Logarithmic Regret for Learning Linear Quadratic Regulators Efficiently

Asaf Cassel [1] **Alon Cohen** [2] **Tomer Koren** [1]

Abstract

We consider the problem of learning in Linear Quadratic Control systems whose transition parameters are initially unknown. Recent results in this setting have demonstrated efficient learning algorithms with regret growing with the square root of the number of decision steps. We present new efficient algorithms that achieve, perhaps surprisingly, regret that scales only (poly)logarithmically with the number of steps in two scenarios: when only the state transition matrix A is unknown, and when only the state-action transition matrix B is unknown and the optimal policy satisfies a certain non-degeneracy condition. On the other hand, we give a lower bound that shows that when the latter condition is violated, square root regret is unavoidable.

1. Introduction

The linear-quadratic regulator model (LQR) is a classic model in optimal control theory. In this model, the dynamics of the environment are given as

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t,$$

where x_t and u_t are the state and the action vectors at time t, $A_\star$ and $B_\star$ are transition matrices, and w_t is a zero-mean i.i.d. Gaussian noise. The cost function is quadratic in both the state and the action. An interesting property of LQR systems is that a linear control policy minimizes the cost while keeping the system at a steady-state (stable) position.

In this work, we study the problem of designing an adaptive controller that regulates the system while learning its parameters. This problem has recently been approached through the lens of regret minimization, beginning in the work of Abbasi-Yadkori and Szepesvári (2011) that established an $O(\sqrt{T})$ regret bound for this setting albeit with a computationally inefficient algorithm. The problem of designing an efficient algorithm that enjoys $O(\sqrt{T})$ was later resolved by Cohen et al. (2019) and Mania et al. (2019). The former work relied on the "optimism in the face of uncertainty" principle and a reduction to an online semi-definite problem, and the latter work used a simpler greedy strategy.

Following this line of work, it has been believed that an $O(\sqrt{T})$ regret is tight for the problem. This appears natural as it is the typical rate for many imperfect information (bandit) optimization problems (e.g., Shamir, 2013).[1] On the other hand, one could suspect that better, polylogarithmic regret bounds, are possible in the LQR setting thanks to the strongly convex structure of the cost functions. Often in optimization, this structure gives rise to faster convergence/regret rates, and indeed, in a recent work, Agarwal et al. (2019b) have demonstrated that such fast rates are attainable in the related, yet full-information online LQR problem endowed with any strongly convex loss functions.

In this paper, we show two interesting scenarios of learning unknown LQR systems in which an expected regret of $O(\log^2 T)$ is, in fact, achievable. In the first, we assume that the matrix $B_\star$ is known and show that polylogarithmic regret can be attained by harnessing the intrinsic noise in the system dynamics for exploration. In the second, we assume that $A_\star$ is known and that the optimal control policy $K_\star$ is given by a full-rank matrix. Both results are attained using simple and efficient algorithms whose runtime per time step is polynomial in the natural parameters of the problem.

We complement our results with a lower bound showing that our assumptions are indeed necessary for obtaining improved regret guarantees. Specifically, we show that when $B_\star$ is unknown and the optimal policy $K_\star$ is near-degenerate (i.e., with very small singular values), any online algorithm, whether efficient or not, must suffer at least $\Omega(\sqrt{T})$ regret. To the best of our knowledge, this is the first $\Omega(\sqrt{T})$ lower bound for learning linear quadratic regulators (that particularly holds even when the learner knows the entire set of system parameters but the matrix $B_\star$, and even in a single-input single-output scenario). Concurrent to this work, Sim-

[1]School of Computer Science, Tel Aviv University [2]Google Research, Tel Aviv. Correspondence to: Asaf Cassel <acassel@mail.tau.ac.il>, Alon Cohen <aloncohen@google.com>, Tomer Koren <tkoren@tauex.tau.ac.il>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]More precisely, this is very often the regret rate in bandit problems with no "gap" assumptions regarding the difference between the best and second-best actions/policies.

chowitz and Foster (2020) suggest a different lower bound construction that relies on uncertainty in both $A_\star$ and $B_\star$, and thus does not contradict our positive findings.

1.1. Setup: Learning in LQR

We consider the problem of regret minimization in the LQR model. At each time step t, a state $x_t \in \mathbb{R}^d$ is observed and action $u_t \in \mathbb{R}^k$ is chosen. The system evolves according to

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t,$$

where the state-state $A_\star \in \mathbb{R}^{d\times d}$ and state-action $B_\star \in \mathbb{R}^{d\times k}$ matrices form the transition model and the w_t are i.i.d. noise terms, each is a zero mean Gaussian with covariance matrix $\sigma^2 I$. At time t, the instantaneous cost is

$$c_t = x_t^T Q x_t + u_t^T R u_t,$$

where $Q, R \succ 0$ are positive definite.

A policy of the learner is a mapping from a state $x \in \mathbb{R}^d$ to an action $u \in \mathbb{R}^k$ to be taken at that state. Classic results in linear control establish that, given the system parameters $A_\star, B_\star, Q$ and R, the optimal policy is a linear mapping from the state space $\mathbb{R}^d$ to the action space $\mathbb{R}^k$ in an infinite-horizon setup. We thus consider policies of the form $u_t = K x_t$ and define the infinite horizon expected cost,

$$J(K) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} x_t^T\big(Q + K^T R K\big)x_t\right],$$

where the expectation is taken with respect to the random noise variables w_t. Let $K_\star = \arg\min_K J(K)$ be an (unique) optimal policy and $J_\star = J(K_\star)$ denote the optimal infinite horizon expected cost, which are both well defined under mild assumptions.[2] We are interested in minimizing the *regret* over T decision rounds, defined as

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

We focus on the setting where the learner does not have a full a-priori description of the transition parameters $A_\star$ and $B_\star$, and has to learn them while controlling the system and minimizing the regret.

Throughout, we assume that the learner has knowledge of the cost matrices Q and R, and that there are constants $\alpha_0, \alpha_1 > 0$ such that

$$\|Q\|, \|R\| \le \alpha_1, \text{ and } \|Q^{-1}\|, \|R^{-1}\| \le \alpha_0^{-1}.$$

We further assume that the learner has bounds on the transition matrices, as well as on the optimal cost; that is, there are known constants $\vartheta, \nu > 0$ such that

$$\|A_\star\|, \|B_\star\| \le \vartheta, \text{ and } J_\star \le \nu.$$

Finally, we assume that there is a known stable (not necessarily optimal) policy K_0 and $\nu_0 > 0$ such that $J(K_0) \le \nu_0$.[3]

1.2. Main results

Our first result focuses on the case where the state-action transition matrix $B_\star$ is known (but the matrix $A_\star$ is unknown).

Theorem 1. *There exists an efficient online algorithm (see Algorithm 1 in Section 3.1) that, given the matrix $B_\star$ as input, has expected regret*

$$\mathbb{E}[R_T] = \mathrm{poly}\big(\alpha_0^{-1}, \alpha_1, \vartheta, \nu, \nu_0, d, k\big)\, O(\log^2 T).$$

Next, we consider the dual setup in which only the state-state matrix $A_\star$ is known. Here we require an additional non-degeneracy assumption for obtaining polylogarithmic regret.

Theorem 2. *Suppose that the optimal policy of the system satisfies $K_\star K_\star^T \succeq \mu_\star I$ for some constant $\mu_\star > 0$ that is unknown to the learner. Then there exists an efficient online algorithm (see Algorithm 2 in Section 3.2) that, given the matrix $A_\star$ as input, has expected regret*

$$\mathbb{E}[R_T] = \mathrm{poly}\big(\mu_\star^{-1}, \alpha_0^{-1}, \alpha_1, \vartheta, \nu, \nu_0, d, k\big)\, O(\log^2 T).$$

Finally, we show that our assumption regarding the non-degeneracy of the optimal policy $K_\star$ is necessary. Our next result shows that without it, the expected regret of any algorithm is unavoidably at least $\Omega(\sqrt{T})$, even in simple one-dimensional (single input, single output) systems.

Theorem 3. *For any learning algorithm and any $\sigma > 0$, there exists an LQR system (in dimensions $d = k = 1$) which is stabilized by the policy $K_0 = 0$ and for which $\alpha_1 = \alpha_0 = 1$, $\vartheta = 1$ and $\nu = 2\sigma^2$, such that the expected regret of the algorithm is at least $\Omega(\sigma^2\sqrt{T})$. This is true even if the algorithm receives the matrix $A_\star$ as input.*

1.3. Discussion

Our results could be interpreted as a proof-of-concept that faster, polylogarithmic rates for learning in LQRs are possible under more limited uncertainty assumptions. This is perhaps surprising in light of the aforementioned work of Shamir (2013), that established $\Omega(\sqrt{T})$ regret lower bounds for online (bandit) optimization, even with quadratic and strongly convex objectives (as is the case in our LQR setup).

[2] These hold under standard, very mild controllablity assumptions (see Bertsekas, 1995) that we implicitly assume throughout.

[3] Regarding the necessity of this assumption, see the discussion in Mania et al. (2019); Cohen et al. (2019).

The questions of whether polylogarithmic regret guarantees are possible under more general uncertainty of the system parameters, as well as whether the squared dependence on $\log T$ is indeed tight, remain open. Our lower bound, however, shows that more assumptions are required for obtaining stronger positive results.

Our results focused on the *expected* regret compared to the *infinite-horizon* performance of the optimal policy $K_\star$. As far as we know, this is the first analysis that bounds the regret in expectation rather than in high-probability (and without additional assumptions, e.g., as in Ouyang et al., 2017). Indeed, in previous analyses we are aware of, there was always a small probability where the algorithm fails and incurs very large (possibly exponentially large) regret. Here, we address this low-probability event by employing a novel "abort procedure" when our algorithms suspect the system has been destabilized; this ensures that the expected regret remains controlled. The question of whether our regret bounds hold with high probability remains for future investigation. We remark that in the analogous multi-armed bandit setting, it is well-known that the logarithmic expected regret bounds of UCB-type algorithms can be converted into high probability ones, and so it is a natural question whether the same holds for LQRs.

We also remark that the infinite-horizon cost of the optimal policy can be easily replaced in the definition of the regret with the finite-time cost of $K_\star$ (up to additional additive low order terms). This is since the expected costs of any (strongly) stable policy converge exponentially fast to its expected steady-state cost. One could also consider a different definition of the regret, akin to that of "pseudo-regret" in multi-armed bandits, where the learner has to commit at each time step to a linear policy K_t and incurs its mean infinite-horizon cost, $J(K_t)$. (This is the type of notion considered in several recent papers, e.g., Fazel et al., 2018; Malik et al., 2019.) We note, however, that in the unbounded LQR setting there are subtleties that make this definition potentially weaker than the actual expected regret that we focus on; for example, the learner could choose K_t so as to deliberately blow up the magnitude of the states and thereby boost the estimation rates of the unknown system parameters, but at the same time, $J(K_t)$ would remain controlled and no significant penalty in the regret will be incurred.

1.4. Related work

The subject of linear quadratic optimal control has been studied extensively in control theory. Importantly, Lai et al. (1982) establish asymptotic convergence rates of system identification, while Polderman (1986a) show the necessity of said identification for optimal control. More generally, it is known that greedy control strategies only converge to a potentially large subset of the parameter space (see e.g., Kumar, 1983; 1985; Polderman, 1986b), and that in the context of our assumptions this subset is a singleton containing only the true system parameters. While this may allude to our positive findings, the asymptotic nature of existing results makes them insufficient for establishing finite-time (polylogarithmic) regret guarantees, which are the focus of this work.

The topic of learning in linear control has been attracting considerable attention in recent years. Since the early work of Abbasi-Yadkori and Szepesvári (2011), a long line of research has focused on obtaining improved regret bounds for learning in LQRs with a variety of algorithms (Ibrahimi et al., 2012; Faradonbeh et al., 2017; Abeille and Lazaric, 2018; Dean et al., 2018; Faradonbeh et al., 2018; Cohen et al., 2019; Abbasi-Yadkori et al., 2019a;b). To the best of our knowledge, our results are the first to exhibit logarithmic regret rates for LQRs albeit in a more restrictive setting.

A closely related line of work considered a non-stochastic variant of online control in which the cost functions can change arbitrarily from round to round (Cohen et al., 2018; Agarwal et al., 2019a;b). Other notable works have studied the sample complexity of estimating the unknown parameters of linear dynamical systems (Dean et al., 2017; Simchowitz et al., 2018; Sarkar and Rakhlin, 2019), improper prediction of linear systems (Hazan et al., 2017; 2018), as well as model-free learning of LQRs via policy gradient methods (Fazel et al., 2018; Malik et al., 2019).

2. Preliminaries

2.1. Linear Quadratic Control

We give a brief background on several basic properties and results in linear quadratic control that we require in the paper. For a given LQR system (A, B) with cost matrices $Q, R \succ 0$, the optimal (infinite horizon) feedback controller is given by

$$\mathcal{K}(A, B, Q, R) = -\left(R + B^T P B\right)^{-1} B^T P A, \tag{1}$$

where P is the positive definite solution to the discrete Riccati equation

$$P = Q + A^T P A - A^T P B\left(R + B^T P B\right)^{-1} B^T P A. \tag{2}$$

In particular, for the system $(A_\star, B_\star)$ we have $K_\star = \mathcal{K}(A_\star, B_\star, Q, R)$. For more background on linear control and derivation of the relations above, see Bertsekas (1995).

The following lemma, proved in Mania et al. (2019), relates the error in estimating a system's parameters to the deviation of the corresponding estimated controller from the optimal one. This relation is given in terms of cost as well as in terms of distance in operator norm.

Lemma 4. *There are explicit constants $C_0, \varepsilon_0 = \mathrm{poly}(\alpha_0^{-1}, \alpha_1, \vartheta, \nu, \nu_0, d, k)$ such that, for any $0 \le \varepsilon \le \varepsilon_0$ and matrices A, B such that $\|A - A_\star\| \le \varepsilon$ and $\|B - B_\star\| \le \varepsilon$, the policy $K = \mathcal{K}(A, B, Q, R)$ satisfies*

$$J(K) - J_\star \le C_0 \varepsilon^2, \quad \textit{and} \quad \|K - K_\star\| \le C_0 \varepsilon.$$

Importantly, the lemma shows that the performance scales *quadratically* in the estimation error. This served Mania et al. (2019) as a key feature in showing that an ε-greedy algorithm obtains $O(\sqrt{T})$ regret. Here, we use this lemma to show that considerably improved regret bounds are achievable in certain scenarios.

Next, we recall the notion of strong stability (Cohen et al., 2018). This is essentially a quantitative version of classic stability notions in linear control.

Definition 5 (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 controller K for the system (A, B) is (κ, γ)–strongly stable if $\|K\| \le \kappa$ and the matrix $A + BK$ is (κ, γ)–strongly stable.

We remark that Cohen et al. (2018) also introduced the notion of sequential strong stability that is an analogous definition for an adaptive strategy that changes its linear policy over time. Here, we avoid this notion by ensuring that each linear policy is played in our algorithms for a sufficiently long duration.

2.2. Confidence bounds for least-squares estimation

Our algorithms use regularized least squares methods in order to estimate the system parameters. An analysis of this method for a general, possibly-correlated sample, was introduced in the context of linear bandit optimization (Abbasi-Yadkori et al., 2011), and was first used in the context of LQRs by Abbasi-Yadkori and Szepesvári (2011). We state the results in terms of a general sequence, since the estimation procedures differ between our two algorithms.

Let $\Theta_\star \in \mathbb{R}^{d \times m}$, $\{y_{t+1}\}_{t=1}^\infty \in \mathbb{R}^d$, $\{z_t\}_{t=1}^\infty \in \mathbb{R}^m$, $\{w_t\}_{t=1}^\infty \in \mathbb{R}^d$ such that $y_{t+1} = \Theta_\star z_t + w_t$, and $\{w_t\}_{t=1}^\infty$ are i.i.d. with distribution $\mathcal{N}(0, \sigma^2 I)$. Denote by

$$\hat{\Theta}_t \in \underset{\Theta \in \mathbb{R}^{d\times m}}{\arg\min} \left\{ \sum_{s=1}^{t-1} \|y_{t+1} - \Theta z_t\|^2 + \lambda \|\Theta\|_F^2 \right\}, \qquad (3)$$

the regularized least squares estimate of $\Theta_\star$ with regularization parameter λ.

Lemma 6 (Abbasi-Yadkori and Szepesvári, 2011)**.** *Let $V_t = \lambda I + \sum_{s=1}^{t-1} z_t z_t^T$ and $\Delta_t = \Theta_\star - \hat{\Theta}_t$. With probability at least $1 - \delta$, we have for all $t \ge 1$*

$$\mathrm{Tr}\left(\Delta_t^T V_t \Delta_t\right) \le 4\sigma^2 d \log\left(\frac{d}{\delta} \frac{\det(V_t)}{\det(V_1)}\right) + 2\lambda \|\Theta_\star\|_F^2.$$

3. Proofs and Algorithms

In this section we present our algorithms and illustrate the main ideas of our upper and lower bounds. The complete versions of the proofs are deferred to Appendices A to C.

3.1. Upper Bound for Unknown $A_\star$

We start with the setting where $A_\star$ is unknown, and show an efficient algorithm that achieves regret at most $O\left(\log^2 T\right)$. To that end, we propose Algorithm 1. The algorithm begins by playing the stable controller K_0 for a τ_0-long warm-up period. It subsequently operates in phases whose length grows exponentially (quadrupling). Each phase begins by estimating the system parameters using Eq. (3) and computing the greedy controller with respect to said parameters using Eq. (1). It then proceeds to play greedily as long as a fail condition is not reached.

Algorithm 1

input: parameters $\tau_0, x_b, \kappa, \lambda$, a strongly stable controller K_0, and the action-state transition matrix $B_\star$.
initialize: $n_T = \lfloor \log_4(T/\tau_0) \rfloor$, $\tau_{n_T+1} = T + 1$
set: $\tau_i \leftarrow \tau_0 4^i$ for all $0 \le i \le n_T$.
for $t = 1, \ldots, \tau_0 - 1$ **do** ▷ warm-up
  play $u_t = K_0 x_t$.
for phase $i = 0, \ldots, n_T$ **do** ▷ main loop
  $A_{\tau_i} = \arg\min_A \sum_{s=1}^{\tau_i - 1} \|(x_{s+1} - B_\star u_s) - A x_s\|^2 + \lambda \|A\|_F^2$
  $K_{\tau_i} = \mathcal{K}\left(A_{\tau_i}, B_\star, Q, R\right)$.
  for $t = \tau_i, \ldots, \tau_{i+1} - 1$ **do**
    if $\|x_t\|^2 > x_b$ **or** $\|K_{\tau_i}\| > \kappa$ **then** ▷ fail, abort
      abort and play K_0 forever.
    play $u_t = K_{\tau_i} x_t$.

We now give a quantified restatement of Theorem 1.

Theorem (Theorem 1 restated)**.** *Suppose Algorithm 1 is run with parameters*

$$\kappa_0 = \sqrt{\frac{\nu_0}{\alpha_0 \sigma^2}}, \quad \kappa = \sqrt{\frac{\nu + \varepsilon_0^2 C_0}{\alpha_0 \sigma^2}}, \quad \tau_0 = \left\lceil \frac{80 d \lambda \left(1 + \vartheta^2\right)}{\sigma^2 \varepsilon_0^2} \right\rceil,$$
$$\lambda = x_b = 135 d \kappa^2 \sigma^2 \max\left\{\kappa_0^6, 4\kappa^6\right\} \log(3T).$$

Then for $T \ge \mathrm{poly}\left(\alpha_0^{-1}, \alpha_1, \vartheta, \nu, \nu_0, d, k\right)$ we have $\mathbb{E}[R_T] \le \mathrm{poly}\left(\alpha_0^{-1}, \alpha_1, \vartheta, \nu, \nu_0, d, k\right) \log^2 T$.

We start by quantifying a high probability event on which the regret of the algorithm is small. The event holds when the error of the algorithm's estimate of $A_\star$ scales as $t^{-1/2}$, the states are bounded, and all control policies generated by the algorithm are strongly-stable. This is formally given by the following lemma.

Lemma 7. *Let $\gamma = 1/2\kappa^2$. With probability at least $1 - T^{-2}$,*

(i) K_{τ_i} *is* (κ, γ)*–strongly stable, for all* $0 \le i \le n_T$*;*
(ii) $\|x_t\|^2 \le x_b$*, for all* $1 \le t \le T$*;*
(iii) $\|\Delta_{A_{\tau_i}}\| \le \varepsilon_0 2^{-i}$*, for all* $0 \le i \le n_T$*.*

Here we give a sketch of the proof of Lemma 7, deferring technical details to the supplementary material.

Proof (sketch). Consider Lemma 6 with $z_t = x_t, y_{t+1} = x_{t+1} - B_\star u_t$, $V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^T$ and $\Delta_{A_t} = A_t - A_\star$, then we have with probability at least $1 - \frac{1}{3}T^{-2}$

$$\mathrm{Tr}\left(\Delta_{A_t}^T V_t \Delta_{A_t}\right) \le 4\sigma d \log\left(3dT^2 \frac{\det(V_t)}{\det(V_1)}\right) + 2\lambda d \vartheta^2, \quad (4)$$

for all $t \ge 1$. Transforming Eq. (4) into the desired bound requires that we bound V_t from above and below. In what follows we show $\|V_t\| \le \lambda t$ on one hand, and $V_t \succeq \frac{\sigma^2 t}{40} I$ on the other hand. Using the upper bound and choice of parameters, one can show that simplifying the right hand side of Eq. (4) yields $\mathrm{Tr}\left(\Delta_{A_t}^T V_t \Delta_{A_t}\right) \le \sigma^2 \varepsilon_0^2 \tau_0 / 40$. Complementing this with the lower bound gets us

$$\|\Delta_{A_t}\|^2 \le \mathrm{Tr}\left(\Delta_{A_t}^T \Delta_{A_t}\right) \le \frac{40}{\sigma^2 t} \mathrm{Tr}\left(\Delta_{A_t}^T V_t \Delta_{A_t}\right) \le \frac{\varepsilon_0^2 \tau_0}{t},$$

and taking the square root, we obtain the desired estimation error bound that indeed scales as $t^{-1/2}$ (up to logarithmic factors).

For a lower bound on V_t, notice that the system noise w_t ensures that we have a sufficient exploration of the state space. Formally, we have

$$\mathbb{E}[V_t] \succeq \lambda I + \sum_{s=1}^{t-1} \mathbb{E}\left[x_s x_s^T\right] \succeq t\sigma^2 I,$$

where we used $\mathbb{E}\left[x_s x_s^T\right] \succeq \mathbb{E}\left[w_s w_s^T\right] \succeq \sigma^2 I$ and $\lambda \ge \sigma^2$. Applying a measure concentration argument yields the sought-after high-probability lower bound on V_t.

Now, for an upper bound on V_t, notice that

$$\|V_t\| \le \lambda + \sum_{s=1}^{t-1} \|x_s\|^2$$

thus it suffices to show that $\|x_t\|^2 \le x_b = \lambda$. The proof of the lemma now follows inductively by the following argument. If the parameter estimation at time τ_i holds then K_{τ_i} is strongly-stable. This implies that the states throughout phase i satisfy $\|x_t\|^2 \le x_b$ which in turn implies the upper bound on $V_{\tau_{i+1}}$. Thus we can bound the parameter estimation error at time τ_{i+1}. We note that the initial parameter estimation, i.e., at time τ_0, follows from the strong-stability of K_0 and by taking the warm-up duration τ_0 to be sufficiently long. ■

Proof of Theorem 1. Let $\mathcal{E}_A$ be the event where Lemma 7 hold, and notice that the algorithm does not abort on $\mathcal{E}_A$. Defining $J_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} x_t^T\left(Q + K_{\tau_i}^T R K_{\tau_i}\right)x_t$, we have the following decomposition of the regret:

$$\mathbb{E}[R_T] = R_1 + R_2 + R_3 - T \cdot J_\star,$$

where

$$R_1 = \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_A\} \sum_{i=0}^{n_T} J_i\right];$$

$$R_2 = \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_A^c\} \sum_{t=\tau_0}^{T} c_t\right];$$

$$R_3 = \mathbb{E}\left[\sum_{t=1}^{\tau_0-1} c_t\right],$$

are the costs due to success, failure, and warm-up respectively. We now bound each of R_1, R_2, R_3 to conclude the proof.

Starting with R_1, the following lemma uses the strong-stability of K_{τ_i} (whenever $\mathcal{E}_A$ holds) to show that J_i is closely related to the steady-state cost of K_{τ_i}.

Lemma 8. *Fix some i such that* $0 \le i \le n_T$*, and define the event* $E_i = \left\{\|\Delta_{A_{\tau_i}}\| \le \varepsilon_0 2^{-i}\right\}$*. We have*

$$\mathbb{E}\left[\mathbb{1}\{\mathcal{E}_A\}J_i\right] \le (\tau_{i+1} - \tau_i)\mathbb{E}\left[\mathbb{1}\{E_i\}J\left(K_{\tau_i}\right)\right] + 4\alpha_1 \kappa^6 x_b.$$

We further relate the lemma's bound to the cost of the optimal policy using Lemma 4. This gets us

$$\begin{aligned}(\tau_{i+1} - \tau_i)\mathbb{E}\left[\mathbb{1}\{E_i\}J\left(K_{\tau_i}\right)\right] &\le (\tau_{i+1} - \tau_i)\left(J_\star + C_0 \varepsilon_0^2 4^{-i}\right) \\ &\le (\tau_{i+1} - \tau_i)J_\star + 3C_0 \varepsilon_0^2 \tau_0.\end{aligned}$$

Next, summing over i, noticing that $\sum_{i=0}^{n_T} \tau_{i+1} - \tau_i \le T$, and simplifying the arguments yields

$$R_1 \le T \cdot J_\star + n_T\left(6C_0 \varepsilon_0^2 \tau_0 + 8\alpha_1 \kappa^6 x_b\right).$$

Moving to R_2, let τ_{abort} be the time when the algorithm decides to abort, formally,

$$\tau_{\text{abort}} = \min\{t \ge \tau_0 \mid \|x_t\|^2 > x_b \text{ or } \|K_t\| > \kappa\},$$

where we treat $\min \emptyset = T + 1$. Then we have the following bound on R_2.

$$R_2 \le \mathbb{E}\left[\mathbb{1}\{\mathcal{E}_A^c\} \sum_{t=\tau_0}^{\tau_{\text{abort}}-1} c_t\right] + \mathbb{E}\left[\sum_{t=\tau_{\text{abort}}}^{T} c_t\right].$$

Now, the state and control policy before τ_{abort} are bounded by x_b and κ respectively hence $c_t \le 2\alpha_1 \kappa^2 x_b$. Further recalling that $\mathbb{P}\left(\mathcal{E}_A^c\right) \le T^{-2}$ bounds the first term. After

τ_{abort} the stable controller K_0 is played for the remaining period. This ensures that the state will not keep growing however some care is required as the state at $\tau_{\text{abort}}, x_{\tau_{\text{abort}}}$, is not bounded. The above is made formal in the following lemma.

Lemma 9. $R_2 \leq J(K_0) + 2\alpha_1\kappa^2 x_b + o(1)$.

Last, for R_3, the strongly stable controller K_0 is played throughout warm-up. Unlike R_2, here the initial state $x_1 = 0$ is clearly bounded and thus it is not difficult to show that R_3 scales linearly with the warm-up duration τ_0. Since the latter behaves as $O(\log T)$, the desired result is obtained. This is made formal in the following lemma.

Lemma 10. $R_3 \leq \tau_0 J(K_0)$.

The final bound now follows by combining the bounds of R_1, R_2, and R_3 and from n_T, x_b, τ_0 being $O(\log T)$. ■

For a full proof of Lemmas 8 to 10, see the supplementary material.

3.2. Upper Bound for Unknown $B_\star$

We move to a setting where $A_\star$ is known, $B_\star$ is unknown, but $K_\star K_\star^T \succeq \mu_\star I$ for some unknown constant $\mu_\star > 0$. We show an efficient algorithm that achieves regret at most $O\big(\mu_\star^{-2}\log^2 T\big)$. We propose Algorithm 2 to that end. The algorithm operates in a similar fashion to Algorithm 1 with warm-up with K_0 and then greedy with fail-safe, but with two main differences:

1. It adds artificial noise to the action during warm-up;
2. The warm-up length is not predetermined and implicitly depends on $\mu_\star$.

The first change ensures that the action space is explored uniformly during warm-up, and the second ensures that exploration continues at the same rate during the main loop where noise is not added. The specifics of these are made clear in what follows.

We now give a quantified restatement of Theorem 2.

Theorem (Theorem 2 restated)**.** *Suppose Algorithm 2 is run with parameters*

$$\kappa_0 = \sqrt{\frac{\nu_0}{\alpha_0\sigma^2}}, \quad \kappa = \sqrt{\frac{\nu + \varepsilon_0^2 C_0}{\alpha_0\sigma^2}}, \quad \tau_0 = \left\lceil \frac{80k\lambda\big(1+\vartheta^2\big)}{\sigma^2\varepsilon_0^2} \right\rceil,$$
$$x_b = 135 d\kappa^2\sigma^2 \max\Big\{(1+\vartheta)^2\kappa_0^6, 4\kappa^6\Big\}\log(4T),$$
$$\lambda = \kappa^2 x_b, \quad \mu_0 = 4\kappa C_0\varepsilon_0.$$

Then for $T \geq \operatorname{poly}\big(\alpha_0^{-1}, \alpha_1, \vartheta, \nu, \nu_0, d, k, \mu_\star^{-1}\big)$ *we have* $\mathbb{E}[R_T] \leq \operatorname{poly}\big(\alpha_0^{-1}, \alpha_1, \vartheta, \nu, \nu_0, d, k, \mu_\star^{-1}\big)\log^2 T$.

Algorithm 2

input: parameters $\tau_0, x_b, \kappa, \lambda, \mu_0$, a strongly stable controller K_0, and the state transition matrix $A_\star$.
initialize: $n_T = \lfloor \log_4(T/\tau_0)\rfloor$, $n_s = n_T + 1$, $\tau_{n_T+1} = T + 1$.
set: $\tau_i \leftarrow \tau_0 4^i$, $\mu_i \leftarrow \mu_0 2^{-i}$ for all $0 \leq i \leq n_T$
for $t = 1, \ldots, \tau_0 - 1$ **do** ▷ initial warm-up
    play $u_t \sim \mathcal{N}(K_0 x_t, \sigma^2 I)$
for phase $i = 0, \ldots, n_T$ **do** ▷ adaptive warm-up
    $B_{\tau_i} = \arg\min_B \sum_{s=1}^{\tau_i - 1} \|(x_{s+1} - A_\star x_s) - Bu_s\|^2 + \lambda\|B\|_F^2$
    $K_{\tau_i} = \mathcal{K}\big(A_\star, B_{\tau_i}, Q, R\big)$.
    if $K_{\tau_i}K_{\tau_i}^T \succeq (3\mu_i/2)I$ **then**
        save $n_s = i$ and **break**.
    for $t = \tau_i, \ldots, \tau_{i+1} - 1$ **do**
        play $u_t \sim \mathcal{N}(K_0 x_t, \sigma^2 I)$
for phase $i = n_s, \ldots, n_T$ **do** ▷ main loop
    $B_{\tau_i} = \arg\min_B \sum_{s=1}^{\tau_i - 1} \|(x_{s+1} - A_\star x_s) - Bu_s\|^2 + \lambda\|B\|_F^2$
    $K_{\tau_i} = \mathcal{K}\big(A_\star, B_{\tau_i}, Q, R\big)$.
    for $t = \tau_i, \ldots, \tau_{i+1} - 1$ **do**
        if $\|x_t\|^2 > x_b$ **or** $\|K_{\tau_i}\| > \kappa$ **then** ▷ fail, abort
            abort and play K_0 forever.
        play $u_t = K_{\tau_i} x_t$.

We provide the main ideas required to prove Theorem 2. As in Algorithm 1, we first quantify the high probability event under which the regret of the algorithm is small. Let us first consider the parameter estimation error during warm-up, which is bounded by the following lemma.

Lemma 11. *With probability at least* $1 - T^{-2}$*, it holds that* $\|\Delta_{B_{\tau_i}}\| \leq \varepsilon_0 2^{-i}$ *for all* $0 \leq i \leq n_s$.

Here we only give a sketch of the proof; for the full technical details, see the supplementary material.

Proof (sketch). Consider Lemma 6 with $z_t = u_t$, $y_{t+1} = x_{t+1} - A_\star x_t$, $V_t = \lambda I + \sum_{s=1}^{t-1} u_t u_t^T$ and $\Delta_{B_t} = B_t - B_\star$, then with probability at least $1 - \frac{1}{4}T^{-2}$

$$\operatorname{Tr}\big(\Delta_{B_t}^T V_t \Delta_{B_t}\big) \leq 4\sigma d \log\left(4dT^2 \frac{\det(V_t)}{\det(V_1)}\right) + 2\lambda k\vartheta^2,$$

for all $t \geq 1$. Hence, bounding V_t from above and below as in Lemma 7 yields the desired parameter estimation error bound.

Now, during warm-up $u_t \sim \mathcal{N}(K_0 x_t, \sigma^2 I)$ which is equivalent to having $u_t = K_0 x_t + \eta_t$ where $\eta_t \sim \mathcal{N}(0, \sigma^2 I)$ are i.i.d. random variables. Note that just as w_t provided exploration for x_t, here η_t provides exploration for u_t. Indeed, for the lower bound, we have

$$\mathbb{E}[V_t] \succeq \lambda I + \sum_{s=1}^{t-1}\mathbb{E}\big[u_s u_s^T\big] \succeq \lambda I + \sum_{s=1}^{t-1}\mathbb{E}\big[\eta_s\eta_s^T\big] \succeq t\sigma^2 I,$$

and thus a measure concentration argument yields the desired high probability lower bound. For the upper bound, notice that

$$\|V_t\| \leq \lambda + \sum_{s=1}^{t-1} \|u_s\|^2 \leq \lambda + 2\sum_{s=1}^{t-1} \left(\|K_0\|^2 \|x_s\|^2 + \|\eta_s\|^2\right),$$

and so the strong-stability of K_0 together with a high probability bound on the system and artificial noises yields the desired upper bound on V_t. Combining both upper and lower bounds concludes the proof. ∎

While the estimation rate during warm-up is desirable, adding constant magnitude noise to the action incurs regret that is linear in the warm-up length, even if $K_0 = K_\star$, and as such we avoid this strategy during the main loop. Nonetheless, the following lemma shows that the estimation rate continues into the main loop albeit with slightly different constants.

Lemma 12. *Let $\gamma = 1/2\kappa^2$. With probability at least $1-T^{-2}$,*

(i) K_{τ_i} is (κ, γ)–strongly stable, $\forall\, n_s \leq i \leq n_T$;
(ii) $\|x_t\|^2 \leq x_b$, $\forall\, 1 \leq t \leq T$;
(iii) $\|\Delta_{B_{\tau_i}}\| \leq \varepsilon_0 \min\{2^{-n_s}, 2\mu_\star^{-1/2} 2^{-i}\}$, $\forall\, n_s < i \leq n_T$.

We proceed with a proof sketch and defer details to the supplementary material.

Proof (sketch). The proof follows inductively by similar arguments to those of Lemma 7, yet with the caveat that the lower bound on V_t may not hold when the controller is rank deficient.

To see this, recall that the algorithm plays $u_t = K_{\tau_i} x_t$ during the main loop as long as the abort state is not triggered, so we have

$$\mathbb{E}\left[u_t u_t^T \mid K_{\tau_i}\right] = K_{\tau_i} \mathbb{E}\left[x_t x_t^T \mid K_t\right] K_{\tau_i}^T \succeq \sigma^2 K_{\tau_i} K_{\tau_i}^T.$$

This means that transforming the exploration of states x_t, provided for by the system noise w_t, into exploration of actions u_t depends on the controller K_{τ_i} being strictly non-degenerate. We show that with high probability, $K_{\tau_i} K_{\tau_i}^T \succeq (\mu_\star/2) I$ thus ensuring the exploration and the parameter estimation rate.

First, suppose that the learner had knowledge of $\mu_\star$ and recall that $\mu_0 = 4\kappa C_0 \varepsilon_0$. Taking $n_s \geq \max\{0, \log_2(\mu_0/\mu_\star)\}$, Lemma 11 implies that $\|\Delta_{B_{\tau_{n_s}}}\| \leq \min\{\varepsilon_0, \frac{\mu_\star}{4\kappa C_0}\}$ and applying Lemma 4 we get that $\|K_{\tau_{n_s}} - K_\star\| \leq \mu_\star/4\kappa$. Further assuming that $\|K_{\tau_{n_s}}\| \leq \kappa$, which is ensured by strong-stability, simple algebra yields that $K_{\tau_{n_s}} K_{\tau_{n_s}}^T \succeq (\mu_\star/2) I$.

Now, when $\mu_\star$ is unknown, we show that the break condition of the warm-up loop ensures that with high probability

$$\max\left\{0, \log_2 \frac{\mu_0}{\mu_\star}\right\} \leq n_s \leq 2 + \max\left\{0, \log_2 \frac{\mu_0}{\mu_\star}\right\}, \quad (5)$$

a proof of which may be found in the supplementary material. The lower bound on n_s ensures the desired non-degeneracy of $K_{\tau_{n_s}}$, and proceeding by induction, the same follows for subsequent controllers. We note that the purpose of the upper bound on n_s is to ensure that the warm-up is not so long as to incur more than $O\left(\mu_\star^{-2} \log^2 T\right)$ regret. ∎

Proceeding from Lemma 12, we obtain a regret decomposition similar to that of Algorithm 1 with an added dependence on the random number of warm-up phases n_s. While this randomness introduces some additional technical challenges, the proof ideas remain largely the same. For the full proof of Theorem 2, see the supplementary material.

3.3. Lower Bound for Degenerate $K_\star$

In this section we prove an $\Omega(\sqrt{T})$ lower bound for systems with a (nearly) degenerate optimal policy, stated in Theorem 3. By Yao's principle, to establish the theorem it is enough to demonstrate a randomized construction of an LQR system such that the expected regret of any deterministic learning algorithm is large.

Fix $d = k = 1$ and consider the system

$$\begin{aligned} x_{t+1} &= a x_t + b u_t + w_t\ ; \\ c_t &= x_t^2 + u_t^2. \end{aligned} \quad (6)$$

Here, $w_t \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Gaussian random variables, $a = 1/\sqrt{5}$ and $b = \chi\sqrt{\epsilon}$ where χ is a Rademacher random variable (drawn initially) and $\epsilon > 0$ is a parameter whose value will be chosen later. For simplicity, we assume that $x_1 = 0$. Notice that for this system, we have the bounds $\alpha_1 = \alpha_0 = 1$, $\vartheta = 1$ and, as we will see below, the optimal cost of the system is bounded by $\nu = 2\sigma^2$. Further, note that the system is controllable and $k_0 = 0$ is a stabilizing policy. Our goal is to lower bound the regret, given by

$$R_T = \sum_{t=1}^{T} \left(x_t^2 + u_t^2 - J(k_\star)\right).$$

Theorem 3 follows directly from the following:

Theorem 13. *Assume that $T \geq 12000$ and set $\epsilon = T^{-1/2}/4$. Then the expected regret of any deterministic learning algorithm on on the system in Eq. (6) satisfies*

$$\mathbb{E}[R_T] \geq \frac{1}{3100} \sigma^2 \sqrt{T} - 4\sigma^2.$$

Here, the expectation is taken with respect to both the stochastic noise terms as well as the random variable χ.

For the proof, we use the following notation. We use $k_\star$ to denote the optimal policy for the system, which (recalling

Eqs. (1) and (2)) is given by

$$k_\star = -\frac{abp_\star}{1+b^2p_\star},$$

where $p_\star > 0$ is a positive solution to the Riccati equation

$$p_\star = 1 + a^2p_\star - \frac{a^2b^2p_\star^2}{1+b^2p_\star} = 1 + \frac{a^2p_\star}{1+b^2p_\star}.$$

Observe that for our choice of $\epsilon \leq 1/400$ we have that $|b| \leq 1/20$, and so

$$\begin{aligned} 1 \leq p_\star \leq 1/(1-a^2) = 5/4, \\ 0.99\sqrt{\epsilon/5} \leq |k_\star| \leq \sqrt{\epsilon/3}. \end{aligned} \tag{7}$$

In particular, this means that the cost of the optimal policy is at most $\sigma^2 p_\star \leq 2\sigma^2$. Further, the sign of $k_\star$ is solely determined by the sign of χ.

Now, fix any deterministic learning algorithm. Let $x^{(t)} = (x_1, \ldots, x_t)$ denote the trajectory generated by the learning algorithm up to and including time step t. Denote by $\mathbb{P}_+$ and $\mathbb{P}_-$ the probability laws with respect to the trajectory generated conditioned on $\chi = 1$ and $\chi = -1$ respectively.

First, we lower bound the expected regret in terms of the cumulative magnitude of the algorithm's actions u_t. The proof first relates the regret to the overall deviation of u_t from the actions of the optimal policy $k_\star$ by using the fact that the action played by $k_\star$ at any state minimizes the Q-function of the system. Since the actions of $k_\star$ are small in expectation, the latter quantity can be in turn related to the total magnitude of the u_t.

Lemma 14. *Suppose $\epsilon \leq 1/400$. The expected regret is lower bounded as*

$$\mathbb{E}[R_T] \geq 0.99\,\mathbb{E}\left[\sum_{t=1}^{T}(u_t - k_\star x_t)^2\right] - 4\sigma^2,$$

and consequently,

$$\mathbb{E}[R_T] \geq \frac{1}{3}\mathbb{E}\left[\sum_{t=1}^{T} u_t^2\right] - \frac{5}{6}\sigma^2 k_\star^2 T - 4\sigma^2.$$

Note that for the last bound to be meaningful, $k_\star$ indeed has to be very small so that the additive term that scales with $k_\star^2 T$ does not dominate the right hand side. The proofs of this as well as subsequent lemmas are deferred to the supplementary material.

Next, by standard information theoretic arguments, we obtain an upper bound on the statistical distance between the probability laws of $x^{(T)}$ under $\mathbb{P}_+$ and $\mathbb{P}_-$, that scales with the total magnitude of the actions u_t.

Lemma 15. *For the trajectory $x^{(T)}$, it holds that*

$$TV(\mathbb{P}_+[x^{(T)}], \mathbb{P}_-[x^{(T)}]) \leq \frac{\sqrt{\epsilon}}{\sigma}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} u_t^2\right]}\,.$$

Our final lemma shows that most of the states visited by the algorithm have a non-trivial (constant) magnitude. This is a straightforward consequence of the added Gaussian noise at each time step.

Lemma 16. *Assume that $T \geq 12000$. With probability $\geq \frac{7}{8}$, at least $\frac{2}{3}T$ of the states $x_1, \ldots, x_T$ satisfy $|x_t| \geq 2\sigma/5$.*

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

Proof of Theorem 13. Notice that if $\mathbb{E}[\sum_{t=1}^T u_t^2] > \frac{1}{4}\sigma^2\sqrt{T}$, then the desired lower bound is directly implied by the second inequality in Lemma 14, as $k_\star^2 \leq \epsilon/3 = T^{-1/2}/12$, so $\mathbb{E}[R_T] \geq \frac{1}{100}\sigma^2\sqrt{T} - 4\sigma^2$. We henceforth assume that $\mathbb{E}[\sum_{t=1}^T u_t^2] \leq \frac{1}{4}\sigma^2\sqrt{T}$. Plugging this into the bound of Lemma 15 for the total variation distance between $\mathbb{P}_+$ and $\mathbb{P}_-$, and using our choice $\epsilon = T^{-1/2}/4$, we obtain that

$$\mathrm{TV}(\mathbb{P}_+[x^{(T)}], \mathbb{P}_-[x^{(T)}]) \leq \sqrt{\frac{\epsilon}{\sigma^2}\cdot\frac{\sigma^2}{4}\sqrt{T}} = \frac{1}{4}.$$

Now, let N_T denote the number of time steps in which $u_t k_\star x_t \leq 0$, i.e., the number of times in which the learner has guessed the sign of χ incorrectly. We claim that $\mathbb{P}[N_T \geq T/2] \geq 3/8$. To see this, denote by N_T' the number of time steps t in which $u_t x_t \leq 0$. Using the fact that N_T' is a deterministic function of the trajectory $x^{(T)}$ together with the bound on the total variation gives

$$\begin{aligned} &|\mathbb{P}_+[N_T' \geq T/2] - \mathbb{P}_-[N_T' \geq T/2]| \\ &\qquad \leq \mathrm{TV}(\mathbb{P}_+[x^{(T)}], \mathbb{P}_-[x^{(T)}]) \leq \frac{1}{4}. \end{aligned}$$

Now, recall that the sign of $k_\star$ is determined by that of χ. Thus, $\mathbb{P}_-[N_T \geq T/2] = \mathbb{P}_-[N_T' < T/2]$ and $\mathbb{P}_+[N_T \geq T/2] = \mathbb{P}_+[N_T' \geq T/2]$ from which

$$\begin{aligned} \mathbb{P}[N_T \geq T/2] &= \tfrac{1}{2}\mathbb{P}_+[N_T \geq T/2] + \tfrac{1}{2}\mathbb{P}_-[N_T \geq T/2] \\ &= \tfrac{1}{2}(1 + \mathbb{P}_+[N_T' \geq T/2] - \mathbb{P}_-[N_T' \geq T/2]) \\ &\geq 3/8. \end{aligned} \tag{8}$$

On the other hand, Lemma 16 implies that with probability at least 7/8, no less than $2T/3$ of the states $x_1, \ldots, x_T$ satisfy $|x_t| > 2\sigma/5$. Then by a union bound, with probability at least 1/4, at least $T/6$ instances of $x_1, \ldots, x_T$ satisfy $|x_t| \geq 2\sigma/5$ and $u_t k_\star x_t \leq 0$. For these instances, we have

$$(u_t - k_\star x_t)^2 \geq k_\star^2 x_t^2 \geq 0.99^2 \frac{4}{125}\epsilon\sigma^2,$$

where we have bounded $k_\star$ as in Eq. (7). Hence, we can lower bound the regret using the first inequality in Lemma 14 as follows:

$$\begin{aligned}\mathbb{E}[R_T] &\geq 0.99 \cdot \mathbb{E}\left[\sum_{t=1}^{T}(u_t - k_\star x_t)^2\right] - 4\sigma^2 \\ &\geq 0.99^3 \cdot \frac{1}{4} \cdot \frac{T}{6} \cdot \frac{4}{125}\epsilon\sigma^2 - 4\sigma^2 \\ &\geq \frac{1}{3100}\sigma^2\sqrt{T} - 4\sigma^2,\end{aligned}$$

where the last transition used our choice of ϵ. ■

Acknowledgements

We thank Yishay Mansour for numerous helpful discussions. This work was partially supported by the Israeli Science Foundation (ISF) grant 2549/19 and by the Yandex Initiative in Machine Learning.

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, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702, 2019a.

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvari. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3108–3117, 2019b.

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.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119, 2019a.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pages 10175–10184, 2019b.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

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

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pages 1–47, 2017.

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

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive regulation and learning. *arXiv preprint arXiv:1811.04258*, 2018.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2018.

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.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pages 6702–6712, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4634–4643, 2018.

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.

Panqanamala Ramana Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23(3):329–380, 1985.

PR Kumar. Optimal adaptive control of linear-quadratic-gaussian systems. *SIAM Journal on Control and Optimization*, 21(2):163–178, 1983.

Tze Leung Lai, Ching Zong Wei, et al. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925, 2019.

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. Control of unknown linear systems with thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205. IEEE, 2017.

Jan Willem Polderman. On the necessity of identifying the true parameter in adaptive lq control. *Systems & control letters*, 8(2):87–91, 1986a.

Jan Willem Polderman. A note on the structure of two subsets of the parameter space in adaptive control problems. *Systems & control letters*, 7(1):25–34, 1986b.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24, 2013.

Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473, 2018.

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.