Optimism-Based Adaptive Regulation of Linear-Quadratic Systems

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis

***Abstract*—The main challenge for adaptive regulation of linear-quadratic systems is the tradeoff between identification and control. An adaptive policy needs to address both the estimation of unknown dynamics parameters (exploration), as well as the regulation of the underlying system (exploitation). To this end, optimism-based methods that bias the identification in favor of optimistic approximations of the true parameter are employed in the literature. A number of asymptotic results have been established, but their finite-time counterparts are few, with important restrictions. This article establishes results for the worst-case regret of optimism-based adaptive policies. The presented high probability upper bounds are optimal up to logarithmic factors. The nonasymptotic analysis of this article requires the following very mild assumptions: stabilizability of the system's dynamics, and limiting the degree of heaviness of the noise distribution. To establish such bounds, certain novel techniques are developed to comprehensively address the probabilistic behavior of dependent random matrices with heavy-tailed distributions.**

***Index Terms*—Certainty equivalence (CE), exploration-exploitation, optimism in the face of uncertainty (OFU), reinforcement learning, regret bounds.**

I. Introduction

Adaptive control of linear-quadratic (LQ) state-space models represents a canonical problem, and is the main focus of this article. Such a model describes the dynamics of the system as follows: starting from the initial state $x(0) \in \mathbb{R}^p$, its temporal evolution and cost are determined by

$$x(t+1) = A_0 x(t) + B_0 u(t) + w(t+1) \tag{1}$$

$$c_t = x(t)' Q x(t) + u(t)' R u(t) \tag{2}$$

for $t = 0, 1, \ldots$. The vector $x(t) \in \mathbb{R}^p$ denotes the output (and state) of the system at time t, $u(t) \in \mathbb{R}^r$ represents the control signal, and the stochastic process of the noise sequence is denoted by $\{w(t)\}_{t=1}^{\infty}$. Furthermore, the quadratic function c_t corresponds to the instantaneous cost of the system (the transpose of the vector v is denoted by v'). The transition matrix $A_0 \in \mathbb{R}^{p \times p}$ and the input matrix $B_0 \in \mathbb{R}^{p \times r}$ that constitute the dynamical parameters of the system are *unknown*, while the positive definite matrices of the cost, $Q \in \mathbb{R}^{p \times p}$, $R \in \mathbb{R}^{r \times r}$ are assumed known.

The main goal is to adaptively regulate the system in order to minimize its long-term average cost. This canonical problem has been thoroughly studied in the literature and a number of asymptotic results have been established, as briefly summarized next. However, finite-time results are scarce and rather incomplete, despite their need in applications (e.g., network systems [1]). Note that the theoretical guarantee for fast stabilization of general linear systems has been recently established [2], but the existing analysis of the regulation problem of cost minimization leads to a remarkable loss of generality, as will be discussed shortly.

Since the system dynamics are unknown, a popular adaptive procedure for regulation is based on the principle of *certainty equivalence* (CE) [3]. Alternating between estimation and regulation, CE applies a control feedback *as if* the identified parameters A_0 and B_0 are the true matrices that drive the system's evolution [4]–[6]. However, it has been shown that the CE-based strategy can lead to wildly incorrect parameter estimates [7]–[9], and thus, suitable modifications have been introduced in the literature [10], [11]. A popular approach, known as *optimism in the face of uncertainty* (OFU) [12], was developed to address the suboptimality of CE. In OFU, after constructing a confidence set for the model parameters, a regulation policy is designed based on the most *optimistic* parameter in the confidence set [13].

The aforementioned references establish the asymptotic convergence of the *average* cost to the optimal value. However, nonasymptotic results on the growth rate of *regret* [i.e., the accumulative deviation from the optimal cost, see (5)] have recently appeared [14], [15]. These papers provide a near-optimal upper bound for the regret of OFU, under the following rather restrictive conditions.

1) The dynamics matrices are assumed to be *controllable* and *observable*. This leads to an excessive complexity in the computation of the adaptive regulator. Furthermore, this assumption restricts the applicability of the analysis since the condition may be violated in many LQ systems.
2) The *operator norm* of the closed-loop matrix is less than one, which excludes a remarkable fraction of systems with stable closed-loop matrices. In fact, a stable matrix can have an arbitrarily large operator norm. Note that condition 1 only implies that the largest closed-loop eigenvalue (not the operator norm) is less than one [16].
3) The noise distribution satisfies a tail condition such as *sub-Gaussianity* [14] or Gaussianity [15]. Moreover, the coordinates of the noise vectors are uncorrelated.

This article aims to address these shortcomings by providing a comprehensive treatment of the problem. We study optimality of OFU policies for an extensive family of LQ systems by establishing upper bounds for the *worst-case* regret, under a minimal set of assumptions. Namely, we remove the aforementioned condition 1, and replace the strict condition 2 with *stabilizability*, which is the necessary assumption for the optimal control problem to be well-defined. Furthermore, the high probability near-optimal upper bound for regret established in

Manuscript received March 28, 2019; revised November 23, 2019; accepted May 15, 2020. Date of publication June 1, 2020; date of current version March 29, 2021. The work of George Michailidis was supported in part by NSF under Grant DMS-1821220 and Grant IIS-1632730. The work of Ambuj Tewari was supported by NSF CAREER under Grant IIS-1452099. Recommended by Associate Editor G. Gu. *(Corresponding author: Mohamad Kazem Shirani Faradonbeh.)*

Mohamad Kazem Shirani Faradonbeh and George Michailidis are with the Department of Statistics, University of Florida, Gainesville, FL 32611-5585 USA, and also with Informatics Institute, University of Florida, Gainesville, FL 32611-5585 USA (e-mail: mfaradonbeh@ufl.edu; gmichail@ufl.edu).

Ambuj Tewari is with the Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107 USA, and also with the Department of Electrical Engineering and Computer Science (by courtesy), University of Michigan, Ann Arbor, MI 48109-1107 USA (e-mail: tewaria@umich.edu).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TAC.2020.2998952

this article holds for a class of *heavy-tailed* noise vectors with arbitrary correlation structures, thus significantly relaxing the condition 3. To the authors' knowledge, this study is the first to address the nonasymptotic analysis of the regret of adaptive policies for general LQ systems.

There are a number of conceptual and technical difficulties one needs to address in order to obtain the results of optimal regulation. First, the existing methodology for analyzing adaptive policies [11], [14], [15] becomes nonapplicable beyond the condition 2. One reason is due to the fact that matrix multiplication preserves the operator norm; i.e., the norm of the product is upper bounded by the product of the norms. However, the product of two stable matrices can have eigenvalues of arbitrarily large magnitude. Furthermore, sub-Weibull distributions assumed in this study do not need to have generating functions [17]. Hence, new tools are required to establish concentration inequalities for random matrices with heavy-tailed probability distributions [18], [19].

In addition, an adaptive strategy is needed to stabilize the system so that the uncertainty about A_0 and B_0 does not lead to instability. Adaptive stabilization methods are proposed before, and their finite-time performance analysis is provided [2]. First, a coarse approximation of the unknown dynamics parameter is shown to be enough for stabilization. Then, it is established that such approximations can be achieved by employing independent random feedbacks in sufficiently many periods. Nevertheless, for nonasymptotic analysis of the performance of regulation policies, a comprehensive study is not currently available, and is adopted as the focus of this study. In case the operator is concerned with stability issues, the algorithm in the aforementioned reference can be applied *a priori* to the regulation algorithms we discuss here.

The remainder of this article is organized as follows. Section II formally defines the problem. Section III addresses the problem of accurate estimation of the closed-loop matrix and includes the analysis of the empirical covariance matrix, as well as a high probability prediction bound. Finally, an optimism-based algorithm for adaptive regulation of the system is presented in Section IV. We show that the regret of *Algorithm 1* is with high-probability optimal, up to a logarithmic factor. Finally, Section VI concludes this article.

The following notation is used throughout this article. For matrix $A \in \mathbb{C}^{p\times q}$, A' is its transpose. When $p = q$, the smallest (respectively, largest) eigenvalue of A (in magnitude) is denoted by $\lambda_{\min}(A)$ (respectively, $\lambda_{\max}(A)$) and the trace of A is denoted by $\mathbf{tr}(A)$. For $\gamma \in \mathbb{R}, \gamma \geq 1, v \in \mathbb{C}^q$, the norm of v is $||v||_\gamma = (\sum_{i=1}^q |v_i|^\gamma)^{1/\gamma}$. Furthermore, when $\gamma = \infty$, the norm is defined according to $||v||_\infty = \max_{1\leq i\leq q} |v_i|$. We also use the following notation for the operator norm of matrices. For $\beta, \gamma \in [1, \infty]$, and $A \in \mathbb{C}^{p\times q}$, define

$$|||A|||_{\gamma\to\beta} = \sup_{v\in\mathbb{C}^q\setminus\{0\}} \frac{||Av||_\beta}{||v||_\gamma}.$$

Whenever $\gamma = \beta$, we simply write $|||A|||_\beta$. Finally, the sigma-field generated by random vectors $X_1, \ldots, X_n$ is denoted by $\sigma(X_1, \ldots, X_n)$. The notation for $\theta, K(\theta), L(\theta)$, and $\widetilde{L}(\theta)$ are provided in Definition 2, (3), (4), and Definition 4, respectively. Finally, log is employed throughout this article to refer to the natural logarithm function.

II. Problem Formulation

First, we formally discuss the problem of adaptive regulation this article is addressing. Equation (1) depicts the dynamics of the system, where $\{w(t)\}_{t=1}^\infty$ are independent mean-zero noise vectors with full rank covariance matrix C as

$$\mathbb{E}\left[w(t)\right] = 0, \;\; \mathbb{E}\left[w(t)w(t)'\right] = C, \;\; |\lambda_{\min}(C)| > 0.$$

The results established also hold if the noise vectors are martingale difference sequences. The true dynamics are assumed to be stabilizable, as defined as follows.

Definition 1 (Stabilizability [16]): $[A_0, B_0]$ is stabilizable if there is $L \in \mathbb{R}^{r\times p}$ such that $|\lambda_{\max}(A_0 + B_0 L)| < 1$. The linear feedback matrix L is called a stabilizer.

Definition 2 (Notation θ): We use θ to denote the dynamics parameter $[A, B]$, where A and B are $p \times p$ and $p \times r$ matrices, respectively. Obviously $\theta \in \mathbb{R}^{p\times q}$, for $q = p + r$. In particular, we frequently refer to $\theta_0 = [A_0, B_0]$ throughout this article.

Here, we consider *perfect observations*, i.e., the output of the system corresponds to the state vector itself. Next, an admissible control policy is a mapping π that designs the control action according to the dynamics matrix θ_0, the cost matrices Q and R, and the history of the system; that is for all $t \geq 0$

$$u(t) = \pi\left(\theta_0, Q, R, \{x(i)\}_{i=0}^t, \{u(j)\}_{j=0}^{t-1}\right).$$

An *adaptive* policy is ignorant about the parameter θ_0. So

$$u(t) = \pi\left(Q, R, \{x(i)\}_{i=0}^t, \{u(j)\}_{j=0}^{t-1}\right).$$

When applying the policy π, the resulting instantaneous quadratic cost at time t defined according to (2) is denoted by $c_t^{(\pi)}$. If there is no superscript, the corresponding policy will be clear from the context. For arbitrary policy π, let $\overline{\mathcal{J}}_\pi(\theta_0)$ be the average cost of the system

$$\overline{\mathcal{J}}_\pi\left(\theta_0\right) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^T c_t^{(\pi)}.$$

Note that the dependence of $\overline{\mathcal{J}}_\pi(\theta_0)$ to the known cost matrices Q and R is suppressed. Then, the optimal average cost is defined by $\mathcal{J}^\star(\theta_0) = \min_\pi \overline{\mathcal{J}}_\pi(\theta_0)$, where the minimum is taken over *all* admissible policies. Furthermore, $\pi^\star$ is called an *optimal* policy for the system θ, if satisfying $\overline{\mathcal{J}}_{\pi^\star}(\theta) = \mathcal{J}^\star(\theta)$. To find $\pi^\star$ for general $\theta \in \mathbb{R}^{p\times q}$, one has to solve a *Riccati* equation. A solution, is a positive semidefinite matrix $K(\theta)$ satisfying

$$\begin{aligned} K\left(\theta\right) &= Q + A'K\left(\theta\right)A \\ &\quad - A'K\left(\theta\right)B\left(B'K\left(\theta\right)B + R\right)^{-1}B'K\left(\theta\right)A. \end{aligned} \tag{3}$$

The following result establishes optimality of the linear feedback provided by $K(\theta)$ according to

$$L\left(\theta\right) = -\left(B'K\left(\theta\right)B + R\right)^{-1}B'K\left(\theta\right)A. \tag{4}$$

Definition 3 (Policy $\pi^\star$): Henceforth, let $\pi^\star$ denote the linear feedback policy $u(t) = L(\theta_0)x(t)$ for all $t \geq 0$.

Lemma 1 (Optimality [2]): If θ_0 is stabilizable, then (3) has a unique solution, $\pi^\star$ is optimal, and $\mathcal{J}^\star(\theta_0) = \mathbf{tr}(K(\theta_0)C)$. Conversely, if $K(\theta_0)$ is a solution of (3), $L(\theta_0)$ is a stabilizer.

Note that in the latter case of Lemma 1, the existence of a solution $K(\theta_0)$ implies that it is unique, $\pi^\star$ is an optimal policy, and $\mathcal{J}^\star(\theta_0) = \mathbf{tr}(K(\theta_0)C)$.

In order to measure the quality of (adaptive) policy π, the resulting cost will be compared to the optimal average cost defined previously. More precisely, letting $c_t^{(\pi)}$ be the resulting instantaneous cost at time t, *regret* at time T is defined as

$$\mathcal{R}(T) = \sum_{t=1}^T \left[c_t^{(\pi)} - \mathcal{J}^\star\left(\theta_0\right)\right]. \tag{5}$$

The comparison between adaptive control policies is made according to regret. The next result describes the asymptotic distribution of the regret. Lemma 2, which is basically a central limit theorem for $\mathcal{R}(T)$,

states that even when applying optimal policy, the regret $\mathcal{R}(T)$ scales as $\mathcal{O}(T^{1/2})$, multiplied by a normal random variable.

Lemma 2: Applying $\pi^\star$, let $D = A_0 + B_0 L(\theta_0)$ be the closed-loop matrix. Then, $T^{-1/2}\mathcal{R}(T)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$ as T grows, where

$$\sigma^2 = 4\,\mathbf{tr}\left(K(\theta_0) C K(\theta_0) \sum_{n=1}^{\infty} D^n C D'^n\right) + \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \mathbf{Var}\left[w(t)' K(\theta_0)\, w(t)\right] > 0.$$

The proof of Lemma 2 based on an application of the martingale central limit theorem [19] is deferred to the supplementary materials. In the sequel, we discuss the result of Lemma 2. In the definition of regret in (5), the cumulative deviation from the optimal average cost can be decomposed into the following two fractions.

1) The *probabilistic* fraction contributed by the stochastic evolution of the system and randomness of $\{w(t)\}_{t=1}^{\infty}$.

2) The *statistical* fraction caused by the uncertainty about the dynamics and unknownness of θ_0 to the operator.

Lemma 2 states that the probabilistic fraction scales with the growth rate $\mathcal{O}(T^{1/2})$. So, trying to push the statistical fraction of the regret (which is due to the error in learning the unknown dynamics) to have a rate less than $\mathcal{O}(T^{1/2})$ is actually *unnecessary*. Furthermore, Lemma 2 provides a lower bound for the worst-case regret of adaptive policies. Since the optimal policy for minimizing the expected cumulative cost $\sum_{t=0}^{T} \mathbb{E}[c_t]$ converges to $\pi^\star$ as T grows [16], the regret of an arbitrary policy cannot be smaller than that of $\pi^\star$. On the other hand, the high probability upper bound of a normal distribution is in magnitude at least $(-\log\delta)^{1/2}$. Therefore, Lemma 2 implies that a high probability regret bound to hold with probability at least $1-\delta$, needs to be at least of the order of magnitude of $T^{1/2}(-\log\delta)^{1/2}$. Note that the aforementioned argument does not necessarily imply *impossibility* of the smaller magnitudes for the statistical fraction of the regret.[1] However, since there are information theoretic limits in learning the unknown parameter θ_0, statistical regret cannot be small. A rigorous derivation of lower bounds for the statistical regret is beyond the scope of this article. Although, later on, we will intuitively discuss efficiency of the rate $T^{1/2}$, based on the decomposition being used in the regret analysis of Section IV.

Definition 4 (Notation $\widetilde{L}(\theta)$): For arbitrary stabilizable θ_1, θ_2, let $\widetilde{L}(\theta_1) = [I_p, L(\theta_1)']'$. So, $\theta_2 \widetilde{L}(\theta_1) = A_2 + B_2 L(\theta_1)$.

III. Closed-Loop Identification

When applying linear feedback $L \in \mathbb{R}^{r\times p}$ to the system, the closed-loop dynamics becomes $x(t+1) = Dx(t) + w(t+1)$, where $D = A_0 + B_0 L$. Subsequently, we present bounds for the time length the user can interact with the system in order to have sufficiently many observations for accurate identification of the closed-loop matrix. The next set of results is used later on to construct the confidence sets being used to design the adaptive policy. Since the focus is on adaptive policies for *regulating* the system, the matrix D is assumed to be stable.

First, we define least-squares estimation for the matrix D, as follows. Observing the state vectors $\{x(t)\}_{t=0}^{n}$, for an arbitrary matrix $M \in \mathbb{R}^{p\times p}$ consider the sum-of-squares loss function

$$\mathcal{L}_n(M) = \sum_{t=0}^{n-1} ||x(t+1) - Mx(t)||_2^2.$$

[1] for example, applying $\pi^\star$, we get $\lim_{T\to\infty} T^{-1/2}\mathbb{E}[\mathcal{R}(T)] = 0$.

Then, the true closed-loop transition matrix D is estimated by $\widehat{D}_n$, which is a minimizer of the aforementioned loss: $\mathcal{L}_n(\widehat{D}_n) = \min_{M\in\mathbb{R}^{p\times p}} \mathcal{L}_n(M)$. Solving for $\widehat{D}_n$, one can easily see that it admits the closed-form expression

$$\widehat{D}_n = \sum_{t=0}^{n-1} x(t+1)x(t)' V_n^{-1}$$

where $V_n = \sum_{t=0}^{n-1} x(t)x(t)'$ denotes the (invertible) empirical covariance matrix of the state process. Therefore, the behavior of V_n needs to be carefully studied. To this end, one needs to tightly examine the state sequence $\{x(t)\}_{t=0}^{n}$, which in turn highly depends on both the spectral properties of the transition matrix D, as well as the noise process $\{w(t)\}_{t=1}^{n}$. The former is reflected through the constant $\boldsymbol{\eta}(D)$, while the latter is indicated by $\boldsymbol{\nu}_n(\delta)$ we shortly define.

To proceed, let $D = P^{-1}\Lambda P$ be the Jordan decomposition of D; i.e., Λ is block diagonal, $\Lambda = \mathbf{diag}(\Lambda_1, \ldots, \Lambda_k)$, where for all $i = 1, \ldots, k$, Λ_i is a Jordan matrix of λ_i as

$$\Lambda_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix} \in \mathbb{C}^{m_i\times m_i}.$$

Definition 5 (Constant $\boldsymbol{\eta}(D)$): Denote the Jordan decomposition described previously by $D = P^{-1}\Lambda P$. Letting

$$\boldsymbol{\eta}_t(\Lambda_i) = \inf_{\rho \geq |\lambda_i|} t^{m_i - 1} \rho^t \sum_{j=0}^{m_i - 1} \frac{\rho^{-j}}{j!}$$

for $t \geq 1$, define $\boldsymbol{\eta}_t(\Lambda) = \max_{1\leq i\leq k} \boldsymbol{\eta}_t(\Lambda_i)$. Then, let $\boldsymbol{\eta}_0(\Lambda) = 1$, and

$$\boldsymbol{\eta}(D) = \left|\left|\left|P^{-1}\right|\right|\right|_{\infty\to 2} |||P|||_\infty \sum_{t=0}^{\infty} \boldsymbol{\eta}_t(\Lambda).$$

Letting $\overline{\lambda} = |\lambda_{\max}(D)|$, if D is *diagonalizable*, then clearly $\boldsymbol{\eta}(D) \leq |||P^{-1}|||_{\infty\to 2} |||P|||_\infty (1-\overline{\lambda})^{-1}$. In general, denoting the dimension of the largest block in the Jordan decomposition of D by $\mu = \max_{1\leq i\leq k} m_i$, we have $\boldsymbol{\eta}_t(\Lambda) \leq t^{\mu-1}\overline{\lambda}^t e^{1/\overline{\lambda}}$, and

$$\boldsymbol{\eta}(D) \leq \left|\left|\left|P^{-1}\right|\right|\right|_{\infty\to 2} |||P|||_\infty e^{1/\overline{\lambda}} \left[\frac{\mu - 1}{-\log\overline{\lambda}} + \frac{(\mu-1)!}{\left(-\log\overline{\lambda}\right)^\mu}\right].$$

Toward studying the effect of the noise vectors on the state process, the following tail condition is assumed.

Assumption 1 (Sub-Weibull distribution [17]): There are positive reals b_1, b_2, and α, such that for all $t \geq 1$; $1 \leq i \leq p$; and $y > 0$

$$\mathbb{P}\left(|w_i(t)| > y\right) \leq b_1 \exp\left(-\frac{y^\alpha}{b_2}\right).$$

Clearly, the smaller the exponent α is, the heavier the tail of $w_i(t)$ will be. Assuming a sub-Weibull distribution for the noise coordinates is more general than the sub-Gaussian (or subexponential) assumption routinely made in the literature of nonasymptotic analysis [14], where $\alpha \geq 2$ ($\alpha \geq 1$). To gain insight into the basic properties of sub-Weibull distributions, consider the setting $\alpha < 1$. It delivers an extensive family of distributions for which moments of all orders are well defined, while the moment generating function does not exist. So, it relaxes more restrictive tail conditions to a minimal framework that finite-time concentration results can be established. Furthermore, Assumption 1

encompasses fundamental distributions that subexponential families fail to capture, such as polynomials of Gaussian random variables. Finally, to obtain analogous results for uniformly bounded noise sequences, it suffices to let $\alpha \to \infty$ in the subsequently presented materials.

In order to study magnitudes of the state vectors over time, define

$$\boldsymbol{\nu}_n(\delta) = \left(b_2 \log\left(\frac{b_1 np}{\delta}\right)\right)^{1/\alpha} \tag{6}$$

$$\boldsymbol{\xi}_n(\delta) = \boldsymbol{\eta}(D)\left(||x(0)||_\infty + \boldsymbol{\nu}_n(\delta)\right). \tag{7}$$

Lemmas 3 and 4 show that $\boldsymbol{\nu}_n(\delta), \boldsymbol{\xi}_n(\delta)$ are the high probability uniform bounds for the size of the noise and the state vectors. As a matter of fact, $\boldsymbol{\nu}_n(\delta)$ and $\boldsymbol{\xi}_n(\delta)$ scale as $\log^{1/\alpha}(n/\delta)$. Hence, for uniformly bounded noise, both of them are fixed constants. Then, recalling that C is the positive definite covariance matrix of the noise vectors, let $N(\epsilon, \delta)$ be large enough, such that the followings hold for all $n \geq N(\epsilon, \delta)$:

$$\frac{n}{\boldsymbol{\nu}_n(\delta)^2} \geq \frac{18\,|\lambda_{\max}(C)| + 2\epsilon}{\epsilon^2} p \log\left(\frac{4p}{\delta}\right) \tag{8}$$

$$\frac{n}{\boldsymbol{\xi}_n(\delta)^2 \boldsymbol{\nu}_n(\delta)^2} \geq \frac{288}{\epsilon^2} p |||D|||_2^2 \log\left(\frac{4p}{\delta}\right) \tag{9}$$

$$\frac{n}{\boldsymbol{\xi}_n(\delta)^2} \geq \frac{6}{\epsilon}\left(|||D|||_2^2 + 1\right). \tag{10}$$

The following result provides a high probability lower bound for the smallest eigenvalue of V_{n+1}. Essentially, Theorem 1 determines the number of state observations needed to ensure that the excitation is persistent enough to identify the closed-loop matrix [20], [21].

Theorem 1 (Empirical covariance): If $n \geq N(\epsilon, \delta)$, then

$$\mathbb{P}\left(|\lambda_{\min}(V_{n+1})| < n\left(|\lambda_{\min}(C)| - \epsilon\right)\right) < 2\delta.$$

Moreover, $\lim_{n\to\infty} n^{-1} V_n = \sum_{i=0}^{\infty} D^i C D'^i$.

Proof: First, for $n \geq 1$, and $0 < \delta < 1$, define the event

$$\mathcal{W} = \left\{\max_{1\leq t\leq n} ||w(t)||_\infty \leq \boldsymbol{\nu}_n(\delta)\right\}. \tag{11}$$

We use the following intermediate results, for which the proofs are delegated to the supplementary materials, due to space limitations.

Lemma 3: Defining $\mathcal{W}$ according to (11), we have $\mathbb{P}(\mathcal{W}) \geq 1 - \delta$.

Lemma 4: The following holds on the event $\mathcal{W}$ in (11):

$$\max_{1\leq t\leq n} ||x(t)||_2 \leq \boldsymbol{\xi}_n(\delta).$$

Lemma 5: Let the event $\mathcal{W}$ be as (11), and define $C_n = n^{-1}\sum_{i=1}^{n} w(i)w(i)'$. Then, on $\mathcal{W}$, we have $\mathbb{P}(|\lambda_{\max}(C_n - C)| > \epsilon) \leq \delta$, if

$$\frac{n}{\boldsymbol{\nu}_n(\delta)^2} \geq \frac{6\,|\lambda_{\max}(C)| + 2\epsilon}{3\epsilon^2} p \log\left(\frac{2p}{\delta}\right). \tag{12}$$

Lemma 6: Let $U_n = n^{-1}\sum_{i=0}^{n-1}[Dx(i)w(i+1)' + w(i+1)x(i)'D']$, and define $\mathcal{W}$ by (11). Then, on $\mathcal{W}$, we have $\mathbb{P}(|\lambda_{\max}(U_n)| > \epsilon) \leq \delta$, if

$$\frac{n}{|||D|||_2^2 \boldsymbol{\nu}_n(\delta)^2 \boldsymbol{\xi}_n(\delta)^2} \geq \frac{32p}{\epsilon^2} \log\left(\frac{2p}{\delta}\right). \tag{13}$$

Next, note that $x(t+1) = Dx(t) + w(t+1)$ implies

$$V_{n+1} = x(0)x(0)' + D\sum_{i=0}^{n-1} x(i)x(i)'D' + nU_n + nC_n$$

Algorithm 1: Adaptive Regulation.

Inputs: $\Omega^{(0)} \subset \mathbb{R}^{p\times q}$, $6\delta > 0$, $\gamma > 1$
Let $\tau_0 = 0$
for $i = 1, 2, \ldots$ **do**
  Define $\widetilde{\theta}^{(i)}$, τ_i according to (20) and (21), respectively
  while $t < \tau_i$ **do**
    Apply control feedback $u(t) = L(\widetilde{\theta}^{(i)})x(t)$
  end while
  Find the estimate $\widehat{D}^{(i)}$ given in (22)
  Using $V^{(i)}$ in (23), construct $\Gamma^{(i)}$ according to (24)
  Update $\Omega^{(i)}$ by (25)
end for

where C_n and U_n are defined in Lemmas 5 and 6. So, we obtain the Lyapunov equation $V_{n+1} = DV_{n+1}D' + nE_n$, for

$$E_n = U_n + C_n + \frac{D\left(x(0)x(0)' - x(n)x(n)'\right)D'}{n} + \frac{x(0)x(0)'}{n}$$

to obtain

$$V_{n+1} = n\sum_{i=0}^{\infty} D^i E_n D'^i. \tag{14}$$

Henceforth, suppose that $\mathcal{W}$ holds. According to Lemma 5, (8) implies that

$$\mathbb{P}\left(|\lambda_{\max}(C_n - C)| > \frac{\epsilon}{3}\right) \leq \frac{\delta}{2}. \tag{15}$$

In addition, by Lemma 6, (9) implies that

$$\mathbb{P}\left(|\lambda_{\max}(U_n)| > \frac{\epsilon}{3}\right) \leq \frac{\delta}{2}. \tag{16}$$

Finally, using Lemma 4, by (10), we get

$$\frac{1}{n}\left(|||D|||_2^2 + 1\right)\left(||x(0)||_2^2 + ||x(n)||_2^2\right) \leq \frac{\epsilon}{3}. \tag{17}$$

Putting (15)–(17) together, on $\mathcal{W}$, with probability at least $1 - \delta$, it holds that $|\lambda_{\min}(E_n)| \geq |\lambda_{\min}(C)| - \epsilon$. Therefore, since (14) implies that $|\lambda_{\min}(V_{n+1})| \geq n|\lambda_{\min}(E_n)|$, we get the desired result. When $n \to \infty$, the conditions hold for arbitrary positive values of ϵ and δ. Thus, we have $|\lambda_{\max}(E_n - C)| \to 0$, which according to (14) implies the desired result. ■

The following corollary provides a high probability confidence set for D, which will be used later in Algorithm 1. Using the bounds $\boldsymbol{\nu}_n(\delta), \boldsymbol{\xi}_n(\delta)$ introduced in (6) and (7), define the prediction bound $\boldsymbol{\beta}_n(\delta)$ according to

$$\boldsymbol{\beta}_n(\delta) = \frac{16np}{(n-1)\,|\lambda_{\min}(C)|} \boldsymbol{\xi}_n(\delta)^2 \boldsymbol{\nu}_n(\delta)^2 \log\left(\frac{2p}{\delta}\right). \tag{18}$$

Corollary 1 (Prediction bound): Define $\boldsymbol{\beta}_n(\delta)$ by (18). Then, $n \geq N(|\lambda_{\min}(C)|/2, \delta) + 1$ implies that

$$\mathbb{P}\left(\left|\left|\left|V_n^{1/2}\left(\widehat{D}_n - D\right)'\right|\right|\right|_2^2 > \boldsymbol{\beta}_n(\delta)\right) \leq 3\delta.$$

Proof: First, since $n \geq N(|\lambda_{\min}(C)|/2, \delta) + 1$, similar to the proof of Theorem 1, on the event $\mathcal{W}$ defined in (11), with probability at least $1 - \delta$, we have $|\lambda_{\min}(V_n)| \geq |\lambda_{\min}(C)|(n-1)/2$. Then, as long as V_n is nonsingular, one can write $\widehat{D}_n - D = (\sum_{t=0}^{n-1} w(t+1)x(t)')V_n^{-1}$, which yields $(\widehat{D}_n - D)V_n(\widehat{D}_n - D)' = U_n' V_n^{-1} U_n$,

where $U_n = \sum_{t=0}^{n-1} x(t)w(t+1)'$. Therefore

$$\left|\left|\left|\left(\widehat{D}_n - D\right) V_n \left(\widehat{D}_n - D\right)'\right|\right|\right|_2 \leq \frac{|||U_n|||_2^2}{|\lambda_{\min}(V_n)|}. \tag{19}$$

To proceed, for the arbitrary matrix $H \in \mathbb{R}^{k\times\ell}$, define the dilation

$$\Phi(H) = \begin{bmatrix} 0_{k\times k} & H \\ H' & 0_{\ell\times\ell} \end{bmatrix} \in \mathbb{R}^{(k+\ell)\times(k+\ell)}.$$

A well-known fact states that the equality $|||H|||_2 = |\lambda_{\max}(\Phi(H))|$ holds [18]. So, letting $Z_t = x(t)w(t+1)'$, apply the following random matrix concentration inequality to $X_t = \Phi(Z_t) \in \mathbb{R}^{2p\times 2p}$.

Lemma 7: [18] Let $\{X_i\}_{i=1}^n$ be a martingale difference sequence of symmetric $p \times p$ matrices adapted to the filtration $\{\mathcal{F}_i\}_{i=0}^n$. Assume for fixed symmetric matrices $\{M_i\}_{i=1}^n$, all matrices $M_i^2 - X_i^2$ are positive semidefinite. Then, letting $\sigma^2 = |\lambda_{\max}(\sum_{i=1}^n M_i^2)|$, for all $y \geq 0$, we have

$$\mathbb{P}\left(\left|\lambda_{\max}\left(\sum_{i=1}^n X_i\right)\right| \geq y\right) \leq 2p\exp\left(-\frac{y^2}{8\sigma^2}\right).$$

Since

$${X_t}^2 = \begin{bmatrix} ||w(t+1)||_2^2 x(t)x(t)' & 0_{p\times p} \\ 0_{p\times p} & ||x(t)||_2^2 w(t+1)w(t+1)' \end{bmatrix}$$

by Lemmas 3 and 4, all matrices ${M_t}^2 - {X_t}^2$ are positive semidefinite on the event $\mathcal{W}$ defined in (11), with $M_t = \Phi(p^{1/2}\boldsymbol{\nu}_n(\delta)\boldsymbol{\xi}_n(\delta)I_p)$. By $\sigma^2 = np\boldsymbol{\nu}_n(\delta)^2\boldsymbol{\xi}_n(\delta)^2$, letting $y = 8^{1/2}\sigma\log^{1/2}(\frac{2p}{\delta})$, Lemma 7 implies $\mathbb{P}(|||U_n|||_2 > y) = \mathbb{P}(|\lambda_{\max}(\Phi(U_n))| > y) \leq \delta$. Plugging in (19), we get the desired result. ■

IV. Design of Adaptive Policy

In this section, we present an algorithm for adaptive regulation of LQ systems. When applying the following algorithm, we assume that a stabilizing set is provided. Construction of such a set with an arbitrary high probability guarantee is addressed in the literature [2]. It is established that the proposed adaptive stabilization procedure returns a stabilizing set in finite time. Nevertheless, if such a set is not available, the operator can apply the proposed method of the random linear feedback [2] in order to stabilize the system before running the following adaptive policy.

In the episodic algorithm described later, estimation will be reinforced at the end of every episode. Indeed, the algorithm is based on a sequence of confidence sets, which are constructed according to Corollary 1. This sequence will be tightened at the end of every episode so that the provided confidence sets become more and more accurate. According to this sequence, the adaptive linear feedback will be updated after every episode. After explaining the algorithm, we present a high probability regret bound.

First, we provide a high level explanation of the algorithm. Starting with the stabilizing set $\Omega^{(0)}$, we select a parameter $\widetilde{\theta}^{(1)} \in \Omega^{(0)}$ based on the OFU principle; $\widetilde{\theta}^{(1)}$ is a minimizer of the optimal average cost over the corresponding confidence set [see (20)].

Then, assuming $\widetilde{\theta}^{(1)}$ is the true parameter, the system evolves according to, during the first episode the algorithm applies the optimal linear feedback $L(\widetilde{\theta}^{(1)})$. Once the observations during the first episode are collected, they are used to improve the accuracy of the high probability confidence set. Therefore, $\Omega^{(0)}$ is tightened to $\Omega^{(1)}$, and the second episode starts by selecting $\widetilde{\theta}^{(2)}$, iterating the aforementioned procedure, and so on. The lengths of the episodes will be increasing, to make every confidence set significantly more accurate than all previous ones.

The intuition behind proficiency of the OFU principle is as follows. Applying a linear feedback L, the closed-loop transition matrix is $A_0 + B_0 L = \theta_0\widetilde{L}$, where $\widetilde{L} = [I_p, L']'$. Importantly, the observed sequence of state vectors accurately identifies the closed-loop matrix $\theta_0\widetilde{L}$. However, an accurate estimation of $\theta_0\widetilde{L}$ does not lead to that of θ_0. Therefore, θ_0 is not guaranteed to be effectively approximable, regardless of the accuracy in the approximation of $\theta_0\widetilde{L}$.

Nevertheless, one has to focus on finding accurate approximations of the feedback matrix $L(\theta_0)$, in order to design an effective adaptive policy for minimizing the average cost. Specifically, as long as θ_1 is available satisfying $L(\theta_1) = L(\theta_0)$, one can apply an optimal linear feedback $L(\theta_1)$, no matter how large $|||\theta_1 - \theta_0|||_2$ is. In general, estimation of such a θ_1 is not possible. Yet, an optimistic approximation in addition to exact knowledge of the closed-loop dynamics lead to an optimal linear feedback, thanks to the OFU principle.

Lemma 8: If $\mathcal{J}^\star(\theta_1) \leq \mathcal{J}^\star(\theta_0)$, and $\theta_1\widetilde{L}(\theta_1) = \theta_0\widetilde{L}(\theta_1)$, then $L(\theta_1)$ is optimal for the system θ_0: $L(\theta_0) = L(\theta_1)$.

In other words, applying linear feedback $L(\theta_1)$, which is designed according to an optimistically selected parameter θ_1, as long as the closed-loop matrix $\theta_0\widetilde{L}(\theta_1)$ is exactly identified, the optimal linear feedback is automatically provided. Recall that the lengths of the episodes are growing so that the estimation of the closed-loop matrix becomes more precise at the end of every episode. Thus, the approximation $\theta_1\widetilde{L}(\theta_1) \approx \theta_0\widetilde{L}(\theta_1)$ is becoming more and more accurate. Rigorous analysis of the aforementioned discussion, leads to the high probability near-optimal regret bound of Theorem 2.

Algorithm 1 takes the stabilizing set $\Omega^{(0)}$, the failure probability 6δ, and the reinforcement rate $\gamma > 1$ as inputs. Indeed, $\Omega^{(0)}$ is a bounded stabilizing set such that for every $\theta \in \Omega^{(0)}$, the system will be stable if the optimal linear feedback of θ is applied; that is, $|\lambda_{\max}(\theta_0\widetilde{L}(\theta))| < 1$. As mentioned before, an algorithmic procedure to obtain a bounded stabilizing set in finite time is available in the literature [2]. Furthermore, $6\delta > 0$ is the highest probability that Algorithm 1 fails to adaptively regulate the system such that the regret will be nearly optimal (see Theorem 2). The reinforcement rate γ determines the growth rate of the lengths of the time intervals (episodes) a specific feedback is applied until being updated [see (21)].

The algorithm provides an adaptive policy as follows. For $i = 1, 2, \ldots$, at the beginning of the ith episode, we apply linear feedback $u(t) = L(\widetilde{\theta}^{(i)})x(t)$, where

$$\widetilde{\theta}^{(i)} \in \arg\min_{\theta\in\Omega^{(i-1)}} \mathcal{J}^\star(\theta). \tag{20}$$

Indeed, based on the OFU principle, at the beginning of every episode, *the most optimistic* parameter among all we are uncertain about is being selected. The length of the episode i, which is the time period we apply the adaptive control policy $u(t) = L(\widetilde{\theta}^{(i)})x(t)$, is designed according to the following equation. Letting $\tau_0 = 0$, we update the control policy at the end of the episode i at the time $t = \tau_i$, defined according to

$$\tau_i = \tau_{i-1} + \gamma^{i/q} N\left(\frac{|\lambda_{\min}(C)|}{2}, \frac{\delta}{i^2}\right) + \gamma^{i/q} \tag{21}$$

where $N(\cdot,\cdot)$ is defined by (8)–(10). After the ith episode, we estimate the closed-loop transition matrix $\theta_0\widetilde{L}(\widetilde{\theta}^{(i)})$ by the following least-squares estimator:

$$\widehat{D}^{(i)} = \arg\min_{M\in\mathbb{R}^{p\times p}} \sum_{t=\tau_{i-1}}^{\tau_i - 1} ||x(t+1) - Mx(t)||_2^2. \tag{22}$$

Letting $V^{(i)}$ be the empirical covariance matrix of the episode i as

$$V^{(i)} = \sum_{t=\lceil \tau_{i-1} \rceil}^{\lceil \tau_i \rceil - 1} x(t)x(t)' \tag{23}$$

define the high probability confidence set as

$$\Gamma^{(i)} = \left\{ \theta \in \mathbb{R}^{p\times q} : \left| \left| \left| V^{(i)^{1/2}} \left(\theta \widetilde{L}\left(\widetilde{\theta}^{(i)}\right) - \widehat{D}^{(i)} \right)' \right| \right|_2^2 \right. \\ \left. \leq \boldsymbol{\beta}_{\tau_i - \tau_{i-1}} \left(\frac{\delta}{i^2} \right) \right\} \tag{24}$$

where $\boldsymbol{\beta}_n(\delta)$ is defined in (18). Note that according to Corollary 1, $\mathbb{P}(\theta_0 \in \Gamma^{(i)}) \geq 1 - 3\delta i^{-2}$. Then, at the end of the episode i, the confidence set $\Omega^{(i-1)}$ will be updated to

$$\Omega^{(i)} = \Omega^{(i-1)} \cap \Gamma^{(i)} \tag{25}$$

and episode $i+1$ starts, finding $\widetilde{\theta}^{(i+1)}$ by (20), and then, iterating all steps described previously.

Remark 1: The choice of $\widetilde{\theta}^{(i)}$ does not need to be as extreme as (20) [14]. In fact, it suffices to satisfy $\mathcal{J}^\star(\widetilde{\theta}^{(i)}) \leq (\tau_i - \tau_{i-1})^{-1/2} + \inf_{\theta \in \Omega^{(i-1)}} \mathcal{J}^\star(\theta)$.

The following result states that performance of the aforementioned adaptive control algorithm is optimal, apart from a logarithmic factor. Theorem 2 also provides the effect of the degree of heaviness of the noise distribution (denoted by α in Assumption 1) on the regret. Compared to $\mathcal{O}(\cdot)$, the notation $\widetilde{\mathcal{O}}(\cdot)$ used in the following, hides the logarithmic factors.

Theorem 2 (Regret bound): For bounded $\Omega^{(0)}$, with probability at least $1 - 6\delta$, the regret of Algorithm 1 satisfies

$$\mathcal{R}(T) \leq \widetilde{\mathcal{O}}\left(T^{1/2} \left(-\log \delta \right)^{1/2+2/\alpha} \right).$$

Proof: The stabilizing set $\Omega^{(0)}$ is bounded

$$\rho_1 = \sup_{\theta \in \Omega^{(0)}} |||\theta'|||_2 < \infty. \tag{26}$$

Suppose that for $t = 1, 2, \ldots$, the parameter θ_t is being used to design the adaptive linear feedback $u(t) = L(\theta_t)x(t)$. So, during every episode, θ_t does not change, and for $\tau_{i-1} \leq t < \tau_i$, we have $\theta_t = \widetilde{\theta}^{(i)}$.

Letting $\mathcal{F}_t = \sigma(w(1), \ldots, w(t))$, the infinite horizon dynamic programming equations [16] are

$$\mathcal{J}^\star(\theta_t) + x(t)'K(\theta_t)x(t) = x(t)'Qx(t) + u(t)'Ru(t) \\ + \mathbb{E}\left[y(t+1)'K(\theta_t)y(t+1) \middle| \mathcal{F}_t \right]$$

where $u(t) = L(\theta_t)x(t)$, and

$$y(t+1) = A_t x(t) + B_t u(t) + w(t+1) = \theta_t \widetilde{L}(\theta_t)x(t) + w(t+1) \tag{27}$$

describes the desired dynamics of the system. Note that since the true evolution of the system is governed by θ_0, the next state is

$$x(t+1) = A_0 x(t) + B_0 u(t) + w(t+1) \\ = \theta_0 \widetilde{L}(\theta_t)x(t) + w(t+1). \tag{28}$$

Substituting (27) and (28) into the dynamic programming equation, and using (2) for the instantaneous cost c_t, we have

$$\mathcal{J}^\star(\theta_t) + x(t)'K(\theta_t)x(t) \\ = c_t + \mathbb{E}\left[w(t+1)'K(\theta_t)w(t+1) \middle| \mathcal{F}_t \right] \\ + x(t)\widetilde{L}(\theta_t)'\theta_t' K(\theta_t)\theta_t \widetilde{L}(\theta_t)x(t) \\ = c_t + \mathbb{E}\left[x(t+1)'K(\theta_t)x(t+1) \middle| \mathcal{F}_t \right] \\ + x(t)\widetilde{L}(\theta_t)'\left[\theta_t' K(\theta_t)\theta_t - \theta_0' K(\theta_t)\theta_0\right]\widetilde{L}(\theta_t)x(t).$$

Adding up the terms for $t = 1, \ldots, T$, we obtain

$$\mathcal{R}(T) = \sum_{t=1}^{T} \left[c_t - \mathcal{J}^\star(\theta_0) \right] = \mathbb{Y}_1 + \mathbb{Y}_2 + \mathbb{Y}_3 + \mathbb{Y}_4 \tag{29}$$

$$\mathbb{Y}_1 = \sum_{t=1}^{T} \left[\mathcal{J}^\star(\theta_t) - \mathcal{J}^\star(\theta_0) \right] \tag{30}$$

$$\mathbb{Y}_2 = \sum_{t=1}^{T} \Big(x(t)'K(\theta_t)x(t) \\ - \mathbb{E}\left[x(t+1)'K(\theta_{t+1})x(t+1) \middle| \mathcal{F}_t \right] \Big) \tag{31}$$

$$\mathbb{Y}_3 = \sum_{t=1}^{T} \mathbb{E}\left[x(t+1)'\left(K(\theta_{t+1}) - K(\theta_t)\right)x(t+1) \middle| \mathcal{F}_t \right] \tag{32}$$

$$\mathbb{Y}_4 = \sum_{t=1}^{T} x(t)'\widetilde{L}(\theta_t)'\left[\theta_0' K(\theta_t)\theta_0 - \theta_t' K(\theta_t)\theta_t\right]\widetilde{L}(\theta_t)x(t). \tag{33}$$

where the expressions for $\mathbb{Y}_1, \mathbb{Y}_2, \mathbb{Y}_3$, and $\mathbb{Y}_4$ are defined in (30)–(33). Let $m(T)$ be the number of episodes considered until time T. Thus

$$\tau_{m(T)} \leq T < \tau_{m(T)+1}.$$

Now, letting $n_i = \lfloor \tau_i - \tau_{i-1} \rfloor$ be the length of the episode i, define the following events:

$$\mathcal{G} = \bigcap_{i=1}^{\infty} \left\{ \max_{\tau_{i-1} \leq t < \tau_i} ||w(t)||_\infty \leq \boldsymbol{\nu}_{n_i}\left(\frac{\delta}{i^2}\right) \right\}$$

$$\mathcal{H} = \bigcap_{i=1}^{\infty} \left\{ \theta_0 \in \Omega^{(i)} \right\}.$$

According to Corollary 1

$$\mathbb{P}(\mathcal{G} \cap \mathcal{H}) \geq 1 - \sum_{i=1}^{\infty} \frac{3\delta}{i^2} \geq 1 - 5\delta. \tag{34}$$

For all $i = 1, 2, \ldots$, as long as $\theta_0 \in \Omega^{(i-1)}$, according to (20), we have $\mathcal{J}^\star(\widetilde{\theta}^{(i)}) \leq \mathcal{J}^\star(\theta_0)$; i.e., $\mathcal{J}^\star(\theta_t) - \mathcal{J}^\star(\theta_0) \leq 0$. Therefore, on $\mathcal{G} \cap \mathcal{H}$, we have

$$\mathbb{Y}_1 \leq 0. \tag{35}$$

To conclude the proof, we leverage some auxiliary results. The proofs of the following lemmas are deferred to supplementary materials due to space limitations.

Lemma 9 (Bounding $\mathbb{Y}_2$): On $\mathcal{G} \cap \mathcal{H}$, the following holds with probability at least $1 - \delta$:

$$\mathbb{Y}_2 \leq \rho_2 + (8T)^{1/2}\rho_3 \left(\log\left(Tm(T)\right)\right)^{2/\alpha} \left(-\log\delta\right)^{1/2+2/\alpha}$$

where $\rho_2, \rho_3 < \infty$ are fixed constants.

Lemma 10 (Bounding $\mathbb{Y}_3$): On $\mathcal{G} \cap \mathcal{H}$, we have

$$\mathbb{Y}_3 \leq \rho_3 \left(\log\left(Tm(T)\right)\right)^{2/\alpha} \left(-\log\delta\right)^{2/\alpha} m(T)$$

where ρ_3 is the same as Lemma 9.

Lemma 11 (Bounding $\mathbb{Y}_4$): On the event $\mathcal{G} \cap \mathcal{H}$, it holds that

$$\mathbb{Y}_4 \leq \rho_4 m(T)^{3/2} \boldsymbol{\beta}_T \left(\frac{\delta}{m(T)^2} \right)^{1/2} T^{1/2}$$

for some fixed constant $\rho_4 < \infty$.

Lemma 12 (Bounding $m(T)$): On the event $\mathcal{G} \cap \mathcal{H}$, the following holds:

$$m(T) \leq \frac{q}{\log \gamma} \log \left(\frac{T\left(\gamma^{1/q} - 1\right)}{\tau_1} + 1 \right).$$

Finally, the definition of $\boldsymbol{\beta}_n(\delta)$ in (18) yields

$$\boldsymbol{\beta}_n(\delta) = \mathcal{O}\left((\log n)^{4/\alpha} (-\log \delta)^{1+4/\alpha} \right).$$

Therefore, plugging (35), and the results of Lemmas 9–12 into (29), we get $\mathcal{R}(T) \leq \widetilde{\mathcal{O}}(T^{1/2}(-\log \delta)^{1/2+2/\alpha})$, with probability at least $1 - \delta$ on $\mathcal{G} \cap \mathcal{H}$. Hence, according to (34), the failure probability is at most 6δ, which completes the proof. ■

To conclude this section, we briefly discuss the behavior of the statistical regret introduced in the discussion after Lemma 2. For this purpose, we use the regret decomposition of (29) into the terms $\mathbb{Y}_1, \ldots, \mathbb{Y}_4$ being defined in (30)–(33). According to Lemma 10, $\mathbb{Y}_3$ scales logarithmically with T. Furthermore, since the martingale $\mathbb{Y}_2$ is bounded in expectation, we have $\limsup_{T\to\infty} \mathbb{E}[\mathbb{Y}_2] < \infty$. Hence, one can approximately study the behavior of the statistical regret by addressing $\mathbb{Y}_1$ and $\mathbb{Y}_4$. First, note that the expression $\theta_0' K(\theta_t)\theta_0 - \theta_t' K(\theta_t)\theta_t$ in (33) can be substituted by $(\theta_0 + \theta_t)' K(\theta_t)(\theta_0 - \theta_t)$. Since $K(\theta_t)$ is positive definite [2], the magnitude of $\mathbb{Y}_4$ is approximately as large as $\sum_{t=1}^{T} |||\theta_t - \theta_0|||_2$. A similar argument applies to $\mathbb{Y}_1$ in the sense that the decay rate of $\mathcal{J}^\star(\theta_t) - \mathcal{J}^\star(\theta_0)$ heavily relies on the error of learning θ_0 through θ_t. Then, the learning accuracy at time t is at best of the order $t^{-1/2}$ [4]. Hence, the statistical regret an adaptive policy needs to incur is at least $\mathcal{O}(T^{1/2})$, because of lack of knowledge about the true parameter. Converting this lower bound sketch into a rigorous proof is beyond the scope of this article, and is left as an interesting problem for future studies.

V. Conclusion

This article investigated adaptive regulation schemes for linear dynamical systems with quadratic costs, focusing on finite-time analysis for regret. Using the OFU principle, we established nonasymptotic efficiency results under the mild condition of stabilizability, and also assuming a fairly general heavy-tailed noise distribution.

Note that implementation of the OFU principle in (20) leads to a nonconvex optimization problem. Thus, from a practical viewpoint, computationally faster algorithms for adaptive regulation are of interest. For this purpose, one can employ randomization methods in order to balance identification and regulation. Analysis of adaptive policies based on dithering the control signal, or randomizing the parameter estimate is provided by Faradonbeh *et al.* [22], [23].

There are a number of interesting extensions of the current work. First, generalizing the nonasymptotic analysis of efficiency to *imperfect* observations of the state vector is a topic of future investigation. Another interesting direction is to specify the sufficient and necessary conditions for the true dynamics which lead to optimality of *CE*. In addition, reexamining the problem for large *network* systems where the dynamics matrices can be sparse is also of interest.

Acknowledgment

The authors would like to thank Prof. T. L. Lai and the anonymous reviewers for helpful discussions.

References

[1] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Optimality of fast matching algorithms for random networks with applications to structural controllability," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 4, pp. 770–780, Dec. 2017.
[2] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time adaptive stabilization of linear systems," *IEEE Trans. Autom. Control*, vol. 64, no. 8, pp. 3498–3505, Aug. 2019.
[3] Y. Bar-Shalom and E. Tse, "Dual effect, certainty equivalence, and separation in stochastic control," *IEEE Trans. Autom. Control*, vol. AC-19, no. 5, pp. 494–500, Oct. 1974.
[4] T. L. Lai, "Asymptotically efficient adaptive control in stochastic regression models," *Advances Appl. Math.*, vol. 7, no. 1, pp. 23–45, 1986.
[5] L. Guo and H. Chen, "Convergence rate of ELS based adaptive tracker," *Syst. Sci Math. Sci*, vol. 1, pp. 131–138, 1988.
[6] T. L. Lai and Z. Ying, "Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems," *SIAM J. Control Optim.*, vol. 29, no. 5, pp. 1091–1127, 1991.
[7] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *Ann. Statist.*, vol. 10, pp. 154–166, 1982.
[8] A. Becker, P. Kumar, and C.-Z. Wei, "Adaptive control with the stochastic approximation algorithm: Geometry and convergence," *IEEE Trans. Autom. Control*, vol. AC-30, no. 4, pp. 330–338, Apr. 1985.
[9] P. Kumar, "Convergence of adaptive control schemes using least-squares parameter estimates," *IEEE Trans. Autom. Control*, vol. 35, no. 4, pp. 416–424, Apr. 1990.
[10] M. C. Campi and P. Kumar, "Adaptive linear quadratic Gaussian control: The cost-biased approach revisited," *SIAM J. Control Optim.*, vol. 36, no. 6, pp. 1890–1907, 1998.
[11] S. Bittanti and M. C. Campi, "Adaptive control of linear time invariant systems: The 'bet on the best' principle," *Commun. Inf. Syst.*, vol. 6, no. 4, pp. 299–320, 2006.
[12] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.
[13] M. C. Campi, "Achieving optimality in adaptive control: The 'bet on the best' approach," in *Proc. 36th IEEE Conf. Decis. Control*, 1997, vol. 5, pp. 4671–4676.
[14] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 1–26.
[15] M. Ibrahimi, A. Javanmard, and B. V. Roy, "Efficient reinforcement learning for high dimensional linear quadratic systems," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 2636–2644.
[16] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Belmont, MA, USA: Athena Scientific, 1995.
[17] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.
[18] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.
[19] B. M. Brown *et al.*, "Martingale central limit theorems," *Ann. Math. Statist.*, vol. 42, no. 1, pp. 59–66, 1971.
[20] M. Green and J. B. Moore, "Persistence of excitation in linear systems," in *Proc. IEEE Amer. Control Conf.*, 1985, pp. 412–417.
[21] T. L. Lai and C.-Z. Wei, "Extended least squares and their applications to adaptive control and prediction in linear systems," *IEEE Trans. Autom. Control*, vol. AC-31, no. 10, pp. 898–906, Oct. 1986.
[22] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.
[23] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On adaptive linear-quadratic regulators," *Automatica*, vol. 117, p. 108982, 2020.