Learning to Control Linear Systems can be Hard

Anastasios Tsiamis ATSIAMIS@CONTROL.EE.ETHZ.CH
Automatic Control Laboratory, ETH Zurich

Ingvar Ziemann ZIEMANN@KTH.SE
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology

Manfred Morari MORARI@SEAS.UPENN.EDU
Department of Electrical and Systems Engineering, University of Pennsylvania

Nikolai Matni NMATNI@SEAS.UPENN.EDU
Department of Electrical and Systems Engineering, University of Pennsylvania

George J. Pappas PAPPASG@SEAS.UPENN.EDU
Department of Electrical and Systems Engineering, University of Pennsylvania

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

In this paper, we study the statistical difficulty of learning to control linear systems. We focus on two standard benchmarks, the sample complexity of stabilization, and the regret of the online learning of the Linear Quadratic Regulator (LQR). Prior results state that the statistical difficulty for both benchmarks scales polynomially with the system state dimension up to system-theoretic quantities. However, this does not reveal the whole picture. By utilizing minimax lower bounds for both benchmarks, we prove that there exist non-trivial classes of systems for which learning complexity scales dramatically, i.e. exponentially, with the system dimension. This situation arises in the case of underactuated systems, i.e. systems with fewer inputs than states. Such systems are structurally difficult to control and their system theoretic quantities can scale exponentially with the system dimension dominating learning complexity. Under some additional structural assumptions (bounding systems away from uncontrollability), we provide qualitatively matching upper bounds. We prove that learning complexity can be at most exponential with the controllability index of the system, that is the degree of underactuation.

1. Introduction

In stochastic linear control, the goal is to design a controller for a system of the form

$$S: \qquad x_{k+1} = Ax_k + Bu_k + Hw_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the system internal state, $u_k \in \mathbb{R}^p$ is some exogenous input, and $w_k \in \mathbb{R}^r$ is some random disturbance sequence. Matrices A, B, H determine the evolution of the state, based on the previous state, control input, and disturbance respectively. Control theory has a long history of studying how to design controllers for system (1) when its model is *known* (Bertsekas, 2017). However, in reality system (1) might be *unknown* and we might not have access to its model. In this case, we have to learn how to control (1) based on data.

Controlling unknown dynamical systems has also been studied from the perspective of Reinforcement Learning (RL). Although the setting of tabular RL is relatively well-understood (Jaksch et al., 2010), it has been challenging to analyze the continuous setting,

where the state and/or action spaces are infinite (Ortner and Ryabko, 2012; Kakade et al., 2020). Recently, there has been renewed interest in learning to control linear systems. Indeed, linear systems are simple enough to allow for an in-depth theoretical analysis, yet exhibit sufficiently rich behavior so that we can draw conclusions about continuous control of more general system classes (Recht, 2019). In this paper we focus on the following two problems.

Regret of online LQR. A fundamental benchmark for continuous control is the Linear Quadratic Regulator (LQR) problem, where the goal is to compute a policy [1] π that minimizes

$$J^*(S) \triangleq \min_{\pi} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{S,\pi}\left[\sum_{t=0}^{T-1}(x_t'Qx_t + u_t'Ru_t) + x_T'Q_Tx_T\right], \tag{2}$$

where $Q \in \mathbb{R}^{n\times n}$, $R \in \mathbb{R}^{p\times p}$ are the state and input penalties respectively; these penalties control the tradeoff between state regulation and control effort. When model (1) is known, LQR enjoys a closed-form solution; the optimal policy is a linear feedback law $\pi_{\star,t}(x_t) = K_\star x_t$, where the control gain $K_\star$ is given by solving the celebrated Algebraic Riccati Equation (ARE) (7). If model (1) is unknown, we have to learn the optimal policy from data. In the online learning setting, the goal of the learner is to find a policy that adapts online and competes with the optimal LQR policy that has access to the true model. The suboptimality of the online learning policy at time T is captured by the *regret*

$$R_T(S) \triangleq \sum_{t=0}^{T-1}(x_t'Qx_t + u_t'Ru_t) + x_T'Q_Tx_T - TJ^*(S). \tag{3}$$

The learning task is to find a policy with as small regret as possible.

Sample Complexity of Stabilization Another important benchmark is the problem of stabilization from data. The goal is to learn a linear gain $K \in \mathbb{R}^{m\times n}$ such that the closed-loop system $A + BK$ is stable, i.e., such that its spectral radius $\rho(A + BK)$ is less than one. Many algorithms for online LQR require the existence of such a stabilizing gain to initialize the online learning policy (Simchowitz and Foster, 2020; Jedra and Proutiere, 2021). Furthermore, stabilization is a problem of independent interest (Faradonbeh et al., 2018b). In this setting, the learner designs an exploration policy π and an algorithm that uses batch state-input data $x_0, \dots, x_N, u_0, \dots, u_{N-1}$ to output a control gain $\hat{K}_N$, at the end of the exploration phase. Here we focus on *sample complexity*, i.e., the minimum number of samples N required to find a stabilizing gain.

Since the seminal papers by Abbasi-Yadkori and Szepesvári (2011) and Dean et al. (2017) both LQR and stabilization have been studied extensively in the literature – see Section 1.1. Current state-of-the-art results state that the regret of online LQR and the sample complexity of stabilization scale at most polynomially with system dimension n

$$R_T(S) \lesssim C_1^{\text{sys}}\text{poly}(n)\sqrt{T}, \quad N \lesssim C_2^{\text{sys}}\text{poly}(n), \tag{4}$$

where C_1^{sys}, C_2^{sys} are system specific constants that depend on several control theoretic quantities of system (1). However, the above statements might not reveal the whole picture.

In fact, system theoretic parameters C_1^{sys}, C_2^{sys} can actually hide dimensional dependence on n. This dependence has been overlooked in prior work. As we show in this paper,

1. A policy decides the current control input u_t based on past state-input values–see Section 2 for details.

there exist non-trivial classes of linear systems for which system theoretic parameters scale dramatically, i.e. exponentially, with the dimension n. As a result, the system theoretic quantities C_1^{sys}, C_2^{sys} might be very large and in fact *dominate* the poly(n) term in the upper bounds (4). This phenomenon especially arises in systems which are structurally difficult to control, such as for example underactuated systems. Then, the upper bounds (4) suggest that learning might be difficult for such instances. This brings up the following questions. *Can learning LQR or stabilizing controllers indeed be hard for such systems? How does system structure affect difficulty of learning?*

To answer the first question, we need to establish lower bounds. As we discuss in Section 1.1, existing lower bounds for online LQR (Simchowitz and Foster, 2020) might not always reveal the dependence on control theoretic parameters. Chen and Hazan (2021) provided exponential lower bounds for the start-up regret of stabilization. Still, to the best of our knowledge, there are no existing lower bounds for the *sample complexity* of stabilization. Recently, it was shown that the sample complexity of system identification can grow exponentially with the dimension n (Tsiamis and Pappas, 2021). However, it is not clear if difficulty of identification translates into difficulty of control. Besides, we do not always need to identify the whole system in order to control it (Gevers, 2005). To answer the second question, we need to provide upper bounds for several control theoretic parameters. Our contributions are the following:

Exp(n) Stabilization Lower Bounds. We prove an information-theoretic lower bound for the problem of learning stabilizing controllers, showing that it can indeed be statistically hard for underactuated systems. In particular, we show that the sample complexity of stabilizing an unknown underactuated linear system can scale exponentially with the state dimension n. To the best of our knowledge this is the first paper to address this issue and consider lower bounds in this setting.

Exp(n) LQR Regret Lower Bounds. We show that the regret of online LQR can scale exponentially with the dimension as $\exp(n)\sqrt{T}$. In fact, even common integrator-like systems can exhibit this behavior. To prove our result, we leverage recent regret lower bounds (Ziemann and Sandberg, 2022), which provide a refined analysis linking regret to system theoretic parameters. Chen and Hazan (2021) first showed that the start-up cost of the regret (terms of low order) can scale exponentially with n. Here, we show that this exponential dependence can also affect multiplicatively the dominant $\sqrt{T}$ term.

Exponential Upper Bounds. Under some additional structural assumptions (bounding systems away from uncontrollability), we provide matching global upper bounds. We show that the sample complexity of stabilization and the regret of online LQR can be at most exponential with the dimension n. In fact, we prove a stronger result, that they can be at most exponential with the *controllability index* of the system, which captures the structural difficulty of control – see Section 3. This implies that if the controllability index is small with respect to the dimension n, then learning is guaranteed to be easy.

1.1. Related Work

System Identification. A related problem is that of system identification, where the learning objective is to recover the model parameters A, B, H from data (Matni and Tu, 2019). The sample complexity of system identification was studied extensively in the setting

of fully observed linear systems (Dean et al., 2017; Simchowitz et al., 2018; Faradonbeh et al., 2018a; Sarkar and Rakhlin, 2018; Fattahi et al., 2019; Jedra and Proutiere, 2019; Wagenmaker and Jamieson, 2020; Efroni et al., 2021) as well as partially-observed systems (Oymak and Ozay, 2018; Sarkar et al., 2019; Simchowitz et al., 2019; Tsiamis and Pappas, 2019; Lee and Lamperski, 2019; Zheng and Li, 2020; Lee, 2020; Lale et al., 2020b). Recently, it was shown that the sample complexity of system identification can grow exponentially with the dimension n (Tsiamis and Pappas, 2021).

Learning Feedback Laws. The problem of learning stabilizing feedback laws from data was studied before in the case of stochastic (Dean et al., 2017; Tu et al., 2017; Faradonbeh et al., 2018b; Mania et al., 2019) as well as adversarial (Chen and Hazan, 2021) disturbances. The standard paradigm has been to perform system identification, followed by a robust control or certainty equivalent gain design. Prior work is limited to sample complexity upper bounds. To the best of our knowledge, there have been no sample complexity lower bounds.

Online LQR. While adaptive control in the LQR framework has a rich history (Matni et al., 2019), the recent line of work on regret minimization in online LQR begins with Abbasi-Yadkori and Szepesvári (2011). They provide a computationally intractable algorithm based on optimism attaining $O(\sqrt{T})$ regret. Algorithms based on optimism have since been improved and made more tractable (Ouyang et al., 2017; Abeille and Lazaric, 2018; Abbasi-Yadkori et al., 2019; Cohen et al., 2019; Abeille and Lazaric, 2020). In a closely related line of work, Dean et al. (2018) provide an $O(T^{2/3})$ regret bound for robust adaptive LQR control, drawing inspiration from classical methods in system identification and robust adaptive control. It has since been shown that certainty equivalent control, without robustness, can attain the (locally) minimax optimal $O(\sqrt{T})$ regret (Mania et al., 2019; Faradonbeh et al., 2020; Lale et al., 2020a; Jedra and Proutiere, 2021). In particular, by providing nearly matching upper and lower bounds, Simchowitz and Foster (2020) refine this analysis and establish that the optimal rate, without taking system theoretic quantities into account, is $R_T = \Theta(\sqrt{p^2nT})$. In this work, we rely on the lower bounds by Ziemann and Sandberg (2022), which provide a refined instance specific analysis and also lower bounds for the partially observed setting. Here, we further refine their lower bounds to reveal a sharper dependence of the regret on control theoretic parameters. Hence, we how that certain non-local minimax complexities can be far worse than $R_T = \Omega(\sqrt{p^2nT})$ and scale exponentially in the problem dimension. Indeed, an exponential start-up cost has already been observed by Chen and Hazan (2021), in the case of adversarial disturbances. Here we show that this exponential dependency can persist multiplicatively even for large T, in the case of stochastic disturbances. Thus, our results complement the results of Chen and Hazan (2021).

1.2. Notation

The transpose of X is denoted by X'. For vectors $v \in \mathbb{R}^d$, $\|v\|_2$ denotes the ℓ_2-norm. For matrices $X \in \mathbb{R}^{d_1 \times d_2}$, the spectral norm is denoted by $\|X\|_2$. For comparison with respect to the positive semi-definite cone we will use $\succeq$ or $\succ$ for strict inequality. By $\mathbb{P}$ we will denote probability measures and by $\mathbb{E}$ expectation. By $\text{poly}(\cdot)$ we denote a polynomial function of its arguments. By $\exp(\cdot)$ we denote a exponential function of its arguments.

2. Problem Statement

System (1) is characterized by the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $H \in \mathbb{R}^{n\times r}$. We assume that $w_k \sim \mathcal{N}(0, I_r)$ is i.i.d. Gaussian with unit covariance. Without loss of generality the initial state is assumed to be zero $x_0 = 0$. In a departure from prior work, we do not necessarily assume that the noise is isotropic. Instead, we consider a more general model, where the noise Hw_k is allowed to be degenerate–see also Remark 6.

Assumption 1 *Matrices A, B, H and the noise dimension $r \leq n$ are all unknown. The unknown matrices are bounded, i.e. $\|A\|_2, \|B\|_2, \|H\|_2 \leq M$, for some positive constant $M \geq 1$. Matrices B, H have full column rank* $\mathrm{rank}(B) = p \leq n$, $\mathrm{rank}(H) = r \leq n$. *We also assume that the system is non-explosive $\rho(A) \leq 1$.*

The boundedness assumption on the state parameters allows us to argue about global sample complexity upper bounds. To simplify the presentation, we make the assumption that the system is non-explosive $\rho(A) \leq 1$. This setting includes marginally stable systems and is rich enough to provide insights about the difficulty of learning more general systems.

A policy is a sequence of functions $\pi = \{\pi_t\}_{t=0}^{N-1}$. Every function π_t maps previous state-input values $x_0, \ldots, x_t, u_0, \ldots, u_{t-1}$ and potentially an auxiliary randomization signal AUX to the new input u_t. Hence all inputs u_t are $\mathcal{F}_t$-measurable, where $\mathcal{F}_t \triangleq \sigma(x_0, \ldots, x_t, u_0, \ldots, u_{t-1}, \mathrm{AUX})$. For brevity we will use the symbol S to denote a system $S = (A, B, H)$. Let $\mathbb{P}_{S,\pi}$ ($\mathbb{E}_{S,\pi}(\cdot)$) denote the probability distribution (expectation) of the input-state data when the true system is equal to S and we apply a policy π.

2.1. Difficulty of Stabilization

In the stabilization problem, the goal is to find a state-feedback control law $u = Kx$, where K renders the closed-loop system $A + BK$ stable with spectral radius less than one, i.e., $\rho(A+BK) < 1$. We assume that we collect data $x_0, \ldots, x_N, u_0, \ldots, u_N$, which are generated by system (1) using any exploration policy π, e.g. white-noise excitation, active learning etc. Since we care only about sample complexity, the policy is allowed to be maximally exploratory. To make the problem meaningful, we restrict the average control energy.

Assumption 2 *The control energy is bounded $\mathbb{E}_{S,\pi}\|u_t\|_2^2 \leq \sigma_u^2$, for some $\sigma_u > 0$.*

Next, we define a notion of learning difficulty for classes of linear systems. By $\mathscr{C}_n$ we will denote a class of systems with dimension n. We will define as easy, classes of linear system that exhibit $\mathrm{poly}(n)$ sample complexity.

Definition 1 (Poly(n)-stabilizable classes) *Let $\mathscr{C}_n$ be a class of systems. Let $\hat{K}_N$ be a function that maps input-state data $(u_0, x_1), \ldots, (u_{N-1}, x_N)$ to a control gain. We call the class $\mathscr{C}_n$ $\mathrm{poly}(n)-$stabilizable if there exists an algorithm $\hat{K}_N$ and an exploration policy π satisfying Assumption 2, such that for any confidence $0 \leq \delta < 1$:*

$$\sup_{S\in\mathscr{C}_n} \mathbb{P}_{S,\pi}\left(\rho(A + B\hat{K}_N) \geq 1\right) \leq \delta, \quad \text{if} \quad N\sigma_u^2 \geq \mathrm{poly}(n, \log 1/\delta, M). \tag{5}$$

Our definition requires both the number of samples and the input energy to be polynomial with the arguments. The above class-specific definition can be turned into a local, instance-specific, definition of sample complexity by considering a neighborhood around an unknown system. The question then arises whether linear systems are generally $\text{poly}(n)$-stabilizable.

Problem 1 *Are there linear system classes which are not* $\text{poly}(n)$*-stabilizable? When can we guarantee* $\text{poly}(n)$*-stabilizability?*

2.2. Difficulty of Online LQR

Consider the LQR objective (2). Let the state penalty matrix $Q \in \mathbb{R}^{n\times n} \succ 0$ be positive definite, with the input penalty matrix $R \in \mathbb{R}^{p\times p}$ also positive definite. When the model is known, the optimal policy is a linear feedback law $\pi_\star = \{K_\star x_k\}_{k=0}^{T-1}$, where $K_\star$ is given by

$$K_\star = -(B'PB + R)^{-1}B'PA, \tag{6}$$

and P is the unique positive definite solution to the Algebraic Riccati Equation (ARE)

$$P = A'PA + Q - A'PB(B'PB + R)^{-1}B'PA. \tag{7}$$

Throughout the paper, we will assume that $Q_T = P$. If the model of (1) is unknown, the goal of the learner is to find an online learning policy π that leads to minimum regret $R_T(S)$. In the setting of online LQR, the data are revealed sequentially, i.e. x_{t+1} is revealed after we select u_t. Contrary to the stabilization problem, here we study regret, i.e. there is a tradoff between exploration and exploitation. We will define a class-specific notion of learning difficulty based on the ratio between the regret and $\sqrt{T}$.

Definition 2 (Poly(n)-Regret) *Let $\mathscr{C}_n$ be a class of systems of dimension n. We say that the class $\mathscr{C}_n$ exhibits poly(n) minimax expected regret if*

$$\min_{\pi} \sup_{S\in\mathscr{C}_n} \mathbb{E}_{S,\pi} R_T(S) \leq \text{poly}(n, M, \log T)\sqrt{T} + \tilde{O}(1), \tag{8}$$

where $\tilde{O}(1)$ hides $\text{poly}\log T$ terms.

Our definition here is based on expected regret, but we could have a similar definition based on high probability regret guarantees – see Dann et al. (2017) for distinctions between the two definitions. Similar to the stabilization problem, we pose the following questions.

Problem 2 *Are there classes of systems for which poly(n)-regret is impossible? When is poly(n)-regret guaranteed?*

3. Classes with Rich Controllability Structure

Before we present our learning guarantees, we need to find classes of systems, where learning is meaningful. To make sure that the stabilization and the LQR problems are well-defined, we assume that system (1) is controllable[2].

2. We can slightly relax the condition to (A, B) stabilizable (Lale et al., 2020a; Simchowitz and Foster, 2020; Efroni et al., 2021). To avoid technicalities we leave that for future work.

Assumption 3 *System (1) is* (A, B) controllable*, i.e. matrix*

$$\mathcal{C}_k(A,B) \triangleq \begin{bmatrix} B & AB & \cdots & A^{k-1}B \end{bmatrix} \tag{9}$$

has full column rank $\text{rank}(\mathcal{C}_k(A,B)) = n$*, for some* $k \leq n$*.*

Unsurprisingly, the class of all controllable systems does not exhibit finite sample complexity/regret, let alone polynomial sample complexity/regret. The main issue is that there exist systems which satisfy the rank condition but are arbitrarily close to uncontrollability. For example, consider the following controllable system, which we want to stabilize

$$x_{k+1} = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k + w_k.$$

The only way to stabilize the system is indirectly by using the second state $x_{k,2}$, via the coupling coefficient α. However, we need to know the sign of α. If α is allowed to be arbitrarily small, i.e. the system is arbitrarily close to uncontrollability, then an arbitrarily large number of samples is required to learn the sign of α, leading to infinite complexity. To obtain classes with finite sample complexity/regret we need to bound the system instances away from uncontrollability. One way is to consider the least singular value of the controllability Gramian $\Gamma_k(A,B)$ at time k:

$$\Gamma_k(A,B) \triangleq \sum_{t=0}^{k-1} A^t B B' (A')^t. \tag{10}$$

An implicit assumption in prior literature is that $\sigma_{\min}^{-1}(\Gamma_k(A,B)) \leq \text{poly}(n)$. We will not assume this here, since it might exclude many systems of interest, such as integrator-like systems, also known as underactuated systems, or networks (Pasqualetti et al., 2014). Instead, we will relax this requirement to allow richer system structures.

To avoid pathologies, we will lower bound the coupling between states in the case of indirectly controlled systems. To formalize this idea, let us review some notions from system theory. The *controllability index* is defined as follows

$$\kappa(A,B) \triangleq \min\left\{k \geq 1 : \text{rank}(\mathcal{C}_k(A,B)) = n\right\}, \tag{11}$$

i.e., it is the minimum time such that the controllability rank condition is satisfied. It captures the degree of underactuation and reflects the structural difficulty of control.

Based on the fact that the rank of the controllability matrix at time κ is n, we can show that the pair (A,B) admits the following canonical representation, under a unitary similarity transformation (Dooren, 2003). It is called the Staircase or Hessenberg form of system (1).

Proposition 3 (Staircase form) *Consider a controllable pair* (A,B) *with controllability index* κ *and controllability matrix* $\mathcal{C}_k$*,* $k \geq 0$*. There exists a unitary similarity transformation* $U \in \mathbb{R}^{n\times n}$ *such that* $U'U = UU' = I$ *and:*

$$U'B = \begin{bmatrix} B_1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad U'AU = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,\kappa-1} & A_{1,\kappa} \\ A_{2,1} & A_{2,2} & \cdots & A_{3,\kappa-1} & A_{2,\kappa} \\ 0 & A_{3,2} & \cdots & A_{3,\kappa-1} & A_{3,\kappa} \\ 0 & 0 & \cdots & A_{4,\kappa-1} & A_{4,\kappa} \\ \vdots & & & \vdots & \\ 0 & 0 & \cdots & A_{\kappa,\kappa-1} & A_{\kappa,\kappa} \end{bmatrix}, \tag{12}$$

where $A_{i,j} \in \mathbb{R}^{p_i \times p_j}$ *are block matrices, with* $p_i = \text{rank}(\mathcal{C}_i) - \text{rank}(\mathcal{C}_{i-1})$, $p_1 = p$, $B_1 \in \mathbb{R}^{p \times p}$. *Matrices* $A_{i+1,i}$ *have full row rank* $\text{rank}(A_{i+1,i}) = p_{i+1}$ *and the sequence* p_i *is decreasing.*

Matrix U is the orthonormal matrix of the QR decomposition of the first n independent columns of $\mathcal{C}_\kappa(A, B)$. It is unique up to sign flips of its columns. The above representation captures the coupling between the several sub-states via the matrices $A_{i+1,i}$. It has been used before as a test of controllability Dooren (2003). This motivates the following definition, wherein we bound the coupling matrices $A_{i+1,i}$ away from zero.

Definition 4 (Robustly coupled systems) *Consider a controllable system* (A, B) *with controllability index* κ. *It is called* $\mu-$*robustly coupled if and only if for some positive* $\mu > 0$*:*

$$\sigma_p(B_1) \geq \mu, \quad \sigma_{p_{i+1}}(A_{i+1,i}) \geq \mu, \ \text{ for all } 1 \leq i \leq \kappa - 1, \tag{13}$$

where B_1, $A_{i+1,i}$ *are defined as in the Staircase form* (12).

In the previous example, by introducing the $\mu-$robust coupling requirement, we enforce a lower bound on the coupling coefficient $\alpha \geq \mu$, thus, avoiding pathological systems.

In the following sections, we connect the controllability index to the hardness/ease of control. We prove rigorously why performance might degrade as the index becomes $\kappa = O(n)$, as, e.g., in the case of integrator-like systems or networks. This cannot be explained based on prior work or based on global lower-bounds on the least singular value of the controllability Gramian. The controllability index and the controllability Gramian are two different measures that are suitable for different types of guarantees. The controllability index captures the structural difficulty of control, so it might be more suitable for class-specific guarantees versus instance-specific local guarantees.

4. Difficulty of Stabilization

In this section, we show that there exist non-trivial classes of linear systems for which the problem of stabilization from data is hard. In fact, the class of robustly coupled systems requires at least an exponential, in the state dimension n, number of samples.

Theorem 5 (Stabilization can be Hard) *Consider the class* $\mathscr{C}^{\mu}_{n,\kappa}$ *of all* μ*-robustly coupled systems* $S = (A, B, H)$ *of dimension* n *and controllability index* κ. *Let Assumption 2 hold and let* $\mu < 1$. *Then, for any stabilization algorithm, the sample complexity is exponential in the index* κ. *For any confidence* $0 \leq \delta < 1/2$ *the requirement*

$$\sup_{S \in \mathscr{C}^{\mu}_{n,\kappa}} \mathbb{P}_{S,\pi} \left(\rho(A + B\hat{K}_N) \geq 1 \right) \leq \delta$$

is satisfied only if

$$N\sigma_u^2 \geq \frac{1}{2} \left(\frac{1}{\mu} \right)^{2\kappa - 2} \left(\frac{1-\mu}{\mu} \right)^2 \log \frac{1}{3\delta}.$$

Theorem 5 implies that system classes with large controllability index, e.g. $\kappa = n$, suffer in general from sample complexity which is exponential with the dimension n. In other words, learning difficulty arises in the case of under-actuated systems. Only a limited number of

system states are directly driven by inputs and the remaining states are only indirectly excited, leading to a hard learning and stabilization problem. Consider now systems

$$S_i: \qquad x_{k+1} = \begin{bmatrix} 1 & \alpha_i\mu & 0 & \cdots & 0 \\ 0 & 0 & \mu & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & 0 & 0 & \cdots & \mu \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \mu \end{bmatrix} u_k + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} w_k,\ i \in \{1, 2\}, \tag{14}$$

where $0 < \mu < 1$, $\alpha_1 = 1$, $\alpha_2 = -1$. Systems S_1, S_2 are almost identical with the exception of element A_{12} where they have different signs. Both systems have one marginally stable mode corresponding to state $x_{k,1}$. The only way to stabilize $x_{k,1}$ with state feedback is indirectly, via $x_{k,2}$. Given system S_1, since $\alpha_1\mu > 0$, it is necessary that the first component of the gain is negative $\hat{K}_{N,1} < 0$. This follows from the Jury stability criterion, a standard stability test in control theory (Fadali and Visioli, 2013, Ch. 4.5). Let $\phi_1(z) = \det(zI - A_1 - B\hat{K}_N)$ be the characteristic polynomial of system S_1. Then one of the necessary conditions in Jury's criterion requires:

$$\phi_1(1) > 0,$$

which can only be satisfied if $\hat{K}_{N,1} < 0$ (see Appendix C for details). On the other hand, we can only stabilize S_2 if $\hat{K}_{N,1} > 0$. Hence, the only way to stabilize the system is to identify the sign of α_i. In other words, we transform the stabilization problem into a system identification problem. However, identification of the correct sign is very hard since the excitation of $x_{k,2} = \mu^{n-1}u_{k-n+1}$ scales with μ^{n-1}. The proof relies on Birgé's inequality (Boucheron et al., 2013). In Section C we construct a slightly more general example with non-zero diagonal elements. Our construction relies on the fact that $\mu < 1$. It is an open question whether we can construct hard learning instances for $\mu \geq 1$.

One insight that we obtain from the above example is that lack of excitation might lead to large sample complexity of stabilization. In particular, this can happen when we have an unstable/marginally stable mode, which can only be controlled via the system identification bottleneck, like $A_{1,2}$ in the above example.

Remark 6 (Singular noise) *Our stabilization lower bound exploits the fact that the constructed system (14) has low-rank noise, such that system identification is hard. It is an open problem whether we can construct examples of systems that are not* poly(n)*−stabilizable even though they are excited by full-rank noise. Nonetheless, in our regret lower bounds, we allow the noise to be full-rank.*

4.1. Sample complexity upper bounds

As we show below, sample complexity cannot be worse than exponential under the assumption of robust coupling. If the exploration policy is a white noise input sequence, then using a least squares identification algorithm (Simchowitz et al., 2018), and a robust control design scheme (Dean et al., 2017), the sample complexity can be upper bounded by a function which is at most exponential with the dimension n. In fact, we provide a more refined result, directly linking sample complexity to the controllability index κ. Our proof relies

on bounding control theoretic quantities like the least singular value of the controllablility Gramian. The details of the proof and the algorithm can be found in Section D.

Theorem 7 (Exponential Upper Bounds) *Consider the class $\mathscr{C}_{n,\kappa}^{\mu}$ of all μ-robustly coupled systems $S = (A, B, H)$ of dimension n and controllability index κ. Let Assumption 2 hold. Then, the sample complexity is at most exponential with κ. There exists an exploration policy π and algorithm $\hat{K}_N$ such that for any $\delta < 1$:*

$$\sup_{S \in \mathscr{C}_{n,\kappa}^{\mu}} \mathbb{P}_{S,\pi}\left(\rho(A + B\hat{K}_N) \geq 1\right) \leq \delta, \quad \text{if} \quad N\sigma_u^2 \geq \text{poly}\left(\left(\frac{M}{\mu}\right)^{\kappa}, M^{\kappa}, n, \log 1/\delta\right).$$

Assume that the constants μ and M are dimensionless. Then, our upper and lower bounds match qualitatively with respect to the dependence on κ. Theorem 7 implies that if the degree of underactuation is mild, i.e. $\kappa = O(\log n)$, then robustly coupled systems are guaranteed to be poly(n)-stabilizable. Our upper bound picks up a dependence on the quantity M/μ. Recall that M upper-bounds the norm of A. Hence, it captures a notion of sensitivity of the dynamics A to inputs/noise. In the lower bounds only the coupling term μ appears. It is an open question to prove or disprove whether the sensitivity of A affects stabilization or it is an artifact of our analysis. Another important open problem is to determine the optimal constant that multiplies κ in the exponent. Our lower bound suggests that the exponent can be at least of the order of 2 times κ. In our upper bounds, by following the proof, we get an exponent which is larger than 2.

5. Difficulty of online LQR

In the following theorem, we prove that classes of robustly coupled systems can exhibit minimax expected regret which grows at least exponentially with the dimension n. Let $\mathscr{C}_{n,\kappa}^{\mu}$ denote the class of μ-robustly coupled systems $S = (A, B, H)$ of state dimension n and controllability index κ. Define the ϵ-dilation $\mathscr{C}_{n,\kappa}^{\mu}(\epsilon)$ of $\mathscr{C}_{n,\kappa}^{\mu}$ as

$$\mathscr{C}_{n,\kappa}^{\mu}(\epsilon) \triangleq \left\{(A, B, H) : \left\| \begin{bmatrix} A - \tilde{A} & B - \tilde{B} \end{bmatrix} \right\|_2 \leq \epsilon, \text{ for some } (\tilde{A}, \tilde{B}, H) \in \mathscr{C}_{n,\kappa}^{\mu}\right\},$$

which consists of every system in $\mathscr{C}_{n,\kappa}^{\mu}$ along with its $\epsilon-$ball around it.

Theorem 8 (Exponential Regret Lower Bounds) *Consider the class $\mathscr{C}_{n,\kappa}^{\mu}$ of all μ-robustly coupled systems $S = (A, B, H)$ of state dimension n and controllability index κ, with $\kappa \leq n - 1$. For every $\epsilon > 0$ define the ϵ-dilation $\mathscr{C}_{n,\kappa}^{\mu}(\epsilon)$. Let $Q_T = P$, the solution to the ARE (7), and assume $\mu < 1$. Let $0 < \alpha < 1/4$. For any policy π*

$$\liminf_{T \to \infty} \sup_{S \in \mathscr{C}_{n,\kappa}^{\mu}(T^{-\alpha})} \mathbb{E}_{S,\pi} \frac{R_T(S)}{\sqrt{T}} \geq \frac{1}{4\sqrt{n}} 2^{\frac{\kappa - 1}{2}}.$$

When the controllability index is large, e.g. $\kappa = n$, then the lower bounds become exponential with n. Hence, achieving poly(n)-regret is impossible in the case of general linear systems. In general, learning difficulty depends on fundamental control theoretic parameters, i.e. on the solution P to the ARE (7) or the steady-state covariance of the closed-loop system, both of which can scale exponentially with the controllability index. Existing regret upper-bounds

depend on such quantities in a transparent way Simchowitz and Foster (2020). Here, we reveal the dependence on such parameters in the regret lower-bounds as well (Lemma 9).

Let us now explain when learning can be difficult. Consider the following 1−strongly coupled system, which consists of two independent subsystems

$$A = \left[\begin{array}{c|ccccc} 0 & 0 & 0 & & 0 & 0 \\ \hline 0 & 1 & 1 & & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & & 1 & 1 \\ 0 & 0 & 0 & & 0 & 1 \end{array}\right], \; B = \left[\begin{array}{c|c} 1 & 0 \\ 0 & 0 \\ \vdots & \\ 0 & 1 \end{array}\right] u_k, \; H = I_n, \; Q = I_n, \; R = I_2, \tag{15}$$

where the first subsystem is a memoryless system, while the second one is the discrete integrator of order $n-1$. Since the sub-systems are decoupled, the optimal LQR controller will also be decoupled and structured

$$K_\star = \left[\begin{array}{cc} 0 & 0 \\ 0 & K_{\star,0} \end{array}\right],$$

where $K_{\star,0}$ is the optimal gain of the second subsystem. The first subsystem (upper-left) is memoryless and does not require any regulation, that is, $[K_\star]_{11} = 0$.

Consider now a perturbed system $\tilde{A} = A - \Delta K_\star$, $\tilde{B} = B + \Delta$, for some $\Delta \in \mathbb{R}^{p\times n}$. Such perturbations are responsible for the $\sqrt{T}$ term in the regret of LQR (Simchowitz and Foster, 2020; Ziemann and Sandberg, 2022); systems (A, B) and $(\tilde{A}, \tilde{B})$ are indistinguishable under the control law $u_t = K_\star x_t$ since $A + BK_\star = \tilde{A} + \tilde{B}K_\star$. Now, informally, to get an $\exp(n)\sqrt{T}$ regret bound it is sufficient to satisfy two conditions: i) the system is sensitive to inputs or noise, in the sense that any exploratory signal can incur extra cost, which grows exponentially with n. ii) the difference $\tilde{A} - A$, $\tilde{B} - B$ is small enough, i.e. polynomial in n, so that identification of Δ requires significant deviation from the optimal policy.

The $n-1$-th integrator is very sensitive to inputs or noises. As inputs $u_{k,2}$ and noises w_k get integrated $(n-1)$-times, this will result in accumulated values that grow exponentially as we move up the integrator chain. Hence, the first informal condition is satisfied. To satisfy the second condition we let the perturbation Δ have the following structure

$$\Delta = \left[\begin{array}{cc} 0 & 0 \\ \Delta_1 & 0 \end{array}\right], \tag{16}$$

where we only perturb the matrix of the first input $u_{k,1}$. By using two subsystems and the above construction, we make it harder to detect Δ. In particular, because of the structure of the system ($[K_\star]_{11} = 0$) and the perturbation Δ, we have $\tilde{A} = A - \Delta K_\star = A$. Hence $\| \left[\begin{array}{cc} A & B \end{array}\right] - \left[\begin{array}{cc} \tilde{A} & \tilde{B} \end{array}\right] \|_2 = \|\Delta\|_2 \leq \text{poly}(n)\|\Delta\|_2$, i.e., the perturbed system does not lie too far away from the nominal one. This last condition might be crucial. If $\|\Delta K_\star\| \geq \exp(n)\|\Delta\|_2$, then it might be possible to distinguish between (A, B) and $(\tilde{A}, \tilde{B})$ without deviating too much from the optimal policy. This may happen if we use only one subsystem, since $\|K_{\star,0}\|_2$ might be large. By using two subsystems, we cancel the effect of $K_{\star,0}$ in $\Delta K_\star$.

In the stabilization problem, we show that the lack of excitation during the system identification stage might hurt sample complexity. Here, we show that if a system is too sensitive to inputs and noises, i.e. some state subspaces are too easy to excite, this can lead to large regret. Both lack of excitation and too much excitation of certain subspaces can hurt learning performance. This was observed before in control (Skogestad et al., 1988).

5.1. Sketch of Lower Bound Proof

Let $S_0 = (A_0, B_0, I_{n-1}) \in \mathscr{C}^{\mu}_{n-1,\kappa}$ be a $\mu-$robustly coupled system of state dimension $n-1$, input dimension $p-1$ and controllability index $\kappa \leq n-1$. Let P_0 be the solution of the Riccati equation for $Q_0 = I_{n-1}$, $R_0 = I_{p-1}$, with $K_{\star,0}$ the corresponding optimal gain. Define the steady-state covariance of the closed-loop system

$$\Sigma_{0,x} = (A_0 + B_0 K_{\star,0}) \Sigma_{0,x} (A_0 + B_0 K_{\star,0})' + I_{n-1}. \tag{17}$$

Now, consider the composite system:

$$A = \begin{bmatrix} 0 & 0 \\ 0 & A_0 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 \\ 0 & B_0 \end{bmatrix}, H = I_n, \tag{18}$$

with $Q = I_n$, $R = I_p$. Let Δ be structured as in (16), for some arbitrary Δ_1 of unit norm $\|\Delta_1\|_2 = 1$. The Riccati matrix of the composite system is denoted by P and the corresponding gain by $K_\star$. Consider the parameterization:

$$A(\theta) = A - \theta \Delta K_\star, \qquad B(\theta) = B + \theta \Delta, \tag{19}$$

for any $\theta \in \mathbb{R}$. Let $\mathcal{B}(\theta, \epsilon)$ denote the open Euclidean ball of radius ϵ around θ. For every $\epsilon > 0$, define the local class of systems around S as $\mathscr{C}_S(\epsilon) \triangleq \{(A(\theta), B(\theta), I_n),\ \theta \in \mathcal{B}(0, \epsilon)\}$. Based on the above construction and Theorem 1 of Ziemann and Sandberg (2022), a general information-theoretic regret lower bound, we prove the following lemma.

Lemma 9 (Two-Subsystems Lower Bound) *Consider the parameterized family of linear systems defined in (19), for $n, p \geq 2$ where Δ is structured as in (16). Let $Q = I_n$, $R = I_p$. Let $Q_T = P(\theta)$, where $P(\theta)$ is the solution to the Riccati equation for $(A(\theta), B(\theta))$. Then, for any policy π and any $0 < a < 1/4$ the expected regret is lower bounded by*

$$\lim \inf_{T \to \infty} \sup_{\hat{S} \in \mathscr{C}_S(T^{-a})} \mathbb{E}_{\hat{S},\pi} \frac{R_T(\hat{S})}{\sqrt{T}} \geq \frac{1}{4\sqrt{n}} \sqrt{\Delta_1' P_0 \left[\Sigma_{0,x} - I_{n-1}\right] P_0 \Delta_1}.$$

Optimizing over Δ_1, we obtain a lower bound on the order of $\|P_0 \left[\Sigma_{0,x} - I_{n-1}\right] P_0\|_2$. What remains to show is that for the $(n-1)$-th order integrator (second subsystem in (15)) the product $\|P_0 \left[\Sigma_{0,x} - I_{n-1}\right] P_0\|_2$ is exponentially large with n.

Lemma 10 (System Theoretic Parameters can be Large) *Consider the $(n-1)-th$ order integrator (second subsystem in (15)). Let P_0 be the Riccati matrix for $Q_0 = I_{n-1}, R_0 = 1$, with $K_{\star,0}$, $\Sigma_{0,x}$ the corresponding LQR control gain and steady-state covariance. Then*

$$\|P_0 \left[\Sigma_{0,x} - I_{n-1}\right] P_0\|_2 \geq \sum_{j=1}^{n-1} \sum_{i=0}^{j} \binom{j}{i}^2 \geq 2^{n-1}$$

Our lemma shows that control theoretic parameters can scale exponentially with the dimension n. The $(n-1)-$th order integrator is a system which is mildly unstable. In Section E.4, we show that **stable** systems can also suffer from the same issue.

5.2. Regret Upper Bounds

Similar to the stabilization problem, we show that under the assumption of robust coupling, the regret cannot be worse than $\exp(\kappa)\sqrt{T}$ with high probability. As we prove in Lemma B.2, the solution P to the Riccati equation has norm $\|P\|_2$ that scales at most exponentially with the index κ in the case of robustly-coupled systems. This result combined with the regret upper bounds of Simchowitz and Foster (2020), give us the following result.

Theorem 11 (Exponential Upper Bounds) *Consider a μ-robustly coupled system $S = (A, B, H)$ of dimension n, controllability index κ. Assume that we are given an initial stabilizing gain K_0. Let $Q = I_n$, $R \succeq I_p$, and $Q_T = 0$. Assume that the noise is non-singular $HH' = I_n$[3]. Let $\delta \in (0, 1/T)$. Using the Algorithm 1 of Simchowitz and Foster (2020) with probability at least $1 - \delta$:*

$$R_T(A, B) \leq \text{poly}(n, (\frac{M}{\mu})^{\kappa}, M^{\kappa}, \log 1/\delta)\sqrt{T} + \text{poly}(n, (\frac{M}{\mu})^{\kappa}, M^{\kappa}, \log 1/\delta, P(K_0)),$$

where $P(K_0) = (A + BK_0)'P(K_0)(A + BK) + Q + K_0'RK_0$.

The result follows immediately by our Lemma B.2 and the upper bounds of Theorem 2 in Simchowitz and Foster (2020). Assuming that the plant sensitivity M and the coupling coefficient μ are dimensionless, then if we have a mild degree of underactuation, i.e. $\kappa = O(\log n)$, we get $\text{poly}(n)$-regret with high probability. Note that the above guarantees are for high probability regret which is not always equivalent to expected regret (Dann et al., 2017). Our upper-bounds are almost global for all robustly coupled systems, in the sense that the dominant $\sqrt{T}$-term is globally bounded. To provide truly global regret guarantees it is sufficient to add an initial exploration phase to Algorithm 1 of Simchowitz and Foster (2020), which first learns a stabilizing gain K_0. For this stage we could use the results of Section 4.1, and Section D. We leave this for future work.

6. Conclusion

We prove that learning to control linear systems can be hard for non-trivial system classes. The problem of stabilization might require sample complexity which scales exponentially with the system dimension n. Similarly, online LQR might exhibit regret which scales exponentially with n. This difficulty arises in the case of underactuated systems. Such systems are structurally difficult to control; they can be very sensitive to inputs/noise or very hard to excite. If the system is robustly coupled and has a mild degree of underactuation (small controllability index), then we can guarantee that learning will be easy.

We stress that system theoretic quantities might not be dimensionless. On the contrary, they might grow very large with the dimension and dominate any $\text{poly}(n)$ terms. Hence, going forward, an important direction of future work is to find policies with optimal dependence on such system theoretic quantities. Although the optimal dependence is known for the problem of system identification (Simchowitz et al., 2018; Jedra and Proutiere, 2019), it is still not clear what is the optimal dependence in the case of control. For example, an interesting open

3. It is possible to relax some of the assumptions on the noise–see Simchowitz and Foster (2020)

problem is to find the optimal dependence of the regret R_T on the Riccati equation solution P. For the problem of stabilization, it is open to find how sample complexity optimally scales with the least singular value of the controllability Gramian.

Acknowledgment

This work was supported by the AFOSR Assured Autonomy grant.

References

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

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-Free Linear Quadratic Control via Reduction to Expert Prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3108–3117, 2019.

Marc Abeille and Alessandro Lazaric. Improved Regret Bounds for Thompson Sampling in Linear Quadratic Control Problems. *Proceedings of Machine Learning Research*, 80, 2018.

Marc Abeille and Alessandro Lazaric. Efficient Optimistic Exploration in Linear-Quadratic Regulators via Lagrangian Relaxation. *arXiv preprint arXiv:2007.06482*, 2020.

B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Dover Publications, 2005.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 4th edition, 2017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Siew Chan, GC Goodwin, and Kwai Sin. Convergence properties of the Riccati difference equation in optimal filtering of nonstabilizable systems. *IEEE Transactions on Automatic Control*, 29(2):110–118, 1984.

Xinyi Chen and Elad Hazan. Black-Box Control for Linear Dynamical Systems. In *Conference on Learning Theory*, pages 1114–1143. PMLR, 2021.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online Linear Quadratic Control. In *International Conference on Machine Learning*, pages 1029–1038. PMLR, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret. *arXiv preprint arXiv:1902.06223*, 2019.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Paul M. Van Dooren. Numerical linear algebra for signals systems and control. *Draft notes prepared for the Graduate School in Systems and Control*, 2003.

Yonathan Efroni, Sham Kakade, Akshay Krishnamurthy, and Cyril Zhang. Sparsity in Partially Controllable Linear Systems. *arXiv preprint arXiv:2110.06150*, 2021.

M Sami Fadali and Antonio Visioli. *Digital Control Engineering: Analysis and Design*. Academic Press, 2013.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite Time Identification in Unstable Linear Systems. *Automatica*, 96:342–353, 2018a.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite-time Adaptive Stabilization of Linear Systems. *IEEE Transactions on Automatic Control*, 64 (8):3498–3505, 2018b.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On Adaptive Linear–Quadratic Regulators. *Automatica*, 117:108982, 2020.

Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Learning sparse dynamical systems from a single sample trajectory. *arXiv preprint arXiv:1904.09396*, 2019.

Michel Gevers. Identification for Control: From the Early Achievements to the Revival of Experiment Design. *European journal of control*, 11(4-5):335–352, 2005.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Yassir Jedra and Alexandre Proutiere. Sample complexity lower bounds for linear system identification. In *IEEE 58th Conference on Decision and Control (CDC)*, pages 2676–2681. IEEE, 2019.

Yassir Jedra and Alexandre Proutiere. Minimal Expected Regret in Linear Quadratic Control. *arXiv preprint arXiv:2109.14429*, 2021.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information Theoretic Regret Bounds for Online Nonlinear Control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Explore more and improve regret in Linear Quadratic Regulators. *arXiv preprint arXiv:2007.12291*, 2020a.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *arXiv preprint arXiv:2003.11227*, 2020b.

Bruce Lee and Andrew Lamperski. Non-asymptotic Closed-Loop System Identification using Autoregressive Processes and Hankel Model Reduction. *arXiv preprint arXiv:1909.02192*, 2019.

Holden Lee. Improved rates for identification of partially observed linear dynamical systems. *arXiv preprint arXiv:2011.10006*, 2020.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of LQR is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Nikolai Matni and Stephen Tu. A tutorial on concentration bounds for system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3741–3749. IEEE, 2019.

Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From Self-Tuning Regulators to Reinforcement Learning and Back Again. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3724–3740. IEEE, 2019.

Ronald Ortner and Daniil Ryabko. Online Regret Bounds for Undiscounted Continuous Reinforcement Learning. *Advances in Neural Information Processing Systems*, 25, 2012.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of Unknown Linear Systems with Thompson Sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE, 2017.

Samet Oymak and Necmiye Ozay. Non-asymptotic Identification of LTI Systems from a Single Trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

Fabio Pasqualetti, Sandro Zampieri, and Francesco Bullo. Controllability Metrics, Limitations and Algorithms for Complex Networks. *IEEE Transactions on Control of Network Systems*, 1(1):40–52, 2014.

Benjamin Recht. A Tour of Reinforcement Learning: The View from Continuous Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279, 2019.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. *arXiv preprint arXiv:1812.01251*, 2018.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-Time System Identification for Partially Observed LTI Systems of Unknown Order. *arXiv preprint arXiv:1902.01848*, 2019.

Max Simchowitz and Dylan Foster. Naive Exploration is Optimal for Online LQR. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification. *arXiv preprint arXiv:1802.08334*, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning Linear Dynamical Systems with Semi-Parametric Least Squares. *arXiv preprint arXiv:1902.00768*, 2019.

Sigurd Skogestad, Manfred Morari, and John C Doyle. Robust Control of Ill-Conditioned Plants: High-Purity Distillation. *IEEE transactions on automatic control*, 33(12):1092–1105, 1988.

Anastasios Tsiamis and George J Pappas. Finite Sample Analysis of Stochastic System Identification. In *IEEE 58th Conference on Decision and Control (CDC)*, 2019.

Anastasios Tsiamis and George J. Pappas. Linear Systems can be Hard to Learn. *arXiv preprint arXiv:2104.01120*, 2021.

Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-Asymptotic Analysis of Robust Control from Coarse-Grained Identification. *arXiv preprint arXiv:1707.04791*, 2017.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2020.

Ingvar Ziemann and Henrik Sandberg. Regret Lower Bounds for Learning Linear Quadratic Gaussian Systems. *arXiv preprint arXiv:2201.01680*, 2022.

Contents

1 Introduction **1**
1.1 Related Work . . . 3
1.2 Notation . . . 4

2 Problem Statement **5**
2.1 Difficulty of Stabilization . . . 5
2.2 Difficulty of Online LQR . . . 6

3 Classes with Rich Controllability Structure **6**

4 Difficulty of Stabilization **8**
4.1 Sample complexity upper bounds . . . 9

5 Difficulty of online LQR **10**
5.1 Sketch of Lower Bound Proof . . . 12
5.2 Regret Upper Bounds . . . 13

6 Conclusion **13**

Appendix **19**

A System Theoretic Preliminaries **19**
A.1 Properties of the Riccati Equation . . . 19

B System Theoretic Bounds for Robustly Coupled Systems **20**

C Lower Bounds for the problem of Stabilization **24**
C.1 Proof of Theorem 5 . . . 24

D Upper Bounds for the problem of Stabilization **26**
D.1 Algorithm . . . 27
D.2 System Identification Analysis . . . 27
D.3 Sensitivity of Stabilization . . . 29
D.4 Proof of Theorem 7 . . . 30

E Regret Lower Bounds **30**
E.1 Proof of Lemma 9 . . . 32
E.2 Proof of Lemma 10 . . . 33
E.3 Proof of Theorem 8 . . . 33
E.4 Stable System Example . . . 34

Appendix A. System Theoretic Preliminaries

In this section, we review briefly some system theoretic concepts. A system $(A, B) \in \mathbb{R}^{n\times(n+p)}$ is **controllable** if and only if the controllability matrix

$$\mathcal{C}_k(A, B) = \begin{bmatrix} B & AB & \cdots & A^{k-1}B \end{bmatrix}$$

has full column rank for some $k \leq n$. The minimum such index κ that the rank condition is satisfied is called the controllability index, and it is always less or equal than the state dimension n. A system (A, B) is called **stabilizable** if and only if there exists a matrix $K \in \mathbb{R}^{p\times n}$ such that $A + BK$ is stable, i.e. has spectral radius $\rho(A + BK)$. Any controllable system is also stabilizable. A system (A', B') is called **observable** if and only if (A, B) is controllable. Similarly (A', B') is **detectable** if and only if (A, B) is stabilizable.

Let A be stable ($\rho(A) < 1$) and consider the transfer matrix $(zI - A)^{-1}, z \in \mathcal{C}$ in the frequency domain. The $\mathcal{H}_\infty$-norm is given by

$$\|(zI - A)^{-1}\|_{\mathcal{H}_\infty} = \sup_{|z|=1} \|(zI - A)^{-1}\|_2.$$

Using the identity $(I - D)^{-1} = I + D + D^2 \dots$ for $\rho(D) < 1$, we can upper bound the $\mathcal{H}_\infty$-norm by

$$\|(zI - A)^{-1}\|_{\mathcal{H}_\infty} \leq \sum_{t=0}^{\infty} \|A^t\|_2.$$

A.1. Properties of the Riccati Equation

Consider the infinite horizon LQR problem defined in (2). Let (A, B) be controllable and assume that $Q \succ 0$ is positive semi-definite and $R \succ 0$ is positive definite. As we stated in Section 2, the optimal policy $K_\star x_k$ has the following closed-form solution

$$K_\star = -(B'PB + R)^{-1}B'PA,$$

where P is the unique positive definite solution to the **Discrete Algebraic Riccati Equation**

$$P = A'PA + Q - A'PB(B'PB + R)^{-1}B'PA.$$

Moreover, $A + BK_\star$ is stable, i.e. $\rho(A + BK_\star) < 1$. The above solution is well-defined under the conditions of (A, B) controllable, $Q \succ 0$, $R \succ 0$. Note that we can relax the conditions to $Q \succeq 0$ being positive semi-definite, $(A, Q^{1/2})$ detectable, and (A, B) stabilizable, which is a well-known result in control theory (Chan et al., 1984, Th. 3.1).

Consider now the **finite-horizon** LQR problem, under the same assumptions of (A, B) controllable, $Q \succ 0$, and $R \succ 0$

$$J_T^*(S) \triangleq \min_\pi \mathbb{E}_{S,\pi} \left[\sum_{t=0}^{T-1} (x_t'Qx_t + u_t'Ru_t) + x_T'Q_Tx_T \right]. \tag{A.1}$$

The optimal policy is a feedback law $K_t x_t$, $t \leq T - 1$, with time varying gains. The gains satisfy the following closed-form expression

$$K_t = -(B'P_{t+1}B + R)^{-1}B'P_{t+1}A,$$

where P_t satisfies the **Riccati Difference Equation**

$$P_t = A'P_{t+1}A + Q - A'P_{t+1}B(B'P_{t+1}B + R)^{-1}B'P_{t+1}A,\ P_T = Q_T.$$

It turns out that as we take the horizon to infinity $T \to \infty$, then we get $\lim_{T\to\infty} P_k = P$ exponentially fast, for any fixed k, where P is the positive definite solution to the Algebraic Riccati Equation. The convergence is true under the conditions of (A, B) controllable, $Q \succ 0$, $R \succ 0$. Again we could relax the conditions to $Q \succeq 0$ being positive semi-definite, $(A, Q^{1/2})$ detectable, and (A, B) stabilizable (Chan et al., 1984, Th. 4.1). Note that if we select the terminal cost $Q_T = P$, then trivially $P_t = P$ for all $t \leq T$, and we recover the same controller as in the infinite horizon case.

Finally, a nice property of the Riccati recursion is that the right-hand side is order-preserving with respect to the matrices P, Q. In particular, define the operator:

$$g(X, Y) = A'XA + Y - A'YB(B'XB + R)^{-1}B'YA.$$

Then, if $X_1 \succeq X_2$, we have that $g(X_1, Y) \succeq g(X_2, Y)$ (Anderson and Moore, 2005, Ch. 4.4). Similarly, if $Y_1 \succeq Y_2$ then $g(X, Y_1) \succeq g(X, Y_2)$.

Appendix B. System Theoretic Bounds for Robustly Coupled Systems

The first result lower bounds the least singular value of the controllability Gramian in terms of the sensitivity M, the coupling coefficient μ, and the controllability index κ of the system.

Theorem B.1 (Gramian lower bound (Tsiamis and Pappas, 2021)) *Consider a system (A, B, H) that satisfies Assumption 1, with κ its controllability index. Assume that (A, B) is μ-robustly coupled. Then, the least singular value of the Gramian $\Gamma_\kappa = \Gamma_\kappa(A, B)$ is lower bounded by:*

$$\sigma_{\min}^{-1}(\Gamma_\kappa) \leq \mu^{-2}\left(\frac{3M}{\mu}\right)^{2\kappa}.$$

Proof The result follows from Theorem 5 in Tsiamis and Pappas (2021). The theorem statement requires a different condition, called robust controllability. However, the proof still goes through if we have $\mu-$robust coupling instead. Recall that $\mathcal{C}_\kappa = \mathcal{C}_\kappa(A, B)$ is the controllability matrix (9) of (A, B) at κ. Following the proof in (Tsiamis and Pappas, 2021), we arrive at

$$\sqrt{\sigma_{\min}(\Gamma_\kappa)} \leq \|\mathcal{C}_\kappa^\dagger\|_2 \leq \|\Xi^{\kappa-1}\|_2\|\alpha\|_2,$$

where

$$\Xi = \begin{bmatrix} 1 & 1 & \mu^{-1} \\ \frac{M}{\mu} & \frac{2+M}{\mu} & \frac{M}{\mu} \\ 0 & 0 & \mu^{-1} \end{bmatrix},\ \alpha = \begin{bmatrix} \frac{1}{\mu} \\ \frac{M}{\mu^2} \\ \frac{1}{\mu} \end{bmatrix}.$$

The result follows from the crude bounds $\|\Xi\|_2 \leq 3M/\mu$, $\|\alpha\|_2 \leq \sqrt{3}M/\mu^{-2}$ where we assumed that $M > 1$. ∎

The following result, upper bounds the solution P to the LQR Riccati equation in terms of the sensitivity M, the coupling coefficient μ, and the controllability index κ of the system.

Lemma B.2 (Riccati Upper Bounds) *Let the system $(A, B) \in \mathbb{R}^{n\times(n+p)}$ be controllable and $\mu-$robustly coupled with controllability index κ. Let $R \in \mathbb{R}^{p\times p}$ be positive definite and $Q \in \mathbb{R}^{n\times n}$ be positive semi-definite. Assume $T > \kappa$ and consider the Riccati difference equation:*

$$P_{k-1} = A'P_kA + Q - A'P_kB(B'P_kB + R)^{-1}B'P_kA,\ P_T = Q.$$

Then, the Riccati matrix evaluated at time 0 is upper-bounded by

$$\|P_0\|_2 \le \mathrm{poly}\Big((\frac{M}{\mu})^{\kappa}, M^{\kappa}, \kappa, \|Q\|_2, \|R\|_2\Big).$$

As a result, if $Q \succ 0$, then the unique positive definite solution P of the algebraic Riccati equation:

$$P = A'PA + Q - A'PB(B'PB + R)^{-1}B'PA$$

satisfies the same bound

$$\|P\|_2 \le \mathrm{poly}\Big((\frac{M}{\mu})^{\kappa}, M^{\kappa}, \kappa, \|Q\|_2, \|R\|_2\Big).$$

Proof The optimal policy of the LQR problem does not depend on the noise. Even for deterministic systems, the optimal policy still have the same form $u_t = K_\star x_t$. This property is known as certainty equivalence (Bertsekas, 2017, Ch. 4). In fact, for deterministic systems, the cost of regulation is given explicitly by $x_0'Px_0$. We leverage this idea to upper bound the stabilizing solution of the Riccati equation P.

Step a) Noiseless system upper bound. Consider the noiseless version of system (1)

$$x_{k+1} = Ax_k + Bu_k, \quad \|x_0\|_2 = 1. \tag{B.1}$$

Let $u_{0:t}$ be the shorthand notation for

$$u_{0:t} = \begin{bmatrix} u_t \\ \vdots \\ u_0 \end{bmatrix}.$$

Consider the deterministic LQR objective

$$\begin{aligned} \min_{u_{0:T-1}} \quad & J(u_{0:T-1}) \triangleq x_T'Qx_T + \sum_{k=0}^{N-1} x_k'Qx_k + u_k'Ru_k \\ \text{s.t.} \quad & \text{dynamics (B.1)}. \end{aligned}$$

The optimal cost of the problem is given by (Bertsekas, 2017, Ch. 4)

$$\min_{u_{0:T-1}} J(u_{0:T-1}) = x_0'P_0x_0,$$

where P_0 is the value of P_t at time $t = 0$. Let $u_{0:T-1}$ be any input sequence. Immediately, by optimality, we obtain an upper bound for the Riccati matrix P_0:

$$x_0'P_0x_0 \le J(u_{0:T-1}). \tag{B.2}$$

Hence, it is sufficient to find a suboptimal policy that incurs a cost which is at most exponential with the controllability index κ.

Step b) Suboptimal Policy. It is sufficient to drive the state x_κ to zero at time κ with minimum energy $u_{0:\kappa-1}$ and then keep $x_{t+1}=0$, $u_t=0$, for $t\geq\kappa$. Recall that $\mathcal{C}_k$ is the controllability matrix at time k. By unrolling the state x_κ:

$$x_\kappa = A^\kappa x_0 + \mathcal{C}_\kappa u_{0:\kappa-1}.$$

To achieve $x_\kappa=0$, it is sufficient to apply the minimum norm control

$$u_{0:\kappa-1} = -\mathcal{C}_\kappa^\dagger A^\kappa x_0,$$

which leads to input penalties

$$\sum_{k=0}^{T-1} u_k' R u_k \leq \|R\|_2 \sigma_{\min}^{-1}(\Gamma_\kappa) M^{2\kappa},$$

where we used the fact that $\|x_0\|_2=1$. For the state penalties, we can write in batch form

$$x_{1:\kappa} \triangleq \begin{bmatrix} x_\kappa \\ \vdots \\ x_1 \end{bmatrix} = \begin{bmatrix} B & AB & \cdots & A^{\kappa-1}B \\ 0 & B & \cdots & A^{\kappa-2}B \\ \vdots & & & \\ 0 & 0 & \cdots & B \end{bmatrix} u_{0:\kappa-1} + \begin{bmatrix} A^\kappa \\ A^{\kappa-1} \\ \vdots \\ A \end{bmatrix} x_0.$$

Exploiting the Toeplitz structure of the first matrix above and by Cauchy-Schwartz

$$\begin{aligned}\sum_{t=0}^{T} x_t' Q x_t &\leq \|Q\|_2(\|x_{1:\kappa}\|_2^2+1) \\ &\leq 2\|Q\|_2\big((\sum_{t=0}^{\kappa-1}\|A^tB\|_2)^2\|u_{0:\kappa-1}\|_2^2 + \sum_{t=0}^{\kappa}\|A^t\|_2\big) \\ &\leq 2\kappa^2\|Q\|_2(M^{4\kappa}\|R\|_2\sigma_{\min}^{-1}(\Gamma_\kappa)+M^{2\kappa}).\end{aligned}$$

Putting everything together and since x_0 is arbitrary, we finally obtain

$$\|P_0\|_2 \leq \frac{\|R\|_2}{\sigma_{\min}(\Gamma_\kappa)}(M^{2\kappa}+2\kappa^2\|Q\|_2M^{4\kappa}) + 2\kappa^2\|Q\|_2M^{2\kappa}. \tag{B.3}$$

The result for P_0 now follows from Theorem B.1.

Step c) Steady State Riccati. If the pair $(A, Q^{1/2})$ is observable, then from standard LQR theory-see Section A.1, $\lim_{T\to\infty} P_0 = P$ and the bound for P follows directly. ■

Similar results have been reported before (Cohen et al., 2018; Chen and Hazan, 2021). However, instead of κ and $(M/\mu)^\kappa$, the least singular value $\sigma_{\min}^{-1}(\Gamma_k)$ shows up in the bounds, for some $k\geq\kappa$.

Finally, based on Lemmas B.10, B.11 of Simchowitz and Foster (2020), we provide some upper bounds on the $\mathcal{H}_\infty-$norm of the closed loop response $(zI-A+BK)^{-1}$, where K is the control gain of the optimal LQR controller for some Q and R.

Lemma B.3 (LQR Robustness Margins) *Let the system $(A, B) \in \mathbb{R}^{n\times(n+p)}$ be controllable and $\mu-$robustly coupled. Let $R = I_p$, $Q = I_n$. Let P be the stabilizing solution of the algebraic Riccati equation:*

$$P = A'PA + Q - A'PB(B'PB + R)^{-1}B'PA$$

with $K_\star$ the respective control gain $K_\star = -(B'PB + R)^{-1}B'PA$. The spectral radius and the $\mathcal{H}_\infty$-norm of the closed loop response are upper bounded by

$$(1 - \rho(A + BK_\star))^{-1} \leq \text{poly}\Big((\frac{M}{\mu})^\kappa, M^\kappa, \kappa\Big) \tag{B.4}$$

$$\|(zI - A - BK_\star)^{-1}\|_{\mathcal{H}_\infty} \leq \text{poly}\Big((\frac{M}{\mu})^\kappa, M^\kappa, \kappa\Big) \tag{B.5}$$

Proof First, note that since $Q = I$, immediately $(A, Q^{1/2})$ is observable and the stabilizing solution P is well-defined. Note that the Riccati solution P also satisfies the Lyapunov equation

$$P = (A + BK_\star)'P(A + BK_\star) + I + K_\star'K_\star \succeq (A + BK_\star)'P(A + BK_\star) + I \succeq I.$$

As a result,

$$(A + BK_\star)'(A + BK_\star) \overset{i)}{\preceq} (A + BK_\star)'P(A + BK_\star) = P - I \overset{ii)}{\preceq} (1 - \|P\|_2^{-1})P, \tag{B.6}$$

where i) follows from $P \succeq I$. To prove ii) observe that $P - I = P^{1/2}(I - P^{-1})P^{1/2}$ and $P^{-1} \succeq \|P\|_2^{-1}I$. Hence

$$P - I \succeq P^{1/2}(I - \|P\|_2^{-1}I)P^{1/2} = (1 - \|P\|_2^{-1})P.$$

Applying inequality (B.6) recursively

$$(A + BK_\star)^{t'}(A + BK_\star)^t = \|(A + BK_\star)^t\|_2^2 \leq \big(1 - \|P\|_2^{-1}\big)^t P.$$

From here, we immediately deduce that

$$\rho(A + BK_\star) \leq \sqrt{1 - \|P\|_2^{-1}},$$

which by Lemma B.2 proves (B.4). For the $\mathcal{H}_\infty$ norm bound

$$\begin{aligned}\|(zI - A - BK_\star)^{-1}\|_{\mathcal{H}_\infty} &\leq \sum_{t\geq 0} \|(A + BK_\star)^t\|_2 \leq \|P\|_2^{1/2} \frac{1}{1 - \sqrt{1 - \|P\|_2^{-1}}} \\ &\leq \|P\|_2^{1/2} \frac{1 + \sqrt{1 - \|P\|_2^{-1}}}{\|P\|_2^{-1}} \leq 2\|P\|_2^{3/2}.\end{aligned}$$

The proof of (B.5) now follows from Lemma B.2. ■

Appendix C. Lower Bounds for the problem of Stabilization

In this section, we prove Theorem 5 by using information theoretic methods. The main idea is to find systems that are nearly indistinguishable from data but require completely different stabilization schemes. We rely on Birgé's inequality (Boucheron et al., 2013), which we review below for convenience.

Definition C.1 (KL divergence) *Let $\mathbb{P}$, $\mathbb{Q}$ be two probability measures on some space $(\Omega, \mathcal{A})$. Let $\mathbb{Q}$ be absolutely continuous with respect to $\mathbb{P}$, that is $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(Y 1_A)$ for some integrable non-negative random variable with $\mathbb{E}_{\mathbb{P}}(Y) = 1$. The KL divergence $D(\mathbb{Q}||\mathbb{P})$ is given by*

$$D(\mathbb{Q}||\mathbb{P}) \triangleq \mathbb{E}_{\mathbb{Q}}(\log Y).$$

Theorem C.2 (Birgé's Inequality (Boucheron et al., 2013)) *Let $\mathbb{P}_0$, $\mathbb{P}_1$ be probability measures on $(\Omega, \mathcal{E})$ and let E_0, $E_1 \in \mathcal{E}$ be disjoint events. If $1-\delta \triangleq \min_{i=0,1} \mathbb{P}_i(E_i) \geq 1/2$ then*

$$(1-\delta)\log\frac{1-\delta}{\delta} + \delta \log \frac{\delta}{1-\delta} \leq D(\mathbb{P}_1||\mathbb{P}_0).$$

The KL divergence between two Gaussian distributions with same variance is given below.

Lemma C.3 (Gaussian KL divergence) *Let $\mathbb{P} = \mathcal{N}(\mu_1, \sigma^2)$ and $\mathbb{Q} = \mathcal{N}(\mu_2, \sigma^2)$ then*

$$D(\mathbb{Q}||\mathbb{P}) = \frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2.$$

C.1. Proof of Theorem 5

It is sufficient to prove it for $\kappa = n$. The proof for $\kappa < n$ is similar. Let $\alpha > 0$ be such that $\alpha + \mu < 1$. Consider the systems:

$$S_1: \quad x_{k+1} = \begin{bmatrix} 1 & \mu & 0 & \cdots & 0 \\ 0 & \alpha & \mu & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & 0 & 0 & \cdots & \mu \\ 0 & 0 & 0 & \cdots & \alpha \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \mu \end{bmatrix} u_k + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} w_k,$$

$$S_2: \quad x_{k+1} = \begin{bmatrix} 1 & -\mu & 0 & \cdots & 0 \\ 0 & \alpha & \mu & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & 0 & 0 & \cdots & \mu \\ 0 & 0 & 0 & \cdots & \alpha \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \mu \end{bmatrix} u_k + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} w_k.$$

By construction, the systems are $\mu-$robustly coupled. Denote the state matrices by A_1, A_2 for S_1, S_2 respectively. Let $\phi_1(z) = \det(zI - A_1 - B\hat{K}_N)$, $\phi_2(z) = \det(zI - A_2 - B\hat{K}_N)$ be the respective characteristic polynomials. By Jury's criterion (Fadali and Visioli, 2013, Ch. 4.5), a necessary (but not sufficient) condition for stability is:

$$\phi_1(1) > 0, \ \phi_2(1) > 0.$$

An direct computation gives:

$$\phi_1(1) = \begin{vmatrix} 0 & -\mu & 0 & \cdots & 0 \\ 0 & 1-\alpha & -\mu & \cdots & 0 \\ & & \ddots & \ddots & \\ 0 & 0 & 0 & \cdots & -\mu \\ -\hat{K}_{N,1} & -\hat{K}_{N,2} & -\hat{K}_{N,3} & \cdots & 1-\alpha-\hat{K}_{N,n} \end{vmatrix} = -\hat{K}_{N,1}\mu^{n-1},\ \phi_2(1) = \hat{K}_{N,1}\mu^{n-1}.$$

As a result, the events:

$$E_1 = \left\{\rho(A_1 + B\hat{K}_N) < 1\right\} \subseteq \left\{\hat{K}_{N,1} < 0\right\}, \quad E_2 = \left\{\rho(A_2 + B\hat{K}_N) < 1\right\} \subseteq \left\{\hat{K}_{N,1} > 0\right\}$$

are disjoint. By Theorem C.2, a necessary condition for stabilizing both systems with probability larger than $1-\delta$ is:

$$D(\mathbb{P}_1||\mathbb{P}_2) \geq (1-2\delta)\log\frac{1-\delta}{\delta} \geq \log\frac{1}{2.4\delta} \geq \log\frac{1}{3\delta}. \tag{C.1}$$

Here $\mathbb{P}_i$ is a shorthand notation for $\mathbb{P}_{S_i,\pi}$, for $i = 1, 2$.

Meanwhile, by the chain rule of KL divergence (see Exercise 4.4 in Boucheron et al. (2013)):

$$\begin{aligned} D(\mathbb{P}_1||\mathbb{P}_2) = \mathbb{E}_{\mathbb{P}_1}\Big(& D(\mathbb{P}_1(\mathrm{AUX})||\mathbb{P}_2(\mathrm{AUX})) \\ &+ \sum_{k=0}^{N} D(\mathbb{P}_1(x_k|x_{0:k-1}, u_{0:k-1}, \mathrm{AUX})||\mathbb{P}_2(x_k|x_{0:k-1}, u_{0:k-1}, \mathrm{AUX})) \\ &+ \sum_{k=0}^{N-1} D(\mathbb{P}_1(u_k|x_{0:k}, u_{0:k-1}, \mathrm{AUX})||\mathbb{P}_2(u_k|x_{0:k}, u_{0:k-1}, \mathrm{AUX})\Big), \end{aligned}$$

where $x_{0:k}$ is a shorthand notation for $x_0, \ldots, x_k$ (same for $u_{0:k}$). By $\mathbb{P}(X|Y)$ we denote the conditional distribution of X given Y. Note that the inputs have the same conditional distributions under both measures hence their KL divergence is zero. As a result

$$\begin{aligned} D(\mathbb{P}_1||\mathbb{P}_2) &= \mathbb{E}_{\mathbb{P}_1} \sum_{k=0}^{N} D(\mathbb{P}_1(x_k|x_{0:k-1}, u_{0:k-1}, \mathrm{AUX})||\mathbb{P}_2(x_k|x_{0:k-1}, u_{0:k-1}, \mathrm{AUX})) \\ &\overset{1)}{=} \mathbb{E}_{\mathbb{P}_1} \sum_{k=0}^{N} D(\mathbb{P}_1(x_k|x_{k-1}, u_{k-1})||\mathbb{P}_2(x_k|x_{k-1}, u_{k-1}) \\ &\overset{2)}{=} \mathbb{E}_{\mathbb{P}_1} \sum_{k=0}^{N} D(\mathbb{P}_1(x_{k,1}|x_{k-1,1}, x_{k-1,2})||\mathbb{P}_2(x_{k,1}|x_{k-1,1}, x_{k-1,2})\Big), \end{aligned}$$

where 1) follows from the Markov property of the linear system and 2) follows from an application of the chain rule, the structure of the dynamics, and the fact that all $x_{k,j}$ have the same distribution for $j \geq 2$. Recall that the normal distribution is denoted by $\mathcal{N}(\mu, \Sigma)$. Now we can explicitly compute the KL divergence:

$$D(\mathbb{P}_1||\mathbb{P}_2) = \mathbb{E}_{\mathbb{P}_1} \sum_{k=1}^{N} D(\mathcal{N}(\alpha x_{k-1,1} + \mu x_{k-1,2}, 1)||\mathcal{N}(\alpha x_{k-1,1} - \mu x_{k-1,2}, 1))$$

$$\overset{i)}{=} \mathbb{E}_{\mathbb{P}_1} \sum_{k=1}^{N} 2\mu^2 x_{k-1,2}^2 = 2\mu^2 \sum_{k=1}^{N} \mathbb{E}_{\mathbb{P}_1} x_{k-1,2}^2, \tag{C.2}$$

where $i)$ follows by Lemma C.3. By (C.1), (C.2), and Lemma C.4, it is necessary to have

$$N\sigma_u^2 \geq \frac{1}{2} \left(\frac{1}{\alpha + \mu} \right)^{2n-2} \left(\frac{1 - a - \mu}{\mu} \right)^2 \log \frac{1}{3\delta}$$

Since we are free to choose α, it is sufficient to choose $\alpha = 0$. ■

Lemma C.4 *Consider system S_1 as defined above. Recall that $\mathbb{P}_1$ is a shorthand notation for $\mathbb{P}_{S_1,\pi}$. Then, under Assumption 2, we have*

$$\mathbb{E}_{\mathbb{P}_1} x_{k,2}^2 \leq \sigma_u^2 (\alpha + \mu)^{2n-2} \left(\frac{1}{1 - (a + \mu)} \right)^2$$

Proof Let e_2 denote the canonical vector $e_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}'$. Then

$$x_{k,2} = \sum_{t=1}^{k} e_2' A^{t-1} B u_{k-t} = \sum_{t=n-1}^{k} e_2' A^{t-1} B u_{k-t},$$

where the second equality follows from the fact that $e_2' A^{t-1} B$, for $t \leq n-1$. Moreover, we can upper bound:

$$\left| e_2' A^{t-1} B \right| \leq (\alpha + \mu)^{t-1},$$

which follows from the fact that the sub-matrix $[A_1]_{2:n,2:n}$ of A_1 if we delete the first row and column is bi-diagonal and Toeplitz hence $\|[A_1]_{2:n,2:n}\|_2 \leq \alpha + \mu$. Define $c_t \triangleq (\alpha + \mu)^{t-1}$. Then, we can upper bound $|x_{k,2}|$ by

$$|x_{k,2}| \leq \sum_{t=n-1}^{k} c_t \left| u_{k-t} \right|.$$

By Cauchy-Schwartz and Assumption 2

$$\mathbb{E}_{S_1,\pi} u_k^2 \leq \sigma_u^2, \quad \mathbb{E}_{S_1,\pi} \left| u_k u_t \right| \leq \sigma_u^2.$$

Finally, combining the above results

$$\mathbb{E}_{S_1,\pi} x_{k,2}^2 \leq \sigma_u^2 (\sum_{t=n-1}^{k} c_t)^2 \leq \sigma_u^2 (\alpha + \mu)^{2n-2} \left(\frac{1}{1 - (a + \mu)} \right)^2,$$

which completes the proof. ■

Appendix D. Upper Bounds for the problem of Stabilization

We employ a naive passive learning algorithm, where we employ a white-noise exploration policy to excite the state. Our gain design proceeds in two parts. First, we perform system identification based on least squares (Simchowitz et al., 2018). Second, we use robust control to design the gain based on the identified model and bounds on the identification error of A and B, similar to Dean et al. (2017).

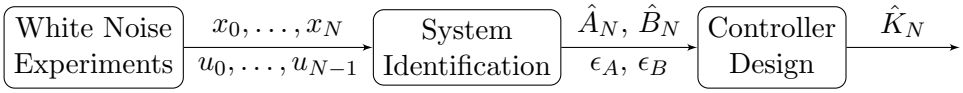

Figure 1: The block diagram of the stabilization scheme. First, we generate white noise inputs $u_t \sim \mathcal{N}(0, \bar{\sigma}_u^2 I)$ to excite the system. Then we perform system identification based on least squares to obtain estimates $\hat{A}_N, \hat{B}_N$ of the true system matrices. Finally, we design a controller gain $\hat{K}_N$, based on the system estimates and upper bounds ϵ_A, ϵ_B on the estimation error.

D.1. Algorithm

The block diagram for the algorithm is shown in Fig. 1. To generate the input data $u_0, \ldots, u_{N-1}$, we employ white noise inputs $u_k \sim \mathcal{N}(0, \bar{\sigma}_u^2 I)$, $\bar{\sigma}_u^2 = \sigma_u^2/p$, where we normalize with p in order to satisfy Assumption 2. For the system identification part, we use a least squares algorithm

$$\begin{bmatrix} \hat{A}_N & \hat{B}_N \end{bmatrix} = \arg \min_{\{F \in \mathbb{R}^{n \times n}, G \in \mathbb{R}^{n \times p}\}} \sum_{t=0}^{N-1} \|x_{t+1} - Fx_t - Gu_t\|_2^2, \tag{D.1}$$

to obtain estimates of the matrices $A\,, B$. Now, let ϵ_A, ϵ_B be large enough constants such that $\|A - \hat{A}_N\|_2 \leq \epsilon_A$, $\|B - \hat{B}_N\|_2 \leq \epsilon_B$. To design the controller gain $\hat{K}_N$, it is sufficient to solve the following problem

$$\begin{aligned} \text{find} \quad & K \in \mathbb{R}^{p \times n} \\ \text{s.t.} \quad & \left\| \begin{bmatrix} \sqrt{2}\epsilon_A (zI - \hat{A}_N - \hat{B}_N K)^{-1} \\ \sqrt{2}\epsilon_B K (zI - \hat{A}_N - \hat{B}_N K)^{-1} \end{bmatrix} \right\|_{\mathcal{H}_\infty} < 1. \end{aligned} \tag{D.2}$$

The idea behind the scheme is the following. Let $\hat{K}_N$ be a gain that stabilizes the estimated plant $(\hat{A}_N, \hat{B}_N)$. To make sure that it also stabilizes the nominal plant (A, B) we impose some additional robustness conditions. In fact, as we show in Theorem D.2, any feasible gain of problem (D.2) will stabilize any plant $(\hat{A}, \hat{B})$ that satisfies $\|\hat{A} - \hat{A}_N\|_2 \leq \epsilon_A$, $\|\hat{B} - \hat{B}_N\|_2 \leq \epsilon_B$, including the nominal one. In this work, we do not study how to efficiently solve (D.2). For efficient implementations one can refer to Dean et al. (2017). Note that the certainty equivalent LQR design (Mania et al., 2019) or the SDP relaxation method (Cohen et al., 2018; Chen and Hazan, 2021) could also work as stabilization schemes.

D.2. System Identification Analysis

Here we review a fundamental system identification result from Simchowitz et al. (2018). The original proof can be easily adapted to the case of singular noise matrices H (Tsiamis and Pappas, 2021).

Theorem D.1 (Identification Sample Complexity) *Consider a system $S = (A, B, H)$ such that Assumption 1 is satisfied. Let (A, B) be controllable with $\Gamma_k = \Gamma_k(A, B)$ the*

respective controllability Gramian and $\kappa = \kappa(A,B)$ *the respective controllability index. Then, under the least squares system identification algorithm* (D.1) *and white noise inputs* $u_k \sim \mathcal{N}(0, \bar{\sigma}_u^2 I_p)$*, we obtain*

$$\mathbb{P}_{S,\pi}(\| \begin{bmatrix} A - \hat{A}_N & B - \hat{B}_N \end{bmatrix} \|_2 \geq \epsilon) \leq \delta$$

if we have a large enough sample size

$$N\bar{\sigma}_u^2 \geq \frac{\text{poly}(n, \log 1/\delta, M)}{\epsilon^2 \sigma_{\min}(\Gamma_\kappa)} \log N.$$

Proof The proof is almost identical to the one of Theorem 4 in Tsiamis and Pappas (2021). The difference is that here we consider only the Gramian and index of (A,B) in the final bound, while in Tsiamis and Pappas (2021) the Gramian and index of $(A \begin{bmatrix} H & B \end{bmatrix})$ appears. We repeat the proof here to avoid notation ambiguity. Our goal is to apply Theorem 2.4 in (Simchowitz et al., 2018). Define the noise-controllability Gramian $\Gamma_t^h = \Gamma_t(A,H)$ as well as the combined controllability Gramian

$$\Gamma_t^c = \Gamma_t(A, \begin{bmatrix} \bar{\sigma}_u B & H \end{bmatrix}) = \bar{\sigma}_u^2 \Gamma_t + \Gamma_t^h.$$

Define $y_k = \begin{bmatrix} x_k' & u_k' \end{bmatrix}'$. It follows that for all $j \geq 0$ and all unit vectors $v \in \mathbb{R}^{(n+p)\times 1}$, the following small-ball condition is satisfied:

$$\frac{1}{2\kappa} \sum_{t=0}^{2\kappa} \mathbb{P}(\left|v' y_{t+j}\right| \geq \sqrt{v' \Gamma_{\text{sb}} v} | \bar{\mathcal{F}}_j) \geq \frac{3}{20}, \tag{D.3}$$

where

$$\Gamma_{\text{sb}} = \begin{bmatrix} \Gamma_\kappa^c & 0 \\ 0 & \bar{\sigma}_u^2 I_p \end{bmatrix}. \tag{D.4}$$

Equation (D.3) follows from the same steps as in Proposition 3.1 in Simchowitz et al. (2018) with the choice $k = 2\kappa$.

Next, we determine an upper bound $\bar{\Gamma}$ for the gram matrix $\sum_{t=0}^{N-1} y_t y_t'$. Using a Markov inequality argument as in (Simchowitz et al., 2018, proof of Th 2.1), we obtain that

$$\mathbb{P}(\sum_{t=0}^{N-1} y_t y_t' \preceq \bar{\Gamma}) \geq 1 - \delta,$$

where

$$\bar{\Gamma} = \frac{n+p}{\delta} N \begin{bmatrix} \Gamma_N^c & 0 \\ 0 & \bar{\sigma}_u^2 I_p \end{bmatrix}.$$

Now, we can apply Theorem 2.4 of Simchowitz et al. (2018). With probability at least $1 - 3\delta$ we have $\| \begin{bmatrix} A - \hat{A}_N & B - \hat{B}_N \end{bmatrix} \|_2 \leq \epsilon$ if:

$$N \geq \frac{\text{poly}(n, \log 1/\delta, M)}{\epsilon^2 \sigma_{\min}(\Gamma_\kappa^c)} \log \det(\bar{\Gamma} \Gamma_{\text{sb}}^{-1}),$$

where we have simplified the expression by including terms in the polynomial term. Using Lemma 1 in Tsiamis and Pappas (2021), we obtain

$$\log\det(\bar{\Gamma}\Gamma_{\text{sb}}^{-1}) = \text{poly}(n, M, \log 1/\delta)\log N.$$

Moreover, we use the lower bound $\Gamma_k^c \succeq \bar{\sigma}_u^2\Gamma_k$, which holds for every $k \geq 0$. ■

We note that we can easily obtain sharper bounds by considering the combined controllability Gramian $\Gamma_k(A, \begin{bmatrix} \bar{\sigma}_u B & H \end{bmatrix})$ for the identification stage. For the economy of the presentation, we omit such an analysis here.

D.3. Sensitivity of Stabilization

Here we prove that when (D.2) is feasible, then $\hat{K}_N$ stabilizes all plants (A, B) such that $\|A - \hat{A}_N\|_2 \leq \epsilon_A$, $\|B - \hat{B}_N\|_2 \leq \epsilon_B$. We also show that feasibility is guaranteed as long as we can achieve small enough error bounds ϵ_A, ϵ_B.

Theorem D.2 *Let $\hat{K}_N$ be a feasible solution to problem (D.2) for some $\epsilon_A, \epsilon_B > 0$. Then for any system (A, B) such that $\|A - \hat{A}_N\|_2 \leq \epsilon_A$, $\|B - \hat{B}_N\|_2 \leq \epsilon_B$ we have that*

$$\rho(A + B\hat{K}_N) < 1.$$

Moreover, there exists an $\epsilon_0 > 0$ such that

$$\epsilon_0 = \text{poly}\Big((\frac{M}{\mu})^{\kappa}, M^{\kappa}, \kappa\Big)$$

and Problem (D.2) is feasible if $\epsilon_A, \epsilon_B \leq \epsilon_0$.

Proof Let $\hat{K}_N$ be a feasible solution to problem (D.2). Define $\mathbf{\Phi}_x = (zI - \hat{A}_N - \hat{B}_N\hat{K}_N)^{-1}$, which is well-defined and stable since $\epsilon_A > 0$ and $\|\mathbf{\Phi}_x\|_{\mathcal{H}_\infty} < 1/(\sqrt{2}\epsilon_A)$. Define the system difference

$$\mathbf{\Delta} \triangleq (\hat{A}_N - A)\mathbf{\Phi}_x + (\hat{B}_N - B)\hat{K}_N\mathbf{\Phi}_x$$

It follows from simple algebra that:

$$\begin{aligned} zI - A - B\hat{K}_N &= zI - \hat{A}_N - \hat{B}_N\hat{K}_N + (\hat{A}_N - A) + (\hat{B}_N - B)\hat{K}_N \\ &= (I + \mathbf{\Delta})(zI - \hat{A}_N - \hat{B}_N\hat{K}_N). \end{aligned}$$

If $(I + \mathbf{\Delta})^{-1}$ is stable then the closed loop response is stable and well-defined

$$(zI - A - B\hat{K}_N)^{-1} = (zI - \hat{A}_N - \hat{B}_N\hat{K}_N)^{-1}(I + \mathbf{\Delta})^{-1}.$$

But $(I + \mathbf{\Delta})^{-1}$ being stable is equivalent to

$$\|(I + \mathbf{\Delta})^{-1}\|_{\mathcal{H}_\infty} < \infty.$$

A sufficient condition for this to occur is to require (Dean et al., 2017)

$$\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} < 1.$$

By Proposition 3.5 (select $\alpha = 1/2$) of (Dean et al., 2017)

$$\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} < \left\| \begin{bmatrix} \sqrt{2}\epsilon_A (zI - \hat{A}_N - \hat{B}_N K)^{-1} \\ \sqrt{2}\epsilon_B K (zI - \hat{A}_N - \hat{B}_N K)^{-1} \end{bmatrix} \right\|_{\mathcal{H}_\infty} < 1.$$

This completes the proof of $\rho(A + B\hat{K}_N) < 1$.

To prove feasibility consider the optimal LQR gain $K_\star$, for $Q = I_n$, $R = I_p$. Following Lemma 4.2 in Dean et al. (2017), if the following sufficient condition holds

$$(\epsilon_A + \epsilon_B \|K_\star\|_2) \|(zI - A - BK_\star)^{-1}\|_{\mathcal{H}_\infty} \leq 1/5,$$

then $K_\star$ is a feasible solution

$$\left\| \begin{bmatrix} \sqrt{2}\epsilon_A (zI - \hat{A}_N - \hat{B}_N K_\star)^{-1} \\ \sqrt{2}\epsilon_B K_\star (zI - \hat{A}_N - \hat{B}_N K_\star)^{-1} \end{bmatrix} \right\|_{\mathcal{H}_\infty} < 1.$$

Hence, we can choose

$$\epsilon_0 = \left(5(1 + \|K_\star\|_2) \|(zI - A - BK_\star)^{-1}\|_{\mathcal{H}_\infty}\right)^{-1}. \tag{D.5}$$

The fact that $\epsilon_0 = \text{poly}\left(\left(\frac{M}{\mu}\right)^\kappa, M^\kappa, \kappa\right)$ follows from Lemmas B.2, B.3. ■

D.4. Proof of Theorem 7

Let $u_t \sim \mathcal{N}(0, \bar{\sigma}_u^2 I)$, with $\bar{\sigma}_u^2 = \sigma_u^2/p$. Consider the stabilization algorithm as described in (D.1), (D.2). Consider the ϵ_0 defined in (D.5). By Theorems D.1, D.2, if

$$N\sigma_u^2 \geq^{\triangleq} \underbrace{\frac{\text{poly}(n, \log 1/\delta, M)}{\epsilon_0^2 \sigma_{\min}(\Gamma_\kappa)}}_{\mathcal{N}} \log N$$

we have with probability at least $1 - \delta$ that $\|A - \hat{A}_N\|_2, \|B - \hat{B}_N\|_2 \leq \epsilon_0$ and problem (D.2) is feasible with $\epsilon_B = \epsilon_A = \epsilon_0$. By Theorems B.1 D.2,

$$\mathcal{N} = \text{poly}\left(\left(\frac{M}{\mu}\right)^\kappa, M^\kappa, n, \log 1/\delta\right).$$

To complete the proof we use the fact that

$$N \geq c \log N \text{ if } N \geq 2c \log 2c.$$

Appendix E. Regret Lower Bounds

First let us state an application of the main result of Ziemann and Sandberg (2022). Consider a system $(A, B, H) \in \mathbb{R}^{n \times (n+p+n)}$, where (A, B) is controllable and $H = I_n$. Let P be the respective Riccati matrix for $Q = I_n$, $R = I_p$, with $K_\star$ the respective optimal LQR gain. Fix a matrix $\Delta \in \mathbb{R}^{p \times n}$ and define the family of systems:

$$A(\theta) = A - \theta B\Delta,\ B(\theta) = B + \theta\Delta,\ H(\theta) = I_n, \tag{E.1}$$

where $\theta \in \mathcal{B}(0,\epsilon)$, for some small ϵ. Assume that ϵ is small enough, such that the Riccati equation has a stabilizing solution for every system in the above family. The respective Riccati matrix is denoted by $P(\theta)$ and the LQR gain by $K(\theta)$. The derivative of $K_\star(\theta)$ with respect to θ at point $\theta = 0$ is given by the following formula.

Lemma E.1 (Lemma 2.1 (Simchowitz and Foster, 2020)) *If the system (A, B) is stabilizable, then*

$$\frac{d}{d\theta} K_\star(\theta)|_{\theta=0} = -(B'PB + R)^{-1}\Delta' P(A + BK_*).$$

Finally, let Σ_x be the solution to the Lyapunov equation:

$$\Sigma_x = (A + BK_\star)\Sigma_x(A + BK_\star)' + I_n. \tag{E.2}$$

Theorem E.2 (Application of Theorem 1 in Ziemann and Sandberg (2022)) *Consider a system $S = (A, B, H) \in \mathbb{R}^{n\times(n+p+n)}$, where (A, B) is controllable and $H = I_n$. Let P be the respective solution of the algebraic Riccati equation for $Q = I_n$, $R = I_p$, with $K_\star$ the respective optimal LQR gain. Recall the definition of Σ_x in (E.2). Define the family of systems $\mathscr{C}_S(\epsilon) \triangleq \{(A(\theta), B(\theta), I_n),\ \theta \in \mathcal{B}(0,\epsilon)\}$ as defined in (E.1), for any $\epsilon > 0$ sufficiently small such that $P(\theta)$ and $K_\star(\theta)$ are well-defined. Let $Q_T = P(\theta)$. Then for any $\alpha \in (0, 1/4)$:*

$$\lim\inf_{T\to\infty} \sup_{\hat{S}\in\mathscr{C}_S(T^{-a})} \mathbb{E}_{\hat{S},\pi} \frac{R_T(\hat{S})}{\sqrt{T}} \geq \frac{1}{2\sqrt{2}}\sqrt{\frac{F}{L}}, \tag{E.3}$$

where

$$F = \operatorname{tr}\left((B'PB + R)^{-1}\Delta' P\left[\Sigma_x - I_n\right] P\Delta\right)$$
$$L = n(\|\Delta K_\star\|_2^2 + \|\Delta\|_2^2)\|(B'PB + R)^{-1}\|_2$$

Proof Note that if $\Delta' P(A + BK_\star) = 0$, then since $\Sigma_x \succeq I_n$ is invertible

$$\Delta' P(A + BK_\star) = 0 \Leftrightarrow \Delta' P(A + BK_\star)\Sigma_x(A + BK_\star)' P\Delta = 0$$
$$\Leftrightarrow \Delta' P(\Sigma_x - I_n)P\Delta = 0.$$

This implies that $F = 0$ and the regret lower bound becomes 0, in which case the claim of the theorem is trivially true. Hence, we will assume that $\Delta' P(A + BK_\star) \neq 0$.

All systems in the family have the same closed-loop response under the control policy $u = K_\star x$. In particular, for all $\theta \in \mathcal{B}(0,\epsilon)$:

$$\frac{d}{d\theta}\begin{bmatrix} A(\theta) & B(\theta)\end{bmatrix}\begin{bmatrix} I_n \\ K_\star\end{bmatrix} = \begin{bmatrix} -\Delta K_\star & \Delta\end{bmatrix}\begin{bmatrix} I_n \\ K_\star\end{bmatrix} = 0.$$

Moreover, by Lemma E.1

$$\frac{d}{d\theta} K_\star(\theta)|_{\theta=0} = (B'PB + R)^{-1}\Delta' P(A + BK_\star) \neq 0.$$

By Proposition 3.4 in Ziemann and Sandberg (2022), the above two conditions imply that the family $\mathscr{C}_S(\epsilon)$ is $\epsilon-$uninformative (see Section 3 in Ziemann and Sandberg (2022) for definition).

Next, by Lemma 3.6 in Ziemann and Sandberg (2022), the family is also $L-$information regret bounded (see Section 3 in Ziemann and Sandberg (2022) for the definition), where

$$L = \mathrm{tr}(I_n)\| \begin{bmatrix} -\Delta K_\star & \Delta \end{bmatrix} \|_2^2 \|(B'PB+R)^{-1}\|_2 \overset{i)}{\leq} n(\|\Delta K_\star\|_2^2 + \|\Delta\|_2^2)\|(B'PB+R)^{-1}\|_2.$$

Inequality $i)$ follows from $\mathrm{tr}(I_n) = n$ and the norm property

$$\| \begin{bmatrix} M_1 & M_2 \end{bmatrix} \|_2^2 = \| \begin{bmatrix} M_1 & M_2 \end{bmatrix} \begin{bmatrix} M_1 & M_2 \end{bmatrix}' \|_2 = \|M_1 M_1' + M_2 M_2'\|_2 \leq \|M_1\|_2^2 + \|M_2\|_2^2.$$

Applying Theorem 1 in Ziemann and Sandberg (2022), we get (E.3), for L defined as above and

$$F = \mathrm{tr}\left(\left[\Sigma_x \otimes (B'P(\theta)B+R)\right] (\frac{d}{d\theta}\mathrm{vec}K_\star(\theta)|_{\theta=0})(\frac{d}{d\theta}\mathrm{vec}K_\star(\theta)|_{\theta=0})' \right),$$

where $\otimes$ is the Kronecker product and vec is the vectorization operator (mapping a matrix into a column vector by stacking its columns). Using the identities:

$$\mathrm{vec}(XYZ) = (Z' \otimes X)\mathrm{vec}(Y), \qquad \mathrm{tr}(\mathrm{vec}(X)\mathrm{vec}(Y)') = \mathrm{tr}(XY'),$$

we can rewrite F as

$$F = \mathrm{tr}\left((B'P(\theta)B+R)\frac{d}{d\theta}K(\theta)|_{\theta=0}\Sigma_x \frac{d}{d\theta}K'(\theta)|_{\theta=0} \right).$$

By Lemma E.1 and the property $\mathrm{tr}(XY) = \mathrm{tr}(YX)$, we finally get

$$F = \mathrm{tr}\left((B'PB+R)^{-1}\Delta' P(A+BK_*)\Sigma_x(A+BK_*)P\Delta \right).$$

The result follows from $(A+BK_*)\Sigma_x(A+BK_*)' = \Sigma_x - I_n$. ■

E.1. Proof of Lemma 9

The result follows by Theorem E.2. We only need to compute and simplify F, L. Due to the structure of system (18), we have

$$P = \begin{bmatrix} 1 & 0 \\ 0 & P_0 \end{bmatrix}, \; K_\star = \begin{bmatrix} 0 & 0 \\ 0 & K_{0,\star} \end{bmatrix}.$$

Moreover, due to the structure of the perturbation Δ in (16)

$$B'PB + R = \begin{bmatrix} 2 & 0 \\ 0 & B_0'P_0B_0 + R_0 \end{bmatrix}, \; P\Delta(B'PB+R)^{-1}\Delta' P = \frac{1}{2}\begin{bmatrix} 0 & 0 \\ 0 & P_0\Delta_1\Delta_1'P_0 \end{bmatrix}.$$

Hence

$$F = \frac{1}{2}\mathrm{tr}\left(\begin{bmatrix} 0 & 0 \\ 0 & P_0\Delta_1\Delta_1'P_0 \end{bmatrix} (\Sigma_x - I_n) \right) = \frac{1}{2}\Delta_1' P_0(\Sigma_{0,x} - I_{n-1})P_0\Delta_1$$

Finally we have $L \leq n$, since $\Delta K_\star = 0$, Δ_1 has unit norm, and $R = I_p$. ■

E.2. Proof of Lemma 10

First note that $P_0 \succeq Q_0 = I_{n-1}$. As a result, we have

$$\|P_0(\Sigma_{0,x} - I_{n-1})P_0\|_2 \geq \|\Sigma_{0,x} - I_{n-1}\|_2.$$

It is sufficient to lower bound $\|\Sigma_{0,x} - I_{n-1}\|_2$. Consider the recursion:

$$\Sigma_k = (A_0 + B_0K_{0,\star})\Sigma_{k-1}(A_0 + B_0K_{0,\star})' + I_{n-1},\ \Sigma_0 = 0.$$

Then $\Sigma_{0,x} = \lim_{k\to\infty} \Sigma_k \succeq \Sigma_{n-1} \succeq I_{n-1}$. The second inequality follows from monotonicity of the Lyapunov operator:

$$g(X) = (A_0 + B_0K_{0,\star})X(A_0 + B_0K_{0,\star})' + I_{n-1},$$

i.e. $g(X) \succeq g(Y)$ if $X \succeq Y$. What remains is to lower bound $\|\Sigma_{n-1} - I_{n-1}\|_2$. Let $e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}'$ be the first canonical vector. Due to the structure of A_0, B_0

$$e_1'(A_0 + B_0K_{0,\star})^i = e_1'(A_0)^i,\ \text{for } i \leq n-1.$$

Hence

$$\begin{aligned}\|\Sigma_{n-1} - I_{n-1}\|_2 &\geq e_1'(\Sigma_{n-1} - I_{n-1})e_1 \\ &= \sum_{k=1}^{n-1} e_1'A_0^k(A_0')^k e_1.\end{aligned}$$

After some algebra we can compute analytically

$$\|\Sigma_{n-1} - I_{n-1}\|_2 \geq \sum_{k=1}^{n-1}\sum_{t=0}^{k}\binom{k}{t}^2 = \sum_{k=1}^{n-1}\binom{2k}{k} \geq \binom{2(n-1)}{n-1} \geq \left(2\frac{n-1}{n-1}\right)^{n-1} = 2^{n-1},$$

which completes the proof. ∎

E.3. Proof of Theorem 8

It is sufficient to prove the result for the class $\mathcal{C}^{\mu}_{n,n-1}$. If $n > \kappa + 1$, then we can consider the system:

$$\tilde{A} = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & A \end{array}\right],\ \tilde{B} = \left[\begin{array}{c|c} I_{n-\kappa-1} & 0 \\ \hline 0 & B \end{array}\right],\ \tilde{H} = \left[\begin{array}{c|c} I_{n-\kappa-1} & 0 \\ \hline 0 & H \end{array}\right]$$

where $(A, B, H) \in \mathscr{C}^{\mu}_{\kappa,\kappa-1}$ and repeat the same arguments.

The proof follows from Lemma 9 and Lemma 10. What remains to show that for every ϵ

$$\mathscr{C}_S(\epsilon) \subseteq \mathscr{C}^{\mu}_{n,n-1}(\epsilon).$$

This follows from the fact that $\Delta K_\star = 0$, hence $A = A(\theta)$ and $\|B - B(\theta)\| = \theta\|\Delta\|_2 = \theta \leq \epsilon$. Thus,

$$\|\begin{bmatrix} A - A(\theta) & B - B(\theta) \end{bmatrix}\|_2 \leq \epsilon.$$

Since $\mathscr{C}_S(\epsilon) \subseteq \mathscr{C}^{\mu}_{n,n-1}(\epsilon)$, we get

$$\liminf_{T\to\infty} \sup_{S\in\mathscr{C}^{\mu}_{n,n-1}(T^{-a})} \mathbb{E}_{\hat{S},\pi}\frac{R_T(\hat{S})}{\sqrt{T}} \geq \liminf_{T\to\infty} \sup_{\hat{S}\in\mathscr{C}_S(T^{-a})} \mathbb{E}_{\hat{S},\pi}\frac{R_T(\hat{S})}{\sqrt{T}}$$ ∎

E.4. Stable System Example

Here we show that the local minimax expected regret can be exponential in the dimension even for stable systems. Using again the two subsystems trick, consider the following stable system

$$S: \qquad x_{k+1} = \left[\begin{array}{c|cccc} 0 & 0 & 0 & & 0 & 0 \\ \hline 0 & \rho & 2 & & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & & \rho & 2 \\ 0 & 0 & 0 & & 0 & \rho \end{array}\right] x_k + \left[\begin{array}{c|c} 1 & 0 \\ 0 & 0 \\ \vdots & \\ 0 & 1 \end{array}\right] u_k + w_k,\ 0 < \rho < 1, \tag{E.4}$$

with $Q = I_n$, $R = I_2$. Following the notation of (18) let:

$$A_0 = \begin{bmatrix} \rho & 2 & 0 & & 0 & 0 \\ 0 & \rho & 2 & & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & & \rho & 2 \\ 0 & 0 & 0 & & 0 & \rho \end{bmatrix},\ B_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},\ Q_0 = I_{n-1},\ R_0 = 1, \tag{E.5}$$

where $A_0 \in \mathbb{R}^{(n-1)\times(n-1)}$ and $B_0 \in \mathbb{R}^{n-1}$. Note that A_0 has spectral radius $\rho < 1$. Let $\Delta = \begin{bmatrix} 0 & 0 \\ \Delta_1 & 0 \end{bmatrix}$. Then, by Lemma 9, the local minimax expected regret for system S, given the perturbation Δ_1 is lower bounded by

$$\liminf_{T\to\infty} \sup_{\hat{S}\in\mathscr{C}_S(T^{-a})} \mathbb{E}_{\hat{S},\pi} \frac{R_T(\hat{S})}{\sqrt{T}} \geq \frac{1}{4\sqrt{n}} \sqrt{\Delta_1' P_0 \left[\Sigma_{0,x} - I_{n-1}\right] P_0 \Delta_1}.$$

As we show in the following lemma, the quantity $\sqrt{\Delta_1' P_0 \left[\Sigma_{0,x} - I_{n-1}\right] P_0 \Delta_1}$ is exponential with n if we choose Δ_1 appropriately. Although the system is stable, it is very sensitive to inputs and noises. Any signal $u_{k,2}$ that we apply gets amplified by 2 as we move up the chain from state $x_{k,n}$ to state $x_{k,2}$. As a result, any suboptimal policy will result in excessive excitation of the state.

Lemma E.3 (Stable systems can be hard to learn) *Consider system (E.5) Let P_0 be the Riccati matrix for $Q_0 = I_{n-1}, R_0 = 1$, with $K_{\star,0}$, $\Sigma_{0,x}$ the corresponding LQR control gain and steady-state covariance, respectively. Then*

$$\left\|P_0 \left[\Sigma_{0,x} - I_{n-1}\right] P_0\right\|_2 \geq 2^{4n-8} + o(1),$$

where $o(1)$ goes to zero as $n \to \infty$.

Proof Let $\Delta_1 = \begin{bmatrix} 0 & 0 & \cdots & 1 & 0 \end{bmatrix}'$. It is sufficient to prove that

$$\Delta_1' P_0 (\Sigma_{0,x} - I_{n-1}) P_0 \Delta_1$$

is exponential. Using the identity $\Sigma_{0,x} - I_{n-1} = (A_0 + B_0 K_{\star,0}) \Sigma_{0,x} (A_0 + B_0 K_{\star,0})'$, $\Sigma_{0,x} \succeq I$, we have:

$$\Delta_1' P_0 (\Sigma_{0,x} - I_{n-1}) P_0 \Delta_1 \geq \|\Delta_1' P_0 (A_0 + B_0 K_{\star,0})\|_2^2.$$

By Lemma E.5 and Lemma E.4 it follows that

$$\|\Delta_1' P_0(A_0 + B_0 K_{\star,0})\|_2^2 \geq 2^{4n-8} + o(1).$$

■

Lemma E.4 (Riccati matrix can grow exponentially) *For system (E.5) we have:*

$$B_0' P_0 B_0 + R_0 \geq 2^{2n-4} + 1.$$

Proof Consider the Riccati operator:

$$g(X, Y) = A_0' X A_0 + Y - A_0' X B_0 (B_0' X B_0 + R_0)^{-1} B_0' X A_0.$$

Based on the above notation, we have $P_0 = g(P_0, Q_0)$. The Riccati operator is monotone (Anderson and Moore, 2005), i.e

$$X_1 \succeq X_2 \Rightarrow g(X_1, Y) \succeq g(X_1, Y).$$

It is also trivially monotone with respect to Y. Let $X_0 = 0$, then the recursion $X_{t+1} = g(X_t, Q_0)$ converges to P_0. By monotonicity

$$P_0 \succeq X_t, \text{ for all } t \geq 0$$

Let e_i denote the i-th canonical vector in $\mathbb{R}^{n-1}$. By monotonicity, we also have:

$$X_1 = g(X_0, Q_0) \succeq g(X_0, e_1 e_1') = \underbrace{e_1 e_1'}_{\tilde{X}_1}$$

Repeating the argument:

$$X_2 = g(X_1, Q_0) \succeq g(\tilde{X}_1, Q_0) \succeq g(\tilde{X}_1, e_1 e_1') = \underbrace{A_0' \tilde{X}_1 A_0 + e_1 e_1'}_{\tilde{X}_2} = A_0' e_1 e_1' A_0 + e_1 e_1'$$

$$= 2^2 e_2 e_2' + \rho^2 e_1 e_1' + 2\rho e_1 e_2' + 2\rho e_2 e_1'$$

Similarly,

$$X_{n-1} = g(X_{n-2}, Q_0) \succeq g(\tilde{X}_{n-2}, e_1 e_1') = (A_0')^{n-2} e_1 e_1' A_0^{n-2} + (A_0')^{n-1} e_1 e_1' A_0^{n-1} + \cdots + e_1 e_1',$$

where we use the fact that every $\tilde{X}_k$ is orthogonal to B_0 for $k \leq n-2$. As a result:

$$\begin{aligned} [P_0]_{n-1,n-1} \geq [X_n]_{n-1,n-1} &\geq e_{n-1}' (A_0')^{n-2} e_1 e_1' A_0^{n-2} e_{n-1} \\ &= (e_1' A_0^{n-2} e_{n-1})^2 = ([A_0^{n-2}]_{1,n-1})^2 \end{aligned} \tag{E.6}$$

What remains is to compute $[A_0^{n-2}]_{1,n-1}$. Define by $J \in \mathbb{R}^{(n-1)\times(n-1)}$ the companion matrix:

$$J = \begin{bmatrix} 0 & 1 & 0 & & 0 & 0 \\ 0 & 0 & 1 & & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & & 0 & 1 \\ 0 & 0 & 0 & & 0 & 0 \end{bmatrix}.$$

Since $A_0 = \rho I + 2J$ and I commutes with J by the binomial expansion formula:

$$A_0^{n-2} = 2^{n-2}J^{n-2} + \sum_{t=0}^{n-3} 2^t \binom{n-2}{t} J^t.$$

Since $e_1' J^{n-1} e_{n-1} = 1$, $e_1' J^t e_{n-1} = 0$, for $t \leq n-2$, we obtain:

$$([A_0^{n-2}]_{1,n-1})^2 = 2^{2n-4}. \tag{E.7}$$

By (E.6) and (E.7) we finally get

$$B_0' P_0 B_0 + R_0 = [P_0]_{n-1,n-1} + 1 \geq 2^{2n-4} + 1$$

■

Lemma E.5 *We have:*

$$\|\Delta_1' P_0 (A_0 + B_0 K_{\star,0})\|_2 \geq (0.5 + o(1))(B_0' P_0 B_0 + R_0),$$

where the $o(1)$ is in the large n regime.

Proof Let e_i denote the i-th canonical vector in $\mathbb{R}^{n-1}$. It is sufficient to show that

$$\left|(B_0' P_0 B_0 + R_0)^{-1} \Delta_1' P_0 (A_0 + B_0 K_{\star,0}) e_{n-1}\right| \geq 0.5 + o(1).$$

For simplicity we will denote:

$$\alpha \triangleq [P_0]_{n-1,n-1}, \quad \beta \triangleq [P_0]_{n-2,n-2}, \quad \gamma \triangleq [P_0]_{n-1,n-2}.$$

Due to the structure of A_0, we have

$$A_0 e_{n-1} = \rho e_{n-1} + 2e_{n-2}.$$

Using this, we obtain

$$\begin{aligned} K_{\star,0} e_{n-1} &= -(B_0' P_0 B_0 + 1)^{-1} B_0' P_0 A_0 e_{n-1} = -(\alpha+1)^{-1} e_{n-1}' P_0 (\rho e_{n-1} + 2e_{n-2}) \\ &= -(\alpha+1)^{-1}(\rho\alpha + 2\gamma). \end{aligned} \tag{E.8}$$

Combining the above results

$$\begin{aligned} &(B_0' P_0 B_0 + R)^{-1} \Delta_1' P_0 (A_0 + B_0 K_{\star,0}) e_{n-1} = (B_0' P_0 B_0 + 1)^{-1} e_{n-2}' P_0 (A_0 + B_0 K_{\star,0}) e_{n-1} \\ &= (\alpha+1)^{-1} \left\{ e_{n-2}' P_0 (\rho e_{n-1} + 2e_{n-2}) - e_{n-2}' P_0 e_{n-1} (\alpha+1)^{-1} (\rho\alpha + 2\gamma) \right\} \\ &= (\alpha+1)^{-1} \left\{ \rho\gamma + 2\beta - \gamma(\alpha+1)^{-1}(\rho\alpha + 2\gamma) \right\} \\ &= 2(\alpha+1)^{-1} \left\{ \beta - (\alpha+1)^{-1}\gamma^2 \right\} - (\alpha+1)^{-2}\rho\gamma \\ &\overset{i)}{=} \frac{2}{\alpha+1} \left\{ \beta - \frac{\gamma^2}{\alpha+1} \right\} + o(1), \end{aligned}$$

where i) follows from Lemma E.6. What remains to show is that

$$\frac{2}{\alpha+1}\left\{\beta-\frac{\gamma^2}{\alpha+1}\right\}=0.5+o(1). \tag{E.9}$$

Using the algebraic Riccati equation:

$$\begin{aligned}\alpha &= e_{n-1}'A_0'P_0A_0e_{n-1}+1-e_{n-1}'A_0'P_0B_0(\alpha+1)^{-1}B_0'P_0A_0e_{n-1}\\ &= (\rho e_{n-1}+2e_{n-2})'P_0(\rho e_{n-1}+2e_{n-2})+1\\ &- (\rho e_{n-1}+2e_{n-2})'P_0e_{n-1}(\alpha+1)^{-1}e_{n-1}'P_0(\rho e_{n-1}+2e_{n-2})\\ &= \rho^2\alpha+4\beta+4\rho\gamma+1-\frac{(\rho\alpha+2\gamma)^2}{\alpha+1}\\ &= 4\beta+\frac{\rho^2\alpha+4\rho\gamma+\alpha+1-4\gamma^2}{\alpha+1}.\end{aligned}$$

Dividing both sides with $\alpha+1$:

$$\frac{\alpha}{1+\alpha}=\frac{4}{\alpha+1}\left\{\beta-\frac{\gamma^2}{\alpha+1}\right\}+\frac{4\rho\gamma}{(\alpha+1)^2}+\frac{1+\rho^2\alpha}{(1+\alpha)^2}$$

Rearranging the terms gives:

$$\frac{2}{\alpha+1}\left\{\beta-\frac{\gamma^2}{\alpha+1}\right\}-0.5=-\frac{0.5}{1+\alpha}-\frac{2\rho\gamma}{(\alpha+1)^2}-\frac{1+\rho^2\alpha}{2(1+\alpha)^2}$$

By Lemma E.6 the second term in the right-hand side is $o(1)$. By Lemma E.4, $\alpha=\Omega(2^{2n})$, hence all remaining terms also go to zero, which completes the proof of (E.9).

■

Lemma E.6 *Recall the notation in the proof of Lemma E.5*

$$\alpha\triangleq[P_0]_{n-1,n-1},\quad \gamma\triangleq[P_0]_{n-1,n-2}.$$

Then, we have:

$$\left|\frac{\gamma}{(\alpha+1)^2}\right|=o(1)$$

Proof We use the relation:

$$P_0=(A_0+B_0K_{\star,0})'P_0(A_0+B_0K_{\star,0})+Q_0+K_{\star,0}'R_0K_{\star,0}\succeq K_{\star,0}'R_0K_{\star,0}.$$

Multiplying from the left and right by e_{n-1} and by invoking (E.8) we obtain:

$$\alpha\geq\left(\frac{\rho\alpha+2\gamma}{\alpha+1}\right)^2=(\xi+\lambda)^2,$$

where for simplicity we define $\xi=\frac{\rho\alpha}{\alpha+1}$, $\lambda=\frac{2\gamma}{\alpha+1}$. We can further lower bound the above expression by:

$$\alpha\geq(\xi+\lambda)^2\geq\xi^2+\lambda^2-2\xi\left|\lambda\right|.$$

This is a quadratic inequality and holds if and only if:

$$\xi - \sqrt{\alpha} \leq |\lambda| \leq \xi + \sqrt{\alpha}.$$

As a result:

$$2\frac{|\gamma|}{\alpha+1} \leq \rho + \sqrt{\alpha+1}$$

which leads to

$$\frac{|\gamma|}{\alpha+1} \leq 0.5\frac{\rho+\sqrt{\alpha+1}}{\alpha+1} = O(1/\sqrt{\alpha}) = o(1)$$

since $\alpha = \Omega(2^{2n})$. ■