

Rate-matching the regret lower-bound in the linear quadratic regulator with unknown dynamics

Feicheng Wang and Lucas Janson

Department of Statistics, Harvard University

Abstract

The theory of reinforcement learning currently suffers from a mismatch between its empirical performance and the theoretical characterization of its performance, with consequences for, e.g., the understanding of sample efficiency, safety, and robustness. The linear quadratic regulator with unknown dynamics is a fundamental reinforcement learning setting with significant structure in its dynamics and cost function, yet even in this setting there is a gap between the best known regret lower-bound of $O_p(\sqrt{T})$ and the best known upper-bound of $O_p(\sqrt{T} \text{polylog}(T))$. The contribution of this paper is to close that gap by establishing a novel regret upper-bound of $O_p(\sqrt{T})$. Our proof is constructive in that it analyzes the regret of a concrete algorithm, and simultaneously establishes an estimation error bound on the dynamics of $O_p(T^{-1/4})$ which is also the first to match the rate of a known lower-bound. The two keys to our improved proof technique are (1) a more precise upper- and lower-bound on the system Gram matrix and (2) a self-bounding argument for the expected estimation error of the optimal controller.

Keywords— reinforcement learning, linear quadratic regulator, rate-optimal, system identification

1 Introduction

We have witnessed great progress in reinforcement learning (RL) beating human professionals in various challenging games like GO (Silver et al., 2016), Starcraft II (Vinyals et al., 2019) and Dota 2 (Berner et al., 2019). Successes in these highly complex simulation environments have led to an increasing drive to apply RL in real world data-driven systems such as self driving cars (Kiran et al., 2021) and automatic robots (Levine et al., 2016). Yet real-world deployment comes with increased risks and costs, and as such has been hindered by the field’s limited understanding of the gap between theoretical bounds and the empirical performance of RL. One line of attack for this problem is to deepen our understanding of relatively simple yet fundamental systems such as the linear quadratic regulator (LQR) with unknown dynamics.

1.1 Problem statement

In the LQR problem, the system obeys the following dynamics starting from $t = 0$:

$$x_{t+1} = Ax_t + Bu_t + \varepsilon_t, \tag{1}$$

where $x_t \in \mathbb{R}^n$ represents the state of the system at time t and starts at some initial state x_0 , $u_t \in \mathbb{R}^d$ represents the action or control applied at time t , $\varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 I_n)$ is the system noise, and $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times d}$ are matrices determining the system’s linear dynamics. The goal is to find an algorithm U that, at each time t , outputs a control $u_t = U(H_t)$ that is computed using the entire thus-far-observed history of the system $H_t = \{x_t, u_{t-1}, x_{t-1}, \dots, u_1, x_1, u_0\}$ to maximize the system’s

function while minimizing control effort. The cost of the LQR problem up to a given finite time T is quadratic:

$$\mathcal{J}(U, T) = \sum_{t=1}^T (x_t^\top Q x_t + u_t^\top R u_t) \quad (2)$$

for some known positive definite matrices $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{d \times d}$. When the system dynamics A and B are also known and $T \rightarrow \infty$, the cost-minimizing algorithm is known: $u_t^* = U^*(H_t) = K x_t$, where $K \in \mathbb{R}^{d \times n}$ is the efficiently-computable solution to a system of equations that only depend on A , B , Q , and R . Like the Gaussian linear model in supervised learning, the aforementioned linear-quadratic problem is foundational to control theory because it is conceptually simple yet it provides a remarkably good description for some real-world systems. In fact, many systems are close to linear over their normal range of operation, and linearity is an important factor in system design (Recht, 2019).

In this paper we consider the case when the system dynamics A and B are unknown. Intuitively, one might hope that after enough time observing a system controlled by almost any algorithm, one should be able to estimate A and B (and hence K) fairly well and thus be able to apply an algorithm quite close to U^* . Indeed the key challenge in LQR with unknown dynamics, as in any reinforcement learning problem, is to trade off *exploration* (actions that help estimate A and B) with *exploitation* (actions that minimize cost). We will quantify the cost of an algorithm by its *regret*, which is the difference in cost achieved by the algorithm and that achieved by the oracle optimal controller U^* :

$$\mathcal{R}(U, T) = \mathcal{J}(U, T) - \mathcal{J}(U^*, T).$$

The best known upper-bound for the regret of LQR with unknown dynamics is $O_p(\sqrt{T} \text{polylog}(T))$, which contains a polylogarithmic factor of T that is not present in the best known lower-bound of $\Omega_p(\sqrt{T})$. This paper closes that rate gap by establishing a novel regret upper-bound of $O_p(\sqrt{T})$, where the improvement comes from a more careful bound of the system Gram matrix combined with a self-bounding argument for the expected estimation error. As part of our proof, we show that the algorithm that achieves our optimal rate of regret also produces data that can be used for system identification (estimation of A and B) at a rate of $\|\hat{A} - A\|_2 = \|\hat{B} - B\|_2 = O_p(T^{-1/4})$, which is also tighter than the best known bounds of $O_p(T^{-1/4} \text{polylog}(T))$ for data from an algorithm achieving $O_p(\sqrt{T} \text{polylog}(T))$ regret, where the tildes hide polylogarithmic terms in T .

1.2 Related works

Many works have studied optimal rates of regret in RL. In bandits, matching upper- and lower-bounds have been found as $\Theta_p(\log(T))$ for the distribution-dependent regret (Lai and Robbins, 1985; Auer et al., 2002; Magureanu et al., 2014; Agrawal and Goyal, 2013; Komiyama et al., 2015; Garivier et al., 2018) and $\Theta_p(\sqrt{T})$ for the distribution-free regret (Agrawal and Goyal, 2013; Osband and Van Roy, 2016; Garivier et al., 2018; Li et al., 2019; Hajiesmaili et al., 2020).

For Markov decision processes (MDPs), most work has considered finite state and action spaces. In this setting, a matching upper- and lower-bound of $\Theta_p(\log(T))$ is known for the distribution-dependent regret (Burnetas and Katehakis, 1997; Tewari and Bartlett, 2007; Ok et al., 2018; Tirinzoni et al., 2021; Xu et al., 2021), while the best known upper-bound of $O_p(\sqrt{T} \text{polylog}(T))$ for the distribution-free regret Jaksch et al. (2010); Azar et al. (2017); Agrawal and Jia (2017); Simchowitz and Jamieson (2019); Xiong et al. (2021) has a polylogarithmic gap with the best-known lower-bound of $\Omega_p(\sqrt{T})$ (Jaksch et al., 2010; Osband and Van Roy, 2016; Azar et al., 2017).

The LQR system is an MDP with *continuous* state and action spaces, and has received increasing interest recently. For the LQR system with unknown dynamics, Simchowitz and Foster (2020) proved a $\Omega_p(\sqrt{T})$ lower-bound for the regret along with an upper-bound of $O_p(\sqrt{T \log(\frac{1}{\delta})})$ with probability $1 - \delta$ under the condition $\delta < 1/T$, so that the upper-bound contains an implicit additional $\log^{1/2}(T)$

term. Other $O_p(\sqrt{T} \text{polylog}(T))$ regret upper-bounds for LQR with unknown dynamics have been established elsewhere (Faradonbeh et al., 2018a,b; Mania et al., 2019; Abbasi-Yadkori and Szepesvári, 2011; Ibrahimi et al., 2012; Faradonbeh et al., 2017; Cohen et al., 2019; Ouyang et al., 2017; Faradonbeh et al., 2018b; Abeille and Lazaric, 2018; Wang and Janson, 2020), but to the best of our knowledge, no existing work has matched the $\Omega_p(\sqrt{T})$ lower-bound until the present paper. Our proof borrows many insightful results and ideas from a number of these prior works, especially Simchowitz et al. (2018); Fazel et al. (2018); Simchowitz and Foster (2020); Wang and Janson (2020).

1.3 Algorithm and assumptions

Throughout the paper, we make only one assumption on the true system parameters:

Assumption 1 (Stability). *Assume the system is stabilizable, i.e., there exists K_0 such that the spectral radius (maximum absolute eigenvalue) of $A + BK_0$ is strictly less than 1.*

Under Assumption 1, it is well known that there is a unique optimal controller $u_t = Kx_t$ (Arnold and Laub, 1984) which can be computed from A and B , where

$$K = -(R + B^\top PB)^{-1} B^\top PA \quad (3)$$

and P is the unique positive definite solution to the discrete algebraic Riccati equation (DARE):

$$P = A^\top PA - A^\top PB(R + B^\top PB)^{-1} B^\top PA + Q. \quad (4)$$

In this paper we will consider the same algorithm as in Wang and Janson (2020), reproduced here as Algorithm 1, which is a *noisy certainty equivalent control* algorithm. In particular, at every round t , we generate an estimate \hat{K}_t for K , and then apply control $u_t = \hat{K}_t x_t + \eta_t$ as a substitute of the optimal unknown control $u_t = Kx_t$, where $\eta_t \sim \mathcal{N}(0, t^{-1/2} I_d)$ is a noise term whose variance shrinks at a carefully chosen rate in t so as to rate-optimally trade off exploration and exploitation. Note that Algorithm 1 is step-wise and online, i.e., it does not rely on independent restarts or episodes of any kind and does not depend on the time horizon T . The two things it does rely on, which are standard in the literature (see, e.g., Dean et al. (2018)), are the knowledge of a stabilizing controller K_0 and an upper-bound C_K on the spectral norm of the optimal controller K ; C_x and σ_η are also inputs but can take any positive numbers.

Algorithm 1 Stepwise Noisy Certainty Equivalent Control

Require: Initial state x_0 , stabilizing control matrix K_0 , scalars $C_x > 0$, $C_K > \|K\|$, $\sigma_\eta > 0$.

1: Let $u_0 = K_0 x_0 + \eta_0$ and $u_1 = K_0 x_1 + \eta_1$, with $\eta_0, \eta_1 \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2 I_d)$.

2: **for** $t = 2, 3, \dots$ **do**

3: Compute

$$(\hat{A}_{t-1}, \hat{B}_{t-1}) \in \underset{(A', B')}{\operatorname{argmin}} \sum_{k=0}^{t-2} \|x_{k+1} - A' x_k - B' u_k\|^2 \quad (5)$$

and if stabilizable, plug them into the DARE (Eqs. (3) and (4)) to compute \hat{K}_t , otherwise set $\hat{K}_t = K_0$.

4: If $\|x_t\| \gtrsim C_x \log(t)$ or $\|\hat{K}_t\| \gtrsim C_K$, reset $\hat{K}_t = K_0$.

5: Let

$$u_t = \hat{K}_t x_t + \eta_t, \quad \eta_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2 t^{-1/2} I_n) \quad (6)$$

6: **end for**

1.4 Notation

Throughout our proofs, we use $X \lesssim Y$ (resp. $X \gtrsim Y$) as shorthand for the inequality $X \leq CY$ (resp. $X \geq CY$) for some constant C . $X \approx Y$ means both $X \lesssim Y$ and $X \gtrsim Y$. We will almost always establish such relations between quantities that (at least may) depend on T and show that they hold with at least some stated probability $1 - \delta$; in such cases, we will always make all dependence on both T and δ explicit, i.e., the hidden constant(s) C will never depend on T or δ , though they may depend on any other parameters of the system or algorithm, including $A, B, Q, R, \sigma_\epsilon^2, \sigma_\eta^2, K_0, C_x, C_K$.

1.5 Outline

In the remainder of this paper, we will present an outline of the proof of our improved regret upper-bound in two parts. First, in Section 2, we will establish a novel $O_p(T^{-1/4})$ bound on the estimation error of \hat{A}_t, \hat{B}_t , and \hat{K}_t from Algorithm 1. Then, in Section 3, we will leverage this tighter estimation error bound to establish our $O_p(\sqrt{T})$ bound on the regret of Algorithm 1.

2 Bounding the estimation error by $O_p(T^{-1/4})$

Our bound on the estimation error starts with a key result from Simchowitz et al. (2018), which relates the estimation error to the system Gram matrix via a lower- and upper-bound for it. The rest of the proof is primarily comprised of two parts. In the first part, we prove a more precise upper- and lower-bound on the system Gram matrix so that the two bounds are almost of the same order, which is crucial in removing the $\text{polylog}(T)$ in the estimation error bound. In the second part, we take the estimation error bound from plugging in the Gram matrix bounds from the first part and transform it into a self-bounding argument for the expected estimation error of the estimated dynamics that yields the $O_p(T^{-1/4})$ final rate for the estimation error.

To streamline notation, define $z_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}$ and $\Theta = [A, B]$, and correspondingly define $\hat{\Theta}_t = [\hat{A}_t, \hat{B}_t]$. Then by Theorem 2.4 of Simchowitz et al. (2018), given a fixed $\delta \in (0, 1)$, $T \in \mathbb{N}$ and $0 \leq \underline{\Gamma} \leq \bar{\Gamma} \in \mathbb{R}^{(n+d) \times (n+d)}$ such that $\mathbb{P}\left(\sum_{t=0}^{T-1} z_t z_t^\top \succeq T\underline{\Gamma}\right) \geq 1 - \delta$ and $\mathbb{P}\left[\sum_{t=0}^{T-1} z_t z_t^\top \preceq T\bar{\Gamma}\right] \geq 1 - \delta$, when

$$T \gtrsim \log\left(\frac{1}{\delta}\right) + 1 + \log \det(\bar{\Gamma}\underline{\Gamma}^{-1}), \quad (7)$$

$\hat{\Theta}_T$ satisfies:

$$\mathbb{P}\left[\left\|\hat{\Theta}_T - \Theta\right\| \gtrsim \sqrt{\frac{1 + \log \det \bar{\Gamma}\underline{\Gamma}^{-1} + \log\left(\frac{1}{\delta}\right)}{T\lambda_{\min}(\underline{\Gamma})}}\right] \leq \delta. \quad (8)$$

Similar upper-bounds to those that already exist in the literature (which contain extra $\text{polylog}(T)$ terms compared to the best known lower-bound) can be achieved by taking $\underline{\Gamma} \approx T^{-1/2}I_{n+d}$ and $\bar{\Gamma} \approx \log^2(T)I_{n+d}$, and we restate this result here (and prove it in Appendix A.1) for completeness.

Lemma 1 (Estimation error bound with $\text{polylog}(T)$ term). *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, when $0 < \delta < 1/2$, for any $T \gtrsim \log(1/\delta)$,*

$$\mathbb{P}\left[\left\|\hat{\Theta}_T - \Theta\right\| \gtrsim T^{-1/4} \sqrt{\left(\log T + \log\left(\frac{1}{\delta}\right)\right)}\right] \leq \delta. \quad (9)$$

In order to improve this $O_p(T^{-1/4} \log^{1/2}(T))$ bound to the desired $O_p(T^{-1/4})$, we need tighter lower- and upper-bounds $\underline{\Gamma}$ and $\bar{\Gamma}$ for $\sum_{t=0}^{T-1} z_t z_t^\top$. The following Lemma is one of the key steps in our proof.

Lemma 2. Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$ and $T \gtrsim \log^3(1/\delta)$, with probability at least $1 - \delta$:

$$\begin{aligned} T\bar{\Gamma} &:= \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} T^{1/2} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \\ &\lesssim \sum_{t=0}^{T-1} z_t z_t^\top \lesssim \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right) := T\bar{\Gamma}, \end{aligned} \quad (10)$$

where $\Delta_t := (\hat{K}_t - K)x_t + \eta_t$.

The complete proof of Lemma 2 can be found at Appendix A.2.

Proof. (sketch) $G_T := \sum_{t=0}^{T-1} z_t z_t^\top$ can be represented as a summation of two parts:

$$G_T = \sum_{t=0}^{T-1} z_t z_t^\top = \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top + \sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix}. \quad (11)$$

We consider the dominating part $\begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top$ (smallest eigenvalue scales with T) and the remainder part $\sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix}$ separately. We then prove in Lemma 7 that with probability at least $1 - \delta$:

$$\begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top \succeq \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \succeq 1/\delta \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top \quad (12)$$

These bounds reflect the intuition that x_t should converge to a stationary distribution, making each of the summands $x_t x_t^\top$ of constant order.

Lower bound Eq. (12) provides a partial lower bound for G_T : with probability at least $1 - \delta$,

$$G_T \succeq \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \gtrsim \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top. \quad (13)$$

This part only covers the subspace spanned by $\begin{bmatrix} I \\ K \end{bmatrix}$; we still need to consider a general bound for the whole matrix G_T . Lemma 34 of Wang and Janson (2020) gives a high probability lower-bound $G_T \gtrsim T^{1/2} I_{n+d}$. Combining this and Eq. (13), with high probability:

$$G_T + G_T \gtrsim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + T^{1/2} I_{n+d} \gtrsim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} T^{1/2} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top.$$

Upper bound The argument for our upper-bound divides \mathbb{R}^{n+d} into two orthogonal subspaces spanned by the columns of $\begin{bmatrix} I_n \\ K \end{bmatrix}$ and $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$, and essentially bounds $\xi^\top G_T \xi$ separately by order T and $\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right)$ for the two subspaces, respectively. In particular, for any ξ_1 in the span of $\begin{bmatrix} I_n \\ K \end{bmatrix}$ and ξ_2 in the span of $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$,

$$(\xi_1 + \xi_2)^\top G_T (\xi_1 + \xi_2) \leq 2\xi_1^\top G_T \xi_1 + 2\xi_2^\top G_T \xi_2$$

$$\begin{aligned}
& \text{(using Eq. (11), because } \xi_2 \text{ is orthogonal to } \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top) \\
& \lesssim 2\xi_1^\top G_T \xi_1 + 2\|\xi_2\|^2 \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \\
& \text{(we show in Appendix A.8 that } \mathbb{P} \left(G_T \lesssim \frac{1}{\delta} T I_{n+d} \right) \geq 1 - \delta.) \\
& \lesssim \frac{1}{\delta} T \|\xi_1\|^2 + \|\xi_2\|^2 \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right),
\end{aligned}$$

where the last inequality holds with high probability. This last expression can in turn be bounded by

$$(\xi_1 + \xi_2)^\top \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right) (\xi_1 + \xi_2),$$

establishing the upper-bound from Eq. (10). \square

In Lemma 2, the upper bound $\bar{\Gamma}$ and lower bound $\underline{\Gamma}$ have similar forms. Plugging them into Eq. (8) gives that when $T \gtrsim \log^3(1/\delta)$,

$$\mathbb{P} \left[\left\| \hat{\Theta}_T - \Theta \right\| \gtrsim \sqrt{\frac{1 + \log \left(\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) T^{-1/2} \right) + \log \left(\frac{1}{\delta} \right)}{T^{1/2}}} \right] \leq \delta. \quad (14)$$

The following Lemmas 3 and 5 will connect the key term $\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right)$ in the estimation error bound of Eq. (14) with the estimation error itself, setting up the self-bounding argument that is key to our main estimation error bound in Theorem 1.

Lemma 3. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$, $T \gtrsim \log^2(1/\delta)$,*

$$\mathbb{P} \left(\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \gtrsim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) + \log^2(1/\delta) + T^{1/2} \right) \right) \leq 2\delta.$$

A complete proof of Lemma 3 can be found at Appendix A.3.

Proof. (sketch) We first define a “stable” event E_δ , which holds with probability $1 - \delta$, on which for large enough T , the estimation errors are uniformly bounded by some small constant. Intuitively, E_δ is the event on which the system remains well-behaved, in the sense that the system is always well controlled after certain time, which makes our analysis much easier. Lemma 4 defines E_δ and proves that it holds with high probability; its proof is deferred to Appendix A.9, but it basically follows from a union bound applied to Eq. (9).

Lemma 4. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for fixed $\epsilon_0 \lesssim 1$ and any $\delta > 0$,*

$$E_\delta := \left\{ \left\| \hat{\Theta}_T - \Theta \right\|, \left\| \hat{K}_T - K \right\| \leq \epsilon_0, \text{ for all } T \gtrsim \log^2(1/\delta) \right\}, \mathbb{P}(E_\delta) \geq 1 - \delta. \quad (15)$$

Then starting with the inequality

$$\|\Delta_t\|^2 = \left\| (\hat{K}_t - K)x_t + \eta_t \right\|^2 \lesssim \left\| \hat{K}_t - K \right\|^2 \|x_t\|^2 + \|\eta_t\|^2,$$

we can show that with probability $1 - \delta$:

$$\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top 1_{E_\delta} \right) \lesssim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 + t^{-1/2} \|x_t\|^4 1_{E_\delta} \right) + T^{1/2} \right).$$

We prove in Lemma 10 that for $t \gtrsim \log^2(1/\delta)$, $\mathbb{E} \left(\|x_t\|^4 1_{E_\delta} \right) \lesssim 1$. For $t \lesssim \log^2(1/\delta)$ we have the bound $\mathbb{E} \|x_t\|^2 \lesssim \log^2(t)$ from Eq. (104) of Wang and Janson (2020). We show the same proof applies if we increase the exponent from 2 to 4:

$$\mathbb{E} \|x_t\|^4 \lesssim \log^4(t). \quad (16)$$

Applying these bounds produces

$$\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top 1_{E_\delta} \right) \lesssim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) + \log^2(1/\delta) + T^{1/2} \right),$$

and we finish the proof by removing 1_{E_δ} on the left hand side and decreasing the probability with which the inequality holds from $1 - \delta$ to $1 - 2\delta$. \square

Lemma 5. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$ and $T \gtrsim \log^3(1/\delta)$,*

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta. \quad (17)$$

A complete proof of Lemma 5 can be found at Appendix A.4.

Proof. (sketch) Combining Lemma 3 and Eq. (14),

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) + \log^2(1/\delta) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

Then we show that the $\log^2(1/\delta)$ can be moved outside and merged with the $\log(1/\delta)$ term:

$$\begin{aligned} & \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) + \log^2(1/\delta) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \\ & \lesssim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right), \end{aligned}$$

completing the proof. \square

We are now able to state the main result of this section:

Theorem 1. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies*

$$\left\| \hat{\Theta}_T - \Theta \right\| = O_p(T^{-1/4}) \text{ and } \left\| \hat{K}_T - K \right\| = O_p(T^{-1/4}). \quad (18)$$

A complete proof of Theorem 1 can be found at Appendix A.5.

Proof. By Proposition 4 of [Simchowitz and Foster \(2020\)](#),

$$\left\| \hat{K}_T - K \right\| \lesssim \left\| \hat{\Theta}_T - \Theta \right\|. \quad (19)$$

as long as $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$, where ϵ_0 is some fixed constant determined by the system parameters. We want to focus on cases where $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$ to transfer $T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2$ to $T^{1/2} \left\| \hat{K}_T - K \right\|^2$ so that Lemma 5 has only estimation error of \hat{K}_T .

We can estimate $\mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0} \right)$ by calculating the integral using the tail bound from Lemma 5 as long as $T \gtrsim \log^3(1/\delta)$. The further tails can be bounded by the constant ϵ_0 .

As a result, when $T \geq T_0$ (T_0 is a large enough constant so that $3e^{-cT^{1/3}}T\epsilon_0^4 \leq 1$):

$$\begin{aligned} & \mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0} \right) \\ & \lesssim \left(\log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} t^{-1/2} \mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right) \right) \right) + 1 \right)^2 + 3e^{-cT^{1/3}}T\epsilon_0^4 \\ & \lesssim \left(\log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} t^{-1/2} \mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right) \right) \right) + 1 \right)^2 + 1. \end{aligned}$$

On the right-hand side, consider the maximum of $\mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right)$ from T_0 to $T_{\max} \geq T$,

$$\begin{aligned} & \mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0} \right) \\ & \quad (\text{Algorithm 1 ensures } \left\| \hat{K}_t \right\| \leq C_K) \\ & \lesssim \left(\log \left(T^{-1/2} \left(\sum_{t=1}^{T_0} t^{-1/2} (C_K + \|K\|)^2 \right) + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1 \\ & \lesssim \left(\log \left(T^{-1/2} T_0^{1/2} \cdot 1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1 \\ & \lesssim \left(\log \left(1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1 \end{aligned}$$

By Eq. (19), we can transfer $\left\| \hat{\Theta}_T - \Theta \right\|$ on the left hand side to $\left\| \hat{K}_T - K \right\|$ as long as $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$.

By Lemma 1, the probability δ that $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$ does not hold can be solved from

$$T^{-1/4} \sqrt{\left(\log T + \log \left(\frac{1}{\delta} \right) \right)} = \epsilon_0,$$

which gives

$$\delta = Te^{-\epsilon_0^2 T^{1/2}}.$$

As a result (all $T \geq T_0$ satisfies $Te^{-\epsilon_0^2 T^{1/2}}T(C_K + \|K\|)^4 \leq 1$)

$$\mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0} \right) \gtrsim \mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0} \right)$$

$$\begin{aligned}
&\geq \mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 \right) - T e^{-\epsilon_0^2 T^{1/2}} T (C_K + \|K\|)^4 \\
&\geq \mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 \right) - 1.
\end{aligned}$$

Now we have

$$\mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 \right) \lesssim \left(\log \left(1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1$$

The right hand side is constant. Taking the maximum over T from T_0 to T_{\max} on the left hand side:

$$\max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \lesssim \left(\log \left(1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1$$

Thus

$$\max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \lesssim 1.$$

The hidden constant only depends on T_0 , and hence the same inequality holds for *any* T_{\max} :

$$\max_{s \geq T_0} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \lesssim 1.$$

Plugging this back to Eq. (17) gives that when $T \gtrsim \log^3(1/\delta)$,

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T_0} t^{-1/2} \mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

Because $\left\| \hat{K}_t \right\| \leq C_K$, the sum over the first T_0 terms is of negligible order, so that the above equation can be simplified to

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

Thus,

$$\left\| \hat{\Theta}_T - \Theta \right\| = O_p(T^{-1/4}),$$

and $\left\| \hat{K}_T - K \right\| = O_p(T^{-1/4})$ is a direct corollary from Eq. (19). \square

3 Bounding the regret by $O_p(\sqrt{T})$

We start this section by stating the main result of this paper, our regret upper-bound that exactly rate-matches the regret lower-bound of $\Omega(\sqrt{T})$ established in [Simchowitz and Foster \(2020\)](#).

Theorem 2. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies*

$$\mathcal{R}(U, T) = O_p(\sqrt{T}). \quad (20)$$

A complete proof of Theorem 2 can be found at Appendix B.

Proof. (sketch) Our first step is to show the following lemma bounding the cumulative costs \mathcal{J} of the system under Algorithm 1 and under the oracle optimal controller.

Lemma 6. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies,*

$$\mathcal{J}(U, T) = \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + \sum_{t=1}^T \eta_t^\top R \eta_t + O_p\left(T^{1/2}\right)$$

and

$$\mathcal{J}(U^*, T) = \sum_{t=1}^T \varepsilon_t^\top P \varepsilon_t + O_p\left(T^{1/2}\right),$$

where ε_t is the system noise and η_t is the exploration noise in Algorithm 1, and $\tilde{\varepsilon}_t = B\eta_t + \varepsilon_t$.

Before sketching the proof of Lemma 6 (a complete proof can be found in Appendix B.1), we show how to finish the proof of Theorem 2 with just a few more steps:

$$\begin{aligned} \mathcal{R}(U, T) &= \mathcal{J}(U, T) - \mathcal{J}(U^*, T) \\ &= \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + \sum_{t=1}^T \eta_t^\top R \eta_t - \sum_{t=1}^T \varepsilon_t^\top P \varepsilon_t + O_p\left(T^{1/2}\right) \\ &= 2 \sum_{t=1}^T \varepsilon_t^\top P(B\eta_t) + \sum_{t=1}^T (B\eta_t)^\top P(B\eta_t) + \sum_{t=1}^T \eta_t^\top R \eta_t + O_p\left(T^{1/2}\right). \end{aligned}$$

The final result follows by bounding the three summations in the last line by $O_p\left(T^{1/2}\right)$: because $\eta_t = O_p\left(t^{-1/4}\right)$, the quadratic summations $\sum_{t=1}^T (B\eta_t)^\top P(B\eta_t)$ and $\sum_{t=1}^T \eta_t^\top R \eta_t$ are both of order $O_p\left(T^{1/2}\right)$ and the cross term $2 \sum_{t=1}^T \varepsilon_t^\top P(B\eta_t) = o_p\left(T^{1/2}\right)$. \square

Proof. (sketch for Lemma 6) We only prove the first equation because the second equation is a special case of the first equation (with $\eta_t = 0$ and $\hat{K}_t = K$). The idea is to consider a new system with system noise $\tilde{\varepsilon}_t = B\eta_t + \varepsilon_t$ and controller $\tilde{u}_t = \hat{K}_t x_t$. One can show that the new system shares the same states x_t as the original system and the cost in the new system is:

$$\sum_{t=1}^T x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t = \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + O_p\left(T^{1/2}\right). \quad (21)$$

The difference between the original cost and transformed cost is

$$\sum_{t=1}^T u_t^\top R u_t - \tilde{u}_t^\top R \tilde{u}_t = \sum_{t=1}^T \eta_t^\top R \eta_t + o\left(T^{1/4} \log^{\frac{3}{2}}(T)\right) \text{ a.s.} \quad (22)$$

The result of the Lemma follows by summing Eqs. (21) and (22); we now briefly sketch the proofs of each equation. Eqs. (21) and (22) are stated in the Appendix as Lemmas 11 and 12 in the Appendix and their complete proofs are given in Appendices B.1.1 and B.1.2.

Eq. (21): After some substitutions to leverage an identity from Lemma 18 of Wang and Janson (2020) and applying bounds to straightforward terms, the cost of the new system can be written as

$$\begin{aligned} &\sum_{t=1}^T x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t \\ &= \sum_{t=1}^T \left[x_t^\top (\hat{K}_t - K)^\top (R + B^\top P B) (\hat{K}_t - K) x_t + 2 \tilde{\varepsilon}_t^\top P (A + B \hat{K}_t) x_t + \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t \right] + \tilde{O}_p(1). \end{aligned}$$

By Theorem 1, $\|\hat{K}_t - K\| = O_p(T^{-1/4})$, and we can show essentially that x_t is of constant order. This gives that the first sum is of order $\sum_{t=1}^T x_t^\top (\hat{K}_t - K)^\top (R + B^\top PB)(\hat{K}_t - K)x_t = O_p(T^{1/2})$. By noting that $\tilde{\varepsilon}_t \perp P(A + B\hat{K}_t)x_t$ and both $\tilde{\varepsilon}_t$ and $P(A + B\hat{K}_t)x_t$ are of constant order, we can use standard properties of martingales to show $\sum_{t=1}^T \tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t = O_p(T^{1/2})$ as well. Eq. (21) is then just the third sum plus the combination of the aforementioned bounds.

Eq. (22): The difference is expressed as

$$\begin{aligned} \sum_{t=1}^T u_t^\top Ru_t - \tilde{u}_t^\top R\tilde{u}_t &= \sum_{t=1}^T (\hat{K}_t x_t + \eta_t)^\top R(\hat{K}_t x_t + \eta_t) - \sum_{t=1}^T (\hat{K}_t x_t)^\top R(\hat{K}_t x_t) \\ &= 2 \sum_{t=1}^T (\hat{K}_t x_t)^\top R\eta_t + \sum_{t=1}^T \eta_t^\top R\eta_t, \end{aligned}$$

and we simply bound the first term by Eq. (83) of Wang and Janson (2020),

$$\sum_{t=1}^T (\hat{K}_t x_t)^\top R\eta_t = o\left(T^{1/4} \log^{\frac{3}{2}}(T)\right) \text{ a.s.},$$

which completes the proof. \square

4 Discussion

Before we can fully understand the practical performance of RL and deploy it in real-world, high-stakes environments, we need to at least understand it well in the simplest, most structured problem settings. This paper provides progress in that direction by, for the LQR problem with unknown dynamics, proving the first regret upper-bound of $O_p(\sqrt{T})$, exactly matching the rate of the best-known lower-bound of $\Omega_p(\sqrt{T})$ established in Simchowitz and Foster (2020). There are related settings such as non-linear LQR (Kakade et al., 2020) and non-stationary LQR (Luo et al., 2021) whose best known regret upper-bounds are $O_p(\sqrt{T} \text{polylog}(T))$, and we hope our work can shed light on removing the $\text{polylog}(T)$ terms in these settings as well. Finally, for the practical deployment of RL algorithms, the constant factor multiplying the regret rate really matters, so it is our hope that now that the LQR rate is tightly characterized the field can move on to characterizing and tightening the constant in the optimal regret, which we expect will lead to algorithmic innovation as well.

Acknowledgement

The authors are grateful for partial support from NSF CBET-2112085.

References

- 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.
- 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.
- Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR.

- Agrawal, S. and Jia, R. (2017). Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194.
- Arnold, W. F. and Laub, A. J. (1984). Generalized eigenproblem algorithms and software for algebraic riccati equations. *Proceedings of the IEEE*, 72(12):1746–1754.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.
- Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR.
- Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.
- Burnetas, A. N. and Katehakis, M. N. (1997). Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255.
- Cohen, A., Koren, T., and Mansour, Y. (2019). Learning linear-quadratic regulators efficiently with only \sqrt{T} regret. In *International Conference on Machine Learning*, pages 1300–1309.
- Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*.
- Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. (2017). Finite time analysis of optimal adaptive policies for linear-quadratic systems. *arXiv preprint arXiv:1711.07230*.
- Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. (2018a). Input perturbations for adaptive regulation and learning,”. *arXiv preprint arXiv:1811.04258*.
- Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. (2018b). On optimality of adaptive linear-quadratic regulators. *arXiv preprint arXiv:1806.10749*.
- Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476.
- Garivier, A., Hadji, H., Menard, P., and Stoltz, G. (2018). Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv preprint arXiv:1805.05071*.
- Hajiesmaili, M., Talebi, M. S., Lui, J., Wong, W. S., et al. (2020). Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm. *Advances in Neural Information Processing Systems*, 33.
- 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.
- Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4).
- Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. (2020). Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325.

- Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*.
- Komiyama, J., Honda, J., Kashima, H., and Nakagawa, H. (2015). Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pages 1141–1154. PMLR.
- Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.
- Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373.
- Li, Y., Wang, Y., and Zhou, Y. (2019). Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR.
- Luo, Y., Gupta, V., and Kolar, M. (2021). Dynamic regret minimization for control of non-stationary linear dynamical systems. *arXiv preprint arXiv:2111.03772*.
- Magureanu, S., Combes, R., and Proutiere, A. (2014). Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR.
- Mania, H., Tu, S., and Recht, B. (2019). Certainty equivalence is efficient for linear quadratic control. In *Advances in Neural Information Processing Systems*, pages 10154–10164.
- Ok, J., Proutiere, A., and Tranos, D. (2018). Exploration in structured reinforcement learning. *arXiv preprint arXiv:1806.00775*.
- Osband, I. and Van Roy, B. (2016). On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*.
- Ouyang, Y., Gagrani, M., and Jain, R. (2017). Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*.
- Recht, B. (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279.
- 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):484–489.
- Simchowitz, M. and Foster, D. J. (2020). Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*.
- Simchowitz, M. and Jamieson, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32:1153–1162.
- Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR.
- Tewari, A. and Bartlett, P. L. (2007). Optimistic linear programming gives logarithmic regret for irreducible mdps. In *NIPS*, pages 1505–1512. Citeseer.
- Tirinzi, A., Pirotta, M., and Lazaric, A. (2021). A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*.

- Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.
- Wang, F. and Janson, L. (2020). Exact asymptotics for linear quadratic adaptive control. *arXiv preprint arXiv:2011.01364*.
- Xiong, Z., Shen, R., and Du, S. S. (2021). Randomized exploration is near-optimal for tabular mdp. *arXiv preprint arXiv:2102.09703*.
- Xu, H., Ma, T., and Du, S. S. (2021). Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*.

A Proofs from Section 2

A.1 Proof of Lemma 1

Lemma (Bound with log term). *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, when $0 < \delta < 1/2$, for any $T \gtrsim \log(1/\delta)$,*

$$\mathbb{P} \left[\left\| \hat{\Theta}_T - \Theta \right\| \gtrsim T^{-1/4} \sqrt{\left(\log T + \log \left(\frac{1}{\delta} \right) \right)} \right] \leq \delta.$$

As a direct corollary:

$$\left\| \hat{\Theta}_T - \Theta \right\| = O_p \left(T^{-1/4} \log^{1/2}(T) \right).$$

Proof. If we can tolerate the additional log terms, the way of upper bounding the Gram matrix is to use the result from Eq. (104) of Wang and Janson (2020): $\mathbb{E} \|z_t\|^2 \lesssim \log^2(t)$. Inspired by the equation before Proposition 3.1 of Simchowitz et al. (2018), we have: for any random positive semi-definite matrix $M \in \mathbb{R}^{d_M \times d_M}$ with $\mathbb{E}(M) \succ 0$,

$$\begin{aligned} & \mathbb{P} \left(M \not\leq \frac{d_M}{\delta} \mathbb{E}(M) \right) \\ &= \mathbb{P} \left(\lambda_{\max} \left((\mathbb{E}(M))^{-1/2} M (\mathbb{E}(M))^{-1/2} \right) \geq \frac{d_M}{\delta} \right) \\ &\leq \mathbb{E} \left(\lambda_{\max} \left((\mathbb{E}(M))^{-1/2} M (\mathbb{E}(M))^{-1/2} \right) \right) / \frac{d_M}{\delta} \\ &\leq \mathbb{E} \left(\text{Tr} \left((\mathbb{E}(M))^{-1/2} M (\mathbb{E}(M))^{-1/2} \right) \right) / \frac{d_M}{\delta} \\ &= \text{Tr} \left((\mathbb{E}(M))^{-1/2} \mathbb{E}(M) (\mathbb{E}(M))^{-1/2} \right) / \frac{d_M}{\delta} \\ &= \text{Tr} (I_{d_M}) / \frac{d_M}{\delta} \\ &= \delta, \end{aligned}$$

which means

$$\mathbb{P} \left(M \preceq \frac{d_M}{\delta} \mathbb{E}(M) \right) \geq 1 - \delta. \quad (23)$$

Take $M = \sum_{t=0}^{T-1} z_t z_t^\top$:

$$\mathbb{P} \left(\sum_{t=0}^{T-1} z_t z_t^\top \preceq \frac{n+d}{\delta} \sum_{t=0}^{T-1} \mathbb{E} z_t z_t^\top \right) \geq 1 - \delta.$$

On the other hand, we have:

$$\sum_{t=0}^{T-1} \mathbb{E} z_t z_t^\top \preceq \sum_{t=0}^{T-1} \mathbb{E} \|z_t\|^2 I_{n+d} \lesssim T \log^2(T) I_{n+d}.$$

As a result, we can take $\bar{\Gamma} \approx \log^2(T) I_{n+d} / \delta$.

By Theorem 2.4 of Simchowitz et al. (2018), the lower bound condition $\mathbb{P} \left(\sum_{t=0}^{T-1} z_t z_t^\top \succeq T \bar{\Gamma} \right) \geq 1 - \delta$ in Eq. (8) could be replaced by the $(k, \underline{\Gamma}, p)$ -BMSB condition on $\{z_t\}_{t \geq 1}$.

Definition 1 (BMSB condition from (Simchowitz et al., 2018)). *Given an $\{\mathcal{F}_t\}_{t \geq 1}$ -adapted random process $\{x_t\}_{t \geq 1}$ taking values in \mathbb{R}^d , we say that it satisfies the $(k, \underline{\Gamma}, p)$ -matrix block martingale small-ball (BMSB) condition for $\underline{\Gamma} \succ 0$ if, for any unit vector w and $j \geq 0$, $\frac{1}{k} \sum_{i=1}^k \mathbb{P}(|\langle w, x_{j+i} \rangle| \geq \sqrt{w^\top \underline{\Gamma} w} | \mathcal{F}_j) \geq p$ a.s.*

The BMSB condition ensures a lower bound on the independent randomness in each entry of an adapted sequence $\{x_t\}_{t \geq 1}$ given all past history. By Lemma 15 of Wang and Janson (2020), the process $\{z_t\}_{t=0}^{T-1}$ satisfies the

$$(k, \underline{\Gamma}, p) = \left(1, \sigma_\eta^2 T^{-1/2} \min\left(\frac{1}{2}, \frac{\sigma_\varepsilon^2}{2\sigma_\varepsilon^2 C_K^2 + \sigma_\eta^2}\right) I_{n+d}, \frac{3}{10}\right) \text{ BMSB condition.}$$

Thus $\bar{\Gamma} \underline{\Gamma}^{-1} \approx \sqrt{T} \log^2(T) I_{n+d}$ Then Eq. (8) becomes

$$\mathbb{P} \left[\left\| \hat{\Theta}_T - \Theta \right\| \gtrsim \sqrt{\frac{1 + \log \det(\log^2(T) \sqrt{T} I_{n+d} / \delta) + \log\left(\frac{1}{\delta}\right)}{T^{1/2}}} \right] \leq \delta.$$

Here $\log^2(T) \sqrt{T}$ is dominated by T when T is large enough. Also we can hide the constant 1 because $\delta < 1/2$ which implies $\log(1/\delta) \gtrsim 1$. Thus,

$$\mathbb{P} \left[\left\| \hat{\Theta}_T - \Theta \right\| \gtrsim \sqrt{\frac{\log(T) + \log\left(\frac{1}{\delta}\right)}{T^{1/2}}} \right] \leq \delta.$$

The condition for the above equation to hold is Eq. (7):

$$\begin{aligned} T &\gtrsim \log\left(\frac{1}{\delta}\right) + 1 + \log \det(\bar{\Gamma} \underline{\Gamma}^{-1}) \\ &\quad (\text{because } \delta < 1/2 \text{ we can hide } 1) \\ &\gtrsim \log\left(\frac{1}{\delta}\right) + \log \det(\bar{\Gamma} \underline{\Gamma}^{-1}), \end{aligned}$$

which by $\bar{\Gamma} \underline{\Gamma}^{-1} \approx \sqrt{T} \log^2(T) I_{n+d}$ becomes

$$T \gtrsim \log(1/\delta) + \log(T).$$

This condition can be simplified to $T \gtrsim \log(1/\delta)$ because T dominates $\log(T)$. \square

A.2 Proof of Lemma 2

Lemma. Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$ and $T \gtrsim \log^3(1/\delta)$, with probability at least $1 - \delta$:

$$\begin{aligned} T \underline{\Gamma} &:= \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} T^{1/2} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \\ &\gtrsim \sum_{t=0}^{T-1} z_t z_t^\top \gtrsim \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right) := T \bar{\Gamma}, \end{aligned} \quad (24)$$

where $\Delta_t := (\hat{K}_t - K)x_t + \eta_t$.

Proof. Consider the matrix

$$\begin{aligned} G_T &:= \sum_{t=0}^{T-1} z_t z_t^\top \\ &= \sum_{t=0}^{T-1} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top \\ &= \sum_{t=0}^{T-1} \begin{bmatrix} x_t \\ Kx_t + (\hat{K}_t - K)x_t + \eta_t \end{bmatrix} \begin{bmatrix} x_t \\ Kx_t + (\hat{K}_t - K)x_t + \eta_t \end{bmatrix}^\top. \end{aligned} \quad (25)$$

Let

$$\Delta_t = (\hat{K}_t - K)x_t + \eta_t,$$

and then the previous equation can be written as

$$G_T = \sum_{t=0}^{T-1} z_t z_t^\top = \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top + \sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix}.$$

We consider the dominating part $\begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top$ (smallest eigenvalue scales with T) and the remainder part $\sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix}$ separately. This separation is useful because the dominating part purely lies in the subspace spanned by $\begin{bmatrix} I \\ K \end{bmatrix}$, and the remainder part is of smaller order than T . For the dominating part we have

Lemma 7. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$ and $T \gtrsim \log^3(1/\delta)$, with probability at least $1 - \delta$:*

$$\begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top \preceq \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \preceq 1/\delta \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top. \quad (26)$$

The proof can be found at Appendix A.6. By Lemma 34 of Wang and Janson (2020), the process $(z_t)_{t=0}^{T-1}$ satisfies the $(1, \sigma_\eta^2 T^{-1/2} \min(\frac{1}{2}, \frac{\sigma_\epsilon^2}{2\sigma_\epsilon^2 C_K^2 + \sigma_\eta^2}) I_{n+d}, \frac{3}{10})$ -BMSB condition, which guarantees us a lower bound $G_T = \sum_{t=0}^{T-1} z_t z_t^\top \gtrsim T^{1/2} I_{n+d}$. Also by the left hand side of Lemma 7,

$$G_T \succeq \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \succeq \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top.$$

Combining these two equations we have:

Lemma 8 (Lower bound of G_T). *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, when $0 < \delta < 1/2$, for any $T \gtrsim \log^3(1/\delta)$, with probability at least $1 - \delta$:*

$$G_T \gtrsim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + T^{1/2} I_{n+d}. \quad (27)$$

See the proof at Appendix A.7. At first glance this seems wrong because the right hand side of Eq. (27) does not have δ . Actually, the role of δ is present in the constraint $T \gtrsim \log^3(1/\delta)$. The result is not surprising because as T grows larger it becomes exponentially unlikely that G_T can be smaller than, for example, $\frac{1}{2} \mathbb{E} G_T$.

Lemma 9 (Upper bound of G_T). *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, when $0 < \delta < 1/2$, for any $T \gtrsim \log^3(1/\delta)$, with probability at least $1 - \delta$:*

$$G_T \lesssim \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right). \quad (28)$$

Proof can be found at Appendix A.8.

Eq. (24) is a direct corollary of Eq. (27) and Eq. (28).

$$\begin{aligned}
& \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} T^{1/2} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \\
& \lesssim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + T^{1/2} I_{n+d} \\
& \lesssim G_T \\
& \lesssim \frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top.
\end{aligned}$$

The first inequality is because

$$\alpha^\top \left(\left\| \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \right\|^2 I_{n+d} - \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right) \alpha \geq 0.$$

for any $\alpha \in \mathbb{R}^{n+d}$. □

A.3 Proof of Lemma 3

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$ and $T \gtrsim \log^2(1/\delta)$,*

$$\mathbb{P} \left(\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \gtrsim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \|\hat{K}_t - K\|^4 \right) + \log^2(1/\delta) + T^{1/2} \right) \right) \leq 2\delta.$$

Proof. Lemma 4 defines a high probability “stable” event, where when T is large enough, the estimation errors are uniformly bounded by some constant. See the proof at Appendix A.9.

Lemma 10 establishes moment bounds in the “stable” event E_δ .

Lemma 10. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$, $k \in \mathbb{N}$ and $T \gtrsim \log^2(1/\delta)$,*

$$\mathbb{E} \left(\|x_t\|^k \mathbf{1}_{E_\delta} \right) \lesssim 1. \tag{29}$$

See the proof at Appendix A.10.

Notice that

$$\|\Delta_t\| = \left\| (\hat{K}_t - K)x_t + \eta_t \right\| \leq \|\hat{K}_t - K\| \|x_t\| + \|\eta_t\|.$$

Then, with probability $1 - \delta$:

$$\begin{aligned}
& \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \mathbf{1}_{E_\delta} \right) \\
& \leq \mathbf{Tr} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \mathbf{1}_{E_\delta} \right) \\
& \text{(by Markov inequality, this holds with probability } 1 - \delta \text{)} \\
& \leq 1/\delta \mathbb{E} \left(\mathbf{Tr} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \mathbf{1}_{E_\delta} \right) \right) \\
& = 1/\delta \mathbb{E} \left(\sum_{t=0}^{T-1} \|\Delta_t\|^2 \mathbf{1}_{E_\delta} \right)
\end{aligned}$$

$$\begin{aligned}
&\leq 1/\delta \sum_{t=0}^{T-1} \mathbb{E} \left(\left(\|\hat{K}_t - K\| \|x_t\| + \|\eta_t\| \right)^2 1_{E_\delta} \right) \\
&\lesssim 1/\delta \sum_{t=0}^{T-1} \mathbb{E} \left(\left(\|\hat{K}_t - K\|^2 \|x_t\|^2 + \|\eta_t\|^2 \right) 1_{E_\delta} \right) \\
&\text{(Bound } t=0 \text{ separately by constant 1 because } t^{-1/2} \text{ is not well defined)} \\
&\lesssim 1/\delta \left(\sum_{t=1}^{T-1} \left(\mathbb{E} \left(\left(t^{1/2} \|\hat{K}_t - K\|^4 + t^{-1/2} \|x_t\|^4 \right) 1_{E_\delta} \right) + t^{-1/2} \right) + 1 \right) \\
&\lesssim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \|\hat{K}_t - K\|^4 + t^{-1/2} \|x_t\|^4 1_{E_\delta} \right) + T^{1/2} \right).
\end{aligned}$$

By Lemma 10, for $t \gtrsim \log^2(1/\delta)$, $\mathbb{E} \left(\|x_t\|^4 1_{E_\delta} \right) \lesssim 1$. Thus

$$\begin{aligned}
&\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top 1_{E_\delta} \right) \\
&\lesssim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \|\hat{K}_t - K\|^4 \right) + \sum_{t=1}^{C \log^2(1/\delta)} \mathbb{E} \left(t^{-1/2} \|x_t\|^4 \right) + \sum_{t=C \log^2(1/\delta)}^{T-1} t^{-1/2} + T^{1/2} \right) \\
&\text{(By Eq. (16), } \mathbb{E} \|x_t\|^4 \lesssim \log^4(t) \text{)} \\
&\lesssim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \|\hat{K}_t - K\|^4 \right) + \sum_{t=1}^{C \log^2(1/\delta)} t^{-1/2} \log^4(C \log^2(1/\delta)) + T^{1/2} + T^{1/2} \right) \\
&\lesssim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \|\hat{K}_t - K\|^4 \right) + (\log^2(1/\delta))^{1/2} * (2 \log \log(1/\delta) + \log(C))^4 + T^{1/2} \right) \\
&\lesssim 1/\delta \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \|\hat{K}_t - K\|^4 \right) + \log^2(1/\delta) + T^{1/2} \right). \tag{30}
\end{aligned}$$

Finally we finish the proof by removing 1_{E_δ} on the left hand side and changing the probability from $1 - \delta$ to $1 - 2\delta$. \square

A.4 Proof of Lemma 5

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$ and $T \gtrsim \log^3(1/\delta)$,*

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \|\hat{K}_t - K\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta. \tag{31}$$

Proof. Now we can use the refined upper and lower bound in Lemma 2.

$$T\bar{\Gamma} \approx \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} T^{1/2} \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top$$

and

$$T\bar{\Gamma} \approx \frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top.$$

Then

$$T\lambda_{\min}(\underline{\Gamma}) \approx T^{1/2}.$$

Also

$$\det(T\underline{\Gamma}) = \det \left(\begin{bmatrix} I_n & -K^T \\ K & I_d \end{bmatrix} \begin{bmatrix} TI_n & 0 \\ 0 & T^{1/2}I_d \end{bmatrix} \begin{bmatrix} I_n & -K^T \\ K & I_d \end{bmatrix}^\top \right) \approx T^n \cdot (T^{1/2})^d.$$

and similarly

$$\det(T\bar{\Gamma}) = \det \left(\begin{bmatrix} I_n & -K^T \\ K & I_d \end{bmatrix} \begin{bmatrix} TI_n & 0 \\ 0 & \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) I_d \end{bmatrix} \begin{bmatrix} I_n & -K^T \\ K & I_d \end{bmatrix}^\top \right) \approx T^n \cdot \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right)^d.$$

With this particular upper and lower bound choices $\underline{\Gamma}$ and $\bar{\Gamma}$, Eq. (8) can be written as

$$\mathbb{P} \left[\left\| \hat{\Theta}_T - \Theta \right\| \gtrsim \sqrt{\frac{\log \left(\lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) T^{-1/2} \right) + \log \left(\frac{1}{\delta} \right)}{T^{1/2}}} \right] \leq \delta.$$

Then we apply Lemma 3:

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) + \log^2(1/\delta) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

Notice that

$$\begin{aligned} & \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) + \log^2(1/\delta) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \\ & \leq \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + \log^2(1/\delta) + 1 \right) + \log \left(\frac{1}{\delta} \right) \\ & \text{(because } 2ab \geq a + b \text{ when both } a \geq 1 \text{ and } b \geq 1\text{)} \\ & \leq \log \left(2 \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) * \log^2(1/\delta) \right) + \log \left(\frac{1}{\delta} \right) \\ & \lesssim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + 2 \log(\log(1/\delta)) + \log(2) + \log \left(\frac{1}{\delta} \right) \\ & \lesssim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right). \end{aligned}$$

Finally we have

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

□

A.5 Proof of Theorem 1

Theorem. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies*

$$\left\| \hat{\Theta}_T - \Theta \right\| = O_p(T^{-1/4}) \text{ and } \left\| \hat{K}_T - K \right\| = O_p(T^{-1/4}).$$

Proof. By Proposition 4 of [Simchowitz and Foster \(2020\)](#),

$$\left\| \hat{K}_T - K \right\| \lesssim \left\| \hat{\Theta}_T - \Theta \right\|. \quad (32)$$

as long as $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$, where ϵ_0 is some fixed constant determined by the system parameters (this is the same ϵ_0 as in Lemma 4). We want to focus on cases where $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$ to transfer $T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2$ to $T^{1/2} \left\| \hat{K}_T - K \right\|^2$ so that Lemma 5 has only estimation error of K .

We can estimate $\mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0} \right)$ by calculating the integral using the tail bound from Lemma 5 as long as $T \gtrsim \log^3(1/\delta)$. The further tails can be bounded by the constant ϵ_0 . Add an extra $1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0}$ on the left hand side of Eq. (31) inside the probability. When $0 < \delta < 1/2$ and $T \gtrsim \log^3(1/\delta)$,

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0} \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

Denote $a_T = T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 1_{\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0}$, $C_T := T^{-1/2} \left(\sum_{t=1}^{T-1} \mathbb{E} \left(t^{1/2} \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1$. When $T \gtrsim \log^3(1/\delta)$, which is $\delta \geq e^{-(cT)^{1/3}}$, we have

$$\mathbb{P} \left(a_T \geq C \left(\log C_T + \log \left(\frac{1}{\delta} \right) \right) \right) \leq 3\delta. \quad (33)$$

Here c and C are two fixed constants which do not depend on δ and T . Denote the tail bound of a_T corresponding to probability $\delta = e^{-(cT)^{1/3}}$ as $U_T := C \left(\log C_T + (cT)^{1/3} \right)$. When $a_T > U_T$, we bound it by the bound $a_T \leq T^{1/2} \cdot \epsilon_0^2$.

$$\begin{aligned} \mathbb{E} a_T^2 &\leq \int_{s=0}^{U_T} s \mathbb{P} (a_T^2 = s) ds + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\leq - \int_{s=0}^{U_T} s d\mathbb{P} (a_T^2 > s) + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\leq -U_T \mathbb{P} (a_T^2 > U_T) + \int_{s=0}^{U_T} \mathbb{P} (a_T^2 > s) ds + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\quad (\text{because } a_T > 0) \\ &\leq \int_{s=0}^{U_T} \mathbb{P} (a_T > \sqrt{s}) ds + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\leq \int_{s=(C \log(2C_T))^2}^{U_T} \mathbb{P} (a_T > \sqrt{s}) ds + (C \log(2C_T))^2 + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \end{aligned}$$

We choose $s = (C \log(2C_T))^2$ because we only know the bound of $\mathbb{P}(a_T > \sqrt{s})$ up to $\delta = 1/2$ with the restriction $\delta < 1/2$. We know that $s = (C \log(2C_T))^2$ corresponds to $\delta = 1/2$ by Eq. (33).

We also need to express probability δ in terms of the tail value s as $\delta(s)$. Solve the equation

$$\begin{aligned} \sqrt{s} &= C \left(\log C_T + \log \left(\frac{1}{\delta(s)} \right) \right) \\ \implies e^{\sqrt{s}/C} &= C_T \cdot \frac{1}{\delta(s)} \end{aligned}$$

$$\implies \delta(s) = e^{-\sqrt{s}/C} C_T.$$

Thus

$$\begin{aligned} \mathbb{E} a_T^2 &\leq \int_{s=(C \log(2C_T))^2}^{U_T} \mathbb{P}(a_T > \sqrt{s}) ds + (C \log(2C_T))^2 + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\leq \int_{s=(C \log(2C_T))^2}^{U_T} 3\delta(s) ds + (C \log(2C_T))^2 + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\leq \int_{s=(C \log(2C_T))^2}^{\infty} 3e^{-\sqrt{s}/C} C_T ds + (C \log(2C_T))^2 + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\text{(By integral calculation)} \\ &= 3C(C \log(2C_T) + C) + (C \log(2C_T))^2 + 3e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\approx \log(2C_T) + (\log(2C_T))^2 + e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &= (\log(2C_T) + 1/2)^2 - 1/4 + e^{-(cT)^{1/3}} T \epsilon_0^4 \\ &\lesssim (\log(C_T) + 1)^2 + e^{-(cT)^{1/3}} T \epsilon_0^4. \end{aligned}$$

As a result, when $T \geq T_0$ (T_0 is a large enough constant so that $e^{-cT^{1/3}} T \epsilon_0^4 \leq 1$):

$$\begin{aligned} \mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 1_{\|\hat{\Theta}_T - \Theta\| \leq \epsilon_0} \right) &\lesssim \left(\log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} t^{-1/2} \mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right) \right) \right) + 1 \right)^2 + e^{-cT^{1/3}} T \epsilon_0^4 \\ &\lesssim \left(\log \left(T^{-1/2} \left(\sum_{t=1}^{T-1} t^{-1/2} \mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right) \right) \right) + 1 \right)^2 + 1. \end{aligned}$$

On the right hand side, consider the maximum of $\mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right)$ from T_0 to $T_{\max} \geq T$,

$$\begin{aligned} &\mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 1_{\|\hat{\Theta}_T - \Theta\| \leq \epsilon_0} \right) \\ &\lesssim \left(\log \left(T^{-1/2} \left(\sum_{t=1}^{T_0} t^{-1/2} \mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right) + \sum_{t=T_0}^{T-1} t^{-1/2} \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) \right) + 1 \right)^2 + 1 \\ &\text{(Algorithm 1 restricted } \left\| \hat{K}_t \right\| \leq C_K) \\ &\lesssim \left(\log \left(T^{-1/2} \left(\sum_{t=1}^{T_0} t^{-1/2} (C_K + \|K\|)^4 \right) + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1 \\ &\lesssim \left(\log \left(T^{-1/2} T_0^{1/2} \cdot 1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1 \\ &\lesssim \left(\log \left(1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1. \end{aligned}$$

By Eq. (32), we can transfer $\left\| \hat{\Theta}_T - \Theta \right\|$ on the right hand side to $\left\| \hat{K}_T - K \right\|$ as long as $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$. By Lemma 1, the upper bound for the probability δ that $\left\| \hat{\Theta}_T - \Theta \right\| \leq \epsilon_0$ does not hold can be solved from

$$T^{-1/4} \sqrt{\left(\log T + \log \left(\frac{1}{\delta} \right) \right)} = \epsilon_0,$$

which gives

$$\delta = Te^{-\epsilon_0^2 T^{1/2}}.$$

As a result, when $T \geq T_0$:

$$\begin{aligned} & \mathbb{E} \left(T \left\| \hat{\Theta}_T - \Theta \right\|^4 \mathbf{1}_{\|\hat{\Theta}_T - \Theta\| \leq \epsilon_0} \right) \\ & \gtrsim \mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 \mathbf{1}_{\|\hat{\Theta}_T - \Theta\| \leq \epsilon_0} \right) \\ & \geq \mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 \right) - Te^{-\epsilon_0^2 T^{1/2}} T(C_K + \|K\|)^4 \\ & \text{(Choose } T_0 \text{ such that for any } T \geq T_0, Te^{-\epsilon_0^2 T^{1/2}} T(C_K + \|K\|)^4 \leq 1) \\ & \geq \mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 \right) - 1. \end{aligned}$$

Now we have, for any $T_0 \leq T \leq T_{\max}$

$$\mathbb{E} \left(T \left\| \hat{K}_T - K \right\|^4 \right) \lesssim \left(\log \left(1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1.$$

Take maximum across T_0 to T_{\max} on the left hand side:

$$\max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \lesssim \left(\log \left(1 + \max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \right) + 1 \right)^2 + 1.$$

Thus

$$\max_{T_0 \leq s \leq T_{\max}} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \lesssim 1.$$

The hidden constant is only related with T_0 . The same inequality hold for any T_{\max} . As a result,

$$\max_{s \geq T_0} \mathbb{E} \left(s \left\| \hat{K}_s - K \right\|^4 \right) \lesssim 1. \quad (34)$$

Plug this back to Eq. (31). When $T \gtrsim \log^3(1/\delta)$,

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(T^{-1/2} \left(\sum_{t=1}^{T_0} t^{-1/2} \mathbb{E} \left(t \left\| \hat{K}_t - K \right\|^4 \right) \right) + 1 \right) + \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

Because Algorithm 1 restricted that $\left\| \hat{K}_t \right\| \leq C_K$, the initial T_0 items is of negligible order. The above equation can be simplified as

$$\mathbb{P} \left[T^{1/2} \left\| \hat{\Theta}_T - \Theta \right\|^2 \gtrsim \log \left(\frac{1}{\delta} \right) \right] \leq 3\delta.$$

Finally,

$$\left\| \hat{\Theta}_T - \Theta \right\| = O_p(T^{-1/4}).$$

$\left\| \hat{K}_T - K \right\| = O_p(T^{-1/4})$ is a direct corollary from Eq. (32). \square

A.6 Proof of Lemma 7

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$ and $T \gtrsim \log^3(1/\delta)$, with probability at least $1 - \delta$:*

$$\begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top \preceq \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \preceq 1/\delta \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top. \quad (35)$$

High probability upper bound For $t \lesssim \log^2(1/\delta)$, we can use the bound from Eq. (104) in Wang and Janson (2020): $\mathbb{E} \|x_t\|^2 \lesssim \log^2(t)$. For $t \gtrsim \log^2(1/\delta)$ part, the total effect is bounded by $\log^2(1/\delta) \log^2(\log^2(1/\delta)) \lesssim \log^3(1/\delta)$. For $t \gtrsim \log^2(1/\delta)$, we can use Lemma 10: $\mathbb{E} \left(\|x_t\|^2 1_{E_\delta} \right) \lesssim 1$. We then combine the bounds for $t \lesssim \log^2(1/\delta)$ and $t \gtrsim \log^2(1/\delta)$ to get

$$\begin{aligned} & \mathbb{E} \left(\sum_{t=0}^{T-1} x_t x_t^\top 1_{E_\delta} \right) \\ &= \sum_{t=0}^{T-1} \mathbb{E} (x_t x_t^\top 1_{E_\delta}) \\ &\preceq \sum_{t=0}^{T-1} \mathbb{E} \left(\|x_t\|^2 1_{E_\delta} \right) I_n \\ &\preceq (T + \log^3(1/\delta)) I_n. \end{aligned}$$

In order to make the formula neat, require $T \gtrsim \log^3(1/\delta)$, which guarantees the simplified formula

$$\mathbb{E} \left(\sum_{t=0}^{T-1} x_t x_t^\top 1_{E_\delta} \right) \lesssim T \cdot I_n. \quad (36)$$

By Eq. (23) we have:

$$\mathbb{P} \left(\sum_{t=0}^{T-1} x_t x_t^\top 1_{E_\delta} \preceq \frac{d}{\delta} \mathbb{E} \left[\sum_{t=0}^{T-1} x_t x_t^\top 1_{E_\delta} \right] \right) \geq 1 - \delta.$$

Further combine this with Eq. (36):

$$\mathbb{P} \left(\sum_{t=0}^{T-1} x_t x_t^\top 1_{E_\delta} \preceq \frac{C}{\delta} T I_n \right) \geq 1 - \delta.$$

Also we can remove the 1_{E_δ} part by subtracting another δ on the right:

$$\mathbb{P} \left(\sum_{t=0}^{T-1} x_t x_t^\top \lesssim \frac{1}{\delta} T I_n \right) \geq 1 - 2\delta$$

or just hide the constant 2 by $\delta \rightarrow \delta/2$. Now we can conclude that when $T \gtrsim \log^3(1/\delta)$:

$$\mathbb{P} \left(\sum_{t=0}^{T-1} x_t x_t^\top \lesssim \frac{1}{\delta} T I_n \right) \geq 1 - \delta. \quad (37)$$

or, with probability $1 - \delta$,

$$\begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \lesssim 1/\delta \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top.$$

High probability lower bound Next we want to show a high probability lower bound of this term, which is a component in G_T . It is sufficient to prove some BMSB condition because when BMSB condition and high probability upper bounds both hold, the lower bound is also guaranteed (we will illustrate this soon).

Following Definition 1, in order to show the process $\{x_t\}_{t \geq 1}$ satisfies the $(k, \underline{\Gamma}, p) = (1, \sigma_\varepsilon^2 I_n, \frac{3}{10})$ -BMSB condition, we only need to prove that for any $w \in \mathcal{S}^{n-1}$, $\mathbb{P}(|\langle w, x_{j+1} \rangle| \geq \sqrt{w^\top \sigma_\varepsilon^2 I_n w} | \mathcal{F}_j) \geq p$ a.s. Let \mathcal{F}_t be the filtration on all history before time t (including x_t and u_t), we know that

$$x_{t+1} | \mathcal{F}_t \sim \mathcal{N}(Ax_t + Bu_t, \sigma_\varepsilon^2 I_n).$$

We also know the distribution of its inner product with any constant vector w :

$$\langle w, x_{t+1} \rangle | \mathcal{F}_t \sim \mathcal{N}(\langle w, Ax_t + Bu_t \rangle, w^\top \sigma_\varepsilon^2 I_n w).$$

We want to lower bound the probability that the absolute value of this inner product (which follows a normal distribution) is larger than its standard error, which is always lower bounded by the case where the normal distribution is centered at zero. More specifically,

$$\begin{aligned} & \mathbb{P}\left(|\langle w, x_{j+1} \rangle| \geq \sqrt{w^\top \sigma_\varepsilon^2 I_n w} | \mathcal{F}_j\right) \\ & \geq \mathbb{P}\left(|\mathcal{N}(0, w^\top \sigma_\varepsilon^2 I_n w)| \geq \sqrt{w^\top \sigma_\varepsilon^2 I_n w}\right) \\ & \geq 3/10. \end{aligned}$$

The last equation is simply a numerical property of the normal distribution. Now we have proved that the process $\{x_t\}_{t \geq 1}$ follows $(k, \underline{\Gamma}, p) = (1, \sigma_\varepsilon^2 I_n, \frac{3}{10})$ -BMSB condition.

The BMSB-condition is useful in deriving high probability lower bounds. Specifically, assume $X = (x_0, x_1, \dots, x_{T-1})$, then if the Gram matrix $\sum_{t=0}^{T-1} x_t x_t^\top$ has a high probability upper bound, then the BMSB-condition can guarantee a high probability lower bound. In the last equation from section D.1 in Simchowit et al. (2018), it is shown that if $\{x_t\}_{t \geq 1}$ satisfies the $(k, \underline{\Gamma}, p)$ -BMSB condition, then

$$\mathbb{P}\left(\left\{\sum_{t=0}^{T-1} x_t x_t^\top \not\preceq \frac{k \lfloor T/k \rfloor p^2 \underline{\Gamma}}{16}\right\} \cap \left\{\sum_{t=0}^{T-1} x_t x_t^\top \preceq T \bar{\Gamma}\right\}\right) \leq \exp\left\{-\frac{Tp^2}{10k} + 2d \log(10/p) + \log \det(\bar{\Gamma} \underline{\Gamma}^{-1})\right\}. \quad (38)$$

Here $\bar{\Gamma}$ comes from the assumption in Eq. (8) which says $\mathbb{P}[\sum_{t=0}^{T-1} z_t z_t^\top \preceq T \bar{\Gamma}] \geq 1 - \delta$. the upper bound $T \bar{\Gamma}$ is guaranteed by Eq. (37) with $\bar{\Gamma} \simeq 1/\delta I_n$, and the lower bound is just shown to be $\underline{\Gamma} = \sigma_\varepsilon^2 I_n$ with $k = 1$ and $p = 3/10$. Put these representations into the previous equation

$$\mathbb{P}\left(\left\{\sum_{t=0}^{T-1} x_t x_t^\top \not\preceq \frac{\lfloor T \rfloor \left(\frac{3}{10}\right)^2 \underline{\Gamma}}{16}\right\} \cap \left\{\sum_{t=0}^{T-1} x_t x_t^\top \preceq T \bar{\Gamma}\right\}\right) \leq \exp\left\{-\frac{T \left(\frac{3}{10}\right)^2}{10} + C + d \log(1/\delta)\right\}.$$

Here C is some constant independent of δ and T . To make the right hand side smaller than δ , the condition is $\exp\left\{-\frac{9T}{1000} + C + d \log(1/\delta)\right\} < \delta$, which means

$$\frac{9T}{1000} - C - d \log(1/\delta) > \log(1/\delta),$$

which is just $T \gtrsim \log(1/\delta)$. With this condition, we have

$$\mathbb{P}\left(\left\{\sum_{t=0}^{T-1} x_t x_t^\top \not\preceq \frac{9 \lfloor T \rfloor \underline{\Gamma}}{1600}\right\} \cap \left\{\sum_{t=0}^{T-1} x_t x_t^\top \preceq T \bar{\Gamma}\right\}\right) \leq \delta.$$

which is

$$\mathbb{P}\left(\left\{\sum_{t=0}^{T-1} x_t x_t^\top \not\preceq \frac{9 \lfloor T \rfloor \sigma_\varepsilon^2 I_n}{1600}\right\} \cap \left\{\sum_{t=0}^{T-1} x_t x_t^\top \preceq \frac{1}{\delta} T I_n\right\}\right) \leq \delta.$$

We can exclude the later event and change the probability on the right hand side to $\delta + \delta = 2\delta$. When $T \gtrsim \log(1/\delta)$,

$$\mathbb{P}\left(\left\{\sum_{t=0}^{T-1} x_t x_t^\top \not\asymp \frac{9\lfloor T \rfloor \sigma_\varepsilon^2 I_n}{1600}\right\}\right) \leq 2\delta.$$

We can change 2δ to δ , and the constraint is still $T \gtrsim \log(1/\delta)$.

$$\mathbb{P}\left(\left\{\sum_{t=0}^{T-1} x_t x_t^\top \not\asymp \frac{9\lfloor T \rfloor \sigma_\varepsilon^2 I_n}{1600}\right\}\right) \leq \delta. \quad (39)$$

Now with probability $1 - 2\delta$ (one δ from upper bound Eq. (37), another δ from lower bound Eq. (39)) we have both upper and lower bound of

$$T I_n \lesssim \sum_{t=0}^{T-1} x_t x_t^\top \lesssim \frac{1}{\delta} T I_n,$$

and

$$\begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top \lesssim \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \lesssim 1/\delta \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top.$$

when $T \gtrsim \log^3(1/\delta)$. We can replace δ with $\delta/2$ so that $1 - 2\delta$ becomes $1 - \delta$. ■

A.7 Proof of Lemma 8

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, when $0 < \delta < 1/2$, for any $T \gtrsim \log^3(1/\delta)$, with probability at least $1 - \delta$,*

$$G_T \lesssim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + T^{1/2} I_{n+d}.$$

By definition Eq. (25), $G_T - \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top \succeq 0$, thus by Lemma 7

$$G_T \lesssim \begin{bmatrix} I \\ K \end{bmatrix} T \begin{bmatrix} I \\ K \end{bmatrix}^\top. \quad (40)$$

This lower bound is growing linearly with T but still is low rank, so we combine this with another lower bound which is full rank but grows sub-linearly with T .

By Lemma 34 of Wang and Janson (2020), the process $(z_t)_{t=0}^{T-1}$ satisfies the $(1, \sigma_\eta^2 T^{-1/2} I_{n+d}, \frac{3}{10})$ -BMSB condition. Here $z_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}$. Now we only need an upper bound to guarantee the lower bound using BMSB condition. Again by Eq. (23), we have

$$\mathbb{P}\left(\sum_{t=0}^{T-1} z_t z_t^\top \not\leq \frac{n+d}{\delta} \mathbb{E}\left[\sum_{t=0}^{T-1} z_t z_t^\top\right]\right) \leq \delta.$$

Also by Eq. (104) from Wang and Janson (2020), $\mathbb{E}\|z_t\|^2 \lesssim \log^2(t)$. Thus $\mathbb{E}\left[\sum_{t=0}^{T-1} z_t z_t^\top\right] \lesssim \log^2(T)T$, we have

$$\mathbb{P}\left(\sum_{t=0}^{T-1} z_t z_t^\top \not\leq \frac{C}{\delta} \log^2(T) T I_{n+d}\right) \leq \delta.$$

Now that we have an upper bound, similar to Eq. (39), we can get the BMSB implied lower bound with $\underline{\Gamma} \approx T^{-1/2}I_{n+d}$: when $T \gtrsim \log(1/\delta)$, with probability at least $1-\delta$, $G_T = \sum_{t=0}^{T-1} z_t z_t^\top \gtrsim T^{1/2}I_{n+d}$. Combining this and Eq. (40), with probability at least $1-\delta$:

$$G_T + G_T \gtrsim \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + T^{1/2}I_{n+d}.$$

We derived the lower bound, and we hope that the upper bound can have a similar form.

A.8 Proof of Lemma 9

Lemma (Upper bound of G_T). *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, when $0 < \delta < 1/2$, for any $T \gtrsim \log^3(1/\delta)$, with probability at least $1-\delta$:*

$$G_T \lesssim \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right). \quad (41)$$

Proof. Any vector $\xi \in \mathbb{R}^{n+d}$ can be represented as the summation of vectors $\xi_1 \in \mathbb{R}^{n+d}$ and $\xi_2 \in \mathbb{R}^{n+d}$ from orthogonal subspaces spanned by the columns of $\begin{bmatrix} I_n \\ K \end{bmatrix}$ and $\begin{bmatrix} -K^\top \\ I_d \end{bmatrix}$. We only need to show that, for any $\xi_1 = \begin{bmatrix} I_n \\ K \end{bmatrix} \alpha_1$ and $\xi_2 = \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \alpha_2$ (with $\alpha_1 \in \mathbb{R}^n$, $\alpha_2 \in \mathbb{R}^d$), for any $T \gtrsim \log^3(1/\delta)$, with probability at least $1-\delta$, we have

$$(\xi_1 + \xi_2)^\top G_T (\xi_1 + \xi_2) \lesssim (\xi_1 + \xi_2)^\top \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right) (\xi_1 + \xi_2).$$

We then show this inequality by the following two inequalities:

1. When $T \gtrsim \log^3(1/\delta)$, with probability at least $1-\delta$:

$$\begin{aligned} & (\xi_1 + \xi_2)^\top G_T (\xi_1 + \xi_2) \\ & \leq 2\xi_1^\top G_T \xi_1 + 2\xi_2^\top G_T \xi_2 \\ & \text{(because } \xi_2 \text{ is orthogonal to } \begin{bmatrix} I \\ K \end{bmatrix} \sum_{t=0}^{T-1} x_t x_t^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top) \\ & = 2\xi_1^\top G_T \xi_1 + 2\xi_2^\top \sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix} \xi_2 \\ & = 2\xi_1^\top G_T \xi_1 + 2 \left(\begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \alpha_2 \right)^\top \sum_{t=0}^{T-1} \begin{bmatrix} 0_n & x_t \Delta_t^\top \\ \Delta_t x_t^\top & \Delta_t \Delta_t^\top + K x_t \Delta_t^\top + \Delta_t x_t^\top K^\top \end{bmatrix} \left(\begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \alpha_2 \right) \\ & = 2\xi_1^\top G_T \xi_1 + 2(\alpha_2)^\top \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) (\alpha_2) \\ & \leq 2\xi_1^\top \sum_{t=0}^{T-1} z_t z_t^\top \xi_1 + 2\|\alpha_2\|^2 \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \\ & \text{(Similar to Eq. (37), when } T \gtrsim \log^3(1/\delta), \mathbb{P} \left(\sum_{t=0}^{T-1} z_t z_t^\top \lesssim \frac{1}{\delta} T I_{n+d} \right) \geq 1-\delta.) \\ & \lesssim \frac{1}{\delta} T \|\xi_1\|^2 + \|\xi_2\|^2 \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right). \end{aligned}$$

In the last step for bounding $\sum_{t=0}^{T-1} z_t z_t^\top$, most of the steps are the same as Eq. (37). The only differences are:

- (a) Replacing the x_t with z_t .
- (b) Replacing the dimension d of x_t with the dimension $n + d$ of z_t .

We also need two other properties

- (a) $\mathbb{E} \|z_t\|^2 \lesssim \log^2(t)$, which is proved by Eq. (104) in Wang and Janson (2020).
- (b) For $t \gtrsim \log^2(1/\delta)$, we can use Lemma 10's conclusion $\mathbb{E} \left(\|x_t\|^2 \mathbf{1}_{E_\delta} \right) \lesssim 1$ to prove $\mathbb{E} \left(\|z_t\|^2 \mathbf{1}_{E_\delta} \right) \lesssim 1$. Recall that $u_t = \hat{K}_t x_t + \eta_t$.

$$\begin{aligned} \|z_t\|^2 &= \|x_t\|^2 + \|u_t\|^2 \\ &\leq \|x_t\|^2 + 2 \left\| \hat{K}_t x_t \right\|^2 + 2 \|\eta_t\|^2 \\ &\leq (1 + 2C_K^2) \|x_t\|^2 + 2 \|\eta_t\|^2. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E} \left(\|z_t\|^2 \mathbf{1}_{E_\delta} \right) &\leq (1 + 2C_K^2) \mathbb{E} \left(\|x_t\|^2 \mathbf{1}_{E_\delta} \right) + 2 \mathbb{E} \left(\|\eta_t\|^2 \right) \\ &\lesssim 1. \end{aligned}$$

2.

$$\begin{aligned} &(\xi_1 + \xi_2)^\top \left(\frac{1}{\delta} \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top + \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \right) (\xi_1 + \xi_2) \\ &= \frac{1}{\delta} \xi_1^\top \begin{bmatrix} I_n \\ K \end{bmatrix} T \begin{bmatrix} I_n \\ K \end{bmatrix}^\top \xi_1 + \xi_2^\top \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \begin{bmatrix} -K^\top \\ I_d \end{bmatrix}^\top \xi_2 \\ &= \frac{1}{\delta} \alpha_1^\top (I_n + K^\top K) T (I_n + K^\top K) \alpha_1 + \alpha_2^\top (I_d + K K^\top) \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) (I_d + K K^\top) \alpha_2 \\ &\text{(because } I_n + K^\top K \succeq I_n) \\ &\geq \frac{1}{\delta} \alpha_1^\top T \alpha_1 + \alpha_2^\top \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \alpha_2 \\ &= \frac{1}{\delta} \|\alpha_1\|^2 T + \|\alpha_2\|^2 \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \\ &\gtrsim \frac{1}{\delta} \|\alpha_1\|^2 \left\| \begin{bmatrix} I_n \\ K \end{bmatrix} \right\|^2 T + \|\alpha_2\|^2 \left\| \begin{bmatrix} -K^\top \\ I_d \end{bmatrix} \right\|^2 \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right) \\ &\gtrsim \frac{1}{\delta} T \|\xi_1\|^2 + \|\xi_2\|^2 \lambda_{\max} \left(\sum_{t=0}^{T-1} \Delta_t \Delta_t^\top \right). \end{aligned}$$

We complete the proof by combining these two inequalities which have identical right hand side. \square

A.9 Proof of Lemma 4

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for fixed $\epsilon_0 \lesssim 1$ and any $\delta > 0$,*

$$E_\delta := \left\{ \left\| \hat{\Theta}_T - \Theta \right\|, \left\| \hat{K}_T - K \right\| \leq \epsilon_0, \text{ for all } T \gtrsim \log^2(1/\delta) \right\}, \mathbb{P}(E_\delta) \geq 1 - \delta. \quad (42)$$

Proof. By Lemma 1, replacing δ by δ/T^2 , the condition on T becomes $T \gtrsim \log(T^2/\delta)$, which is $T \gtrsim \log(1/\delta)$, we have

$$\mathbb{P} \left[\left\| \hat{\Theta}_T - \Theta \right\| \gtrsim T^{-1/4} \sqrt{\log T + \log(T^2/\delta)} \right] \leq \delta/T^2.$$

Since $\sum_{T=2}^{\infty} 1/T^2 < \infty$, we can sum $T \gtrsim \log(1/\delta)$ these equations up:

$$\mathbb{P} \left[\text{Exists } T \gtrsim \log(1/\delta), \left\| \hat{\Theta}_T - \Theta \right\| \gtrsim T^{-1/4} \sqrt{3 \log t + \log(1/\delta)} \right] \leq \delta \sum_{T=2}^{\infty} 1/T^2.$$

Let new $\delta = 3\delta \sum_{T=2}^{\infty} 1/T^2$, and we can hide the constants. As a result, we still have

$$\mathbb{P} \left[\text{Exists } T \gtrsim \log(1/\delta), \left\| \hat{\Theta}_T - \Theta \right\| \gtrsim T^{-1/4} \sqrt{\log T + \log(1/\delta)} \right] \leq \delta.$$

We need a uniform upper bound ϵ_0 on $\left\| \hat{\Theta}_T - \Theta \right\|$. Take ϵ_0 that satisfies: $\|B\| \epsilon_0 < 1 - \frac{1+\rho(A+BK)}{2}$. Here $\rho(\cdot)$ is the spectral radius function. This choice of ϵ_0 is to make sure even after perturbation, the system controlled by \hat{K} is still “stable”. Here we use quotes on stable as stability is not satisfied by the perturbation bound itself because $\rho(A + B\hat{K}_T) \leq \rho(A + BK) + \rho(B(\hat{K}_T - K))$ does not hold, but this condition serves similar utility as stability as we will see in Appendix A.10. We need to find the condition for T to satisfy:

$$\begin{aligned} T^{-1/4} \sqrt{\log t + \log(1/\delta)} &\lesssim \epsilon_0. \\ T^{-1/2} (\log t + \log(1/\delta)) &\lesssim \epsilon_0^2. \\ \epsilon_0^{-4} (\log T + \log(1/\delta))^2 &\lesssim T. \end{aligned}$$

Because T dominates $\log(T)$, and we can hide constant ϵ_0^4 , the final equation can be simplified to

$$T \gtrsim \log^2(1/\delta).$$

Now we replace $T^{-1/4} \sqrt{\log T + \log(1/\delta)}$ with ϵ_0 , and take the complement of the whole event:

$$\mathbb{P} \left[\text{For all } T \gtrsim \log^2(1/\delta), \left\| \hat{\Theta}_T - \Theta \right\| \lesssim \epsilon_0 \right] \leq 1 - \delta.$$

By Eq. (19) we can also control $\left\| \hat{K}_T - K \right\|$ along with $\left\| \hat{\Theta}_T - \Theta \right\|$. With probability $1 - \delta$, we have the following event holds:

$$E_\delta := \left\{ \left\| \hat{\Theta}_T - \Theta \right\|, \left\| \hat{K}_T - K \right\| \leq \epsilon_0, \text{ for all } T \gtrsim \log^2(1/\delta) \right\}.$$

□

A.10 Proof of Lemma 10

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies, for any $0 < \delta < 1/2$, $k \in \mathbb{N}$ and $T \gtrsim \log^2(1/\delta)$,*

$$\mathbb{E} \left(\|x_t\|^k 1_{E_\delta} \right) \lesssim 1.$$

Proof. We know from the proof of Lemma 19 from Wang and Janson (2020) that for any $m > 0$:

$$x_{t+m} = \sum_{p=t}^{t+m-1} (A + B\hat{K}_{t+m-1}) \cdots (A + B\hat{K}_{p+1})(B\eta_p + \varepsilon_p) + (A + B\hat{K}_{t+m-1}) \cdots (A + B\hat{K}_t)x_t. \quad (43)$$

By Lemma 43 from Wang and Janson (2020), as long as \hat{K}_t is consistent, the norm of such product $(A + B\hat{K}_{t+m-1}) \cdots (A + B\hat{K}_{p+1})$ is decaying exponentially fast. More specifically, denote $L := A + BK$, which by Assumption 1 satisfies $\rho(L) < 1$. Further define

$$\tau(L, \rho) := \sup \{ \|L^k\| \rho^{-k} : k \geq 0 \}.$$

The proof of Lemma 43 of Wang and Janson (2020) showed that

$$\begin{aligned} & \left\| (A + B\hat{K}_{t+m-1}) \cdots (A + B\hat{K}_{p+1}) \right\| \\ & \leq \tau \left(L, \frac{1 + \rho(L)}{2} \right) \left(\frac{1 + \rho(L)}{2} + \|B(\hat{K}_{t+m-1} - K)\| \right) \cdots \left(\frac{1 + \rho(L)}{2} + \|B(\hat{K}_p - K)\| \right). \end{aligned}$$

By Eq. (15), under event E_δ , when $t \gtrsim \log^2(1/\delta)$, the difference $\hat{K}_t - K$ is uniformly bounded by ϵ_0 . Denote $\rho_0 = \frac{1 + \rho(L)}{2} + \|B\| \epsilon_0 < 1$. When $p \gtrsim \log^2(1/\delta)$,

$$\left\| (A + B\hat{K}_{t+m-1}) \cdots (A + B\hat{K}_{p+1}) 1_{E_\delta} \right\| \lesssim \rho_0^{t+m-p}.$$

Apply this equation to Eq. (43):

$$\begin{aligned} & \mathbb{E} \left(\|x_{t+m}\|^k 1_{E_\delta} \right) \\ & \lesssim \mathbb{E} \left(\left(\left\| \sum_{p=t}^{t+m-1} \rho_0^{t+m-p} (B\eta_p + \varepsilon_p) \right\| + \rho_0^m \|x_t\| \right)^k 1_{E_\delta} \right) \\ & \quad (\text{By Holder's inequality}) \\ & \lesssim 2^{k-1} \mathbb{E} \left(\left\| \sum_{p=t}^{t+m-1} \rho_0^{t+m-p} (B\eta_p + \varepsilon_p) \right\|^k 1_{E_\delta} + \rho_0^{km} \|x_t\|^k 1_{E_\delta} \right) \\ & \lesssim \mathbb{E} \left(\left\| \sum_{p=t}^{t+m-1} \rho_0^{t+m-p} (B\eta_p + \varepsilon_p) \right\|^k \right) + \rho_0^{km} \mathbb{E} \left(\|x_t\|^k 1_{E_\delta} \right). \end{aligned}$$

Consider the variance of $\sum_{p=t}^{t+m-1} \rho_0^{t+m-p} (B\eta_p + \varepsilon_p)$:

$$\begin{aligned} & \text{Var} \left(\sum_{p=t}^{t+m-1} \rho_0^{t+m-p} (B\eta_p + \varepsilon_p) \right) \\ & = \sum_{p=t}^{t+m-1} \rho_0^{2(t+m-p)} \text{Var}((B\eta_p + \varepsilon_p)) \end{aligned}$$

$$\lesssim \sum_{p=t}^{t+m-1} \rho_0^{2(t+m-p)} = \sum_{i=1}^m \rho_0^{2i} \lesssim 1.$$

Since $\sum_{p=t}^{t+m-1} \rho_0^{t+m-p} (B\eta_p + \varepsilon_p)$ is Gaussian with finite variance, the first item is of constant order for any m . Thus

$$\mathbb{E} \left(\|x_{t+m}\|^k \mathbf{1}_{E_\delta} \right) \lesssim 1 + \rho_0^{km} \mathbb{E} \left(\|x_t\|^k \mathbf{1}_{E_\delta} \right).$$

Replace $t \leftarrow m$, and $m \leftarrow t - m$, then for $m \gtrsim \log^2(1/\delta)$,

$$\mathbb{E} \left(\|x_t\|^k \mathbf{1}_{E_\delta} \right) \lesssim 1 + \rho_0^{k(t-m)} \mathbb{E} \left(\|x_m\|^k \mathbf{1}_{E_\delta} \right).$$

Since the \hat{K}_t in Algorithm 1 cannot have norm greater than C_K , we have

$$\begin{aligned} & \mathbb{E} \left(\|x_m\|^k \mathbf{1}_{E_\delta} \right) \\ & \leq \mathbb{E} \|x_m\|^k \\ & \leq \mathbb{E} \left((\|A\| + \|B\| \|\hat{K}_m\|) \|x_{m-1}\| + \|B\| \|\eta_m\| + \|\varepsilon_m\| \right)^k \\ & \text{(By Holder's inequality)} \\ & \leq 3^{k-1} \left((\|A\| + \|B\| C_K)^k \mathbb{E} \|x_{m-1}\|^k + \|B\|^k \mathbb{E} \|\eta_m\|^k + \|\varepsilon_m\|^k \right) \\ & \leq 3^{k-1} \left((\|A\| + \|B\| C_K)^k \mathbb{E} \|x_{m-1}\|^k + \|B\|^k \sigma_\eta^k + \sigma_\varepsilon^k \right). \end{aligned}$$

By iterating this inequality down to $\|x_0\|^2$, we know that $\mathbb{E} \|x_m\|^k \lesssim C^{km}$ for some constant C . Thus, we know

$$\mathbb{E} \left(\|x_t\|^k \mathbf{1}_{E_\delta} \right) \lesssim 1 + \rho_0^{k(t-m)} C^{km}.$$

Since $\rho_0 < 1$, we can take $t \geq (\log_{1/\rho_0}(C) + 1)m$ which satisfies $\rho_0^{k(t-m)} C^{km} \leq 1$. Because we require $m \gtrsim \log^2(1/\delta)$, the condition for $t \geq (\log_{1/\rho_0}(C) + 1)m$ is still $t \gtrsim \log^2(1/\delta)$, which satisfies

$$\mathbb{E} \left(\|x_t\|^k \mathbf{1}_{E_\delta} \right) \lesssim 1.$$

□

B Proof of Theorem 2

Theorem. Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies

$$\mathcal{R}(U, T) = O_p \left(\sqrt{T} \right). \quad (44)$$

Proof. Recall from Lemma 6 that

$$\mathcal{J}(U, T) = \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + \sum_{t=1}^T \eta_t^\top R \eta_t + O_p \left(T^{1/2} \right),$$

and

$$\mathcal{J}(U^*, T) = \sum_{t=1}^T \varepsilon_t^\top P \varepsilon_t + O_p \left(T^{1/2} \right).$$

Thus

$$\begin{aligned}
\mathcal{R}(U, T) &= \mathcal{J}(U, T) - \mathcal{J}(U^*, T) \\
&= \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + \sum_{t=1}^T \eta_t^\top R \eta_t - \sum_{t=1}^T \varepsilon_t^\top P \varepsilon_t + O_p\left(T^{1/2}\right) \\
&= 2 \sum_{t=1}^T \varepsilon_t^\top P(B\eta_t) + \sum_{t=1}^T (B\eta_t)^\top P(B\eta_t) + \sum_{t=1}^T \eta_t^\top R \eta_t + O_p\left(T^{1/2}\right).
\end{aligned} \tag{45}$$

Recall that $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2 t^{-1/2} I_n)$.

$$\mathbb{E} \sum_{t=1}^T \eta_t^\top R \eta_t = \sum_{t=1}^T \mathbf{Tr}(R \mathbb{E} \eta_t \eta_t^\top) = \sum_{t=1}^T \mathbf{Tr}(R \sigma_\eta^2 t^{-1/2}) = O\left(T^{1/2}\right).$$

$$\text{Var} \left(\sum_{t=1}^T \eta_t^\top R \eta_t \right) = \sum_{t=1}^T \text{Var}(\eta_t^\top R \eta_t) = \sum_{t=1}^T O(t^{-1}) = O(\log(T)).$$

The standard error is of smaller order than the expectation. Thus, $\sum_{t=1}^T \eta_t^\top R \eta_t = O_p\left(T^{1/2}\right)$. Similarly, $\sum_{t=1}^T (B\eta_t)^\top P(B\eta_t) = O_p\left(T^{1/2}\right)$.

It remains to consider the order of $\sum_{t=1}^T \varepsilon_t^\top P(B\eta_t)$. Its expectation is 0.

$$\mathbb{E} \sum_{t=1}^T \varepsilon_t^\top P(B\eta_t) = 0.$$

The variance is

$$\text{Var} \left(\sum_{t=1}^T \varepsilon_t^\top P(B\eta_t) \right) = \sum_{t=1}^T \text{Var}(\varepsilon_t^\top P(B\eta_t)) = \sum_{t=1}^T O(t^{-1/2}) = O\left(T^{1/2}\right).$$

The standard error is of order $T^{1/4}$. Thus, $\sum_{t=1}^T \varepsilon_t^\top P(B\eta_t) = O_p\left(T^{1/4}\right) = o_p\left(T^{1/2}\right)$. Using these results, Eq. (45) becomes:

$$\begin{aligned}
\mathcal{R}(U, T) &= 2 \sum_{t=1}^T \varepsilon_t^\top P(B\eta_t) + \sum_{t=1}^T (B\eta_t)^\top P(B\eta_t) + \sum_{t=1}^T \eta_t^\top R \eta_t + O_p\left(T^{1/2}\right) \\
&= O_p\left(T^{1/2}\right).
\end{aligned}$$

□

B.1 Proof of Lemma 6

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies,*

$$\mathcal{J}(U, T) = \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + \sum_{t=1}^T \eta_t^\top R \eta_t + O_p\left(T^{1/2}\right).$$

and

$$\mathcal{J}(U^*, T) = \sum_{t=1}^T \varepsilon_t^\top P \varepsilon_t + O_p\left(T^{1/2}\right),$$

where ε_t is the system noise and η_t is the exploration noise in Algorithm 1, and $\tilde{\varepsilon}_t = B\eta_t + \varepsilon_t$.

We only prove the first equation because the second equation is a simplified version of the first equation (with $\eta_t = 0$ and $\hat{K}_t = K$).

Recursively applying system equations $x_{t+1} = Ax_t + Bu_t + \varepsilon_t$ and $u_t = \hat{K}_t x_t + \eta_t$ we have:

$$x_t = \sum_{p=0}^{t-1} (A + B\hat{K}_{t-1}) \cdots (A + B\hat{K}_{p+1})(B\eta_p + \varepsilon_p) + (A + B\hat{K}_{t-1}) \cdots (A + BK_0)x_0. \quad (46)$$

Notice that the state x_t has the same expression as if the system had noise $\tilde{\varepsilon}_t = B\eta_t + \varepsilon_t$ and controller $\tilde{u}_t = \hat{K}_t x_t$. We wish to switch to the new system because there are some existing tools with controls in the form of $\tilde{u}_t = \hat{K}_t x_t$.

We are interested in the cost

$$\mathcal{J}(U, T) = \sum_{t=1}^T x_t^\top Q x_t + u_t^\top R u_t \quad \text{with } u_t = \hat{K}_t x_t + \eta_t.$$

We will first show in Appendix B.1.1 the new system cost is

Lemma 11. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies,*

$$\sum_{t=1}^T x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t = \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + O_p\left(T^{1/2}\right).$$

and then prove in Appendix B.1.2 that the difference between the original cost and new cost is

Lemma 12. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies,*

$$\sum_{t=1}^T u_t^\top R u_t - \tilde{u}_t^\top R \tilde{u}_t = \sum_{t=1}^T \eta_t^\top R \eta_t + o\left(T^{1/4} \log^{\frac{3}{2}}(T)\right) \quad a.s.$$

Combining the above two equations, we conclude that

$$\begin{aligned} \mathcal{J}(U, T) &= \left[\sum_{t=1}^T x_t^\top Q x_t + u_t^\top R u_t \right] \\ &= \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + \sum_{t=1}^T \eta_t^\top R \eta_t + O_p\left(T^{1/2}\right). \end{aligned}$$

The optimal controller U^* is a simplified version of U from Algorithm 1 with $\eta_t = 0$ and $\hat{K}_t - K = 0$. With the same proof we can show that

$$\mathcal{J}(U^*, T) = \sum_{t=1}^T \varepsilon_t^\top P \varepsilon_t + O_p\left(T^{1/2}\right).$$

B.1.1 Cost of new system

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies,*

$$\sum_{t=1}^T x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t = \sum_{t=1}^T \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t + O_p\left(T^{1/2}\right).$$

Proof. Next we proceed as if our system was x_t with system noise $\tilde{\varepsilon}_t = B\eta_t + \varepsilon_t$ and controller $\tilde{u}_t = \hat{K}_t x_t$. The key idea of the following proof is from Appendix C of [Fazel et al. \(2018\)](#).

We are interested in the cost

$$\sum_{t=1}^T x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t \quad \text{with } \tilde{u}_t = \hat{K}_t x_t,$$

which can be written as

$$\begin{aligned} \sum_{t=1}^T x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t &= \sum_{t=1}^T x_t^\top Q x_t + (\hat{K}_t x_t)^\top R \hat{K}_t x_t \\ &= \sum_{t=1}^T x_t^\top (Q + \hat{K}_t^\top R \hat{K}_t) x_t \\ &= \sum_{t=1}^T \left[x_t^\top (Q + \hat{K}_t^\top R \hat{K}_t) x_t + x_{t+1}^\top P x_{t+1} - x_t^\top P x_t \right] + x_1^\top P x_1 - x_{T+1}^\top P x_{T+1} \\ &= \sum_{t=1}^T \left[x_t^\top (Q + \hat{K}_t^\top R \hat{K}_t) x_t + ((A + B \hat{K}_t) x_t + \tilde{\varepsilon}_t)^\top P ((A + B \hat{K}_t) x_t + \tilde{\varepsilon}_t) - x_t^\top P x_t \right] \\ &\quad + x_1^\top P x_1 - x_{T+1}^\top P x_{T+1} \\ &\quad \text{(by Lemma 18 in [Wang and Janson \(2020\)](#))} \\ &= \sum_{t=1}^T \left[x_t^\top (Q + \hat{K}_t^\top R \hat{K}_t) x_t + x_t^\top (A + B \hat{K}_t)^\top P (A + B \hat{K}_t) x_t - x_t^\top P x_t \right. \\ &\quad \left. + 2\tilde{\varepsilon}_t^\top P (A + B \hat{K}_t) x_t + \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t \right] + \tilde{O}_p(1). \end{aligned} \tag{47}$$

We constructed the specific form of the first term on purpose. The following lemma translates the first term into a quadratic term with respect to $\hat{K}_t - K$. We use the Lemma 25 from [Wang and Janson \(2020\)](#):

Lemma 13 (Lemma 25 from [Wang and Janson \(2020\)](#)). *For any \hat{K} with suitable dimension,*

$$\begin{aligned} x^\top (Q + \hat{K}^\top R \hat{K}) x + x^\top (A + B \hat{K})^\top P (A + B \hat{K}) x - x^\top P x \\ = x^\top (\hat{K} - K)^\top (R + B^\top P B) (\hat{K} - K) x. \end{aligned}$$

As a result

$$\begin{aligned} \sum_{t=1}^T x_t^\top Q x_t + \tilde{u}_t^\top R \tilde{u}_t &= \sum_{t=1}^T \left[x_t^\top (\hat{K}_t - K)^\top (R + B^\top P B) (\hat{K}_t - K) x_t \right. \\ &\quad \left. + 2\tilde{\varepsilon}_t^\top P (A + B \hat{K}_t) x_t + \tilde{\varepsilon}_t^\top P \tilde{\varepsilon}_t \right] + \tilde{O}_p(1). \end{aligned}$$

Now we have three terms, and we will estimate the order of each of these three terms.

1. The first term we consider is $\sum_{t=1}^T x_t^\top (\hat{K}_t - K)^\top (R + B^\top P B) (\hat{K}_t - K) x_t$. For any $0 < \delta < 1/2$,

$$\begin{aligned} &\mathbb{E} \left(\sum_{t=1}^T x_t^\top (\hat{K}_t - K)^\top (R + B^\top P B) (\hat{K}_t - K) x_t \mathbf{1}_{E_\delta} \right) \\ &\leq \mathbb{E} \left(\sum_{t=1}^T \|x_t\|^2 \|\hat{K}_t - K\|^2 \|R + B^\top P B\| \mathbf{1}_{E_\delta} \right) \end{aligned}$$

$$\begin{aligned}
&\lesssim \mathbb{E} \left(\sum_{t=1}^T \|x_t\|^2 \|\hat{K}_t - K\|^2 \mathbf{1}_{E_\delta} \right) \\
&\lesssim \mathbb{E} \left(\sum_{t=1}^T t^{1/2} \|\hat{K}_t - K\|^4 + t^{-1/2} \|x_t\|^4 \mathbf{1}_{E_\delta} \right) \\
&\text{(By Eq. (34) and the same inequalities as in Eq. (30))} \\
&\lesssim \sum_{t=T_0}^T t^{-1/2} + \sum_{t=1}^{T_0-1} t^{1/2} (C_K + \|K\|)^4 + \log^2(1/\delta) + T^{1/2} \\
&\lesssim T^{1/2} + \log^2(1/\delta).
\end{aligned}$$

Then for any $0 < \delta < 1/2$, we have

$$\begin{aligned}
&\mathbb{P} \left(\sum_{t=1}^T x_t^\top (\hat{K}_t - K)^\top (R + B^\top P B) (\hat{K}_t - K) x_t \mathbf{1}_{E_\delta} \gtrsim \frac{1}{\delta} (T^{1/2} + \log^2(1/\delta)) \right) \leq \delta. \\
&\mathbb{P} \left(\sum_{t=1}^T x_t^\top (\hat{K}_t - K)^\top (R + B^\top P B) (\hat{K}_t - K) x_t \gtrsim \frac{1}{\delta} (\log^2(1/\delta) + 1) T^{1/2} \right) \leq 2\delta.
\end{aligned}$$

By big O in probability notation, this implies

$$\sum_{t=1}^T x_t^\top (\hat{K}_t - K)^\top (R + B^\top P B) (\hat{K}_t - K) x_t = O_p(T^{1/2}).$$

2. The second term we consider is $\sum_{t=1}^T \tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t$. Notice that $\tilde{\varepsilon}_t = \varepsilon_t + B\eta_t \perp (A + B\hat{K}_t)x_t$. Then

$$\mathbb{E} \sum_{t=1}^T \tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t = 0.$$

$$\begin{aligned}
&\mathbb{E} \left(\sum_{t=1}^T \tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t \mathbf{1}_{E_\delta} \right)^2 \\
&= \sum_{t=1}^T \mathbb{E} \left(\tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t \right)^2 \mathbf{1}_{E_\delta} \\
&\leq \sum_{t=1}^T \mathbb{E} \left(\|\tilde{\varepsilon}_t\|^2 \|P\|^2 \|(A + B\hat{K}_t)\|^2 \|x_t\|^2 \mathbf{1}_{E_\delta} \right) \\
&\quad \left(\|\hat{K}_t\| \leq C_K \text{ based on Algorithm 1 design} \right) \\
&\leq \sum_{t=1}^T \|P\|^2 (\|A\| + \|B\| C_K)^2 \mathbb{E} \|\tilde{\varepsilon}_t\|^2 \mathbb{E} (\|x_t\|^2 \mathbf{1}_{E_\delta}) \\
&\lesssim \sum_{t=1}^T \mathbb{E} (\|x_t\|^2 \mathbf{1}_{E_\delta}) \\
&\text{(By the inequalities in Eq. (30))} \\
&\lesssim T + \log^2(1/\delta) \log \log(1/\delta) \\
&\lesssim T + \log^3(1/\delta).
\end{aligned}$$

Then for any $0 < \delta < 1/2$, we have

$$\mathbb{P} \left(\sum_{t=1}^T \tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t \mathbf{1}_{E_\delta} \gtrsim \sqrt{\frac{1}{\delta}(T + \log^3(1/\delta))} \right) \leq \delta$$

and

$$\mathbb{P} \left(\sum_{t=1}^T \tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t \gtrsim \sqrt{\frac{1}{\delta}(1 + \log^3(1/\delta))T} \right) \leq 2\delta.$$

By big O in probability notation, this implies

$$\sum_{t=1}^T \tilde{\varepsilon}_t^\top P(A + B\hat{K}_t)x_t = O_p \left(T^{1/2} \right). \quad (48)$$

3. The third term is $\sum_{t=1}^T \tilde{\varepsilon}_t^\top P\tilde{\varepsilon}_t$ and we leave that in the equation.

Summing up the three parts we have:

$$\sum_{t=1}^T x_t^\top Qx_t + \tilde{u}_t^\top R\tilde{u}_t = \sum_{t=1}^T \tilde{\varepsilon}_t^\top P\tilde{\varepsilon}_t + O_p \left(T^{1/2} \right).$$

□

B.1.2 Cost difference induced by transformation

Lemma. *Algorithm 1 applied to a system described by Eq. (1) under Assumption 1 satisfies,*

$$\sum_{t=1}^T u_t^\top Ru_t - \tilde{u}_t^\top R\tilde{u}_t = \sum_{t=1}^T \eta_t^\top R\eta_t + o \left(T^{1/4} \log^{\frac{3}{2}}(T) \right) \text{ a.s.}$$

Proof. The difference is expressed as

$$\begin{aligned} \sum_{t=1}^T u_t^\top Ru_t - \tilde{u}_t^\top R\tilde{u}_t &= \sum_{t=1}^T (\hat{K}_t x_t + \eta_t)^\top R(\hat{K}_t x_t + \eta_t) - \sum_{t=1}^T (\hat{K}_t x_t)^\top R(\hat{K}_t x_t) \\ &= 2 \sum_{t=1}^T (\hat{K}_t x_t)^\top R\eta_t + \sum_{t=1}^T \eta_t^\top R\eta_t. \end{aligned}$$

Eq. (83) of Wang and Janson (2020) shows that

$$\sum_{t=1}^T (\hat{K}_t x_t)^\top R\eta_t = o \left(T^{1/4} \log^{\frac{3}{2}}(T) \right) \text{ a.s.}$$

As a conclusion,

$$\sum_{t=1}^T u_t^\top Ru_t - \tilde{u}_t^\top R\tilde{u}_t = \sum_{t=1}^T \eta_t^\top R\eta_t + o \left(T^{1/4} \log^{\frac{3}{2}}(T) \right) \text{ a.s.}$$

□