Automatica 117 (2020) 108982

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

On adaptive Linear–Quadratic regulators☆

Mohamad Kazem Shirani Faradonbeh [a,*], Ambuj Tewari [b], George Michailidis [a]

[a] Informatics Institute, University of Florida, Gainesville, FL, 32611, USA
[b] Department of Statistics, University of Michigan, Ann Arbor, MI, 48109, USA

ARTICLE INFO

Article history:
Received 27 June 2018
Received in revised form 12 December 2019
Accepted 20 March 2020
Available online 11 April 2020

Keywords:
Regret analysis
Certainty Equivalence
Randomized algorithms
Thompson sampling
System identification
Adaptive policies

ABSTRACT

Performance of adaptive control policies is assessed through the regret with respect to the optimal regulator, which reflects the increase in the operating cost due to uncertainty about the dynamics parameters. However, available results in the literature do not provide a quantitative characterization of the effect of the unknown parameters on the regret. Further, there are problems regarding the efficient implementation of some of the existing adaptive policies. Finally, results regarding the accuracy with which the system's parameters are identified are scarce and rather incomplete.

This study aims to comprehensively address these three issues. First, by introducing a novel decomposition of adaptive policies, we establish a *sharp* expression for the regret of an *arbitrary* policy in terms of the deviations from the optimal regulator. Second, we show that adaptive policies based on slight modifications of the Certainty Equivalence scheme are efficient. Specifically, we establish a regret of (nearly) square-root rate for two families of randomized adaptive policies. The presented regret bounds are obtained by using *anti-concentration* results on the random matrices employed for randomizing the estimates of the unknown parameters. Moreover, we study the minimal additional information on dynamics matrices that using them the regret will become of logarithmic order. Finally, the rates at which the unknown parameters of the system are being identified are presented.

1. Introduction

This work studies the problem of designing *adaptive* policies for the following Linear–Quadratic (LQ) system. Given an initial state $x(0) \in \mathbb{R}^p$, the system evolves as

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

for $t \geq 0$, where the vector $x(t) \in \mathbb{R}^p$ corresponds to the state (and also output) of the system at time t, $u(t) \in \mathbb{R}^r$ is the control input, and $\{w(t)\}_{t=1}^{\infty}$ denotes a sequence of random disturbances. Further, the instantaneous quadratic cost of the control law π is denoted by

$$c_t(\pi) = x(t)'Qx(t) + u(t)'Ru(t), \tag{2}$$

where $Q \in \mathbb{R}^{p\times p}$, $R \in \mathbb{R}^{r\times r}$ are symmetric positive definite matrices, and $x(t)'$, $u(t)'$ denote the transpose of the vectors $x(t)$, $u(t)$. The dynamics of the system, i.e., both the transition matrix $A_0 \in \mathbb{R}^{p\times p}$, as well as the input matrix $B_0 \in \mathbb{R}^{p\times r}$, are fixed and *unknown*, while Q, R are assumed known. The overall objective is to adaptively regulate the system in order to minimize its long-term average cost.

Although regulation of LQ systems represents a canonical problem in optimal control, adaptive policies have not been adequately studied in the literature. In fact, a large number of classical papers focuses on the setting of adaptive tracking, where the objective is to steer the system to track a reference trajectory (Bercu, 1995; Chen & Zhang, 1989; Guo, 1995; Guo & Chen, 1988, 1991; Kumar, 1990; Lai, 1986; Lai & Wei, 1986; Lai & Ying, 1991). So, because the operating cost is not directly a function of the control signal (i.e., $R = 0$), analysis of adaptive regulators becomes different and less technically involved. Therefore, existing results are not applicable to general LQ systems, wherein both the state and the control input impact the operating cost. The adaptive Linear–Quadratic Regulators (LQR) problem has been studied in the literature (Abbasi-Yadkori & Szepesvári, 2011; Abeille & Lazaric, 2017; Bittanti & Campi, 2006; Campi & Kumar, 1998; Duncan, Guo, & Pasik-Duncan, 1999; Faradonbeh, Tewari, & Michailidis, 2017b; Ibrahimi, Javanmard, & Roy, 2012; Ouyang, Gagrani, & Jain, 2017), but there are still gaps that the present work aims to fill by addressing cost optimality, parameter estimation, and the trade-off between identification and control.

Since the system's dynamics are unknown, learning the key parameters A_0, B_0 is needed for designing an optimal regulation policy. However, the system operator needs to apply some

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Martin Guay under the direction of Editor Miroslav Krstic.

* Corresponding author.
E-mail addresses: mfaradonbeh@ufl.edu (M.K. Shirani Faradonbeh), tewaria@umich.edu (A. Tewari), gmichail@ufl.edu (G. Michailidis).

https://doi.org/10.1016/j.automatica.2020.108982
0005-1098/

control inputs, in order to collect data (observations) for parameter estimation. A popular approach to design an adaptive regulator is Certainty Equivalence (CE) (Bar-Shalom & Tse, 1974). Intuitively, its prescription is to apply a control policy *as if* the estimated parameters are the true ones guiding the system's evolution. In general, the inefficiency (as well as the inconsistency) of CE (Becker, Kumar, & Wei, 1985; Bittanti & Campi, 2006; Lai & Wei, 1982) has led researchers to consider several modifications of the CE approach.

One idea is to use the principle of Optimism in the Face of Uncertainty (OFU) (Abbasi-Yadkori & Szepesvári, 2011; Faradonbeh et al., 2017b; Ibrahimi et al., 2012) (also known as bet on the best (Bittanti & Campi, 2006), and the cost-biased approach (Campi & Kumar, 1998)). OFU recommends to apply the optimal regulators by treating *optimistic* approximations of the unknown matrices as the true dynamics (Lai & Robbins, 1985). Another idea is to replace the point estimate of the system parameters by a posterior distribution which is obtained through Bayes law by integrating a prior distribution and the likelihood of the data collected so far. One then draws a sample from this posterior distribution and applies the optimal policy, *as if* the system evolves according to the sampled dynamics matrices. This approach is known as Thompson (or posterior) sampling (Abeille & Lazaric, 2017; Ouyang et al., 2017).

Note that most of the existing work in the literature is purely asymptotic in nature so that it establishes the convergence of the adaptive *average* cost to the optimal value. It includes adaptive LQRs based on the OFU principle (Bittanti & Campi, 2006; Campi & Kumar, 1998), as well as those based on the method of random perturbations being applied to continuous time Ito processes (Duncan et al., 1999). However, results on the speed of convergences are rare and rather incomplete. On the other hand, from the identification viewpoint, consistency of parameter estimates is lacking for general dynamics matrices (Polderman, 1986a, 1986b). Moreover, accuracy rates for estimation of system parameters are only provided for minimum-variance problems (Bercu, 1995; Guo, 1995). Indeed, the estimation rate for matrices describing the system's dynamics is not currently available for general LQ systems.

Since in many applications the effective horizon is finite, the aforementioned asymptotic analyses are practically less relevant. Thus, addressing the optimality of an adaptive strategy under more sensitive criteria is needed. For this purpose, one needs to comprehensively examine the *regret*; i.e., the cumulative deviation from the optimal policy. Regret analyses are thus far limited to recent work addressing OFU adaptive policies (Abbasi-Yadkori & Szepesvári, 2011; Faradonbeh et al., 2017b; Ibrahimi et al., 2012), and results for TS obtained under restricted conditions (Abeille & Lazaric, 2017; Ouyang et al., 2017). One issue with OFU is the computational intractability of finding an optimistic approximation of the true parameters, since it needs to solve lots of non-convex matrix optimization problems. More importantly, we show that the existing regret bounds (Abbasi-Yadkori & Szepesvári, 2011; Abeille & Lazaric, 2017; Faradonbeh et al., 2017b; Ibrahimi et al., 2012; Ouyang et al., 2017) can be achieved or improved through simpler adaptive regulators.

A key contribution of this work is a remarkably general result to address the performance of control policies. Namely, tailoring a novel method for regret decomposition, we utilize some results from martingale theory to establish Theorem 1. It provides a *sharp* expression for the regret of arbitrary regulators in terms of the deviations from the optimal feedback. Leveraging Theorem 1, we analyze two families of CE-based adaptive policies.

First, we show that the growth rate of the regret is (nearly) square-root in time (of the interaction with the system), if the CE regulator is properly *randomized*. Performance analyses are presented for both common approaches of additive randomization and posterior sampling. Then, the adaptive LQR problem is discussed when additional information (regarding the unknown dynamics parameters of the system) is available. In this case, a *logarithmic* rate for the regret of generalizations of CE adaptive policies is established, assuming that the available side information satisfies an identifiability condition. Examples of side information include constraints on the rank or the support of dynamics matrices, that in turn lead to optimality of the linear feedback regulator, if the closed-loop matrix is accurately estimated. Further, the identification performance of the corresponding adaptive regulators is also addressed. To the best of our knowledge, this work provides the first comprehensive study of CE-based adaptive LQRs, for both the identification and the regulation problem.

The remainder of the paper is organized as follows. The problem is formulated in Section 2. Then, we provide an expression for the regret of general adaptive policies in Section 3.1. Subsequently, the consistency of estimating the dynamics parameter is given in Section 3.2. In Section 4, we study the growth rate of the regret, as well as the accuracy of parameter estimation, for two randomization schemes. Finally, in Section 5 we study a general condition which leads to significant performance improvements in both regulation and identification.

Remark 1 (*Stochastic Statements*)**.** All probabilistic equalities and inequalities throughout this paper hold almost surely, unless otherwise explicitly mentioned.

The following notation will be used throughout this paper. For a matrix $A \in \mathbb{C}^{k\times\ell}$, A' denotes its transpose. When $k = \ell$, the smallest (respectively largest) eigenvalue of A (in magnitude) is denoted by $\lambda_{\min}(A)$ (respectively $\lambda_{\max}(A)$). For $v \in \mathbb{C}^d$, define the norm $\|v\| = \left(\sum_{i=1}^{d} |v_i|^2\right)^{1/2}$. We also use the following notation for the operator norm of matrices. For $A \in \mathbb{C}^{k\times\ell}$ let $|||A||| = \sup_{\|v\|=1} \|Av\|$. In order to show the dimension of the manifold $\mathcal{M}$ we employ $\dim(\mathcal{M})$. Finally, to indicate the order of magnitude, we use $a_n = O(b_n)$ whenever $\limsup_{n\to\infty} |a_n/b_n| < \infty$.

2. Problem formulation

We start by defining the adaptive LQR problem this work is addressing. The stochastic evolution of the system is governed by the dynamics (1), where for all $t \geq 1$, $w(t)$ is the vector of random disturbances satisfying: $\mathbb{E}[w(t)] = 0$, $\mathbb{E}\left[w(t)w(t)'\right] = C$, and $|\lambda_{\min}(C)| > 0$.

For the sake of simplicity, the noise vectors $\{w(t)\}_{t=1}^{\infty}$ are assumed to be independent over time t. The latter assumption is made to simplify the presentation, and generalization to martingale difference sequences (adapted to a filtration) is straightforward.[1] Further, the following moment condition for the noise process is assumed.

Assumption 1 (*Moment Condition*)**.** There is $\alpha > 4$, such that α-th moments exist: $\sup_{t\geq 1} \mathbb{E}\left[\|w(t)\|^{\alpha}\right] < \infty$.

In addition, we assume that the true dynamics of the underlying system are stabilizable, a minimal assumption for the optimal control problem to be well-posed.

Assumption 2 (*Stabilizability*)**.** The true dynamics $[A_0, B_0]$ is stabilizable: there exists a stabilizing feedback $L \in \mathbb{R}^{r\times p}$ such that $|\lambda_{\max}(A_0 + B_0 L)| < 1$.

[1] It suffices to replace the involved terms with those consisting of the conditional expressions (w.r.t. the corresponding filtration).

Note that Assumption 2 implies stabilizability in the average sense: $\limsup_{n\to\infty} n^{-1}\sum_{t=0}^{n}\|x(t)\|^2 < \infty$.

Definition 1. Henceforth, for $A \in \mathbb{R}^{p\times p}, B \in \mathbb{R}^{p\times r}$, we use θ to denote $[A, B]$. So, $\theta \in \mathbb{R}^{p\times q}$, where $q = p + r$.

We assume *perfect* observations; i.e., the output of the system corresponds to the state vector $x(t)$. Next, an admissible control policy is a mapping π that designs the input according to the dynamics matrices A_0, B_0, the cost matrices Q, R, and the history of the system:

$$u(t) = \pi\left(A_0, B_0, Q, R, \{x(i)\}_{i=0}^{t}, \{u(j)\}_{j=0}^{t-1}\right),$$

for all $t \geq 0$. An *adaptive* policy such as $\widehat{\pi}$, is oblivious to the dynamics parameter θ_0; i.e.,

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

When applying the policy π, the resulting instantaneous quadratic cost at time t defined in (2) is denoted by $c_t(\pi)$. For an arbitrary policy π, let $\overline{\mathcal{J}}_\pi(A_0, B_0)$ denote the expected average cost of the system: $\overline{\mathcal{J}}_\pi(A_0, B_0) = \limsup_{n\to\infty} n^{-1}\sum_{t=0}^{n-1}\mathbb{E}[c_t(\pi)]$. Note that the dependence of $\overline{\mathcal{J}}_\pi(\theta_0)$ to the known cost matrices Q, R is suppressed. Then, the optimal expected average cost is defined as $\mathcal{J}^\star(A_0, B_0) = \min_\pi \overline{\mathcal{J}}_\pi(A_0, B_0)$, where the minimum is taken over *all* admissible control policies. The following proposition provides an optimal policy for minimizing the average cost, based on the Riccati equations:

$$K(\theta) = Q + A'K(\theta)A - A'K(\theta)B\left(B'K(\theta)B + R\right)^{-1}B'K(\theta)A, \tag{3}$$

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

Accordingly, define the linear time-invariant policy $\pi^\star$:

$$\pi^\star: \quad u(t) = L(\theta_0)x(t), \quad t = 0, 1, 2, \ldots. \tag{5}$$

Proposition 1 (*Optimal Policy (Chan, Goodwin, & Sin, 1984); (De Souza, Gevers, & Goodwin, 1986; Faradonbeh, Tewari, & Michailidis, 2019)*). *If $[A_0, B_0]$ is stabilizable, (3) has a unique solution, and $\pi^\star$ defined in (5) is an optimal regulator. Conversely, if $K(\theta_0)$ is a solution of (3), $L(\theta_0)$ defined by (4) is a stabilizer.*

In the latter case of Proposition 1, the solution $K(\theta_0)$ is unique and $\pi^\star$ is an optimal regulator. Note that although $\pi^\star$ is the only optimal policy among the time-invariant feedback regulators, there are uncountably many time varying optimal controllers.

To rigorously set the stage, we denote the linear regulator $u(t) = L_t x(t)$ by $\pi = \{L_t\}_{t=0}^{\infty}$, where L_t is a $r\times p$ matrix determined according to $A_0, B_0, Q, R, \{x(i)\}_{i=0}^{t}, \{u(j)\}_{j=0}^{t-1}$. For time-invariant policy $\pi_0 = \{L_0\}_{t=0}^{\infty}$, we use π_0 and L_0 interchangeably. For an adaptive operator, the dynamics matrices A_0, B_0 are unknown. Hence, adaptive policy $\widehat{\pi} = \left\{\widehat{L}_t\right\}_{t=0}^{\infty}$ constitutes the linear feedbacks $u(t) = \widehat{L}_t x(t)$, where $\widehat{L}_t \in \mathbb{R}^{r\times p}$ is required to be determined according to $Q, R, \{x(i)\}_{i=0}^{t}, \{u(j)\}_{j=0}^{t-1}$. In order to measure the efficiency of an arbitrary regulator π, the resulting instantaneous cost will be compared to that of the optimal policy $\pi^\star$ defined in (5). Specifically, the *regret* of policy π at time n is defined as

$$\mathcal{R}_n(\pi) = \sum_{t=0}^{n-1}\left[c_t(\pi) - c_t\left(\pi^\star\right)\right]. \tag{6}$$

The comparison between adaptive control policies is made according to regret, which is the cumulative deviation of the instantaneous cost of the corresponding adaptive policy from that of the optimal controller $\pi^\star$.

An analogous expression for regret is previously used for the problem of adaptive tracking (Lai, 1986; Lai & Wei, 1986). An alternative definition of the regret that has been used in the existing literature (Abbasi-Yadkori & Szepesvári, 2011; Abeille & Lazaric, 2017; Faradonbeh et al., 2017b; Ibrahimi et al., 2012; Ouyang et al., 2017) is the cumulative deviations from the optimal *average* cost: $\sum_{t=0}^{n-1}[c_t(\pi) - \mathcal{J}^\star(\theta_0)]$. The expression above differs from $\mathcal{R}_n(\pi)$ by the term $\sum_{t=0}^{n-1}c_t(\pi^\star) - n\mathcal{J}^\star(\theta_0)$, which is studied in the following result.

Proposition 2. *We have*

$$\limsup_{n\to\infty}\frac{\sum_{t=0}^{n-1}c_t(\pi^\star) - n\mathcal{J}^\star(\theta_0)}{n^{1/2}\log n} < \infty.$$

Therefore, the aforementioned definitions for the regret are *indifferent*, as long as one can establish an upper bound of $O\left(n^{1/2}\right)$ magnitude (modulo a logarithmic factor) for either definition. However, defining the regret by (6) leads to more accurate analyses and tighter results (e.g. the regret specification of Theorem 1, and the logarithmic rate of Theorem 5). To proceed, we introduce the following definition.

Definition 2. For a stabilizable parameter $\theta \in \mathbb{R}^{p\times q}$, define $\widetilde{L}(\theta) = \left[I_p, L(\theta)'\right]' \in \mathbb{R}^{q\times p}$.

We can then express the closed-loop matrices based on $\theta, \widetilde{L}(\theta)$. For arbitrary stabilizable θ_1, θ_2, if one applies the optimal feedback matrix $L(\theta_1)$ to a system with dynamics parameter θ_2, the resulting closed-loop matrix is $A_2 + B_2L(\theta_1) = \theta_2\widetilde{L}(\theta_1)$.

3. General adaptive policies

Next, we study the properties of general adaptive regulators. First, we study the regulation viewpoint in Section 3.1, and examine the regret of arbitrary linear policies. Then, from an identification viewpoint, consistency of parameter estimation is considered in Section 3.2.

3.1. Regulation

The main result of this subsection provides an expression for the regret of an arbitrary (i.e., either adaptive or non-adaptive) policy. According to the following theorem, the regret of the regulator $\{L_t\}_{t=0}^{\infty}$ is of the same order as the summation of the squares of the deviations of the linear feedbacks L_t from $L(\theta_0)$. Note that it is stronger than the previously known result that expressed the regret as the summation of the deviations from $L(\theta_0)$ (not squared) (Abbasi-Yadkori & Szepesvári, 2011; Abeille & Lazaric, 2017; Faradonbeh et al., 2017b; Ibrahimi et al., 2012; Ouyang et al., 2017). As will be shown shortly, this difference changes the nature of both the lower-bound, as well as the upper-bound of the regret.

Theorem 1 (*Regret Specification*). *Suppose that $\pi = \{L_t\}_{t=0}^{\infty}$ is a linear policy. Letting $\{x^\star(t)\}_{t=0}^{\infty}$ be the trajectory under the optimal policy $\pi^\star$, we have*

$$0 < \liminf_{n\to\infty}\frac{\mathcal{R}_n(\pi)}{\chi_n + \varrho_n} \leq \limsup_{n\to\infty}\frac{\mathcal{R}_n(\pi)}{\chi_n + \varrho_n} < \infty,$$

where $\varrho_n = x^\star(n)'K(\theta_0)x^\star(n) - x(n)'K(\theta_0)x(n)$, and $\chi_n = \sum_{t=0}^{n-1}\|(L(\theta_0) - L_t)x(t)\|^2$.

The above specification for the regret is remarkably general, since policy π does not need to satisfy any condition. Even for destabilized systems, the exponential growth of the state (and so the regret) is captured by χ_n. Conceptually, χ_n captures the effect of the *past* sub-optimality $\{L_t\}_{t=0}^{n-1}$ on the regret, while the influence of the sub-optimal feedback $\{L_t\}_{t=n}^{\infty}$ to be applied henceforth is reflected in ϱ_n. This is formally stated in the following

result, which also addresses the magnitude of $\|x^\star(n)\|$. According to Assumption 1, Corollary 1 shows that $\limsup_{n\to\infty} n^{-1/2}\varrho_n = 0$.

Corollary 1. *We have* $\limsup_{n\to\infty} n^{-\beta}\|x^\star(n)\| = 0$, *for all* $\beta > 1/\alpha$. *Further, letting* $L_t = L(\theta_0)$ *for* $t \geq n$, *and* $\pi = \{L_t\}_{t=0}^{\infty}$, *we get* $0 < \mathcal{R}_\infty(\pi)/\chi_\infty < \infty$.

Theorem 1 can be used for the sharp specification of the performance of adaptive regulators. The immediate consequence of Theorem 1 provides a tight upper bound for the regret of an adaptive policy, in terms of the linear feedbacks. Indeed, since the presented result is bidirectional and not just an upper bound, it will also provide a general information theoretic lower bound for the regret of an adaptive regulator. For stabilized dynamics, it is shown that the smallest estimation error when using a sample of size t is at least of the order $t^{-1/2}$ (Simchowitz, Mania, Tu, Jordan, & Recht, 2018). Thus, at time t, the error in the identification of the unknown dynamics parameter θ_0 is *at least* of the same order. Therefore, for the minimax growth rate of the regret, Theorem 1 implies the lower bound $\log n$.

In other words, for an arbitrary adaptive policy $\widehat{\pi}$, it holds that $\liminf_{n\to\infty}(\log n)^{-1}\mathcal{R}_n(\widehat{\pi}) > 0$. In general, the information theoretic lower bound above is not known to be *operationally* achievable because of the common trade-off between estimation and control. We will discuss the reasoning behind the presence of such a gap in Section 4, which leads to the operational lower bound $\liminf_{n\to\infty} n^{-1/2}\mathcal{R}_n(\widehat{\pi}) > 0$. Nevertheless, in Section 5 we discuss settings where availability of some side information leads to an achievable regret of logarithmic order.

Next, we provide some intuition behind Theorem 1 and Corollary 1. The expression is in nature similar to the concept of memorylessness, as discussed below. The dynamics of the system in (1) indicate that the influence of non-optimal control inputs lasts forever. That is, if $L_{t_1}x(t_1) \neq L(\theta_0)x(t_1)$, then for all $t > t_1$, the state vector $x(t)$ deviates from the optimal trajectory $\{x^\star(t)\}_{t=0}^{\infty}$, and future control inputs $\{u(t)\}_{t=t_1+1}^{\infty}$ cannot fully *compensate* this deviation. However, according to Theorem 1, the regret is dominated by the magnitude of the square of the deviations of the non-optimal feedbacks from $L(\theta_0)$. In other words, if switching to the optimal feedback $L(\theta_0)$ occurs, then the regret remains of the same order of the effect of the non-optimal control inputs previously applied, and so is memoryless.

3.2. Identification

Another consideration for an adaptive policy is the estimation (learning) problem. Since in general the operator has *no knowledge* regarding the dynamics parameter θ_0, a natural question to address is that of identifying θ_0, in addition to examining cost optimality. In this subsection, we address the asymptotic estimation consistency of general adaptive policies. That is, a rigorous formulation of the relationship between the *estimable* information (through observing the state of the system), and the *desired* optimality manifold is provided.

On one hand, for a linear feedback L, the best one can do by observing the state vectors is "closed-loop identification" (Faradonbeh et al., 2017b; Kumar, 1990); i.e., estimating the closed-loop matrix $A_0 + B_0L$ accurately. On the other hand, an adaptive policy is *at least* desired to provide a sub-linear regret;

$$\limsup_{n\to\infty} \frac{\mathcal{R}_n(\widehat{\pi})}{n} = 0. \tag{7}$$

The above two aspects of an adaptive policy provide the properties of the asymptotic uncertainty about the true dynamics parameter θ_0. By the uniqueness of $L(\theta_0)$ according to Proposition 1, the linear feedbacks of the adaptive policy $\widehat{\pi} = \{\widehat{L}_t\}_{t=0}^{\infty}$ require to converge to $L(\theta_0)$. Further, $\widehat{\pi}$ uniquely identifies the asymptotic closed-loop matrix $\lim_{t\to\infty} A_0 + B_0\widehat{L}_t$. This matrix according to (7) is supposed to be $\theta_0\widetilde{L}(\theta_0)$. Putting the above together, the asymptotic uncertainty is reduced to the set of parameters θ_∞ that satisfy

$$L(\theta_\infty) = L(\theta_0), \quad \theta_\infty\widetilde{L}(\theta_0) = \theta_0\widetilde{L}(\theta_0). \tag{8}$$

To rigorously analyze this uncertainty, we introduce some additional notation. First, for an arbitrary stabilizable θ_1, introduce the shifted null-space of the linear transformation $\widetilde{L}(\theta_1) : \mathbb{R}^{p\times q} \to \mathbb{R}^{p\times p}$ by $\mathcal{N}(\theta_1)$ as:

$$\mathcal{N}(\theta_1) = \left\{\theta \in \mathbb{R}^{p\times q} : \theta\widetilde{L}(\theta_1) = \theta_1\widetilde{L}(\theta_1)\right\}. \tag{9}$$

So, $\mathcal{N}(\theta_1)$ is indeed the set of parameters θ, such that the closed-loop transition matrix of two systems with dynamics parameters θ, θ_1 will be the same, if applying the optimal linear regulator in (4) calculated for θ_1. Hence, if the operator regulates the system by feedback $L(\theta_1)$, one cannot identify θ, θ_1. In other words, $\mathcal{N}(\theta_1)$ is the *learning capability* of adaptive regulators. Then, we define the *desired planning* of adaptive policies as follows. For an arbitrary stabilizable θ_1, define $\mathcal{S}(\theta_1)$ as the level-set of the optimal controller function (4), which maps $\theta \in \mathbb{R}^{p\times q}$ to $L(\theta) \in \mathbb{R}^{r\times p}$:

$$\mathcal{S}(\theta_1) = \left\{\theta \in \mathbb{R}^{p\times q} : L(\theta) = L(\theta_1)\right\}. \tag{10}$$

Therefore, $\mathcal{S}(\theta_1)$ is in fact the set of parameters θ, such that the calculation of optimal linear regulator (4) provides the same feedback matrix for both θ, θ_1. Intuitively, $\mathcal{N}(\theta_0)$ reflects the identification aspect of the adaptive regulators by specifying the accuracy of the parameter estimation procedure. Similarly, $\mathcal{S}(\theta_0)$ reflects the control aspect, and specifies the regulation performance in terms of optimality of the cost minimization procedure. Hence, the asymptotic uncertainty about the true parameter θ_0 is according to (8) limited to the set

$$\mathcal{P}_0 = \mathcal{S}(\theta_0) \cap \mathcal{N}(\theta_0). \tag{11}$$

The system theoretic interpretation is as follows. Assuming (7), $\mathcal{P}_0$ is the smallest subset of dynamics parameters θ that one can identify according to the state and the input sequences. Thus, the consistency of identifying the true dynamics parameter θ_0 is equivalent to $\mathcal{P}_0 = \{\theta_0\}$. The following result establishes the properties of $\mathcal{P}_0$, and will be used later to discuss the operational optimality of adaptive regulators. It generalizes some results in the literature (Polderman, 1986a, 1986b).

Theorem 2 (*Consistency*). *The set* $\mathcal{P}_0$ *defined in* (11) *is a shifted linear subspace of dimension* $\dim(\mathcal{P}_0) = (p - \operatorname{rank}(A_0))\, r$.

Therefore, consistency of estimating θ_0 is *automatically* guaranteed for an adaptive policy with a sublinear regret, only if A_0 is a full-rank matrix. In other words, effective control (exploitation) suffices for consistent estimation (exploration) only if $\operatorname{rank}(A_0) = p$. For example, the sublinear regret bounds of OFU (Abbasi-Yadkori & Szepesvári, 2011; Faradonbeh et al., 2017b) imply consistency, assuming A_0 is of the full rank. Intuitively, a singular A_0 precludes unique identification of both of A_0, B_0 by (8). Note that the converse is always true: consistency of parameter estimation implies the sublinearity of the regret. Clearly, full-rankness of A_0 holds for almost all θ_0 (with respect to Lebesgue measure).

4. Randomized adaptive policies

The classical idea to design an adaptive policy is the following procedure known as CE. At every time n, its prescription is to apply the optimal regulator provided by (4), as if the estimated parameter $\widehat{\theta}_n$ coincides exactly with the truth θ_0. According to (1),

a natural estimation procedure is to linearly regress $x(t+1)$ on the covariates $x(t), u(t)$, using all observations collected so far; $0 \leq t \leq n-1$. Formally, the CE policy is $\{L(\widehat{\theta}_n)\}_{n=1}^{\infty}$, where $\widehat{\theta}_n$ is a solution of the least-squares estimator using the data observed until time n. That is,

$$\widehat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^{p\times q}} \sum_{t=0}^{n-1} \left\| x(t+1) - \theta \widetilde{L}(\widehat{\theta}_t)\, x(t) \right\|^2.$$

The issue with CE is that it is capable of *adapting* to a non-optimal regulation. Technically, CE possibly fails to *falsify* an incorrect estimation of the true parameter (Bittanti & Campi, 2006). Suppose that at time n, the hypothetical estimate of the true parameter is $\widehat{\theta}_n \neq \theta_0$. When applying the linear feedback $L(\widehat{\theta}_n)$, the true closed-loop transition matrix will be $\theta_0 \widetilde{L}(\widehat{\theta}_n)$. Then, if this matrix is the same as the (falsely) assumed closed-loop transition matrix $\widehat{\theta}_n \widetilde{L}(\widehat{\theta}_n)$, the estimation procedure can fail to falsify $\widehat{\theta}_n$. So, if $L(\widehat{\theta}_n) \neq L(\theta_0)$, the adaptive policy is not guaranteed to tend toward a better control feedback, and a non-optimal regulator will be persistently applied.

Fortunately, if slightly modified, CE can avoid *unfalsifiable* approximations of the true parameters. More precisely, we show that the set of unfalsifiable parameters defined below is of zero Lebesgue measure;

$$\mathcal{U}(\theta_0) = \left\{ \theta \in \mathbb{R}^{p\times q} : \theta_0 \widetilde{L}(\theta) = \theta \widetilde{L}(\theta) \right\}. \tag{12}$$

Note that by (9), $\theta_1 \in \mathcal{U}(\theta_2)$ if and only if $\theta_2 \in \mathcal{N}(\theta_1)$. Recalling the discussion in the previous section, $\mathcal{N}(\theta_1)$ captures the estimation ability of adaptive regulators. That is, the set $\mathcal{U}(\theta_0)$ contains the matrices θ for which the hypothetically assumed closed-loop matrix is indistinguishable from the true one. The next lemma sets the stage for the subsequent results which show that CE can be efficient, if it is suitably randomized.

Lemma 1 (*Unfalsifiable Set*)**.** *The set $\mathcal{U}(\theta_0)$ defined in (12) has Lebesgue measure zero.*

4.1. Randomized certainty equivalence

According to Lemma 1, we can avoid the pathological set $\mathcal{U}(\theta_0)$. As subsequently explained, it suffices to randomize the least-squares estimates of θ_0, with a small (diminishing) perturbation. First, such perturbations are chosen to be continuously distributed over the parameter space $\mathbb{R}^{p\times q}$, in order to evade $\mathcal{U}(\theta_0)$. Further, since the linear transformation $\widetilde{L}(\widehat{\theta}_n)$ is randomly perturbed, we can estimate the unknown dynamics parameter θ_0. Note that as discussed in the previous section, the sequence $\{\widetilde{L}(\widehat{\theta}_n)\}_{n=0}^{\infty}$ relates the estimation of θ_0 to the accurate identification of the closed-loop matrix $\theta_0 \widetilde{L}(\widehat{\theta}_n)$. Finally, according to Theorem 1, the magnitude of the random perturbation needs to diminish sufficiently fast. Indeed, while a larger magnitude perturbation helps to the improvement of estimation, an efficient regulation requires it to be sufficiently small. Addressing this trade-off is the common dilemma of adaptive control. At the end of this section, we will examine this trade-off based on properties of estimation methods and the tight specification of the regret in Theorem 1.

In the sequel, we present the *Randomized Certainty Equivalence* (RCE) adaptive regulator. RCE is an episodic algorithm as follows. First, when identifying a linear dynamical system using n observations, the estimation accuracy scales at rate $n^{-1/2}$. Therefore, one can defer updating of the parameter estimates until collecting sufficiently more data. This leads to the episodic adaptive policies, where the linear feedbacks are updated *only* after episodes of exponentially growing lengths (Faradonbeh et al., 2017b). In RCE, the randomization of the parameter estimate is episodic as well. Thus, calculation of the linear feedbacks $L(\widehat{\theta}_n)$ by (4) will occur sparsely (only $O(\log n)$ times, instead of n times), which remarkably reduces the computational cost of the algorithm.

Algorithm 1: RCE

```
Input: γ > 1, and σ_0 > 0
Let L(θ̂_0) be a stabilizer
for m = 0, 1, 2, ··· do
    while n < ⌊γ^m⌋ do
        Apply u(n) = L(θ̂_n) x(n)
        θ̂_{n+1} = θ̂_n
    end while
    Update the estimate θ̂_n by (13)
end for
```

To formally define RCE, let $\{\phi_m\}_{m=0}^{\infty}$ be a sequence of i.i.d. $p\times q$ random matrices with independent $\boldsymbol{N}(0, \sigma_0^2)$ entries, for a fixed $\sigma_0 > 0$. This sequence will be used to randomize the estimates. RCE has an arbitrary parameter $\gamma > 1$ for determining the lengths of the episodes, and starts by an arbitrary initial estimate $\widehat{\theta}_0$ such that $L(\widehat{\theta}_0)$ stabilizes the system. To find such initial estimates, one can employ the existing adaptive algorithm to stabilize the system in a short period (Faradonbeh et al., 2019). Later on, we will briefly discuss the aforementioned stabilization algorithm. Then, for each time $n \geq 0$, we apply the linear feedback $L(\widehat{\theta}_n)$. If n satisfies $n = \lfloor \gamma^m \rfloor$ for some $m \geq 0$, we update the estimate by

$$\widehat{\theta}_n = \widetilde{\theta}_n + \arg\min_{\theta \in \mathbb{R}^{p\times q}} \sum_{t=0}^{n-1} \left\| x(t+1) - \theta \widetilde{L}(\widehat{\theta}_t)\, x(t) \right\|^2, \tag{13}$$

where $\widetilde{\theta}_n = \left(n^{-1/4} \log^{1/4} n\right) \phi_m$ is the random perturbation. Otherwise, for $n \neq \lfloor \gamma^m \rfloor$, the policy does not update the estimates: $\widehat{\theta}_n = \widehat{\theta}_{n-1}$. Note that since the distribution of $\widetilde{\theta}_n$ over $p\times q$ matrices is absolutely continuous with respect to Lebesgue measure, $\widehat{\theta}_n$ is stabilizable (as well as controllable (Bertsekas, 1995; Faradonbeh, Tewari, & Michailidis, 2017a)). Therefore, by Proposition 1, the adaptive feedback $L(\widehat{\theta}_n)$ is well defined.

Remark 2 (*Non-Gaussian Randomization*)**.** In general, it suffices to draw $\{\phi_m\}_{m=0}^{\infty}$ from an arbitrary distribution with bounded probability density functions on $\mathbb{R}^{p\times q}$ such that $\sup_{m\geq 1} \mathbb{E}\left[\|\phi_m\|^{4+\epsilon} \right] < \infty$, for some $\epsilon > 0$.

As mentioned before, the rate γ determines the lengths of the episodes during which the algorithm uses $\widehat{\theta}_n$, before updating the estimate. Smaller values of γ correspond to shorter episodes and thus more updates and additional randomization; i.e., the smaller γ is, the better the estimation performance of RCE is. Although we will shortly see that such an improvement will not provide a better asymptotic rate for the regret, it speeds up the convergence and so is suitable if the actual time horizon is not very large. Further, it increases the number of times the Riccati equation (4) needs to be computed. Therefore, in practice the operator can decide γ according to the time length of interacting with the system, and the desired computational complexity. It is important especially if the evolution of the real-world plant under control requires the feedback policy to be updated fast (compared to the time the operator needs to calculate the linear feedback). The following theorem addresses the behavior of RCE, and shows that adaptive policies based on OFU (Abbasi-Yadkori & Szepesvári, 2011; Faradonbeh et al., 2017b; Ibrahimi et al., 2012) do not provide a better rate for the regret, while they impose a large computational burden by requiring solving a matrix optimization problem.

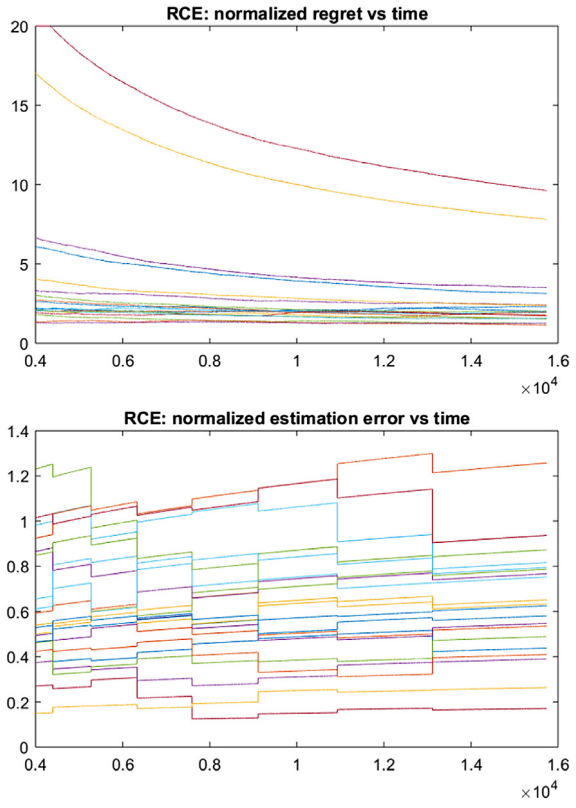

Fig. 1. RCE performance: normalized regret $\left(n^{-1/2}\log^{-1} n\right)\mathcal{R}_n\left(\widehat{\pi}\right)$ vs n (top), and normalized estimation error $\left(n^{1/4}\log^{-1/2} n\right)\left\|\left|\widehat{\theta}_n-\theta_0\right|\right\|$ vs n (bottom).

Theorem 3 (*RCE Rates*). *Suppose that $\widehat{\pi}$ is RCE, and $\widehat{\theta}_n$ is the parameter estimate at time n. Then, we have*

$$\limsup_{n\to\infty}\frac{\mathcal{R}_n\left(\widehat{\pi}\right)}{n^{1/2}\log n}<\infty,\ \limsup_{n\to\infty}\frac{\left\|\left|\widehat{\theta}_n-\theta_0\right|\right\|^2}{n^{-1/2}\log n}<\infty.$$

Note that the analysis of RCE strongly leverages the specification of the regret presented in Theorem 1. Fig. 1 illustrates the results of Theorem 3 by depicting the performance of RCE for $\gamma=1.2$, and the dynamics and cost matrices

$$A_0=\begin{bmatrix}1.04 & 0 & -0.27\\ 0.52 & -0.81 & 0.83\\ 0 & 0.04 & -0.90\end{bmatrix},\quad B_0=\begin{bmatrix}-0.47 & 0.61 & -0.29\\ -0.50 & 0.58 & 0.25\\ 0.29 & 0 & -0.72\end{bmatrix},$$
$$Q=\begin{bmatrix}0.65 & -0.08 & -0.14\\ -0.08 & 0.57 & 0.26\\ -0.14 & 0.26 & 2.50\end{bmatrix},\quad R=\begin{bmatrix}0.20 & 0.05 & 0.08\\ 0.05 & 0.14 & 0.04\\ 0.08 & 0.04 & 0.24\end{bmatrix}. \tag{14}$$

Curves of the normalized values of both the regret and the estimation error are depicted as a function of time, with the colors of the various curves corresponding to different replicates of the stochastic dynamics, as well as the adaptive policy RCE.

4.2. Thompson sampling

Another approach in existing literature is *Thompson Sampling* (TS), which has the following Bayesian interpretation. Applying an initial stabilizing linear feedback, TS updates the estimate $\widehat{\theta}_n$ through posterior sampling. That is, the operator draws a realization $\widehat{\theta}_n$ of the Gaussian posterior for which the mean and the covariance matrix are determined by the data observed to date.

Formally, let $\Sigma_0\in\mathbb{R}^{q\times q}$ be a fixed positive definite (PD) matrix, and choose a coarse approximation $\mu_0\in\mathbb{R}^{p\times q}$ of the truth θ_0. We will shortly explain an algorithmic procedure for computing such coarse approximations. Further, similar to RCE, fix the rate $\gamma>1$. Then, at each time $n\geq 0$, we apply $L\left(\widehat{\theta}_n\right)$, where $\widehat{\theta}_n$ is designed as follows. If n satisfies $n=\lfloor\gamma^m\rfloor$ for some $m\geq 0$, $\widehat{\theta}_n$ is drawn from a Gaussian distribution $\boldsymbol{N}\left(\mu_m,\Sigma_m^{-1}\right)$, where

$$\mu_m=\arg\min_{\mu\in\mathbb{R}^{p\times q}}\sum_{t=0}^{\lfloor\gamma^m\rfloor-1}\left\|x(t+1)-\mu\widetilde{L}\left(\widehat{\theta}_t\right)x(t)\right\|^2, \tag{15}$$

$$\Sigma_m=\Sigma_0+\sum_{t=0}^{\lfloor\gamma^m\rfloor-1}\widetilde{L}\left(\widehat{\theta}_t\right)x(t)x(t)'\widetilde{L}\left(\widehat{\theta}_t\right)'. \tag{16}$$

Algorithm 2: TS

Input: $\gamma>1$
Let $\Sigma_0\in\mathbb{R}^{q\times q}$ be PD, and $L\left(\widehat{\theta}_0\right)$ be a stabilizer
for $m=0,1,2,\cdots$ **do**
  while $n<\lfloor\gamma^m\rfloor$ **do**
    Apply $u(n)=L\left(\widehat{\theta}_n\right)x(n)$
    $\widehat{\theta}_{n+1}=\widehat{\theta}_n$
  end while
  Calculate μ_m, Σ_m by (15), (16)
  Draw all rows of $\widehat{\theta}_n$ from $\boldsymbol{N}\left(\mu_m,\Sigma_m^{-1}\right)$
end for

Namely, for $1\leq i\leq p$, the ith row of $\widehat{\theta}_n$ is drawn independently from a multivariate Gaussian distribution of mean $\mu_m^{(i)}$ (the ith row of μ_m), and covariance matrix Σ_m^{-1}. Otherwise, for $n\neq\lfloor\gamma^m\rfloor$ the policy does not update: $\widehat{\theta}_n=\widehat{\theta}_{n-1}$. Clearly, μ_m is the least-squares estimate and Σ_m is the (unnormalized) empirical covariance of the data observed by the end of episode m. Note that unlike RCE, the randomization in TS is based on the state and control signals. The following result establishes the performance rates for TS.

Theorem 4 (*TS Rates*). *Let the adaptive policy $\widehat{\pi}$ be TS, and the parameter estimate be $\widehat{\theta}_n$. Then, we have*

$$\limsup_{n\to\infty}\frac{\mathcal{R}_n\left(\widehat{\pi}\right)}{n^{1/2}\log^2 n}<\infty,\ \limsup_{n\to\infty}\frac{\left\|\left|\widehat{\theta}_n-\theta_0\right|\right\|^2}{n^{-1/2}\log^2 n}<\infty.$$

Note that the above upper-bounds differ by those of Theorem 3 by a logarithmic factor. The performance of TS for $\gamma=1.2$, and the matrices A_0,B_0,Q,R in (14) is depicted in Fig. 2. Clearly, the curves of the normalized regret and the normalized estimation error in Fig. 2 fully reflect the rates of Theorem 4. For TS based adaptive LQRs, the *Bayesian regret* (i.e., the *expected* value of the regret, wherein the expectation is taken under the assumed prior) has been shown to be of a similar magnitude (Ouyang et al., 2017). Of course, this heavily relies on a Gaussian prior imposed on the true θ_0, and the (non-Bayesian) regret is known to be of $O\left(n^{2/3}\right)$ magnitude (Abeille & Lazaric, 2017). Therefore, Theorem 4 provides an improved regret bound for TS, thanks to Theorem 1. By assuming stronger assumptions (e.g. boundedness of the state), a similar result has been recently established for the case $p=1$, which holds uniformly over time (Abeille & Lazaric, 2018).

For the sake of completeness, we briefly discuss an existing adaptive stabilization procedure that one can employ before utilizing RCE or TS. First, in the work of Faradonbeh et al. (Faradonbeh et al., 2019), it is shown that for some fixed $\epsilon_0>0$, a coarse approximation $\widehat{\theta}_0$ that satisfies $\left\|\left|\widehat{\theta}_0-\theta_0\right|\right\|\leq\epsilon_0$, is sufficient for stabilizing the system (Faradonbeh et al., 2019).

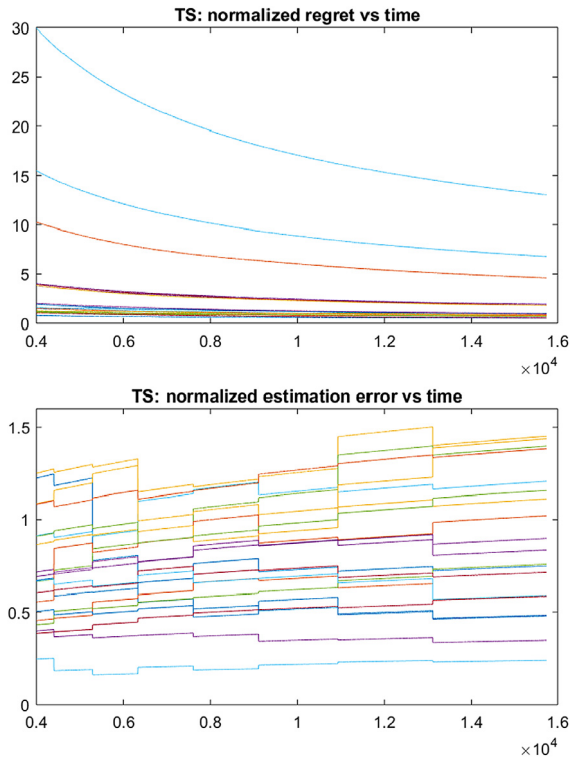

Fig. 2. TS performance: normalized regret $\left(n^{-1/2}\log^{-1}n\right)\mathcal{R}_n\left(\widehat{\pi}\right)$ vs n (top), and normalized estimation error $\left(n^{1/4}\log^{-1/2}n\right)\left\|\left|\widehat{\theta}_n-\theta_0\right|\right\|$ vs n (bottom).

Note that the closed-loop matrix can be unstable before termination of an stabilization procedure. On the other hand, there exists a pathological subset of unstable matrices such that if the closed-loop transition matrix belongs to that subset, it is not feasible to be accurately estimated (Faradonbeh, Tewari, & Michailidis, 2018a). Specifically, in order to ensure consistency, the true unstable closed-loop transition matrix during the stabilization period needs to be *regular*, as defined below (Faradonbeh et al., 2018a). The unstable square matrix D is *regular* if the eigenspaces corresponding to the eigenvalues of D outside the unit circle are one dimensional (Faradonbeh et al., 2018a). Then, it is established that random linear feedback matrices preclude the closed-loop irregularity (Faradonbeh et al., 2019). Therefore, the method of random feedback matrices guarantees that a coarse approximation of θ_0 is achievable in finite time, and a stabilization set can be constructed (Faradonbeh et al., 2019). Thus, we assume that the initial linear feedback matrix $L\left(\widehat{\theta}_0\right)$ is a stabilizer (i.e., $\left|\lambda_{\max}\left(\theta_0\widetilde{L}\left(\widehat{\theta}_0\right)\right)\right|<1$), and the system remains stable when RCE or TS is being employed. More details for establishing finite time adaptive stabilization are provided in the aforementioned Ref. Faradonbeh et al. (2019). As a matter of fact, closed-loop regularity is not guaranteed, if only the control signals $\{u(t)\}_{t=0}^{\infty}$ are randomized. Further, the classical framework of persistent excitation is not applicable due to the possible instability of the closed-loop matrix (Anderson, 1985; Faradonbeh et al., 2018a; Johnstone & Anderson, 1983; Zhang, 1990).

4.3. Optimality

Next, we discuss the reason for the presence of a significant gap between the operational regrets of Theorems 3 and 4, and the information theoretic lower bound mentioned in Section 3.1. In fact, the following discussion shows that the logarithmic lower bound is *not* practically achievable. Nevertheless, in the next section we show how using additional information for the true dynamics parameter yields a regret of logarithmic order. In the sequel, we discuss an argument that leads to the following conjecture: the regret is operationally of order $n^{1/2}$. For this purpose, we first state the following lemma about the level-set manifold $\mathcal{S}(\theta_0)$ defined in (10). It is a generalization of a previously established result for full-rank matrices (Polderman, 1986a, 1986b).

Lemma 2 (*Optimality Manifold*). *The optimality level-set* $\mathcal{S}(\theta_0)$ *is a manifold of dimension* $\dim(\mathcal{S}(\theta_0)) = p^2+(p-\text{rank}(A_0))(r-\text{rank}(B_0))$ *at point* θ_0.

By Theorem 2, we have $\dim(\mathcal{S}(\theta_0))-\dim(\mathcal{P}_0)=k$, where $k=p^2-(p-\text{rank}(A_0))\,\text{rank}(B_0)$. The tangent space of the manifold $\mathcal{S}(\theta_0)$ at point θ_0, shares $(p-\text{rank}(A_0))\,r$ of its dimensions with $\mathcal{N}(\theta_0)$, and the other k dimensions are apart from $\mathcal{N}(\theta_0)$. Intuitively, $\mathcal{N}(\theta_0)$ reflects the constraint of estimating the dynamics parameter, and $\mathcal{S}(\theta_0)$ is the desired information to design an optimal policy. Thus, those k dimensions of $\mathcal{S}(\theta_0)$ which are not in $\mathcal{N}(\theta_0)$, *cannot* be estimated unless the subspace $\mathcal{N}(\theta_0)$ is sufficiently perturbed. Such a perturbation is available only through applying non-optimal feedbacks, which yields a larger regret than the logarithmic rate mentioned in Section 3.1.

Next, we carefully analyze the regret based on the limits in falsifying the parameters not belonging to $\mathcal{S}(\theta_0)$. First, inefficiency of an adaptive regulator compared to the optimal feedback $L(\theta_0)$ is determined by the uncertainty for the exact specification of the optimality manifold $\mathcal{S}(\theta_0)$. As an extreme example, suppose that $\mathcal{S}(\theta_0)$ is provided to an operator who does not know θ_0. Then, denoting the adaptive policy above by $\widehat{\pi}$, we have $\mathcal{R}_n(\widehat{\pi})=0$. Theorem 1 states that if at time n the adaptive regulator approximates $\mathcal{S}(\theta_0)$ with error ϵ_n, the growth in the regret is in magnitude ϵ_n^2. Thus, it suffices to examine the estimation accuracy ϵ_n that in turn depends both on the identification accuracy of the closed-loop transition matrix, as well as the falsification of dynamics parameters $\theta\notin\mathcal{S}(\theta_0)$.

Now, suppose that the objective is to falsify $\theta_1\in\mathcal{N}(\theta_0)$, such that $\left\|\left|\theta_1-\theta_0\right|\right\|=\sigma_n$, and $\theta_1-\theta_0$ is orthogonal to the linear manifold $\mathcal{P}_0$ defined in (11). The latter property of θ_1 dictates $\liminf_{n\to\infty}\sigma_n^{-1}\left\|\left|L(\theta_1)-L(\theta_0)\right|\right\|>0$. The key point is that in order to falsify θ_1, *non-optimal* linear feedbacks need to be applied *sufficiently many* times. For instance, if applying $L(\theta_0)$, the estimation provides $\mathcal{N}(\theta_0)$, i.e., θ_1 can never get falsified. More generally, assume that L is a δ_n-perturbation of the optimal feedback: $\left\|\left|L-L(\theta_0)\right|\right\|=\delta_n$. The shifted subspace of uncertainty when applying L deviates from $\mathcal{N}(\theta_0)$ by at most $O(\delta_n)$ (in the sense of inner products of the unit vectors). Next, assume that the operator applies L (or a similar δ_n-perturbed feedback) for a duration of n time points. Note that the closed-loop estimation error is at least of the order of $n^{-1/2}$ (Simchowitz et al., 2018). Thus, the operator can falsify θ_1 only if $\liminf_{n\to\infty}n^{1/2}\delta_n\sigma_n>0$. In other words, the adaptive regulator can avoid applying control feedbacks of distance at least $n^{-1/2}\delta_n^{-1}$ from the optimal feedback, only if control feedbacks of distance δ_n are in advance applied for a period of length n. Hence, we obtain $\liminf_{n\to\infty}\sigma_n^{-2}\left(\mathcal{R}_{n+1}(\widehat{\pi})-\mathcal{R}_n(\widehat{\pi})\right)>0$ by using Theorem 1, which also implies that such perturbed feedbacks impose a regret of the order $n\delta_n^2$. Putting together, we get $\liminf_{n\to\infty}\mathcal{R}_n(\widehat{\pi})\left(\mathcal{R}_{n+1}(\widehat{\pi})-\mathcal{R}_n(\widehat{\pi})\right)>0$. It leads to the following conjecture which constitutes an interesting direction for future work.

Conjecture 1 (*Lower Bound*). *For an arbitrary adaptive policy* $\widehat{\pi}$ *we have* $\liminf_{n\to\infty}n^{-1/2}\mathcal{R}_n(\widehat{\pi})>0$.

Note that if the above conjecture is true, RCE and TS provide a nearly optimal bound for the regret. Even the logarithmic gap between the lower and upper bounds is inevitable, due to the existence of an analogous gap in the closed-loop identification of linear systems (Simchowitz et al., 2018). Further, the above discussion explains the intuition behind the design of RCE. Specifically, the magnitude of the perturbation $\|\widetilde{\theta}_n\|$ according to the above discussion is optimally selected, since it satisfies $0 < \liminf_{n\to\infty} \sigma_n^{-1}\delta_n \leq \limsup_{n\to\infty} \sigma_n^{-1}\delta_n < \infty$, modulo a logarithmic factor. Indeed, if randomization is (significantly) smaller in magnitude than $n^{-1/4}$, the portion of the regret due to such a perturbation will reduce. However, it also reduces the accuracy of the parameter estimate. Thus, the other portion of the regret due to estimation error will increase. A similar discussion holds for larger magnitudes of the perturbation $\widetilde{\theta}_n$. On the other hand, the magnitude of randomization in TS is determined by the collected observations. As one can see in the proof of Theorem 4, a similar magnitude of randomization is automatically imposed by the structure of TS adaptive LQR.

5. Generalized certainty equivalence

It is possible that the operator has additional information on the dynamics. Examples of such information are the set of non-zero entries of θ_0, the rank of θ_0, or a plant whose subsystems evolve independently of each other. Another example comes from large network systems, where a substantial portion of the matrix θ_0 entries are zero (Faradonbeh et al., 2017a). Further, it is easy to see that the transition matrix of a system whose dynamics exhibit longer memory has a specific form (Faradonbeh et al., 2018a; Guo & Chen, 1991).

In such cases, this additional structural information on θ_0 can be used by the operator in order to obtain a smaller regret for the adaptive regulation of the system. Nevertheless, a comprehensive theory needs to formalize how this side information can provide theoretical sharp bounds for the regret. In this section, we provide an identifiability condition that ensures that the adaptive LQRs attain the informational lower bound of logarithmic order. In addition to the classical CE adaptive regulator, we also consider the family of CE-based schemes which provide a logarithmic order of magnitude for the regret.

First, we introduce the *Generalized Certainty Equivalence* (GCE) adaptive regulator. GCE is an episodic algorithm with exponentially growing duration of episodes. Instead of randomizing the parameter estimate similar to RCE and TS, in GCE the least-squares estimate is perturbed with an arbitrary matrix $\widetilde{\theta}_n$. Suppose that the operator knows that $\theta_0 \in \Gamma_0$, based on side information $\Gamma_0 \subset \mathbb{R}^{p\times q}$. Then, fixing the rate $\gamma > 1$, at time $n \geq 0$, we apply the controller $L\left(\widehat{\theta}_n\right)$. If n satisfies $n = \lfloor \gamma^m \rfloor$ for some $m \geq 0$, we update the estimate by

$$\widehat{\theta}_n = \widetilde{\theta}_n + \arg\min_{\theta\in\Gamma_0} \sum_{t=0}^{n-1} \left\|x(t+1) - \theta \widetilde{L}\left(\widehat{\theta}_t\right) x(t)\right\|^2, \tag{17}$$

where $\widetilde{\theta}_n$ is *arbitrary*, and satisfies $\limsup_{n\to\infty} n^{1/2}\|\widetilde{\theta}_n\| < \infty$. For $n \neq \lfloor \gamma^m \rfloor$ the policy does not update: $\widehat{\theta}_n = \widehat{\theta}_{n-1}$. Note that if $\widetilde{\theta}_n = 0$, we get the episodic CE adaptive regulator. To proceed, we define the following condition.

Definition 3 (*Identifiability*)**.** Suppose that there is $\Gamma_0 \subset \mathbb{R}^{p\times q}$ such that $\theta_0 \in \Gamma_0$. Then, θ_0 is identifiable, if for some $\beta_0 < \infty$ and all stabilizable $\theta_1, \theta_2 \in \Gamma_0$:

$$\|L(\theta_2) - L(\theta_0)\| \leq \beta_0 \left\|(\theta_2 - \theta_0)\widetilde{L}(\theta_1)\right\|. \tag{18}$$

Algorithm 3: GCE

Inputs: $\gamma > 1$, $\Gamma_0 \subset \mathbb{R}^{p\times q}$
Let $L\left(\widehat{\theta}_0\right)$ be a stabilizer
for $m = 0, 1, 2, \cdots$ **do**
  while $n < \lfloor \gamma^m \rfloor$ **do**
    Apply $u(n) = L\left(\widehat{\theta}_n\right) x(n)$
    $\widehat{\theta}_{n+1} = \widehat{\theta}_n$
  end while
  Update the estimate $\widehat{\theta}_n$ by (17)
end for

Intuitively, the definition above describes settings where side information Γ_0 is sufficient in the sense that an ϵ-accurate identification of the closed-loop matrix (the RHS of (18)) provides an $O(\epsilon)$-accurate approximation of the optimal linear feedback (the LHS of (18)). Subsequently, we provide concrete examples of Γ_0, such as presence of sparsity or low-rankness in θ_0. Essentially, a finite union of manifolds of proper dimension in the space $\mathbb{R}^{p\times q}$ suffices for identifiability. To see that, we use the critical subsets $\mathcal{N}(\theta_0)$, $\mathcal{S}(\theta_0)$, and $\mathcal{P}_0$ defined in (9), (10), and (11), respectively.

First, note that $\mathcal{P}_0 \subset \mathcal{S}(\theta_0)$ provides the optimal linear feedback $L(\theta_0)$. Hence, for $\theta_1 \in \mathcal{N}(\theta_0)$, $\|L(\theta_1) - L(\theta_0)\|$ and $\inf_{\theta\in\mathcal{P}_0} \|\theta_1 - \theta\|$ are of the same order of magnitude. Then, according to Theorem 2, both $\mathcal{N}(\theta_0)$ and $\mathcal{P}_0$ are shifted linear subspaces passing through θ_0. Since $\dim(\mathcal{N}(\theta_0)) = pr$, the null-space $\mathcal{N}(\theta_0)$ shares $(p - \text{rank}(A_0))\, r$ dimensions with $\mathcal{P}_0$, and has $\dim(\mathcal{N}(\theta_0)) - \dim(\mathcal{P}_0) = \text{rank}(A_0)\, r$ dimensions orthogonal to $\mathcal{P}_0$. The regret of an adaptive regulator $\widehat{\pi}$ becomes larger than a logarithmic function of time, because of the uncertainty $\mathcal{N}(\theta_0)/\mathcal{P}_0$. In other words, although the RHS of (18) is estimated accurately, the aforementioned uncertainty precludes obtaining an accurate approximation for the LHS of (18). In Definition 3, additional knowledge about θ_0 removes such uncertainty. Thus, a manifold (or a finite union of manifolds) of dimension $pq - \text{rank}(A_0)\, r$ implies the aforementioned identifiability condition. Below, we provide some examples of Γ_0.

(i) Optimality manifold: obviously, a trivial example is $\Gamma_0 = \mathcal{S}(\theta_0)$. In this case, the LHS of (18) vanishes.

(ii) Support condition: let Γ_0 be the set of $p \times q$ matrices with a priori known support $\mathcal{I}$. That is, for some set of indices $\mathcal{I} \subset \{(i,j) : 1 \leq i \leq p, 1 \leq j \leq q\}$, entries of all matrices $\theta \in \Gamma_0$ are zero outside of $\mathcal{I}$; $\Gamma_0 = \left\{\theta = \left[\theta_{ij}\right] : \theta_{ij} = 0 \text{ for } (i,j) \notin \mathcal{I}\right\}$. Then, Γ_0 is a (basic) subspace of $\mathbb{R}^{p\times q}$ and can satisfy the identifiability condition (18). Note that it is necessary to have $\dim(\Gamma_0) = |\mathcal{I}| \leq pq - \text{rank}(A_0)\, r$.

(iii) Sparsity condition: let Γ_0 be the set of all $p \times q$ matrices with at most $pq - \text{rank}(A_0)\, r$ non-zero entries. Then, Γ_0 is the union of the matrices with support $\mathcal{I}$ for different sets $\mathcal{I}$. Hence, the previous case implies that Γ_0 is a finite union of manifolds of proper dimension.

(iv) Rank condition: let Γ_0 be the set of $p \times q$ matrices θ such that $\text{rank}(\theta) \leq d$. Then, Γ_0 is a finite union of manifolds of dimension at most $d(p + q - d)$ (Shalit, Weinshall, & Chechik, 2012). Hence, if $d(p + q - d) \leq pq - \text{rank}(A_0)\, r$, and (18) holds, θ_0 is identifiable.

(v) Subspace condition: for $k = \text{rank}(A_0)\, r$, let $\{\theta_i\}_{i=1}^k$ be $p \times q$ matrices such that $\theta_i \widetilde{L}(\theta_0) = 0$. Suppose that $\theta_1, \ldots, \theta_k$ are linearly independent: if $\sum_{i=1}^k a_i\theta_i = 0$, then $a_1 = \cdots = a_k = 0$. Define $\Gamma_0 = \left\{\theta + \theta_0 : \text{tr}\left(\theta'\theta_i\right) = 0 \text{ for all } 1 \leq i \leq k\right\}$. If for all $1 \leq i \leq k$ it holds that $\theta_0 + \theta_i \notin \mathcal{P}_0$, then Γ_0 satisfies the identifiability condition of Definition 3.

The following Theorem establishes the optimality of GCE under the identifiability assumption. As mentioned in Section 4, a logarithmic gap between the lower and upper bounds for the

regret is inevitable due to similar limitations in system identification (Simchowitz et al., 2018).

Theorem 5 (*GCE Rates*). *Suppose that θ_0 is identifiable and the adaptive policy $\widehat{\pi}$ corresponds to GCE. Defining $\mathcal{P}_0$ by (11), let $\widehat{\theta}_n$ be the parameter estimate at time n. Then, we have*

$$\limsup_{n\to\infty} \frac{\mathcal{R}_n(\widehat{\pi})}{\log^2 n} < \infty, \limsup_{n\to\infty} \frac{\inf_{\theta\in\mathcal{P}_0} \left\|\left|\widehat{\theta}_n - \theta\right|\right\|^2}{n^{-1}\log n} < \infty.$$

Comparing the above result with Theorems 3 and 4, the identifiability assumption leads to significant improvements in rates of both the regret and the estimation error. Moreover, if $\text{rank}(A_0) = p$, then $\mathcal{P}_0 = \{\theta_0\}$. Thus, the estimation accuracy in Theorem 5 becomes: $\limsup_{n\to\infty} n\left(\log^{-1} n\right)\left\|\left|\widehat{\theta}_n - \theta_0\right|\right\|^2 < \infty$. Finally, Theorem 5 improves an existing result for identifiable systems. That is, under stronger assumptions, Ibrahimi et al. (Ibrahimi et al., 2012) show the regret bound $O\left(n^{1/2}\log^2 n\right)$ for adaptive policies based on OFU. However, according to Theorem 5, the regret of GCE is $O\left(\log^2 n\right)$.

6. Concluding remarks

The performances of adaptive policies for LQ systems is addressed in this work, including both aspects of regulation and identification. First, we established a general result which specifies the regret of an arbitrary adaptive regulator in terms of the deviations from the optimal feedback. This *tight* bidirectional result provides a powerful tool to analyze the subsequently presented policies. That is, we show that slight modifications of CE provide a regret of (nearly) square-root magnitude. The modifications consist of two basic approaches of randomization: additive randomness, and Thompson sampling. In addition, we formulated a condition which leads to *logarithmic* regret. The rates of *identification* are also discussed for the corresponding adaptive regulators.

Rigorous establishment of the proposed operational *lower bound* for the regret is an interesting direction for future works. Besides, extending the developed framework to other settings such as *switching* systems, or those with *imperfect* observations are topics of interest. On the other hand, extensions to the dynamical models illustrating *network* systems (e.g., high-dimensional sparse dynamics matrices) is a challenging problem for further investigation.

7. Proofs of main results

The proofs of the main theorems are given next. Due to space limitations, proofs of auxiliary lemmas are deferred to the supplementary material (Faradonbeh, Tewari, & Michailidis, 2018b).

7.1. Proof of Theorem 1 and Corollary 1

Given $n \geq 1$, and the linear policy $\pi = \{L_t\}_{t=0}^{n-1}$, define the sequence of policies $\pi_0, \ldots, \pi_n$ as follows.

$$\begin{aligned}
\pi_0 &= \{L(\theta_0), \ldots, L(\theta_0)\},\\
\pi_1 &= \{L_0, L(\theta_0), \ldots, L(\theta_0)\},\\
&\vdots\\
\pi_n &= \{L_0, L_1, \ldots, L_{n-1}\}.
\end{aligned}$$

Indeed, the policy π_i applies the same feedback as π at every time $t < i$, and then for $t \geq i$ switches to the optimal policy $\pi^\star$. Clearly, $\pi_0 = \pi^\star$, and $\pi_n = \pi$. Since

$$\mathcal{R}_n(\pi) = \sum_{k=1}^{n}\sum_{t=0}^{n-1}\left[c_t(\pi_k) - c_t(\pi_{k-1})\right], \tag{19}$$

it suffices to find $c_t(\pi_k) - c_t(\pi_{k-1})$, for $1 \leq k \leq n$, and $0 \leq t \leq n-1$. Fixing k, let $\{x(t)\}_{t=0}^{n-1}, \{y(t)\}_{t=0}^{n-1}$ be the state trajectories under π_k, π_{k-1}, respectively. So, letting $\mathbb{D} = A_0 + B_0 L(\theta_0)$ and $D_{k-1} = A_0 + B_0 L_{k-1}$, we have $x(t) = y(t)$ for $0 \leq t \leq k-1$, as well as $c_t(\pi_k) = c_t(\pi_{k-1})$ for $0 \leq t \leq k-2$, and $x(k) = D_{k-1}x(k-1) + w(k)$. Further, if $k \leq t \leq n-1$, then

$$\begin{aligned}
y(t) &= \mathbb{D}^{t-k+1}x(k-1) + \sum_{j=k}^{t}\mathbb{D}^{t-j}w(j),\\
x(t) &= \mathbb{D}^{t-k}D_{k-1}x(k-1) + \sum_{j=k}^{t}\mathbb{D}^{t-j}w(j).
\end{aligned}$$

Therefore, we have $x(t) = y(t) + \mathbb{D}^{t-k}\Delta_{k-1}x(k-1)$, for $k \leq t < n$, where

$$\Delta_{k-1} = D_{k-1} - \mathbb{D} = B_0\left(L_{k-1} - L(\theta_0)\right).$$

Thus, for we obtain

$$\begin{aligned}
&c_{k-1}(\pi_k) - c_{k-1}(\pi_{k-1})\\
&= x(k-1)'\left(L_{k-1}'RL_{k-1} - L(\theta_0)'RL(\theta_0)\right)x(k-1).
\end{aligned}$$

Similarly, denote $P_0 = Q + L(\theta_0)'RL(\theta_0)$, and replace for $x(t)$ to see that if $k \leq t < n$, then

$$\begin{aligned}
&c_t(\pi_k) - c_t(\pi_{k-1})\\
&= \left(2y(t) + \mathbb{D}^{t-k}\Delta_{k-1}x(k-1)\right)'P_0\mathbb{D}^{t-k}\Delta_{k-1}x(k-1).
\end{aligned}$$

To proceed, plug-in for $y(t)$ to get $c_t(\pi_k) - c_t(\pi_{k-1}) = x(k-1)'F_{k-1}(t)x(k-1) + \eta_{k-1}(t)$, where $\Delta_{k-1} = D_{k-1} - \mathbb{D}$ leads to

$$\begin{aligned}
\eta_{k-1}(t) &= 2x(k-1)'\Delta_{k-1}'\mathbb{D}'^{t-k}P_0\sum_{j=k}^{t}\mathbb{D}^{t-j}w(j),\\
F_{k-1}(t) &= D_{k-1}'\mathbb{D}'^{t-k}P_0\mathbb{D}^{t-k}D_{k-1} - \mathbb{D}'^{t-k+1}P_0\mathbb{D}^{t-k+1}.
\end{aligned}$$

Next, letting $z_k = \sum_{t=k}^{n-1}\eta_{k-1}(t)$, and $G_k = L_{k-1}'RL_{k-1} - L(\theta_0)'RL(\theta_0) + \sum_{t=k}^{n-1}F_{k-1}(t)$, clearly

$$\sum_{t=0}^{n-1}\left[c_t(\pi_k) - c_t(\pi_{k-1})\right] = x(k-1)'G_kx(k-1) + z_k. \tag{20}$$

To proceed, for $0 \leq j \leq n$ let $K_j = \sum_{\ell=n-j}^{\infty}\mathbb{D}'^{\ell}P_0\mathbb{D}^{\ell}$. So,

$$\sum_{t=k}^{n-1}F_{k-1}(t) = D_{k-1}'\left(K_n - K_k\right)D_{k-1} - \mathbb{D}'\left(K_n - K_k\right)\mathbb{D}$$

implies $G_k = E_k + H_k$, where

$$\begin{aligned}
E_k &= -D_{k-1}'K_kD_{k-1} + \mathbb{D}'K_k\mathbb{D},\\
H_k &= L_{k-1}'RL_{k-1} - L(\theta_0)'RL(\theta_0)\\
&\quad - \mathbb{D}'K_n\mathbb{D} + D_{k-1}'K_nD_{k-1}.
\end{aligned}$$

The Lyapunov equation (see Faradonbeh et al. (2019))

$$K(\theta_0) - \mathbb{D}'K(\theta_0)\mathbb{D} = P_0, \tag{21}$$

leads to $K_n = K(\theta_0)$. Thus, letting $X = L_{k-1} - L(\theta_0)$, $M = B_0'K(\theta_0)B_0 + R$, since $ML(\theta_0) = -B_0'K(\theta_0)A_0$, after doing some algebra we get

$$\begin{aligned}
H_k &= L(\theta_0)'RX + X'RL(\theta_0) + \mathbb{D}'K(\theta_0)B_0X\\
&\quad + X'RX + X'B_0'K(\theta_0)\mathbb{D} + X'B_0'K(\theta_0)B_0X\\
&= X'MX
\end{aligned}$$

Hence, adding up the terms in (20), (19) implies that

$$\mathcal{R}_n(\pi) = Z_n + S_n + T_n, \tag{22}$$

where $Z_n = \sum_{k=1}^{n} z_k$, $S_n = \sum_{k=0}^{n-1} x(k)' E_{k+1} x(k)$, and $T_n = \sum_{k=0}^{n-1} \left\| M^{1/2} \left(L_k - L(\theta_0)\right) x(k) \right\|^2$. In order to investigate S_n, we use the dynamics $x(k) = D_{k-1} x(k-1) + w(k)$, as well as $\mathbb{D}' K_{k+1} \mathbb{D} = K_k$, to get

$$x(k)' \mathbb{D}' \; K_{k+1} \; \mathbb{D} x(k) = x(k-1)' D_{k-1}' K_k D_{k-1} x(k-1) + \; w(k)' K_k w(k) + 2 w(k)' K_k D_{k-1} x(k-1),$$

for $0 < k < n$. Substituting in the expression for S_n, and denoting $w(0) = x(0)$, the telescopic differences vanish:

$$S_n + x(n)' K(\theta_0) x(n) = \sum_{k=0}^{n-1} 2 w(k+1)' K_{k+1} D_k x(k) + \sum_{k=0}^{n} w(k)' K_k w(k). \tag{23}$$

Plugging

$$D_k x(k) = \sum_{j=0}^{k} \left(\mathbb{D}^{j+1} w(k-j) + \mathbb{D}^j \Delta_{k-j} x(k-j) \right),$$

as well as $x^\star(n) = \sum_{j=0}^{n} \mathbb{D}^{n-j} w(j)$, in (23), we have $\widetilde{S}_n = S_n + x(n)' K_n x(n) - x^\star(n)' K_n x^\star(n) = \sum_{k=1}^{n} w(k)' K_k \xi_k$, where $\xi_k = 2 \sum_{\ell=1}^{k} \mathbb{D}^{\ell-1} \Delta_{k-\ell} x(k-\ell)$. Moreover, it is straightforward to show that $Z_n = \sum_{j=1}^{n-1} \zeta_j' w(j)$, where $\zeta_j = 2 \sum_{\ell=1}^{j} \sum_{t=j}^{n-1} \mathbb{D}'^{t-j} P_0 \mathbb{D}^{t-\ell} \Delta_{\ell-1} x(\ell-1)$.

Hence, $\zeta_j = \left(K_n - K_j\right) \xi_j$ implies $\widetilde{S}_n + Z_n = \sum_{k=1}^{n} w(k)' K_n \xi_k$. Next, we use the following lemma.

Lemma 3 (*Lai & Wei, 1982*). *Suppose that for all $t \geq 0$, $y(t+1)$, $v(t)$ are $\mathcal{G}_t$ measurable, $\mathcal{G}_t \subseteq \mathcal{G}_{t+1}$, and $\mathbb{E}[v(t+1)|\mathcal{G}_t] = 0$. Define the martingale $\psi_n = \sum_{t=1}^{n} y(t)' v(t)$, and let $\varphi_n = \sum_{t=1}^{n} \|y(t)\|^2$. If $\sup_{t \geq 0} \mathbb{E}\left[\|v(t+1)\|^2 \middle| \mathcal{G}_t \right] < \infty$, then*

$$\limsup_{n \to \infty} |\psi_n| < \infty \quad \text{on} \quad \varphi_\infty < \infty,$$

$$\limsup_{n \to \infty} \frac{\psi_n}{\varphi_n^{1/2} \log \varphi_n} = 0 \quad \text{on} \quad \varphi_\infty = \infty.$$

Taking $\mathcal{G}_t = \sigma\left(\{w(i)\}_{i=1}^{t}, \{x(i)\}_{i=0}^{t}, \{L_i\}_{i=0}^{t} \right)$, and $v(t) = w(t)$, $y(t) = \xi_t$, we can use Lemma 3 since Assumption 1 holds. So, stability of $\mathbb{D}$ (Proposition 1), and $|\lambda_{\min}(M)| > 0$, lead to $\sum_{k=1}^{n} \|\xi_k\|^2 = O(T_n)$. Thus, by (22), we get the desired result since $\widetilde{S}_n + Z_n = O\left(T_n^{1/2} \log T_n\right)$.

Next, the first statement in Corollary 1 follows from Theorem 1 in the work of Lai and Wei (Lai & Wei, 1985). To prove the second result, first observe that $S_\infty = S_n$, $T_\infty = T_n$, and $Z_\infty = Z_n$. Furthermore, note that for $t \geq n$ we have $c_t(\pi) = x(t)' P_0 x(t)$, $c_t(\pi^\star) = x^\star(t)' P_0 x^\star(t)$, as well as

$$x(t) = \mathbb{D}^{t-n} x(n) + \sum_{j=n+1}^{t} \mathbb{D}^{t-j} w(j),$$

$$x^\star(t) = \mathbb{D}^{t-n} x^\star(n) + \sum_{j=n+1}^{t} \mathbb{D}^{t-j} w(j).$$

So, letting

$$\delta_n = 2 \sum_{t=n}^{\infty} \sum_{j=n+1}^{t} \left(x(n) - x^\star(n) \right)' \mathbb{D}'^{t-n} P_0 \mathbb{D}^{t-j} w(j),$$

by (21) the following holds:

$$\sum_{t=n}^{\infty} \left[c_t(\pi) - c_t\left(\pi^\star\right) \right] = x(n)' K_n x(n) - x^\star(n)' K_n x^\star(n) + \delta_n.$$

Finally,

$$\left\| x(n) - x^\star(n) \right\|^2 = \left\| \sum_{j=0}^{n-1} \mathbb{D}^{n-1-j} \Delta_j x(j) \right\|^2 = O(T_n), \tag{24}$$

together with Lemma 3 imply $\delta_n = O\left(T_n^{1/2} \log T_n\right)$.

7.2. Proof of Theorem 2

First, for an arbitrary $\theta \in \mathcal{P}_0$, since $\theta \in \mathcal{N}(\theta_0)$, we have

$$A + BL(\theta_0) = A_0 + B_0 L(\theta_0) = D_0. \tag{25}$$

Next, for an arbitrary fixed unit matrix (in the Frobenius norm) $X \in \mathbb{R}^{r \times p}$, let $L = L(\theta_0) + \epsilon X$ be a linear feedback matrix which stabilizes the system of dynamics parameters θ. Note that according to Proposition 1, $\theta \in \mathcal{S}(\theta_0)$ leads to $\left|\lambda_{\max}\left(\theta \widetilde{L}(\theta_0)\right)\right| < 1$. Thus, $|\lambda_{\max}(A + BL)| < 1$, as long as ϵ is sufficiently small.

Then, applying L to the system θ, we get $\overline{\mathcal{J}}_L(\theta) = \operatorname{tr}(P(\epsilon) C)$, where $P(\epsilon)$ is the unique solution of the Lyapunov equation

$$P(\epsilon) - (A + BL)' P(\epsilon) (A + BL) = Q + L' R L. \tag{26}$$

Note that according to (21) and (25), it holds that $P(0) = K(\theta_0)$. Letting $\Delta(X) = \lim_{\epsilon \to 0} \epsilon^{-1} (P(\epsilon) - P(0))$, (26) leads to

$$\Delta(X) - D_0' \Delta(X) D_0 = X' N + N' X, \tag{27}$$

where $N = RL(\theta_0) + B' K(\theta_0) D_0$. Next, $\theta \in \mathcal{S}(\theta_0)$ implies that $L(\theta_0)$ is an optimal linear feedback for the system of dynamics parameter θ. So, the directional derivative of $\overline{\mathcal{J}}_L(\theta)$ with respect to L is zero in all directions. In the direction of X, the derivative is $\operatorname{tr}(\Delta(X) C)$. Since all above statements hold regardless of the positive definite matrix C, (27) and $\operatorname{tr}(\Delta(X) C) = 0$ imply $N = 0$;

$$D_0' K(\theta_0) B = -L(\theta_0)' R. \tag{28}$$

Therefore, (28) is a necessary condition for $\theta \in \mathcal{P}_0$. Note that according to (21) and (25), the necessary condition (28) implies the necessity of $D_0' K(\theta_0) A = K(\theta_0) - Q$. Further, for every input matrix B which satisfies (28), the transition matrix A will be uniquely determined by (25) as $A = D_0 - BL(\theta_0)$.

Conversely, suppose that B is an arbitrary matrix which satisfies (28). Letting $A = D_0 - BL(\theta_0)$, we show that $[A, B] = \theta \in \mathcal{P}_0$. For this purpose, since the above definition of A automatically leads to $\theta \in \mathcal{N}(\theta_0)$, it suffices to show $\theta \in \mathcal{S}(\theta_0)$. Writing $Y = B - B_0$, we get $A = A_0 - YL(\theta_0)$. Moreover, define $G = A' K(\theta_0) A$, $H = B' K(\theta_0) A$, $M = B_0' K(\theta_0) B_0 + R$, and $S = B_0' K(\theta_0) Y + Y' K(\theta_0) B_0 + Y' K(\theta_0) Y$. Then, we calculate the matrix

$$V = Q + G - H' (M + S)^{-1} H = Q + A' K(\theta_0) A - A' K(\theta_0) B \left(B' K(\theta_0) B + R \right)^{-1} B' K(\theta_0) A.$$

Writing A, B, G, H in terms of A_0, B_0, M, S, Y, we have

$$V = Q + A_0' K(\theta_0) A_0 + L(\theta_0)' S L(\theta_0) - \left[B_0' K(\theta_0) A_0 - S L(\theta_0) \right]' (M + S)^{-1} \left[B_0' K(\theta_0) A_0 - S L(\theta_0) \right]$$

Then, using $(M + S)^{-1} = M^{-1} - (M + S)^{-1} S M^{-1}$, (3), and $ML(\theta_0) = -B_0' K(\theta_0) A_0$, V can be written as $V = K(\theta_0) + L(\theta_0)' S W$, where

$$W = L(\theta_0) - (M + S)^{-1} \left(S L(\theta_0) + B_0' K(\theta_0) A_0 \right) = L(\theta_0) - (M + S)^{-1} (S + M) L(\theta_0) = 0;$$

i.e., $V = K(\theta_0)$ is a solution of the Riccati equation (3) for θ. According to Proposition 1, the solution is unique; which is $K(\theta) = K(\theta_0)$. Moreover, $L(\theta) = -(M + S)^{-1} H = L(\theta_0)$ shows that $\theta \in \mathcal{S}(\theta_0)$. So far, we have shown that $\theta \in \mathcal{P}_0$, if and only if (25) and (28) hold. Next, (28) is essentially stating

that every column of $B - B_0$ (which is a vector in $\mathbb{R}^p$), is orthogonal to the all columns of $K(\theta_0) D_0$. This verifies that (28) specifies a shifted linear subspace. To find the dimension, since B has r columns, and (25) uniquely determines A in terms of B, we get $\dim(\mathcal{P}_0) = (p - \text{rank}(K(\theta_0) D_0)) r$. Finally, by positive definiteness of Q, (21) implies $\text{rank}(K(\theta_0)) = p$. Further, since $D_0 = \left[I_p - B_0 M^{-1} B_0' K(\theta_0)\right] A_0$, it suffices to show

$$\text{rank}\left(I_p - B_0 M^{-1} B_0' K(\theta_0)\right) = p. \tag{29}$$

If (29) does not hold, there exists $v \in \mathbb{R}^p$ such that $v \neq 0$ and $v = B_0 M^{-1} B_0' K(\theta_0) v$. So, $v = B_0 \widetilde{v}$ where $\widetilde{v} = M^{-1} B_0' K(\theta_0) v \in \mathbb{R}^r$. Thus,

$$B_0' K(\theta_0) B_0 \widetilde{v} = B_0' K(\theta_0) v = M \widetilde{v} = \left[B_0' K(\theta_0) B_0 + R\right] \widetilde{v},$$

or equivalently, $R\widetilde{v} = 0$. Positive definiteness of R implies that $\widetilde{v} = 0$, which contradicts $B_0 \widetilde{v} \neq 0$. This proves (29), which completes the proof.

7.3. Proof of Theorem 3

The proof is based on a sequence of intermediate results. First, for $i \geq 1$, let V_i be the (unnormalized) state covariance during the ith episode: $V_i = \sum_{t=\lfloor \gamma^{i-1} \rfloor}^{\lfloor \gamma^i \rfloor - 1} x(t) x(t)'$.

Lemma 4. *For the matrix V_i defined above, the followings hold: $|\lambda_{\max}(V_m)| = O(\gamma^m)$, $\liminf_{m\to\infty} \gamma^{-m} |\lambda_{\min}(V_m)| \geq (\gamma - 1) |\lambda_{\min}(C)|$.*

Then, in order to study the behavior of the least-squares estimate in (13), define

$$U_i = \sum_{t=0}^{\lfloor \gamma^i \rfloor - 1} \widetilde{L}\left(\widehat{\theta}_t\right) x(t) x(t)' \widetilde{L}\left(\widehat{\theta}_t\right)'.$$

Note that since the parameter $\widehat{\theta}_t$ remains set (not changing) during each episode, U_i can be written in terms of $V_1, \ldots, V_i$ as follows. First, for all $\lfloor \gamma^{i-1} \rfloor \leq t \leq \lfloor \gamma^i \rfloor - 1$, the parameter estimate $\widehat{\theta}_t$ does not change. So, if t belongs to the ith episode, define the linear feedback matrix is $L_i = L\left(\widehat{\theta}_t\right)$. Letting $\widetilde{L}_i = \widetilde{L}\left(\widehat{\theta}_t\right)$, we have $U_i = \sum_{j=1}^{i} \widetilde{L}_j V_j \widetilde{L}_j'$. Then, the smallest eigenvalue of U_i follows a different lower bound compared to that of V_i:

Lemma 5. *Define U_m as above. Then, we have $\liminf_{m\to\infty} \gamma^{-m/2} |\lambda_{\min}(U_m)| > 0$, and $|\lambda_{\max}(U_m)| = O(\gamma^m)$.*

Next, the following result states that the estimation accuracy is determined by the eigenvalues of U_i.

Lemma 6 (*Lai & Wei, 1982*). *For $n = \lfloor \gamma^m \rfloor$, define $\widehat{\theta}_n, \widetilde{\theta}_n$ according to (13). Then, we have*

$$\left\|\widehat{\theta}_n - \widetilde{\theta}_n - \theta_0\right\|^2 = O\left(\frac{\log |\lambda_{\max}(U_m)|}{|\lambda_{\min}(U_m)|}\right).$$

Therefore, Lemma 5 leads to $\left\|\widehat{\theta}_n - \widetilde{\theta}_n - \theta_0\right\| = O\left(n^{-1/4} \log^{1/2} n\right)$. Using the moment condition in Remark 2, Markov's inequality gives $\mathbb{P}\left(\|\phi_m\| > m^{1/4}\right) = O\left(m^{-1-\epsilon/4}\right)$. Thus, an application of the Borel–Cantelli Lemma leads to $\|\phi_m\| = O\left(m^{1/4}\right)$; i.e., $\left\|\widetilde{\theta}_n\right\| = O\left(n^{-1/4} \log^{1/2} n\right)$. So, we get the desired result about the identification rate: $\left\|\widehat{\theta}_n - \theta_0\right\| = O\left(n^{-1/4} \log^{1/2} n\right)$. To proceed, we present the following auxiliary result which shows that a similar rate holds for the deviations from the optimal linear feedback.

Lemma 7 (*Faradonbeh et al., 2017b*). *There exist $0 < \epsilon_0, \beta_L < \infty$, such that for all stabilizable θ satisfying $\|\theta - \theta_0\| < \epsilon_0$, the following holds: $\|L(\theta) - L(\theta_0)\| \leq \beta_L \|\theta - \theta_0\|$.*

So, utilizing Lemma 7, we have

$$\left\|L\left(\widehat{\theta}_n\right) - L(\theta_0)\right\| = O\left(\frac{\log^{1/2} n}{n^{1/4}} + \left\|\widetilde{\theta}_n\right\|\right). \tag{30}$$

On the other hand, since the policy is not being updated during each episode, we can write down the regret in terms of the matrices V_i. Henceforth in the proof, suppose that the time n belongs to the mth episode: $\lfloor \gamma^{m-1} \rfloor \leq n < \lfloor \gamma^m \rfloor$. Then, applying Theorem 1 and Corollary 1, we get

$$\begin{aligned}\mathcal{R}_n(\widehat{\pi}) &= O\left(\sum_{i=0}^{m} (L_i - L(\theta_0)) V_i (L_i - L(\theta_0))' + \gamma^{m/2}\right) \\ &= O\left(\sum_{i=0}^{m} \gamma^i \|L_i - L(\theta_0)\|^2 + \gamma^{m/2}\right),\end{aligned}$$

where in the last equality above we applied Lemma 4. Based on the definition of the perturbation $\widetilde{\theta}_n$ in terms of the random matrix ϕ_m, define

$$S_m = \sum_{i=0}^{m} i^{1/2} \gamma^{i/2} \|\phi_i\|^2, \qquad T_m = \sum_{i=0}^{m} i^{3/4} \gamma^{i/2} \|\phi_i\|.$$

So, by (30), the regret is in magnitude dominated by S_m, T_m, and $m\gamma^m/2$: $\mathcal{R}_n(\widehat{\pi}) = O\left(S_m + T_m + m\gamma^{m/2}\right)$. Note that as m and n grow, the magnitudes of $n^{1/2} \log n$ and $m\gamma^{m/2}$ are the same. Finally, the following lemma leads to the desired result:

Lemma 8. *For the terms S_m, T_m defined above the followings hold: $S_m = O\left(m\gamma^{m/2}\right)$, $T_m = O\left(m\gamma^{m/2}\right)$.*

7.4. Proof of Theorem 4

In this proof, we use the following result.

Lemma 9. *For the matrix Σ_m defined in (16) we have $\liminf_{m\to\infty} \gamma^{-m/2} m^{1/2} |\lambda_{\min}(\Sigma_m)| > 0$, $|\lambda_{\max}(\Sigma_m)| = O(\gamma^m)$.*

Hence, since μ_m is the least-squares estimate, and Σ_m is the unnormalized empirical covariance matrix, Lemma 6 leads to $\|\mu_m - \theta_0\| = O\left(\gamma^{-m/4} m\right)$. Then, because every row of $\widehat{\theta}_{\lfloor \gamma^m \rfloor} - \mu_m$ is a mean zero Gaussian with covariance matrix Σ_m^{-1}, by Lemma 9 we have

$$\sum_{m=0}^{\infty} \mathbb{P}\left(\left\|\widehat{\theta}_{\lfloor \gamma^m \rfloor} - \mu_m\right\| > \gamma^{-m/4} m\right) < \infty.$$

Thus, Borel–Cantelli Lemma leads to the desired result about the identification rate: $\left\|\widehat{\theta}_{\lfloor \gamma^m \rfloor} - \theta_0\right\| = O\left(\gamma^{-m/4} m\right)$. By Lemma 7, a similar rate holds for the linear feedbacks: $\left\|L\left(\widehat{\theta}_{\lfloor \gamma^m \rfloor}\right) - L(\theta_0)\right\| = O\left(\gamma^{-m/4} m\right)$. Finally, plugging in the expression of Theorem 1, and utilizing Corollary 1, we get the desired result for the regret:

$$\begin{aligned}\mathcal{R}_{\lfloor \gamma^m \rfloor}(\widehat{\pi}) &= O\left(\sum_{i=0}^{m} \gamma^i \left\|L\left(\widehat{\theta}_{\lfloor \gamma^m \rfloor}\right) - L(\theta_0)\right\|^2 + \gamma^{m/2}\right) \\ &= O\left(\sum_{i=0}^{m} \gamma^{i/2} i^2\right) = O\left(\gamma^{m/2} m^2\right).\end{aligned}$$

7.5. Proof of Theorem 5

Define $V_i, U_i, L_i, \widetilde{L}_i$ similar to the proof of Theorem 3. Further, for $i \geq 1$, let $n_i = \lfloor \gamma^i \rfloor - 1$ be the end time of episode i, and denote $\mathcal{L}_i(\theta) = \sum_{t=0}^{n_i - 1} \left\|x(t+1) - \theta \widetilde{L}\left(\widehat{\theta}_t\right) x(t)\right\|^2$.

Letting $\theta_\star = \arg\min_{\theta \in \mathbb{R}^{p\times q}} \mathcal{L}_i(\theta)$ for a fixed i, it is straightforward to show that

$$\mathcal{L}_i(\theta) = \text{tr}\left((\theta - \theta_\star) U_i (\theta - \theta_\star)'\right) - \text{tr}\left(\theta_\star U_i \theta_\star'\right).$$

Therefore, since $\theta_0 \in \Gamma_0$, (17) implies that $\mathcal{L}_i\left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i}\right) \leq \mathcal{L}_i(\theta_0)$. So, the triangle inequality leads to

$$\begin{aligned}&\mathrm{tr}\left(\left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i} - \theta_0\right) U_i \left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i} - \theta_0\right)'\right)\\ &\leq 4\mathrm{tr}\left((\theta_\star - \theta_0)\, U_i\, (\theta_\star - \theta_0)'\right)\end{aligned}$$

Hence, the normal equation $(\theta_\star - \theta_0)\, U_i = \sum_{t=0}^{n_i-1} w(t+1)x(t)'$ $\widetilde{L}\left(\widehat{\theta}_t\right)'$, in addition to Lemma 5 and Lemma 6 imply that

$$\mathrm{tr}\left(\left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i} - \theta_0\right) U_i \left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i} - \theta_0\right)'\right) = O(i)$$

Applying Lemma 4, we obtain

$$\sum_{j=0}^{i} \gamma^j \left\|\left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i} - \theta_0\right)\widetilde{L}_j\right\|^2 = O(i)\,. \tag{31}$$

Since $\widetilde{\theta}_{n_j} = O\left(n_j^{-1/2}\right)$, by Lemma 7 we have $\left\|L_j - L\left(\widehat{\theta}_{n_j} - \widetilde{\theta}_{n_j}\right)\right\| = O\left(\gamma^{-j/2}\right)$. Hence,

$$\sum_{j=0}^{i} \gamma^j \left\|\left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i} - \theta_0\right)\widetilde{L}\left(\widehat{\theta}_{n_j} - \widetilde{\theta}_{n_j}\right)\right\|^2 = O(i)\,.$$

Using $\widehat{\theta}_{n_j} - \widetilde{\theta}_{n_j} \in \Gamma_0$, (18) leads to $\left\|L\left(\widehat{\theta}_{n_i} - \widetilde{\theta}_{n_i}\right) - L(\theta_0)\right\| = O\left(i^{1/2}\gamma^{-i/2}\right)$, which by Lemma 7 implies that

$$\left\|L\left(\widehat{\theta}_{n_i}\right) - L(\theta_0)\right\| = O\left(i^{1/2}\gamma^{-i/2}\right)\,. \tag{32}$$

Thus, we have

$$\begin{aligned}\sum_{t=0}^{n_m-1} \left\|\left(L(\theta_0) - L_t\right)x(t)\right\|^2 &= O\left(\sum_{i=0}^{m} \gamma^i \left\|L_i - L(\theta_0)\right\|^2\right)\\ &= O\left(m^2\right)\,.\end{aligned} \tag{33}$$

Moreover, putting Assumption 1, Corollary 1, (24), and (32) together, we obtain $\|x(n_m) - x^\star(n_m)\|\,\|x^\star(n_m)\| = O(m)$, which in turn leads to

$$x^\star(n_m)'\, K(\theta_0)\, x^\star(n_m) - x(n_m)' K(\theta_0)\, x(n_m) = O(m)\,. \tag{34}$$

Then, (33) and (34) lead to the desired result for the regret: $\mathcal{R}_{n_m}(\widehat{\pi}) = O\left(m^2\right)$. Further, (31) and (32) imply that

$$\left\|\left(\widehat{\theta}_{n_m} - \widetilde{\theta}_{n_m} - \theta_0\right)\widetilde{L}(\theta_0)\right\| = O\left(\gamma^{-m/2}m^{1/2}\right);$$

i.e., $\inf_{\theta\in\mathcal{N}(\theta_0)} \left\|\widehat{\theta}_{n_m} - \theta\right\| = O\left(\gamma^{-m/2}m^{1/2}\right)$. Finally, since (32) implies a similar result for $\mathcal{S}(\theta_0)$, the desired result for $\mathcal{P}_0$ holds.

Acknowledgments

The authors would like to thank Professor Tze Leung Lai for helpful discussions, and the reviewers for their constructive comments. The work of M. K. S. F. was partially supported by the University of Florida Informatics Institute, USA. A. T. acknowledges the support of NSF, USA via CAREER grant IIS-1452099. G.M.'s work was supported in part by NSF, USA grants DMS-1821220 and DMS-1632730.

References

Abbasi-Yadkori, Y., & Szepesvári, C. (2011). Regret bounds for the adaptive control of linear quadratic systems. In *COLT* (pp. 1–26).

Abeille, M., & Lazaric, A. (2017). Thompson sampling for linear-quadratic control problems. In AISTATS 2017-20th international conference on artificial intelligence and statistics.

Abeille, M., & Lazaric, A. (2018). Improved regret bounds for Thompson sampling in linear quadratic control problems. In International conference on machine learning (pp. 1–9).

Anderson, B. D. (1985). Adaptive systems, lack of persistency of excitation and bursting phenomena. *Automatica*, *21*(3), 247–258.

Bar-Shalom, Y., & Tse, E. (1974). Dual effect, certainty equivalence, and separation in stochastic control. *IEEE Transactions on Automatic Control*, *19*(5), 494–500.

Becker, A., Kumar, P., & Wei, C.-Z. (1985). Adaptive control with the stochastic approximation algorithm: geometry and convergence. *IEEE Transactions on Automatic Control*, *30*(4), 330–338.

Bercu, B. (1995). Weighted estimation and tracking for ARMAX models. *SIAM Journal on Control and Optimization*, *33*(1), 89–106.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control: Vol. 1, no. 2*. MA: Athena Scientific Belmont.

Bittanti, S., & Campi, M. C. (2006). Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, *6*(4), 299–320.

Campi, M. C., & Kumar, P. (1998). Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, *36*(6), 1890–1907.

Chan, S., Goodwin, G., & Sin, K. (1984). Convergence properties of the riccati difference equation in optimal filtering of nonstabilizable systems. *IEEE Transactions on Automatic Control*, *29*(2), 110–118.

Chen, H.-F., & Zhang, J.-F. (1989). Convergence rates in stochastic adaptive tracking. *International Journal of Control*, *49*(6), 1915–1935.

De Souza, C., Gevers, M., & Goodwin, G. (1986). Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices. *IEEE Transactions on Automatic Control*, *31*(9), 831–838.

Duncan, T. E., Guo, L., & Pasik-Duncan, B. (1999). Adaptive continuous-time linear quadratic gaussian control. *IEEE Transactions on Automatic Control*, *44*(9), 1653–1662.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2017a). Optimality of fast matching algorithms for random networks with applications to structural controllability. *IEEE Transactions on Control of Network Systems*, *4*(4), 770–780.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2017b). Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, arXiv:1711.07230.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2018a). Finite time identification in unstable linear systems. *Automatica*, *96*, 342–353.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2018b). On adaptive linear-quadratic regulators. arXiv preprint arXiv:1806.10749.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2019). Finite time adaptive stabilization of linear systems. *IEEE Transactions on Automatic Control*, *64*(8), 3498–3505.

Guo, L. (1995). Convergence and logarithm laws of self-tuning regulators. *Automatica*, *31*(3), 435–450.

Guo, L., & Chen, H. (1988). Convergence rate of els based adaptive tracker. *Journal of Systems Science and Mathematical Sciences*, *1*, 131–138.

Guo, L., & Chen, H.-F. (1991). The åstrom-wittenmark self-tuning regulator revisited and els-based adaptive trackers. *IEEE Transactions on Automatic Control*, *36*(7), 802–812.

Ibrahimi, M., Javanmard, A., & Roy, B. V. (2012). Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in neural information processing systems* (pp. 2636–2644).

Johnstone, R., & Anderson, B. (1983). Global adaptive pole placement: detailed analysis of a first-order system. *IEEE Transactions on Automatic Control*, *28*(8), 852–855.

Kumar, P. (1990). Convergence of adaptive control schemes using least-squares parameter estimates. *IEEE Transactions on Automatic Control*, *35*(4), 416–424.

Lai, T. L. (1986). Asymptotically efficient adaptive control in stochastic regression models. *Advances in Applied Mathematics*, *7*(1), 23–45.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22.

Lai, T. L., & Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 154–166.

Lai, T. L., & Wei, C. Z. (1985). Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. *Multivariate Analysis VI*, 375–393.

Lai, T. L., & Wei, C.-Z. (1986). Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control*, *31*(10), 898–906.

Lai, T. L., & Ying, Z. (1991). Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems. *SIAM Journal on Control and Optimization*, *29*(5), 1091–1127.

Ouyang, Y., Gagrani, M., & Jain, R. (2017). Control of unknown linear systems with thompson sampling. In *Communication, control, and computing (Allerton), 2017 55th annual Allerton conference on* (pp. 1198–1205). IEEE.

Polderman, J. W. (1986a). A note on the structure of two subsets of the parameter space in adaptive control problems. *Systems & Control Letters*, *7*(1), 25–34.

Polderman, J. W. (1986b). On the necessity of identifying the true parameter in adaptive LQ control. *Systems & Control Letters*, *8*(2), 87–91.

Shalit, U., Weinshall, D., & Chechik, G. (2012). Online learning in the embedded manifold of low-rank matrices. *Journal of Machine Learning Research (JMLR)*, *13*(Feb), 429–458.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., & Recht, B. (2018). Learning without mixing: towards a sharp analysis of linear system identification. arXiv preprint arXiv:1802.08334.

Zhang, H.-M. (1990). Further comments on nonstationarity identification problems for autoregressive models. In *29th IEEE conference on decision and control* (pp. 3204–3205). IEEE.

Mohamad Kazem Shirani Faradonbeh received the Ph.D. degree in Statistics from the University of Michigan, Ann Arbor, MI, USA, in 2017, and the B.S. degree in Electrical Engineering from Sharif University of Technology, Tehran, Tehran, Iran, in 2012. He is currently a Post-Doctoral Fellow with the Informatics Institute and with the Department of Statistics at the University of Florida, Gainesville, FL, USA.

Ambuj Tewari received the Ph.D. degree in computer science at the University of California at Berkeley, Berkeley, CA, USA, in 2007. He is an associate professor in the Department of Statistics and the Department of Electrical Engineering and Computer Science (by courtesy) at the University of Michigan, Ann Arbor, MI, USA. He is also affiliated with the Michigan Institute for Data Science (MIDAS), Ann Arbor, MI, USA. His research interests lie in machine learning including statistical learning theory, online learning, reinforcement learning and control theory, network analysis, and optimization for machine learning. He collaborates with scientists to seek novel applications of machine learning in mobile health, learning analytics, and computational chemistry. His research has been recognized with paper awards at COLT 2005, COLT 2011, and AISTATS 2015. He was the recipient of an NSF CAREER award in 2015 and a Sloan Research Fellowship in 2017.

George Michailidis received the Ph.D. degree in mathematics from the University of California, Los Angeles, CA, USA, in 1996. He was a Post-Doctoral Fellow with the Department of Operations Research, Stanford University, Stanford, CA, from 1996 to 1998. He joined the University of Michigan, Ann Arbor, MI, USA, in 1998, where he spent 17 years a faculty members in Statistics, Electrical Engineering, and Computer Science. He is currently the Director of the Informatics Institute at the University of Florida and a faculty in the Departments of Statistics and Computer and Information Science and Engineering. He is a Fellow with the Institute of Mathematical Statistics, American Statistical Association, and the International Statistical Institute. His current research interests include stochastic network modeling and performance evaluation, queuing analysis and congestion control, statistical modeling and analysis of internet traffic, network tomography, and analysis of high dimensional data with network structure. Prof. Michailidis served as the Editor-in-Chief of the Electronic Journal of Statistics from 2013–2015 and serves on a number of editorial boards. He served as Symposium Chair for the Support for Storage, Renewable Sources, and Microgrids Symposium for SmartGridComm 2012 and is the Secretary of the IEEE Subcommittee on SmartGrid Communications.