(\sum_{k'=1}^{k-1}\mathbb{E}_{k'}\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Uw_{k,h}w_{k,h}^\top U^\top\right)\right)}{H}}\\
\leq&m\left[\frac{\exp\left\{-k^{-1/4}\right\}}{\left(1-k^{-1/4}\right)^{\left(1-k^{-1/4}\right)}}\right]^{\frac{(k-1)\lambda_- + (k-1)(H-1)\sigma^2}{H}}. && (43)
\end{aligned}
$$

We combine (39) and (43) to get

$$\begin{aligned}
&\mathbb{P}\left[\lambda_m(S_k) \le \left(1-k^{-1/4}\right)(k-1)\lambda_-\right] \\
=&\mathbb{P}\left[\lambda_{\min}\left(\sum_{k'=1}^{k-1}\sum_{h=1}^{H-1}\left(b_{k',h}b_{k',h}^\top + Lw_{k,h}w_{k,h}^\top L^\top\right)\right)\right. \\
&\qquad\left. \le \left(1-k^{-1/4}\right)\left[(k-1)\lambda_- + (k-1)(H-1)\sigma^2\right]\right] \\
\le& m\left[\frac{\exp\left\{-k^{-1/4}\right\}}{\left(1-k^{-1/4}\right)^{\left(1-k^{-1/4}\right)}}\right]^{\frac{(k-1)\lambda_- + (k-1)(H-1)\sigma^2}{H}}.
\end{aligned}$$

Since $\left[\frac{\exp\left\{-k^{-1/4}\right\}}{\left(1-k^{-1/4}\right)^{\left(1-k^{-1/4}\right)}}\right]^{\frac{(k-1)\lambda_- + (k-1)(H-1)\sigma^2}{H}} \le \exp\left\{-\frac{k^{1/4}\left(\lambda_- + (H-1)\sigma^2\right)}{H}\right\}$, which can be verified by calculus, after some computation and rearrangement, we get

$$\begin{aligned}
&\mathbb{P}\left[\lambda_m(S_k) \le \left(1-k^{-1/4}\right)\left[(k-1)\lambda_- + (k-1)(H-1)\sigma^2\right]\right] \\
\le& m\exp\left\{-\frac{k^{1/4}\left(\lambda_- + (H-1)\sigma^2\right)}{H}\right\}.
\end{aligned}$$

Thus for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\lambda_m(S_k) > (k-1)\lambda_- + (k-1)(H-1)\sigma^2 - k^{3/4}H\log\frac{m}{\delta}. \tag{44}$$

Step 2: High probability upper bound on $\left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2$ $\left(\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]$ defined below$\right)$

Note: the $\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]$ term is defined below in (45).*

For simplicity, we define

$$\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right] := \sum_{k'=1}^{k-1}\mathbb{E}_{k'}\left[\sum_{h=1}^{H-1}\hat{x}_{k',h}\hat{x}_{k',h}^\top - x_{k',h}x_{k',h}^\top - LL^\top w_{k',h}w_{k1,h}^\top LL^\top\right], \tag{45}$$

where $\mathbb{E}_{k'}$ is the expectation conditioning on all randomness before episode k'.

High level sketch for this step: *Apply the Matrix Azuma Inequality (Lemma 7) to bound the term* $\left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2$.

By definition, we have

$$\hat{S}_k - S_k = W_kX_k^\top + X_kW_k^\top + W_kW_k^\top - LL^\top W_kW_k^\top LL^\top$$

$$\begin{aligned}
\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right] =& \mathbb{E}_k^\Sigma\left[W_kX_k^\top + X_kW_k^\top + W_kW_k^\top - LL^\top W_kW_k^\top LL^\top\right] \\
=& (H-1)(k-1)\sigma^2(I_d - LL^\top),
\end{aligned} \tag{46}$$

where the notation $\mathbb{E}[w_{k,h}] = 0$ is defined in (45), $\mathbb{E}_k[w_{k',h}w_{k',h}^\top] = \sigma^2 I_d$ (Assumption 3), and the last equation uses that $\mathbb{E}_k w_{k',h} = 0$ (Assumption 3).

Thus we have,

$$\begin{aligned}
\left\|\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2^2 =& \left\|(H-1)(k-1)\sigma^2(I_d - LL^\top)\right\|_2 \\
=& (H-1)(k-1)\sigma^2, \\
\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right] =& W_kX_k^\top + X_kW_k^\top + W_kW_k^\top - LL^\top W_kW_k^\top LL^\top \\
&- (H-1)(k-1)\sigma^2(I_d - LL^\top).
\end{aligned}$$

We write $P_*^\perp := I_d - P_*$. Since $P_* = LL^\top$, from the above expression for $\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]$, we have

$$\begin{aligned}\left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2 =& \left\|W_k X_k^\top + X_k W_k^\top + P_*^\perp W_k W_k^\top P_*^\perp - (H-1)(k-1)\sigma^2 P_*^\perp\right\|_2 \\ \leq& \lambda_{\max}\left(P_*^\perp W_k W_k^\top P_*^\perp - \sigma^2(H-1)(k-1)P_*^\perp\right) \\ &+ \lambda_{\max}\left(W_k X_k^\top + X_k W_k^\top\right), \end{aligned} \tag{47}$$

where on the last step we use the triangle inequality.

Since, for any (k,h),

$$\mathbb{E}_k\left[x_{k,h} w_{k,h}^\top\right] = 0, \quad \text{and} \quad \mathbb{E}_k\left[P_*^\perp w_{k,h} w_{k,h}^\top P_*^\perp - \sigma^2 P_*^\perp\right] = 0, \tag{48}$$

the sequences $\left\{P_*^\perp W_k W_k^\top P_*^\perp - \sigma^2(H-1)(k-1)P_*^\perp\right\}_k$ and $\left\{W_k X_k^\top + X_k W_k^\top\right\}_k$ are both martingale sequences of symmetric matrices.

From there we use the Matrix Azuma inequality to get: for any $\delta \in (0,1)$,

$$\mathbb{P}\left\{\lambda_{\max}\left(P_*^\perp W_k W_k^\top P_*^\perp - \sigma^2(H-1)(k-1)P_*^\perp\right) \geq \sqrt{8C_w^4 Hk\log\frac{d}{\delta}}\right\} \leq \delta, \tag{49}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(W_k X_k^\top + X_k W_k^\top\right) \geq 4C_w\sqrt{Hk\log\frac{d}{\delta}}\right\} \leq \delta. \tag{50}$$

Then by a union bound and (47), we get, for any $\delta \in (0,1)$, with probability at least $1-2\delta$,

$$\left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2 \leq C_{\max}\sqrt{Hk\log\frac{d}{\delta}}, \tag{51}$$

where $C_{\max} = 4C_w + 2\sqrt{2}C_w^2$.

Step 3: High probability lower bound on $\lambda_m(S_k) - \lambda_{m+1}(\hat{S}_k)$ when k is larger than a constant.

High level sketch for this step: *link* $\lambda_{m+1}(\hat{S}_k)$ *to* $\left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2$ *and apply results from Step 1 and Step 2.*

Since $\lambda_{m+1}(S_k) = 0$, by Weyl's inequality, we have

$$\lambda_{m+1}(\hat{S}_k) \leq \lambda_{m+1}(S_k) + \lambda_1(\hat{S}_k - S_k) = \left\|\hat{S}_k - S_k\right\|_2. \tag{52}$$

By triangle inequality,

$$\begin{aligned}\left\|\hat{S}_k - S_k\right\|_2 \leq& \left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2 + \left\|\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2 \\ \leq& \left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2 + (k-1)(H-1)\sigma^2, \end{aligned} \tag{53}$$

where the last step uses $\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right] = \sigma^2(k-1)(H-1)(I_d - P_*)$. (Recall $\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]$ is defined in Eq. 45.)

From step 1, we have, with probability at least $1-\delta$,

$$\lambda_m(S_k) \geq (k-1)\lambda_- + (k-1)(H-1)\sigma^2 - k^{3/4}H\log\frac{m}{\delta}. \tag{54}$$

With probability at least $1-2\delta$, the results in both Step 1 and Step 2 holds, which gives,

$$\begin{aligned}&\lambda_m(S_k) - \lambda_{m+1}(\hat{S}_k) \\ \geq& (k-1)\lambda_- + (k-1)(H-1)\sigma^2 - k^{3/4}H\log\frac{m}{\delta} - \lambda_{m+1}(\hat{S}_k) && (55)\\ \geq& (k-1)\lambda_- + (k-1)(H-1)\sigma^2 - k^{3/4}H\log\frac{m}{\delta} - \left\|\hat{S}_k - S_k\right\|_2 && (56)\\ \geq& (k-1)\lambda_- + (k-1)(H-1)\sigma^2 - k^{3/4}H\log\frac{m}{\delta} \\ &- \left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2 - (k-1)(H-1)\sigma^2 && (57)\\ =& (k-1)\lambda_- - k^{3/4}H\log\frac{m}{\delta} - \left\|\hat{S}_k - S_k - \mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right]\right\|_2 \\ \geq& (k-1)\lambda_- - k^{3/4}H\log\frac{m}{\delta} - C_{\max}\sqrt{Hk\log\frac{d}{\delta}}, && (58)\end{aligned}$$

where (55) uses Step 1, (56) uses (52), (57) uses (53), and (58) uses Step 2.

When $k > K_{\min} := 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$, the expression in (58) is positive. Thus, with probability at least $1-2\delta$, for all $k > K_{\min} := 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$,

$$\lambda_m(S_k) - \lambda_{m+1}(\hat{S}_k) \geq (k-1)\lambda_- - k^{3/4}H\log\frac{m}{\delta} - C_{\max}\sqrt{Hk\log\frac{d}{\delta}} > 0. \tag{59}$$

Step 4: Final step.

High level sketch of this step: *Apply Theorem 2 and combine previous steps.*

We use P_* to denote the true projection matrix and it is clear that $P_* = LL^\top$. Then, with probability at least $1-3\delta$, for any $k > K_{\min}$

$$\|P_{k+1} - P_*\|_2 \overset{①}{\leq} \frac{\left\|\hat{S}_k L - L\Lambda\right\|_2}{\lambda_m(S_k) - \lambda_{m+1}(\hat{S}_k)} \overset{②}{=} \frac{\left\|(\hat{S}_k - S_k)L\right\|_2}{\lambda_m(S_k) - \lambda_{m+1}(\hat{S}_k)},$$

where ① uses Theorem 2 (we can use Theorem 2 since by Step 3, $\lambda_m(S) > \lambda_{m+1}(\hat{S})$ with high probability for all $k > K_{\min}$), ② uses $S = L\Lambda L^\top$.

Applying the triangle inequalities to the above and get:

$$\begin{aligned}
\|P_{k+1} - P_*\|_2 \leq & \frac{\left\|(\hat{S}_k - S_k)L\right\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S}_k)} \\
= & \frac{\left\|\mathbb{E}_k^\Sigma(\hat{S}_k - S_k)L + (\hat{S}_k - S_k)L - \mathbb{E}_k^\Sigma(\hat{S}_k - S_k)L\right\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S}_k)} \\
\leq & \frac{\left\|\mathbb{E}_k^\Sigma(\hat{S}_k - S_k)L\right\|_2 + \left\|(\hat{S}_k - S_k)L - \mathbb{E}_k^\Sigma(\hat{S}_k - S_k)L\right\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S}_k)} \\
\leq & \frac{\left\|\mathbb{E}_k^\Sigma(\hat{S}_k - S_k)L\right\|_2 + \left\|(\hat{S}_k - S_k) - \mathbb{E}_k^\Sigma(\hat{S}_k - S_k)\right\|_2 \|L\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S}_k)} \\
\leq & \frac{\left\|\mathbb{E}_k^\Sigma(\hat{S}_k - S_k)L\right\|_2 + \left\|(\hat{S}_k - S_k) - \mathbb{E}_k^\Sigma(\hat{S}_k - S_k)\right\|_2}{\lambda_m(S_k) - \lambda_{m+1}(\hat{S}_k)},
\end{aligned} \tag{60}$$

where in the last step we use $\|L\|_2 = 1$.

We use (46) to get,

$$\mathbb{E}_k^\Sigma\left[\hat{S}_k - S_k\right] L = (H-1)(k-1)\sigma^2(I_d - LL^\top)L = 0. \tag{61}$$

We then plug (61) into (60) to get, with probability at least $1-3\delta$,

$$\begin{aligned}
\|P_{k+1} - P_*\|_2 \leq & \frac{\left\|\mathbb{E}_k^\Sigma(\hat{S}_k - S_k)L\right\|_2 + \left\|(\hat{S}_k - S_k) - \mathbb{E}_k^\Sigma(\hat{S}_k - S_k)\right\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S}_k)} \\
\leq & \frac{\left\|(\hat{S}_k - S_k) - \mathbb{E}_k^\Sigma(\hat{S}_k - S_k)\right\|_2}{\lambda_m(S) - \lambda_{m+1}(\hat{S}_k)}.
\end{aligned}$$

Then by Step 2 (Eq. 51) and Step 3 (Eq. 59), we have with probability at least $1-3\delta$, for any $k > K_{\min} := 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\}$,

$$\|P_{k+1} - P_*\|_2 \leq \frac{C_{\max}\sqrt{Hk\log\frac{d}{\delta}}}{(k-1)\lambda_- - k^{3/4}H\log\frac{m}{\delta} - C_{\max}\sqrt{Hk\log\frac{d}{\delta}}}. \tag{62}$$

This concludes the proof.

□

E Rank-deficit Self-normalized Processes (for Lemma 3)

In this section, we prove Lemma 9, which is used to prove Lemma 3. We first need the following result.

Lemma 8 (Lemma 8 in Abbasi-Yadkori et al. (2011))**.** *Let $Y_1, Y_2, \cdots \in \mathbb{R}^d$ be vector-valued random variables, and let $\{\eta_s\}_s$ be real-valued random variables. Let $\mathcal{F}_t := \sigma(Y_1, Y_2, \cdots, Y_{t-1}, \eta_1, \eta_2, \cdots, \eta_{t-1})$, and let $\{\eta_s\}_s$ be conditionally R-sub-Gaussian:*

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}[\lambda \eta_t | \mathcal{F}_t] \leq \exp\left\{\frac{R^2\lambda^2}{2}\right\}.$$

Let $\lambda \in \mathbb{R}^d$ be arbitrary and consider for any $t \geq 0$,

$$M_t^\lambda := \exp\left(\sum_{s=1}^{t}\left[\frac{\eta_s \langle \lambda, Y_s\rangle}{R} - \frac{1}{2}\langle \lambda, Y_s\rangle^2\right]\right). \tag{63}$$

Then for τ a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^\infty$, M_τ^λ is almost surely well-defined and

$$\mathbb{E}[M_\tau^\lambda] \leq 1.$$

Lemma 9. *Consider a process $\{y_s\}_s$ in $\mathbb{R}^d$. Let $\{\eta_s\}_s$ be a mean-zero process on the real line. Define $\mathcal{F}_t = \sigma(y_1, \cdots, y_{t-1}, \eta_1, \cdots, \eta_{t-1})$ and let η be conditionally R-sub-Gaussian (with respect to $\mathcal{F}_t$). Let $Y_t := [y_s]_{s\in[1,t]}$ be the matrix whose columns are y_s. Let t be a stopping time and let $\{P_s\}_s$ be a sequence of rank-m projection matrices such that P_t is $\mathcal{F}_t$-measurable. Let $\bar{Y}_t := P_t Y_t$, let $\bar{V}_t := \bar{Y}_t\bar{Y}_t^\top$, $V_t := \bar{Y}_t\bar{Y}_t^\top + P_t$ and let $S_t := \bar{Y}_t\xi_t$, where $\xi_t = [\eta_s]_{s\in[1,t]}^\top$ is the vector formed by η_s. Then, for any $\delta > 0$,*

$$\mathbb{P}\left(\|S_t\|^2_{V_t^\dagger} > 2R^2 \log\left(\delta^{-1}\sqrt{\frac{\det^* V_t}{\det^* P_t}}\right)\right) \leq \delta, \tag{64}$$

where $\|\cdot\|_{V_t^\dagger}$ is the semi-norm induced by $V_t^\dagger$, and $\det^*$ is the pseudo-determinant operator.

Proof. To prove this lemma, we first need Lemma 8 by Abbasi-Yadkori et al. (2011), and use the techniques by de la Peña et al. (2009); Abbasi-Yadkori et al. (2011).

For a rank-deficit symmetric positive semi-definite matrix $\Sigma^\dagger$ and vector $\mu \in span(\Sigma)$, consider the singular (or degenerate) multi-variate Gaussian distribution $\mathcal{N}(\mu, \Sigma^\dagger)$ whose mean is μ and covariance is $\Sigma^\dagger$. The density of this distribution is defined only over $span(\Sigma)$. For $x \in span(\Sigma)$, the density function of this degenerate multi-variate Gaussian is

$$\begin{aligned} f_{\mu,\Sigma}(x) &= \frac{1}{\sqrt{\det^*(2\pi\Sigma^\dagger)}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma(x-\mu)\right) \\ &= \frac{1}{\sqrt{\det^*(2\pi\Sigma^\dagger)}}\exp\left(-\frac{1}{2}\|x-\mu\|_\Sigma\right), \end{aligned} \tag{65}$$

where $\det^*$ is the pseudo-determinant. (For PSD matrices, pseudo-determinant gives the product of positive eigenvalues.) For $\lambda \in span(\bar{V}_t)$, let

$$M_t^\lambda = \exp\left(\sum_{s=1}^{t}\left[\frac{\eta_s\langle\lambda, P_t y_s\rangle}{R} - \frac{1}{2}\langle\lambda, P_t y_s\rangle^2\right]\right),$$

as defined above in (63).

Let P_t be the projection matrix at t. Let f_{0,P_t} be the density for $\mathcal{N}(0, P_t)$. With respect to this density

f_{0,P_t}, we have

$$
\begin{aligned}
&\int_{span(P_t)} M_t^{\lambda} f_{0,P_t}(\lambda) d\lambda \\
=&\int_{span(P_t)} \exp\left(\frac{\lambda^\top S_t}{R} - \frac{1}{2}\|\lambda\|_{\bar{V}_t}\right) f_{0,P_t}(\lambda)\, d\lambda \\
=&\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{\bar{V}_t^\dagger S_t}{R}\right\|_{\bar{V}_t}^2 + \frac{1}{2}\left\|\frac{S_t}{R}\right\|_{\bar{V}_t^\dagger}^2\right) f_{0,P_t}(\lambda)\, d\lambda \qquad (66)\\
=&\exp\left(\frac{1}{2R^2}\|S_t\|_{\bar{V}_t^\dagger}^2\right)\cdot\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{\bar{V}_t^\dagger S_t}{R}\right\|_{\bar{V}_t}^2\right) f_{0,P_t}(\lambda)\, d\lambda \\
=&\frac{\exp\left(\frac{1}{2R^2}\|S_t\|_{\bar{V}_t^\dagger}^2\right)}{\left(\det{}^*\left(2\pi P_t^\dagger\right)\right)^{1/2}}\cdot\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{\bar{V}_t^\dagger S_t}{R}\right\|_{\bar{V}_t}^2 - \frac{1}{2}\|\lambda\|_{P_t}^2\right) d\lambda, \qquad (67)
\end{aligned}
$$

where (66) can be verified by expanding all terms and compare, and (67) is from inserting $f_{0,P_t}(\lambda)$ (defined in Eq. 65).

We also have the following computational identity

$$
\left\|\lambda - \frac{\bar{V}_t^\dagger S_t}{R}\right\|_{\bar{V}_t}^2 + \|\lambda\|_{P_t}^2 = \left\|\lambda - \frac{V_t^\dagger S_t}{R}\right\|_{V_t}^2 + \left\|\frac{S_t}{R}\right\|_{\bar{V}_t^\dagger}^2 - \left\|\frac{S_t}{R}\right\|_{V_t^\dagger}^2, \qquad (68)
$$

which can be verified by expanding all terms and using $span(\bar{V}_t) = span(S_t)$.

We can then use (68) in (67) to get

$$
\begin{aligned}
&\frac{\exp\left(\frac{1}{2R^2}\|S_t\|_{\bar{V}_t^\dagger}^2\right)}{\left(\det{}^*\left(2\pi P_t^\dagger\right)\right)^{1/2}}\cdot\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{\bar{V}_t^\dagger S_t}{R}\right\|_{\bar{V}_t}^2 - \frac{1}{2}\|\lambda\|_{P_t}^2\right) d\lambda \\
=&\frac{\exp\left(\frac{1}{2R^2}\|S_t\|_{\bar{V}_t^\dagger}^2\right)}{\left(\det{}^*\left(2\pi P_t^\dagger\right)\right)^{1/2}}\cdot\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{V_t^\dagger S_t}{R}\right\|_{V_t}^2 - \frac{1}{2}\left\|\frac{S_t}{R}\right\|_{\bar{V}_t^\dagger}^2 + \frac{1}{2}\left\|\frac{S_t}{R}\right\|_{V_t^\dagger}^2\right) d\lambda \qquad (69)\\
=&\frac{\exp\left(\frac{1}{2R^2}\|S_t\|_{\bar{V}_t^\dagger}^2\right)}{\left(\det{}^*\left(2\pi P_t^\dagger\right)\right)^{1/2}}\cdot\frac{\exp\left(\frac{1}{2R^2}\|S_t\|_{V_t^\dagger}^2\right)}{\exp\left(\frac{1}{2R^2}\|S_t\|_{\bar{V}_t^\dagger}^2\right)}\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{V_t^\dagger S_t}{R}\right\|_{V_t}^2\right) d\lambda \\
=&\frac{\exp\left(\frac{1}{2R^2}\|S_t\|_{V_t^\dagger}^2\right)}{\left(\det{}^*\left(2\pi P_t^\dagger\right)\right)^{1/2}}\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{V_t^\dagger S_t}{R}\right\|_{V_t}^2\right) d\lambda \\
=&\frac{\exp\left(\frac{1}{2R^2}\|S_t\|_{V_t^\dagger}^2\right)}{\left(\det{}^*\left(2\pi P_t^\dagger\right)\right)^{1/2}}\cdot\left(\det{}^*\left(2\pi V_t^\dagger\right)\right)^{1/2} \\
&\quad\cdot\left[\frac{1}{\left(\det{}^*\left(2\pi V_t^\dagger\right)\right)^{1/2}}\int_{span(P_t)} \exp\left(-\frac{1}{2}\left\|\lambda - \frac{V_t^\dagger S_t}{R}\right\|_{V_t}^2\right) d\lambda\right] \qquad (70)\\
=&\exp\left(\frac{1}{2R^2}\|S_t\|_{V_t^\dagger}^2\right)\frac{(\det{}^*(2\pi V_t^\dagger))^{1/2}}{(\det{}^*(2\pi P_t^\dagger))^{1/2}}, \qquad (71)
\end{aligned}
$$

where (69) uses (68), and the last equation is due to $V_t^\dagger S_t \in span(P_t) = span(V_t)$, and the density of the singular multivariate Gaussian $\mathcal{N}\left(\frac{V_t^\dagger S_t}{R}, V_t^\dagger\right)$ (in Eq. 70) integrates to 1 over $span(V_t)$ (or $span(P_t)$).

Next, let Λ be the singular multi-variate normal random variable $\Lambda \sim \mathcal{N}\left(\frac{V_t^\dagger S_t}{R}, V_t^\dagger\right)$, such that Λ is independent of $\mathcal{F}_\infty$ ($\mathcal{F}_\infty := \sigma(y_1, \eta_1, y_2, \eta_2, y_3, \eta_3, \cdots)$).

Since, by Lemma 8, $\mathbb{E}\left[M_t^{\lambda}\right] \leq 1$ for arbitrary λ, we have

$$\mathbb{E}\left[M_t^{\Lambda}\right] = \mathbb{E}\left[\int_{span(\bar{V}_t)} M_t^{\lambda} f_{0,P_t}(\lambda) d\lambda\right] = \mathbb{E}\left[\exp\left(\frac{1}{2R^2}\left\|S_t\right\|_{V_t^{\dagger}}^2\right)\frac{(\det{}^*(2\pi V_t^{\dagger}))^{1/2}}{(\det{}^*(2\pi P_t^{\dagger}))^{1/2}}\right] \leq 1. \tag{72}$$

Since $V_t^{\dagger}$ (resp. $P_t^{\dagger}$) is positive semi-definite and symmetric, the pseudo-determinant of $V_t^{\dagger}$ (resp. $P_t^{\dagger}$) is the product of the non-zero eigenvalues of $V_t^{\dagger}$ (resp. $P_t^{\dagger}$). Since rank $rank(V_t) = rank(P_t) = m$, we have

$$\frac{\left(\det{}^*\left(2\pi V_t^{\dagger}\right)\right)^{1/2}}{\left(\det{}^*\left(2\pi P_t^{\dagger}\right)\right)^{1/2}} = \sqrt{\frac{\det{}^* P_t}{\det{}^* V_t}} \tag{73}$$

Next, for any $\delta > 0$, we use Markov inequality to get

$$\begin{aligned}
&\mathbb{P}\left(\left\|S_t\right\|_{V_t^{\dagger}}^2 > 2R^2 \log\left(\delta^{-1}\sqrt{\frac{\det{}^* V_t}{\det{}^* P_t}}\right)\right) \\
&= \mathbb{P}\left(\exp\left(\frac{1}{2R^2}\left\|S_t\right\|_{V_t^{\dagger}}^2\right)\frac{\delta(\det{}^*(2\pi V_t^{\dagger}))^{1/2}}{(\det{}^*(2\pi P_t^{\dagger}))^{1/2}} > 1\right) \\
&\leq \mathbb{E}\left[\exp\left(\frac{1}{2R^2}\left\|S_t\right\|_{V_t^{\dagger}}^2\right)\frac{\delta(\det{}^*(2\pi V_t^{\dagger}))^{1/2}}{(\det{}^*(2\pi P_t^{\dagger}))^{1/2}}\right] \\
&= \delta\mathbb{E}\left[\exp\left(\frac{1}{2R^2}\left\|S_t\right\|_{V_t^{\dagger}}^2\right)\frac{(\det{}^*(2\pi V_t^{\dagger}))^{1/2}}{(\det{}^*(2\pi P_t^{\dagger}))^{1/2}}\right] \\
&\leq \delta,
\end{aligned}$$

where the last inequality is due to (72).

□

F Proof of Lemma 3

The proof of Lemma 3 needs two auxiliary results, which are in Appendices F.1 and F.2.

Lemma 3. *Under Assumptions 1-4, with probability at least* $1 - 4K\delta$*, for* $K > K_{\min}$*, where*

$$K_{\min} = 2\max\left\{\frac{\left(H\log\frac{m}{\delta}\right)^4}{\lambda_-^4}, \frac{C_{\max}^2 H\log\frac{d}{\delta}}{\lambda_-^2}\right\},$$

the event $\mathcal{E}_{K,\delta} := \left\{M_* \in \mathcal{C}_* \cap \mathcal{C}_1^{(k)} \cap \mathcal{C}_2^{(k)}, \quad \forall k \in (K_{\min}, K]\right\}$ *holds.*

Proof. To prove Lemma 3, we need Lemma 2, Lemma 10 and Proposition 8. Lemma 2 is proved in Appendix D. Lemma 10 and Proposition 8 are in Appendix F.1 and Appendix F.2 respectively.

Step 1: $M_* \in \mathcal{C}_1^{(k)}$ with high probability. By Lemma 2 and a union bound, with probability at least $1 - 3K\delta$,

$$\left\|P_* - P_k\right\|_2 \leq G_{k,\delta}, \tag{74}$$

for all $k = K_{\min} + 1, \cdots, K$.

By triangle inequality, we have

$$\begin{aligned}
\left\|\left(M_* - M_k\right)\left(I - P_k^{\text{aug}}\right)\right\|_2 &= \left\|\left(M_* - \bar{X}_k^{\text{next}}\bar{Z}_k^{\top}V_k^{\dagger}\right)\left(I - P_k^{\text{aug}}\right)\right\|_2 \\
&\leq \left\|M_*\left(I - P_k^{\text{aug}}\right)\right\|_2 + \left\|\bar{X}_k^{\text{next}}\bar{Z}_k^{\top}V_k^{\dagger}\left(I - P_k^{\text{aug}}\right)\right\|_2.
\end{aligned} \tag{75}$$

Since $V_k(I_{d+d'} - P_k^{\text{aug}}) = 0$, the second term in (75) is zero and we have:

$$\left\|\left(M_* - M_k\right)\left(I - P_k^{\text{aug}}\right)\right\|_2 \leq \left\|M_*\left(I - P_k^{\text{aug}}\right)\right\|_2 \overset{①}{=} \left\|M_*\left(P_*^{\text{aug}} - P_k^{\text{aug}}\right)\right\|_2, \tag{76}$$

where ① uses Proposition 5.

Since $\|M_*\| \leq 1$ (Assumption 1), combining (74) and (76) gives

$$\|(M_* - M_k)(I - P_k^{\text{aug}})\|_2 \leq \|M_*(P_*^{\text{aug}} - P_k^{\text{aug}})\|_2 \leq G_{k,\delta}.$$

Step 2: $M_* \in \mathcal{C}_2^k$ with high probability.

By Proposition 8 (in Appendix F.2) and Lemma 10 (in Appendix F.1), we have, with probability at least $1 - K\delta$,

$$\begin{aligned}
&\left\|(M_* - M_k)V_k^{1/2}\right\|_2^2 \\
\leq &4C^2\|P_* - P_k\|_2^2 Hk + 1 + \left\|\left(P_k W_k \bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_2^2. &(77)\\
\leq &4C^2\|P_* - P_k\|_2^2 Hk + 1 + 2m^2 C_w^2 \log\left(\delta^{-1}\left(1 + \frac{HkC^2}{m}\right)^{1/2}\right), &(78)\\
:= &\beta_{k,\delta}^2
\end{aligned}$$

where (77) uses Proposition 8 (in Appendix F.2) and (78) uses Lemma 10 (in Appendix F.1). This concludes Step 1.

Step 3: $M_* \in \mathcal{C}_*$. By item (1) in Assumption 1 and property (**A**), $M_* \in \mathcal{C}_*$.

The above three steps together conclude the proof.

□

F.1

Lemma 10. *Under Assumption 1, with probability at least* $1 - K\delta$*, for all* $k = 1, 2, 3, \cdots, K$*,*

$$\left\|\left(P_k W_k \bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_F^2 \leq 2m^2 C_w^2 \log\left(\delta^{-1}\left(1 + \frac{HkC^2}{m}\right)^{1/2}\right).$$

Proof. Recall that $P_k = \bar{L}_k \bar{L}_k^\top$. For $j = 1, 2, \cdots, m$, let $l_k^{(j)}$ be the j-th column of $\bar{L}_k$. Note that

$$\begin{aligned}
&\left\|\left(P_k W_k \bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_F^2 &(79)\\
= &tr\left[P_k W_k \bar{Z}_k^\top V_k^\dagger \bar{Z}_k W_k^\top P_k\right] \\
= &\sum_{j=1}^m tr\left[l_k^{(j)}\left(l_k^{(j)}\right)^\top W_k \bar{Z}_k^\top V_k^\dagger \bar{Z}_k W_k^\top l_k^{(j)}\left(l_k^{(j)}\right)^\top\right] \\
= &\sum_{j=1}^m tr\left[\left(l_k^{(j)}\right)^\top W_k \bar{Z}_k^\top V_k^\dagger \bar{Z}_k W_k^\top l_k^{(j)}\right] \\
= &\sum_{j=1}^m \left\|\left(l_k^{(j)}\right)^\top W_k \bar{Z}_k^\top\right\|_{V_k^\dagger}^2. &(80)
\end{aligned}$$

Since $\left\|l_k^{(j)}\right\|_2 = 1$ and $\|w_{k,h}\|_2 \leq C_w$, we know that the entries of $\left(l_k^{(j)}\right)^\top W_k$ are also C_w-sub-Gaussian.

Then by Lemma 9 and union bound, with probability at least $1 - K\delta$, for all $j = 1, 2, \cdots, d$ and all $k = 1, 2, \cdots, K$,

$$\begin{aligned}
\left\|\left(l_k^{(j)}\right)^\top W_k \bar{Z}_k^\top\right\|_{V_k^\dagger}^2 &\le 2C_w^2 \log\left(\delta^{-1}\sqrt{\frac{\det^* V_k}{\det^* P_k}}\right) \\
&\le 2C_w^2 \log\left(\delta^{-1}\sqrt{\det{}^* V_k}\right) \\
&\le 2mC_w^2 \log\left(\delta^{-1}\sqrt{\left(\frac{tr(V_k)}{m}\right)^m}\right) \\
&\le 2mC_w^2 \log\left(\delta^{-1}\left(1+\frac{HkC^2}{m}\right)^{1/2}\right),
\end{aligned}$$

where the second last inequality uses the AM-GM inequality, and that V_k is PSD and has m non-zero eigenvalues, and the last inequality uses $\|\hat{z}_{k,h}\|_2 \le C$ (Proposition 7).

Plugging the above result give into (80) gives: with probability at least $1 - K\delta$,

$$\begin{aligned}
\left\|\left(P_k W_k \bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_F^2 &= \sum_{j=1}^m \left\|\left(l_k^{(j)}\right)^\top W_k \bar{Z}_k^\top\right\|_{V_k^\dagger}^2 \\
&\le 2m^2 C_w^2 \log\left(\delta^{-1}\left(1+\frac{HkC^2}{m}\right)^{1/2}\right),
\end{aligned}$$

for all $k = 1, 2, 3 \cdots, K$. □

F.2

Proposition 8. *Under Assumptions 1 and 2, we have*

$$\left\|(M_* - M_k)V_k^{1/2}\right\|_2^2 \le 4C^2 \left\|P_* - P_k\right\|_2^2 Hk + 1 + \left\|\left(P_k W_k \bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_2^2.$$

Proof. Using

$$\begin{aligned}
M_k^\top &= V_k^\dagger \bar{Z}_k \left(\bar{X}_k^{next}\right)^\top \\
\hat{X}_k^{\text{next}} &= M_* \hat{Z}_k + W_k, \\
\bar{X}_k^{\text{next}} &= P_k \hat{X}_k^{\text{next}}
\end{aligned}$$

we have

$$M_k = \bar{X}_k^{\text{next}} \bar{Z}_k^\top V_k^\dagger = P_k(M_* \hat{Z}_k + W_k)\bar{Z}_k^\top V_k^\dagger.$$

Thus,

$$\begin{aligned}
&(M_* - M_k)V_k(M_* - M_k)^\top \\
&= \left(M_* - P_k(M_* \hat{Z}_k + W_k)\bar{Z}_k^\top V_k^\dagger\right) V_k \left(M_* - P_k(M_* \hat{Z}_k + W_k)\bar{Z}_k^\top V_k^\dagger\right)^\top \\
&\overset{①}{=} \left(M_* V_k - P_k M_* \hat{Z}_k \bar{Z}_k^\top - P_k W_k \bar{Z}_k^\top\right) V_k^\dagger \left(M_* V_k - P_k M_* \hat{Z}_k \bar{Z}_k^\top - P_k W_k \bar{Z}_k^\top\right)^\top,
\end{aligned}$$

where ① can be verified by expanding all terms and comparing.

We can expand the above and get

$$
\begin{aligned}
&\left\|(M_* - M_k)\, V_k^{1/2}\right\|_2^2 \\
=&\left\|(M_* - M_k)\, V_k\, (M_* - M_k)^\top\right\|_2 \\
=&\left\|\left(M_* - P_k(M_*\hat{Z}_k + W_k)\bar{Z}_k^\top V_k^\dagger\right) V_k \left(M_* - P_k(M_*\hat{Z}_k + W_k)\bar{Z}_k^\top V_k^\dagger\right)^\top\right\|_2 \\
=&\Big\| \left(M_* \left(\bar{Z}_k\bar{Z}_k^\top + P_k^{\text{aug}}\right) - P_k M_*\hat{Z}_k\bar{Z}_k^\top - P_k W_k\bar{Z}_k^\top\right) V_k^\dagger \\
&\qquad \cdot \left(M_* \left(\bar{Z}_k\bar{Z}_k^\top + P_k^{\text{aug}}\right) - P_k M_*\hat{Z}_k\bar{Z}_k^\top - P_k W_k\bar{Z}_k^\top\right)^\top \Big\|_2 \\
\overset{①}{\leq}&\left\|\left(M_*\bar{Z}_k\bar{Z}_k^\top - P_k M_*\hat{Z}_k\bar{Z}_k^\top\right) V_k^\dagger \left(M_*\bar{Z}_k\bar{Z}_k^\top - P_k M_*\hat{Z}_k\bar{Z}_k^\top\right)^\top\right\|_2 \\
&\quad + \left\|(M_* P_k^{\text{aug}})\, V_k^\dagger\, (M_* P_k^{\text{aug}})^\top\right\|_2 + \left\|\left(P_k W_k\bar{Z}_k^\top\right) V_k^\dagger \left(P_k W_k\bar{Z}_k^\top\right)^\top\right\|_2 \\
=&\left\|\left(M_*\bar{Z}_k\bar{Z}_k^\top - P_k M_*\hat{Z}_k\bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_2^2 \qquad (81)\\
&\quad + \left\|(M_* P_k^{\text{aug}})\left(V_k^\dagger\right)^{1/2}\right\|_2^2 + \left\|\left(P_k W_k\bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_2^2, \qquad (82)
\end{aligned}
$$

where ① uses a triangle inequality.

For the term in (81), using $M_* = M_* P_*^{\text{aug}}$ (Proposition 5), and $\bar{Z}_k = P_k^{\text{aug}}\hat{Z}_k$, we have

$$
\begin{aligned}
\left\|\left(M_*\bar{Z}_k\bar{Z}_k^\top - P_k M_*\hat{Z}_k\bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_2^2 &= \left\|\left(M_* P_k^{\text{aug}}\hat{Z}_k\bar{Z}_k^\top - P_k M_*\hat{Z}_k\bar{Z}_k^\top\right)\left(V_k^\dagger\right)^{1/2}\right\|_2^2 \\
&\leq \left\|\left(M_* P_k^{\text{aug}} - P_k M_*\right)\hat{Z}_k\bar{Z}_k^\top\left(V_k^\dagger\right)^{1/2}\right\|_2^2 \\
&\leq \left\|M_* P_k^{\text{aug}} - P_k M_*\right\|_2 \left\|\hat{Z}_k\bar{Z}_k^\top\left(V_k^\dagger\right)^{1/2}\right\|_2^2 \qquad (83)
\end{aligned}
$$

Since $V_k = \bar{Z}_k\bar{Z}_k + P_k^{\text{aug}}$, we have $\left\|\bar{Z}_k\left(V_k^\dagger\right)^{1/2}\right\| \leq 1$. Also,

$$
\left\|\hat{Z}_k\right\|_2^2 = \left\|\hat{Z}_k\hat{Z}_k^\top\right\|_2 = \left\|\sum_{k'=1}^{k}\sum_{h=1}^{H}\hat{z}_{k',h}\hat{z}_{k',h}^\top\right\|_2 \leq C^2 Hk,
$$

where the last inequality uses a triangle inequality and $\left\|\hat{z}_{k',h}^\top\right\|_2 \leq C$ (Proposition 7).

Under Assumption 1, $\|M_*\|_2 \leq 1$. Thus we have $\left\|M_* P_k^{\text{aug}}\left(V_k^\dagger\right)^{1/2}\right\|_2^2 \leq 1$.

Collecting relevant terms, we can continue from (82) to get

$$
\left\|(M_* - M_k)\, V_k^{1/2}\right\|_2^2 \leq C^2\left\|M_* P_k^{\text{aug}} - P_k M_*\right\|_2^2 Hk + 1 + \left\|P_k W_k\bar{Z}_k^\top\left(V_k^\dagger\right)^{1/2}\right\|_2^2. \qquad (84)
$$

Then we use $\|M_*\|_2 \leq 1$ (Assumption 1) and Proposition 5 to get

$$
\begin{aligned}
&\left\|M_* P_k^{\text{aug}} - P_k M_*\right\|_2 \\
=&\left\|P_* M_* P_k^{\text{aug}} - P_* M_* P_*^{\text{aug}} + P_* M_* P_*^{\text{aug}} - P_k M_* P_*^{\text{aug}}\right\|_2 \\
=&\left\|P_* M_*(P_k^{\text{aug}} - P_*^{\text{aug}}) + (P_* - P_k) M_* P_*^{\text{aug}}\right\|_2 \\
\leq&2\left\|P_k - P_*\right\|_2.
\end{aligned}
$$

Collecting terms yields the expression in the lemma statement. □

G Proof of Proposition 2

Proposition. *Let $\widetilde{\Psi}_{k,h} := \Psi_h(\widetilde{M}_k)$ computed by (4). Under event $\mathcal{E}_{K,\delta}$ ($K > K_{\min}^2$), we have*

$$\mathrm{Reg}(K) \le \mathcal{O}\left(H\sqrt{K}\right) + \sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left(\Delta_{k,h} + \Delta'_{k,h} + \Delta''_{k,h}\right), \tag{85}$$

where $\Delta_{k,h} := \mathbb{E}\left[J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1})\,|\mathcal{F}_{k,h}\right] - J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1})$, $\Delta'_{k,h} := \|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} - \mathbb{E}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}}|\mathcal{F}_{k,h}\right]$, $\Delta''_{k,h} := \|M_*\hat{z}_{k,h}\|_{\widetilde{\Psi}_{k,h+1}} - \left\|\widetilde{M}_k\hat{z}_{k,h}\right\|_{\widetilde{\Psi}_{k,h+1}}$, and $\mathcal{F}_{k,h}$ is all randomness before time (k,h).

Proof. By boundedness results in Proposition 7, we have the costs $c_{k,h}$ satisfies

$$c_{k,h} = \hat{x}_{k,h}^\top Q_h \hat{x}_{k,h} + u_{k,h}^\top R_h u_{k,h} \le 2C. \tag{86}$$

Thus by (86), for the first $\left\lceil\sqrt{K}\right\rceil$ episodes, we have

$$Reg\left(\left\lceil\sqrt{K}\right\rceil\right) \le \sum_{k=1}^{\lceil\sqrt{K}\rceil} J_h^{\pi_k}(M_*, \hat{x}_{k,h}) \le \mathcal{O}\left(H\sqrt{K}\right). \tag{87}$$

Under the event $\mathcal{E}_{K,\delta}$, we have the optimistic rule (for $k > K_{\min}$):

$$J_1^*\left(\widetilde{M}_k, \hat{x}_{k,1}\right) \le J_1^*(M_*, \hat{x}_{k,1}). \tag{88}$$

Write $\Theta_{k,h} := J_h^{\pi_k}(M_*, \hat{x}_{k,h}) - J_h^*\left(\widetilde{M}_k, \hat{x}_{k,h}\right)$. Then under event $\mathcal{E}_{K,\delta}$, since $K > K_{\min}^2$, the regret (Eq. 6) can be bounded by:

$$\begin{aligned}
\mathrm{Reg}(K) &= \sum_{k=1}^{K} J_1^{\pi_k}(M_*, \hat{x}_{k,1}) - J_1^*(M_*, \hat{x}_{k,1}) \\
&= \sum_{k=1}^{\lceil\sqrt{K}\rceil}\left[J_1^{\pi_k}(M_*, \hat{x}_{k,1}) - J_1^*(M_*, \hat{x}_{k,1})\right] + \sum_{k=\lceil K\rceil+1}^{K}\left[J_1^{\pi_k}(M_*, \hat{x}_{k,1}) - J_1^*(M_*, \hat{x}_{k,1})\right] \\
&\overset{①}{\le} \mathcal{O}\left(\sqrt{HK}\right) + \sum_{k=\lceil K\rceil+1}^{K} J_1^{\pi_k}(M_*, \hat{x}_{k,h}) - J_1^*(M_*, \hat{x}_{k,h}) \\
&\overset{②}{\le} \mathcal{O}\left(\sqrt{HK}\right) + \sum_{k=\lceil K\rceil+1}^{K}\Theta_{k,1},
\end{aligned} \tag{89}$$

where ① uses (87) and ② uses the optimism property (88).

Let $\mathcal{F}_{k,h}$ be the σ-algebra generated by all randomness up to (k,h).

Note. *Next we focus on $k \ge K_{\min} + 1$.*

For simplicity, let $\widetilde{\Psi}_{k,h} := \Psi_h(\widetilde{M}_k)$ and $\widetilde{\psi}_{k,h}$ be the Ψ_h and ψ_h quantities computed with $\widetilde{M}_k$ using (4). Using (1) and (2), we can compute the term $\Delta_{k,h}$ as follows.

$$\begin{aligned}
\Theta_{k,h} = &\|\hat{x}_{k,h}\|_{Q_h} + \|a_{k,h}\|_{R_h} \\
&+ \mathbb{E}\left[J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1})\,|\mathcal{F}_{k,h}\right] \\
&- \|\hat{x}_{k,h}\|_{Q_h} - \|a_{k,h}\|_{R_h} \\
&- \mathbb{E}\left[\hat{x}_{k,h+1}^\top \widetilde{\Psi}_{k,h+1}\hat{x}_{k,h+1}\big|\mathcal{F}_{k,h}\right] - \widetilde{\psi}_{k,h+1},
\end{aligned}$$

which gives,

$$\begin{aligned}
\Theta_{k,h} =&\mathbb{E}\left[J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1})\,|\mathcal{F}_{k,h}\right] - \mathbb{E}\left[\left\|\widetilde{M}_k\hat{z}_{k,h} + w_{k,h+1}\right\|_{\widetilde{\Psi}_{k,h+1}}\middle|\mathcal{F}_{k,h}\right] - \widetilde{\psi}_{k,h+1} \\
=&\Delta_{k,h} + J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1}) - \mathbb{E}\left[\left\|\widetilde{M}_k\hat{z}_{k,h} + w_{k,h+1}\right\|_{\widetilde{\Psi}_{k,h+1}}\middle|\mathcal{F}_{k,h}\right] - \widetilde{\psi}_{k,h+1},
\end{aligned} \tag{90}$$

where

$$\Delta_{k,h} := \mathbb{E}\left[J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right)|\mathcal{F}_{k,h}\right] - J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right).$$

Since noise is independent and mean zero (Assumption 3), we have $\mathbb{E}\left[(\widetilde{M}_k\hat{z}_{k,h})^\top \widetilde{\Psi}_{k,h+1} w_{k,h}|\mathcal{F}_{k,h}\right] = 0$. We then have

$$\begin{aligned}
\Theta_{k,h} =& \Delta_{k,h} + J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right) - \left\|\widetilde{M}_k\hat{z}_{k,h}\right\|_{\widetilde{\Psi}_{k,h+1}} \\
&- \mathbb{E}\left[\|w_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}}|\mathcal{F}_{k,h}\right] - \widetilde{\psi}_{k,h+1} \\
=& \Delta_{k,h} + J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right) - \left\|\widetilde{M}_k\hat{z}_{k,h}\right\|_{\widetilde{\Psi}_{k,h+1}} - \widetilde{\psi}_{k,h+1} \\
&- \mathbb{E}\left[\|\hat{x}_{k,h+1} - M_*\hat{z}_{k,h}\|_{\widetilde{\Psi}_{k,h+1}}|\mathcal{F}_{k,h}\right] \\
=& \Delta_{k,h} + J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right) - \left\|\widetilde{M}_k\hat{z}_{k,h}\right\|_{\widetilde{\Psi}_{k,h+1}} - \widetilde{\psi}_{k,h+1} \\
&- \mathbb{E}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}}|\mathcal{F}_{k,h}\right] + \|M_*\hat{z}_{k,h}\|_{\widetilde{\Psi}_{k,h+1}}. \qquad (91)
\end{aligned}$$

From (5), we know

$$J_h^*\left(\widetilde{M}_k, \hat{x}_{k,h+1}\right) = \|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} + \widetilde{\psi}_{k,h+1}.$$

Thus we can rewrite (91) into

$$\begin{aligned}
\Theta_{k,h} =& \Delta_{k,h} + J_{h+1}^{\pi_k}\left(M_*, \hat{x}_{k,h+1}\right) - J_{h+1}^*\left(\widetilde{M}_k, \hat{x}_{k,h+1}\right) \\
&+ \|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} - \mathbb{E}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}}|\mathcal{F}_{k,h}\right] \\
&+ \|M_*\hat{z}_{k,h}\|_{\widetilde{\Psi}_{k,h+1}} - \left\|\widetilde{M}_k\hat{z}_{k,h}\right\|_{\widetilde{\Psi}_{k,h+1}} \\
=& \Delta_{k,h} + \Delta'_{k,h} + \Delta''_{k,h} + \Theta_{k,h+1},
\end{aligned}$$

where

$$\begin{aligned}
\Delta'_{k,h} &:= \|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} - \mathbb{E}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}}|\mathcal{F}_{k,h}\right], \\
\Delta''_{k,h} &:= \|M_*\hat{z}_{k,h}\|_{\widetilde{\Psi}_{k,h+1}} - \left\|\widetilde{M}_k\hat{z}_{k,h}\right\|_{\widetilde{\Psi}_{k,h+1}}
\end{aligned}$$

Since the cost for $H+1$ and beyond are always zero, we know that the regret can be bounded as:

$$\sum_{k=\lceil\sqrt{K}\rceil+1}^{K} \Theta_{k,1} \leq \sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H}\left(\Delta_{k,h} + \Delta'_{k,h} + \Delta''_{k,h}\right).$$

Plug the above expression into (89) concludes the proof.

□

H Proof of Lemma 4

Lemma 4. *Under Assumptions 1-4, conditioning on $\mathcal{E}_{K,\delta}$ being true, with probability at least $1-2\delta$, we have both*

$$\left|\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\Delta_{k,h}\right| \leq \mathcal{O}\left(\sqrt{KH^3\log\frac{2}{\delta}}\right) \quad \textit{and} \quad \left|\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\Delta'_{k,h}\right| \leq \mathcal{O}\left(\sqrt{HK\log\frac{2}{\delta}}\right).$$

Proof. Fistly, we show that $|\Delta_{k,h}|$ and $|\Delta'_{k,h}|$ are bounded.

Recall $\widetilde{\Psi}_{k,h+1} := \Psi_h(\widetilde{M}_k)$ (computed from Eq. 4). From Assumption 1, Propositions 6 and 7, we know that

$$|\Delta'_{k,h}| = \left| \|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} - \mathbb{E}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} | \mathcal{F}_{k,h} \right] \right| \leq 2C.$$

For the bound of $|\Delta_{k,h}|$, we use an induction argument to bound it. By Proposition 7 and $\|Q_h\|_2 \leq C$ (Assumption 1), we have

$$|J_H^{\pi_k}(M_*, \hat{x}_{k,H})| = \hat{x}_{k,H}^\top Q_H \hat{x}_{k,H} \leq C. \tag{92}$$

Inductively,

$$\begin{aligned}
|J_h^{\pi_k}(M_*, \hat{x}_{k,h})| \leq& \left| \hat{x}_{k,h}^\top Q_h \hat{x}_{k,h} + u_{k,h}^\top R_h u_{k,h} \right| + \left| \mathbb{E}\left[J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1}) | \mathcal{F}_{k,h} \right] \right| \\
\leq& \hat{z}_{k,h}^\top \begin{bmatrix} Q_h & 0 \\ 0 & R_h \end{bmatrix} \hat{z}_{k,h} + (H-h)C \\
\leq& C^3 + (H-h)C \\
=& (H - h + C^2)C
\end{aligned}$$

Next, since the costs at $H+1$ are always zero,

$$|\Delta_{k,h}| = \left| \mathbb{E}\left[J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1}) | \mathcal{F}_{k,h} \right] - J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1}) \right| \leq 2\left(H - h + C^2\right) C.$$

Let $\mathcal{F}_{k,h}$ be all randomness before time (k,h). Then,

$$\begin{aligned}
\mathbb{E}\left[\Delta_{k,h} | \mathcal{F}_{k,h}\right] &= \mathbb{E}\left[J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1}) | \mathcal{F}_{k,h} \right] - \mathbb{E}\left[J_{h+1}^{\pi_k}(M_*, \hat{x}_{k,h+1}) | \mathcal{F}_{k,h} \right] = 0, \\
\mathbb{E}\left[\Delta'_{k,h} | \mathcal{F}_{k,h}\right] &= \mathbb{E}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} | \mathcal{F}_{k,h} \right] - \mathbb{E}\left[\|\hat{x}_{k,h+1}\|_{\widetilde{\Psi}_{k,h+1}} | \mathcal{F}_{k,h} \right] = 0.
\end{aligned}$$

Thus $\{\Delta_{k,h}\}$ and $\{\Delta_{k,h}\}$ are two bounded martingale difference sequence. We can then apply the Azuma's inequality to get the lemma statement. □

I Proof of Lemma 5

Lemma 5. *Let Assumptions 1 and 2 hold. Under event $\mathcal{E}_{K,\delta}$ ($K > K_{\min}$), we have*

$$\sum_{k=K_{\min}+1}^{K} \sum_{h=1}^{H-1} \Delta''_{k,h} \leq \widetilde{\mathcal{O}}\left(\left(H^{5/2} + m^{3/2} H \right) \sqrt{K} \right).$$

Lemma 5 is direct result of Lemma 11 and Lemma 12, which are presented below in Appendices I.1 and I.2 respectively.

I.1

Lemma 11. *Let Assumptions 1 and 2 hold. Under event $\mathcal{E}_{K,\delta}$ ($K > K_{\min}$), we have*

$$\sum_{k=\lceil\sqrt{K}\rceil+1}^{K} \sum_{h=1}^{H} \Delta''_{k,h} \leq \widetilde{\mathcal{O}}\left((H+m)\, \Gamma_K \sqrt{HK} \right),$$

where

$$\Gamma_K := \left[\sum_{k=\sqrt{K}}^{K} \sum_{h=1}^{H} \left\| \left(V_k^\dagger \right)^{1/2} \hat{z}_{k,h} \right\|_2^2 \right]^{1/2}.$$

Proof. We compute

$$\begin{aligned}
&\sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H-1}\Delta''_{k,h} \le \sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H-1}\left|\Delta''_{k,h}\right| \\
=&\sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H-1}\left|\left\|\widetilde{\Psi}_{k,h+1}^{1/2}M_*\hat{z}_{k,h}\right\|_2^2-\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\widetilde{M}_k\hat{z}_{k,h}\right\|_2^2\right| \\
=&\sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H-1}\left|\left\|\widetilde{\Psi}_{k,h+1}^{1/2}M_*\hat{z}_{k,h}\right\|_2-\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\widetilde{M}_k\hat{z}_{k,h}\right\|_2\right|\cdot\left|\left\|\widetilde{\Psi}_{k,h+1}^{1/2}M_*\hat{z}_{k,h}\right\|_2+\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\widetilde{M}_k\hat{z}_{k,h}\right\|_2\right| \\
\le&\left[\sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H-1}\left(\left\|\widetilde{\Psi}_{k,h+1}^{1/2}M_*\hat{z}_{k,h}\right\|_2-\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\widetilde{M}_k\hat{z}_{k,h}\right\|_2\right)^2\right]^{1/2} \qquad (93)\\
&\cdot\left[\sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H-1}\left(\left\|\widetilde{\Psi}_{k,h+1}^{1/2}M_*\hat{z}_{k,h}\right\|_2+\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\widetilde{M}_k\hat{z}_{k,h}\right\|_2\right)^2\right]^{1/2}, \qquad (94)
\end{aligned}$$

where in the last step we use the Cauchy-Schwarz inequality. Under Assumption 1, we have

$$\left[\sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H-1}\left(\left\|\widetilde{\Psi}_{k,h+1}^{1/2}M_*\hat{z}_{k,h}\right\|_2+\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\widetilde{M}_k\hat{z}_{k,h}\right\|_2\right)^2\right]^{1/2}\le 2C^3\sqrt{HK}. \qquad (95)$$

Next, under event $\mathcal{E}_{K,\delta}$ ($K^2 > K_{\min}$), we have

$$\begin{aligned}
&\left|\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\left(M_*\hat{z}_{k,h}\right)\right\|_2-\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\left(\widetilde{M}_k\hat{z}_{k,h}\right)\right\|_2\right|^2 \\
\le&\left\|\widetilde{\Psi}_{k,h+1}^{1/2}\left(\widetilde{M}_k-M_*\right)\hat{z}_{k,h}\right\|_2^2 \\
\overset{①}{\lesssim}&\left\|\left(\widetilde{M}_k-M_*\right)\hat{z}_{k,h}\right\|_2^2 \\
\le&\left\|\left(\widetilde{M}_k-M_*\right)P_k^{\mathrm{aug}}\hat{z}_{k,h}\right\|_2^2+\left\|\left(\widetilde{M}_k-M_*\right)(I-P_k^{\mathrm{aug}})\hat{z}_{k,h}\right\|_2^2, \qquad (96)
\end{aligned}$$

where ① uses boundedness of $\widetilde{\Psi}_{k,h+1}$ (Proposition 6).

Under event $\mathcal{E}_{K,\delta}$, the second term in (96) can be taken care of by the confidence region $\mathcal{C}_2^{(k)}$: for $k > K_{\min}$,

$$\left\|\left(\widetilde{M}_k-M_*\right)(I-P_k^{\mathrm{aug}})\hat{z}_{k,h}\right\|_2^2=\mathcal{O}\left(G_{k,\delta}^2\right)=\mathcal{O}\left(\frac{H}{k}\log(d/\delta)\right), \qquad (97)$$

which sums to $\widetilde{\mathcal{O}}(H^2)$ over k and h.

We now focus on the first term in (96). For this term, we have, under event $\mathcal{E}_{K,\delta}$,

$$\begin{aligned}
\left\|\left(\widetilde{M}_k-M_*\right)P_k^{\mathrm{aug}}\hat{z}_{k,h}\right\|_2^2=&\left\|\left(\widetilde{M}_k-M_*\right)V_k^{1/2}\left(V_k^{\dagger}\right)^{1/2}P_k^{\mathrm{aug}}\hat{z}_{k,h}\right\|_2^2 \\
\le&\left\|\left(\widetilde{M}_k-M_*\right)V_k^{1/2}\right\|_2^2\left\|\left(V_k^{\dagger}\right)^{1/2}P_k^{\mathrm{aug}}\hat{z}_{k,h}\right\|_2^2 \\
\le&2\beta_{k,\delta}\left\|\left(V_k^{\dagger}\right)^{1/2}P_k^{\mathrm{aug}}\hat{z}_{k,h}\right\|_2^2 \qquad (98)\\
\le&2\beta_{k,\delta}\left\|\left(V_k^{\dagger}\right)^{1/2}\hat{z}_{k,h}\right\|_2^2, \qquad (99)
\end{aligned}$$

where in (98) we bound $\left\|\left(\widetilde{M}_k-M_*\right)V_k^{1/2}\right\|_2^2$ using the confidence region $\mathcal{C}_1^{(k)}$.

When $k \in \left[\left\lceil\sqrt{K}\right\rceil, K\right]$, we have

$$\begin{aligned}\beta_{k,\delta} &= 1 + 4C^2 G_{k,\delta}^2 Hk + G'_{k,\delta} \\ &\le 1 + 4C^2 G_{\lceil\sqrt{K}\rceil,\delta}^2 H\left\lceil\sqrt{K}\right\rceil + G'_{K,\delta} \qquad (100) \\ &= \mathcal{O}\left(H^2\log(d/\delta) + m^2\log(HK/\delta)\right), \qquad (101)\end{aligned}$$

where (100) is due to $G_{k,\delta}^2 Hk$ decreases with k for $k \in \left[\left\lceil\sqrt{K}\right\rceil, K\right]$, and $G'_{k,\delta}$ increases with k for $k \in \left[\left\lceil\sqrt{K}\right\rceil, K\right]$.

Combining the above results (94), (95), (97), (101) yields the lemma statement. □

I.2

Lemma 12. *Assume event $\mathcal{E}_{K,\delta}$ holds. Under Assumption 1, we have*

$$\sum_{k=\sqrt{K}}^{K}\sum_{h=1}^{H}\left\|V_k^{-1/2}\hat{z}_{k,h}\right\|_2^2 \le \widetilde{\mathcal{O}}\left(mH + H^2\right),$$

where $V_k^{-1/2} := \left(V_k^{\dagger}\right)^{1/2}$.

Proof. Recall $V_k := P_k^{\text{aug}}\widetilde{V}_k P_k^{\text{aug}}$. We have

$$\begin{aligned}&\left\|V_k^{-1/2}\hat{z}_{k,h}\right\|_2^2 \\ =&\left\|P_k^{\text{aug}}\widetilde{V}_k^{-1/2}P_k^{\text{aug}}\hat{z}_{k,h}\right\|_2^2 \\ =&\Big\|P_k^{\text{aug}}\widetilde{V}_k^{-1/2}P_k^{\text{aug}}\hat{z}_{k,h} - P_K^{\text{aug}}\widetilde{V}_k^{-1/2}P_k^{\text{aug}}\hat{z}_{k,h} + P_K^{\text{aug}}\widetilde{V}_k^{-1/2}P_k^{\text{aug}}\hat{z}_{k,h} \\ &- P_K^{\text{aug}}\widetilde{V}_k^{-1/2}P_K^{\text{aug}}\hat{z}_{k,h} + P_K^{\text{aug}}\widetilde{V}_k^{-1/2}P_K^{\text{aug}}\hat{z}_{k,h}\Big\|_2^2 \\ \le&\left\|\left(P_k^{\text{aug}} - P_K^{\text{aug}}\right)\widetilde{V}_k^{-1/2}P_k^{\text{aug}}\hat{z}_{k,h}\right\|_2^2 + \left\|P_K^{\text{aug}}\widetilde{V}_k^{-1/2}\left(P_k^{\text{aug}} - P_K^{\text{aug}}\right)\hat{z}_{k,h}\right\|_2^2 \\ &+ \left\|P_K^{\text{aug}}\widetilde{V}_k^{-1/2}P_K^{\text{aug}}\hat{z}_{k,h}\right\|_2^2, \qquad (102)\end{aligned}$$

where (the sum over) the third term is bounded by Lemma 13.

Also, under event $\mathcal{E}_{K,\delta}$, Lemma 2 tells us

$$\begin{aligned}\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\|P_K^{\text{aug}} - P_*\|_2^2 &= \mathcal{O}\left(H^2\log(K/\delta)\right), \\ \sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\|P_k^{\text{aug}} - P_*\|_2^2 &= \mathcal{O}\left(H^2\log^2(K/\delta)\right),\end{aligned}$$

which gives

$$\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\|P_K^{\text{aug}} - P_k\|_2^2 = \widetilde{\mathcal{O}}\left(H^2\right).$$

This means

$$\begin{aligned}
&\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left\|\left(P_k^{\text{aug}}-P_K^{\text{aug}}\right)\widetilde{V}_k^{-1/2}P_k^{\text{aug}}\hat{z}_{k,h}\right\|_2^2\\
\leq&\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left\|P_K^{\text{aug}}-P_*\right\|_2^2=\mathcal{O}\left(H^2\log(K/\delta)\right),\\
&\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left\|P_K^{\text{aug}}\widetilde{V}_k^{-1/2}\left(P_k^{\text{aug}}-P_K^{\text{aug}}\right)\hat{z}_{k,h}\right\|_2^2\\
\leq&\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left\|P_k^{\text{aug}}-P_*\right\|_2^2=\mathcal{O}\left(H^2\log^2(K/\delta)\right),
\end{aligned}$$

We can apply the above results to (102), and use Lemma 13 (proved after this) to get

$$\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H}\left\|V_k^{-1/2}\hat{z}_{k,h}\right\|_2^2\leq\widetilde{\mathcal{O}}\left(mH+H^2\right).$$

□

Lemma 13. *Recall*

$$\widetilde{V}_k:=I_{d+d_u}+\sum_{h=1}^{H-1}\sum_{k'=1}^{k-1}\hat{z}_{k',h}\left(\hat{z}_{k',h}\right)^\top.$$

For $k\in[1,K]$, let $V_{K,k}:=P_K^{\text{aug}}\widetilde{V}_kP_K^{\text{aug}}$. Then under Assumption 1, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|V_{K,k}^{-1/2}\hat{z}_{k,h}\right\|_2^2\leq\mathcal{O}\left(mH\log(HK)\right),$$

where $V_{K,k}^{-1/2}:=\left(V_{K,k}^\dagger\right)^{1/2}$.

Proof. Notationally, for any matrix S, we will use $S^{-1/2}$ to denote $\left(S^\dagger\right)^{1/2}$ (when the matrix S is not invertible.) Recall: $\widetilde{V}_K:=I_{d+d_u}+\sum_{h=1}^{H-1}\sum_{k'=1}^{K-1}\hat{z}_{k',h}\hat{z}_{k',h}$.

For the projection matrix P_K^{aug}, let L_K^{aug} be the matrix of m orthonormal columns such that $P_K^{aug}=L_K^{\text{aug}}\left(L_K^{\text{aug}}\right)^\top$.

Define $D_{K,k}:=\left(L_K^{\text{aug}}\right)^\top\widetilde{V}_kL_K^{\text{aug}}$.

Let $\det^*$ be the pseudo-determinant operator, since $P_K^{\text{aug}}\widetilde{V}_kP_K^{\text{aug}}$ is a PSD matrix, $\det^*P_K^{\text{aug}}\widetilde{V}_kP_K^{\text{aug}}$ is the product of positive eigenvalues of $P_K^{\text{aug}}\widetilde{V}_kP_K^{\text{aug}}$, which is the determinant of $\left(L_K^{\text{aug}}\right)^\top\widetilde{V}_kL_K^{\text{aug}}$.

Thus for $1\leq k\leq K$

$$\det{}^*P_K^{\text{aug}}\widetilde{V}_kP_K^{\text{aug}}=\det\left(\left(L_K^{\text{aug}}\right)^\top\widetilde{V}_kL_K^{\text{aug}}\right)=\det\left(\left(L_K^{\text{aug}}\right)^\top\left(\widetilde{V}_{k-1}+\sum_{h=1}^{H-1}\hat{z}_{k-1,h}\hat{z}_{k-1,h}^\top\right)L_K^{\text{aug}}\right).\tag{103}$$

Define

$$\Omega_{K,k}:=D_{K,k-1}^{-1/2}\left(L_K^{\text{aug}}\right)^\top\left(\sum_{h=1}^{H-1}\hat{z}_{k-1,h}\hat{z}_{k-1,h}^\top\right)L_K^{\text{aug}}D_{K,k-1}^{-1/2}.$$

Then (103) can be written as

$$\det{}^*P_K^{\text{aug}}\widetilde{V}_kP_K^{\text{aug}}=\det\left[D_{K,k-1}^{1/2}\left(I_{m+d'}+\Omega_{K,k}\right)D_{K,k-1}^{1/2}\right]=\det\left(D_{K,k-1}\right)\det\left(I_{m+d'}+\Omega_{K,k}\right).\tag{104}$$

Let $\{\lambda_i\}_{i=1,2,\cdots,m+d'}$ be the eigenvalues of $I_{m+d'}+\Omega_{K,k}$, which are at least 1, since $\Omega_{K,k}$ are positive semi-definite. Let $V_{K,k}:=P_K^{\text{aug}}\tilde{V}_kP_K^{\text{aug}}$. From this definition, we have

$$V_{K,k}=P_K^{\text{aug}}V_{K,k}P_K^{\text{aug}}=L_K^{\text{aug}}D_{K,k}\left(L_K^{\text{aug}}\right)^\top \tag{105}$$

For any $k, 1\le k\le K$, we have

$$\begin{aligned}
&\sum_{i=1}^{m+d'}(\lambda_i-1)\\
=&tr\left(I_{m+d'}+\Omega_{K,k}\right)-m-d'\\
=&\sum_{h=1}^{H-1}tr\Big[D_{K,k-1}^{-1/2}\left(L_K^{\text{aug}}\right)^\top\hat{z}_{k-1,h}\hat{z}_{k-1,h}^\top L_K^{\text{aug}}D_{K,k-1}^{-1/2}\Big],
\end{aligned}$$

where the last line uses definition of $\Omega_{K,k-1}$ and linearity of trace operator. Next, by circular property of trace operator,

$$\begin{aligned}
&\sum_{i=1}^{m+d'}(\lambda_i-1)\\
=&\sum_{h=1}^{H-1}tr\Big[L_K^{\text{aug}}D_{K,k-1}^{-1}\left(L_K^{\text{aug}}\right)^\top\hat{z}_{k-1,h}\hat{z}_{k-1,h}^\top\Big]\\
=&\sum_{h=1}^{H-1}tr\left[\widetilde{V}_{K,k-1}^{\dagger}\hat{z}_{k-1,h}\hat{z}_{k-1,h}^\top\right]\\
=&\sum_{h=1}^{H-1}\left\|V_{K,k-1}^{-1/2}\hat{z}_{k-1,h}\right\|_2^2.
\end{aligned}$$

Putting the above results together, we have

$$\begin{aligned}
&\det\left(I_{m+d'}+\Omega_{K,k}\right)=\prod_{i=1}^{m+d'}\lambda_i=\prod_{i=1}^{m+d'}\left[(\lambda_i-1)+1\right]\\
\ge&1+\sum_{i=1}^{m+d'}(\lambda_i-1)=1+\sum_{h=1}^{H-1}\left\|V_{K,k-1}^{-1/2}\hat{z}_{k-1,h}\right\|_2^2,
\end{aligned}$$

which gives

$$\begin{aligned}
&1+\sum_{h=1}^{H-1}\left\|V_{K,k-1}^{-1/2}\hat{z}_{k-1,h}\right\|_2^2\\
=&\frac{\sum_{h=1}^{H-1}\left(1+(H-1)\left\|V_{K,k-1}^{-1/2}\hat{z}_{k-1,h}\right\|_2^2\right)}{H-1}\\
\ge&\prod_{h=1}^{H-1}\left(1+(H-1)\left\|V_{K,k-1}^{-1/2}\hat{z}_{k-1,h}\right\|_2^2\right)^{1/(H-1)}
\end{aligned}$$

Put the above results into (104) to get

$$\begin{aligned}
&\det{}^*P_K^{\text{aug}}\widetilde{V}_KP_K^{\text{aug}}=\det\left(L_K^{\text{aug}}\right)^\top\widetilde{V}_KL_K^{\text{aug}}\\
=&\det\left(D_{K,K-1}\right)\det\left(I_{m+d'}+\Omega_{K,k-1}\right)\\
\ge&\det\left(D_{K,K-1}\right)\prod_{h=1}^{H-1}\left(1+\left\|V_{K,K-1}^{-1/2}\hat{z}_{k-1,h}\right\|_2^2\right)^{1/(H-1)}\\
\ge&\cdots\\
\ge&\prod_{k=1}^{K-1}\prod_{h=1}^{H-1}\left(1+\left\|V_{K,k-1}^{-1/2}\hat{z}_{k-1,h}\right\|_2^2\right)^{1/(H-1)}
\end{aligned}\tag{106}$$

Thus, under Assumption 1, we have

$$\begin{aligned}
\sum_{k=1}^{K-1}\sum_{h=1}^{H-1}\left\|V_{K,k}^{-1/2}\hat{z}_{k,h}\right\|_2^2 \leq & C^2\sum_{k=1}^{K-1}\sum_{h=1}^{H-1}\min\left\{1,\left\|V_{K,k}^{-1/2}\hat{z}_{k,h}\right\|_2^2\right\} \\
\leq & 2C^2\sum_{k=1}^{K-1}\sum_{h=1}^{H-1}\log\left(1+\left\|V_{K,k}^{-1/2}\hat{z}_{k,h}\right\|_2^2\right) \\
= & 2C^2\log\prod_{k=1}^{K-1}\prod_{h=1}^{H-1}\left(1+\left\|V_{K,k}^{-1/2}\hat{z}_{k,h}\right\|_2^2\right) \\
\leq & 2C^2(H-1)\log\det{}^*\left(P_K^{\text{aug}}\widetilde{V}_K P_K^{\text{aug}}\right) \qquad (107) \\
\leq & 2C^2(H-1)\log\left(\frac{tr\left(P_K^{\text{aug}}\widetilde{V}_K P_K^{\text{aug}}\right)}{m+d_u}\right)^{m+d_u} \qquad (108) \\
\leq & 2C^2H(m+d_u)\log\left(1+\frac{CHK}{m+d_u}\right), \qquad (109)
\end{aligned}$$

where (107) uses (106), and (108) uses Cauchy-Swarchz inequality. Finally, since $m=\Theta(d_u)$, we arrive at the lemma statement. □

J Proof of Theorem 1

Theorem 1. *Under Assumptions 1-4, for any $\delta>0$, with probability at least $1-(4K+2)\delta$, the regret for the first K ($K>K_{\min}^2$) rounds satisfies*

$$Reg(K)\leq\widetilde{\mathcal{O}}\left(\left(H^{5/2}+m^{3/2}H\right)\sqrt{K}\log\left(K/\delta\right)\right), \qquad (110)$$

where $\mathcal{O}$ omits (poly)-logarithmic terms.

Proof. Recall by Proposition 2 we have

$$\text{Reg}(K)\leq\mathcal{O}\left(H\sqrt{K}\right)+\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left(\Delta_{k,h}+\Delta'_{k,h}+\Delta''_{k,h}\right)$$

By Lemma 4, we know, condition on $\mathcal{E}_{K,\delta}$ ($K>K_{\min}^2$) being true, with probability at least $1-2\delta$,

$$\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\left(\Delta_{k,h}+\Delta'_{k,h}\right)\leq\mathcal{O}\left(\sqrt{KH^3\log\frac{2}{\delta}}+\sqrt{2HK\log\frac{2}{\delta}}\right).$$

Next, from Lemma 5 we know, under event $\mathcal{E}_{K,\delta}$,

$$\sum_{k=\lceil\sqrt{K}\rceil+1}^{K}\sum_{h=1}^{H-1}\Delta''_{k,h}\leq\widetilde{\mathcal{O}}\left(\left(H^{5/2}+m^{3/2}H\right)\sqrt{K}\right).$$

The above facts together prove the theorem statement. □

K Experiment Details

The state features fed to the systems are (down-sampled) image representations of the state, together with cart velocity and pole tip velocity. The pole tip velocity is clipped to within a certain range. The Q matrix penalizes large velocities, bad cart position and bad pole positions. Since the states contain image information, the state space is high dimensional, while the image clearly has a low-rank representation. A random search is used when finding the optimistic estimate $\widetilde{M}_k$.

All experiments are conducted on a `Dell Precision T3620 Mini Tower` machine with the following configuration.

- Processor: 6th Gen Intel Core i7-6700 (Quad Core 3.40GHz, 4.0Ghz Turbo, 8MB).
- Memory: 32GB (4x8GB) 2133MHz DDR4 Non-ECC.
- Video Card: NVIDIA Quadro K620, 2GB.
- HDD: 256GB SATA Class 20 Solid State Drive.

For software environment, the experiments are conducted using Python 3.6.3 and `OpenAI Gym 0.15.4` Brockman et al. (2016a).