Learning Linear-Quadratic Regulators Efficiently with only $\sqrt{T}$ Regret

Alon Cohen* Tomer Koren† Yishay Mansour‡

February 26, 2019

Abstract

We present the first computationally-efficient algorithm with $\widetilde{O}(\sqrt{T})$ regret for learning in Linear Quadratic Control systems with unknown dynamics. By that, we resolve an open question of Abbasi-Yadkori and Szepesvári (2011) and Dean, Mania, Matni, Recht, and Tu (2018).

1 Introduction

Optimal control theory dates back to the 1950s, and has been applied successfully to numerous real-world engineering problems (e.g., Bermúdez and Martinez, 1994; Chen and Islam, 2005; Lenhart and Workman, 2007; Geering, 2007). Classical results in control theory pertain to asymptotic convergence and stability of dynamical systems, and recently, there has been a renewed interest in such problems from a learning-theoretic perspective with a focus on finite-time convergence guarantees and computational tractability.

Perhaps the most well-studied model in optimal control is Linear-Quadratic (LQ) control. In this model, both the state and the action are real-valued vectors. The dynamics of the environment are linear in both the state and action, and are perturbed by Gaussian noise; the cost is quadratic in the state and action vectors. When the costs and dynamics are known, the optimal control policy, which minimizes the steady-state cost, selects its actions as a linear function of the state vector, and can be derived by solving the algebraic Ricatti equations (e.g., Bertsekas et al., 2005).

Among the most challenging problems in LQ control is that of adaptive control: regulating a system with parameters which are initially unknown and have to be learned while incurring the associated costs. This problem is exceptionally challenging since the system might become unstable. Specifically, the controller must control the magnitude of the state vectors; otherwise, its cost might grow arbitrarily large.

Abbasi-Yadkori and Szepesvári (2011) were the first to address the adaptive control problem from a learning-theoretic perspective. In their setting, there is a learning agent who knows the quadratic costs, yet has no knowledge regarding the dynamics of the system. The agent acts for T rounds; at each round she observes the current state then chooses an action. Her goal is to minimize her regret, defined as the difference between her total cost and T times the steady-state cost of the optimal policy—one that is computed using complete knowledge of the dynamics.

Abbasi-Yadkori and Szepesvári (2011) gave $O(\sqrt{T})$-type regret bounds for LQ control where the dependency on the dimensionality is exponential, which was later improved by Ibrahimi et al. (2012)

*Technion—Israel Inst. of Technology and Google Tel Aviv; alon.cohen@campus.technion.ac.il.

†Google Brain, Mountain View; tkoren@google.com.

‡Tel-Aviv University and Google Tel Aviv; mansour.yishay@gmail.com.

to a polynomial dependence. However, the algorithms given in these works are not computationally efficient and require solving a complex non-convex optimization problem at each step. Developing an *efficient* algorithm with $O(\sqrt{T})$ regret has been a long standing open problem. Recently, Dean et al. (2018) proposed a computationally-efficient algorithm attaining an $O(T^{2/3})$ regret bound, and stated as an open problem providing an $O(\sqrt{T})$ regret efficient algorithm.

In this paper, we give the first computationally-efficient algorithm that attains $\widetilde{O}(\sqrt{T})$ regret for learning LQ systems, thus resolving the open problem of Abbasi-Yadkori and Szepesvári (2011) and Dean et al. (2018). The key to the efficiency of our algorithm is in reformulating the LQ control problem as a *convex* semi-definite program. Our algorithm solves a sequence of semi-definite relaxations of the infinite horizon LQ problem, the solutions of which are used to compute "optimistic" policies for the underlying unknown LQ system. As time progresses and the algorithm receives more samples from the system, these relaxations become tighter and serve as a better approximation of the actual LQ system. In this context, an optimistic policy is one that balances between exploration and exploitation; that is, between myopically utilizing its current information about the system parameters versus collecting new samples in order to obtain better estimates for subsequent predictions.

1.1 Related work

The techniques used in Abbasi-Yadkori and Szepesvári (2011); Ibrahimi et al. (2012) as well as those in this paper, draw inspiration from the UCRL algorithm (Jaksch et al., 2010) for learning in unknown Markov Decision Processes (MDPs). The main methodology is that of "optimism in the face of uncertainty" that has been highly influential in the reinforcement learning literature (Lai and Robbins, 1985; Brafman and Tennenholtz, 2002).

Over the years, techniques from reinforcement learning have been applied extensively in control theory. In particular, many recent works were published on the topic of learning LQ systems; these are Abbasi-Yadkori and Szepesvári (2011); Ibrahimi et al. (2012); Faradonbeh et al. (2017); Abbasi-Yadkori et al. (2018); Arora et al. (2018); Fazel et al. (2018); Malik et al. (2018) to name a few.

It is also worth noting an orthogonal line of works that attempts to adaptively control LQ systems using Thompson sampling, most notably Abeille and Lazaric (2017); Ouyang et al. (2017); Abeille and Lazaric (2018). Unfortunately, these works are also concerned with the statistical aspects of the problem, and none of them present computationally-efficient algorithms.

2 Preliminaries

Notation. The following notation will be used throughout the paper. We use $\|\cdot\|$ to denote the operator norm, that is, $\|M\| = \max_{x:\|x\|=1} \|Mx\|$ is the maximum singular value of a matrix M, and $\|\cdot\|_*$ to denote the trace norm, $\|M\|_* = \mathrm{Tr}(\sqrt{M^\mathsf{T} M})$. The notation $\rho(M)$ refers to the spectral radius of a matrix M, i.e., $\rho(M)$ is the largest absolute value of its eigenvalues.[1] Finally, we use the $A \bullet B$ to denote the entry-wise dot product between matrices, namely $A \bullet B = \mathrm{Tr}(A^\mathsf{T} B)$.

2.1 Problem Setting and Background

Linear-Quadratic Control. We consider the problem of adaptively controlling an unknown discrete-time *Linear-Quadratic Regulator* (LQR) over T rounds. At time t, a learner observes the

[1] Note that for a non-symmetric matrix M (as would often be the case in the sequel), the spectral radius can be very different from the operator norm of M. In particular, it could be the case that $\rho(M) < 1$ yet $\|M\| \gg 1$.

current state of the system, which is a vector $x_t \in \mathbb{R}^d$, and chooses an action $u_t \in \mathbb{R}^k$. Thereafter, the learner incurs a cost c_t, and the system transitions to the next state x_{t+1}, both of which are defined as follows:

$$\begin{aligned} c_t &= x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t \ ; \\ x_{t+1} &= A_\star x_t + B_\star u_t + w_t \ . \end{aligned} \tag{1}$$

Here, $Q \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{k\times k}$ are positive-definite matrices, $w_t \sim \mathcal{N}(0, W)$ is an i.i.d. zero-mean Gaussian vector with covariance W, and $A_\star \in \mathbb{R}^{d\times d}$ and $B_\star \in \mathbb{R}^{d\times k}$ are real valued matrices. We henceforth denote $n = d + k$, so that the augmented matrix $(A_\star\, B_\star)$ is of dimension $d \times n$.

A (stationary and deterministic) policy $\pi : \mathbb{R}^d \mapsto \mathbb{R}^k$ maps the current state x_t to an action u_t. The cost of the policy after T time steps is

$$J_T(\pi) = \sum_{t=1}^{T} \big(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t\big) \ ,$$

where $u_1, \ldots, u_T$ are chosen according to π starting from some fixed state x_1. In the infinite-horizon version of the problem, the goal is to minimize the steady-state cost $J(\pi) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[J_T(\pi)]$.

As is standard in the literature, we assume that the system (1) is controllable,[2] in which case the optimal policy that minimizes $J(\pi)$ is *linear*, i.e., has the form $\pi^\star(x) = K_\star x$ for some matrix $K_\star \in \mathbb{R}^{k\times d}$. For the optimal policy $\pi^\star$ we denote $J(\pi^\star) = J^\star$.

A policy $\pi(x) = Kx$ is *stable* if the matrix $A_\star + B_\star K$ is stable, that is, if $\rho(A_\star + B_\star K) < 1$. For a stable policy π we can define a cost-to-go function $x_1 \mapsto x_1^\mathsf{T} P x_1$ that maps a state x_1 to the total additional expected cost of π when starting from x_1. Concretely, we have $x_1^\mathsf{T} P x_1 = \sum_{t=1}^{\infty}\big(\mathbb{E}[c_t] - J(\pi)\big)$. For the optimal policy $\pi^\star(x) = K_\star x$, a classic result (Whittle, 1996; Bertsekas et al., 2005) states that the matrix $P^\star$ associated with its cost-to-go function is a positive definite matrix that satisfies:

$$P^\star \preceq Q + K^\mathsf{T} R K + (A_\star + B_\star K)^\mathsf{T} P^\star (A_\star + B_\star K) \tag{2}$$

for any matrix $K \in \mathbb{R}^{k\times d}$, with equality when $K = K_\star$:

$$P^\star = Q + K_\star^\mathsf{T} R K_\star + (A_\star + B_\star K_\star)^\mathsf{T} P^\star (A_\star + B_\star K_\star) \ . \tag{3}$$

Furthermore, the optimal steady-state cost $J^\star$ equals $P^\star \bullet W$.

Problem definition. We henceforth consider a learning setting in which the learner is uninformed about the dynamics of the system. Namely, the matrices $A_\star$ and $B_\star$ in Eq. (1) are fixed but unknown to the learner. For simplicity, we assume that the cost matrices Q and R are fixed and known; a straightforward yet technical adaptation of our approach can handle uncertainties in these matrices as well.

A learning algorithm is a mapping from the current state x_t and previous observations $\{x_s, u_s\}_{s=1}^{t-1}$ to an action u_t at time t. An algorithm is measured by its T-round regret, defined as the difference between its total cost over T rounds and T times the steady-state cost of the optimal policy which knows both $A_\star$ and $B_\star$. That is,

$$R_T = \sum_{t=1}^{T} \big(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J^\star\big) \ ,$$

where $u_1, \ldots, u_T$ are the actions chosen by the algorithm and $x_1, \ldots, x_T$ are the resulting states.

[2] The system (1) is said to be controllable when the matrix $\big(B_\star\ A_\star B_\star\ \cdots\ A_\star^{d-1} B_\star\big)$ is of full rank.

Our assumptions. We make the following assumptions about the LQ system (1):

(i) there are known positive constants $\alpha_0, \alpha_1, \sigma, \vartheta, \nu > 0$ such that

$$\alpha_0 I \preceq Q \preceq \alpha_1 I, \quad \alpha_0 I \preceq R \preceq \alpha_1 I, \quad W = \sigma^2 I, \quad \|(A_\star\; B_\star)\| \leq \vartheta, \quad J^\star \leq \nu;$$

(ii) there is a policy $K_0 \in \mathbb{R}^{k \times d}$, known to the learner, which is stable for the LQR (1).

Assumption (i) is rather mild and only requires having upper and lower bounds on the unknown system parameters. We remark that the assumption $W = \sigma^2 I$ is made only for simplicity, and in fact, our analysis only requires upper and lower bounds on the eigenvalues of W. Assumption (ii), which has already appeared in the context of learning in LQRs (Dean et al., 2018), is also not very restrictive. In realistic systems, it is reasonable that one knows how to "reset" the dynamics and prevent them from reaching unbounded states. Further, in many cases a stabilizing policy can be found efficiently (Dean et al., 2017).

2.2 SDP Formulation of LQR

A key step in our approach towards the design of an *efficient* learning algorithm is in reformulating the planning problem in LQRs as a convex optimization problem. To this end, we make use of a semidefinite formulation introduced in Cohen et al. (2018) that would allow us to find the optimal cost of the LQ system (1):

$$\begin{aligned} &\text{minimize} && \left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right) \bullet \Sigma \\ &\text{subject to} && \Sigma_{xx} = \big(A_\star\; B_\star\big)\Sigma\big(A_\star\; B_\star\big)^{\mathsf{T}} + W \;, \\ & && \Sigma \succeq 0 \;. \end{aligned} \tag{4}$$

Here, Σ is an $n \times n$ PSD matrix, with $n = d + k$, that has the following block structure:

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xu} \\ \Sigma_{ux} & \Sigma_{uu} \end{pmatrix},$$

where $\Sigma_{xx} \in \mathbb{R}^{d \times d}$, $\Sigma_{ux} = \Sigma_{xu}^{\mathsf{T}} \in \mathbb{R}^{k \times d}$ and $\Sigma_{uu} \in \mathbb{R}^{k \times k}$. The matrix Σ represents the covariance matrix of the joint distribution of (x, u) when the system is in its steady-state.

As was established in Cohen et al. (2018), the optimal value of the program is exactly the infinite-horizon optimal cost $J^\star$. Moreover, when $W \succ 0$, the optimal policy of the system $K_\star$ can be extracted from an optimal Σ via $K = \mathcal{K}(\Sigma)$ where $\mathcal{K}(\Sigma) = \Sigma_{ux}\Sigma_{xx}^{-1}$. In fact, when the LQ system follows *any* stable policy K, the state vectors converge to a steady-state distribution whose covariance matrix is denoted by $X = \mathbb{E}[xx^{\mathsf{T}}]$, and the matrix $\mathcal{E}(K) = \left(\begin{smallmatrix} X & XK^{\mathsf{T}} \\ KX & KXK^{\mathsf{T}} \end{smallmatrix}\right)$ is feasible for the SDP. This particularly implies that the optimal solution $\Sigma^\star$ is of rank d and has the form $\Sigma^\star = \mathcal{E}(K_\star)$. This is formalized as follows.

Theorem (Cohen et al., 2018)**.** *Let Σ be any feasible solution to the SDP (4), and let $K = \mathcal{K}(\Sigma)$. Then the policy $\pi(x) = Kx$ is stable for the LQR (1), and it holds that $\mathcal{E}(K) \preceq \Sigma$. In particular, $\mathcal{E}(K)$ is also feasible for the SDP and its cost is at most that of Σ.*

2.3 Strong Stability

The quadratic cost function is unbounded. Indeed, it might be that the norms of the state vectors $x_1, x_2, \ldots$ grow exponentially fast resulting in poor regret for the learner.

To alleviate this issue we rely on the notion of a strongly-stable policy, introduced by Cohen et al. (2018). Intuitively, strongly-stable policies are ones in which the norms of the state vectors remain controlled.

Definition 1 (strong stability)**.** A matrix M is (κ, γ)-*strongly stable* (for $\kappa \geq 1$ and $0 < \gamma \leq 1$) if there exists matrices $H \succ 0$ and L such that $M = HLH^{-1}$, with $\|L\| \leq 1 - \gamma$ and $\|H\|\|H^{-1}\| \leq \kappa$.

A policy K for the linear system (1) is (κ, γ)-*strongly stable* (for $\kappa \geq 1$ and $0 < \gamma \leq 1$) if $\|K\| \leq \kappa$ and the matrix $A_\star + B_\star K$ is (κ, γ)-strongly stable.

We note that, in particular, any stable policy K is in fact (κ, γ)-strongly stable for some $\kappa, \gamma > 0$ (see Cohen et al., 2018 for a proof). Our analysis requires a stronger notion that pertains to the stability of a sequence of policies, also borrowed from Cohen et al. (2018).

Definition 2 (sequential strong stability)**.** A sequence of policies $K_1, K_2, \ldots$ for the linear dynamics in Eq. (1) is (κ, γ)-strongly stable (for $\kappa > 0$ and $0 < \gamma \leq 1$) if there exist matrices $H_1, H_2, \ldots \succ 0$ and $L_1, L_2, \ldots$ such that $A_\star + B_\star K_t = H_t L_t H_t^{-1}$ for all t, with the following properties:

(i) $\|L_t\| \leq 1 - \gamma$ and $\|K_t\| \leq \kappa$;

(ii) $\|H_t\| \leq B_0$ and $\|H_t^{-1}\| \leq 1/b_0$ with $\kappa = B_0/b_0$;

(iii) $\|H_{t+1}^{-1} H_t\| \leq 1 + \gamma/2$.

For a sequentially strongly stable sequence of policies one can show that the expected magnitude of the state vectors remains controlled; for completeness, we include a proof in Appendix A.3.

Lemma 3. *Let $x_1, x_2, \ldots$ be a sequence of states starting from a deterministic state x_1, and generated by the dynamics in Eq. (1) following a (κ, γ)-strongly stable sequence of policies $K_1, K_2, \ldots$. Then, for all $t \geq 1$ we have*

$$\|x_t\| \leq \kappa e^{-\gamma(t-1)/2} \|x_1\| + \frac{2\kappa}{\gamma} \max_{1 \leq s < t} \|w_t\|.$$

3 Efficient Algorithm for Learning in LQRs

In this section we describe our efficient online algorithm for learning in LQRs; see pseudo-code in Algorithm 1. The algorithm receives as input the parameters α_0, ν, σ^2 and ϑ, further requires an initial estimate $(A_0 \, B_0)$ that approximates the true parameters $(A_\star \, B_\star)$ within an error ϵ. As we later show, this estimate only needs to be accurate to within $\epsilon = O(1/\sqrt{T})$ of the true parameters, and we can make sure this is satisfied by employing a known stabilizing policy K_0 for exploration over $O(\sqrt{T})$ rounds.

We next describe in detail the main steps of the algorithm. The algorithm maintains estimates $(A_t \, B_t)$ of the true parameters $(A_\star \, B_\star)$ that improve from round to round, as well as a PD matrix $V_t \succ 0$ that represents a confidence ellipsoid around the current estimates $(A_t \, B_t)$. The algorithm proceeds in epochs, each starting whenever the volume of the ellipsoid is halved and consists of the following steps.

3.1 Estimating parameters

The first step of each epoch is standard: we employ a least-squares estimator (in line 7) to construct a new approximation $(A_t \, B_t)$ of the parameters $(A_\star \, B_\star)$ based on the observations z_t collected so far. The confidence bounds of this estimator are given in terms of the covariance matrix V_t of the vectors $z_1, \ldots, z_{t-1}$.

Algorithm 1 OSLO: Optimistic Semi-definite programming for Lq cOntrol

1: **input**: parameters $\alpha_0, \sigma^2, \vartheta, \nu > 0$; confidence $\delta \in (0,1)$; and an initial estimate $(A_0\, B_0)$ such that $\|(A_0\, B_0) - (A_\star\, B_\star)\|_\mathsf{F}^2 \le \epsilon$.

2: **initialize**: $\mu = 5\vartheta\sqrt{T}$, $V_1 = \lambda I$ where

$$\lambda = \frac{2^{11}\nu^5\vartheta\sqrt{T}}{\alpha_0^5\sigma^{10}} \quad \text{and} \quad \beta = \frac{2^{18}\nu^4 n^2}{\alpha_0^4\sigma^6}\log\frac{T}{\delta} \ .$$

3: **for** $t = 1, \ldots, T$ **do**

4: **receive** state x_t.

5: **if** $\det(V_t) > 2\det(V_\tau)$ or $t = 1$ **then**

6: **start new episode**: $\tau = t$.

7: **estimate system parameters**: Let $(A_t\, B_t)$ be a minimizer of

$$\frac{1}{\beta}\sum_{s=1}^{t-1}\|(A\, B)\, z_s - x_{s+1}\|^2 + \lambda\|(A\, B) - (A_0\, B_0)\,\|_\mathsf{F}^2$$

over all matrices $(A\, B) \in \mathbb{R}^{d\times n}$.

8: **compute policy**: let $\Sigma_t \in \mathbb{R}^{n\times n}$ be an optimal solution to the SDP program:

$$\begin{aligned} &\min\ \Sigma \bullet \left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right) \\ &\text{s.t.}\ \ \Sigma_{xx} \succeq (A_t\ B_t)\Sigma(A_t\ B_t)^\mathsf{T} + W - \mu(\Sigma \bullet V_t^{-1})I \ , \\ &\qquad \Sigma \succeq 0 \ . \end{aligned}$$

9: **set** $K_t = (\Sigma_t)_{ux}\,(\Sigma_t)_{xx}^{-1}$.

10: **else**

11: **set** $K_t = K_{t-1}$, $A_t = A_{t-1}$, $B_t = B_{t-1}$.

12: **end if**

13: **play** $u_t = K_t x_t$.

14: **update** $z_t = \left(\begin{smallmatrix} x_t \\ u_t \end{smallmatrix}\right)$ and $V_{t+1} = V_t + \beta^{-1} z_t z_t^\mathsf{T}$.

15: **end for**

3.2 Computing a policy via an SDP

The main step of the algorithm takes place in line 8 of Algorithm 1, where we form a "relaxed" SDP program based on the current estimates $(A_t\, B_t)$ and the corresponding confidence matrix V_t, and solve it in order to compute a stable policy for the underlying LQR system. The idea here is to adapt the SDP formulation (4) of the LQR system, whose description needs the true underlying parameters, to an SDP program that only relies on estimates of the true parameters and accounts for the uncertainty associated with them. Once the relaxed SDP is solved, extracting a (deterministic) policy K_t from the solution Σ_t is done in the same way as in the case of the exact SDP (4).

The relaxed SDP incorporates a relaxed form of the inequality constraint in (4); as we show in the analysis, this program is a relaxation of the "exact" SDP (4) provided that the estimates $(A_t\, B_t)$ are sufficiently accurate (this is one place where having fairly accurate initial estimates as input to the algorithm is useful). In other words, the relaxed SDP always underestimates the steady-state cost of the optimal policy of the LQR (1). In this sense, Algorithm 1 is "optimistic in the face of uncertainty" (e.g., Brafman and Tennenholtz, 2002; Jaksch et al., 2010).

3.3 Exploring, exploiting, and updating confidence

After retrieving a policy K_t, the algorithm takes action: it computes $u_t = K_t x_t$, which is the action recommended by policy K_t at state x_t, and then plays u_t and updates the confidence matrix V_t with the new observations at step t. The policy K_t therefore serves and balances two goals—*exploitation* and *exploration*—as it is used both as a "best guess" to the optimal policy (based on past observations), as well as means to collect new samples and obtain better estimates of the system parameters in subsequent steps of the algorithm.

4 Overview of Analysis

We now formally state our main result: a high-probability $\widetilde{O}(\sqrt{T})$ regret bound for the efficient algorithm given in Algorithm 1.

Theorem 4. *Suppose that Algorithm 1 is initialized so that the initial estimation error $\|(A_0\, B_0) - (A_\star\, B_\star)\|_{\mathsf{F}}^2 \le \epsilon$ satisfies*

$$\epsilon \le \frac{1}{4\lambda} = \frac{\alpha_0^5 \sigma^{10}}{2^{13}\nu^5\vartheta\sqrt{T}} \ .$$

Assume $T \ge \mathrm{poly}(n, \nu, \vartheta, \alpha_0^{-1}, \sigma^{-1}, \|x_1\|)$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ the regret of Algorithm 1 satisfies

$$R_T = O\bigg(\frac{\nu^5 n^3 \vartheta}{\alpha_0^4 \sigma^8}\sqrt{T \log^4 \frac{T}{\delta}} + \nu\sqrt{T\log^3\frac{T}{\delta}}\bigg) \ .$$

Furthermore, the run-time per round of the procedure is polynomial in these factors.

Remark. At first glance it may appear that the regret bound of Theorem 4 becomes worse as the noise variance σ^2 becomes smaller. This seems highly counter-intuitive and, indeed, is not true in general. This is because when σ is small we also expect the bound on the optimal loss ν to be small. In particular, suppose that $K_\star$ is $(\kappa_\star, \gamma_\star)$-strongly stable; then, one can show that $J^\star \le \sigma^2\alpha_1\kappa_\star^2/\gamma_\star$. Plugging this as ν into the bound of Theorem 4 reveals a linear dependence in σ^2.

In Section 6 we show how to set up the initial conditions of Theorem 4; we utilize a stable (but otherwise arbitrary) policy given as input and show the following.

Corollary 5. *Suppose we are provided a policy K_0 which is known to be (κ_0, γ_0)-strongly stable for the LQR (1). Assume $T \ge \mathrm{poly}(n, \nu, \vartheta, \alpha_0^{-1}, \sigma^{-1}, \kappa_0, \gamma_0^{-1}, \log(\delta^{-1}))$. Suppose at first we utilize K_0 in the warm-up procedure of Algorithm 2 for*

$$T_0 = \Theta\bigg(\frac{n^2\nu^5\vartheta}{\alpha_0^5\sigma^{10}}\sqrt{T\log^2\frac{T}{\delta}}\bigg)$$

rounds; thereafter, we run Algorithm 1. Then, the initial conditions of Theorem 4 hold by the end of the warm-up phase, and with probability at least $1-\delta$ the regret of the overall procedure is bounded as

$$R_T = O\bigg(\frac{\alpha_1 n^2 \nu^5 \vartheta \kappa_0^4}{\alpha_0^5\sigma^8\gamma_0^2}(n + k\vartheta^2\kappa_0^2)\sqrt{T\log^4\frac{T}{\delta}} + \nu\sqrt{T\log^3\frac{T}{\delta}}\bigg) \ .$$

Furthermore, the runtime per round of the procedure is polynomial in these factors and in $T, \log(1/\delta)$.

In the remainder of the section, we give an overview of the main steps in the analysis, delegating the technical proofs to later sections and appendices.

4.1 Parameters estimation

Algorithm 1 repeatedly computes least-square estimates of $(A_\star\ B_\star)$. The next theorem, similar to one shown in Abbasi-Yadkori and Szepesvári (2011), yields a high-probability bound on the error of this least-squares estimate.

Lemma 6. *Let $\Delta_t = (A_t\ B_t) - (A_\star\ B_\star)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{Tr}(\Delta_t V_t \Delta_t^\mathsf{T}) \le \frac{4\sigma^2 d}{\beta} \log\left(\frac{d}{\delta}\frac{\det(V_t)}{\det(V_1)}\right) + 2\lambda \left\|\Delta_0\right\|_\mathsf{F}^2 .$$

In particular, when $\|\Delta_0\|_\mathsf{F}^2 \le 1/(4\lambda)$ and $\sum_{s=1}^t \|z_s\|^2 \le 2\beta T$, one has $\mathrm{Tr}(\Delta_t V_t \Delta_t^\mathsf{T}) \le 1$.

We see that the boundness of the states z_t (specifically, the fact that they do not grow exponentially with t) is crucial for the estimation. Below, we will show how the policies computed by the algorithm ensure this condition.

The proof of Lemma 6 is based on a self-normalized martingale concentration inequality due to Abbasi-Yadkori et al. (2011); for completeness, we include a proof in Appendix B.3.

4.2 Policy computation via a relaxed SDP

Next, assume that the estimates A_t, B_t of $A_\star, B_\star$ computed in the previous step are indeed such that the error $\Delta_t = (A_t\ B_t) - (A_\star\ B_\star)$ has $\mathrm{Tr}(\Delta_t V_t \Delta_t^\mathsf{T}) \le 1$ for the confidence matrix $V_t = \lambda I + \beta^{-1}\sum_{s=1}^{t-1} z_s z_s^\mathsf{T}$.

Consider the relaxed SDP program solved by the algorithm in line 8. The following lemma follows from the optimality conditions of the SDP and will be used to extract a stable policy from the SDP solution, and to relate the cost of actions taken by this policy to properties of the SDP solutions. This lemma, together with Lemma 9 below, summarize the key consequences of the relaxed SDP formulation that central to our approach; we elaborate more on the relaxed SDP and its properties in Section 5 below.

Lemma 7. *Assume the conditions of Theorem 4, and further that $\|V_t\| \le 4T$. Then the SDP solved in line 8 of the algorithm is a relaxation of the exact SDP (4), and we have:*
(i) the value of the optimal solution is at most $J^\star \le \nu$ which implies $\|\Sigma_t\|_* \le J^\star/\alpha_0$;
(ii) $(\Sigma_t)_{xx}$ is invertible and so the policy $K_t = (\Sigma_t)_{ux}(\Sigma_t)_{xx}^{-1}$ is well defined;
(iii) there exists a positive semi-definite matrix $P_t \succeq 0$ with $\|P_t\|_* \le J^\star/\sigma^2$ such that

$$P_t \succeq Q + K_t^\mathsf{T} P_t K_t + (A_\star + B_\star K_t)^\mathsf{T} P_t (A_\star + B_\star K_t) - 2\mu\|P_t\|_* \begin{pmatrix} I \\ K_t \end{pmatrix}^\mathsf{T} V_t^{-1} \begin{pmatrix} I \\ K_t \end{pmatrix} .$$

The positive definite matrix P_t in the above lemma is in fact the dual variable corresponding to the optimal solution Σ_t of the (primal) SDP, and the equality involving P_t follows from the complementary slackness conditions of the SDP. This equality can be viewed as an approximate version of the Ricatti equation that applies to policies computed based on estimates of the system parameters (as opposed to the "exact" Ricatti equation, which is relevant only for *optimal* policies of the actual LQR, that can only be computed based on the true parameters).

4.3 Boundness of states

Next, we show that the policies computed by the algorithm keep the underlying system stable, and that state vectors visited by the algorithm are uniformly bounded with high probability. To this end, consider the following sequence of "good events" $\mathcal{E}_1 \supseteq \mathcal{E}_2 \supseteq \cdots \supseteq \mathcal{E}_T$, where for each t,

$$\mathcal{E}_t = \left\{ \forall\, s = 1, \ldots, t, \quad \mathrm{Tr}(\Delta_s V_s \Delta_s^\mathsf{T}) \leq 1 \ , \ \|z_s\|^2 \leq 4\kappa^4 e^{-\gamma(s-1)} \|x_1\|^2 + \beta \right\} .$$

That is, $\mathcal{E}_t$ is the event on which everything worked as planned up to round t: our estimations were sufficiently accurate and the norms of $\{z_s\}_{s=1}^t$ were properly bounded. We show that the events $\mathcal{E}_1, \ldots, \mathcal{E}_T$ hold with high probability; this would ensure that V_t is appropriately bounded.

Lemma 8. *Under the conditions of Theorem 4, the event $\mathcal{E}_T$ occurs with probability $\geq 1 - \delta/2$.*

4.4 Sequential strong stability

Crucially, Lemma 8 above holds true since the sequence of policies extracted by Algorithm 1 from repeated solutions to the relaxed SDP is *sequentially* strongly stable.

Lemma 9. *Assume the conditions of Theorem 4, and further that for any t, $\|V_s\| \leq 4T$ for all $s = 1, \ldots, t$. Then the sequence of policies $K_1, \ldots, K_t$ is (κ, γ)-strongly stable for $\kappa = \sqrt{2\nu/\alpha_0\sigma^2}$ and $\gamma = 1/2\kappa^2$.*

This follows from a stability property of solutions to the relaxed SDP: we show that as the relaxed constraint becomes tighter, the optimal solutions of the SDP do not change by much (see Section 5). This, in turn, can be used to show that the policies extracted from these solutions are not drastically different from each other, and so the sequence of policies generated by the algorithm keeps the system stable. Lemma 8 is then implied via a simple inductive argument: suppose that the state-vector norms are bounded up until round t; then the sequence of policies generated until time t is strongly-stable thus keeping the norms of future states bounded with high probability.

We remark that stability of the individual policies does not suffice, and the stronger *sequential* strong stability condition is in fact required for our analysis. Indeed, even if we guarantee the (non-sequential) strong stability of each individual policy, the system's state might blow up exponentially in the number of times the algorithm switches between policies: after switching to a new policy there is an initial burn-in period in which the norm of the state can increase by a constant factor (and thereafter stabilize). Thus, even if we ensure that there are as few as $O(\log T)$ policy switches, the states might become polynomially large in T and deteriorate our regret guarantee. Sequential strong stability wards off against such a blow up in the magnitude of states.

4.5 Regret analysis

Let us now connect the dots and sketch how our main result (Theorem 4) is derived; for the formal proof, see Appendix B.1. Consider the instantaneous regret $r_t = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J^\star$ and let $\widetilde{R}_T = \sum_{t=1}^T r_t \mathbb{I}\{\mathcal{E}_t\}$. We will bound $\widetilde{R}_T$ with high probability, and since $R_t = \widetilde{R}_T$ with high probability due to Lemma 8, this would imply a high-probability bound on R_T from which the theorem would follow.

To bound the random variable $\widetilde{R}_T$, we appeal to Lemma 7 that can be used to relate the instantaneous regret of the algorithm to properties of the SDP solutions it computes. Conditioned on the good event $\mathcal{E}_t$, the boundness of the visited states ensures that the confidence matrix V_t is bounded as the lemma requires. The lemma then implies that

$$Q + K_t^\mathsf{T} R K_t \preceq P_t - (A_\star + B_\star K_t)^\mathsf{T} P_t (A_\star + B_\star K_t) + 2\mu \|P_t\|_* \left(\begin{smallmatrix} I \\ K_t \end{smallmatrix}\right)^\mathsf{T} V_t^{-1} \left(\begin{smallmatrix} I \\ K_t \end{smallmatrix}\right) .$$

On the other hand, as $u_t = K_t x_t$ and $J^\star \geq \sigma^2 \|P_t\|_*$ (which is also a consequence of Lemma 7), we have

$$r_t = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J^\star \leq x_t^\mathsf{T} \big(Q + K_t^\mathsf{T} R K_t\big) x_t - \sigma^2 \|P_t\|_*.$$

Combining the inequalities and summing over $t = 1, \ldots, T$, gives via some algebraic manipulations the following bound:

$$\begin{aligned} \widetilde{R}_T \leq \quad & \sum_{t=1}^{T} \big(x_t^\mathsf{T} P_t x_t - x_{t+1}^\mathsf{T} P_t x_{t+1}\big) \mathbb{I}\{\mathcal{E}_t\} \\ & + \sum_{t=1}^{T} w_t^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) x_t \mathbb{I}\{\mathcal{E}_t\} \\ & + \sum_{t=1}^{T} \big(w_t^\mathsf{T} P_t w_t - \sigma^2 \|P_t\|_*\big) \mathbb{I}\{\mathcal{E}_t\} \\ & + \frac{4\nu\mu}{\sigma^2} \sum_{t=1}^{T} \big(z_t^\mathsf{T} V_t^{-1} z_t\big) \mathbb{I}\{\mathcal{E}_t\} \ . \end{aligned} \tag{5}$$

We now proceed to bounding each of the sums in the above. The first sum above telescopes over consecutive rounds in which Algorithm 1 uses the same policy and thus the matrix P_t remains unchanged. Therefore, the number of remaining terms, each of which is bounded by a constant, is exactly the number of times that Algorithm 1 computes a new policy. We show that when the good events occur, the number of policy switches is at most $O(n \log T)$, which gives rise to the following.

Lemma 10. *It holds that*

$$\sum_{t=1}^{T} \big(x_t^\mathsf{T} P_t x_t - x_{t+1}^\mathsf{T} P_t x_{t+1}\big) \mathbb{I}\{\mathcal{E}_t\} \leq \frac{4\nu}{\sigma^2} \big(4\kappa^4 \|x_1\|^2 + \beta\big) n \log T \ .$$

The next two terms in the bound above are sums of martingale difference sequences, as the noise terms w_t are i.i.d., and each w_t is independent of P_t, K_t and x_t. Using standard concentration arguments, we show that both are bounded by $\widetilde{O}(\sqrt{T})$ with high probability.

Lemma 11. *With probability at least $1 - \delta/4$, it holds that*

$$\sum_{t=1}^{T} w_t^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) x_t \mathbb{I}\{\mathcal{E}_t\} \leq \frac{\nu\vartheta}{\sigma} \sqrt{3\beta T \log \frac{4}{\delta}} \ .$$

Lemma 12. *With probability at least $1 - \delta/4$, it holds that*

$$\sum_{t=1}^{T} \big(w_t^\mathsf{T} P_t w_t - \sigma^2 \|P_t\|_*\big) \mathbb{I}\{\mathcal{E}_t\} \leq 8\nu \sqrt{T \log^3 \frac{4T}{\delta}} \ .$$

Finally, using the elementary identity $z^{\mathsf{T}}V^{-1}z \leq 2\log(\det(V + zz^{\mathsf{T}})/\det(V))$ for $V \succ 0$ and any vector z such that $z^{\mathsf{T}}V^{-1}z \leq 1$, we show that the final sum in the bound telescopes and can be bounded in terms of $\log(\det(V_{T+1})/\det(V_1))$; in turn, the latter quantity can be bounded by $O(n \log T)$ using the fact that the z_t are uniformly bounded on the event $\mathcal{E}_T$. This argument results with:

Lemma 13. *We have* $\sum_{t=1}^{T}(z_t^{\mathsf{T}}V_t^{-1}z_t)\mathbb{I}\{\mathcal{E}_t\} \leq 4\beta n \log T$.

Our main theorem now follows by plugging-in the bounds into Eq. (5), using a union bound to bound the failure probability, and applying some algebraic simplification.

5 The relaxed SDP program

In this section we present useful properties of the relaxed SDP program repeatedly solved by Algorithm 1, which are used to prove Lemmas 7 and 9 discussed above and are central to our development.

The relaxed SDP program takes the following form. Let $\mu > 0$ be a fixed parameter, and assume A, B and V are matrices such that the error matrix $\Delta = (A\,B) - (A_\star\,B_\star)$ satisfies $\mathrm{Tr}(\Delta V \Delta^{\mathsf{T}}) \leq 1$.

$$\begin{aligned} &\text{minimize} && \Sigma \bullet \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \\ &\text{subject to} && \Sigma_{xx} \succeq (A\ B)\Sigma(A\ B)^{\mathsf{T}} + W - \mu(\Sigma \bullet V^{-1})I, \\ &&& \Sigma \succeq 0,\ \Sigma \in \mathbb{R}^{n\times n}\ . \end{aligned} \tag{6}$$

For this section, the dual program to (6) will be useful:

$$\begin{aligned} &\text{maximize} && P \bullet W \\ &\text{subject to} && \begin{pmatrix} Q-P & 0 \\ 0 & R \end{pmatrix} + (A\,B)^{\mathsf{T}}P(A\,B) \succeq \mu\|P\|_* V^{-1}, \\ &&& P \succeq 0,\ P \in \mathbb{R}^{d\times d}\ . \end{aligned} \tag{7}$$

We now aim at proving Lemma 7 which states that SDP (6) is a relaxation of the original exact SDP (4). It follows directly from Lemmas 15 and 16 given below; see Appendix B.4. First, we present a matrix-perturbation lemma also proven in Appendix C.1.

Lemma 14. *Let X and Δ be matrices of matching sizes and assume $\Delta^{\mathsf{T}}\Delta \preceq V^{-1}$ for some matrix $V \succ 0$. Then for any $\Sigma \succeq 0$ and $\mu \geq 1 + 2\|X\|\|V\|^{1/2}$,*

$$\|(X+\Delta)\Sigma(X+\Delta)^{\mathsf{T}} - X\Sigma X^{\mathsf{T}}\| \leq \mu\Sigma \bullet V^{-1}\ .$$

Lemma 15. *Assume $\mu \geq 1 + 2\vartheta\|V\|^{1/2}$. Then the optimal value of SDP (6) is at most $J^\star$. Consequently, for a primal-dual optimal solution Σ, P we have $\|\Sigma\|_* \leq J^\star/\alpha_0$ and $\|P\|_* \leq J^\star/\sigma^2$.*

Proof. It suffices to show that $\Sigma^\star$, the solution to the original SDP (4), is feasible for the relaxed SDP. Indeed, $\Sigma^\star \succeq 0$, and combining Eq. (4) and Lemma 14 (note that $\mathrm{Tr}(\Delta V \Delta^{\mathsf{T}}) \leq 1$ implies that $\Delta^{\mathsf{T}}\Delta \preceq V^{-1}$) yields Eq. (6) due to

$$\Sigma^\star_{xx} = \big(A_\star\ B_\star\big)\Sigma^\star\big(A_\star\ B_\star\big)^{\mathsf{T}} + W \succeq \big(A\ B\big)\Sigma^\star\big(A\ B\big)^{\mathsf{T}} + W - \mu(\Sigma^\star \bullet V^{-1})I\ .$$

Therefore, it is feasible for SDP (6). $\square$

The next lemma shows how to extract a policy from the relaxed SDP. Somewhat surprisingly, this policy is deterministic and has the linear form $x \mapsto Kx$, as is the case in the original SDP.

Lemma 16. *Assume that $V \succeq (\nu\mu/\alpha_0\sigma^2)I$, and $\mu \geq 1 + 2\vartheta\|V\|^{1/2}$. Let Σ and P be primal and dual optimal solutions to the relaxed SDP. Then Σ_{xx} is invertible, and for $K = \Sigma_{ux}\Sigma_{xx}^{-1}$ we have*

$$P = Q + K^\mathsf{T} PK + (A + BK)^\mathsf{T} P(A + BK) - \mu\|P\|_* \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right) V^{-1} \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T}.$$

Proof. Denote

$$Z = \left(\begin{smallmatrix} Q-P & 0 \\ 0 & R \end{smallmatrix}\right) + \left(A\ B\right)^\mathsf{T} P \left(A\ B\right) - \mu\|P\|_* V^{-1} \ .$$

Recall the complementary-slackness conditions of the SDP, that read $\Sigma Z = 0$. We now show that $\Sigma_{xx} \succ 0$ and $\operatorname{rank}(\Sigma) = d$ as this would entail that

$$\Sigma = \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)\Sigma_{xx}\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T} \quad \text{for} \quad K = \Sigma_{ux}\Sigma_{xx}^{-1}.$$

Thus the span of Σ is the span of $\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)$ whence $\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T} Z \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right) = 0$ as required.

To that end, we begin by stating the following basic fact about matrices: For any two n-dimensional symmetric matrices, X, Y, that satisfy $XY = 0$, it must be that $\operatorname{rank}(X) + \operatorname{rank}(Y) \leq n$. Then it suffices to show $\Sigma_{xx} \succ 0$ and $\operatorname{rank}(Z) \geq k$. Indeed, using Lemma 15,

$$\mu(\Sigma \bullet V^{-1})I \preceq \mu\|\Sigma\|_*\|V^{-1}\|I \prec \mu \frac{\nu}{\alpha_0} \frac{\alpha_0\sigma^2}{\nu\mu} I = W, \tag{8}$$

$$\mu\|P\|_* V^{-1} \preceq \mu\|P\|_*\|V^{-1}\|I \prec \mu \frac{\nu}{\sigma^2} \frac{\alpha_0\sigma^2}{\nu\mu} I \preceq \left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right), \tag{9}$$

as $W = \sigma^2 I$ and $\left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right) \succeq \alpha_0 I$. Plugging Eq. (8) into Eq. (6) and using $\Sigma \succeq 0$, shows that $\Sigma_{xx} \succ 0$. Moreover, Z is the difference of

$$\left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right) + \left(A\ B\right)^\mathsf{T} P \left(A\ B\right) - \mu\|P\|_* V^{-1},$$

which is of rank $d + k$ in light of Eq. (9) and since $P \succeq 0$, and $\left(\begin{smallmatrix} P & 0 \\ 0 & 0 \end{smallmatrix}\right)$ which is of rank at most d. Therefore, $\operatorname{rank}(Z) \geq k$ as required. ∎

We continue with proving the main result of this section that would imply Lemma 9 (see Appendix B.6). We show that the sequence of policies generated by solving a certain series of relaxed SDPs is strongly-stable.

Theorem 17. *Let $P_1, P_2, \ldots$ be optimal solutions to the relaxed dual SDP; each P_t associated with $(A_t\ B_t)$ and V_t respectively. Let $\kappa = \sqrt{2\nu/\alpha_0\sigma^2}$, $\gamma = 1/2\kappa^2$, and suppose that $\mu \geq 1 + 2\vartheta\|V_t\|^{1/2}$ and $V_t \succeq 16\kappa^{10}\mu I$ for all t. Moreover, let K_t be the policy associated with P_t (as in Lemma 16). Then the sequence $K_1, K_2, \ldots$ is (κ, γ)-strongly stable.*

The proof is given by combining the following two lemmas. Indeed, in Appendix C.2 we show that each policy K_t is strongly stable.

Lemma 18. *K_t is (κ, γ)-strongly stable for $A_\star + B_\star K_t = H_t L_t H_t^{-1}$ where $H_t = P_t^{1/2}$ and $\|L_t\| \leq 1 - \gamma$. Moreover, $(\alpha_0/2)I \preceq P_t \preceq (\nu/\sigma^2)I$.*

Furthermore, having established strong stability, the next lemma shows that P_t is "close" to P_{t+1} (see Appendix C.3 for a proof).

Lemma 19. *$P_t \preceq P^\star \preceq P_{t+1} + (\alpha_0\gamma/2)I$ for all $t \geq 1$.*

Proof of Theorem 17. We show that the conditions for sequential strong-stability hold. Notice that not only does Lemma 18 show that for all t, K_t is (κ,γ)-strongly stable, it also gives us uniform upper and lower bounds on $H_t = P_t^{1/2}$ as $\|P_t\| \le \|P_t\|_* \le \nu/\sigma^2$ (Lemma 15), and $\|P_t^{-1}\| \le 2/\alpha_0$. Together with $P_{t+1} \succeq (\alpha_0/2)I$, the lemma implies

$$\begin{aligned}
\|H_{t+1}^{-1}H_t\|^2 &= \|P_{t+1}^{-1/2}P_t^{1/2}\|^2 \\
&= \|P_{t+1}^{-1/2}P_tP_{t+1}^{-1/2}\| \\
&\le \|I + \tfrac{1}{2}\alpha_0\gamma P_{t+1}^{-1}\| \\
&\le 1 + \tfrac{1}{2}\alpha_0\gamma\|P_{t+1}^{-1}\| \\
&\le 1+\gamma \ .
\end{aligned}$$

Thus $\|H_{t+1}^{-1}H_t\| \le \sqrt{1+\gamma} \le 1+\frac{1}{2}\gamma$ which provides sequential strong-stability. □

6 Warm-up Using a Stable Policy

In this section we give a simple warm-up scheme that can be used in an initial exploration phase, after which the conditions of our main algorithm are met. Here we assume that we are given a policy K_0 which is known to be (κ_0,γ_0)-strongly stable for the LQR (1).

Starting from $x_1 = 0$ and over T_0 rounds, the warm-up procedure samples actions $u_t \sim \mathcal{N}(K_0x_t, 2\sigma^2\kappa_0^2 I)$ independently; this is summarized in Algorithm 2.

Algorithm 2 Warm-up procedure

input: (κ_0,γ_0)-strongly stable policy K_0, horizon T_0.
for $t = 1,\ldots,T_0$ **do**
  observe state x_t.
  play $u_t \sim \mathcal{N}(K_0x_t, 2\sigma^2\kappa_0^2 I)$.
end for

Let $V_0 = \sum_{t=1}^{T_0} z_tz_t^\mathsf{T}$ be the empirical covariance matrix corresponding to the samples z_t collected during warm-up, where $z_t = \left(\begin{smallmatrix} x_t \\ u_t \end{smallmatrix}\right)$ for all t. The main result of this section gives upper and lower bounds on the matrix V_0.

Theorem 20. *Let $\delta \in (0,1)$. Provided that $T_0 \ge \mathrm{poly}(\sigma, n, \vartheta, \kappa_0, \gamma_0^{-1}, \log(\delta^{-1}))$, we have with probability at least $1-\delta$ that*

$$\begin{aligned}
\mathrm{Tr}(V_0) &\le T_0 \cdot \frac{300\sigma^2\kappa_0^4}{\gamma_0^2}\big(n + k\vartheta^2\kappa_0^2\big)\log\frac{T_0}{\delta} \ , \\
\|x_{T_0}\|^2 &\le \frac{150\sigma^2\kappa_0^2}{\gamma_0}\big(n + k\vartheta^2\kappa_0^2\big)\log\frac{T_0}{\delta} \ , \\
V_0 &\succeq \frac{T_0\sigma^2}{80} I \ ,
\end{aligned}$$

and for $V = V_0 + \sigma^2\vartheta^{-2}I$ and initial estimates $\big(A_0 \; B_0\big) = \sum_{t=1}^{T_0-1} x_{t+1}z_t^\mathsf{T} V^{-1}$ we have

$$\mathrm{Tr}(\Delta_0 V \Delta_0^\mathsf{T}) \le 20n^2\sigma^2 \log\frac{T_0}{\delta} \ .$$

With Theorem 20 in hand, the proof of Corollary 5 readily follows; see details in Appendix B.2. The proof of Theorem 20 itself is based on adaptations of techniques developed in Simchowitz et al. (2018) in the context of identification of Linear Dynamical Systems, and is given in Appendix D.1.

Acknowledgments

We thank Yoram Singer and Kunal Talwar for many helpful discussions. YM was supported in part by a grant from the Israel Science Foundation (ISF). AC thanks Lotem Peled for her assistance and support.

References

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

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvari. Regret bounds for model-free linear quadratic control. *arXiv preprint arXiv:1804.06021*, 2018.

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. *arXiv preprint arXiv:1703.08972*, 2017.

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

Sanjeev Arora, Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Towards provable control for unknown linear dynamical systems, 2018. URL https://openreview.net/forum?id=BygpQlbA-.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Alfredo Bermúdez and Aurea Martinez. A state constrained optimal control problem related to the sterilization of canned foods. *Automatica*, 30(2):319–329, 1994.

Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 2005.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Ping Chen and Sardar MN Islam. *Optimal control models in finance*. Springer, 2005.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1029–1038, 2018.

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

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*, 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time analysis of optimal adaptive policies for linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1466–1475, 2018.

Hans P Geering. Optimal control with engineering applications. *Berlin Heidelberg*, 2007.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Morteza Ibrahimi, Adel Javanmard, and Benjamin V Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2012.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Suzanne Lenhart and John T Workman. *Optimal control applied to biological models.* Crc Press, 2007.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *arXiv preprint arXiv:1812.08305*, 2018.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Peter Whittle. *Optimal control: basics and beyond.* John Wiley & Sons, Inc., 1996.

Farrol Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, pages 1068–1070, 1973.

A Preliminaries

A.1 Concentration inequalities

First, we state a variant of the Hanson-Wright inequality (Hanson and Wright, 1971; Wright, 1973), which can be found in Hsu et al. (2012).

Theorem 21 (Hanson-Wright inequality)**.** *Let $x \sim \mathcal{N}(0, I_n)$ be a Gaussian random vector and let $A \in \mathbb{R}^{m\times n}$. For all $z > 0$,*

$$\mathbb{P}\big[\|Ax\|^2 - \|A\|_{\mathsf{F}}^2 > 2\|A\|_{\mathsf{F}}\|A\|\sqrt{z} + 2\|A\|^2 z\big] < e^{-z}.$$

In particular, if $x \sim \mathcal{N}(0, \Sigma)$ *then* $\mathbb{E}\|Ax\|^2 = \mathrm{Tr}(A\Sigma A^\mathsf{T})$ *and for any* $z \geq 1$,

$$\mathbb{P}\big[\|Ax\|^2 - \mathbb{E}\|Ax\|^2 > 4z\|\Sigma\|\|A\|\|A\|_\mathsf{F}\big] < e^{-z}.$$

The following is Azuma's inequality for concentration of martingales with bounded differences.

Theorem 22 (Azuma, 1967)**.** *Let* $X_1, \ldots, X_N$ *be a martingale difference sequence such that* $|X_i| \leq c$ *for all* $i = 1, \ldots, n$. *Then,*

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i > t\right] \leq \exp\left(-\frac{t^2}{2nc^2}\right) .$$

The following is is a self-normalized concentration inequality for vector-valued martingales useful for guaranteeing generalization in linear regression.

Theorem 23 (Abbasi-Yadkori et al., 2011)**.** *Let* $(\mathcal{F}_t)_{t=0}^{\infty}$ *be a filtration and let* $(\eta_t)_{t=1}^{\infty}$ *be a real-valued martingale difference sequence adapted to* $(\mathcal{F}_t)$ *such that* η_t *is R-sub-Gaussian conditioned on* $\mathcal{F}_{t-1}$*, that is,*

$$\mathbb{E}\big[e^{\lambda\eta_t} \bigm| \mathcal{F}_{t-1}\big] \leq e^{\lambda^2 R^2/2}, \qquad \forall\, t \geq 1 .$$

Further, let $(u_t)_{t=1}^{\infty}$ *be an* $\mathbb{R}^n$*-valued stochastic process adapted to* $(\mathcal{F}_{t-1})_{t=1}^{\infty}$*, let* $V \in \mathbb{R}^{n\times n}$ *positive definite matrix, and define*

$$U_t = \sum_{s=1}^{t-1} \eta_s u_s \ , \qquad V_t = V + \sum_{s=1}^{t-1} u_s u_s^\mathsf{T} \ , \qquad t = 1, 2, \ldots .$$

Then, for any $\delta \in (0,1)$ *we have with probability at least* $1 - \delta$ *that*

$$U_t^\mathsf{T} V_t^{-1} U_t \leq 2R^2 \log\left(\frac{1}{\delta}\frac{\det(V_t)}{\det(V)}\right) \qquad \forall\, t = 1, 2, \ldots \ .$$

A.2 Technical Lemmas

Lemma 24. *Let* X *and* Δ *be matrices of matching sizes and assume* $\Delta^\mathsf{T}\Delta \preceq V^{-1}$ *for some matrix* $V \succ 0$. *Then for any* $P \succeq 0$ *and* $\mu \geq 1 + 2\|X\|\|V\|^{1/2}$,

$$-\mu\|P\|_* V^{-1} \ \preceq\ (X+\Delta)^\mathsf{T} P (X+\Delta) - X^\mathsf{T} P X \ \preceq\ \mu\|P\|_* V^{-1} .$$

Proof. Note that $(X+\Delta)^\mathsf{T} P(X+\Delta) - X^\mathsf{T} PX = X^\mathsf{T} P\Delta + \Delta^\mathsf{T} PX + \Delta^\mathsf{T} P\Delta$. Let $\epsilon > 0$. We have

$$X^\mathsf{T} P\Delta + \Delta^\mathsf{T} PX \preceq \epsilon^{-1} X^\mathsf{T} PX + \epsilon\Delta^\mathsf{T} P\Delta;$$

this can be seen by expanding the inequality $(\epsilon^{-1/2}X - \epsilon^{1/2}\Delta)^\mathsf{T} P(\epsilon^{-1/2}X - \epsilon^{1/2}\Delta) \succeq 0$. Setting $\epsilon = \|X\|\|V\|^{1/2}$ and using our assumption that $\Delta^\mathsf{T}\Delta \preceq V^{-1}$ yields

$$\begin{aligned}
X^\mathsf{T} P\Delta + \Delta^\mathsf{T} PX &\preceq \epsilon^{-1} X^\mathsf{T} PX + \epsilon\Delta^\mathsf{T} P\Delta \\
&\preceq \epsilon^{-1}\|X\|^2\|P\| I + \epsilon\|P\| V^{-1} \\
&= \|X\|\|P\|\Big(\|V\|^{-1/2} I + \|V\|^{1/2} V^{-1}\Big) && (\epsilon = \|X\|\|V\|^{1/2}) \\
&\preceq \|X\|\|P\|\|V\|^{1/2} V^{-1} . && (\|V\|^{-1/2} I \preceq \|V\|^{1/2} V^{-1})
\end{aligned}$$

This, together with $\Delta^\mathsf{T} P\Delta \preceq \|P\|\Delta^\mathsf{T}\Delta \preceq \|P\|V^{-1}$, proves one direction of the inequality. For the other direction, a similar argument shows

$$(X+\Delta)^\mathsf{T} P(X+\Delta) - X^\mathsf{T} PX \succeq X^\mathsf{T} P\Delta + \Delta^\mathsf{T} PX \succeq -\|P\|\|X\|\|V\|^{1/2} V^{-1} .$$

□

Lemma 25. *Let X, Z be symmetric matrices of equal sizes and Y a (κ, γ)-strongly stable matrix such that $X \preceq Y^{\mathsf{T}} XY + Z$. Then $X \preceq (\kappa^2/\gamma)\|Z\| I$.*

Proof. The inequality $X \preceq Y^{\mathsf{T}} XY + Z$ implies there exists a matrix M such that $M \succeq 0$, and $X = Y^{\mathsf{T}} XY + Z - M$. As Y is stable, the equation has a unique solution that satisfies:

$$X = \sum_{t=0}^{\infty} (Y^t)^{\mathsf{T}} (Z - M) Y^t \preceq \sum_{t=0}^{\infty} (Y^t)^{\mathsf{T}} Z Y^t.$$

Let us proceed in bounding the norm of the right-hand side of this inequality. As Y is (κ, γ)-strongly stable, $Y = HLH^{-1}$ with $\|L\| \leq 1 - \gamma$ and $\|H\| \|H^{-1}\| \leq \kappa$. Therefore,

$$\left\| \sum_{t=0}^{\infty} (Y^t)^{\mathsf{T}} Z Y^t \right\| \leq \sum_{t=0}^{\infty} \|Y^t\|^2 \|Z\| \ ,$$

and we have

$$\|Y^t\| = \|HL^t H^{-1}\| \leq \|H\| \|H^{-1}\| \|L\| \leq \kappa(1 - \gamma) \ .$$

This implies that

$$\left\| \sum_{t=0}^{\infty} (Y^t)^{\mathsf{T}} Z Y^t \right\| \leq \|Z\| \sum_{t=0}^{\infty} \kappa^2 (1 - \gamma)^{2t} \leq \|Z\| \kappa^2 \sum_{t=0}^{\infty} (1 - \gamma)^t = \frac{\kappa^2}{\gamma} \|Z\|.$$ □

Lemma 26. *For $M \succ 0$ and a vector z such that $z^{\mathsf{T}} M^{-1} z \leq 1$,*

$$z^{\mathsf{T}} M^{-1} z \leq 2 \log \frac{\det(M + zz^{\mathsf{T}})}{\det M} \ .$$

Proof. Observe that $\det(M + zz^{\mathsf{T}}) = \det(M) \det(I + M^{-1/2} zz^{\mathsf{T}} M^{-1/2}) = (1 + z^{\mathsf{T}} M^{-1} z) \det(M)$ by the determinant lemma, and so

$$\log(1 + z^{\mathsf{T}} M^{-1} z) = \log \frac{\det(M + zz^{\mathsf{T}})}{\det(M)} \ .$$

The proof is finished using the concavity of $x \mapsto \log(1 + x)$ and the fact that $0 \leq z^{\mathsf{T}} M^{-1} z \leq 1$:

$$\log(1 + z^{\mathsf{T}} M^{-1} z) \geq (1 - z^{\mathsf{T}} M^{-1} z) \log 1 + (z^{\mathsf{T}} M^{-1} z) \log 2 \geq \tfrac{1}{2} z^{\mathsf{T}} M^{-1} z \ .$$ □

Lemma 27. *If $N \succeq M \succ 0$, then for any vector v one has*

$$v^{\mathsf{T}} N v \leq \frac{\det N}{\det M} v^{\mathsf{T}} M v.$$

Proof. Note that the claimed inequality is equivalent to $N \preceq (\det(N)/\det(M))M$, which in turn is equivalent to $\|M^{-1/2} N M^{-1/2}\| \leq \det(M^{-1/2} N M^{-1/2})$. The latter is true because $R = M^{-1/2} N M^{-1/2} \succeq I$, and so the product of the eigenvalues of R (all of which are ≥ 1) is no smaller than the maximal eigenvalue of R. □

A.3 Proof of Lemma 3

Proof. Following the sequence $K_1, K_2, \ldots$ induces updates $x_{t+1} = (A_\star + B_\star K_t)x_t + w_t$. Thus

$$x_t = M_1 x_1 + \sum_{s=1}^{t-1} M_{s+1} w_s \ ,$$

where

$$M_t = I; \quad M_s = M_{s+1}\big(A_\star + B_\star K_s\big) = \prod_{j=s}^{t-1}(A_\star + B_\star K_j), \quad \forall\, 1 \le s \le t-1.$$

Since the sequence $K_1, K_2, \ldots$ is sequential strong stable, there exist matrices $H_1, H_2, \ldots$ and $L_1, L_2, \ldots$ such that $A_\star + B_\star K_j = H_j L_j H_j^{-1}$ with the properties specified in Definition 2. Thus, we have for all $1 \le s < t$ that

$$\begin{aligned}
\|M_s\| &= \left\| \prod_{j=s}^{t-1} H_j L_j^\mathsf{T} H_j^{-1} \right\| \\
&\le \|H_{t-1}\| \left(\prod_{j=s}^{t-1} \|L_j\| \right) \left(\prod_{j=s}^{t-2} \|H_{j+1}^{-1} H_j\| \right) \|H_s^{-1}\| \\
&\le B_0 (1-\gamma)^{t-s} (1+\gamma/2)^{t-s} (1/b_0) \\
&\le \kappa (1-\gamma/2)^{t-s} \ .
\end{aligned}$$

As $\kappa \ge 1$, the same holds for M_t.

Thus, for all $t \ge 1$,

$$\begin{aligned}
\|x_t\| &\le \|M_1\| \|x_1\| + \sum_{s=1}^{t-1} \|M_{s+1}\| \|w_s\| \\
&\le \kappa (1-\gamma/2)^{t-1} \|x_1\| + \kappa \sum_{s=1}^{t-1} (1-\gamma/2)^{t-s-1} \|w_s\| \\
&\le \kappa e^{-\gamma(t-1)/2} \|x_1\| + \kappa \max_{1 \le s < t} \|w_s\| \sum_{t=1}^{\infty} (1-\gamma/2)^t \\
&= \kappa e^{-\gamma(t-1)/2} \|x_1\| + \frac{2\kappa}{\gamma} \max_{1 \le s < t} \|w_s\| \ .
\end{aligned}$$

□

B Proofs of Section 4

For the proofs in this section, we require the following two simple lemmas.

Lemma 28. *Assume $T \ge 2$, $\lambda \ge 1$ and $\sum_{s=1}^{t} \|z_s\|^2 \le 2\beta t$. Let $V_t = \lambda I + \beta^{-1} \sum_{t=1}^{t-1} z_t z_t^\mathsf{T}$. Then*

$$\log \frac{\det(V_t)}{\det(V_1)} \le 2n \log T$$

.

Proof. We have

$$\begin{aligned}
\log \frac{\det(V_t)}{\det(V_1)} &= \log \det(V_1^{-1/2} V_t V_1^{-1/2}) \\
&\le n \log \|V_1^{-1/2} V_t V_1^{-1/2}\| \\
&\le n \log \left(1 + \frac{1}{\beta\lambda} \sum_{s=1}^{t-1} \|z_s\|^2\right) && (V_1 = \lambda I;\ V_t = \beta^{-1} \textstyle\sum_{s=0}^{t-1} z_s z_s^\mathsf{T} + V_1) \\
&\le n \log(1 + 2T) && (\textstyle\sum_{s=1}^{t} \|z_s\|^2 \le 2\beta T;\ \lambda \ge 1) \\
&\le 2n \log T\ . && (T \ge 2)
\end{aligned}$$

□

Lemma 29. *Assume that* $\|z_s\|^2 \le 4\kappa^4 e^{-\gamma(t-1)} + \beta$ *for* $s = 1, \ldots, t$, *and* $\kappa = \sqrt{2\nu/\alpha_0\sigma^2}$, $\gamma = 1/2\kappa^2$. *Also suppose that* $t \ge \|x_1\|^2$. *Then* $\sum_{s=1}^{t} \|z_s\|^2 \le 2\beta t$.

Proof.

$$\sum_{s=1}^{t} \|z_s\|^2 \le \frac{8\kappa^4}{\gamma} \|x_1\|^2 + \beta t = 16\kappa^6 \|x_1\|^2 + \beta t \le 2\beta t\ .$$

□

B.1 Proof of Main Theorem (Theorem 4)

Proof. Consider the instantaneous regret $r_t = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J^\star$ and let $\widetilde{R}_T = \sum_{t=1}^{T} r_t \mathbb{I}\{\mathcal{E}_t\}$. We will bound $\widetilde{R}_T$ with high probability, and due to Lemma 8 this would imply a high-probability bound on R_T from which the theorem would follow.

To bound the random variable $\widetilde{R}_T$, we appeal to Lemma 7. The lemma requires that, at round s, the confidence matrix V_s is well-conditioned. Indeed, assuming $\mathcal{E}_t$ holds, then on the one hand $V_s \succeq \lambda I$ for $\lambda \ge (10\nu\vartheta/\alpha_0\sigma^2)\sqrt{T}$, and on the other hand, $\|V_s\| \le \lambda + \beta^{-1} \sum_{r=1}^{s-1} \|z_r\|^2 \le T + 2T \le 4T$ thanks to Lemma 29. Now, for any time t, let $\tau(t)$ denote the last time before round t in which Algorithm 1 updated its policy, so that $A_t = A_{\tau(t)}$, $B_t = B_{\tau(t)}$, $K_t = K_{\tau(t)}$ and $P_t = P_{\tau(t)}$ for all t. Lemma 7 then implies

$$Q + K_t^\mathsf{T} R K_t = P_t - (A_t + B_t K_t)^\mathsf{T} P_t (A_t + B_t K_t) + \mu \|P_t\|_* \left(\begin{smallmatrix} I \\ K_t \end{smallmatrix}\right)^\mathsf{T} V_{\tau(t)} \left(\begin{smallmatrix} I \\ K_t \end{smallmatrix}\right)\ .$$

On the other hand, as $u_t = K_t x_t$ and $J^\star \ge \sigma^2 \|P_t\|_*$, we have

$$r_t = x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J^\star \le x_t^\mathsf{T} (Q + K_t^\mathsf{T} R K_t) x_t - \sigma^2 \|P_t\|_*.$$

Thus, given that $\mathcal{E}_t$ holds,

$$\begin{aligned}
r_t &\le x_t^\mathsf{T} P_t x_t - x_t^\mathsf{T} \big(A_t + B_t K_t\big)^\mathsf{T} P_t \big(A_t + B_t K_t\big) x_t + \mu \|P_t\|_* z_t^\mathsf{T} V_{\tau(t)}^{-1} z_t - \sigma^2 \|P_t\|_* \\
&= x_t^\mathsf{T} P_t x_t - x_{t+1}^\mathsf{T} P_t x_{t+1} \\
&\quad + x_t^\mathsf{T} \big(A_\star + B_\star K_t\big)^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) x_t - x_t^\mathsf{T} \big(A_t + B_t K_t\big)^\mathsf{T} P_t \big(A_t + B_t K_t\big) x_t \\
&\quad + w_t^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) x_t \\
&\quad + w_t^\mathsf{T} P_t w_t - \sigma^2 \|P_t\|_* \\
&\quad + \mu \|P_t\|_* (z_t^\mathsf{T} V_{\tau(t)}^{-1} z_t)\ .
\end{aligned}$$

Lemma 24 now gives

$$x_t^\mathsf{T}\big(A_\star + B_\star K_t\big)^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) x_t - x_t^\mathsf{T}\big(A_t + B_t K_t\big)^\mathsf{T} P_t \big(A_t + B_t K_t\big) x_t \le \mu \|P_t\|_* z_t^\mathsf{T} V_{\tau(t)}^{-1} z_t \ ,$$

and since Algorithm 1 maintains that $\det(V_t) \le 2\det(V_{\tau(t)})$, we have $z_t^\mathsf{T} V_{\tau(t)}^{-1} z_t \le 2 z_t^\mathsf{T} V_t^{-1} z_t$ as a result of Lemma 27. This, along with the fact that $\|P_t\|_* \le \nu/\sigma^2$ on $\mathcal{E}_t$ (recall Lemma 7), yields

$$\begin{aligned}
\widetilde{R}_T \le \ & \sum_{t=1}^{T} \big(x_t^\mathsf{T} P_t x_t - x_{t+1}^\mathsf{T} P_t x_{t+1}\big) \mathbb{I}\{\mathcal{E}_t\} \\
& + \sum_{t=1}^{T} w_t^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) x_t \mathbb{I}\{\mathcal{E}_t\} \\
& + \sum_{t=1}^{T} \big(w_t^\mathsf{T} P_t w_t - \sigma^2 \|P_t\|_*\big) \mathbb{I}\{\mathcal{E}_t\} \\
& + \frac{4\nu\mu}{\sigma^2} \sum_{t=1}^{T} \big(z_t^\mathsf{T} V_t^{-1} z_t\big) \mathbb{I}\{\mathcal{E}_t\} \ .
\end{aligned}$$

The theorem now follows by plugging in the bounds of Lemmas 10 to 13, using a union bound to bound the failure probability, and applying some algebraic simplifications. □

B.2 Proof of Corollary 5

Proof. First, let us show that if Theorem 20 holds then the initial conditions of Theorem 4 are satisfied. Indeed, using $V \succeq V_0 \succeq (T_0\sigma^2/80) I$ gives

$$\|\Delta_0\|_F \le 40 n \sqrt{\frac{\log(T_0/\delta)}{T_0}} \ ,$$

which, by our choice of T_0, is at most $\frac{\alpha_0^5 \sigma^{10}}{2^{13}\nu^5 \vartheta \sqrt{T}}$. This means that the conditions of Theorem 4 hold. Now, by a union bound, with probability at least $1-\delta$ Theorems 4 and 20 and Lemma 6 hold, each with probability at least $1-\delta/3$. Then the regret of this procedure is

$$\begin{aligned}
\sum_{t=1}^{T_0} \big(x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t - J^\star\big) &\le \left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right) \bullet V_0 \\
&\le \alpha_1 T_0 \frac{300\sigma^2\kappa_0^4}{\gamma_0^2} \big(n + k\vartheta^2\kappa_0^2\big) \log\frac{T_0}{\delta} \\
&\le \frac{2^{35}\alpha_1 n^2 \nu^5 \vartheta \kappa_0^4}{\alpha_0^5 \sigma^8 \gamma_0^2} \big(n + k\vartheta^2\kappa_0^2\big) \sqrt{T} \log^2\frac{T}{\delta} \ ,
\end{aligned}$$

and regret on the remaining rounds is bounded by virtue of Theorem 4. □

B.3 Proof of Lemma 6

Proof. Denote $\Theta_\star = (A_\star \, B_\star)$ and $\Theta_t = (A_t \, B_t)$. Note that the solution to the least-square estimate is given as:

$$\Theta_t = \left(\lambda\Theta_0 + \frac{1}{\beta}\sum_{s=1}^{t-1} x_{s+1} z_s^\mathsf{T}\right) V_t^{-1} \ . \tag{10}$$

Plugging $x_{s+1} = \Theta_\star z_s + w_s$ into Eq. (10) and denoting $S_t = \sum_{s=1}^{t} w_s z_s^\mathsf{T}$, we have

$$\Theta_t = \Theta_\star \cdot \frac{1}{\beta}\sum_{s=1}^{t-1} z_s z_s^\mathsf{T} V_t^{-1} + \frac{1}{\beta} S_t V_t^{-1} + \lambda \Theta_0 V_t^{-1} = \Theta_\star + \frac{1}{\beta} S_t V_t^{-1} + \lambda \Delta_0 V_t^{-1} ,$$

whence

$$\begin{aligned}\operatorname{Tr}(\Delta_t V_t \Delta_t^\mathsf{T}) &\le \frac{2}{\beta^2}\operatorname{Tr}(S_t V_t^{-1} S_t^\mathsf{T}) + 2\lambda^2 \operatorname{Tr}(\Delta_0 V_t^{-1}\Delta_0^\mathsf{T}) \\ &\le \frac{2}{\beta^2}\operatorname{Tr}(S_t V_t^{-1} S_t^\mathsf{T}) + 2\lambda\|\Delta_0\|_\mathsf{F}^2. && (\because\ V_t \succeq \lambda I)\end{aligned}$$

To get the result we need to bound the first term. Denote $S_t(i) = \sum_{s=1}^{t-1} w_s(i) z_s$ for all $i = 1, \ldots, d$. For each i, applying Theorem 23 yields that, with probability at least $1 - \delta/d$,

$$S_t(i)^\mathsf{T} V_t^{-1} S_t(i) \le 2\sigma^2\beta \log\left(\frac{d}{\delta}\frac{\det(V_t)}{\det(\beta V_1)}\right) .$$

By additionally applying a union bound, the above holds with probability at least $1-\delta$ for all $i = 1, \ldots, d$ simultaneously, and then

$$\operatorname{Tr}(S_t V_t^{-1} S_t^\mathsf{T}) = \sum_{i=1}^{d} S_t(i)^\mathsf{T} V_t^{-1} S_t(i) \le 2\sigma^2 \beta d \log\left(\frac{d}{\delta}\frac{\det(V_t)}{\det(\beta V_1)}\right) .$$

Plugging this to the inequality above, and using $\beta \ge 1$, gives the main statement of the lemma.

To show $\operatorname{Tr}(\Delta_t V_t \Delta_t) \le 1$ under the conditions of Algorithm 1, note that $\|\Delta_0\|_F^2 \le \epsilon \le 1/4\lambda$ by assumption. Thus it remains to prove $\log\left(\frac{d}{\delta}\frac{\det(V_t)}{\det(V_1)}\right) \le \frac{\beta}{8\sigma^2 d}$. Indeed, in view of Lemma 28 and the definition of β:

$$\begin{aligned}\log\left(\frac{d}{\delta}\frac{\det(V_t)}{\det(V_1)}\right) &\le \log\frac{d}{\delta} + 2n\log T \\ &\le 4n\log\frac{T}{\delta} && (T \ge d) \\ &\le \frac{\beta}{8\sigma^2 d} . && \square\end{aligned}$$

B.4 Proof of Lemma 7

Proof. To prove the lemma, we aim to apply Lemmas 15 and 16.

First, we show that $\mu \ge 1 + 2\vartheta\|V_t\|$. Indeed, assuming $\|V_t\| \le 4T$ and $T \ge \vartheta^{-2}$ implies that

$$1 + 2\vartheta\|V_t\| \le \vartheta\sqrt{T} + 2\vartheta\sqrt{4T} \le 5\vartheta\sqrt{T} = \mu .$$

Consequently, Lemma 15 gives item (i) as well as that the dual solution of the SDP, P_t, is bounded as $\|P_t\|_* \le \nu/\sigma^2$.

Next, note that $V_t \succeq \lambda I$ as well as $\lambda \ge \nu\mu/\alpha_0\sigma^2$, where we have used the fact that $\nu \ge J^\star \ge P^\star \bullet W \ge \left(\begin{smallmatrix} Q & 0 \\ 0 & R\end{smallmatrix}\right) \bullet W \ge \alpha_0\sigma^2$. This gives $V_t \succeq (\nu\mu/\alpha_0\sigma^2)I$. Thus we apply Lemma 16 that shows item (ii).

To show item (iii), we have that P_t is positive semi-definite immediately from the dual formulation of the SDP (7). Moreover, notice that Lemma 16 also gives

$$P_t = Q + K_t^\mathsf{T} P_t K_t + (A_t + B_t K_t)^\mathsf{T} P_t (A_t + B_t K_t) - \mu\|P_t\|_* \left(\begin{smallmatrix} I \\ K_t\end{smallmatrix}\right) V_t^{-1} \left(\begin{smallmatrix} I \\ K_t\end{smallmatrix}\right)^\mathsf{T} ,$$

which we link with the true parameters $(A_\star\ B_\star)$ by combining the equation with Lemma 24. $\square$

B.5 Proof of Lemma 8

Proof. With probability at least $1-\delta/2$, Lemma 6 holds. Also, for any $t = 1, \ldots, T$ with probability at least $1 - \delta/2T$, by the Hanson-Wright concentration inequality (Theorem 21),

$$\|w_t\| \le 5\sigma\sqrt{d \log \frac{2T}{\delta}} \le 10\sigma\sqrt{d \log \frac{T}{\delta}} \ ,$$

as $T \ge 2$. Thus, via a union bound, both statements hold simultaneously with probability $1 - \delta$.

Next, we show by induction on t that $\mathrm{Tr}(\Delta_t V_t \Delta_t^{\mathsf{T}}) \le 1$ and $\|V_t\| \le 4T$. This will particularly ensure that the policies generated by Algorithm 1 are sequentially strongly-stable which will give us $\|z_t\|^2 \le 4\kappa^4 e^{-\gamma(t-1)}\|x_1\|^2 + \beta$ for all $t = 1, \ldots, T$.

For the base case, $t = 1$, we have by assumption

$$\mathrm{Tr}(\Delta_0 V_1 \Delta_0^{\mathsf{T}}) = \lambda\|\Delta_0\|_F^2 \le \frac{1}{4} \le 1 \ , \quad \|V_1\| = \lambda \le 4T \ , \qquad (T \ge \tfrac{2^{30}\nu^{10}\vartheta^2}{\alpha_0^{10}\sigma^{20}} \implies T \ge \lambda)$$

and by definition of β:

$$\|z_1\|^2 \le 2\kappa^2\|x_1\|^2 \le 2\beta T \ .$$

Now, assume that for all $s = 1, \ldots, t-1$, $\mathrm{Tr}(\Delta_s V_s \Delta_s^{\mathsf{T}}) \le 1$ and $\|V_s\| \le 4T$. We show that $\mathrm{Tr}(\Delta_t V_t \Delta_t^{\mathsf{T}}) \le 1$ and $\|V_t\| \le 4T$.

To that end we first show that $\|z_s\|^2 \le 4\kappa^4 e^{-(t-1)\gamma}\|x_1\|^2 + \beta$ for all $s = 1, \ldots, t$. Indeed, by $V_t \succeq \lambda I \succeq 16\kappa^{10}\mu I$, Lemma 9 implies that policies generated by Algorithm 1 up to round t form a (κ, γ)-strongly stable sequence for $\kappa = \sqrt{2\nu/\alpha_0\sigma^2}$ and $\gamma = \frac{1}{2}\kappa^{-2}$. Consequently, Lemma 3 yields for all $s = 1, \ldots, t$

$$\|x_t\| \le \kappa e^{-\gamma(t-1)/2}\|x_1\| + \frac{20\kappa}{\gamma}\sigma\sqrt{d \log \frac{T}{\delta}} \ ,$$

which entails that

$$\begin{aligned}
\|z_s\|^2 &\le 2\kappa^2\|x_s\|^2 && (z_s = \left(\begin{smallmatrix} I \\ K_s \end{smallmatrix}\right)x_s,\ \|\left(\begin{smallmatrix} I \\ K_s \end{smallmatrix}\right)\|^2 \le 2\kappa^2) \\
&\le 2\kappa^2\left(2\kappa^2 e^{-\gamma(t-1)}\|x_1\|^2 + \frac{800\sigma^2\kappa^2 d}{\gamma^2}\log\frac{T}{\delta}\right) \\
&= 4\kappa^4 e^{-\gamma(t-1)}\|x_1\|^2 + \beta \ . && (\gamma = \tfrac{1}{2}\kappa^{-2},\ \kappa = \sqrt{2\nu/\alpha_0\sigma^2})
\end{aligned}$$

In particular, $\sum_{s=1}^{t}\|z_s\|^2 \le 2\beta T$ in view of Lemma 29. This, along with assuming $T \ge \lambda$, immediately gives

$$\|V_t\| \le \lambda + \beta^{-1}\sum_{s=1}^{t-1}\|z_s\|^2 \le 4T.$$

Finally, as we've shown $\sum_{s=1}^{t}\|z_s\|^2 \le 2\beta T$, Lemma 6 additionally provides $\mathrm{Tr}(\Delta_t V_t \Delta_t^{\mathsf{T}}) \le 1$. ☐

B.6 Proof of Lemma 9

Proof. The proof follows by applying Theorem 17 over the sequence $K_1, \ldots, K_T$ of policies generated by Algorithm 1. To that end, define $\tau(t)$ as the last round before t in which Algorithm 1 updates its policy. Note that each policy K_t is associated with A_t, B_t, P_t and $V_{\tau(t)}$.

Thus, to apply Theorem 17 it suffice to show that $\mu \ge 1 + 2\vartheta\|V_t\|^{1/2}$ and $V_t \succeq 16\kappa^{10}\mu I$ for all rounds $t \ge 1$. Indeed, as we assume $T \ge \vartheta^{-2}$ and $\|V_t\| \le 4T$, we have

$$1 + 2\vartheta\|V_t\| \le \vartheta\sqrt{T} + 2\vartheta\sqrt{4T} \le 5\vartheta\sqrt{T} = \mu \ .$$

Furthermore, using $\kappa = \sqrt{2\nu/\alpha_0\sigma^2}$, we have $V_t \succeq \lambda I$ where $\lambda \ge 2^9\nu^5 \cdot 5\vartheta\sqrt{T}/\alpha_0^5\sigma^{10} = 16\kappa^{10}\mu$ as required. ☐

B.7 Proof of Lemma 10

Proof. Let N the last round t such that $\mathcal{E}_t$ holds. Let $\tau_1 < \cdots < \tau_M$ be the time instances in which Algorithm 1 changes policy up to round N, and let $\tau_0 = 1$, $\tau_{M+1} = N+1$. By Lemma 28, as $\mathcal{E}_N$ holds,

$$M = \left\lfloor \log_2 \frac{\det(V_N)}{\det(V_1)} \right\rfloor \le 2n \log T \ .$$

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \big(x_t^\mathsf{T} P_t x_t - x_{t+1}^\mathsf{T} P_t x_{t+1}\big)\mathbb{I}\{\mathcal{E}_t\} &= \sum_{t=1}^{N} \big(x_t^\mathsf{T} P_t x_t - x_{t+1}^\mathsf{T} P_t x_{t+1}\big) \\
&= \sum_{i=0}^{M} x_{\tau_i}^\mathsf{T} P_{\tau_i} x_{\tau_i} - x_{\tau_{i+1}-1}^\mathsf{T} P_{\tau_i} x_{\tau_{i+1}-1} \\
&\le \sum_{i=0}^{M} x_{\tau_i}^\mathsf{T} P_{\tau_i} x_{\tau_i} \ .
\end{aligned}$$

Since $\|P_t\|_* \le \nu/\sigma^2$ and $\|x_t\|^2 \le \|z_t\|^2 \le 4\kappa^4\|x_1\|^2 + \beta$ on $\mathcal{E}_t$, we can bound

$$\begin{aligned}
\sum_{i=0}^{M} x_{\tau_i}^\mathsf{T} P_{\tau_i} x_{\tau_i} &\le \sum_{i=0}^{M} \|P_{\tau_i}\| \|x_{\tau_i}\|^2 \\
&\le (1+M)\cdot \frac{\nu}{\sigma^2} \cdot (32\kappa^8\|x_1\|^2 + \beta) \\
&\le \frac{4\nu(4\kappa^4\|x_1\|^2 + \beta)}{\sigma^2} n \log T \ ,
\end{aligned}$$

and the lemma follows. □

B.8 Proof of Lemma 11

The lemma would follow directly from the following.

Lemma 30. *Let $\delta \in (0,1)$. Let $(\mathcal{F}_t)_{t=1}^\infty$ be a filtration. Let $w_1, w_2, \ldots \sim \mathcal{N}(0, \sigma^2 I)$ be i.i.d Gaussian random variables. Let $v_1, v_2, \ldots$ be a sequence of vectors such that v_t is $\mathcal{F}_{t-1}$-measurable and $\sum_{t=1}^T \|v_t\|^2 \le D^2$ almost surely for each t. Then with probability $1-\delta$,*

$$\sum_{t=1}^{T} v_t^\mathsf{T} w_t \le 2\sigma D \sqrt{\log \frac{2}{\delta}} \ .$$

Proof. Denote $Y_t = v_t^\mathsf{T} w_t$. Note that, conditioned on the randomness before round t, each Y_t is a zero-mean Gaussian random variable. Thus we can write $Y_t = \eta_t m_t$, where m_t^2 is the variance of Y_t given $\mathcal{F}_{t-1}$, and $\eta_t \sim \mathcal{N}(0,1)$.

Let $\lambda > 0$. Using the observation above, we apply Theorem 23 with $V = \lambda$ to obtain that with probability $1-\delta$

$$\frac{\big(\sum_{t=1}^T \eta_t m_t\big)^2}{\lambda + \sum_{t=1}^T m_t^2} \le 2 \log \frac{1 + \lambda^{-1}\sum_{t=1}^T m_t^2}{\delta} \ . \tag{11}$$

We now proceed by upper bounding $\sum_{t=1}^T m_t^2$. The variance of Y_t given $\mathcal{F}_{t-1}$ is:

$$\sum_{t=1}^T m_t^2 = \sigma^2 \sum_{t=1}^T \|v_t\|^2 \le \sigma^2 D^2 \ .$$

Set $\lambda = \sigma^2 D^2$. Plugging the bound above into Eq. (11) and rearranging gets us this lemma's statement. $\square$

of Lemma 11. Apply Lemma 30 with $v_t = P_t\big(A_\star\ B_\star\big) z_t \mathbb{I}\{\mathcal{E}_t\}$ and failure probability $\frac{\delta}{2}$. Note that we have $\sum_{t=1}^T \|v_t\|^2 \le \vartheta^2(\nu/\sigma^2)^2 2\beta T$. We obtain the bound

$$\sum_{t=1}^T w_t^\mathsf{T} P_t\big(A_\star + B_\star K_t\big) x_t \mathbb{I}\{\mathcal{E}_t\} \le \frac{\nu\vartheta}{\sigma}\sqrt{3\beta T \log\frac{4}{\delta}} \ .$$

$\square$

B.9 Proof of Lemma 12

The lemma is an immediate consequence of the following.

Lemma 31. *Let $(\mathcal{F}_t)_{t=1}^\infty$ be a filtration, and let $M_1, M_2, \ldots$ be a sequence of symmetric positive semi-definite matrices such that M_t is $\mathcal{F}_{t-1}$-measurable and $\|M_t\|_* \le D$ almost surely for each t. Further, let $w_1, w_2, \ldots \sim \mathcal{N}(0, \sigma^2 I)$ be a sequence of i.i.d. Gaussian random variables. Then for $T \ge 2$ and for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$\sum_{t=1}^T \big(w_t^\mathsf{T} M_t w_t - \sigma^2\|M_t\|_*\big) \le 8D\sigma^2\sqrt{T\log^3\frac{4T}{\delta}} \ .$$

Proof. Define the random variables $X_t = w_t^\mathsf{T} M_t w_t - \sigma^2\|M_t\|_*$ for all $t \ge 1$. Observe that

$$\mathbb{E}_t[X_t] = M_t \bullet \mathbb{E}_t[w_t^\mathsf{T} w_t - \sigma^2 I] = 0.$$

That is, $\{X_t\}_{t\ge1}$ is a martingale difference sequence with respect to the filtration. Moreover, $X_t \ge -\sigma^2\|M_t\|_* \ge -\sigma^2 D$ for all t with probability one, However, X_t is not bounded from above almost surely. Therefore, consider the truncated random variables $\widetilde{X}_t = X_t\mathbb{I}\{X_t \le \Gamma\}$ with threshold $\Gamma = 5D\sigma^2\log(4T/\delta)$. By Azuma's inequality (Theorem 22), we have with probability at least $1-\delta/2$ that

$$\sum_{t=1}^T \widetilde{X}_t - \sum_{t=1}^T \mathbb{E}_t[\widetilde{X}_t] \le \Gamma\sqrt{2T\log\frac{1}{\delta}} \le 8\sigma^2 D\sqrt{T\log^3\frac{4T}{\delta}}.$$

On the other hand, let us show that $\widetilde{X}_t = X_t$ for all t with probability at least $1-\delta/2$. Indeed, note that $\mathbb{E}_t[w_t^\mathsf{T} M_t w_t] = \sigma^2\|M_t\|_*$, and using the Hanson-Wright inequality (Theorem 21), for any fixed t and any $\delta' \in (0, 1/e)$ we have

$$w_t^\mathsf{T} M_t w_t \le \sigma^2\|M_t\|_* + 4\sigma^2\|M_t^{1/2}\|\|M_t^{1/2}\|_\mathsf{F}\log(1/\delta') \le 5\sigma^2\|M_t\|_*\log(1/\delta') \tag{12}$$

with probability at least $1-\delta'$. Since $\|M_t\|_* \le D$, this implies that $X_t \le \Gamma$ for all t with probability at least $1-\delta/2T$, which in turn means that $\widetilde{X}_t = X_t$ for all t with the same probability.

Finally, since $\sum_{t=1}^T \mathbb{E}_t[\widetilde{X}_t] \le \sum_{t=1}^T \mathbb{E}_t[X_t] = 0$ (as $\widetilde{X}_t \le X_t$ for all t), we obtain that with probability at least $1-\delta$,

$$\sum_{t=1}^T \big(w_t^\mathsf{T} M_t w_t - \sigma^2 \|M_t\|_*\big) = \sum_{t=1}^T X_t = \sum_{t=1}^T \widetilde{X}_t \le 8\sigma^2 D \sqrt{T \log^3 \frac{4T}{\delta}}. \qquad \square$$

of Lemma 12. Apply Lemma 31 with failure probability $\delta/2$ and define $M_t = P_t \mathbb{I}\{\mathcal{E}_t\}$, and note that $\|P_t\|_* \le \nu/\sigma^2$ on $\mathcal{E}_t$. $\square$

B.10 Proof of Lemma 13

Proof. Note that for any t we have on $\mathcal{E}_t$ that

$$\frac{1}{\beta} z_t V_t^{-1} z_t \le \frac{1}{\beta} z_t V_1^{-1} z_t = \frac{1}{\beta\lambda}\|z_t\|^2 \le \frac{1+\|x_1\|^2}{\vartheta\sqrt{T}} \le 1 \ ,$$

using $\lambda \ge \vartheta\sqrt{T}$ as $\kappa \ge 1$, and $T \ge \vartheta^{-2}(1+\|x_1\|^2)^2$. Let N the last round t such that $\mathcal{E}_t$ holds, then

$$\begin{aligned}
\sum_{t=1}^T z_t^\mathsf{T} V_t^{-1} z_t \mathbb{I}\{\mathcal{E}_t\} &= \beta \sum_{t=1}^N \frac{1}{\beta} z_t^\mathsf{T} V_t^{-1} z_t \\
&\le \beta \sum_{t=1}^N 2 \log \frac{\det(V_{t+1})}{\det(V_t)} && \text{(Lemma 26)} \\
&= 2\beta \log \frac{\det(V_{N+1})}{\det(V_1)} \\
&\le 4\beta n \log T \ . && \text{(Lemmas 28 and 29)}
\end{aligned}$$

$\square$

C Proofs of Section 5

C.1 Proof of Lemma 14

Proof. We have

$$\|(X+\Delta)\Sigma(X+\Delta)^\mathsf{T} - X\Sigma X^\mathsf{T}\| \le \Sigma \bullet \big((X+\Delta)^\mathsf{T}(X+\Delta) - X^\mathsf{T} X\big) \ ,$$

and by Lemma 24,

$$(X+\Delta)^\mathsf{T}(X+\Delta) - X^\mathsf{T} X \preceq \mu V^{-1} \ . \qquad \square$$

C.2 Proof of Lemma 18

Proof. Lemmas 16 and 24 imply

$$P_t \succeq Q + K_t^\mathsf{T} R K_t + \big(A_\star + B_\star K_t\big)^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) - 2\mu\|P_t\|_* \big(\begin{smallmatrix} I \\ K_t \end{smallmatrix}\big)^\mathsf{T} V^{-1} \big(\begin{smallmatrix} I \\ K_t \end{smallmatrix}\big).$$

Now, recall that, by assumption, $V \succeq 2\kappa^2\mu I$. Therefore, by Lemma 14,

$$\mu\|P_t\|_* V^{-1} \preceq \mu \cdot \frac{\nu}{\sigma^2} \cdot \frac{1}{2\kappa^2\mu} I = \frac{\nu}{2\sigma^2\kappa^2} I = \frac{\alpha_0}{4} I \ .$$

Hence, as $Q \succeq \alpha_0 I$ and $R \succeq \alpha_0 I$,

$$\begin{aligned} P_t &\succeq \tfrac{1}{2}\alpha_0 I + \tfrac{1}{2}\alpha_0 K_t^\mathsf{T} K_t + \big(A_\star + B_\star K_t\big)^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) \\ &\succeq \tfrac{1}{2}\alpha_0 I + \big(A_\star + B_\star K_t\big)^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) \ . \end{aligned} \tag{13}$$

In particular, this shows that $P_t \succeq \frac{1}{2}\alpha_0 I$. Further, using again the fact that $\|P_t\|_* \leq \nu/\sigma^2$ (Lemma 15) to bound $P_t - \frac{1}{2}\alpha_0 I \preceq (1-\kappa^{-2})P_t$ and rearranging yields

$$P_t^{-1/2}\big(A_\star + B_\star K_t\big)^\mathsf{T} P_t \big(A_\star + B_\star K_t\big) P_t^{-1/2} \preceq (1-\kappa^{-2}) I \ .$$

Letting $H_t = P_t^{1/2}$ and $L_t = P_t^{-1/2}\big(A_\star + B_\star K_t\big)P_t^{1/2}$, we have established that $\|L_t\| \leq \sqrt{1-\kappa^{-2}} \leq 1 - \frac{1}{2}\kappa^{-2}$, as well as $\|H_t\| \leq \sqrt{\nu/\sigma^2}$ and $\|H_t^{-1}\| \leq \sqrt{2/\alpha_0}$. To bound the norm of K_t, observe that Eq. (13) implies $P_t \succeq \frac{1}{2}\alpha_0 K_t^\mathsf{T} K_t$ hence

$$\|K_t\| \leq \sqrt{\frac{2}{\alpha_0}\|P\|} \leq \sqrt{\frac{2\nu}{\alpha_0\sigma^2}} = \kappa \ .$$

As $A_\star + B_\star K_t = H_t L_t H_t^{-1}$, this shows that K_t is (κ, γ)-strongly stable. ☐

C.3 Proof of Lemma 19

Proof. It suffices to show that $P_t \preceq P^\star \preceq P_t + \frac{\alpha_0\gamma}{2} I$ for all $t \geq 1$.

For $P_t \preceq P^\star$, let $K_\star$ denote the optimal policy corresponding to $P^\star$. As $P^\star$ is the solution to the Riccati equation:

$$P^\star = Q + K_\star^\mathsf{T} R K_\star + (A_\star + BK_\star)^\mathsf{T} P^\star (A_\star + BK_\star). \tag{Eq. (3)}$$

On the other hand, applying Lemma 24 over Eq. (7) gives

$$\left(\begin{smallmatrix} Q-P_t & 0 \\ 0 & R \end{smallmatrix}\right) + \big(A_\star \ B_\star\big)^\mathsf{T} P_t \big(A_\star \ B_\star\big) \succeq 0 \ ,$$

which particularly implies

$$P_t \preceq Q + K_\star^\mathsf{T} R K_\star + (A_\star + B_\star K_\star)^\mathsf{T} P_t (A_\star + B_\star K_\star).$$

Subtracting the two inequalities gets us

$$P_t - P^\star \preceq \big(A_\star + B_\star K_\star\big)^\mathsf{T} \big(P - P^\star\big) \big(A_\star + B_\star K_\star\big) \ ,$$

and, as $K_\star$ is a (strongly) stable policy, Lemma 25 implies $P - P^\star \preceq 0$.

For the converse inequality, Eq. (2) implies

$$P^\star \preceq Q + K^\mathsf{T} R K + \big(A_\star + B_\star K\big)^\mathsf{T} P^\star \big(A_\star + B_\star K\big).$$

On the other hand, combining Lemmas 16 and 24 yields

$$P_t \succeq Q + K^\mathsf{T} R K + \big(A_\star + B_\star K\big)^\mathsf{T} P_t \big(A_\star + B_\star K\big) - 2\mu\|P_t\|_* \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T} V_t^{-1} \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right).$$

Subtracting the two matrix inequalities gets us

$$P^\star - P_t \preceq \big(A_\star + B_\star K\big)^\mathsf{T} \big(P^\star - P_t\big) \big(A_\star + B_\star K\big) + 2\mu\|P_t\|_* \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T} V_t^{-1} \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right) \ .$$

Applying Lemma 25 shows

$$P^\star - P_t \preceq \frac{2\kappa^2\mu}{\gamma}\|P_t\|_* \|\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T} V_t^{-1} \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)\| I \ .$$

Moreover, $\|K\| \le \kappa$ provides $\|\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)\|^2 \le 1+\kappa^2 \le 2\kappa^2$, thus $\|\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T} V_t^{-1} \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)\| \le 2\kappa^2 \|V_t^{-1}\|$. Finally, by Lemma 15 and the lower bound on V_t,

$$\frac{4\kappa^4}{\gamma}\|P_t\|_* \|V_t^{-1}\| \le \frac{4\kappa^4}{\gamma} \cdot \frac{\nu}{\sigma^2} \cdot \frac{1}{16\kappa^{10}\mu} = \frac{\alpha_0 \gamma}{2}$$

where we have used $\kappa = \sqrt{2\nu/\sigma^2\alpha_0}$ and $\gamma = \frac{1}{2}\kappa^{-2}$. ∎

D Proofs of Section 6

D.1 Proof of Theorem 20

We first require the following lemma.

Lemma 32. *Assume $x_1 = 0$. Let $\delta \in (0, 1/e)$. With probability at least $1-\delta$, for all $t = 1, \ldots, T_0+1$ it holds that*

$$\|x_t\| \le \frac{4\sigma\kappa_0}{\gamma_0}\sqrt{\left(d + k\vartheta^2\kappa_0^2\right)\log\frac{T_0}{\delta}} \ .$$

Proof. We begin by upper bounding the norm of x_t using the strong stability of K_0. Let $u_t = K_0 x_t + \eta_t$ where $\eta_t \sim \mathcal{N}(0, 2\sigma^2\kappa_0^2 I)$. We have,

$$x_{t+1} = (A_\star + B_\star K_0)x_t + B_\star \eta_t + w_t \ ,$$

and, as η_t is independent of x_t, we can think about the state transitions as if they are done according the another LQR system that is exactly the same as the original one except that the noise term is now $B_\star\eta_t + w_t$ instead of w_t. Thus, applying Lemma 3:

$$\|x_t\| \le \frac{\kappa_0}{\gamma_0}\max_{0\le s\le t-1}\|B_\star\eta_s + w_s\| \ .$$

Next, $B_\star\eta_s + w_s$ is a Gaussian random variable with zero mean and covariance $C = 2\sigma^2\kappa_0^2 B_\star B_\star^\mathsf{T} + \sigma^2 I$. Using the Hanson-Wright inequality (Theorem 21) and a union bound, with probability $1-\delta$, for all $t = 1, \ldots, T_0+1$,

$$\begin{aligned}
\|B_\star\eta_t + w_t\|^2 &\le 5\,\mathrm{Tr}(C)\log(T_0/\delta) \\
&= 5\sigma^2\left(d + 2\kappa_0^2\|B_\star\|_F^2\right)\log(T_0/\delta) \\
&\le 10\sigma^2\left(d + k\kappa_0^2\vartheta^2\right)\log(T_0/\delta) \ .
\end{aligned}$$ ∎

For the lower bound, we also require the next lemma.

Lemma 33. *Let $\delta \in (0,1)$, and let $n \in \mathbb{R}^n$ be any unit vector. Suppose that $T_0 \ge 200\log(1/\delta)$. Then with probability at least $1-\delta$ we have $n^\mathsf{T} V n \ge T_0 \cdot \sigma^2/40$.*

The proof relies on a couple of technical results. In what follows, we let $(\mathcal{F}_t)_{t=1}^\infty$ be the filtration with respect to which $\{w_t, u_t\}_{t=1}^\infty$ is adapted.

Lemma 34. *For all t we have $\mathbb{E}_t\big[z_t z_t^\mathsf{T} \mid \mathcal{F}_{t-1}\big] \succeq (\sigma^2/2)I$.*

Proof. Note that since $W = \sigma^2 I$ we have $\mathbb{E}[x_t x_t^\mathsf{T} \mid \mathcal{F}_{t-1}] \succeq \sigma^2 I$ for each $t \geq 1$, and so

$$
\begin{aligned}
\mathbb{E}\big[z_t z_t^\mathsf{T} \mid \mathcal{F}_{t-1}\big] &= \left(\begin{smallmatrix} I \\ K_0 \end{smallmatrix}\right) \mathbb{E}\big[x_t x_t^\mathsf{T} \mid \mathcal{F}_t\big] \left(\begin{smallmatrix} I \\ K_0 \end{smallmatrix}\right)^\mathsf{T} + \left(\begin{smallmatrix} 0 & 0 \\ 0 & 2\sigma^2\kappa_0^2 I \end{smallmatrix}\right) \\
&\succeq \sigma^2 \left(\begin{smallmatrix} I & K_0^\mathsf{T} \\ K_0 & K_0 K_0^\mathsf{T} + 2\kappa_0^2 I \end{smallmatrix}\right) \\
&\succeq \sigma^2 \left(\begin{smallmatrix} I & K_0^\mathsf{T} \\ K_0 & \frac{1}{2} I + 2K_0 K_0^\mathsf{T} \end{smallmatrix}\right) && (\|K_0\| \leq \kappa_0,\ \kappa_0 \geq 1) \\
&= \frac{\sigma^2}{2} I + \sigma^2 \left(\begin{smallmatrix} \sqrt{2}^{-1} I \\ \sqrt{2} K_0 \end{smallmatrix}\right) \left(\begin{smallmatrix} \sqrt{2}^{-1} I \\ \sqrt{2} K_0 \end{smallmatrix}\right)^\mathsf{T} \\
&\succeq \frac{\sigma^2}{2} I \ .
\end{aligned}
$$

The lemma now follows by taking expectations. ☐

Lemma 35. *Denote $S_t = n^\mathsf{T} z_t$, and let E_t be an indicator random variable that equals 1 if $S_t^2 > \sigma^2/4$ and 0 otherwise. Then $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \geq 1/5$.*

Proof. We have,

$$
\begin{aligned}
\mathbb{P}[E_t = 1 \mid \mathcal{F}_{t-1}] &= \mathbb{P}[S_t^2 > \sigma^2/4 \mid \mathcal{F}_{t-1}] \\
&= \mathbb{P}[|S_t| > \sigma/2 \mid \mathcal{F}_{t-1}] \\
&\geq \mathbb{P}[S_t - \mathbb{E}S_t > \sigma/2 \mid \mathcal{F}_{t-1}] && (S_t - \mathbb{E}S_t \text{ is a symmetric r.v.}) \\
&= \mathbb{P}[\sqrt{\mathrm{Var}[S_t|\mathcal{F}_{t-1}]} Z > \sigma/2] && (Z \text{ is a standard Gaussian r.v.}) \\
&\geq \mathbb{P}[\sqrt{\sigma^2/2} Z > \sigma/2] && (\text{Lemma } 34) \\
&= \mathbb{P}[Z > 1/\sqrt{2}] \\
&\geq 2^{-3/2} e^{-(1/\sqrt{2})^2} && (\text{Standard Gaussian tail lower bound}) \\
&\geq 1/5 \ .
\end{aligned}
$$
☐

of Lemma 33. Let $U_t = E_t - \mathbb{E}_t[E_t \mid \mathcal{F}_{t-1}]$. Then U_t is a martingale difference sequence with $|U_t| \leq 1$ almost surely. Applying Azuma's inequality, we have that with probability at least $1 - \delta$,

$$
\sum_{t=1}^{T_0} U_t \geq -\sqrt{2T_0 \log \frac{1}{\delta}} \geq -\frac{T_0}{10} \ , \qquad (T_0 \geq 200 \log(1/\delta))
$$

which means that, by Lemma 35,

$$
\sum_{t=1}^{T_0} E_t \geq \sum_{t=1}^{T_0} \mathbb{E}_t[E_t] - \frac{T_0}{10} \geq \sum_{t=1}^{T_0} \frac{1}{5} - \frac{T_0}{10} = \frac{T_0}{10} \ .
$$

Now, by definition of E_t, $S_t^2 \geq E_t \cdot \sigma^2/4$. Therefore, with probability at least $1 - \delta$,

$$
n^\mathsf{T} V n = \sum_{t=1}^{T_0} S_t^2 \geq \sum_{t=1}^{T_0} E_t \cdot \sigma^2/4 = \frac{T_0 \sigma^2}{40} \ .
$$
☐

We are now ready to prove the main theorem of this section.

of Theorem 20. We first prove the upper bound. Let $u_t = K_0 x_t + \eta_t$ where $\eta_t \sim \mathcal{N}(0, 2\sigma^2\kappa_0^2 I)$. Then,

$$\mathrm{Tr}(V_0) = \sum_{t=1}^{T_0} \|z_t\|^2 \ ,$$

and, as $\|K_0\| \le \kappa_0$ and $\kappa_0 \ge 1$:

$$\|z_t\| \le \|x_t\| + \|u_t\| \le \|x_t\| + \|K_0 x_t\| + \|\eta_t\| \le 2\kappa_0\|x_t\| + \|\eta_t\| \ .$$

Now, using a union bound, with probability $1-\delta/2$ we have for all $t = 1, \ldots, T_0+1$ by Lemma 32

$$\|x_t\| \le \frac{4\sigma\kappa_0}{\gamma_0}\sqrt{\left(d + k\vartheta^2\kappa_0^2\right)\log(4T_0/\delta)} \ ,$$

and by the Hanson-Wright inequality $\|\eta_t\|^2 \le 10\sigma^2\kappa_0^2 k \log(4T_0/\delta)$ for all t. Therefore,

$$\begin{aligned}
\|z_t\| &\le 2\kappa_0 \cdot \frac{4\sigma\kappa_0}{\gamma_0}\sqrt{\left(d + k\vartheta^2\kappa_0^2\right)\log(4T_0/\delta)} + 4\kappa_0\sigma\sqrt{k\log(4T_0/\delta)} \\
&\le \frac{12\sigma\kappa_0^2}{\gamma_0}\sqrt{\left(n + k\vartheta^2\kappa_0^2\right)\log(4T_0/\delta)} \ .
\end{aligned}$$

We next turn to lower bounding the smallest eigenvalue of V; we will actually prove that $\|V^{-1}\| \le 80/(T_0\sigma^2)$. Let $\mathcal{N}(1/4)$ be a minimal $1/4$-net of $\mathbb{S}^{n-1}$, and define the set $M = \{V^{-1/2}u/\|V^{-1/2}u\| : u \in \mathcal{N}(1/4)\}$. Suppose that $T_0 \ge 200\log(|M|/\delta)$. Applying a union bound, we get that with probability at least $1-\delta$ simultaneously for all $n \in M$

$$n^{\mathsf{T}} V n \ge \frac{T_0\sigma^2}{40} \ . \qquad \text{(Lemma 33)}$$

Using the definition of M, this entails that for all $u \in \mathcal{N}(1/4)$

$$u^{\mathsf{T}} V^{-1} u \le \frac{40}{T_0\sigma^2} \ . \qquad (14)$$

Next, let z be the eigenvector corresponding to the minimum eigenvalue of V, and let $u_z \in \mathcal{N}(1/4)$ be such that $\|z - u_z\| \le 1/4$. Then,

$$\begin{aligned}
\|V^{-1}\| &= z^{\mathsf{T}} V^{-1} z \\
&\le u_z^{\mathsf{T}} V^{-1} u_z + (z-u_z)^{\mathsf{T}} V^{-1}(z+u_z) \\
&\le u_z^{\mathsf{T}} V^{-1} u_z + \|z-u_z\|\|V^{-1}\|\big(\|z\| + \|u_z\|\big) \\
&\le \frac{40}{T_0\sigma^2} + \frac{1}{4}\|V^{-1}\| \cdot 2 \ . && \text{(Eq. (14). } z \text{ and } u_z \text{ are unit vectors)}
\end{aligned}$$

Rearranging gets us $\|V^{-1}\| \le 80/(T_0\sigma^2)$ as required. Note that $|M| = |\mathcal{N}(1/4)|$, and by standard bounds on the size of ϵ-nets, $|\mathcal{N}(1/4)| \le 12^n$. That is, for T_0 to be larger that $200\log(|M|/\delta)$ it suffices to have $T_0 \ge 400(n + \log(1/\delta))$.

To show that a bound on the estimation error of $\big(A_0\ B_0\big)$, set $V = V_0 + \sigma^2\vartheta^{-2} I$, and

$$\big(A_0\ B_0\big) = \left(\sum_{t=1}^{T_0} x_{t+1} z_t^{\mathsf{T}}\right) V^{-1} \ .$$

Applying Lemma 6 with these parameters and $\begin{pmatrix} A_0 & B_0 \end{pmatrix} = 0$, shows that with probability $1 - \delta/2$

$$\begin{aligned}
\operatorname{Tr}(\Delta_0 V \Delta_0^\mathsf{T}) &\leq 4\sigma^2 d \log\left(\frac{d}{\delta} \det\bigl(I + \vartheta^2 \sigma^{-2} V_0\bigr)\right) + 2\sigma^2 \vartheta^{-2} \|\begin{pmatrix} A_\star & B_\star \end{pmatrix}\|_F^2 \\
&\leq 4\sigma^2 d \log\frac{d}{\delta} + 4\sigma^2 dn \log\left(1 + T_0 \cdot \frac{300\vartheta^2\kappa_0^4}{\gamma_0^2}(1 + \vartheta^2\kappa_0^2) \log\frac{T_0}{\delta}\right) + 2\sigma^2 d \\
&\leq 20 n^2 \sigma^2 \log(T_0/\delta) \ ,
\end{aligned}$$

using $\log\det X \leq n \log(\operatorname{Tr}(X)/n)$ for a positive-definite $X \in \mathbb{R}^{n\times n}$, by our choice of T_0 and the lower bound on T. $\square$