Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator

Sarah Dean **Horia Mania** **Nikolai Matni** **Benjamin Recht** **Stephen Tu**

University of California, Berkeley

Abstract

We consider adaptive control of the Linear Quadratic Regulator (LQR), where an unknown linear system is controlled subject to quadratic costs. Leveraging recent developments in the estimation of linear systems and in robust controller synthesis, we present the first provably polynomial time algorithm that provides high probability guarantees of sub-linear regret on this problem. We further study the interplay between regret minimization and parameter estimation by proving a lower bound on the expected regret in terms of the exploration schedule used by any algorithm. Finally, we conduct a numerical study comparing our robust adaptive algorithm to other methods from the adaptive LQR literature, and demonstrate the flexibility of our proposed method by extending it to a demand forecasting problem subject to state constraints.

1 Introduction

The problem of adaptively controlling an unknown dynamical system has a rich history, with classical asymptotic results of convergence and stability dating back decades [12, 13]. Of late, there has been a renewed interest in the study of a particular instance of such problems, namely the adaptive Linear Quadratic Regulator (LQR), with an emphasis on *non-asymptotic* guarantees of stability and performance. Initiated by Abbasi-Yadkori and Szepesvári [1], there have since been several works analyzing the regret suffered by various adaptive algorithms on LQR– here the regret incurred by an algorithm is thought of as a measure of deviations in performance from optimality over time. These results can be broadly divided into two categories: those providing high-probability guarantees for a single execution of the algorithm [1, 4, 8, 11], and those providing bounds on the expected *Bayesian* regret incurred over a family of possible systems [2, 16]. As we discuss in more detail, these methods all suffer from one or several of the following limitations: restrictive and unverifiable assumptions, limited applicability, and computationally intractable subroutines. In this paper, we provide, to the best of our knowledge, the first polynomial-time algorithm for the adaptive LQR problem that provides high probability guarantees of sub-linear regret, and that does not require unverifiable or unrealistic assumptions.

Related Work. There is a rich body of work on the estimation of linear systems as well as on the robust and adaptive control of unknown systems. We target our discussion to works on non-asymptotic guarantees for the LQR control of an unknown system, broadly divided into three categories.

Offline estimation and control synthesis: In a non-adaptive setting, i.e., when system identification can be done offline prior to controller synthesis and implementation, the first work to provide end-to-end guarantees for the LQR optimal control problem is that of Fiechter [10], who shows that the *discounted* LQR problem is PAC-learnable. Dean et al. [6] improve on this result, and provide the first end-to-end sample complexity guarantees for the infinite horizon average cost LQR problem.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Optimism in the Face of Uncertainty (OFU): Abbasi-Yadkori and Szepesvári [1], Faradonbeh et al. [8], and Ibrahimi et al. [11] employ the *Optimism in the Face of Uncertainty* (OFU) principle [5], which optimistically selects model parameters from a confidence set by choosing those that lead to the *best* closed-loop (infinite horizon) control performance, and then plays the corresponding optimal controller, repeating this process online as the confidence set shrinks. While OFU in the LQR setting has been shown to achieve optimal regret $\widetilde{\mathcal{O}}(\sqrt{T})$, its implementation requires solving a non-convex optimization problem to precision $\widetilde{\mathcal{O}}(T^{-1/2})$, for which no provably efficient implementation exists.

Thompson Sampling (TS): To circumvent the computational roadblock of OFU, recent works replace the intractable OFU subroutine with a random draw from the model uncertainty set, resulting in *Thompson Sampling* (TS) based policies [2, 4, 16]. Abeille and Lazaric [4] show that such a method achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret with high-probability for scalar systems. However, their proof does not extend to the non-scalar setting. Abbasi-Yadkori and Szepesvári [2] and Ouyang et al. [16] consider expected regret in a Bayesian setting, and provide TS methods which achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. Although not directly comparable to our result, we remark on the computational challenges of these algorithms. Whereas the proof of Abbasi-Yadkori and Szepesvári [2] was shown to be incorrect [15], Ouyang et al. [16] make the restrictive assumption that there exists a (known) initial compact set Θ describing the uncertainty in the system parameters, such that for any system $\theta_1 \in \Theta$, the optimal controller $K(\theta_1)$ is stabilizing when applied to any other system $\theta_2 \in \Theta$. No means of constructing such a set are provided, and there is no known tractable algorithm to verify if a given set satisfies this property. Also, it is implicitly assumed that projecting onto this set can be done efficiently.

Contributions. To develop the first polynomial-time algorithm that provides high probability guarantees of sub-linear regret, we leverage recent results from the estimation of linear systems [17], robust controller synthesis [14, 19], and coarse-ID control [6]. We show that our robust adaptive control algorithm: (i) guarantees stability and near-optimal performance at all times; (ii) achieves a regret up to time T bounded by $\widetilde{\mathcal{O}}(T^{2/3})$; and (iii) is based on finite-dimensional semidefinite programs of size logarithmic in T.

Furthermore, our method estimates the system parameters at $\widetilde{\mathcal{O}}(T^{-1/3})$ rate in operator norm. Although system parameter identification is not necessary for optimal control performance, an accurate system model is often desirable in practice. Motivated by this, we study the interplay between regret minimization and parameter estimation, and identify fundamental limits connecting the two. We show that the expected regret of our algorithm is lower bounded by $\Omega(T^{2/3})$, proving that our analysis is sharp up to logarithmic factors. Moreover, our lower bound suggests that the estimation rate achievable by any algorithm with $\mathcal{O}(T^{\alpha})$ regret is $\Omega(T^{-\alpha/2})$.

Finally, we conduct a numerical study of the adaptive LQR problem, in which we implement our algorithm, and compare its performance to heuristic implementations of OFU and TS based methods. We show on several examples that the regret incurred by our algorithm is comparable to that of the OFU and TS based methods. Furthermore, the infinite horizon cost achieved by our algorithm at any given time on the true system is consistently lower than that attained by OFU and TS based algorithms. Finally, we use a demand forecasting example to show how our algorithm naturally generalizes to incorporate environmental uncertainty and safety constraints. The full version of this paper is [7].

2 Problem Statement and Preliminaries

In this work we consider adaptive control of the following discrete-time linear system

$$x_{k+1} = A_\star x_k + B_\star u_k + w_k \ , \ \ w_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2 I) \ , \tag{2.1}$$

where $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^p$ is the control input, and $w_k \in \mathbb{R}^n$ is the process noise. We assume that the state variables are observed exactly and, for simplicity, that $x_0 = 0$. We consider the *Linear Quadratic Regulator* optimal control problem, given by cost matrices $Q \succeq 0$ and $R \succ 0$,

$$J_\star = \min_u \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{k=1}^{T} x_k^\top Q x_k + u_k^\top R u_k \right] \text{ s.t. dynamics (2.1)} \ , \tag{2.2}$$

where the minimum is taken over measurable functions $u = \{u_k(\cdot)\}_{k\geq 1}$, with each u_k adapted to the history $x_k, x_{k-1}, \ldots, x_1$, and possibe additional randomness independent of future states. Given knowledge of $(A_\star, B_\star)$, the optimal policy is a static state-feedback law $u_k = K_\star x_k$, where $K_\star$ is derived from the solution to a discrete algebraic Riccati equation.

We are interested in algorithms which operate without knowledge of the true system transition matrices $(A_\star, B_\star)$. We measure the performance of such algorithms via their regret, defined as

$$\mathsf{Regret}(T) := \sum_{k=1}^{T} (x_k^\top Q x_k + u_k^\top R u_k - J_\star) \; . \tag{2.3}$$

The regret of any algorithm is lower-bounded by $\Omega(\sqrt{T})$, a bound matched by OFU up to logarithmic factors [8]. However, after each epoch, OFU requires optimizing a non-convex objective to $\mathcal{O}(T^{-1/2})$ precision. Instead, our method uses a subroutine based on quasi-convex optimization and robust control.

2.1 Preliminaries: System Level Synthesis

We briefly describe the necessary background on robust control and System Level Synthesis [19] (SLS). These tools were recently used by Dean et al. [6] to provide non-asymptotic bounds for LQR in the offline "estimate-and-then-control" setting. In the appendix of the full version [7] we expand on these preliminaries.

Consider the dynamics (2.1), and fix a static state-feedback control policy K, i.e., let $u_k = K x_k$. Then, the closed loop map from the disturbance process $\{w_0, w_1, \ldots\}$ to the state x_k and control input u_k at time k is given by

$$\begin{aligned} x_k &= \textstyle\sum_{t=1}^{k} (A_\star + B_\star K)^{k-t} w_{t-1} \; , \\ u_k &= \textstyle\sum_{t=1}^{k} K(A_\star + B_\star K)^{k-t} w_{t-1} \; . \end{aligned} \tag{2.4}$$

Letting $\Phi_x(k) := (A_\star + B_\star K)^{k-1}$ and $\Phi_u(k) := K(A_\star + B_\star K)^{k-1}$, we can rewrite Eq. (2.4) as

$$\begin{bmatrix} x_k \\ u_k \end{bmatrix} = \sum_{t=1}^{k} \begin{bmatrix} \Phi_x(k-t+1) \\ \Phi_u(k-t+1) \end{bmatrix} w_{t-1} \; , \tag{2.5}$$

where $\{\Phi_x(k), \Phi_u(k)\}$ are called the *closed loop system response elements* induced by the controller K. The SLS framework shows that for any elements $\{\Phi_x(k), \Phi_u(k)\}$ constrained to obey

$$\Phi_x(k+1) = A_\star \Phi_x(k) + B_\star \Phi_u(k) \; , \;\; \Phi_x(1) = I \; , \;\; \forall k \geq 1 \; , \tag{2.6}$$

there exists some controller that achieves the desired system responses (2.5). The state-feedback parameterization result in Theorem 1 of Wang et al. [19] formalizes this observation: the SLS framework therefore allows for any optimal control problem over linear systems to be cast as an optimization problem over elements $\{\Phi_x(k), \Phi_u(k)\}$, constrained to satisfy the affine equations (2.6). Comparing equations (2.4) and (2.5), we see that the former is non-convex in the controller K, whereas the latter is affine in the elements $\{\Phi_x(k), \Phi_u(k)\}$, enabling solutions to previously difficult optimal control problems.

As we work with infinite horizon problems, it is notationally more convenient to work with *transfer function* representations of the above objects, which can be obtained by taking a z-transform of their time-domain representations. The frequency domain variable z can be informally thought of as the time-shift operator, i.e., $z\{x_k, x_{k+1}, \ldots\} = \{x_{k+1}, x_{k+2}, \ldots\}$, allowing for a compact representation of LTI dynamics. We use boldface letters to denote such transfer functions, e.g., $\mathbf{\Phi}_x(z) = \sum_{k=1}^{\infty} \Phi_x(k) z^{-k}$. Then, the constraints (2.6) can be rewritten as

$$\begin{bmatrix} zI - A_\star & -B_\star \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I \; , \tag{2.7}$$

and the corresponding (not necessarily static) control law $\mathbf{u} = \mathbf{K}\mathbf{x}$ is given by $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$.

Although other approaches to optimal controller design exists, we argue now that the SLS parameterization has some appealing properties when applied to the control of uncertain systems. In particular, suppose that rather than having access to the true system transition matrices $(A_\star, B_\star)$, we instead only have access to estimates $(\widehat{A}, \widehat{B})$. The SLS framework allows us to characterize the system responses achieved by a controller, computed using only the estimates $(\widehat{A}, \widehat{B})$, on the true system $(A_\star, B_\star)$. Specifically, if we denote $\widehat{\mathbf{\Delta}} := (\widehat{A} - A_\star)\mathbf{\Phi}_x + (\widehat{B} - B_\star)\mathbf{\Phi}_u$, simple algebra shows that

$$\begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I \quad \text{if and only if} \quad \begin{bmatrix} zI - A_\star & -B_\star \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I + \widehat{\mathbf{\Delta}} \, .$$

The robust stability result in Theorem 2 of Matni et al. [14] shows that if $(I + \widehat{\mathbf{\Delta}})^{-1}$ exists, then the controller $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$, computed using only the estimates $(\widehat{A}, \widehat{B})$, achieves the following response on the true system $(A_\star, B_\star)$:

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} (I + \widehat{\mathbf{\Delta}})^{-1} \mathbf{w} \, .$$

Further, if $\mathbf{K}$ stabilizes the system $(\widehat{A}, \widehat{B})$, and $(I + \widehat{\mathbf{\Delta}})^{-1}$ is stable (simple sufficient conditions can be derived to ensure this, see [6]), then $\mathbf{K}$ is also stabilizing for the true system. It is this transparency between system uncertainty and controller performance that we exploit in our algorithm.

We end this discussion with the definition of a function space that we use extensively throughout:

$$\mathcal{S}(C, \rho) := \left\{ \mathbf{M} = \sum_{k=1}^{\infty} M(k) z^{-k} \mid \|M(k)\| \le C\rho^k \ , \ \ k = 1, 2, ... \right\} . \tag{2.8}$$

The space $\mathcal{S}(C, \rho)$ consists of (strictly proper) stable transfer functions that satisfy a certain decay rate in the spectral norm of their impulse response elements. We denote the restriction of $\mathcal{S}(C, \rho)$ to the space of F-length finite impulse response (FIR) filters by $\mathcal{S}_F(C, \rho)$, i.e., $\mathbf{M} \in \mathcal{S}_F(C, \rho)$ if $\mathbf{M} \in \mathcal{S}(C, \rho)$, and $M(k) = 0$ for all $k > F$.

We equip $\mathcal{S}(C, \rho)$ with the $\mathcal{H}_\infty$ and $\mathcal{H}_2$ norms, which are infinite horizon analogs of the spectral and Frobenius norms of a matrix, respectively: $\|\mathbf{M}\|_{\mathcal{H}_\infty} = \sup_{\|\mathbf{w}\|_2 = 1} \|\mathbf{M}\mathbf{w}\|_2$ and $\|\mathbf{M}\|_{\mathcal{H}_2} = \sqrt{\sum_{k=1}^{\infty} \|M(k)\|_F^2}$. The $\mathcal{H}_\infty$ and $\mathcal{H}_2$ norm have distinct interpretations. The $\mathcal{H}_\infty$ norm of a system $\mathbf{M}$ is equal to its $\ell_2 \mapsto \ell_2$ operator norm, and can be used to measure the robustness of a system to unmodelled dynamics [20]. The $\mathcal{H}_2$ norm has a direct interpretation as the energy transferred to the system by a white noise process, and is hence closely related to the LQR optimal control problem. Unsurprisingly, the $\mathcal{H}_2$ norm appears in the objective function of our optimization problem, whereas the $\mathcal{H}_\infty$ norm appears in the constraints to ensure robust stability and performance.

3 Algorithm and Guarantees

Our proposed robust adaptive control algorithm for LQR is shown in Algorithm 1. We note that while Line 8 of Algorithm 1 is written as an infinite-dimensional optimization problem, it can be formulated in terms of finite-dimensional decision variables $\{\Phi_x(k), \Phi_u(k)\}_{k=1}^F$ due to the restriction to FIR filters. In this formulation, the $\mathcal{H}_2$ cost can be written as a Frobenius norm and the $\mathcal{H}_\infty$ constraint reduces to a linear matrix inequality. Therefore, the inner optimization can be equivalently written as a semidefinite program over $\mathcal{O}(F_i(n^2 + np))$ decision variables. We describe this transformation in detail in appendix Section G of the full version [7]. We also note that the outer optimization over γ can be performed efficiently by bisection search because the objective is jointly quasi-convex in the decision variables and is smooth with respect to γ in the feasible domain.

Some remarks on practice are in order. First, in Line 6, only the trajectory data collected during the i-th epoch is used for the least squares estimate. Second, the epoch lengths we use grow exponentially in the epoch index. These settings are chosen primarily to simplify the analysis; in practice all the data collected should be used, and it may be preferable to use a slower growing epoch schedule (such as $T_i = C_T(i + 1)$). Additionally, for storage considerations, instead of performing a batch least squares update of the model, a recursive least squares (RLS) estimator rule can be used to update the parameters in an online manner. Furthermore, many constants in Algorithm 1 depend on the unknown system to be consistent with our data-independent analysis. In practice, these parameters can be estimated from collected data.

Finally, we note that the proofs for all results in this section can be found in the full version [7].

Algorithm 1 Robust Adaptive Control Algorithm

Require: Stabilizing controller $\mathbf{K}^{(0)}$, failure probability $\delta \in (0,1)$, and constants $(C_\star, \rho_\star, \|K_\star\|)$.

Set $C_x \leftarrow \frac{\mathcal{O}(1)C_\star}{(1-\rho_\star)^3}$, $C_u \leftarrow \|K_\star\|C_x$, and $\rho \leftarrow .999 + .001\rho_\star$.

Set $C_T \leftarrow \widetilde{\mathcal{O}}\left((n+p)\frac{C_\star^4(1+\|K_\star\|)^4}{(1-\rho_\star)^8}\right)$.

for $i = 0, 1, 2, \ldots$ **do**

Set $T_i \leftarrow C_T 2^i$ and $\sigma_{\eta,i}^2 \leftarrow \sigma_w^2 (T_i/C_T)^{-1/3}$.

Set $D_i = \{(x_k^{(i)}, u_k^{(i)})\}_{k=1}^{T_i} \leftarrow$ evolve system forward T_i steps, where each action $u_k^{(i)}$ is obtained from the controller $\mathbf{K}^{(i)}$ plus an additional noise term for exploration. More precisely, $\mathbf{u}^{(i)} = \mathbf{K}^{(i)}\mathbf{x}^{(i)} + \boldsymbol{\eta}^{(i)}$, where each entry of $\boldsymbol{\eta}^{(i)}$ is drawn i.i.d. from $\mathcal{N}(0, \sigma_{\eta,i}^2 I_p)$.

Set $(\widehat{A}_i, \widehat{B}_i) \leftarrow \arg\min_{A,B} \sum_{k=1}^{T_i-1} \frac{1}{2}\|x_{k+1}^{(i)} - Ax_k^{(i)} - Bu_k^{(i)}\|_2^2$.

Set $\varepsilon_i \leftarrow \widetilde{\mathcal{O}}\left(\frac{\sigma_w \|K_\star\| C_\star}{\sigma_{\eta,i}(1-\rho_\star)^3}\sqrt{\frac{n+p}{T_i}}\right)$ and $F_i \leftarrow \frac{\widetilde{\mathcal{O}}(1)(i+1)}{1-\rho_\star}$.

Set $\mathbf{K}^{(i+1)} = \boldsymbol{\Phi}_u \boldsymbol{\Phi}_x^{-1}$, where $(\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u)$ are the solution to

$$
\begin{aligned}
&\text{minimize}_{\gamma \in [0,1)} \frac{1}{1-\gamma} \min_{\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u, V} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_x \\ \boldsymbol{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \\
&\qquad \text{s.t.} \begin{bmatrix} zI - \widehat{A}_i & -\widehat{B}_i \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_x \\ \boldsymbol{\Phi}_u \end{bmatrix} = I + \frac{1}{z^{F_i}} V \,, \quad \frac{\sqrt{2}\varepsilon_i}{1 - C_x \rho^{F_i+1}} \left\| \begin{bmatrix} \boldsymbol{\Phi}_{\mathbf{x}} \\ \boldsymbol{\Phi}_{\mathbf{u}} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \gamma \,, \\
&\qquad \|V\| \leq C_x \rho^{F_i+1} \,, \quad \boldsymbol{\Phi}_x \in \mathcal{S}_{F_i}(C_x, \rho) \,, \quad \boldsymbol{\Phi}_u \in \mathcal{S}_{F_i}(C_u, \rho) \,.
\end{aligned}
$$

end for

3.1 Regret Upper Bounds

Our guarantees for Algorithm 1 are stated in terms of certain system specific constants, which we define here. We let $K_\star$ denote the static feedback solution to the LQR problem for $(A_\star, B_\star, Q, R)$. Next, we define $(C_\star, \rho_\star)$ such that the closed loop system $A_\star + B_\star K_\star$ belongs to $\mathcal{S}(C_\star, \rho_\star)$. Our main assumption is stated as follows.

Assumption 3.1. *We are given a controller $\mathbf{K}^{(0)}$ that stabilizes the true system $(A_\star, B_\star)$. Furthermore, letting $(\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_u)$ denote the response of $\mathbf{K}^{(0)}$ on $(A_\star, B_\star)$, we assume that $\boldsymbol{\Phi}_x \in \mathcal{S}(C_x, \rho)$ and $\boldsymbol{\Phi}_u \in \mathcal{S}(C_u, \rho)$, where the constants C_x, C_u, ρ are defined in Algorithm 1.*

The requirement of an initial stabilizing controller $\mathbf{K}^{(0)}$ is not restrictive; Dean et al. [6] provide an offline strategy for finding such a controller. Furthermore, in practice Algorithm 1 can be initialized with no controller, with random inputs applied instead to the system in the first epoch to estimate $(A_\star, B_\star)$ within an initial confidence set for which the synthesis problem becomes feasible.

Our first guarantee is on the rate of estimation of $(A_\star, B_\star)$ as the algorithm progresses through time. This result builds on recent progress [17] for estimation along trajectories of a linear dynamical system. For what follows, the notation $\widetilde{\mathcal{O}}(\cdot)$ hides absolute constants and $\mathrm{polylog}\left(T, \frac{1}{\delta}, C_\star, \frac{1}{1-\rho_\star}, n, p, \|B_\star\|, \|K_\star\|\right)$ factors.

Theorem 3.2. *Fix a $\delta \in (0,1)$ and suppose that Assumption 3.1 holds. With probability at least $1-\delta$ the following statement holds. Suppose that T is at an epoch boundary. Let $(\widehat{A}(T), \widehat{B}(T))$ denote the current estimate of $(A_\star, B_\star)$ computed by Algorithm 1 at the end of time T. Then, this estimate satisfies the guarantee*

$$
\max\{\|\widehat{A}(T) - A_\star\|, \|\widehat{B}(T) - B_\star\|\} \leq \widetilde{\mathcal{O}}\left(\frac{C_\star \|K_\star\|}{(1-\rho_\star)^3} \frac{\sqrt{n+p}}{T^{1/3}}\right) .
$$

Theorem 3.2 shows that Algorithm 1 achieves a consistent estimate of the true dynamics $(A_\star, B_\star)$, and learns at a rate of $\widetilde{\mathcal{O}}(T^{-1/3})$. We note that consistency of parameter estimates is not a guarantee provided by OFU or TS based approaches.

Next, we state an upper bound on the regret incurred by Algorithm 1.

Theorem 3.3. *Fix a $\delta \in (0, 1)$ and suppose that Assumption 3.1 holds. With probability at least $1-\delta$ the following statement holds. For all $T \geq 0$ we have that Algorithm 1 satisfies*

$$\mathsf{Regret}(T) \leq \widetilde{\mathcal{O}}\left((n+p) \frac{C_\star^4 (1+\|K_\star\|)^4 (1+\|B_\star\|)^2 J_\star}{(1-\rho_\star)^{16}} T^{2/3} \right) .$$

Here, the notation $\widetilde{\mathcal{O}}(\cdot)$ also hides $o(T^{2/3})$ terms.

Our proof strategy works as follows. We first decompose regret by epochs as follows:

$$\mathsf{Regret}(T) = \sum_{i=0}^{\mathcal{O}(\log_2 T)} \sum_{k=1}^{T_i} ((x_k^{(i)})^\top Q x_k^{(i)} + (u_k^{(i)})^\top R u_k^{(i)} - J_\star) ,$$

where $x_k^{(i)}$ denotes the state at the k-th timestep in the i-th epoch (and similarly for $u_k^{(i)}$). By standard concentration of measure arguments, we can upper bound w.h.p. the per-epoch regret $\sum_{k=1}^{T_i} ((x_k^{(i)})^\top Q x_k^{(i)} + (u_k^{(i)})^\top R u_k^{(i)} - J_\star)$ by its expected value plus a deviation term that involves the norm of $x_0^{(i)}$. Because we constrain the impulse response coefficients of the SLS response $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\}$ in Algorithm 1, this allows to easily bound $\|x_0^{(i)}\|_2$ w.h.p. again by using standard concentration arguments. We then use the SLS machinery to quantify the difference between the expected cost over the horizon T_i minus $J_\star$, which yields that the regret incurred during epoch i is bounded by $\widetilde{\mathcal{O}}(T_i(\sigma_{\eta,i}^2/\sigma_w^2 + \varepsilon_{i-1})J_\star)$, where ε_{i-1} is the estimation error, and the $\mathcal{O}(\sigma_{\eta,i}^2/\sigma_w^2)$ contribution is the additional cost incurred from injecting exploration noise. We then bound our estimation error by $\varepsilon_i = \widetilde{\mathcal{O}}((\sigma_w/\sigma_{\eta,i})T_i^{-1/2})$ using Theorem 3.2. Setting $\sigma_{\eta,i}^2 = \sigma_w^2 T_i^{-\alpha}$, we have the per-epoch regret is bounded by $\widetilde{\mathcal{O}}(T_i^{1-\alpha} + T_i^{1-(1-\alpha)/2})$. Choosing $\alpha = 1/3$ to balance these competing powers of T_i and summing over logarithmic number of epochs, we obtain a final regret of $\widetilde{\mathcal{O}}(T^{2/3})$.

The main difficulty in the proof is ensuring that the transient behavior of the resulting controllers is uniformly bounded when applied to the true system. Prior works sidestep this issue by assuming that the true dynamics lie within a (known) compact set for which the Heine-Borel theorem asserts the existence of finite constants that capture this behavior. We go a step further and work through the perturbation analysis which allows us to give a regret bound that depends only on simple quantities of the true system $(A_\star, B_\star)$. The full proof is given in the appendix.

Finally, we remark that the dependence on $1/(1-\rho_\star)$ in our results is an artifact of our perturbation analysis, and we leave sharpening this dependence to future work.

3.2 Regret Lower Bounds and Parameter Estimation Rates

We saw that Algorithm 1 achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret with high probability. Now we provide a matching algorithmic lower bound on the expected regret, showing that the analysis presented in Section 3.1 is sharp as a function of T. Moreover, our lower bound characterizes how much regret must be accrued in order to achieve a specified estimation rate for the system parameters $(A_\star, B_\star)$.

Theorem 3.4. *Let the initial state x_0 be distributed according to the steady state distribution $\mathcal{N}(0, P_\infty)$ of the optimal closed loop system, and let $\{u_t\}_{t\geq 0}$ be any sequence of inputs as in Section 2. Furthermore, let $f : \mathbb{R} \to \mathbb{R}$ be any function such that with probability $1-\delta$ we have*

$$\lambda_{\min}\left(\sum_{k=0}^{T-1} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \begin{bmatrix} x_k^\top & u_k^\top \end{bmatrix} \right) \geq f(T) . \tag{3.1}$$

Then, there exist positive values T_0 and C_0 such that for all $T \geq T_0$ we have

$$\sum_{k=0}^{T} \mathbb{E}\left[x_k^\top Q x_k + u_k^\top R u_k - J_\star \right] \geq \frac{1}{2}(1-\delta)\lambda_{\min}(R)(1+\sigma_{\min}(K_\star)^2) f(T-T_0) - C_0 ,$$

where T_0 and C_0 are functions of $A_\star$, $B_\star$, Q, R, σ_w^2, and n. We note the specific form of T_0 and C_0 are given in the proof.

The proof of the estimation error Theorem 3.2 shows that Algorithm 1 satisfies Eq. (3.1) with $f(T) = \widetilde{\mathcal{O}}(T\sigma^2_{\eta,\Theta(\log_2(T))})$. Since the exploration variance $\sigma^2_{\eta,i}$ used by Algorithm 1 during the i-th epoch is given by $\sigma^2_{\eta,i} = \mathcal{O}(\sigma^2_w T^{-i/3})$, we obtain the following corollary which demonstrates the sharpness of our regret analysis with respect to the scaling of T.

Corollary 3.5. *For $T > C_1(n, \delta, \sigma^2_w, A_\star, B_\star, Q, R)$ the expected regret of Algorithm 1 satisfies*

$$\sum_{k=1}^{T} \mathbb{E}\left[x_k^\top Q x_k + u_k^\top R u_k - J_\star\right] \geq \widetilde{\Omega}(\lambda_{\min}(R)(1 + \sigma_{\min}(K_\star)^2) T^{2/3}) .$$

A natural question to ask is how much regret does any algorithm accrue in order to achieve estimation error $\|\widehat{A} - A_\star\| \leq \varepsilon$ and $\|\widehat{B} - B_\star\| \leq \varepsilon$. From Theorem 3.2 we know that Algorithm 1 estimates $(A_\star, B_\star)$ at rate $\widetilde{\mathcal{O}}(T^{-1/3})$. Therefore, in order to achieve ε estimation error, T must be $\widetilde{\Omega}(\varepsilon^{-3})$. Hence, Theorem 3.3 implies that the regret of Algorithm 1 to achieve ε estimation error is $\widetilde{\mathcal{O}}(\varepsilon^{-2})$.

Interestingly, let us consider any other Algorithm achieving $\mathcal{O}(T^\alpha)$ regret for some $0 < \alpha < 1$. Then, Theorem 3.4 suggests that the best rate achievable by such an algorithm is $\mathcal{O}(T^{-\alpha/2})$, since the minimum eigenvalue condition Eq. (3.1) governs the signal-to-noise ratio. In the case of linear-regression with independent data it is known that the minimax estimation rate is lower bounded by square root of the inverse of the minimum eigenvalue (3.1). We conjecture that the same results holds in our case. Therefore, to achieve ε estimation error, any Algorithm would likely require $\Omega(\varepsilon^{-2})$ regret, showing that Algorithm 1 is optimal up to logarithmic factors in this sense. Finally, we note that while Algorithm 1 estimates $(A_\star, B_\star)$ at a rate $\widetilde{\mathcal{O}}(T^{-1/3})$, Theorem 3.4 suggests that any algorithm achieving the $\mathcal{O}(\sqrt{T})$ regret would estimate $(A_\star, B_\star)$ at a rate $\Omega(T^{-1/4})$.

4 Experiments

Regret Comparison. We illustrate the performance of several adaptive schemes empirically. We compare the proposed robust adaptive method with non-Bayesian Thompson sampling (TS) as in Abeille and Lazaric [4] and a heuristic projected gradient descent (PGD) implementation of OFU. As a simple baseline, we use the nominal control method, which synthesizes the optimal infinite-horizon LQR controller for the estimated system and injects noise with the same schedule as the robust approach. Computational burden varies across adaptive methods due to differences in both cost and frequency of controller synthesis; implementation details and computational considerations for all methods are in Section G of the full version [7].

The comparison experiments are carried out on the following LQR problem:

$$A_\star = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, \quad B_\star = I, \quad Q = 10I, \quad R = I, \quad \sigma_w = 1 . \tag{4.1}$$

This system corresponds to a marginally unstable Laplacian system where adjacent nodes are weakly connected; these dynamics were also studied by [3, 6, 18]. The cost is such that input size is penalized relatively less than state. This problem setting is amenable to robust methods due to both the cost ratio and the marginal instability, which are factors that may hurt optimistic methods. In Section H of the full version [7], we show similar results for an unstable system with large transients.

To standardize the initialization of the various adaptive methods, we use a rollout of length $T_0 = 100$ where the input is a stabilizing controller plus Gaussian noise with fixed variance $\sigma_u = 1$. This trajectory is not counted towards the regret, but the recorded states and inputs are used to initialize parameter estimates. In each experiment, the system starts from $x_0 = 0$ to reduce variance over runs. For all methods, the actual errors $\widehat{A}_t - A_\star$ and $\widehat{B}_t - B_\star$ are used rather than bounds or bootstrapped estimates. The effect of this choice on regret is small, as examined empirically in Section H of [7].

The performance of the various adaptive methods over time is compared in Figure 1. The median and 90th percentile cumulative regret over 500 instances is displayed in Figure 1a, which gives an idea of both typical and worst-case behavior. The regret of the optimal LQR controller for the true system is displayed as a baseline. Overall, the methods have very similar performance. One benefit of robustness is the guaranteed stability and bounded infinite-horizon cost at every point during

(a) Cumulative Regret **(b)** Infinite Horizon LQR Cost

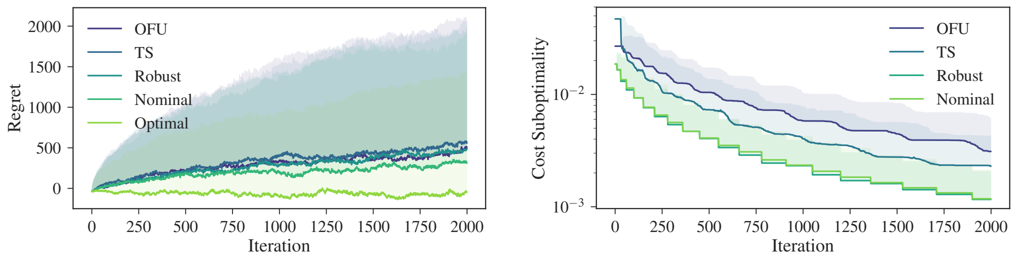

Figure 1: A comparison of different adaptive methods on 500 experiments of the marginally unstable Laplacian example in 4.1. In (a), the median and 90th percentile cumulative regret is plotted over time. In (b), the median and 90th percentile infinite-horizon LQR cost of the epoch's controller.

(a) Demand Forecasting **(b)** Constraint Satisfaction

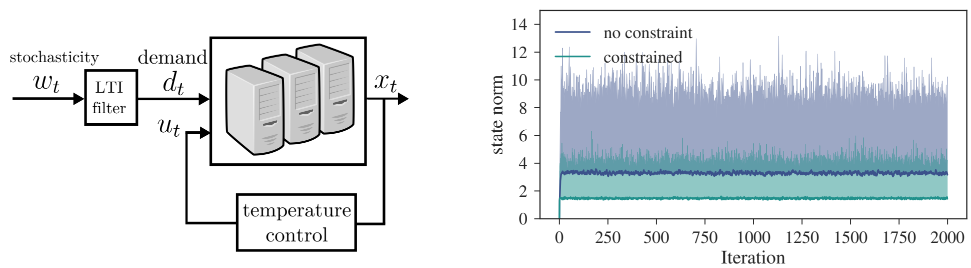

Figure 2: The addition of constraints in the robust synthesis problem can guarantee the safe execution of adaptive systems. We consider an example inspired by demand forecasting, as illustrated in (a), where the left hand side of the diagram represents unknown dynamics. The median and maximum values of $\|x_t\|_\infty$ over 500 trials are plotted for both the unconstrained and constrained synthesis problems in (b).

operation. In Figure 1b, this infinite-horizon LQR cost is plotted for the controllers played during each epoch. This value measures the cost of using each epoch's controller indefinitely, rather than continuing to update its parameters. The robust adaptive method performs relatively better than other adaptive algorithms, indicating that it is more amenable to early stopping, i.e., to turning off the adaptive component of the algorithm and playing the current controller indefinitely.

Extension to Uncertain Environment with State Constraints. The proposed robust adaptive method naturally generalizes beyond the standard LQR problem. We consider a disturbance forecasting example which incorporates environmental uncertainty and safety constraints. Consider a system with known dynamics driven by stochastic disturbances that are now correlated in time. We model the disturbance process as the output of an unknown autonomous LTI system, as illustrated in Figure 2a. This setting can be interpreted as a demand forecasting problem, where, for example, the system is a server farm and the disturbances represent changes in the amount of incoming jobs. If the dynamics of the correlated disturbance process are known, this knowledge can be used for more cost-effective temperature control.

We let the system $(A_\star, B_\star)$ with known dynamics be described by the graph Laplacian dynamics as in Eq. (4.1). The disturbance dynamics are unknown and are governed by a stable system transition matrix A_d, resulting in the following dynamics for the full system:

$$\begin{bmatrix} x_{t+1} \\ d_{t+1} \end{bmatrix} = \begin{bmatrix} A_\star & I \\ 0 & A_d \end{bmatrix} \begin{bmatrix} z_t \\ d_t \end{bmatrix} + \begin{bmatrix} B_\star \\ 0 \end{bmatrix} u_t + \begin{bmatrix} 0 \\ I \end{bmatrix} w_t \,, \quad A_d = \begin{bmatrix} 0.5 & 0.1 & 0 \\ 0 & 0.5 & 0.1 \\ 0 & 0 & 0.5 \end{bmatrix}.$$

The costs are set to model expensive inputs, with $Q = I$ and $R = 1 \times 10^3 I$. The controller synthesis problem in Line 8 of Algorithm 1 is modified to reflect the problem structure, and crucially, we add a constraint on the system response $\mathbf{\Phi}_x$. Further details of the formulation are explained in Section H of [7]. Figure 2b illustrates the effect. While the unconstrained synthesis results in trajectories with large state values, the constrained synthesis results in much more moderate behavior.

5 Conclusions and Future Work

We presented a polynomial-time algorithm for the adaptive LQR problem that provides high probability guarantees of sub-linear regret. In contrast to other approaches to this problem, our robust adaptive method guarantees stability, robust performance, and parameter estimation. We also explored the interplay between regret minimization and parameter estimation, identifying fundamental limits connecting the two.

Several questions remain to be answered. It is an open question whether a polynomial-time algorithm can achieve a regret of $\widetilde{\mathcal{O}}(\sqrt{T})$. In our implementation of OFU, we observed that PGD performed quite effectively. Interesting future work is to see if the techniques of Fazel et al. [9] for policy gradient optimization on LQR can be applied to prove convergence of PGD on the OFU subroutine, which would provide an optimal polynomial-time algorithm. Moreover, we observed that OFU and TS methods in practice gave estimates of system parameters that were comparable with our method which explicitly adds excitation noise. It seems that the switching of control policies at epoch boundaries provides more excitation for system identification than is currently understood by the theory. Furthermore, practical issues that remain to be addressed include satisfying safety constraints and dealing with nonlinear dynamics; in both settings, finite-sample parameter estimation/system identification and adaptive control remain an open problem.

Acknowledgments

We thank the anonymous reviewers for their feedback, which improved the clarity of our presentation. SD is supported by an NSF Graduate Research Fellowship under Grant No. DGE 1752814. As part of the RISE lab, HM is generally supported in part by NSF CISE Expeditions Award CCF-1730628, DHS Award HSHQDC-16-3-00083, and gifts from Alibaba, Amazon Web Services, Ant Financial, CapitalOne, Ericsson, GE, Google, Huawei, Intel, IBM, Microsoft, Scotiabank, Splunk and VMware. BR is generously supported in part by NSF award CCF-1359814, ONR awards N00014-17-1-2191, N00014-17-1-2401, and N00014-17-1-2502, the DARPA Fundamental Limits of Learning (Fun LoL) and Lagrange Programs, and an Amazon AWS AI Research Award.

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret Bounds for the Adaptive Control of Linear Quadratic Systems. In *Conference on Learning Theory*, 2011.

[2] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian Optimal Control of Smoothly Parameterized Systems: The Lazy Posterior Sampling Algorithm. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

[3] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-Free Linear Quadratic Control via Reduction to Expert Prediction. *arXiv:1804.06021*, 2018.

[4] Marc Abeille and Alessandro Lazaric. Thompson Sampling for Linear-Quadratic Control Problems. In *AISTATS*, 2017.

[5] S. Bittanti and M. C. Campi. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information and Systems*, 6(4), 2006.

[6] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the Sample Complexity of the Linear Quadratic Regulator. *arXiv:1710.01688*, 2017.

[7] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator. *arXiv:1805.09388*, 2018.

[8] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite Time Analysis of Optimal Adaptive Policies for Linear-Quadratic Systems. *arXiv:1711.07230*, 2017.

[9] Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator. In *International Conference on Machine Learning*, 2018.

[10] Claude-Nicolas Fiechter. PAC Adaptive Control of Linear Systems. In *Conference on Learning Theory*, 1997.

[11] Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy. Efficient Reinforcement Learning for High Dimensional Linear Quadratic Systems. In *Neural Information Processing Systems*, 2012.

[12] Petros A Ioannou and Jing Sun. *Robust adaptive control*, volume 1. PTR Prentice-Hall Upper Saddle River, NJ, 1996.

[13] Miroslav Krstic, Ioannis Kanellakopoulos, and Peter V Kokotovic. *Nonlinear and adaptive control design*. Wiley, 1995.

[14] Nikolai Matni, Yuh-Shyang Wang, and James Anderson. Scalable system level synthesis for virtually localizable systems. In *IEEE Conference on Decision and Control*, 2017.

[15] Ian Osband and Benjamin Van Roy. Posterior Sampling for Reinforcement Learning Without Episodes. *arXiv:1608.02731*, 2016.

[16] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based Control of Unknown Linear Systems with Thompson Sampling. *arXiv:1709.04047*, 2017.

[17] Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification. In *Conference on Learning Theory*, 2018.

[18] Stephen Tu and Benjamin Recht. Least-Squares Temporal Difference Learning for the Linear Quadratic Regulator. In *International Conference on Machine Learning*, 2018.

[19] Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. A System Level Approach to Controller Synthesis. *arXiv:1610.04815*, 2016.

[20] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. 1995.

A Background on System Level Synthesis

We begin by defining two function spaces which we use extensively throughout:

$$\mathcal{RH}_\infty = \{\mathbf{M} : \mathbb{C} \longrightarrow \mathbb{C}^{n\times p} \mid \mathbf{M}(z) \text{ is rational }, \; \mathbf{M}(z) \text{ is analytic on } \mathbb{D}^c\} \,, \tag{A.1}$$

$$\mathcal{RH}_\infty(C,\rho) = \{\mathbf{M} \in \mathcal{RH}_\infty \mid \|\mathbf{M}[k]\| \le C\rho^k \,, \;\; k = 1,2,...\} \,. \tag{A.2}$$

Note that we use $\mathcal{S}(C,\rho)$ to denote $\mathcal{RH}_\infty(C,\rho)$ in the main body of the text.

Recall that our main object of interest is the system

$$x_{k+1} = Ax_k + Bu_k + w_k \,,$$

and our goal is to design a LTI feedback control policy $\mathbf{u} = \mathbf{K}\mathbf{x}$ such that the resulting closed loop system is stable. For a given $\mathbf{K}$, we refer to the closed loop transfer functions from $\mathbf{w} \mapsto \mathbf{x}$ and $\mathbf{w} \mapsto \mathbf{u}$ as the *system response*. Symbolically, we denote these maps as $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$. Simple algebra shows that given $\mathbf{K}$, these maps take on the form

$$\mathbf{\Phi}_x = (zI - A - B\mathbf{K})^{-1} \,, \;\; \mathbf{\Phi}_u = \mathbf{K}(zI - A - B\mathbf{K})^{-1} \,. \tag{A.3}$$

We then have the following theorem parameterizing the set of such stable closed-loop transfer functions that are achievable by a stabilizing controller $\mathbf{K}$.

Theorem A.1 (State-Feedback Parameterization [19]). *The following are true:*

- *The affine subspace defined by*

$$\begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I, \;\; \mathbf{\Phi}_x, \mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty \tag{A.4}$$

  parameterizes all system responses (**??**) *from* $\mathbf{w}$ *to* $(\mathbf{x}, \mathbf{u})$*, achievable by an internally stabilizing state-feedback controller* $\mathbf{K}$.

- *For any transfer matrices* $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\}$ *satisfying* (**??**)*, the controller* $\mathbf{K} = \mathbf{\Phi}_u\mathbf{\Phi}_x^{-1}$ *is internally stabilizing and achieves the desired system response* (**??**).

If $\mathbf{K}$ stabilizes (A, B), then the LQR cost of $\mathbf{K}$ on (A, B) can be written by Parseval's identity as

$$J(A,B,\mathbf{K};\sigma_w^2 I) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{k=1}^{T} x_k^\top Q x_k + u_k^\top R u_k\right] = \sigma_w^2 \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2}^2 \,. \tag{A.5}$$

More generally, we will define $J(A, B, \mathbf{K}; \Sigma)$ to be the LQR cost when the process noise is driven by $w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$. When we omit the last argument, we mean $\sigma_w^2 = 1$, i.e. $J(A, B, \mathbf{K}) = J(A, B, \mathbf{K}; I)$.

In [6], the authors use SLS to study how uncertainty in the true parameters $(A_\star, B_\star)$ affect the LQR objective cost. Our analysis relies on these tools, which we briefly describe below.

The starting point for the theory is a characterization of all *robustly* stabilizing controllers.

Theorem A.2 ([14]). *Suppose that the transfer matrices* $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\} \in \frac{1}{z}\mathcal{RH}_\infty$ *satisfy*

$$\begin{bmatrix} zI - A & -B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I + \mathbf{\Delta}. \tag{A.6}$$

Then the controller $\mathbf{K} = \mathbf{\Phi}_u\mathbf{\Phi}_x^{-1}$ *stabilizes the system described by* (A, B) *if and only if* $(I+\mathbf{\Delta})^{-1} \in \mathcal{RH}_\infty$*. Furthermore, the resulting system response is given by*

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} (I + \mathbf{\Delta})^{-1}\mathbf{w}. \tag{A.7}$$

This robustness result is used to derive a cost perturbation result for LQR.

Lemma A.3 ([6]). *Let the controller $\mathbf{K}$ stabilize $(\widehat{A}, \widehat{B})$ and $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ be its corresponding system response on system $(\widehat{A}, \widehat{B})$. Then if $\mathbf{K}$ stabilizes (A, B), it achieves the following LQR cost*

$$\sqrt{J(A,B,\mathbf{K})} = \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \left(I + [\Delta_A \quad \Delta_B] \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right)^{-1} \right\|_{\mathcal{H}_2} . \tag{A.8}$$

Furthermore, letting

$$\widehat{\mathbf{\Delta}} := [\Delta_A \quad \Delta_B] \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} . \tag{A.9}$$

a sufficient condition for $\mathbf{K}$ to stabilize (A, B) is that $\|\widehat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} < 1$. An upper bound on $\|\widehat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty}$ is given by, for any $\alpha \in (0, 1)$,

$$\|\widehat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} \leq \left\| \begin{bmatrix} \frac{\varepsilon_A}{\sqrt{\alpha}} \mathbf{\Phi}_x \\ \frac{\varepsilon_B}{\sqrt{1-\alpha}} \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} , \tag{A.10}$$

where we assume that $\|A - \widehat{A}\|_2 \leq \varepsilon_A$ and $\|B - \widehat{B}\|_2 \leq \varepsilon_B$.

B Synthesis Results

We first study the following infinite-dimensional synthesis problem.

$$\begin{aligned} \text{minimize}_{\gamma \in [0,1)} \frac{1}{1-\gamma} \min_{\mathbf{\Phi}_x, \mathbf{\Phi}_u} & \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \\ \text{s.t.} & \begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I, \quad \left\| \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \tfrac{\gamma}{\sqrt{2}\varepsilon} \\ & \mathbf{\Phi}_x \in \frac{1}{z}\mathcal{RH}_\infty(C_x, \rho), \ \mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty(C_u, \rho). \end{aligned} \tag{B.1}$$

We will conduct our analysis assuming that this infinite-dimensional problem is solvable. Later on, we will show how to relax this problem to a finite-dimension one via FIR truncation, and show the minor modifications needed to the analysis for the guarantees to hold.

We now prove a sub-optimality guarantee on the solution to (**??**) which holds for certain choices of ε and the coefficients (C_x, ρ_x) and (C_u, ρ_u). This result also establishes an important technical consideration, which is when the problemmmm (**??**) is feasible.

Theorem B.1. *Let $J_\star$ denote the minimal LQR cost achievable by any controller for the dynamical system with transition matrices $(A_\star, B_\star)$, and let $K_\star$ denote its optimal static feedback contoller. Suppose that $\mathfrak{R}_{A_\star + B_\star K_\star} \in \mathcal{RH}_\infty(C_\star, \rho_\star)$ and that (wlog) $\rho_\star \geq 1/e$. Suppose furthermore that ε is small enough to satisfy the following conditions:*

$$\begin{aligned} \varepsilon(1 + \|K_\star\|)\|\mathfrak{R}_{A_\star+B_\star K_\star}\|_{\mathcal{H}_\infty} &\leq 1/5 , \\ \varepsilon(1 + \|K_\star\|)C_\star &\leq 1 - \rho_\star . \end{aligned}$$

Let $(\widehat{A}, \widehat{B})$ be any estimates of the transition matrices such that $\max\{\|\Delta_A\|, \|\Delta_B\|\} \leq \varepsilon$. Then, if (C_x, ρ) and (C_u, ρ) are set as,

$$\begin{aligned} C_x &= \frac{\mathcal{O}(1)C_\star}{1-\rho_\star} , \\ C_u &= \frac{\mathcal{O}(1)\|K_\star\|C_\star}{1-\rho_\star} , \\ \rho &= (1/4)\rho_\star + 3/4 , \end{aligned}$$

*we have that (a) the program (**??**) is feasible, (b) letting $\mathbf{K}$ denote an optimal solution to (**??**), the relative error in the LQR cost is*

$$J(A_\star, B_\star, \mathbf{K}) \leq (1 + 5\varepsilon(1 + \|K_\star\|)\|\mathfrak{R}_{A_\star+B_\star K_\star}\|_{\mathcal{H}_\infty})^2 J_\star , \tag{B.2}$$

and (c) if furthermore $\varepsilon(C_x + C_u) \le 2(1-\rho_\star)$, the response $\{\widehat{\mathbf{\Phi}}_x, \widehat{\mathbf{\Phi}}_u\}$ of $\mathbf{K}$ on the true system $(A_\star, B_\star)$ satisfies

$$\widehat{\mathbf{\Phi}}_x \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)C_\star}{(1-\rho_\star)^2}, 7/8 + (1/8)\rho_\star\right) ,$$
$$\widehat{\mathbf{\Phi}}_u \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)\|K_\star\|C_\star}{(1-\rho_\star)^2}, 7/8 + (1/8)\rho_\star\right) .$$

Proof. The proof of (a) and (b) is nearly identical to that given in [6], which works by showing that $\mathbf{\Phi}_x = \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}$ and $\mathbf{\Phi}_u = K_\star\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}$ is a feasible response which gives the desired sub-optimality guarantee. The only modification is that we need to find constants C_x, C_u, ρ for which $\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} \in \frac{1}{z}\mathcal{RH}_\infty(C_x,\rho)$ and $K_\star\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} \in \frac{1}{z}\mathcal{RH}_\infty(C_u,\rho)$. We do this by writing

$$\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} = \mathfrak{R}_{A_\star+B_\star K_\star}(I-\mathbf{\Delta})^{-1} \ , \ \ \mathbf{\Delta} = (\Delta_A + \Delta_B K_\star)\mathfrak{R}_{A_\star+B_\star K_\star} \ .$$

By the definition of $\mathbf{\Delta}$ and our assumptions, we have that

$$\mathbf{\Delta} \in \mathcal{RH}_\infty(\varepsilon(1+\|K_\star\|)C_\star, \rho_\star) \ , \ \ \|\mathbf{\Delta}\|_{\mathcal{H}_\infty} < 1 \ .$$

This places us in a position to apply Lemma **??**, from which we conclude that

$$(I-\mathbf{\Delta})^{-1} \in \mathcal{RH}_\infty\left(\mathcal{O}(1), \mathsf{Avg}(\rho_\star, 1)\right) \ .$$

Now applying Lemma **??** to $\mathfrak{R}_{A_\star+B_\star K_\star}(I-\mathbf{\Delta})^{-1}$, we conclude that

$$\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)C_\star}{1-\rho_\star}, (1/4)\rho_\star + 3/4\right) \ .$$

The claims of (a) and (b) now follows.

Now for the proof of (c). Let $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\}$ be the solution to (**??**). We have that

$$\begin{bmatrix}\widehat{\mathbf{\Phi}}_x \\ \widehat{\mathbf{\Phi}}_u\end{bmatrix} = \begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix}(I+\widehat{\mathbf{\Delta}})^{-1} \ , \ \ \widehat{\mathbf{\Delta}} = \begin{bmatrix}\Delta_A & \Delta_B\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_x \\ \mathbf{\Phi}_u\end{bmatrix} \ .$$

We know that $\|\widehat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} < 1$ by the constraints of the optimization problem (**??**) and furthermore,

$$\widehat{\mathbf{\Delta}} \in \mathcal{RH}_\infty(\varepsilon(C_x+C_u), \rho) \ .$$

By assumption we have $\varepsilon(C_x + C_u) \le 2$, from which we conclude using Lemma **??** that

$$(I+\widehat{\mathbf{\Delta}})^{-1} \in \mathcal{RH}_\infty\left(\mathcal{O}(1), \mathsf{Avg}(\rho, 1)\right) \ .$$

Furthermore, from Lemma **??**, we conclude that

$$\mathbf{\Phi}_x(I+\widehat{\mathbf{\Delta}})^{-1} \in \mathcal{RH}_\infty\left(\frac{C_x}{1-\rho}, 3/4 + (1/4)\rho\right) \ ,$$
$$\mathbf{\Phi}_u(I+\widehat{\mathbf{\Delta}})^{-1} \in \mathcal{RH}_\infty\left(\frac{C_u}{1-\rho}, 3/4 + (1/4)\rho\right) \ .$$

The claim now follows by plugging in the values of C_x, C_u, and ρ. □

B.1 Suboptimality bounds for FIR truncated SLS

Optimization problem (**??**) is convex but infinite dimensional, and as far as we are aware does not admit an efficient solution. In Algorithm 1, we instead propose solving the following FIR approximation to problem (**??**):

$$\begin{aligned}
&\text{minimize}_{\gamma\in[0,1)} \frac{1}{1-\gamma} \min_{\mathbf{\Phi_x},\mathbf{\Phi_u},V} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2}\end{bmatrix}\begin{bmatrix}\mathbf{\Phi_x} \\ \mathbf{\Phi_u}\end{bmatrix}\right\|_{\mathcal{H}_2} \\
&\text{s.t.} \begin{bmatrix} zI - \widehat{A} & -\widehat{B}\end{bmatrix}\begin{bmatrix}\mathbf{\Phi_x} \\ \mathbf{\Phi_u}\end{bmatrix} = I + \frac{1}{z^F}V \ , \frac{\sqrt{2}\varepsilon}{1-C_x\rho^{F+1}}\left\|\begin{bmatrix}\mathbf{\Phi_x} \\ \mathbf{\Phi_u}\end{bmatrix}\right\|_{\mathcal{H}_\infty} \le \gamma \\
&\|V\|_2 \le C_x\rho^{F+1} \ , \ \mathbf{\Phi_x} \in \frac{1}{z}\mathcal{RH}_\infty^F(C_x,\rho), \mathbf{\Phi_u} \in \frac{1}{z}\mathcal{RH}_\infty^F(C_u,\rho) \ .
\end{aligned} \tag{B.3}$$

where here F denotes the FIR truncation length used. This optimization problem can be posed as a finite dimensional semidefinite program (see Section **??**). Let $\mathbf{K}(F)$ denote the resulting controller. We begin with a lemma identifying conditions under which optimization problem (**??**) is feasible. to ease notation going forward, we let $\zeta := \varepsilon(1+\|K_\star\|_2)\|\mathfrak{R}_{A_\star+B_\star K_\star}\|_{\mathcal{H}_\infty}$.

Lemma B.2. *Let the assumptions of Theorem **??** hold, and further assume that*

$$F_0 \geq \frac{\log(2C_x)}{\log(1/\rho)} - 1 \,.$$

Then optimization problem (**??**) *is feasible for any* $F \geq F_0$.

Proof. We construct a feasible solution as follows. Let $\mathbf{\Phi}_x = \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}(1:F)$, $\mathbf{\Phi}_u = K_\star \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}(1:F)$, $V = \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}(F+1)$, and $\gamma = \frac{2\sqrt{2}\zeta}{1-\zeta}$. First, the proposed $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ are FIR of length F, and hence, using the same arguments as in the proof of Theorem **??**, $\mathbf{\Phi}_x \in \mathcal{RH}_\infty^F(C_x, \rho)$ and $\mathbf{\Phi}_u \in \mathcal{RH}_\infty^F(C_u, \rho)$. It then also follows immediately that $\|V\|_2 = \|\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}(F+1)\|_2 \leq C_x \rho^{F+1}$.

Note that the affine constraint

$$\begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} = I + \frac{1}{z^F} V \tag{B.4}$$

is equivalent to

$$\Phi_x(t+1) = \widehat{A}\Phi_x(t) + \widehat{B}\Phi_u(t), \; \Phi_x(1) = I,$$

for $1 \leq t < F$. We have by construction that the proposed $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_u$ satisfy this constraint. Further, the combination of the FIR constraints and the affine constraint (**??**) impose that

$$\Phi_x(F+1) = \widehat{A}\Phi_x(F) + \widehat{B}\Phi_u(F) - V = 0.$$

Now notice that for the proposed $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$, we have that $\widehat{A}\Phi_x(F) + \widehat{B}\Phi_u(F) = (\widehat{A} + \widehat{B}K_\star)\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}(F) = \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}(F+1)$, where the last equality follows from the fact that $\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}(t+1) = (\widehat{A} + \widehat{B}K_\star)^t$. It follows that $\Phi_x(F+1) = 0$, as desired.

It remains to prove that

$$\frac{\sqrt{2}\varepsilon}{1 - C_x\rho^{F+1}} \left\| \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \frac{2\sqrt{2}\zeta}{1-\zeta} < 1.$$

The final inequality follows immediately from the assumption that $\zeta \leq 1/5$. Further, note that

$$\frac{\sqrt{2}\varepsilon}{1 - C_x\rho^{F+1}} \left\| \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq 2\sqrt{2}\varepsilon \left\| \begin{bmatrix} \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} \\ K_\star \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \frac{2\sqrt{2}\zeta}{1-\zeta},$$

where the first inequality follows from the assumption on on F_0 and that the proposed $\mathbf{\Phi}_x$ is a truncation of $\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}$ and that the proposed $\mathbf{\Phi}_u$ is a truncation of $K_\star\mathfrak{R}_{\widehat{A}+\widehat{B}K_\star}$, and final inequality follows by applying the triangle inequality and the definition of ζ. This proves the result. ☐

Next, we use this to bound the suboptimality gap of the performance achieved by the controller implemented using the solutions of optimization problem (**??**).

Lemma B.3. *Let the assumptions of Lemma **??** hold. Fix any* $C_J > 0$*, and further let*

$$F \geq \frac{\log((1 + C_J^{-1})C_x)}{\log(1/\rho)} - 1 \,.$$

Denote by $(\mathbf{\Phi}_x(F), \mathbf{\Phi}_u(F), V(F), \gamma(F))$ *the optimal solution to optimization problem* (**??**)*, and let* $\mathbf{K}(F) = \mathbf{\Phi}_u(F)\mathbf{\Phi}_x^{-1}(F)$*. Then*

$$J(A_\star, B_\star, \mathbf{K}(F)) \leq (1 + C_J)^2(1 + \mathcal{O}(1)\varepsilon(1 + \|K_\star\|_2)\|\mathfrak{R}_{A_\star+B_\star K_\star}\|_{\mathcal{H}_\infty})^2 J_\star. \tag{B.5}$$

Proof. Let

$$\widehat{\mathbf{\Delta}} := \begin{bmatrix} \Delta_A & \Delta_B \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x(F) \\ \mathbf{\Phi}_u(F) \end{bmatrix} \left(I + \frac{1}{z^F} V(F) \right)^{-1}.$$

Further note that using a similar argument to that in the proof of Lemma 4.2 of [6], one can verify that

$$\|\widehat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} \leq \frac{\sqrt{2}\varepsilon}{1 - C_x\rho^{F+1}} \left\| \begin{bmatrix} \mathbf{\Phi}_x(F) \\ \mathbf{\Phi}_u(F) \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \gamma(F),$$

where we have exploited that $(\mathbf{\Phi}_x(F), \mathbf{\Phi}_u(F), V(F), \gamma(F))$ form a feasible solution to optimization problem (**??**).

Then, repeated application of Theorem **??** tells us that the performance achieved by $K(F)$ on the true system is given by

$$\begin{aligned}\sqrt{J(A_\star, B_\star, \mathbf{K}(F))} &= \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x(F) \\ \mathbf{\Phi}_u(F) \end{bmatrix} \left(I + \frac{1}{z^F} V(F) \right)^{-1} (I + \widehat{\mathbf{\Delta}})^{-1} \right\|_{\mathcal{H}_2} \\ &\leq \frac{1}{1 - C_x \rho^{F+1}} \frac{1}{1-\gamma(F)} \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x(F) \\ \mathbf{\Phi}_u(F) \end{bmatrix} \right\|_{\mathcal{H}_2},\end{aligned}$$

where the inequality follows from $\|\widehat{\mathbf{\Delta}}\|_{\mathcal{H}_\infty} \leq \gamma(F) < 1$, and $\|V(F)\|_2 \leq 1/2$ (by the assumption of $F \geq F_0$.

Denote by $(\mathbf{\Phi}_x, \mathbf{\Phi}_u, V, \gamma_0)$ the feasible solution constructed in the proof of Lemma **??**. Then,

$$\begin{aligned}\frac{1}{1 - C_x \rho^{F+1}} \frac{1}{1-\gamma(F)} \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x(F) \\ \mathbf{\Phi}_u(F) \end{bmatrix} \right\|_{\mathcal{H}_2} &\leq \frac{1}{1 - C_x \rho^{F+1}} \frac{1}{1-\gamma_0} \left\| \begin{bmatrix} Q^{\frac{1}{2}} & 0 \\ 0 & R^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \\ &= \frac{1}{1 - C_x \rho^{F+1}} \frac{1}{1-\gamma_0} \sqrt{J_F(\widehat{A}, \widehat{B}, K_\star)} \\ &\leq \frac{1}{1 - C_x \rho^{F+1}} \frac{1}{1-\gamma_0} \sqrt{J(\widehat{A}, \widehat{B}, K_\star)} \\ &\leq \frac{1}{1 - C_x \rho^{F+1}} \frac{1}{1-\gamma_0} \frac{1}{1-\zeta} \sqrt{J_\star},\end{aligned}$$

where the first inequality follows from the optimality of $(\mathbf{\Phi}_x(F), \mathbf{\Phi}_u(F), V(F), \gamma(F))$, the equality and second inequality from the fact that $(\mathbf{\Phi}_x, \mathbf{\Phi}_u)$ are truncations of the response of $K_\star$ on $(\widehat{A}, \widehat{B})$ to the first F time steps, and the final inequality by following similar arguments to the proof of Theorem 4.1 in [6] in applying Theorem **??** and noting that

$$\left\| \begin{bmatrix} \Delta_A & \Delta_B \end{bmatrix} \begin{bmatrix} \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} \\ K_\star \mathfrak{R}_{\widehat{A}+\widehat{B}K_\star} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \zeta < 1.$$

We therefore have that

$$\sqrt{J(A_\star, B_\star, \mathbf{K}(F))} \leq \frac{1}{1 - C_x \rho^{F+1}} \frac{1}{1-\gamma_0} \frac{1}{1-\zeta} \sqrt{J_\star} \leq (1 + C_J) \frac{1}{1-\gamma_0} \frac{1}{1-\zeta} \sqrt{J_\star},$$

where the last inequality follows from the assumptions on F stated in the Lemma. Finally, by assumption $\zeta \leq 1/5 < .8(1+2\sqrt{2})^{-1}$, from which it follows that $(1-\gamma_0)^{-1}(1-\zeta)^{-1} \leq 1 + 20\zeta$, leading to the bound

$$\sqrt{J(A_\star, B_\star, \mathbf{K}(F))} \leq (1 + C_J)(1 + 20\zeta)\sqrt{J_\star}.$$

Squaring both sides proves the result. □

The following Theorem is then immediate.

Theorem B.4. *Let $J_\star$ denote the minimal LQR cost achievable by any controller for $(A_\star, B_\star)$. Let $K_\star$ denote the optimal controller and suppose that $\mathfrak{R}_{A_\star + B_\star K_\star} \in \mathcal{RH}_\infty(C_\star, \rho_\star)$. Fix a $C_J > 0$, and suppose that F_0 and ε satisfy the assumptions of Lemmas* **??** *and* **??**. *Let $(\widehat{A}, \widehat{B})$ be any estimates of the transition matrices such that $\max\{\|\Delta_A\|, \|\Delta_B\|\} \leq \varepsilon$. Then, if (C_x, ρ) and (C_u, ρ) are set as in Lemma* **??**, *we have that (a) the program* (**??**) *is feasible for any truncation length $F \geq F_0$, (b) letting $\mathbf{K}(F)$ denote an optimal solution to* (**??**) *for truncation length F, the relative error in the LQR cost is*

$$J(A_\star, B_\star, \mathbf{K}(F)) \leq (1 + C_J)^2 (1 + \mathcal{O}(1)\varepsilon(1 + \|K_\star\|_2)\|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty})^2 J_\star\,, \qquad \text{(B.6)}$$

and (c) if furthermore $\varepsilon(C_x + C_u) \le \mathcal{O}(1)(1-\rho_\star)^2$, the response $\{\widehat{\mathbf{\Phi}}_x, \widehat{\mathbf{\Phi}}_u\}$ of $\mathbf{K}$ on the true system $(A_\star, B_\star)$ satisfies

$$\widehat{\mathbf{\Phi}}_x \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)C_\star}{(1-\rho_\star)^3}, .999 + .001\rho_\star\right) ,$$
$$\widehat{\mathbf{\Phi}}_u \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)\|K_\star\|_2 C_\star}{(1-\rho_\star)^3}, .999 + .001\rho_\star\right) .$$

Proof. Claims (a) and (b) follow immediately from Lemmas **??** and **??**.

Now for the proof of (c). Let $\{\mathbf{\Phi}_x(F), \mathbf{\Phi}_u(F)\}$ be the solution to (**??**). Then as argued in the proof of Lemma **??**, the response achieved on the true system $(A_\star, B_\star)$ is given by

$$\begin{bmatrix}\mathbf{\Phi}_x(F)\\ \mathbf{\Phi}_u(F)\end{bmatrix}\left(I + \frac{1}{z^F}V(F)\right)^{-1}(I+\widehat{\mathbf{\Delta}})^{-1},$$

where $\widehat{\mathbf{\Delta}}$ is defined as in the proof of Lemma **??**.

We start by noting that $\mathbf{\Phi}_x(F) \in \mathcal{RH}_\infty(C_x, \rho)$, and by the assumption on $F \ge F_0$, it holds that $z^{-F}V(F) \in \mathcal{RH}_\infty(2, \rho^{1/2})$. This allows us to apply Lemma **??** to conclude that $(I + z^{-F}V(F))^{-1} \in \mathcal{RH}_\infty\left(\mathcal{O}(1)(1-\rho^{1/2})^{-1}, \mathsf{Avg}(\rho^{1/2}, 1)\right)$. Thus, applying Lemma **??** we conclude that

$$\mathbf{\Phi}_x(F)\left(I + \frac{1}{z^F}V(F)\right)^{-1} \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)C_x}{1-\rho^{1/2}}, \mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1)\right).$$

A similar argument yields

$$\mathbf{\Phi}_u(F)\left(I + \frac{1}{z^F}V(F)\right)^{-1} \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)C_u}{1-\rho^{1/2}}, \mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1)\right).$$

Now note that

$$\widehat{\mathbf{\Delta}} = (\Delta_A \mathbf{\Phi}_x(F) + \Delta_B \mathbf{\Phi}_u(F))(I + z^{-F}V(F))^{-1}.$$

From the previous argument, we have that

$$\Delta_A \mathbf{\Phi}_x(F)(I + z^{-F}V(F))^{-1} \in \mathcal{RH}_\infty\left(\varepsilon\frac{\mathcal{O}(1)C_x}{1-\rho^{1/2}}, \mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1)\right),$$
$$\Delta_B \mathbf{\Phi}_u(F)(I + z^{-F}V(F))^{-1} \in \mathcal{RH}_\infty\left(\varepsilon\frac{\mathcal{O}(1)C_u}{1-\rho^{1/2}}, \mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1)\right),$$

from which it follows that

$$\widehat{\mathbf{\Delta}} \in \mathcal{RH}_\infty\left(\varepsilon\frac{\mathcal{O}(1)(C_x+C_u)}{1-\rho^{1/2}}, \mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1)\right).$$

By the assumptions of the Theorem, we have that $\varepsilon\frac{\mathcal{O}(1)(C_x+C_u)}{1-\rho^{1/2}} \le 2$, allowing us to apply Lemma **??** to conclude that

$$(I+\widehat{\mathbf{\Delta}})^{-1} \in \mathcal{RH}_\infty\left(\mathcal{O}(1), \mathsf{Avg}(\mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1),1)\right).$$

Applying Lemma **??**, we see that

$$\mathbf{\Phi}_x(F)(I + z^{-F}V(F))^{-1}(I+\widehat{\mathbf{\Delta}})^{-1} \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)C_x}{1-\rho^{1/2}}, \mathsf{Avg}(\mathsf{Avg}(\mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1),1)1)\right)$$
$$\mathbf{\Phi}_u(F)(I + z^{-F}V(F))^{-1}(I+\widehat{\mathbf{\Delta}})^{-1} \in \mathcal{RH}_\infty\left(\frac{\mathcal{O}(1)C_u}{1-\rho^{1/2}}, \mathsf{Avg}(\mathsf{Avg}(\mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1),1)1)\right)$$

Finally, to simplify these bounds to those in the Theorem statement, notice first that for $\rho \ge .4$, we have that $(1-\rho^{1/2}) > (1-\rho)^2$. Then, we also have that

$$\mathsf{Avg}(\mathsf{Avg}(\mathsf{Avg}(\mathsf{Avg}(\rho^{1/2},1),1),1)1) = \frac{31}{32} + \frac{1}{32}\rho^{1/2} = \frac{31}{32} + \frac{1}{32}(\frac{1}{4}\rho_\star + \frac{3}{4})^{1/2}.$$

Finally, one can check that for $\rho_\star \ge .4$, we have that $(\frac{1}{4}\rho_\star + \frac{3}{4})^{1/2} \le .95 + .05\rho_\star$, leading to the bound

$$\frac{31}{32} + \frac{1}{32}(\frac{1}{4}\rho_\star + \frac{3}{4})^{1/2} \le \frac{31.95}{32} + \frac{.05}{32}\rho_\star \le .999 + .001\rho_\star.$$

We note that these constants are by no means optimized. $\square$

C Estimation

Recall that Algorithm 1 proceeds in epochs and that we denote by $x_t^{(i)}$ and $u_t^{(i)}$ the state and input at time t during epoch i, respectively. The i-th epoch has length T_i. Note that $x_{T_i}^{(i)}$, the last state of epoch i, is equal to $x_0^{(i+1)}$, the first state of epoch $i+1$.

At the end of each epoch our method estimates the parameters $(A_\star, B_\star)$ from the trajectory observed during that epoch, i.e.

$$(\widehat{A}, \widehat{B}) \in \arg\min_{A,B} \sum_{t=0}^{T_i-1} \frac{1}{2}\|x_{t+1}^{(i)} - Ax_t^{(i)} - Bu_t^{(i)}\|_2^2. \tag{C.1}$$

The goal of this section is to offer high probability confidence bounds on the estimation error of (**??**). For the rest of the section we suppress the dependence on the epoch index i because we prove a statistical rate for a fixed epoch.

Algorithm 1 generates the inputs u_t using a feedback controller $\mathbf{K}$ which stabilizes the true system $(A_\star, B_\star)$. Let $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\}$ denote the response of $\mathbf{K}$ on the true system $(A_\star, B_\star)$, and suppose that $\mathbf{\Phi}_x \in \frac{1}{z}\mathcal{RH}_\infty(C_x, \rho)$ and $\mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty(C_u, \rho)$. More precisely, if $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2 I_p)$ is the process noise at time t and $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\eta^2 I_p)$ is the input noise added at time t, then we can write

$$x_t = \Phi_x(t+1)x_0 + \sum_{k=0}^{t-1} \Phi_x(t-k)(B_\star \eta_k + w_k) \tag{C.2}$$

$$u_t = \eta_t + \Phi_u(t+1)x_0 + \sum_{k=0}^{t-1} \Phi_u(t-k)(B_\star \eta_k + w_k). \tag{C.3}$$

For the statistical analysis it is useful to consider the stochastic process $z_t = [x_t^\top, u_t^\top]^\top$. Also, we denote the filtration $\mathcal{F}_t = \sigma(x_0, \eta_0, w_0 \ldots, \eta_{t-1}, w_{t-1}, \eta_t)$. It is clear that the process $\{z_t\}_{t\geq 0}$ is $\{\mathcal{F}_t\}_{t\geq 0}$-adapted. Throughout this section we assume that $C_u, C_x \geq 1$ and denote $C_K^2 := nC_x^2 + pC_u^2$.

C.1 Estimation after one epoch

Throughout this section we assume that $\sigma_\eta \leq \sigma_w$. This condition is not needed for achieving the necessary statistical rate of estimation of (A, B), but it aids in simplifying several algebraic quantities.

Proposition C.1. *Let $x_0 \in \mathbb{R}^n$ be any initial state, let $\sigma_\eta \leq \sigma_w$, and assume that a trajectory $\{(x_t, u_t)\}_{t=0}^{T-1}$ is observed. Furthermore, suppose the inputs $u_t \in \mathbb{R}^p$ are generated by a feedback controller $\mathbf{K}$ which stabilizes and achieves a response $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\}$ on $(A_\star, B_\star)$ with $\mathbf{\Phi}_x \in \frac{1}{z}\mathcal{RH}_\infty(C_x, \rho)$ and $\mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty(C_u, \rho)$. Then, the error of the OLS estimator $(\widehat{A}, \widehat{B})$ from Eq.* **??** *satisfies with probability $1-\delta$ the guarantee*

$$\max\left\{\begin{smallmatrix}\|\widehat{A}-A_\star\|, \\ \|\widehat{B}-B_\star\|\end{smallmatrix}\right\} \lesssim \frac{\sigma_w C_u}{\sigma_\eta}\sqrt{\frac{(n+p)}{T}\log\left(1 + \frac{pC_u}{\delta} + \frac{\sigma_w}{\sigma_\eta}\frac{\rho C_u C_K}{\delta(1-\rho^2)}\left(1 + \|B_\star\| + \frac{\|x_0\|_2}{\sigma_w\sqrt{T}}\right)\right)},$$

as long as

$$T \gtrsim (n+p)\log\left(1 + \frac{pC_u^2}{\delta} + \frac{\sigma_w^2}{\sigma_\eta^2}\frac{\rho^2 C_u^2 C_K^2}{\delta(1-\rho^2)}\left(1 + \|B_\star\|^2 + \frac{\|x_0\|_2^2}{\sigma_w^2 T}\right)\right). \tag{C.4}$$

The proof of this result follows from a result by Simchowitz et al. [17] on the estimation of linear response time-series. We present that result in the context of our problem. Let $M_\star = [A_\star, B_\star]$, and recall that $z_t = [x_t^\top, y_t^\top]^\top$. Then, the OLS estimator (**??**) can be written in the form

$$\widehat{M} \in \arg\min_{M} \sum_{t=0}^{T-1} \frac{1}{2}\|x_{t+1} - Mz_t\|_2^2. \tag{C.5}$$

The process $\{z_t\}_{t\geq 0}$ is said to satisfy the (k,ν,β)-*block martingale small-ball* (BMSB) condition if for any $j\geq 0$ and $v\in\mathbb{R}^{n+p}$, one has that

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{P}\left(|\langle v, z_{j+i}\rangle|\geq \nu\right)\geq \beta \text{ almost surely.}$$

This condition is used for characterizing the size of the minimum eigenvalue of the covariance matrix $\sum_{t=0}^{T-1} z_t z_t^\top$. A larger ν guarantees a larger lower bound of the minimum eigenvalue. In the context of our problem the result by Simchowitz et al. [17] translates as follows.

Theorem C.2 (Simchowitz et al. [17])**.** *Fix $\epsilon,\delta\in(0,1)$. For every T, k, ν, and β such that $\{z_t\}_{t\geq 0}$ satisfies the (k,ν,β)-BMSB and*

$$\left\lfloor\frac{T}{k}\right\rfloor \gtrsim \frac{n+p}{\beta^2}\log\left(1+\frac{\sum_{t=0}^{T-1}\mathbf{Tr}(\mathbb{E}z_t z_t^\top)}{k\lfloor T/k\rfloor\beta^2\nu^2\delta}\right),$$

the estimate $\widehat{M}$ defined in Eq. **??** *satisfies the following statistical rate*

$$\mathbb{P}\left(\|\widehat{M}-M\|_2 > \frac{\mathcal{O}(1)\sigma_w}{\beta\nu}\sqrt{\frac{n+p}{k\lfloor T/k\rfloor}\log\left(1+\frac{\sum_{t=0}^{T-1}\mathbf{Tr}(\mathbb{E}z_t z_t^\top)}{k\lfloor T/k\rfloor\beta^2\nu^2\delta}\right)}\right)\leq\delta.$$

Therefore, in order to apply this result we need to find k, ν, and β such that $\{z_t\}_{t\geq 0}$ satisfies the (k,ν,β)-BMSB condition, and we also have to upper bound the trace of the covariance of z_t. The next two lemmas address these two issues.

Lemma C.3. *Let x_0 be any initial state in $\mathbb{R}^n$ and let $\{u_t\}_{t\geq 0}$ be the sequence of inputs generated according to (**??**), and assume $\sigma_\eta\leq\sigma_w$. Then, the process $z_t=[x_t^\top, u_t^\top]^\top$ satisfies the*

$$\left(1,\frac{\sigma_\eta}{2C_u},\frac{3}{20}\right) \textit{ BMSB condition.}$$

Proof. For all $t\geq 1$, denote

$$\begin{aligned}\xi_t &= u_t-\eta_t-\Phi_u(1)w_{t-1}\\ &= \Phi_u(t+1)x_0+\sum_{k=0}^{t-2}\Phi_u(t-k)(B_\star\eta_k+w_k)+\Phi_u(1)B_\star\eta_{t-1}.\end{aligned}$$

Therefore, we have

$$\begin{bmatrix}x_{t+1}\\ u_{t+1}\end{bmatrix}=\begin{bmatrix}A_\star x_t+B_\star u_t\\ \xi_{t+1}\end{bmatrix}+\begin{bmatrix}I_n & 0\\ \Phi_u(1) & I_p\end{bmatrix}\begin{bmatrix}w_t\\ \eta_{t+1}\end{bmatrix},$$

and hence

$$\begin{bmatrix}x_{t+1}\\ u_{t+1}\end{bmatrix}|\mathcal{F}_t\sim\mathcal{N}\left(\begin{bmatrix}A_\star x_t+B_\star u_t\\ \xi_{t+1}\end{bmatrix},\begin{bmatrix}\sigma_w^2 I_n & \sigma_w^2\Phi_u(1)^\top\\ \sigma_w^2\Phi_u(1) & \sigma_w^2\Phi_u(1)\Phi_u(1)^\top+\sigma_\eta^2 I_p\end{bmatrix}\right).$$

Denote by $\mu_{z,t}$ and Σ_z the mean and covariance of this multivariate normal distribution. Recall that we denote $z_t=[x_t^\top,u_t^\top]^\top$. Let $v\in\mathbb{R}^{n+p}$ and then $\langle v,z_t\rangle\sim\mathcal{N}(\langle v,\mu_{z,t}\rangle, v^\top\Sigma_z v)$. Therefore,

$$\begin{aligned}\mathbb{P}\left(|\langle v,z_t\rangle|\geq\sqrt{\lambda_{\min}(\Sigma_z)}\right)&\geq\mathbb{P}\left(|\langle v,z_t\rangle|\geq\sqrt{v^\top\Sigma_z v}\right)\\ &\geq\mathbb{P}\left(|\langle v,z_t-\mu_{z,t}\rangle|\geq\sqrt{v^\top\Sigma_z v}\right)\geq 3/10,\end{aligned}$$

where the last two inequalities follow because for any $\mu,\sigma^2\in\mathbb{R}$ and $\omega\sim\mathcal{N}(0,\sigma^2)$ we have

$$\mathbb{P}(|\mu+\omega|\geq\sigma)\geq\mathbb{P}(|\omega|\geq\sigma)\geq 3/10.$$

Since $\Phi_u\in\frac{1}{z}\mathcal{RH}_\infty(C_u,\rho)$ we have $\|\Phi_u(1)\|\leq C_u$. Then, by a simple argument based on a Schur complement (detailed in Lemma **??**) it follows that

$$\lambda_{\min}(\Sigma_z)\geq\sigma_\eta^2\min\left(\frac{1}{2},\frac{\sigma_w^2}{2\sigma_w^2C_u^2+\sigma_\eta^2}\right).$$

The conclusion follows since $C_u\geq 1$. □

Lemma C.4. *Let $\sigma_\eta \le \sigma_w$. Then, the process $z_t = [x_t^\top, u_t^\top]^\top$ satisfies*

$$\sum_{t=0}^{T-1} \mathbf{Tr}\left(\mathbb{E} z_t z_t^\top\right) \le \sigma_\eta^2 pT + \sigma_w^2 \frac{\rho^2 C_K^2 T}{(1-\rho^2)}\left(1 + \|B_\star\|^2 + \frac{\|x_0\|_2^2}{\sigma_w^2 T}\right).$$

Proof. Now, note that

$$\begin{aligned}\mathbb{E} z_t z_t^\top &= \begin{bmatrix}\Phi_x(t+1)\\ \Phi_u(t+1)\end{bmatrix} x_0 x_0^\top \begin{bmatrix}\Phi_x(t+1)\\ \Phi_u(t+1)\end{bmatrix}^\top + \begin{bmatrix}0 & 0\\ 0 & \sigma_\eta^2 I_p\end{bmatrix}\\ &\quad + \sum_{k=0}^{t-1}\begin{bmatrix}\Phi_x(t-k)\\ \Phi_u(t-k)\end{bmatrix}(\sigma_\eta^2 B_\star B_\star^\top + \sigma_w^2 I_n)\begin{bmatrix}\Phi_x(t-k)\\ \Phi_u(t-k)\end{bmatrix}.\end{aligned}$$

Since for all $j \ge 1$ we have $\|\Phi_x(j)\| \le C_x \rho^j$ and $\|\Phi_u(j)\| \le C_u \rho^j$, we obtain

$$\mathbf{Tr}\,\mathbb{E} z_t z_t^\top \le p\sigma_\eta^2 + (nC_x^2 + pC_u^2)\left(\rho^{2t+2}\|x_0\|_2^2 + (\sigma_w^2 + \sigma_\eta^2\|B_\star\|^2)\sum_{k=1}^{t}\rho^{2k}\right)$$

Therefore, we get that

$$\sum_{t=0}^{T-1}\mathbf{Tr}\,\mathbb{E} z_t z_t^\top \le p\sigma_\eta^2 T + \frac{\rho^2 T}{1-\rho^2}(nC_x^2 + pC_u^2)(\sigma_w^2 + \sigma_\eta^2\|B_\star\|^2) + \frac{\rho^2}{1-\rho^2}(nC_x^2 + pC_u^2)\|x_0\|_2^2,$$

and the conclusion follows by simple algebra. □

Proposition **??** follows from Theorem **??**, Lemma **??**, Lemma **??**, and simple algebra.

C.2 Stitching the epochs together

We start by bounding with high probability the size of the initial states of the epochs. Recall that epoch i has length T_i and that we denote by $x_{T_i}^{(i)}$ the last state of the epoch i, which is equal to the first state $x_0^{(i+1)}$ of the epoch $i+1$. For simplicity we assume that $x_0^{(0)} = 0$, an assumption that is not restrictive in any way.

Lemma C.5. *Fix $\delta \in (0,1)$, $r > 0$, and an epoch i. Assume that for all $k \le i$ the epoch length T_k is large enough so that $C_x \rho^{T_k} \le \rho^r$. Then, for any $t \ge 0$ we have*

$$\mathbb{P}\left(\|x_0^{(i+1)}\|_2 \ge \sigma_w(\sqrt{n} + t)\frac{C_x\rho(1+\|B_\star\|)}{(1-\rho^r)(1-\rho^2)}\right) \le \exp\left(-\frac{t^2}{2}\right).$$

Proof. From Eq. (**??**) we have that

$$x_0^{(i+1)} = \Phi_x^{(i)}(T_i+1)x_0^{(i)} + \underbrace{\sum_{j=0}^{T_i-1}\Phi_x^{(i)}(T_i-1-j)(B_\star\eta_j^{(i)} + w_j^{(i)})}_{\xi_i},$$

where we denoted the sum over disturbances during the epoch i by ξ_i. Therefore,

$$\begin{aligned}\|x_0^{(i+1)}\|_2 &\le C_x\rho^{T_i}\|x_0^{(i)}\|_2 + \|\xi_i\|_2\\ &\le \rho^r\|x_0^{(i)}\|_2 + \|\xi_i\|_2\\ &\le \sum_{k=0}^{i}\rho^{r(i-k)}\|\xi_k\|_2.\end{aligned}$$

By definition ξ_k is a zero-mean multivariate Gaussian random vector with covariance

$$\Sigma_{x,k} := \sum_{j=0}^{T_k-1}\Phi_x^{(k)}(T_k-1-j)(\sigma_w^2 + \sigma_{\eta,k}^2 B_\star B_\star^\top)\Phi_x^{(k)}(T_k-1-j)^\top,$$

whose top eigenvalue is upper bounded by

$$\sum_{j=0}^{T_k-2} C_x^2(\sigma_w^2 + \sigma_{\eta,k}^2\|B_\star\|^2)\rho^{2(T_k-1-j)} \le (\sigma_w^2 + \sigma_{\eta,k}^2\|B_\star\|^2)\frac{C_x^2\rho^2}{1-\rho^2}$$
$$\le \sigma_w^2(1+\|B_\star\|^2)\frac{C_x^2\rho^2}{1-\rho^2}, \tag{C.6}$$

where the last inequality follows because $\sigma_{\eta,k} \le \sigma_w$.

Then, we can write $\|\xi_k\|_2$ as $\|\Sigma_{x,k}^{1/2}\omega_k\|_2$, where ω_k is a standard Gaussian random vector distributed according to $\mathcal{N}(0, I_n)$, and hence $\|\xi_i\|_2$ is a Lipschitz function of ω_i with Lipschitz constant equal to squared root of (**??**). Hence, $\|x_0^{(i)}\|_2$ is a Lipschitz function of standard normal random variables with the Lipschitz constant

$$\sqrt{\sigma_w^2\frac{(1+\|B_\star\|^2)}{1-\rho^r}\frac{C_x^2\rho^2}{1-\rho^2}}.$$

By the concentration of Lipschitz functions of isotropic Gaussians, for $\nu \ge 0$, we have that

$$\mathbb{P}\left(\|x_0^{(i+1)}\|_2 \ge \mathbb{E}\|x_0^{(i+1)}\|_2 + \nu\right) \le \exp\left(-\frac{\nu^2(1-\rho^2)(1-\rho^r)}{2\sigma_w^2\rho^2(1+\|B_\star\|^2)C_x}\right).$$

By Jensen's inequality we have that

$$\mathbb{E}\|x_0^{(i+1)}\|_2 \le \sqrt{\mathbb{E}\|x_0^{(i+1)}\|_2^2} \le \sqrt{\sum_{k=0}^{i}\rho^{r(i-k)}\,\mathbf{Tr}\left(\mathbb{E}\xi_k\xi_k^\top\right)}$$
$$\le \sqrt{n\sigma_w^2\frac{(1+\|B_\star\|^2)}{1-\rho^r}\frac{C_x^2\rho^2}{1-\rho^2}}.$$

The conclusion follows. $\square$

We are now ready to prove that the statistical rate holds across epochs. In order to achieve this, we need the statistical rate after the first epoch to be small enough to satisfy the feasibility constraints on ε given in Theorem **??** for the IIR case and given in Theorem **??** for the FIR truncated case. Once this occurs, we immediately have feasibility at the next epoch (w.h.p.), and iterating the argument gives us recursive feasibility (w.h.p.).

Theorem C.6. *Fix a $\delta \in (0,1)$. For the IIR case, let C_x, C_u, ρ be defined as*

$$C_x = \frac{\mathcal{O}(1)C_\star}{(1-\rho_\star)^2},$$
$$C_u = \frac{\mathcal{O}(1)\|K_\star\|C_\star}{(1-\rho_\star)^2},$$
$$\rho = (1/8)\rho_\star + (7/8),$$

and for the FIR case, let C_x, C_u, ρ be defined as

$$C_x = \frac{\mathcal{O}(1)C_\star}{(1-\rho_\star)^3},$$
$$C_u = \frac{\mathcal{O}(1)\|K_\star\|C_\star}{(1-\rho_\star)^3},$$
$$\rho = 0.001\rho_\star + .999,$$

where $(C_\star, \rho_\star)$ are as defined in Theorem **??** *(resp. Theorem* **??***), and suppose (wlog) that $C_x \ge 1$ and $C_u \ge 1$. Let the length of epoch $i \in \{0,1,2,...\}$ be $T_i = C_T 2^i$ time steps and let the injected* *noise variance at epoch i be $\sigma_{\eta,i}^2 = \sigma_w^2 2^{-i/3}$. Suppose the constant C_T is large enough to satisfy the following inequalities,*

$$C_T \geq \frac{\log(2C_x)}{\log(1/\rho)} \,, \tag{C.7}$$

$$C_T \gtrsim \frac{1}{2^i}\left(n + \log\left(\frac{i+1}{\delta}\right)\right) \text{ for all } i = 0,1,2,\dots\,, \tag{C.8}$$

$$C_T \gtrsim \frac{(n+p)}{2^i}\log\left(1 + (i+1)^2\frac{pC_u^2}{\delta} + (i+1)^2 2^{i/3}\frac{\rho^2 C_u^2 C_K^2}{\delta(1-\rho^2)}\left(\frac{C_x^2(1+\|B_\star\|)^2}{(1-\rho)^2}\right)\right) \tag{C.9}$$
$$\text{for all } i = 0,1,2,\dots\,,$$

$$C_T \gtrsim \frac{(n+p)}{2^{2i/3}}\frac{C_u^2(C_x+C_u)^2}{(1-\rho_\star)^\alpha} \tag{C.10}$$
$$\times \log\left(1 + (i+1)\frac{pC_u}{\delta} + (i+1)2^{i/6}\frac{\rho C_u C_K}{\delta(1-\rho^2)}\left(\frac{C_x(1+\|B_\star\|)}{1-\rho}\right)\right)$$
$$\text{for all } i = 0,1,2,\dots\,,$$

where above $\alpha = 2$ for the IIR case and $\alpha = 4$ for the FIR case. Then, with probability $1-\delta$, the following two statements hold. First, for all epochs i, the norm of the first state at the beginning of each epoch satisfies

$$\|x_0^{(i)}\|_2 \lesssim \sigma_w\left(\sqrt{n} + \sqrt{\log\left(\frac{i+1}{\delta}\right)}\right)\frac{C_x\rho(1+\|B_\star\|)}{1-\rho^2} \,. \tag{C.11}$$

Second, for all epochs i, the OLS estimate $(\widehat{A}^{(i)}, \widehat{B}^{(i)})$ satisfies the statistical rate

$$\max\left\{\begin{matrix}\|\widehat{A}^{(i)}-A\|,\\ \|\widehat{B}^{(i)}-B\|\end{matrix}\right\} \lesssim \frac{\sigma_w C_u}{\sigma_{\eta,i}}\sqrt{\frac{(n+p)}{T_i}\log\left(1 + (i+1)\frac{pC_u}{\delta} + (i+1)\frac{\sigma_w}{\sigma_{\eta,i}}\frac{\rho C_u C_K}{\delta(1-\rho^2)}\left(\frac{C_x(1+\|B_\star\|)}{1-\rho}\right)\right)} \,. \tag{C.12}$$

Proof. For this proof, we set $r = \log(2)/\log(1/\rho)$.

By Theorem **??** for the IIR case and Theorem **??** for the FIR case, we know that the true responses $\{\mathbf{\Phi}_x, \mathbf{\Phi}_u\}$ of the synthesized controllers $\mathbf{K}_i$ on $(A_\star, B_\star)$ at every epoch satisfy $\mathbf{\Phi}_x \in \mathcal{RH}_\infty(C_x, \rho)$ and $\mathbf{\Phi}_u \in \mathcal{RH}_\infty(C_u, \rho)$.

Because of the assumption (**??**) on C_T we have $C_x\rho^{T_i} \leq \rho^r$. Therefore, we can apply Lemma **??** with $t^2 = \log(\mathcal{O}(1)(i+1)^2/\delta)$ to obtain that with probability at least $1-\delta/2$ the norm of $x_0^{(i)}$ for all epochs i satisfies

$$\|x_0^{(i)}\|_2 \lesssim \sigma_w\left(\sqrt{n} + \sqrt{\log\left(\frac{i+1}{\delta}\right)}\right)\frac{C_x\rho(1+\|B_\star\|)}{(1-\rho^r)(1-\rho^2)}.$$

Furthermore, by the assumption (**??**) on C_T we have that with probability at least $1-\delta/2$,

$$\frac{\|x_0^{(i)}\|_2^2}{\sigma_w^2 T_i} \leq \frac{C_x^2(1+\|B_\star\|)^2}{(1-\rho)^2} \,.$$

Our assumption (**??**) means that condition (**??**) is satisfied for each epoch i and therefore under the assumption the SLS program is feasible at every iteration, we can invoke Proposition **??** with $\delta = \mathcal{O}(1)\delta/(i+1)^2$ and reach the desired conclusions.

To show feasibility of the SLS at every epoch, Theorem **??** for the IIR case requires that

$$\varepsilon(i) \leq \mathcal{O}(1)\frac{1-\rho_\star}{C_x + C_u} \,,$$

and Theorem **??** for the FIR case requires that

$$\varepsilon(i) \le \mathcal{O}(1)\frac{(1-\rho_\star)^2}{C_x + C_u} \,,$$

where $\varepsilon(i)$ is our statistical upper bound on the errors $\max\left\{\substack{\|\widehat{A}^{(i)}-A\|, \\ \|\widehat{B}^{(i)}-B\|}\right\}$. This condition is ensured by our assumption (**??**) on C_T. ☐

We now remark on the satisfiability of the constraints on C_T given by (**??**), (**??**), and (**??**). For (**??**) and (**??**) (resp. (**??**)), the RHS grows like $\mathrm{poly}(i)/2^i$ (resp. $\mathrm{poly}(i)/2^{2i/3}$) and hence the supremum of the RHS (as a function of i) is achieved for some finite i. Therefore, we have that C_T satisfies in the IIR case

$$\begin{aligned} C_T &= \widetilde{\mathcal{O}}\left(\max\left\{\frac{1}{1-\rho_\star}, n, (n+p)\frac{C_\star^4(1+\|K_\star\|)^4}{(1-\rho_\star)^8}\right\}\right) \\ &= \widetilde{\mathcal{O}}\left((n+p)\frac{C_\star^4(1+\|K_\star\|)^4}{(1-\rho_\star)^8}\right) \,, \end{aligned} \tag{C.13}$$

and that C_T satisfies in the FIR case

$$\begin{aligned} C_T &= \widetilde{\mathcal{O}}\left(\max\left\{\frac{1}{1-\rho_\star}, n, (n+p)\frac{C_\star^4(1+\|K_\star\|)^4}{(1-\rho_\star)^{10}}\right\}\right) \\ &= \widetilde{\mathcal{O}}\left((n+p)\frac{C_\star^4(1+\|K_\star\|)^4}{(1-\rho_\star)^{16}}\right) \,. \end{aligned} \tag{C.14}$$

D Regret Decomposition and Analysis

We use the following regret decomposition, and for simplicity we assume that T is such that $T_0 + T_1 + ... + T_{E-1} = T$ for some E. Note that $E = O(\log_2 T)$.

$$\mathsf{Regret}(T) = \sum_{k=1}^{T}(x_k^\top Q x_k + u_k^\top R u_k - J_\star) = \sum_{i=0}^{E-1}\sum_{j=1}^{T_i}(x_{i,j}^\top Q x_{i,j} + u_{i,j}^\top R u_{i,j} - J_\star) \,. \tag{D.1}$$

Here, we let $x_{i,j}$ denote the j-th state at the i-th epoch (and similarly for $u_{i,j}$). Our definition of regret is defined for a given realization, as opposed to in expectation. However, in our analysis so far we have considered sub-optimality guarantees in expectation. Hence, our first concern is going from a realization to expectation.

Denote by $J_T(A, B, \mathbf{K}; \Sigma)$ the expected cost incurred by a (stabilizing) feedback policy $\mathbf{K}$ over a finite horizon T on system (A, B) being driven by process noise $w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ and starting from an initial condition of $x_0 = 0$, i.e.,

$$J_T(A, B, \mathbf{K}; \Sigma) := \sum_{k=1}^{T}\mathbb{E}\left[x_k^\top Q x_k + u_k^\top R u_k\right] \,. \tag{D.2}$$

Recall also that $J(A, B, \mathbf{K}; \Sigma)$ is the infinite-horizon LQR cost of $\mathbf{K}$ in feedback with (A, B). We now state some basic properties of J_T and J. We omit the proofs of these properties as they are standard.

Lemma D.1. *The following are true*

(i) $J_T(A, B, \mathbf{K}; \Sigma) \le T J(A, B, \mathbf{K}; \Sigma)$*,*

(ii) $J(A, B, \mathbf{K}; \Sigma_1 + \Sigma_2) = J(A, B, \mathbf{K}; \Sigma_1) + J(A, B, \mathbf{K}; \Sigma_2)$*,*

(iii) $J(A, B, \mathbf{K}; \alpha\Sigma) = \alpha J(A, B, \mathbf{K}; \Sigma)$ *for* $\alpha > 0$*,*

(iv) $J(A, B, \mathbf{K}; \Sigma_1) \le J(A, B, \mathbf{K}; \Sigma_2)$ *if* $\Sigma_1 \preceq \Sigma_2$*.*

From these properties, we immediately conclude that

$$J_T(A, B, \mathbf{K}; \sigma_w^2 I + \sigma_\eta^2 BB^\top) \le T \left(1 + \frac{\sigma_\eta^2 \|B\|^2}{\sigma_w^2}\right) J(A, B, \mathbf{K}; \sigma_w^2 I) \,, \tag{D.3}$$

a fact we will make use of later on.

The following lemma relates the finite horizon cost to its expectation.

Lemma D.2. *Let $\mathbf{K}$ be a feedback policy that stabilizes (A, B) and that induces system responses $\mathbf{\Phi}_x \in \mathcal{RH}_\infty(C_x, \rho)$ and $\mathbf{\Phi}_u \in \mathcal{RH}_\infty(C_u, \rho)$. Suppose that the system (A, B) is started at $x_0 = x$ and is driven by process noise $w \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$ and $\|\Sigma\| \le \sigma^2$. Then with probability at least $1 - \frac{1}{\delta}$ over the randomness of the process noise,*

$$\sum_{k=1}^{T} x_k^\top Q x_k + u_k^\top R u_k \le J_T(A, B, \mathbf{K}; \Sigma) + C_c \cdot \mathcal{O}\left(\|x\|_2^2 + \sigma^2 (\sqrt{nT \log\left(\tfrac{2}{\delta}\right)} + \log\left(\tfrac{2}{\delta}\right))\right) , \tag{D.4}$$

for $C_c := (1-\rho)^{-2}(\|Q\| C_x^2 + \|R\| C_u^2)$.

Proof. Writing $\mathbf{\Phi}_x$ as $\mathbf{\Phi}_x = \sum_{k=1}^\infty \Phi_x(k) z^{-k}$, we define the following finite-horizon truncations of its block-Toeplitz representation:

$$\mathbf{\Phi}_{x,T} := \begin{bmatrix} \Phi_x(1) & & \\ \vdots & \ddots & \\ \Phi_x(T) & \dots & \Phi_x(1) \end{bmatrix} \quad \mathbf{\Phi}_{x,+} := \begin{bmatrix} \Phi_x(2) \\ \Phi_x(3) \\ \vdots \\ \Phi_x(T+1) \end{bmatrix} .$$

We let $\mathbf{\Phi}_{u,T}$ and $\mathbf{\Phi}_{u,T,+}$ define similar matrices for $\mathbf{\Phi}_u$. Using these definitions, we can write

$$\sum_{k=1}^{T} x_k^\top Q x_k + u_k^\top R u_k = \begin{bmatrix} x \\ \omega \end{bmatrix}^\top \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\top & M_{22} \end{bmatrix} \begin{bmatrix} x \\ \omega \end{bmatrix} ,$$

for

$$\begin{aligned}
\omega^\top &= \begin{bmatrix} w_0^\top & w_1^\top & \dots & w_{T-1}^\top \end{bmatrix} \\
M_{11} &= \begin{bmatrix} \mathbf{\Phi}_{x,+} \\ \mathbf{\Phi}_{u,+} \end{bmatrix}^\top \begin{bmatrix} \mathcal{Q} & \\ & \mathcal{R} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{x,+} \\ \mathbf{\Phi}_{u,+} \end{bmatrix} \\
M_{12} &= \begin{bmatrix} \mathbf{\Phi}_{x,+} \\ \mathbf{\Phi}_{u,+} \end{bmatrix}^\top \begin{bmatrix} \mathcal{Q} & \\ & \mathcal{R} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{x,T} \\ \mathbf{\Phi}_{u,T} \end{bmatrix} \\
M_{22} &= \begin{bmatrix} \mathbf{\Phi}_{x,T} \\ \mathbf{\Phi}_{u,T} \end{bmatrix}^\top \begin{bmatrix} \mathcal{Q} & \\ & \mathcal{R} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{x,T} \\ \mathbf{\Phi}_{u,T} \end{bmatrix} ,
\end{aligned}$$

where $\mathcal{Q} := \mathrm{blkdiag}(Q)$ and $\mathcal{R} := \mathrm{blkdiag}(R)$ are block-diagonal matrices of compatible dimension. With these definitions, one can then check that $\mathbf{Tr}\, M_{22} \mathrm{blkdiag}(\Sigma) = J_T(A, B, \mathbf{K}; \Sigma)$.

Finally, given that $\mathbf{\Phi}_{x,+}, \mathbf{\Phi}_{x,T}$ are sub-matrices of the block-Toeplitz representation of $\mathbf{\Phi}_x$, it follows that $\max\{\|\mathbf{\Phi}_{x,+}\|, \|\mathbf{\Phi}_{x,T}\|\} \le \|\mathbf{\Phi}_x\|_{\mathcal{H}_\infty} \le \frac{C_x}{1-\rho}$, where the last inequality follows from Lemma **??**. Similarly, we have that $\max\{\|\mathbf{\Phi}_{u,+}\|, \|\mathbf{\Phi}_{u,T}\|\} \le \|\mathbf{\Phi}_u\|_{\mathcal{H}_\infty} \le \frac{C_u}{1-\rho}$. The result then follows by using these bounds, noting that $\omega \sim \mathcal{N}(0, \mathrm{blkdiag}(\Sigma))$, and applying Lemma **??** with the inequality $\|M\|_F \le \sqrt{\mathrm{rank}(M)} \|M\| \le \sqrt{\max(n_1, n_2)} \|M\|$ for an $n_1 \times n_2$ matrix M. □

We now proceed to prove our main regret upper bounds, for both the IIR and FIR case.

Let $\mathcal{E}_{\mathrm{est},i}$ denote the event that the conclusions of Theorem **??** hold up to and including epoch i. Let $\{\widehat{\mathbf{\Phi}}_{i,x}\}_{i \ge 0}$ and $\{\widehat{\mathbf{\Phi}}_{i,u}\}_{i \ge 0}$ denote the closed loop SLS responses on the true system $(A_\star, B_\star)$. When

$\mathcal{E}_{\mathrm{est},i}$ holds, Theorem **??** in the IIR case and Theorem **??** in the FIR case state that uniformly for all epochs i we have

$$\widehat{\mathbf{\Phi}}_{i,x} \in \mathcal{RH}_\infty(\widehat{C}, \widehat{\rho}) \ , \ \ \widehat{\mathbf{\Phi}}_{i,u} \in \mathcal{RH}_\infty(\|K_\star\|\widehat{C}, \widehat{\rho}) \ ,$$

for

$$\begin{aligned} \widehat{C} &= \frac{\mathcal{O}(1) C_\star}{(1-\rho_\star)^2} \ , \\ \widehat{\rho} &= 7/8 + (1/8)\rho_\star \ , \end{aligned}$$

in the IIR case and

$$\begin{aligned} \widehat{C} &= \frac{\mathcal{O}(1) C_\star}{(1-\rho_\star)^3} \ , \\ \widehat{\rho} &= 0.999 + 0.001\rho_\star \ , \end{aligned}$$

in the FIR case. For ease of notation, define $\widehat{C}_c^2 := \frac{(\|Q\|+\|R\|\|K_\star\|)\widehat{C}^2}{(1-\widehat{\rho})^2}$.

Now fix an epoch $i \geq 1$ (the epoch $i = 0$ will be dealt with separately) and let $\mathbf{K}_i$ denote the controller that is active during epoch i. We invoke Lemma **??** conditioned on $\mathcal{E}_{\mathrm{est},i}$ and $x_{i,0}$ with $\delta \leftarrow \mathcal{O}(1)\delta/(i+1)^2$, $\Sigma \leftarrow \sigma_w^2 I + \sigma_{\eta,i}^2 B_\star B_\star^\top$, $C_x \leftarrow \widehat{C}$, $C_u \leftarrow \|K_\star\|\widehat{C}$, and $\rho \leftarrow \widehat{\rho}$. The conclusion is that with (conditional) probability at least $1 - \mathcal{O}(1)\delta/(i+1)^2$,

$$\begin{aligned} &\sum_{k=1}^{T_i} x_{i,k}^\top Q x_{i,k} + u_{i,k}^\top R u_{i,k} \\ &\quad \leq J_T(A_\star, B_\star, \mathbf{K}_i; \sigma_w^2 I + \sigma_{\eta,i}^2 B_\star B_\star^\top) \\ &\qquad + \widehat{C}_c^2 \mathcal{O}\left(\|x_{i,0}\|_2^2 + (\sigma_w^2 + \sigma_{\eta,i}^2 \|B_\star\|^2)(\sqrt{nT_i \log((i+1)/\delta)} + \log((i+1)/\delta)) \right) \\ &\quad \leq T_i \left(1 + \frac{\sigma_{\eta,i}^2 \|B_\star\|^2}{\sigma_w^2} \right) J(A_\star, B_\star, \mathbf{K}_i; \sigma_w^2 I) \\ &\qquad + \widehat{C}_c^2 \mathcal{O}\Bigg(\sigma_w^2 (n + \log((i+1)/\delta)) \frac{\widehat{C}^2 \widehat{\rho}^2 (1 + \|B_\star\|)^2}{(1-\widehat{\rho})^2} \\ &\qquad\quad + (\sigma_w^2 + \sigma_{\eta,i}^2 \|B_\star\|^2)(\sqrt{nT_i \log((i+1)/\delta)} + \log((i+1)/\delta)) \Bigg) \ . \end{aligned}$$

For the second inequality, we used the bound (**??**) and the bound on $\|x_{i,0}\|_2$ from (**??**).

Furthermore, (**??**) and Theorem **??** in the IIR case (Theorem **??** in the FIR case) tell us that on $\mathcal{E}_{\mathrm{est},i}$, we have the sub-optimality bound

$$\begin{aligned} J(A_\star, B_\star, \mathbf{K}_i; \sigma_w^2 I) &\leq (1 + C_{J_{i-1}})^2 (1 + \mathcal{O}(1)\varepsilon_{i-1}(1 + \|K_\star\|)\|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty})^2 J_\star \ , \\ \varepsilon_i &= \widetilde{\mathcal{O}}\left(\frac{\sigma_w \|K_\star\| \widehat{C}}{\sigma_{\eta,i}} \sqrt{\frac{n+p}{T_i}} \right) \ . \end{aligned}$$

Above, in the IIR case, we set $C_{J_i} = 0$ for all i, and in the FIR case we choose $C_{J_i} = 1/2^{i+1}$. Since $C_{J_i} \leq 1$, we have that $(1 + C_{J_i})^2 \leq 1 + 3C_{j_i}$. Recalling that $\sigma_{\eta,i}/\sigma_w = 2^{-i/6}$ and that $T_i = C_T 2^i$, we simplify $\varepsilon_i = \widetilde{\mathcal{O}}\left(\|K_\star\|\widehat{C}\sqrt{\frac{n+p}{C_T}} 2^{-i/3} \right) := \widetilde{\mathcal{O}}(\frac{D_1}{\sqrt{C_T}} 2^{-i/3})$ which gives us

$$\begin{aligned} &(1 + \mathcal{O}(1)\varepsilon_{i-1}(1 + \|K_\star\|)\|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty})^2 \\ &\quad = 1 + \widetilde{\mathcal{O}}\left(\frac{D_1}{\sqrt{C_T}} (1 + \|K_\star\|)\|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty} 2^{-i/3} + \frac{D_1^2}{C_T} (1 + \|K_\star\|)^2 \|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty}^2 2^{-2i/3} \right) \\ &\quad := 1 + \widetilde{\mathcal{O}}\left(\frac{D_2}{\sqrt{C_T}} 2^{-i/3} + \frac{D_2^2}{C_T} 2^{-2i/3} \right) \ . \end{aligned}$$

This means that

$$
\begin{aligned}
&T_i \left(1 + \frac{\sigma_{\eta,i}^2 \|B_\star\|^2}{\sigma_w^2}\right) J(A_\star, B_\star, \mathbf{K}_i; \sigma_w^2 I) \\
&\quad \le T_i \left(1 + 2^{-i/3}\|B_\star\|^2\right)(1 + 3C_{J_{i-1}}) \left(1 + \widetilde{\mathcal{O}}\left(\frac{D_2}{\sqrt{C_T}} 2^{-i/3} + \frac{D_2^2}{C_T} 2^{-2i/3}\right)\right) J_\star \\
&\quad \le T_i \left(1 + \widetilde{\mathcal{O}}\left(\left(\frac{D_2}{\sqrt{C_T}} + \|B_\star\|^2\right) 2^{-i/3} + \left(\frac{D_2^2}{C_T} + \frac{D_2 \|B_\star\|^2}{\sqrt{C_T}}\right) 2^{-2i/3} + \frac{D_2^2 \|B_\star\|^2}{C_T} 2^{-i}\right)\right) J_\star \\
&\qquad + \widetilde{\mathcal{O}}((1 + \|B_\star\|^2)(C_T + D_2\sqrt{C_T} + D_2^2) J_\star) \\
&\quad = T_i J_\star + \widetilde{\mathcal{O}}(\sqrt{C_T} D_2 + C_T \|B_\star\|^2) J_\star 2^{2i/3} + \widetilde{\mathcal{O}}(D_2^2 + \sqrt{C_T} D_2 \|B_\star\|^2) J_\star 2^{i/3} \\
&\qquad + \widetilde{\mathcal{O}}((1 + \|B_\star\|^2)(C_T + D_2\sqrt{C_T} + D_2^2) J_\star) \,.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\sum_{k=1}^{T_i} (x_{i,k}^\top Q x_{i,k} + u_{i,k}^\top R u_{i,k} - J_\star) \\
&\quad \le \widetilde{\mathcal{O}}(\sqrt{C_T} D_2 + C_T \|B_\star\|^2) J_\star 2^{2i/3} + \widetilde{\mathcal{O}}(D_2^2 + \sqrt{C_T} D_2 \|B_\star\|^2) J_\star 2^{i/3} \\
&\qquad + \widetilde{\mathcal{O}}\left(\frac{\widehat{C}_c^2 \sigma_w^2 n \widehat{C}^2 (1 + \|B_\star\|)^2}{(1 - \widehat{\rho})^2}\right) + \widetilde{\mathcal{O}}(\widehat{C}_c^2 \sigma_w^2 \sqrt{n C_T} 2^{i/2}) + \widetilde{\mathcal{O}}(\widehat{C}_c^2 \sigma_w^2 \|B_\star\|^2 \sqrt{n C_T} 2^{i/6}) \\
&\qquad + \mathcal{O}(C_T 2^{i/2} (1 + \|B_\star\|^2)) + \widetilde{\mathcal{O}}((1 + \|B_\star\|^2)(C_T + D_2\sqrt{C_T} + D_2^2) J_\star) \,.
\end{aligned}
$$

On the other hand, when epoch $i = 0$, we have that

$$
\begin{aligned}
\sum_{k=1}^{T} x_{0,k}^\top Q x_{0,k} + u_{0,k}^\top R u_{0,k} &\le J_T(A_\star, B_\star, \mathbf{K}_0, \sigma_w^2 I + \sigma_{\eta,0}^2 B_\star B_\star^\top) + \widetilde{\mathcal{O}}(\widehat{C}_c^2 \sigma_w^2 (1 + \|B_\star\|^2)\sqrt{n C_T}) \\
&\le C_T (1 + \|B_\star\|^2) J(A_\star, B_\star, \mathbf{K}_0, \sigma_w^2 I) + \widetilde{\mathcal{O}}(\widehat{C}_c^2 \sigma_w^2 (1 + \|B_\star\|^2)\sqrt{n C_T}) \,.
\end{aligned}
$$

Summing over all the epochs,

$$
\begin{aligned}
\mathsf{Regret}(T) &= \sum_{i=0}^{O(\log_2 T)} \sum_{k=1}^{T_i} (x_{i,k}^\top Q x_{i,k} + u_{i,k}^\top R u_{i,k} - J_\star) \\
&\le \widetilde{\mathcal{O}}((\sqrt{C_T} D_2 + C_T \|B_\star\|^2) J_\star T^{2/3}) + \widetilde{\mathcal{O}}(\widehat{C}_c^2 \sigma_w^2 \sqrt{n C_T} T^{1/2}) \\
&\quad + \widetilde{\mathcal{O}}(D_2^2 + \sqrt{C_T} D_2 \|B_\star\|^2 J_\star T^{1/3}) + \widetilde{\mathcal{O}}(\widehat{C}_c^2 \sigma_w^2 \|B_\star\|^2 \sqrt{n C_T} T^{1/6}) \\
&\quad + \widetilde{\mathcal{O}}\left(\frac{\widehat{C}_c^2 \sigma_w^2 n \widehat{C}^2 (1 + \|B_\star\|)^2}{(1 - \widehat{\rho})^2} + C_T (1 + \|B_\star\|^2) J(A_\star, B_\star, \mathbf{K}_0, \sigma_w^2 I)\right) \\
&\quad + \widetilde{\mathcal{O}}((1 + \|B_\star\|^2)(C_T + D_2\sqrt{C_T} + D_2^2) J_\star) \\
&\quad + \widetilde{\mathcal{O}}(\widehat{C}_c^2 \sigma_w^2 (1 + \|B_\star\|^2)\sqrt{n C_T}) + \mathcal{O}(C_T (1 + \|B_\star\|^2)\sqrt{T}) \,.
\end{aligned}
$$

Using the bound on C_T from (**??**), recalling that

$$
D_2 = \sqrt{n + p} \|K_\star\| \widehat{C} (1 + \|K_\star\|) \|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty} \,,
$$

and ignoring the $o(T^{2/3})$ terms in the regret bound, we have that the regret is bounded by in the IIR case

$$
\widetilde{\mathcal{O}}\left((n + p) \|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty} \frac{C_\star^3 (1 + \|K_\star\|)^4}{(1 - \rho_\star)^6} J_\star T^{2/3} + (n + p) \frac{C_\star^4 (1 + \|K_\star\|)^4 \|B_\star\|^2}{(1 - \rho_\star)^8} J_\star T^{2/3}\right) \,.
$$

By using Lemma **??**, we have that $\|\mathfrak{R}_{A_\star + B_\star K_\star}\|_{\mathcal{H}_\infty} \le \frac{C_\star}{1 - \rho_\star}$, and hence the bound in the IIR case simplifies to

$$
\widetilde{\mathcal{O}}\left((n + p) \frac{C_\star^4 (1 + \|K_\star\|)^4 (1 + \|B_\star\|)^2 J_\star}{(1 - \rho_\star)^8} T^{2/3}\right) \,.
$$

Now for the FIR case, we use the bound (**??**) and ignoring the $o(T^{2/3})$ terms, the regret is bounded by

$$\widetilde{\mathcal{O}}\left((n+p)\|\mathfrak{R}_{A_\star+B_\star K_\star}\|_{\mathcal{H}_\infty}\frac{C_\star^3(1+\|K_\star\|)^4}{(1-\rho_\star)^{11}}J_\star T^{2/3}+(n+p)\frac{C_\star^4(1+\|K_\star\|)^4\|B_\star\|^2}{(1-\rho_\star)^{16}}J_\star T^{2/3}\right).$$

Using the same bound on $\|\mathfrak{R}_{A_\star+B_\star K_\star}\|_{\mathcal{H}_\infty}$ as before, we obtain the FIR regret bound

$$\widetilde{\mathcal{O}}\left((n+p)\frac{C_\star^4(1+\|K_\star\|)^4(1+\|B_\star\|)^2}{(1-\rho_\star)^{16}}J_\star T^{2/3}\right).$$

E Lower bound

This section is dedicated to proving Theorem 3.4. Throughout this section we assume the following setup and notation. We consider the LQR problem defined by

$$\min_{u_0,u_1,\ldots,u_{T-1}}\mathbb{E}\left[x_T^\top P x_T+\sum_{t=0}^{T-1}u_t^\top R u_t+x_t^\top Q x_t\right],$$
$$\text{s.t. } x_{t+1}=A_\star x_t+B_\star u_t+w_t.$$

where u_t is allowed to be any random variable taking values in $\mathbb{R}^p$ that is independent of the sigma algebra $\sigma(w_t,w_{t+1},\ldots)$. In particular, u_t can be a measurable function of x_0, w_0, w_1, …, w_{t-1}, and possibly other exogenous randomness.

We assume that Q and R are both positive definite matrices. Throughout this section we denote by P the solution to the discrete algebraic Riccati equation:

$$P_\star=A^\top P_\star A-A^\top P_\star B(R+B^\top P_\star B)^{-1}B^\top P_\star A+Q.$$

Moreover, we denote by $K_\star$ the optimal controller for the infinite horizon LQR problem, namely $K_\star=-(R+B^\top P_\star B)^{-1}B^\top P_\star A$. Hence, the optimal closed loop matrix is given by $M=A_\star+B_\star K_\star$. Throughout this section we assume that the system (A,B) is controllable and hence $\rho(M)<1$. Therefore, there exist $C>0$ and $\rho\in(0,1)$ such that $\|M^k\|_2\le C\rho^k$ for all $k\ge 1$.

The initial state x_0 for the LQR problem defined above is assumed to have distribution $\mathcal{N}(0,P_\infty)$, where P_∞ is the unique solution to the Lyapunov equation

$$P_\infty=(A_\star+B_\star K_\star)P_\infty(A_\star+B_\star K_\star)^\top+\sigma_w^2 I_n.$$

The distribution $\mathcal{N}(0,P_\infty)$ corresponds to the stationary distribution of the optimal closed loop system $x_{t+1}=(A_\star+B_\star K_\star)x_t+w_t$. In particular, if $x_t\sim\mathcal{N}(0,P_\infty)$, then $x_{t+1}\sim\mathcal{N}(0,P_\infty)$.

We consider the objective

$$J_T(\nu_0,\nu_1,\ldots,\nu_{T-1})=\mathbb{E}\left[x_T^\top P_\star x_T+\sum_{t=0}^{T-1}u_t^\top R u_t+x_t^\top Q x_t\right], \tag{E.1}$$

where $u_t=K_\star x_t+\nu_t$ for the optimal controller $K_\star$. Then, since the terminal cost is given by $P_\star$, we know that the minimum of objective (**??**) over ν_0, ν_1, …, ν_{T-1} such that ν_t is independent of $\sigma(w_t,w_{t+1},\ldots)$ is achieved when all ν_t are identically zero. The random variables ν_t should be thought of as deviations from the optimal inputs $K_\star x_t$ for the infinite horizon LQR. Finally, since $x_0\sim\mathcal{N}(0,P_\infty)$ we have that the optimal objective value is $J_T^\star=J_T(0)=TJ_\star+\mathbf{Tr}(P_\star P_\infty)$, where $J_\star=\sigma_w^2\,\mathbf{Tr}(P_\star)$ is the optimal objective value of the infinite horizon LQR.

The proof of Theorem 3.4 follows an argument inspired from the field of strongly convex optimization. We show that under the minimum eigenvalue condition of the process $z_t=[x_t^\top,u_t^\top]^\top$, the process $\{\nu_t\}_{t\ge0}$ is bounded away from zero. Moreover, we show that the expected regret at time T is a strongly convex function of $\nu_0,\nu_1,\ldots,\nu_{T-1}$, leading us to the desired conclusion. We proceed by proving a sequence of technical result, followed by the proof of Theorem 3.4.

Lemma E.1. . *Suppose that*

$$\lambda_{\min}\left(\sum_{t=0}^{T-1}\begin{bmatrix}x_t\\u_t\end{bmatrix}\begin{bmatrix}x_t^\top & u_t^\top\end{bmatrix}\right)\ge\tau, \tag{E.2}$$

with $u_t = K_\star x_t + \nu_t$. Then

$$\sum_{t=0}^{T-1} \|\nu_t\|_2^2 \geq \left(1 + \sigma_{\min}(K_\star)^2\right) \tau \tag{E.3}$$

Proof. Consider $v = [v_1^\top, v_2^\top]^\top \in \mathbb{R}^{n+p}$ such that $\|v\|_2 = 1$ and $v_1 + K^\top v_2 = 0$ (such v exists because $[I, K^\top]$ is an $n \times (n+p)$ matrix and hence has a non-trivial null space). Moreover, $\|v_2\|_2^2 \leq (1 + \sigma_{\min}(K_\star)^2)^{-1}$. Then, by assumption we have

$$\begin{aligned} \tau &\leq \sum_{t=0}^{T-1} (\langle x_t, v_1\rangle + \langle u_t, v_2\rangle)^2 = \sum_{t=0}^{T-1} (\langle x_t, v_1\rangle + \langle K x_t + \nu_t, v_2\rangle)^2 = \sum_{t=0}^{T-1} \langle \nu_t, v_2\rangle^2 \\ &\leq \|v_2\|_2^2 \sum_{t=0}^{T-1} \|\nu_t\|_2^2 \leq \frac{1}{1 + \sigma_{\min}(K)^2} \sum_{t=0}^{T-1} \|\nu_t\|_2^2. \end{aligned}$$

□

Lemma E.2. *Denote by M the optimal closed loop matrix $A_\star + B_\star K_\star$. Then*

$$\begin{aligned} J_T(\nu_0, \nu_1, \ldots, \nu_{T-1}) - J_T^\star = \mathbb{E}\left[\sum_{j=0}^{T-1} \nu_j^\top (B_\star^\top P_\star B_\star + R)\nu_j\right] \\ + 2\mathbb{E}\left[\sum_{0 \leq i < j \leq T-1} \nu_i^\top B_\star^\top (M^\top)^{j-i} P_\star B_\star \nu_j\right]. \end{aligned}$$

Proof. We know that

$$J_T^* = \mathbb{E}\left[\sum_{t=0}^{T-1} x_{\star,t}^\top (Q + K_\star^\top R K_\star) x_{\star,t}\right] + \mathbb{E}\left[x_{\star,T}^\top P_\star x_{\star,T}\right],$$

where $x_{\star,t} = \sum_{j=-1}^{t-1} M^{t-1-j} w_j$. Here, $w_{-1} = x_0$ for convenience, and $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2 I_n)$ for convenience. Also,

$$J_T(\nu_0, \nu_1, \ldots, \nu_{T-1}) = \mathbb{E}\left[\sum_{t=0}^{T-1} x_t^\top Q x_t + (K_\star x_t + \nu_t)^\top R (K_\star x_t + \nu_t)\right] + \mathbb{E}\left[x_T^\top P_\star x_T\right],$$

where $x_t = \sum_{j=-1}^{t-1} M^{t-1-j} w_j + M^{t-1-j} B \nu_j$ and $\nu_{-1} = 0$. Recall that ν_t is independent of any w_i with $i \geq t$. Hence, for any matrix N we have that $\mathbb{E}\left[w_i^\top N \nu_t\right] = 0$ if $i \geq t$. Therefore

$$\begin{aligned} J_T - J_T^\star = \mathbb{E}&\left[\sum_{t=0}^{T-1} \sum_{0 \leq i < j \leq t-1} 2 w_i^\top (M^\top)^{t-1-i} (Q + K_\star^\top R K_\star) M^{t-1-j} B_\star \nu_j\right] \\ &+ \mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{i,j=0}^{t-1} \nu_i^\top B_\star^\top (M^\top)^{t-1-i} (Q + K_\star^\top R K_\star) M^{t-1-j} B_\star \nu_j\right] + \mathbb{E}\left[\sum_{t=0}^{T-1} \nu_t^\top R \nu_t\right] \\ &+ \mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{i=0}^{t-1} 2 w_i^\top (M^\top)^{t-1-i} K_\star^\top R \nu_t\right] \\ &+ \mathbb{E}\left[\sum_{0 \leq i < j \leq T-1} 2 w_i^\top (M^\top)^{T-1-i} P_\star M^{T-1-j} B_\star \nu_j\right] \\ &+ \mathbb{E}\left[\sum_{i,j=0}^{T-1} \nu_i^\top B_\star^\top (M^\top)^{T-1-i} P_\star M^{T-1-j} B_\star \nu_j\right]. \end{aligned}$$

Now, we note that the sum of the terms that depend linearly on ν_t is equal to zero, otherwise the optimum of J_T would not be achieved at $\nu_t = 0$ for all t. Indeed, this can be checked through direct computation by remarking that the optimal controller $K_\star$ satisfies $K_\star^\top R = -M^\top P_\star B_\star$, and recalling that $P_\star$ satisfies the Lyapunov equation

$$P_\star = M^\top P_\star M + Q + K_\star^\top R K_\star. \tag{E.4}$$

Hence, we have

$$\begin{aligned}
J_T - J_T^\star = \mathbb{E}&\left[\sum_{j=0}^{T-2} \nu_j^\top B_\star^\top \left(\sum_{t=j+1}^{T-1} (M^\top)^{t-1-j}(Q + K^\top RK)M^{t-1-j}\right) B_\star \nu_j\right] \\
&+ 2\mathbb{E}\left[\sum_{0\le i<j\le T-2} \nu_i^\top B_\star^\top (M^\top)^{j-i} \left(\sum_{t=j+1}^{T-1} (M^\top)^{t-1-j}(Q + K_\star^\top RK_\star)M^{t-1-j}\right) B_\star \nu_j\right] \\
&+ \mathbb{E}\left[\sum_{t=0}^{T-1} \nu_t^\top R \nu_t\right] + \mathbb{E}\left[\sum_{j=0}^{T-1} \nu_j^\top B_\star^\top (M^\top)^{T-1-j} P_\star M^{T-1-j} B_\star \nu_j\right] \\
&+ 2\mathbb{E}\left[\sum_{0\le i<j\le T-1} \nu_i^\top B_\star^\top (M^\top)^{T-1-i} P_\star M^{T-1-j} B_\star \nu_j\right].
\end{aligned}$$

The conclusion follows by using the Lyapunov equation (**??**) and simple algebra. ☐

Lemma E.3. *Let M and N be any matrices in $\mathbb{R}^{n\times n}$, with N positive definite, and let T be any positive integer. Also, consider the $(nT)\times(nT)$ block matrix $D(T)$ with blocks $D(T)_{i,j}$ equal to*

$$D_{i,j} = \begin{cases} (M^\top)^{j-i}\left(\sum_{k=0}^{T-j}(M^\top)^k N M^k\right) & \text{if } i<j, \\ \sum_{k=0}^{T-j}(M^\top)^k N M^k & \text{if } i=j, \\ \left(\sum_{k=0}^{T-i}(M^\top)^k N M^k\right) M^{i-j} & \text{if } i<j, \end{cases}$$

where $1 \le i, j \le T$. The matrix D is positive definite.

Proof. We proceed by induction. Let $T = 2$. Then the matrix of interest is

$$D(2) = \begin{bmatrix} N + M^\top N M & M^\top N \\ NM & N \end{bmatrix}.$$

Since $N \succ 0$, we see that $D(T)$ is positive definite because its Schur complement is

$$N + M^\top NM - M^\top N N^{-1} NM = N \succ 0.$$

For $T > 2$ we proceed similarly. We consider the matrix $D(T)$ and take its Schur complement with respect to bottom right corner, i.e.

$$\begin{bmatrix} D(T)_{1,1} & \dots & D(T)_{1,T-1} \\ \vdots & \ddots & \vdots \\ D(T)_{T-1,1} & \dots & D(T)_{T-1,T-1} \end{bmatrix} - \begin{bmatrix} D(T)_{1,T} \\ \vdots \\ D(T)_{T-1,T} \end{bmatrix} D(T)_{T,T}^{-1} \begin{bmatrix} D_{T,1} & \dots & D_{T,T-1} \end{bmatrix}$$

Let $i \le j < T$. Then, the (i, j) block of the Schur complement of $D(T)$ is

$$\begin{aligned}
D(T)_{i,j} - D(T)_{i,T} D(T)^{-1} D(T)_{T,j} &= (M^\top)^{j-i}\left(\sum_{k=0}^{T-j}(M^\top)^k N M^k\right) - (M^\top)^{T-i} N N^{-1} N M^{T-j} \\
&= (M^\top)^{j-i}\left(\sum_{k=0}^{T-1-j}(M^\top)^k N M^k\right) = D(T-1)_{i,j}.
\end{aligned}$$

Similarly, if $j \le i < T$ we have that $D(T)_{i,j} - D(T)_{i,T}D(T)^{-1}D(T)_{T,j} = D(T-1)_{i,j}$. Hence, we have shown that the Schur complement of $D(T)$ with respect to the entry $D(T)_{T,T}$ is $D(T-1)$. By induction this matrix is positive definite and the conclusion follows. ☐

Lemma E.4. *As before, $P_\star$ is the solution to the algebraic Riccati equation and $M = A_\star + B_\star K_\star$ is the optimal closed loop matrix. For any vectors v_0, v_1, ..., v_{T-1} in $\mathbb{R}^p$ we have*

$$\sum_{j=0}^{T-1} v_j^\top (B_\star^\top P_\star B_\star + R) v_j + 2 \sum_{0 \le i < j \le T-1} v_i^\top B^\top (M^\top)^{j-i} P_\star B_\star v_j \ge \lambda_{\min}(R) \sum_{j=0}^{T-1} \|v_j\|_2^2.$$

Proof. It suffices to prove that the following matrix is positive semi-definite:

$$\begin{bmatrix} P_\star & M^\top P_\star & (M^\top)^2 P_\star & \dots & (M^\top)^{T-1} P_\star \\ P_\star M & P_\star & M^\top P_\star & \dots & (M^\top)^{T-2} P_\star \\ P_\star M^2 & P_\star M & P_\star & \dots & (M^\top)^{T-2} P_\star \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P_\star M^{T-1} & P_\star M^{T-2} & P_\star M^{T-3} & \dots & P_\star \end{bmatrix}.$$

The Schur complement of this matrix around the bottom right corner $P_\star$ has the form $D(T-1)$ with $N = Q + K_\star^\top R K_\star$, where $D(T-1)$ is defined as in Lemma **??**. To see this recall that $P_\star$ satisfies the Lyapunov equation (**??**). The conclusion follows. □

Lemma E.5. *Fix a horizon $T_0 > 0$, and suppose the inputs are of the form $u_t = K_\star x_t + \nu_t$. Recall that there exists constants $C > 0$ and $\rho \in (0, 1)$ such that $\|M^k\|_2 \le C\rho^k$ for all $k \ge 1$. Then*

$$\mathbb{E}\|x_{T_0}\|_2^2 \le 3C^2 \rho^{2T_0} \mathbb{E}\|x_0\|_2^2 + 3\frac{n\sigma_w^2 C^2}{1-\rho^2} + 3\frac{C^2}{1-\rho^2}\mathbb{E}\left[\sum_{t=0}^{T_0} \|\nu_t\|_2^2\right].$$

Proof. Recall that we denote by M the closed loop matrix $A_\star + B_\star K_\star$. We have that

$$x_{T_0} = M^{T_0} x_0 + \sum_{t=0}^{T_0-1} M^{T_0-1-t}(B_\star \nu_t + w_t).$$

Then

$$\|x_{T_0}\|_2^2 \le 3\|M^{T_0} x_0\|_2^2 + 3\|\sum_{t=0}^{T_0-1} M^{T_0-1-t} w_t\|_2^2 + 3\|\sum_{t=0}^{T_0-1} M^{T_0-1-t} B \nu_t\|_2^2.$$

Recall that $\|M^t\|_2 \le C\rho^t$. Then

$$\mathbb{E}\|x_{T_0}\|_2^2 \le 3C^2 \rho^{2T_0} \mathbb{E}\|x_0\|_2^2 + 3\frac{n\sigma_w^2 C^2}{1-\rho^2} + 3\frac{C^2}{1-\rho^2}\mathbb{E}\left[\sum_{t=0}^{T_0} \|\nu_t\|_2^2\right].$$

□

Lemma E.6. *Let Q and R be positive definite matrices, and $P_0 = 0$. Consider the Riccati recursion*

$$P_{t+1} = A^\top P_t A - A^\top P_t B (R + B^\top P_t B)^{-1} B^\top P_t A + Q.$$

Then, if $P_\star$ is the unique solution of the Riccati equation, we have

$$\|P_t - P_\star\|_2 \le \left(1 + \frac{1}{\nu}\right)^{-t}, \text{ where } \nu = 2\|P_\star\|_2 \max\left\{\frac{\|A_\star\|_2^2}{\lambda_{\min}(Q)}, \frac{\|B_\star\|_2^2}{\lambda_{\min}(R)}\right\}.$$

Moreover, we have that

$$\sum_{t=0}^{\infty} \mathbf{Tr}(P_t) - \mathbf{Tr}(P_\star) \ge -n\left(1 + \nu\right).$$

Proof. The first part follows from Proposition 1 of **?**] on value iteration. The second part follows by bounding

$$\mathbf{Tr}(P_t) - \mathbf{Tr}(P_\star) \ge -n\|P_t - P_\star\|_2 \ge -n\left(1 + \frac{1}{\nu}\right)^{-t},$$

and summing up these inequalities. □

Lemma E.7. *Fix a horizon $T_0 > 0$ and denote $\hat{x}_t = x_{t+T_0}$ and $\hat{u}_t = \hat{u}_{t+T_0}$. Then*

$$\mathbb{E}\left[\hat{x}_0^\top P \hat{x}_0\right] \leq \min_{\hat{u}_0, \hat{u}_1, \ldots} \mathbb{E}\left[\sum_{t=0}^{T-1} \hat{x}_t^\top Q \hat{x}_t + \hat{u}_t^\top R \hat{u}_t\right] - T J_\star + n\sigma_w^2(1+\nu) + \left(1 + \frac{1}{\nu}\right)^{-T} \mathbb{E}\|\hat{x}_0\|_2^2\,,$$

where

$$\nu = 2\|P_\star\|_2 \max\left\{\frac{\|A_\star\|_2^2}{\lambda_{\min}(Q)}, \frac{\|B_\star\|_2^2}{\lambda_{\min}(R)}\right\}.$$

Proof. Let us consider the Ricatti recursion

$$P_{t+1} = A^\top P_t A - A^\top P_t B(R + B^\top P_t B)^{-1} B^\top P_t A + Q,$$

where $P_0 = 0$. Then

$$\min_{\hat{u}_0, \hat{u}_1, \ldots} \mathbb{E}\left[\sum_{t=0}^{T} \hat{x}_t^\top Q \hat{x}_t + \hat{u}_t^\top R \hat{u}_t\right] = \mathbb{E}\hat{x}_0^\top P_T \hat{x}_0 + \sigma_w^2 \sum_{t=0}^{T-1} \mathbf{Tr}\left(P_t\right).$$

From the first part of Lemma **??** we know that

$$\|P_T - P_\star\|_2 \leq \left(1 + \frac{1}{\nu}\right)^{-T},$$

while from the second part of that Lemma we know that

$$\sum_{t=0}^{T-1} \left[\mathbf{Tr}\left(P_t\right) - \mathbf{Tr}\left(P_\star\right)\right] \geq -n(1+\nu).$$

The conclusion follows once we recall that $J_\star = \sigma_w^2\, \mathbf{Tr}(P_\star)$. □

Proof of Theorem 3.4. Let $T_0 > 0$ to be chosen later and let $T \geq T_0$. We decompose the regret as the sum of the regret from 0 to $T - T_0 - 1$ and the regret from $T - T_0$ to $T - 1$, and we write the first component in terms terms of the expected cost J_{T-T_0} defined in Eq. (**??**). We have

$$\begin{aligned}\sum_{t=0}^{T-1} \mathbb{E}\left[x_t^\top Q x_t + u_t^\top R u_t - J_\star\right] &= \mathbb{E}\left[x_{T-T_0}^\top P_\star x_{T-T_0} + \sum_{t=0}^{T-T_0-1} x_t^\top Q x_t + u_t^\top R u_t\right] - J_{T-T_0}^\star \\ &\quad + \sum_{t=T-T_0}^{T-1} \mathbb{E}\left[x_t^\top Q x_t + u_t^\top R u_t - J_\star\right] - T_0 J_\star \\ &\quad + \mathbf{Tr}(P_\star P_\infty) - \mathbb{E}x_{T-T_0}^\top P_\star x_{T-T_0},\end{aligned}$$

where we used $J_{T-T_0}^\star = (T - T_0)J_\star + \mathbf{Tr}(P_\star P_\infty)$. The term $\mathbf{Tr}(P_\star P_\infty)$ we an simply lower bound by zero since P_∞ and $P_\star$ are positive semi-definite matrices. From Lemmas **??** and **??** we have

$$\mathbb{E}\left[x_{T-T_0}^\top P_\star x_{T-T_0} + \sum_{t=0}^{T-T_0-1} x_t^\top Q x_t + u_t^\top R u_t\right] - J_{T-T_0}^\star \geq \lambda_{\min}(R) \sum_{t=0}^{T-T_0-1} \|\nu_t\|_2^2.$$

By Lemma **??** we have that

$$\begin{aligned}&\sum_{t=T-T_0}^{T-1} \mathbb{E}\left[x_t^\top Q x_t + u_t^\top R u_t - J_\star\right] - T_0 J_\star - \mathbb{E}x_{T-T_0}^\top P_\star x_{T-T_0} \\ &\qquad \geq -n\sigma_w^2(1+\nu) - \left(1 + \frac{1}{\nu}\right)^{-T_0} \mathbb{E}\|x_{T-T_0}\|_2^2\,.\end{aligned}$$

Then, from Lemma **??** we get

$$\sum_{t=0}^{T-1} \mathbb{E}\left[x_t^\top Q x_t + u_t^\top R u_t - J_\star\right] \geq \frac{1}{2}\lambda_{\min}(R) \sum_{t=0}^{T-T_0} \|\nu_t\|_2^2 - \underbrace{\left(3C\rho^{2T_0}\,\mathbf{Tr}(P_\infty) + n\sigma_w^2 \frac{\lambda_{\min}(R)}{2}\right)}_{C_0},$$

by choosing

$$T_0 \geq \frac{\log\left(\frac{2C^2}{(1-\rho^2)\lambda_{\min}(R)}\right)}{\log(1+\nu^{-1})}.$$

The conclusion follows by Lemma **??**. □

F Miscellaneous Results

First we state some results for the function class $\mathcal{RH}_\infty(C,\rho)$.

Lemma F.1. *Let* $\mathbf{G}_i \in \mathcal{RH}_\infty(C_i,\rho_i)$ *for* $i = 1,2$ *and Then* $\mathbf{H} = \mathbf{G}_1\mathbf{G}_2 \in \mathcal{RH}_\infty(C,\rho)$ *for any* $\rho \in (\max(\rho_1,\rho_2),1)$ *and* $C = \max\left\{1, \frac{1}{e\log\left(\frac{\rho}{\max(\rho_1,\rho_2)}\right)}\frac{\rho}{\max(\rho_1,\rho_2)}\right\}C_1C_2$. *Note for simplicity if we assume* $\rho \geq 1/4$ *we can take* $C = \frac{6C_1C_2}{1-\rho}$ *and* $\rho = \mathsf{Avg}(\max(\rho_1,\rho_2),1)$.

Proof. Assume wlog that $\rho_1 \geq \rho_2$. Note that $H(k) = \sum_{t=0}^{k} G_1(t)G_2(k-t)$, and therefore for all $k \geq 0$ we have that

$$\begin{aligned}\|H(k)\| = \left\|\sum_{t=0}^{k} G_1(t)G_2(k-t)\right\| &\leq C_1C_2\sum_{t=0}^{k}\rho_1^t\rho_2^{k-t}\\ &\leq C_1C_2\sum_{t=0}^{k}\rho_1^k = C_1C_2(k+1)\rho_1^k,\end{aligned}$$

Fix a $\rho \in (\rho_1,1)$. Define $g(k) = (k+1)(\rho_1/\rho)^k$ and $h(k) = \log g(k)$. We see that $h'(k) = 0$ only for $k = k_* = \frac{1}{\log(\rho/\rho_1)} - 1$. Furthermore, $h(k_*) = \log(1/\log(\rho/\rho_1)) - 1 + \log(\rho/\rho_1)$. Hence, $g(k_*) = \frac{1}{e\log(\rho/\rho_1)}(\rho/\rho_1)$.

The claim now follows since for any $k \geq 0$,

$$(k+1)\rho_1^k = (k+1)(\rho_1/\rho)^k\rho^k \leq \left[\sup_{k=0,1,\ldots}(k+1)(\rho_1/\rho)^k\right]\rho^k \leq \max\{1,g(k_*)\}\rho^k\ .$$

We also use the inequality $\log(1+x) \geq x/2$ for $x \in [0,2.5]$. □

Lemma F.2. *Let* $\mathbf{G}_i \in \mathcal{RH}_\infty(C_i,\rho_i)$ *for* $i=1,2$. *Then* $\mathbf{G}_1+\mathbf{G}_2 \in \mathcal{RH}_\infty(C_1+C_2,\max\{\rho_1,\rho_2\})$.

Proof. Straightforward from triangle inequality and the definitions. □

Lemma F.3. *Suppose that* $\mathbf{\Delta} \in \mathcal{RH}_\infty(C,\rho)$ *with* $C \leq 2$ *and* $\rho \geq 1/e$, *and furthermore* $\|\mathbf{\Delta}\|_{\mathcal{H}_\infty} < 1$. *Then we have*

$$(I \pm \mathbf{\Delta})^{-1} \in \mathcal{RH}_\infty\left(1 + \frac{\mathcal{O}(1)C}{1-\rho}, \mathsf{Avg}(\rho,1)\right)\ .$$

Proof. The function $f(x) = \frac{x}{e\log(x)}$ is monotonically decreasing on the interval $(1, 1/\rho)$. Hence for any $x \in (1, 1/\rho)$, we have $f(x) \geq f(1/\rho) \geq f(e) = 1$. Applying the composition lemma (Lemma **??**) to the system $\mathbf{\Delta} \circ \mathbf{\Delta}$, we have that for $c_1 \in (1, 1/\rho)$,

$$\mathbf{\Delta}^2 \in \mathcal{RH}_\infty \left(\frac{c_1}{e\log(c_1)} C^2, c_1\rho \right) .$$

Now if we recursively set $c_k \in (c_{k-1}, 1/\rho)$ for $k = 2, 3, ...$, repeated applications of the composition lemma yield that

$$\mathbf{\Delta}^n \in \mathcal{RH}_\infty \left(C^n \prod_{i=1}^{n-1} \frac{c_i}{e\log(c_i)}, c_{n-1}\rho \right) .$$

Let $c_\infty = \lim_{k\to\infty} c_k$, which exists and is finite because the sequence c_k is monotonically increasing and bounded above. Furthermore, we have that for any $n \geq 2$,

$$C^n \prod_{i=1}^{n-1} \frac{c_i}{e\log(c_i)} \leq \left(\frac{Cc_\infty}{e} \right)^{n-1} \frac{C}{\log(\sum_{i=1}^{n-1} c_i)} \leq \left(\frac{Cc_\infty}{e} \right)^{n-1} \frac{C}{\log(c_1)} .$$

Now choose any strictly increasing sequence such that $c_\infty = \mathsf{Avg}(1, 1/\rho) = (1/2)(1/\rho + 1)$ and $c_1 = \mathsf{Avg}(1, c_\infty) = (1/4)(3 + 1/\rho)$. By the addition lemma (Lemma **??**), the assumption on C, and a simple limiting argument,

$$\sum_{n=0}^{\infty} \mathbf{\Delta}^n \in \mathcal{RH}_\infty (C', c_\infty\rho) ,$$

where C' is given as

$$C' \leq 1 + C + \frac{C}{\log(c_1)} \frac{1}{1 - Cc_\infty/e} \leq 1 + C + \frac{2C}{\log(c_1)} .$$

The claim now follows by using the inequality $\log(1 + x) \geq x/2$ for $x \in [0, 2.5]$ and the assumed bound $C \leq 2$. $\square$

Lemma F.4. *Suppose that* $\mathbf{G} \in \mathcal{RH}_\infty(C, \rho)$. *Then* $\|\mathbf{G}\|_{\mathcal{H}_\infty} \leq \frac{C}{1-\rho}$.

Proof. We have that

$$\|\mathbf{G}\|_{\mathcal{H}_\infty} = \sup_{z\in\mathbb{T}} \|\mathbf{G}(z)\| = \sup_{z\in\mathbb{T}} \left\| \sum_{k=0}^{\infty} G(k) z^{-k} \right\| \leq C \sum_{k=0}^{\infty} \rho^k = \frac{C}{1-\rho}.$$

$\square$

Next, a probabilistic lemma which we use to control the LQR cost on a finite horizon.

Lemma F.5. *Let x and M be fixed, and $w \sim \mathcal{N}(0, \Sigma)$, with $\Sigma \succ 0$ and $\|\Sigma\| = \sigma^2$. Then there exists a universal constant $c > 0$ such that with probability at least $1 - \delta$*

$$\begin{aligned} \begin{bmatrix} x \\ w \end{bmatrix}^\top M \begin{bmatrix} x \\ w \end{bmatrix} &\leq x^\top M_{11} x + 2\sqrt{2}\sigma\|x\|\|M_{12}\|\sqrt{\log\left(\tfrac{2}{\delta}\right)} \\ &\quad + \mathbf{Tr}\, M_{22}\Sigma + c\sigma^2\|M_{22}\|_F\sqrt{\log\left(\tfrac{2}{\delta}\right)} + c\sigma^2\|M_{22}\|\log\left(\tfrac{2}{\delta}\right) . \end{aligned} \tag{F.1}$$

Proof. Expanding the quadratic we have

$$\begin{bmatrix} x \\ w \end{bmatrix}^\top M \begin{bmatrix} x \\ w \end{bmatrix} = x^\top M_{11} x + 2x^\top M_{12} w + w^\top M_{22} w.$$

Noting that $x^\top M_{12} w \sim \mathcal{N}(0, x^\top M_{12} \Sigma M_{12}^\top x)$, by standard Gaussian concentration we have with probability at least $1 - \frac{\delta}{2}$ that

$$\begin{aligned} x^\top M_{12} w &\leq \sqrt{2 x^\top M_{12} \Sigma M_{12}^\top x \log\left(\tfrac{2}{\delta}\right)} \\ &\leq \sqrt{2} \|x\| \|M_{12}\| \|\Sigma\|^{\frac{1}{2}} \sqrt{\log\left(\tfrac{2}{\delta}\right)} \\ &= \sqrt{2} \|x\| \sigma \|M_{12}\| \sqrt{\log\left(\tfrac{2}{\delta}\right)} . \end{aligned}$$

On the other hand, by the Hanson-Wright inequality [**?**], we have that with probability at least $1 - \frac{\delta}{2}$ that

$$\begin{aligned} w^\top M_{22} w &\leq \mathbf{Tr}\, M_{22} \Sigma + c \sqrt{\|\Sigma^{\frac{1}{2}} M_{22} \Sigma^{\frac{1}{2}}\|_F^2 \log\left(\tfrac{2}{\delta}\right)} + c \|\Sigma^{\frac{1}{2}} M_{22} \Sigma^{\frac{1}{2}}\|^2 \log\left(\tfrac{2}{\delta}\right) \\ &\leq \mathbf{Tr}\, M_{22} \Sigma + c\sigma^2 \|M_{22}\|_F \sqrt{\log\left(\tfrac{2}{\delta}\right)} + c\sigma^2 \|M_{22}\| \log\left(\tfrac{2}{\delta}\right) . \end{aligned}$$

□

Lemma F.6. *Let Σ be a $n \times n$ positive-definite matrix and let K be a real $p \times n$ matrix. Then, for any $\sigma_u \in \mathbb{R}$ we have that*

$$\lambda_{\min}\left(\begin{bmatrix} \Sigma & \Sigma K^\top \\ K\Sigma & K\Sigma K^\top + \sigma_u^2 I \end{bmatrix} \right) \geq \sigma_u^2 \min\left(\frac{1}{2}, \frac{\lambda_{\min}(\Sigma)}{2\|K\Sigma K^\top\|_2 + \sigma_u^2} \right) .$$

Proof. We find $0 < \gamma_1 < 1$ and $\gamma_2 > 0$ such that the following condition holds

$$\begin{bmatrix} \Sigma & \Sigma K^\top \\ K\Sigma & K\Sigma K^\top + \sigma_u^2 I \end{bmatrix} \succeq \begin{bmatrix} \gamma_1 \Sigma & 0 \\ 0 & \gamma_2 I \end{bmatrix} .$$

By Schur complements, this condition is equivalent to

$$\begin{aligned} 0 &\preceq K\Sigma K^\top + (\sigma_u^2 - \gamma_2) I - K\Sigma((1-\gamma_1)\Sigma)^{-1} \Sigma K^\top \\ &= -\frac{\gamma_1}{1-\gamma_1} K\Sigma K^\top + (\sigma_u^2 - \gamma_2) I . \end{aligned}$$

Now set $\gamma_2 = \sigma_u^2/2$ and $\gamma_1 = \frac{\sigma_u^2}{2\|K\Sigma K^\top\|_2 + \sigma_u^2}$. □

G Implementation of Adaptive Methods

We consider several adaptive methods for numerical comparison. This section described the relevant implementation details.

G.1 Optimism in the Face of Uncertainty

At the start of each epoch, the OFU method computes a confidence set around the dynamics and then finds the (A, B) that would achieve the smallest LQR cost. The method then plays the associated optimal controller.

The confidence sets at epoch i are of the form

$$\begin{aligned} C_i(\varepsilon) &= \{\Theta \in \mathbb{R}^{n \times (n+p)} : \mathbf{Tr}((\Theta - \widehat{\Theta}_i) Z_{T_i} (\Theta - \widehat{\Theta}_i)^\top) \leq \varepsilon\} , \\ Z_{T_i} &= \lambda I + \sum_{i=1}^{T_i} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top . \end{aligned} \tag{G.1}$$

Here, $\widehat{\Theta}_i$ denotes the (regularized) least squares estimate of the true parameters $\Theta_* = (A_\star, B_\star)$. For our experiments, we set $\lambda = 10^{-5}$ and $\varepsilon = \mathbf{Tr}((\widehat{\Theta}_i - \Theta_*) Z_{T_i} (\widehat{\Theta}_i - \Theta_*)^\top)$ using the true and estimation values of (A, B).

Then controller is selected by finding the "best" dynamics. To be precise, let $J(A, B) = \mathbf{Tr}(P(A, B))$, where $P(A, B)$ is the solution to the discrete algebraic Riccati solution

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

Then for every epoch of OFU, it is necessary to solve to the non-convex optimization problem

$$[\widetilde{A}, \widetilde{B}] = \arg \min_{[A,B]\in C_i(\varepsilon)} J(A, B)\ . \tag{G.2}$$

up to an absolute error of at most $O(1/\sqrt{T_i})$.

As in Section 5.4 of [**?**], we heuristically solve this optimization problem using projected gradient descent (PGD). An expression for the gradient of $\Theta \mapsto J(A, B)$ is derived in [**?**] (see also [4]) by use of the implicit function theorem. Specifically, $\nabla_\Theta \mathbf{Tr}(P(A, B))$ evaluated at a point $\Theta = (A, B)$ is an $n \times (n + p)$ matrix D. The i, j-th entry is given by $\mathbf{Tr}(E_{ij})$, where E_{ij} is the solution to the Lyapunov equation

$$E_{ij} = A_c^\top E_{ij} A_c + 2\mathrm{Sym}\left(A_c^\top P(A, B) e_i e_j^\top \begin{bmatrix} I \\ K \end{bmatrix}\right)\ ,$$

with K as the optimal LQR controller for (A, B), $A_c = A + BK$, and $\mathrm{Sym}(A) = \frac{1}{2}(A + A^\top)$. Finally, the projection of Θ onto the set $C_i(\varepsilon)$ can be solved by a eigendecomposition of Z_{T_i} followed by a scalar root-finding search. The details of this are also found in Section 5.4 of [**?**].

We determine the end of an epoch using a switching rule based on a slight modification of the determinant condition of [1]. We switch an epoch when both (a) $T - T_i \geq 10$ and (b) $\det(Z_T) > 2\det(Z_{T_i})$ hold. The first condition is to ensure that the switches are not too frequent in the beginning of the algorithm.

G.2 Thompson Sampling

The Thompson sampling algorithm is nearly identical to the OFU algorithm, except the optimization problem (**??**) is replaced by sampling. While the description of Thompson sampling in the Bayesian setting of [2] and [16] requires sampling from the posterior distribution, we follow the more frequentist setting of [4] and sample a point $\widetilde{\Theta}$ uniformly at random from the confidence set $C_i(\varepsilon)$ as in (**??**).

We implement this uniform samping by first drawing a $U \sim \mathrm{Unif}([0, 1])$ and a $\eta \in \mathbb{R}^{n\times(n+p)}$ with each $\eta_{ij} \sim \mathcal{N}(0, 1)$, and setting

$$\widetilde{\Theta} = \widehat{\Theta} + \sqrt{\varepsilon}\left(\frac{U^{1/(n(n+p))}}{\|\eta\|_F}\eta\right) Z_{T_i}^{-1/2}\ .$$

For the epoch switching rule, we follow the suggestion of [4] to force exploration after τ iterations, where we set $\tau = 500$. Specifically, we switch an epoch when the following predicate holds:

$$(T - T_i \geq \tau) \text{ or } ((T - T_i \geq 10) \text{ and } (\det(Z_T) > 2\det(Z_{T_i})))\ .$$

G.3 Robust Adaptive Control with FIR truncation

We now describe how to turn the infinite-dimensional optimization problem in Algorithm 1 into a finite-dimensional problem. First, recall the problem we want to solve,

$$\begin{aligned} &\text{minimize}_{\gamma\in[0,1)} \frac{1}{1-\gamma} \min_{\mathbf{\Phi}_x, \mathbf{\Phi}_u, V} \left\| \begin{bmatrix} Q^{1/2} & 0 \\ 0 & R^{1/2} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} \right\|_{\mathcal{H}_2} \\ &\text{s.t.} \begin{bmatrix} zI - \widehat{A} & -\widehat{B} \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_x \\ \mathbf{\Phi}_u \end{bmatrix} = I + \frac{1}{z^F} V\ ,\ \frac{\sqrt{2}\varepsilon}{1 - C_x\rho^{F+1}} \left\| \begin{bmatrix} \mathbf{\Phi_x} \\ \mathbf{\Phi_u} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq \gamma\ , \\ &\|V\| \leq C_x \rho^{F+1}\ ,\ \mathbf{\Phi}_x \in \frac{1}{z}\mathcal{RH}_\infty^F(C_x, \rho)\ ,\ \mathbf{\Phi}_u \in \frac{1}{z}\mathcal{RH}_\infty^F(C_u, \rho)\ . \end{aligned} \tag{G.3}$$

Ignoring the outer minimization over γ (which can be solved with bisection), the inner minimization is convex. Truncating the system responses to be FIR of length F means that

$$\mathbf{\Phi}_x = \sum_{k=1}^{F} \Phi_x(k) z^{-k}\ ,\ \ \mathbf{\Phi}_u = \sum_{k=1}^{F} \Phi_u(k) z^{-k}\ .$$

All pieces of the infinite dimensional problem can be written in terms of these variables. First, consider the $\mathcal{H}_2$ cost in the objective. By Parseval's identity, we can simply add the second order cone constraint

$$\left\| \begin{bmatrix} Q^{1/2}\Phi_x(1) \\ \vdots \\ Q^{1/2}\Phi_x(F) \\ R^{1/2}\Phi_u(1) \\ \vdots \\ R^{1/2}\Phi_u(F) \end{bmatrix} \right\|_F \leq t \, , \tag{G.4}$$

and minimize t. Next, we consider the constraints of the original optimization. The function space constraints reduce to the requirement that

$$\|\Phi_x(k)\| \leq C_x \rho^k \, , \ \ \|\Phi_u(k)\| \leq C_u \rho^k \, , \ \ k = 1, ..., F \, . \tag{G.5}$$

Next, to rewrite the subspace constraint, we first consider that

$$z\mathbf{\Phi}_x = \sum_{k=0}^{F-1} \Phi_x(k+1) z^{-k} \, ,$$

then the subspace constraint yields the following equality constraints,

$$\begin{aligned} \Phi_x(1) &= I \, , \\ \Phi_x(k+1) &= \widehat{A}\Phi_x(k) + \widehat{B}\Phi_u(k) \, , \ \ k = 1, ..., F-1 \, , \\ V &= \widehat{A}\Phi_x(F) + \widehat{B}\Phi_u(F) \, . \end{aligned} \tag{G.6}$$

The only constraint that remains is the $\mathcal{H}_\infty$ constraint, for which we use the following result.

Theorem G.1 (Theorem 5.8, [**?**])**.** *Consider the T-length FIR filter*

$$\mathbf{H}(z) = \sum_{k=0}^{T} H_k z^{-k} \, , H_k \in \mathbb{R}^{p \times m} \, .$$

Define the matrix

$$\overline{H} = \begin{bmatrix} H_0 \\ \vdots \\ H_T \end{bmatrix} \in \mathbb{R}^{p(T+1) \times m} \, .$$

We have that $\|\mathbf{H}(z)\|_{\mathcal{H}_\infty} \leq \gamma$ iff there exists $Q = Q^\top \succeq 0$ with $Q \in \mathbb{R}^{p(T+1) \times p(T+1)}$ satisfying

$$Q = \begin{bmatrix} Q_{00} & Q_{01} & \dots & Q_{0T} \\ * & Q_{11} & \dots & Q_{1T} \\ * & * & \ddots & \vdots \\ * & * & * & Q_{TT} \end{bmatrix} \, , \ \ Q_{ij} \in \mathbb{R}^{p \times p} \, ,$$

$$\sum_{t=0}^{T} Q_{tt} = \gamma^2 I_p \, , \ \ \sum_{t=0}^{T-k} Q_{t(t+k)} = 0_{p \times p} \, , k = 1, ..., T \, , \ \ \begin{bmatrix} Q & \overline{H} \\ \overline{H}^\top & I_m \end{bmatrix} \succeq 0 \, .$$

For the SLS problem, the $\mathcal{H}_\infty$ constraint on is the filter

$$\mathbf{H}(z) = \sum_{k=1}^{F} \begin{bmatrix} \Phi_x(k) \\ \Phi_u(k) \end{bmatrix} z^{-k} \, .$$

The constraint can be rewritten using the LMI in Theorem **??**. To avoid a decision variable of size $(n+p)(F+1) \times (n+p)(F+1)$, we instead consider the transpose system $\mathbf{H}^\top$ which has the same $\mathcal{H}_\infty$ norm and coefficients of size $n \times (n+p)$.

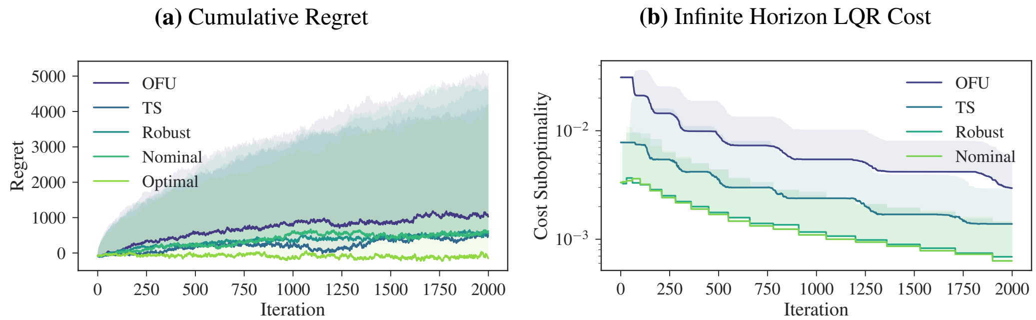

Figure 3: A comparison of different adaptive methods on 500 experiments of the large-transient system example (**??**). In (a), the median and 90th percentile cumulative regret is plotted over time. In (b), the median and 90th percentile infinite-horizon LQR cost of the epoch's controller.

Putting this together, we arrive at the following SDP, which can be solved using an off the shelf solver,

$$\begin{aligned}
\min_{\substack{\mathbf{\Phi}_x[k]\in\mathbb{R}^{n\times n},\ \mathbf{\Phi}_u[k]\in\mathbb{R}^{p\times n},\ V\in\mathbb{R}^{n\times n}\\ P\in\mathbb{R}^{n(F+1)\times n(F+1)},\ t\in\mathbb{R}}} \quad & t \\
\text{s.t.} \quad & (\textbf{??})\,,\ (\textbf{??})\,,\ (\textbf{??})\,, \\
& \sum_{t=0}^{F} P_{tt} = \gamma^2 I\,,\ \sum_{t=0}^{F-k} P_{t(t+k)} = 0\,,\ \ k=1,...,F\,, \\
& \overline{H} = \frac{\sqrt{2}\varepsilon}{1-C_x\rho^{F+1}}\begin{bmatrix} 0_{n\times n} & 0_{n\times p} \\ \Phi_x(1)^\top & \Phi_u(1)^\top \\ \vdots & \vdots \\ \Phi_x(F)^\top & \Phi_u(F)^\top \end{bmatrix}\,,\ \begin{bmatrix} P & \overline{H} \\ \overline{H}^\top & I_m \end{bmatrix} \succeq 0\,, \\
& \|V\| \leq C_x\rho^{F+1}\,.
\end{aligned}$$

For our experiments, we used the SCS solver [**?**] via CVXPY [**?**].

Finally, once the FIR responses $\{\Phi_x(k)\}_{k=1}^F$ and $\{\Phi_u(k)\}_{k=1}^F$ are found, we need a way to implement the system responses as a controller. We represent the dynamic controller $\mathbf{K} = \mathbf{\Phi}_u \mathbf{\Phi}_x^{-1}$ by finding an equivalent state-space realization (A_K, B_K, C_K, D_K) via Theorem 2 of [**?**].

As a final note, the adaptive method as described in Algorithm 1 requires several constants to be specified. For the numerical experiments, we set $\sigma_{\eta,i} = C_\eta \sigma_w T_i^{-1/3}$ where we vary C_η for different experiments, fix $\gamma = 0.98$, and use a fixed FIR trunction length of $F = 12$. For the experiments in Section 4, we set $C_\eta = 0.1$.

H Additional Experiments

H.1 Large-Transient Dynamics

We present the regret comparison results using another system

$$A_\star = \begin{bmatrix} 2 & 0 & 0 \\ 4 & 2 & 0 \\ 0 & 4 & 2 \end{bmatrix}\,,\ \ B_\star = I,\ \ Q = 10I,\ \ R = I\,. \tag{H.1}$$

The system is both unstable and has large transients. Each state receives direct input, and the cost is such that input size is penalized relatively less than state. This problem setting is amenable to robust methods due to both the cost ratio and the large transients, which are factors that may hurt optimistic methods. For this experiment, we ran all adaptive methods as described in Appendix **??**, and used an initialization with a horizon of length $T_0 = 250$ and $C_\eta = 2$.

The performance of the various adaptive methods is compared in Figure **??**. The median and 90th percentile regret over 500 instances is displayed in Figure **??**a, which gives an idea of both "average"

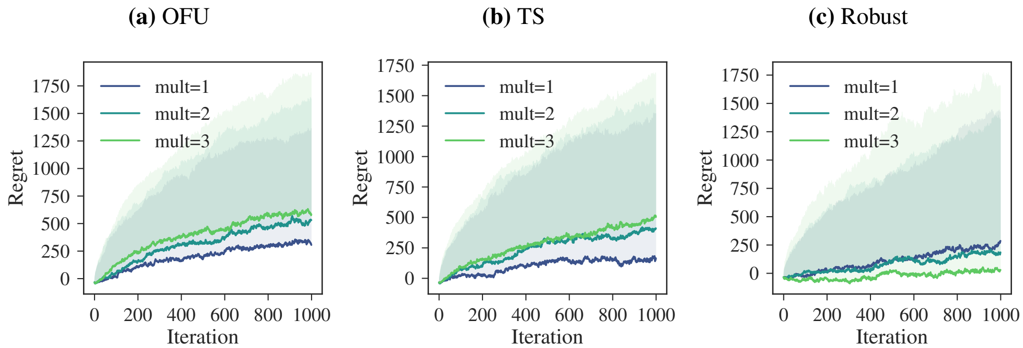

Figure 4: A comparison of cumulative regret when enlarged error bounds are used for synthesis, rather than the true errors. Both the median over 500 trials and the 90th percentile regret are plotted. In (a) is OFU, in (b) is TS, in (c) is robust. The plots show modest if any degradation in performance.

and worst-case behavior. Overall, the methods have very similar performance. One benefit of robustness is the guaranteed stability and therefore bounded infinite-horizon cost at every point during operation. In Figure **??**b, this infinite-horizon cost of the controller in each epoch is plotted. This measures the cost of using each epoch's controller indefinitely, rather than continuing to update its parameters. Especially for small numbers of iterations, the robust method performs relatively better than other adaptive algorithms, indicating that it is more amenable to early stopping.

H.2 Error Scaling

In our experiments, we use the actual estimation errors for controller synthesis. To examine the effect of this choice, we artificially inflate the estimation errors by various multipliers, and plot the regret for various methods in Figure **??**. These experiments were run on the the graph Laplacian example in (4.1) with an initialization with a horizon of length $T_0 = 300$ and $C_\eta = 1$.

The adaptive methods were run as described in Appendix **??**. The error term ε for OFU and TS appears in the computation of the uncertainty set as in (**??**). The errors ε_A and ε_B for the robust adaptive method appear in (**??**). The plot shows a modest degradation in regret as these terms are increased.

H.3 Learning the Disturbance Process

We consider the problem of regulating a known system which is subject to disturbances correlated in time. These disturbances are modeled as the output of a LTI filter driven by white noise. In other words,

$$x_{k+1} = A_\star x_k + B_\star u_k + d_k \;, \qquad d_{k+1} = A_d d_k + w_k,$$

where x_k is the state to drive to zero, and d_k are the disturbances. We will take $(A_\star, B_\star)$ to be known and A_d unknown. This setting models many phenomenon related to demand forecasting, in which the dynamics of e.g. a server farm is known, and the changes in demand are stochastic but correlated in time, and can thus be approximated by the output of an LTI filter.

The plant inputs u_k are designed for regulation. The controller design problem can be formulated as an optimization problem by defining the augmented system as

$$\begin{bmatrix} x_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} A_\star & I \\ 0 & A_d \end{bmatrix} \begin{bmatrix} x_k \\ d_k \end{bmatrix} + \begin{bmatrix} B_\star \\ 0 \end{bmatrix} u_k + \begin{bmatrix} 0 \\ I \end{bmatrix} w_k \;. \tag{H.2}$$

We will denote the augmented state $z_k = [x_k; d_k]$. Then the control actions can be designed using an adaptive LQR strategy. In many situations, inputs are relatively more costly, corresponding for example to energy usage. Defining an LQR cost directly related to the economics of the system can be unwise, due to the resulting tendency for states to become large, which may correspond to unsafe execution. While tuning the quadratic cost to represent a mixture of economic and safety considerations can often achieve good behavior in practice, the method is heuristic and lacks guarantees. Instead, consider the explicit addition of a constraint on the state, $\|x_k\|_\infty \leq a$ for $0 \leq k \leq H$ for some horizon (which may be infinite).

To state the necessary modification to the controller synthesis problem, we define the norm

$$\|\mathbf{M}\|_{\mathcal{L}_1} = \sup_{\|\mathbf{w}\|_\infty = 1} \|\mathbf{M}\mathbf{w}\|_\infty ,$$

for both system responses and state matrices. This norm corresponds to the $\ell_\infty \mapsto \ell_\infty$ operator norm.

Proposition H.1. *For the system described in (**??**), let $\mathbf{\Phi}_z$ denote a closed-loop state response. Then consider constraints*

$$\begin{aligned} \|(\mathbf{\Phi}_z)_{22}\|_{\mathcal{L}_1} &\le \gamma/\tilde{\varepsilon}_A , \\ \|(\mathbf{\Phi}_z)_{12}\|_{\mathcal{L}_1} &\le \frac{a}{b} \cdot (1-\gamma) := c \end{aligned} \tag{H.3}$$

*where $(\mathbf{\Phi}_z)_{ij}$ denotes the blocks defined by the partition of z_t into x_t and d_t, and $\|\widehat{A}_d - A_d\|_{\mathcal{L}_1} \le \tilde{\varepsilon}_A$. The addition of these constraints to the synthesis problem in (**??**) ensures that the resulting closed loop system has $\|x_k\|_\infty \le a$ for $0 \le k \le H$ as long as $\|w_k\|_\infty \le b$ for $0 \le k \le H$.*

Proof. In transfer function notation, the state of the plant can be described by

$$\mathbf{x} = [I \quad 0]\, \mathbf{z} = [I \quad 0]\, \mathbf{\Phi}_z (I + \widehat{\mathbf{\Delta}})^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix} \mathbf{w} .$$

Furthermore, due to the known structure of the dynamics,

$$(I + \widehat{\mathbf{\Delta}})^{-1} = \left(I + \begin{bmatrix} 0 & 0 \\ 0 & \Delta_A \end{bmatrix} \mathbf{\Phi}_z \right)^{-1} = \begin{bmatrix} I & 0 \\ X & (I + \Delta_A (\mathbf{\Phi}_z)_{22})^{-1} \end{bmatrix} ,$$

where $X = (I + \Delta_A(\mathbf{\Phi}_z)_{22})^{-1} \Delta_A (\mathbf{\Phi}_z)_{21}$. Then we have, letting $\widehat{\mathbf{\Delta}}_{22} = \Delta_A (\mathbf{\Phi}_z)_{22}$,

$$\mathbf{x} = [I \quad 0]\, \mathbf{\Phi}_z \begin{bmatrix} 0 \\ (I + \widehat{\mathbf{\Delta}}_{22})^{-1} \end{bmatrix} \mathbf{w} = (\mathbf{\Phi}_z)_{12} (I + \widehat{\mathbf{\Delta}}_{22})^{-1} \mathbf{w} .$$

Finally, to bound the size of the state,

$$\|\mathbf{x}\|_\infty \le \|(\mathbf{\Phi}_z)_{12} (I + \widehat{\mathbf{\Delta}}_{22})^{-1}\|_{\mathcal{L}_1} \|\mathbf{w}\|_\infty \le \frac{1}{1 - \|\widehat{\mathbf{\Delta}}_{22}\|_{\mathcal{L}_1}} \|(\mathbf{\Phi}_z)_{12}\|_{\mathcal{L}_1} \|\mathbf{w}\|_\infty .$$

Then we have that $\|\widehat{\mathbf{\Delta}}_{22}\|_{\mathcal{L}_1} \le \tilde{\varepsilon}_A \|(\mathbf{\Phi}_z)_{22}\|_{\mathcal{L}_1}$, so the result follows from the constraints and the assumption on w_k. □

Therefore, with either a bounded noise assumption on w_k or a high-probability bound over a finite time horizon, we can apply the previous result to synthesize safe controllers. In the example displayed in Figure 2, the constraint as in (**??**) is added to the controller synthesis procedure with $c = 0.1$ and $\gamma = 0.98$.