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

Alon Cohen [1 2] **Tomer Koren** [3] **Yishay Mansour** [1 4]

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 or its cost might grow arbitrarily.

[1]Google Research, Tel-Aviv [2]Technion—Israel Inst. of Technology [3]Google Brain, Mountain View [4]Tel-Aviv University. Correspondence to: Alon Cohen <alon.cohen@technion.ac.il>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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) 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).

Our work builds upon the foundations laid in Cohen et al. (2018) who study LQ control with fixed and known dynamics and adversarially changing costs. Indeed, both our SDP formulation (Section 2.2) and the definition of strong stability (a quantification of the notion of stability; see Section 2.3) are borrowed from that paper.

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.

Finally, we note that after the submission of this work, Mania et al. (2019) published an alternative approach that could be used to obtain a result similar to ours.

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, $\|\cdot\|_{\mathsf{F}}$ to denote the Frobenius norm, $\|M\|_{\mathsf{F}} = \sqrt{\mathrm{Tr}(M^{\mathsf{T}}M)}$, 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 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

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

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 $x_1^{\mathsf{T}} P x_1 = \sum_{t=1}^{\infty} \left(\mathbb{E}[c_t] - J(\pi) \right)$. 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;

[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 $\|M\| \gg 1$.

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

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,

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

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

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

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

(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):

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

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 $\Sigma^\star = \mathcal{E}(K_\star)$.

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 $\kappa \ge 1$ and $0 < \gamma \le 1$) if there exists matrices $H \succ 0$ and L such that $M = HLH^{-1}$, with $\|L\| \le 1 - \gamma$ and $\|H\|\|H^{-1}\| \le \kappa$.

A policy K for the linear system (1) is (κ, γ)-*strongly stable* (for $\kappa \ge 1$ and $0 < \gamma \le 1$) if $\|K\| \le \kappa$ 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 $0 < \gamma \le 1$) 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) $\|L_t\| \le 1 - \gamma$ and $\|K_t\| \le \kappa$;
(ii) $\|H_t\| \le B_0$ and $\|H_t^{-1}\| \le 1/b_0$ with $\kappa = B_0/b_0$;
(iii) $\|H_{t+1}^{-1} H_t\| \le 1 + \gamma/2$.

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 the full version of the paper (Cohen et al., 2019).

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*

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

Algorithm 1 OSLO: Optimistic Sdp 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 $\|(A_0\ B_0) - (A_\star\ B_\star)\|_F^2 \leq \varepsilon$.
2: **initialize**: $\mu = 5\vartheta\sqrt{T}$, $V_1 = \lambda I$ where
$$\lambda = \frac{2^{11}\nu^5\vartheta\sqrt{T}}{\alpha_0^5\sigma^{10}} \text{ and } \beta = \frac{2^{18}\nu^4 n^2}{\alpha_0^4\sigma^6}\log\frac{T}{\delta}\,.$$
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
$$\frac{1}{\beta}\sum_{s=1}^{t-1} \|(A\ B)\, z_s - x_{s+1}\|^2 + \lambda\|(A\ B) - (A_0\ B_0)\|_F^2$$
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:
$$\begin{aligned} &\min \Sigma \bullet \left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right) \\ &\text{s.t. } \Sigma_{xx} \succeq (A_t\ B_t)\Sigma(A_t\ B_t)^\top + W - \mu(\Sigma \bullet V_t^{-1})I\,, \\ &\quad \Sigma \succeq 0\,. \end{aligned}$$
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 $V_{t+1} = V_t + \beta^{-1} z_t z_t^\top$.
15: **end for**

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 $\varepsilon = O(1/\sqrt{T})$ 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 $z_1, \ldots, z_{t-1}$.

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 $\|(A_0\, B_0)-(A_\star\, B_\star)\|_{\mathsf{F}}^2 \le \varepsilon$ satisfies*

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

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

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

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*

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

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

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

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 the full version of the paper (Cohen et al., 2019).

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 $1-\delta$,*

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

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 boundedness 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 the full version of the paper (Cohen et al., 2019).

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. *Let t be an arbitrary time. 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 inequality involving P_t follows from the complementary slackness conditions of the SDP. This inequality 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. Boundedness 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,

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

where $\kappa = \sqrt{2\nu/\alpha_0\sigma^2}$ and $\gamma = 1/2\kappa^2$. 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 $\ge 1 - \delta/2$.*

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, $\|V_s\| \le 4T$ 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 the full version of the paper (Cohen et al., 2019). 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 boundedness 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 $J^\star \ge \sigma^2 \|P_t\|_*$ (which is also a consequence of Lemma 7), we have

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

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

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

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*

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

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*

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

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

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

Finally, using the elementary identity $z^\mathsf{T} V^{-1} z \le 2 \log(\det(V + zz^\mathsf{T})/\det(V))$ for $V \succ 0$ and any vector z such that $z^\mathsf{T} V^{-1} z \le 1$, 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* $\sum_{t=1}^{T} (z_t^\mathsf{T} V_t^{-1} z_t) \mathbb{I}\{\mathcal{E}_t\} \le 4\beta n \log T$.

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 $\mathrm{Tr}(\Delta V \Delta^\mathsf{T}) \le 1$.

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

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

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

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 the full version of the paper (Cohen et al., 2019). First, we present a matrix-perturbation lemma also proven in the full version of the paper (Cohen et al., 2019).

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 $\mu \ge 1 + 2\|X\|\|V\|^{1/2}$,*

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

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

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 $\mathrm{Tr}(\Delta V \Delta^\mathsf{T}) \le 1$ implies that $\Delta^\mathsf{T}\Delta \preceq V^{-1}$) yields Eq. (6) due to

$$\begin{aligned} \Sigma_{xx}^\star &= \left(A_\star\ B_\star\right)\Sigma^\star\left(A_\star\ B_\star\right)^\mathsf{T} + W \\ &\succeq \left(A\ B\right)\Sigma^\star\left(A\ B\right)^\mathsf{T} + W - \mu\left(\Sigma^\star \bullet V^{-1}\right)I\,. \end{aligned}$$

Therefore, it is feasible for SDP (6). □

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 $\mu \ge 1 + 2\vartheta\|V\|^{1/2}$. 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*

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

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 $\mathrm{rank}(\Sigma) = d$ as this would entail that $\Sigma = \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)\Sigma_{xx}\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\mathsf{T}$ for $K = \Sigma_{ux}\Sigma_{xx}^{-1}$.

Thus the span of Σ is the span of $\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)$ whence $\left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right)^\top Z \left(\begin{smallmatrix} I \\ K \end{smallmatrix}\right) = 0$ 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 $\mathrm{rank}(X) + \mathrm{rank}(Y) \le n$. Then it suffices to show $\Sigma_{xx} \succ 0$ and $\mathrm{rank}(Z) \ge k$. Indeed, using Lemma 15,

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

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

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 $\left(\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right) + \big(A \; B\big)^\top P \big(A \; B\big) - \mu \|P\|_* V^{-1}$, 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 $\le d$. Therefore $\mathrm{rank}(Z) \ge k$. □

We continue with proving the main result of this section that would imply Lemma 9 (see the full version of the paper (Cohen et al., 2019)). 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 $\mu \ge 1 + 2\vartheta \|V_t\|^{1/2}$ 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 the full version of the paper (Cohen et al., 2019) 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 $\|L_t\| \le 1 - \gamma$. 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 the full version of the paper (Cohen et al., 2019) for a proof).

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

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

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

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 $u_t \sim \mathcal{N}(K_0 x_t, 2\sigma^2 \kappa_0^2 I)$ independently; this is summarized in Algorithm 2.

Algorithm 2 Warm-up

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 $u_t \sim \mathcal{N}(K_0 x_t, 2\sigma^2 \kappa_0^2 I)$.
end for

Let $V_0 = \sum_{t=1}^{T_0} z_t z_t^\top$ 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 $\big(A_0 \; B_0\big) = \sum_{t=1}^{T_0 - 1} x_{t+1} z_t^\top V^{-1}$ we have

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

With Theorem 20 in hand, the proof of Corollary 5 readily follows; see details in the full version of the paper (Cohen et al., 2019). 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 the full version of the paper (Cohen et al., 2019).

Acknowledgments

We thank Nevena Lazic, Yoram Singer and Kunal Talwar for many helpful discussions and comments. YM was supported in part by a grant from the Israel Science Foundation (ISF).

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.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

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.

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

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.