2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Regret Analysis of Learning-Based Linear Quadratic Gaussian Control with Additive Exploration*

Archith Athrey[1], Othmane Mazhar[2], Meichen Guo[1], Bart De Schutter[1] and Shengling Shi[1]

Abstract— **In this paper, we analyze the regret incurred by a computationally efficient exploration strategy, known as naive exploration, for controlling unknown partially observable systems within the Linear Quadratic Gaussian (LQG) framework. We introduce a two-phase control algorithm called LQG-NAIVE, which involves an initial phase of injecting Gaussian input signals to obtain a system model, followed by a second phase of an interplay between naive exploration and control in an episodic fashion. We show that LQG-NAIVE achieves a regret growth rate of $\tilde{\mathcal{O}}(\sqrt{T})$, i.e., $\mathcal{O}(\sqrt{T})$ up to logarithmic factors after T time steps, and we validate its performance through numerical simulations. Additionally, we propose LQG-IF2E, which extends the exploration signal to a 'closed-loop' setting by incorporating the Fisher Information Matrix (FIM). We provide compelling numerical evidence of the competitive performance of LQG-IF2E compared to LQG-NAIVE.**

I. INTRODUCTION

In this work, we address adaptive control of *unknown* partially observable linear dynamical systems in the Linear Quadratic Gaussian (LQG) setting. Adaptive control caters to the control of unknown systems, where the controller is updated online from the collected data, to optimize some performance measures [1]. The LQG control problem is one of the key issues in adaptive control [2]. The seemingly benign difference of not being able to measure the true states will in fact pose a significant challenge when controlling a system with unknown dynamics [3]. The errors in the state estimates due to approximate models could potentially accumulate to have a significant impact on the control performance.

A metric called regret quantifies an adaptive controller's performance, i.e., its capability to balance exploration and exploitation [4]. The regret measures the cumulative performance gap over a finite time horizon between the adaptive controller and the ideal controller having full knowledge of the true system dynamics. For the adaptive control in the LQG setting, several works have contributed to statistical guarantees on both learning and control [3], [5]–[9].

The adaptive control algorithm in [3] uses an exploration approach called optimism in the face of uncertainty, which utilizes model parameter uncertainty to engender optimism in its deployed policy. Although this scheme is shown to guarantee a regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$, i.e., $\mathcal{O}(\sqrt{T})$ after T time steps up to logarithmic factors in T, finding optimistic model parameters involves non-convex optimization. In [5], the performance of the Certainty Equivalence Controller (CEC) is analyzed, but it is analyzed without any exploration or online model updates. A Thompson-sampling-based approach is adopted in [7], which exploits parameter uncertainty and promises computational efficiency. This approach is also shown to guarantee a regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$. The results in [8] show that the best regret upper bound that one can achieve is $\tilde{\mathcal{O}}(\sqrt{T})$, for the LQ setting. In [6], under an additional assumption that the optimal controller persistently excites the true underlying system, a convex reparametrization of a linear dynamical controller is considered, which guarantees a polylogarithmic regret upper bound. However, this assumption can be restrictive since the optimal controller in the LQ setting typically cannot persistently excite the true underlying system [8].

In the Linear Quadratic Regulator (LQR) setting, it has been shown that naive exploration, which involves a simple CEC with an additive excitation signal, whose covariance diminishes at a rate $\mathcal{O}(1/\sqrt{t})$ for intermediate time step $t \leq T$, is sufficient to guarantee a regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$ [5], [10]. Whereas, naive-exploration-based control with regret guarantee in the LQG setting is still an open problem.

In the present work, we investigate naive exploration in the LQG setting. We propose two adaptive control algorithms, LQG-NAIVE and LQG-IF2E, which operate in an episodic fashion and conduct exploration by injecting additive Gaussian signals. While the covariance of the Gaussian exploration signal in LQG-NAIVE decreases over episodes, the covariance of the exploration signal in LQG-IF2E adjusts adaptively to the data informativity by exploiting the Fisher Information Matrix (FIM). The latter exploration strategy is inspired by the approach in [11] designed for the LQR setting.

The structure of the adaptive control algorithms proposed in this work is similar to the one in [3]; however, the main difference lies in the exploration strategy: the algorithm in [3] employs optimism in the face of uncertainty and thus requires solving non-convex optimization problems online for exploration, whereas in this work, we consider an additive Gaussian signal for exploration. This additive exploration avoids solving optimization problems online and is thus much more computationally efficient. This difference in the exploration strategy poses a major challenge to the

*This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No.101018826 - CLariNet).

[1]Archith Athrey, Meichen Guo, Bart De Schutter, and Shengling Shi are with the Delft Center for Systems and Control, Delft University of Technology, Mekelweg 5, 2628 CD Delft, The Netherlands a.athrey@student.tudelft.nl, {m.guo, S.Shi-3, b.deschutter}@tudelft.nl

[2]Othamne Mazhar is with Université Paris Cité - Grands Moulins Campus, 5 Rue Thomas Mann, 75013 Paris, France omazhar@lpsm.paris

regret analysis of the proposed algorithms. In this work, we establish a regret growth rate of $\tilde{\mathcal{O}}(\sqrt{T})$ for LQG-NAIVE. This is achieved by exploiting and extending the analysis techniques for the naive exploration in the LQR setting [10] and the techniques for analyzing the CEC in the LQG setting [5]. Moreover, the performance of LQG-NAIVE and LQG-IF2E is validated in numerical simulations.

The contributions of this work can be summarized as follows:

- A novel regret guarantee of $\tilde{\mathcal{O}}(\sqrt{T})$ is established for a naive-exploration-based adaptive control algorithm in the LQG setting.
- A novel adaptive control algorithm that exploits the FIM for exploration is proposed for the LQG setting and is validated in simulations.

Due to space constraints, only the sketches of the proofs are presented. Full proofs can be found in the extended version of this paper in [12].

II. Preliminaries

A. Notations

The Euclidean norm of a vector x is denoted by $||x||$. For a matrix $X \in \mathbb{R}^{n\times m}$, $||X||$ denotes the spectral norm, $||X||_{\mathrm{F}}$ denotes the Frobenius norm, $\rho(X)$ denotes the spectral radius, and $\mathrm{Tr}(X)$ denotes the trace. The j^{th} singular value of a matrix X is denoted by $\sigma_j(X)$, where $\sigma_{\max}(X) := \sigma_1(X) \geq \sigma_2(X) \geq ... \geq \sigma_{\min}(X) := \sigma_{\min(n,m)}(X) \geq 0$. Similarly, $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ have analogous meanings for eigenvalues of a square matrix X. In this work, we use $\hat{X}_t$ to denote an estimate of X at time step t. The notation $\mathrm{diag}(\cdot)$ denotes a block diagonal matrix with the arguments as the blocks along the main diagonal. Given two functions $f(\cdot)$ and $g(\cdot)$, whose domain and co-domain are subsets of non-negative real numbers, we write $f(x) = \mathcal{O}(g(x))$ if $\exists\ c > 0$ and $\tilde{x} \geq 0$ such that $f(x) \leq cg(x)$ for all $x \geq \tilde{x}$. We write $f(x) = \Omega(g(x))$ if $\exists\ c > 0$ and $\tilde{x} \geq 0$ such that $f(x) \geq cg(x)$ for all $x \geq \tilde{x}$. The notations $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ ignore constants that depend on the true system parameters and poly-logarithmic factors that depend on the number of time steps T. The inequality $f(x) \lesssim g(x)$ denotes $f(x) \leq cg(x)$ for a universal constant $c > 0$. The notation $\mathrm{poly}(\cdot)$ denotes a polynomial function.

B. Problem setting

A discrete-time Linear Time-Invariant (LTI) system is characterized by the state-space equation

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t, \;\; w_t \sim \mathcal{N}(0, \sigma_w^2 I), \\ y_t &= Cx_t + z_t, \;\; z_t \sim \mathcal{N}(0, \sigma_z^2 I), \end{aligned} \tag{1}$$

for $t = 0, 1, 2, \ldots$, $A \in \mathbb{R}^{n_x\times n_x}$, $B \in \mathbb{R}^{n_x\times n_u}$, and $C \in \mathbb{R}^{n_y\times n_x}$. At time step t, $u_t \in \mathbb{R}^{n_u}$ is the input, $x_t \in \mathbb{R}^{n_x}$ is the state, $w_t \in \mathbb{R}^{n_x}$ is the process noise, $y_t \in \mathbb{R}^{n_y}$ is the system output, and $z_t \in \mathbb{R}^{n_y}$ is the measurement noise. Let the model parameter of the true system be $\Theta = (A, B, C)$. To measure the performance of a controller, the cost c_t incurred at time step t is defined as

$$c_t = y_t^\top Q y_t + u_t^\top R u_t,$$

where $Q \in \mathbb{R}^{n_y\times n_y}$ is positive semi-definite and $R \in \mathbb{R}^{n_u\times n_u}$ is positive-definite. In this work, the infinite-horizon setting is considered, wherein the goal is to design an input signal such that the long-term average expected cost is minimized. The long-term average expected cost in this setting is given by

$$J = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} c_t\right] \;\text{ s.t. (1).}$$

Let J_* denote the optimal long-term average expected cost for the true system with parameter Θ. The true system is assumed to be controllable and observable to ensure that J_* exists [2].

An optimal feedback control law minimizing J has the following form:

$$u_t = -K\hat{x}_{t|t,\Theta}, \tag{2}$$

where $\hat{x}_{t|t,\Theta}$ is the state estimate given the true parameter value Θ and the observations until time step t, and

$$K = (B^\top PB + R)^{-1}B^\top PA,$$

with P solving the following Discrete-Time Algebraic Riccati Equation (DARE) [2]:

$$P = C^\top QC + A^\top PA - A^\top PB(B^\top PB + R)^{-1}B^\top PA. \tag{3}$$

The state estimate $\hat{x}_{t|t,\Theta}$ in (2) is obtained from the Kalman filter:

$$\begin{aligned} \hat{x}_{t|t,\Theta} &= (I - LC)\hat{x}_{t|t-1,\Theta} + Ly_t, \\ \hat{x}_{t+1|t,\Theta} &= A\hat{x}_{t|t,\Theta} + Bu_t, \\ L &= \Sigma C^\top (C\Sigma C^\top + \sigma_z^2 I)^{-1}, \end{aligned} \tag{4}$$

where L is the Kalman gain, and Σ is the solution to the following DARE:

$$\Sigma = \sigma_w^2 I + A\Sigma A^\top - A\Sigma C^\top (C\Sigma C^\top + \sigma_z^2 I)^{-1} C\Sigma A^\top.$$

In (4), the two expressions concerning $\hat{x}_{t|t,\Theta}$ and $\hat{x}_{t+1|t,\Theta}$ can be combined to obtain the innovation form [13]:

$$\begin{aligned} \hat{x}_{t+1|t,\Theta} &= A\hat{x}_{t|t-1,\Theta} + Bu_t + Fe_t, \\ e_t &= C\left(x_t - \hat{x}_{t|t-1,\Theta}\right) + z_t, \\ e_t &\sim \mathcal{N}(0, \Sigma_e), \end{aligned} \tag{5}$$

where $\Sigma_e = C\Sigma C^\top + \sigma_z^2 I$ and $F = AL$ is the Kalman gain in the innovations form. Equation (5) can be expanded to obtain the one-step-ahead prediction model [13]:

$$\begin{aligned} \hat{x}_{t+1|t,\Theta} &= (A - FC)\hat{x}_{t|t-1,\Theta} + Bu_t + Fy_t, \\ \hat{y}_{t+1|t,\Theta} &= C\hat{x}_{t+1|t,\Theta}, \end{aligned} \tag{6}$$

where the Kalman gain here ensures that $A - FC$ is asymptotically stable. There exists a closed-form expression for the optimal long-term average expected cost when applying the optimal feedback control law (2) [14]:

$$J_* = \mathrm{Tr}\left(C^\top Q C \bar{\Sigma}\right) + \sigma_z^2 \mathrm{Tr}\left(Q\right) + \mathrm{Tr}\left(P(\Sigma - \bar{\Sigma})\right), \quad (7)$$

where

$$\bar{\Sigma} = \Sigma - \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1} C\Sigma.$$

In this work, the true model parameter Θ is unknown, whereas Q and R are user-defined (known). The main problem considered in this work is to design a controller that computes an input u_t based on the past observations $\mathcal{I}_t$:

$$\mathcal{I}_t = \{(y_k, u_k) \mid k = 0, 1, \ldots, t-1\} \cup \{y_t\}. \quad (8)$$

Moreover, the designed controller should perform optimally with respect to specific metrics. Following the literature [4], we consider the notion of regret as our metric. Given a finite time horizon of length T, the cumulative regret $\mathrm{Regret}(T)$ is given by

$$\mathrm{Regret}(T) = \sum_{t=0}^{T-1} (c_t - J_*), \quad (9)$$

where c_t is the cost incurred by the controller at time step t. It is desired to have a controller whose $\mathrm{Regret}(T)$ grows sub-linearly, i.e., $\lim_{T\to\infty} R(T)/T \to 0$ with high probability (w.h.p.). In this case, the average performance of the adaptive controller converges to the optimal average performance $J_\star$.

As Θ is unknown, given another parameter value $\Theta' = (A', B', C')$, the notations $K(\Theta')$, $L(\Theta')$, and $P(\Theta')$ denote the control gain in (2), the Kalman gain in (4), and the solution of the DARE in (3) respectively, obtained from the parameter value Θ'. The main assumptions of this work are summarized as follows and are considered to hold throughout the entire paper.

Assumption 2.1: Q, R are positive-definite, $x_0 \sim \mathcal{N}(0, \Sigma)$, and $\hat{x}_{0|-1,\Theta} = 0$. The state dimension n_x is known.

The positive definiteness of Q is required to quantify the sub-optimality in the long-term average cost [5, Th. 3]. Since the convergence of the Kalman filter gain is exponentially fast, the assumption on x_0 is not restrictive [3].

Assumption 2.2: The unknown model parameter Θ is an element in a set $\mathcal{S}$ satisfying

$$\mathcal{S} \subseteq \left\{ \Theta' = (A', B', C') \;\middle|\; \begin{array}{l} \rho(A') < 1, \\ (A', B') \text{ is controllable,} \\ (A', C') \text{ is observable,} \\ (A', F') \text{ is controllable.} \end{array} \right\}.$$

Assumption 2.2 is standard in the literature of finite-sample system identification and regret minimization [3], [5], [15]–[17]. The stability of the open-loop plant is assumed in Assumption 2.2 to avoid explosive behavior during the initial system identification phase [3].

Definition 2.1 ([7]): Given an invertible matrix $\mathbf{T} \in \mathbb{R}^{n_x \times n_x}$ and $\hat{\Theta}_t = (\hat{A}_t, \hat{B}_t, \hat{C}_t)$, the estimated model parameters at time step t, we define the following model mismatch pseudo-metric:

$$\tau(\hat{\Theta}_t, \Theta) := \min_{\mathbf{T}\in \mathrm{GL}_n} \max \left\{ \begin{array}{l} ||\hat{A}_t - \mathbf{T}^\top A \mathbf{T}||, \\ ||\hat{B}_t - \mathbf{T}^\top B||, \\ ||\hat{C}_t - C\mathbf{T}|| \end{array} \right\},$$

which is invariant under similarity transformations.

C. Closed-loop system identification

In this work, we adopt a model-based control approach, in which an estimate of the unknown parameter Θ is obtained and continuously updated online using a system identification technique. We specifically use the subspace identification [6]. We consider the predictor form in (6), and for the sake of brevity, we introduce the notation $\bar{A} = (A - FC)$. At time step t, we examine the system's evolution over the last H time steps, with the condition that $t \geq H$. Then we obtain

$$y_t = \mathbf{M}\phi_t + e_t + C\bar{A}^H \hat{x}_{t-H|t-H-1,\Theta}, \quad (10)$$

where

$$\mathbf{M} := \begin{bmatrix} M^{(0)} & \cdots & M^{(H-1)} \end{bmatrix} \in \mathbb{R}^{n_y \times (n_y+n_u)H}, \quad (11)$$

with $M^{(i)} := C\bar{A}^i[F\ B]$, and $\phi_t \in \mathbb{R}^{(n_y+n_u)H}$ is defined as

$$\phi_t := \begin{bmatrix} y_{t-1}^\top & \cdots & y_{t-H}^\top & u_{t-1}^\top & \cdots & u_{t-H}^\top \end{bmatrix}^\top. \quad (12)$$

Since $\bar{A}$ is stable, the last term in (10) becomes negligible for a large enough H, specifically for $H \geq \bar{H}$ with some $\bar{H} = \Omega(\log T)$. The exact expression of $\bar{H}$ can be found in [12, eq. 63]. Now with $\{y_i\}_{i=0}^t$ and $\{u_i\}_{i=0}^{t-1}$, we have

$$\begin{aligned} Y_t &= \Phi_t \mathbf{M}^\top + E_t + N_t \\ \implies Y_t &\approx \Phi_t \mathbf{M}^\top + E_t, \end{aligned} \quad (13)$$

where $Y_t = [y_H\ y_{H+1} \ldots y_t]^\top, \Phi_t = [\phi_H\ \phi_{H+1} \ldots \phi_t]^\top$, $N_t = [C\bar{A}^H \hat{x}_{0|-1,\Theta} \ldots C\bar{A}^H \hat{x}_{t-H|t-H-1,\Theta}]$, and $E_t = [e_H\ e_{H+1} \ldots e_t]^\top$. The approximation in (13) comes from the fact that N_t becomes negligible for a large enough H. Therefore, from (13), the Markov parameters $\mathbf{M}$ of the unknown true system can be estimated using regularized least squares [3]:

$$\hat{\mathbf{M}}_t^\top = (\Phi_t^\top \Phi_t + \lambda I)^{-1} \Phi_t^\top Y_t, \quad (14)$$

where $\lambda > 0$ is a regularization parameter. Define $V_t = \Phi_t^\top \Phi_t + \lambda I$.

Following [6], from $\hat{\mathbf{M}}_t$, a subroutine called SYSID will be deployed in the control algorithms of this work to obtain an estimate of the model parameters $\hat{A}_t, \hat{B}_t, \hat{C}_t, \hat{L}_t$. This subroutine is a variation of the classical Ho-Kalman realization algorithm [18], and details of this identification approach are found in [6].

III. Adaptive control with additive exploration

A. Naive exploration

This section presents LQG-NAIVE (Algorithm 1), which provides a computationally efficient method to address regret minimization in the LQG setting. Overall, the algorithm consists of two phases: the warm-up phase and the adaptive-control phase.

Warm-up phase: To obtain an initial CEC that can stabilize the unknown true system, an initial model parameter estimate is obtained. This is achieved through pure exploration by injecting Gaussian input signals for T_w time steps to effectively excite the system and then conducting system identification. The length T_{w} of this phase depends on how accurate the initial estimate needs to be [3]. Moreover, we let $T_{\text{w}} \geq H$.

Adaptive-control phase: Following the warm-up phase, the algorithm proceeds in an episodic fashion. The number of time steps l_k of the k^{th} episode satisfies $l_k = 2^k T_{\text{w}}$, for $k = 0, 1, 2, \ldots$ It holds that the time step at the beginning of the k^{th} episode equals l_k. Since $l_{k+1} = 2l_k$, the total number of episodes k_{fin} within a time horizon of length T is $\lfloor \log_2(T/T_{\text{w}}) \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function.

At the beginning of the k^{th} episode, LQG-NAIVE updates the parameter estimate to $\hat{\Theta}_{l_k}$. Then, within this episode, the corresponding CEC with an additive Gaussian excitatory signal is deployed:

$$\begin{aligned} u_t &= -K(\hat{\Theta}_{l_k})\hat{x}_{t|t,\hat{\Theta}_{l_k}} + \eta_t, \\ \eta_t &\sim \left(\gamma/\sqrt{l_k}\right)^{1/2} \mathcal{N}(0, I), \end{aligned} \tag{15}$$

where $\gamma > 0$ is a tuning parameter, and $K(\hat{\Theta}_{l_k})$ is the optimal feedback gain computed from $\hat{\Theta}_{l_k} = (\hat{A}_{l_k}, \hat{B}_{l_k}, \hat{C}_{l_k})$.

Algorithm 1 LQG-NAIVE

```
 1: Initialize $Q, R, \gamma > 0$, $H$, $T_{\text{w}}$, $n_x$, $n_y$, $n_u$, $\sigma_u^2$
 2: procedure WARM-UP
 3:     for $t = 0, 1, ..., T_{\text{w}} - 1$ do
 4:         Inject $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$
 5:     end for
 6:     Store $\{(y_t, u_t)\}_{t=0}^{T_{\text{w}}-1}$
 7: end procedure
 8: procedure ADAPTIVE CONTROL
 9:     for $k = 0, 1, ...$ do
10:         Let $l_k = 2^k T_{\text{w}}$
11:         Calculate $\hat{\mathbf{M}}_{l_k}$ using $\mathcal{I}_{l_k}$ and (14)
12:         Perform SYSID [6] to obtain $\hat{A}_{l_k}, \hat{B}_{l_k}, \hat{C}_{l_k}, \hat{L}_{l_k}$
13:         for $t = 2^k T_{\text{w}}, ..., 2^{k+1} T_{\text{w}} - 1$ do
14:             Inject $u_t$ as in (15)
15:         end for
16:     end for
17: end procedure
```

B. FIM-based exploration (LQG-IF2E)

The adaptive control algorithm based on the FIM employs the same structure as in Algorithm 1, with the only difference at step 14 for exploration. To present the algorithm, called the LQG-IF2E, we first provide an intuition for using the FIM. The recent work [11] uses the FIM to explicitly design the exploration signal for the LQR setting. The usage of FIM is particularly advantageous when adapting the magnitude of the additive excitation signal to unexpected disturbances [11]. Moreover, the FIM reflects the informativity of data and thus can be exploited to generate informative data for model re-estimation.

Definition 3.1 ([8]): For a family of parameterized probability densities $\{p_\theta, \theta \in \bar{\mathcal{S}}\}$ of a random variable $x \in \mathbb{R}^n$, where $\bar{\mathcal{S}} \subseteq \mathbb{R}^d$, the FIM $\bar{I}_p(\theta) \in \mathbb{R}^{d \times d}$ is given by

$$\bar{I}_p(\theta) = \int_{\mathbb{R}^n} \nabla_\theta \log p_\theta(x) \left(\nabla_\theta \log p_\theta(x)\right)^\top p_\theta(x) dx, \tag{16}$$

whenever the integral exists.

In the present work, the FIM is constructed with respect to the Markov parameters $(\mathbf{M})$ that govern the dynamics of the approximate model (13). In this case, the FIM under any policy π after collecting the observations $\{(y_i, u_i)\}_{i=0}^{t-1}$ for $t \geq H$, is given by

$$I_{H,t} = \sum_{i=H}^{t} \mathbb{E}\left[\phi_i \phi_i^\top \otimes \Sigma_e^{-1}\right]. \tag{17}$$

The FIM cannot be constructed for the first H time steps since the ϕ_t vector is not defined during this period. This is acceptable because after the warm-up phase, sufficient data is collected, i.e., $T_{\text{w}} \geq H$, to construct the FIM, which is then used in the adaptive control phase. The derivation of (17) is presented in the extended version of this paper [12, Lemma 6.7] and is an extension of [8, Lemma 3.3] from the LQR setting to the LQG setting.

There is however a caveat in using the FIM: the FIM requires knowledge of the unknown true parameter Θ, as in (17). To circumvent this issue, we evaluate the FIM instead at $\hat{\Theta}_{l_k}$. Even if $\hat{\Theta}_{l_k}$ can only converge to a similarity transformation of Θ, the eigenvalues of a matrix are preserved under similarity transformation, and thus one can evaluate the FIM with $\hat{\Theta}_{l_k}$. For the simplicity of notations, we use $\hat{\Theta}_t$ here to denote the estimated parameter at time step t to estimate the FIM, and note that in Algorithm 1, $\hat{\Theta}_t = \hat{\Theta}_{l_k}$ when $t \in [l_k,\, l_{k+1})$. This holds for other estimates as well, e.g., $\hat{\mathbf{M}}_t = \hat{\mathbf{M}}_{l_k}$ when $t \in [l_k,\, l_{k+1})$. Therefore, we can estimate the 'true' FIM as

$$\hat{I}_{H,t} = \sum_{i=H}^{t} \phi_i \phi_i^\top \otimes \hat{\Sigma}_{e,i}^{-1}, \tag{18}$$

and

$$\hat{\Sigma}_{e,i} = \frac{1}{i+1} \sum_{j=0}^{i} \left(y_j - \hat{y}_{j|j-1,\hat{\Theta}_{j-1}}\right)\left(y_j - \hat{y}_{j|j-1,\hat{\Theta}_{j-1}}\right)^\top,$$

with $\hat{y}_{j|j-1,\hat{\Theta}_{j-1}} = \hat{C}_{j-1}\hat{x}_{j|j-1,\hat{\Theta}_{j-1}}$.

To ensure that the FIM is not ill-conditioned, the exploration strategy in (15) is used until $\lambda_{\text{min}}\left(\hat{I}_{H,t}\right)$ is larger than some tolerance value. After achieving this tolerance, the FIM-based exploration strategy is deployed. That is, given $c_{\text{tol}} > 0$, if $\lambda_{\text{min}}\left(\hat{I}_{H,t}\right) \geq c_{\text{tol}}$,

$$\begin{aligned} u_t &= -K(\hat{\Theta}_{l_k})\hat{x}_{t|t,\hat{\Theta}_{l_k}} + \eta_t, \\ \eta_t &\sim \left(\alpha/\lambda_{\text{min}}\left(\hat{I}_{H,t}\right)\right)^{1/2} \mathcal{N}(0, I), \end{aligned} \tag{19}$$

where $\alpha > 0$ is a tuning parameter. Given that $\hat{I}_{H,t}$ depends on past inputs and outputs through the vector ϕ_t, the FIM-based exploration strategy is a type of 'closed-loop' exploration strategy capable of adaptively changing the magnitude of the exploration signal to the 'degree' of informativity.

IV. Regret guarantee

We now establish a theoretical guarantee on the regret growth of LQG-NAIVE. To this end, this section presents a finite-time guarantee on the persistency of excitation of the data, which is necessary for parameter estimation, and a guarantee that the closed-loop system is stabilized during the adaptive control phase. From these two guarantees, we ensure that the model parameter estimation error is upper-bounded by a monotonically decreasing rate of $\tilde{\mathcal{O}}(1/\sqrt{t})$.

Warm-up phase: The modeling error of the initial parameter estimate after the warm-up phase can be bounded as shown in [3], [19] under the same setting. From [3, Lemma 3.1], the input-output data persistently excites the underlying system during the warm-up period, i.e., $\sigma_{\min}(V_{T_\mathrm{w}}) = \Omega(T_\mathrm{w})$. Further, [3, Th. 3.3] shows that

$$||\hat{\mathbf{M}}_{T_\mathrm{w}} - \mathbf{M}|| \leq \frac{\beta_{T_\mathrm{w}}}{\sqrt{\sigma_{\min}(V_{T_\mathrm{w}})}} = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T_\mathrm{w}}}\right), \tag{20}$$

with a probability of at least $1-\delta$, where

$$\beta_{T_\mathrm{w}} := \sqrt{n_y||\Sigma_e|| \log\left(\frac{\det(V_{T_\mathrm{w}})^{1/2}}{\delta\det(\lambda I)^{1/2}}\right)} + ||\mathbf{M}||_\mathrm{F}\sqrt{\lambda} + \frac{\sqrt{H}}{T_\mathrm{w}}$$

for $\delta \in (0,1)$ and $T_\mathrm{w} \geq H \geq \bar{H}$. [3, Th. 3.4] shows that if $H \geq \bar{H}$, then

$$\tau(\hat{\Theta}_t, \Theta) = \mathcal{O}\left(||\hat{\mathbf{M}}_t - \mathbf{M}||\right) w.h.p. \tag{21}$$

Combining the above result with (20) shows that $\tau(\hat{\Theta}_{T_\mathrm{w}}, \Theta) = \tilde{\mathcal{O}}(1/\sqrt{T_\mathrm{w}})$ w.h.p. If $T_\mathrm{w} \geq \bar{T}_\mathrm{w}$ for some positive integer $\bar{T}_\mathrm{w}$, we have $\tau(\hat{\Theta}_{T_\mathrm{w}}, \Theta) \leq \epsilon_\mathrm{w}$, where ϵ_w is a positive constant and the exact formulation of $\bar{T}_\mathrm{w}$ can be found in [12, Appendix].

Adaptive Control Period

During the adaptive control period, it is imperative to guarantee that the input and output signals remain bounded with high probability, to ensure the safe operation of the closed-loop system. Such guarantee is provided with LQG-NAIVE, as shown in the following lemma:

Lemma 4.1: For all $t \geq T_\mathrm{w}$ with $T_\mathrm{w} \geq \bar{T}_\mathrm{w}$, LQG-NAIVE satisfies the following with a probability of at least $1-\delta$ for $\delta \in (0,1)$:

$$\begin{aligned} &||\hat{x}_{t|t,\hat{\Theta}_t}|| \leq \bar{\mathcal{X}}, \;\; ||\hat{x}_{t|t-1,\hat{\Theta}_{t-1}}|| \leq X_\text{est,ac}, \\ &||y_t|| \leq Y_\mathrm{ac}, \;\; ||u_t|| \leq U_\mathrm{ac}, \;\; ||x_t|| \leq X_\mathrm{ac}, \end{aligned} \tag{22}$$

for some $\bar{\mathcal{X}}, X_\text{est,ac}, U_\mathrm{ac}, Y_\mathrm{ac}, X_\mathrm{ac} = \mathcal{O}(\sqrt{\log(T/\delta)})$.

Another important pre-requirement for the regret guarantee is to ensure the informativity of the data. To guarantee informativity, we consider a sufficient number of time steps T_ac after the warm-up phase. The detailed formulation of T_ac can be found in [12, eq. 60] of the extended version of this paper. Then the guarantee for informativity is presented in the following result:

Lemma 4.2: If $T_\mathrm{w} \geq \bar{T}_\mathrm{w}$, we have the following with probability of at least $1-\delta$ for $\delta \in (0,1)$: for all $t \geq T_\mathrm{ac} + T_\mathrm{w}$ and for some constant $\sigma_\mathrm{c} > 0$,

$$\sigma_{\min}\left(\sum_{i=T_\mathrm{w}}^{t} \phi_i\phi_i^\top\right) \geq (t - T_\mathrm{w} + 1)\frac{\sigma_\mathrm{c}^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_{\eta_{t-1}}^2\}}{8}. \tag{23}$$

From the persistence of excitation property in (23), we can now provide a bound on the parameter estimation error during the adaptive control phase.

Lemma 4.3: If $T_\mathrm{w} \geq \bar{T}_\mathrm{w}$, for any $t \geq \max\{T_\mathrm{ac} + T_\mathrm{w}, 2T_\mathrm{w}\}$, the estimate of the Markov parameters, $\hat{\mathbf{M}}_t$, obeys the following bound with a probability of at least $1-\delta$ for $\delta \in (0,1)$:

$$||\hat{\mathbf{M}}_t - \mathbf{M}|| \leq \frac{\bar{\beta}_\mathrm{ac}}{\sqrt{\sigma_{\min}(V_t)}} = \tilde{\mathcal{O}}(1/\sqrt{t}), \tag{24}$$

for some $\bar{\beta}_\mathrm{ac} = \mathrm{poly}(n_y, \Sigma_e, \delta, Y_\mathrm{ac}, U_\mathrm{ac})$.

The complete proof can be found in the extended version of this paper [12, Lemma 4.3]. From (21) and Lemma 4.3, we have $\tau(\hat{\Theta}_t, \Theta) = \tilde{\mathcal{O}}(1/\sqrt{t})$. Therefore, the model parameter estimation error is monotonically decreasing.

The final piece in establishing the regret upper bound requires bounding the sub-optimality gap $\Delta_{\hat{\Theta}_t} := J(\hat{\Theta}_t) - J_*$. This inherently requires a way to represent $J(\hat{\Theta}_t)$. It is a standard procedure to write the long-term average expected cost as a function of the solution to a Lyapunov equation [10], and thus we connect $J(\hat{\Theta}_t)$ to a Lyapunov equation as follows.

Consider the true system (1), and another model parameter $\tilde{\Theta} = (\tilde{A}, \tilde{B}, \tilde{C}) \in \mathcal{S}$. Let $\tilde{K} = K(\tilde{\Theta})$ and $\tilde{L} = L(\tilde{\Theta})$. Now, define an alternative formulation of the LQG cost function as

$$\begin{aligned} J_s(\tilde{\Theta}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}&\left[\sum_{t=0}^{T-1} x_t^\top Q_c x_t + \hat{x}_{t|t,\tilde{\Theta}}^\top \tilde{K}^\top R\tilde{K}\hat{x}_{t|t,\tilde{\Theta}}\right], \\ &\text{s.t. (1)}, \\ &\hat{x}_{t|t,\tilde{\Theta}} = (I - \tilde{L}\tilde{C})\hat{x}_{t|t-1,\tilde{\Theta}} + \tilde{L}y_t, \\ &\hat{x}_{t+1|t,\tilde{\Theta}} = \tilde{A}\tilde{x}_{t|t,\tilde{\Theta}} + \tilde{B}u_t, \\ &u_t = -\tilde{K}\hat{x}_{t|t,\tilde{\Theta}}, \end{aligned}$$

where $Q_c = C^\top QC$, $\tilde{K}$ stabilises the true system, and $\tilde{A} - \tilde{F}\tilde{C}$ is asymptotically stable. This alternative formulation of the quadratic cost shows up when upper bounding the cumulative cost in the regret analysis. Further, consider the following closed-loop state-space equation with extended states:

$$\begin{bmatrix} x_t \\ \hat{x}_{t|t,\tilde{\Theta}} \end{bmatrix} = \tilde{\mathbf{G}}_\mathbf{1} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\tilde{\Theta}} \end{bmatrix} + \tilde{\mathbf{G}}_\mathbf{2} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix},$$

where

$$\tilde{\mathbf{G}}_\mathbf{1} = \begin{bmatrix} A & -B\tilde{K} \\ \tilde{L}CA & \left(I - \tilde{L}\tilde{C}\right)\left(\tilde{A} - \tilde{B}\tilde{K}\right) - \tilde{L}CB\tilde{K} \end{bmatrix},$$

$$\tilde{\mathbf{G}}_\mathbf{2} = \begin{bmatrix} I & 0 \\ \tilde{L}C & \tilde{L} \end{bmatrix}.$$

Consider the discrete Lyapunov equation $\tilde{S} = \tilde{\mathbf{G}}_\mathbf{1}^\top \tilde{S}\tilde{\mathbf{G}}_\mathbf{1} + \mathrm{diag}(Q_c, \tilde{K}^\top R\tilde{K})$ with $\tilde{S}$ being its positive semi-definite solution. Then, we have

$$J_s(\tilde{\Theta}) = \mathrm{Tr}\left(\tilde{\mathbf{G}}_{\mathbf{2}}^{\top} \tilde{S} \tilde{\mathbf{G}}_{\mathbf{2}} \mathrm{diag}(\sigma_w^2 I, \sigma_z^2 I)\right). \tag{25}$$

Moreover, it holds that $J(\tilde{\Theta}) = J_s(\tilde{\Theta}) + \mathrm{Tr}(Q\sigma_z^2 I)$ [5, Th. 3]. This property can aid in quantifying the sub-optimality gap $\Delta_{\hat{\Theta}_t}$. Now, we are ready to state the regret upper bound.

Theorem 4.1: If $T_{\mathrm{w}} \geq \bar{T}_{\mathrm{w}}$, with a probability of at least $1-\delta$ for $\delta \in (0,1)$, we have for any $T \geq \max\{T_{\mathrm{ac}}+T_{\mathrm{w}}, 2T_{\mathrm{w}}\}$ that the regret of LQG-NAIVE is bounded as

$$\begin{aligned}
\mathrm{Regret}(T) &\lesssim \sum_{k=0}^{k_{\mathrm{fin}}-1} l_k \left(J_s(\hat{\Theta}_{l_k}) - J_*\right) + l_k n_y \sigma_z^2 \mathrm{Tr}\left(Q\right) \\
&+ l_k \sigma^2_{\eta_{l_k}} \mathrm{poly}\left(\tau(\hat{\Theta}_{l_k}, \Theta)\right) \\
&+ \sqrt{l_k} \mathrm{poly}\left(\tau(\hat{\Theta}_{l_k}, \Theta), X_{\mathrm{ac}}, \bar{\mathcal{X}}, ||Q||, ||R||\right) \qquad (26) \\
&= \tilde{\mathcal{O}}(\sqrt{T})
\end{aligned}$$

where l_k is the number of time steps in the k^{th} episode, k_{fin} is the total number of episodes and $\sigma^2_{\eta_{l_k}} = \gamma/\sqrt{l_k}$.

The result in Theorem 4.1 confirms that a naive-exploration-based adaptive control strategy is sufficient to guarantee a $\sqrt{T}$-regret growth, which is the optimal rate of regret growth up to logarithmic terms in the control of unknown partially observable linear systems. The proof is presented in the extended version of this paper [12, Th. 4.1]. Now, an intuition is provided on how the regret bound is derived. As Algorithm 1 operates in an episodic fashion, the regret is also analyzed episode-wise. First, an upper bound on the cumulative cost incurred by LQG-NAIVE in any arbitrary episode is obtained. From this, we can obtain an upper bound on the regret for any episode. This episode-wise regret bound is then summed over the number of episodes to obtain the final regret upper bound incurred by LQG-NAIVE during the adaptive control phase as shown in (26).

In (26), the sub-optimality gap $J_s(\hat{\Theta}_{l_k}) - J_*$ and the exploration cost $\sigma^2_{\eta_{l_k}} \mathrm{poly}\left(\tau(\hat{\Theta}_{l_k}, \Theta)\right)$ have significant contributions towards the regret, as they are linearly dependent on l_k. To provide a bound on the sub-optimality gap, we exploit an earlier result [5, Th. 4], which essentially bounds the contribution from the sub-optimality gap by $\tilde{\mathcal{O}}(\log_2(T/T_{\mathrm{w}}))$. On the other hand, the exploration cost is bounded by $\tilde{\mathcal{O}}(\sqrt{T})$ as $\sigma^2_{\eta_{l_k}} = \gamma/\sqrt{l_k}$. We can further see from the 3rd and 4th terms in (26) that the model parameter estimation error along with the established bounds on the state and its estimate, also influence the regret upper bound. From (21) and Lemma 4.3, we have that the model parameter estimation errors are monotonically decreasing, and as a consequence, $\tau(\hat{\Theta}_t, \Theta) \leq \tau(\hat{\Theta}_{T_{\mathrm{w}}}, \Theta)$ with a high probability. This result is used in deriving the regret upper bound (26).

V. Numerical simulations

In this section, we validate the performance of LQG-NAIVE and LQG-IF2E through numerical simulations. For the simulation, we consider a linearized version of the web server control problem [20]. Different from [20], we consider the partial observability case, i.e., the inclusion of the C matrix and the measurement noise. The true system under consideration is given by

$$\begin{aligned}
x_{t+1} &= \begin{bmatrix} 0.54 & -0.11 \\ -0.026 & 0.63 \end{bmatrix} x_t + \begin{bmatrix} -85 & 4.4 \\ -2.5 & 2.8 \end{bmatrix} u_t + w_t, \\
y_t &= \begin{bmatrix} 0.2 & 0.3 \\ 0.3 & 0.2 \end{bmatrix} x_t + z_t,
\end{aligned}$$

where $w_t, z_t \sim \mathcal{N}(0, 0.01I)$. The cost matrices for the control problem are given by [20]:

$$Q = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} \frac{1}{50^2} & 0 \\ 0 & \frac{1}{10^6} \end{bmatrix}.$$

The optimal long-term average expected cost calculated from (7) is 0.0707.

To implement the adaptive control algorithm, the length of the warm-up phase is set to $T_{\mathrm{w}} = 25$. During the warm-up phase, Gaussian excitatory signals are injected, where $u_t \sim \mathcal{N}(0, 0.1I)$. For the adaptive control phase, the number of episodes is taken to be $k_{\mathrm{fin}} = 11$. The hyper-parameters for the adaptive control policies (15) and (19) are $\gamma = \frac{\sqrt{T_{\mathrm{w}}}}{10}$ and $\alpha = 1$ respectively. To avoid ill-conditioned FIM, we select $c_{\mathrm{tol}} = 1$. Finally, the length of the input-output data for constructing the ϕ vector in system identification is $H = 12$. Each of the algorithms LQG-NAIVE and LQG-IF2E are run 100 times to report the mean and the standard deviation of the regret growth.

Fig. 1 shows the regret growth of the 100 simulations. The bold red line represents the mean regret of LQG-NAIVE, whereas the bold blue line represents the mean regret of LQG-IF2E. LQG-NAIVE incurs a long-term average cost of 0.0744 and LQG-IF2E incurs a long-term average cost of 0.0742, averaged over the 100 simulations. The LQG-IF2E algorithm switches to the FIM-based exploration strategy at the 35th time step, on average. This means that with a delay of approximately one episode, the algorithm is able to deploy the FIM-based exploration strategy. The hyper-parameter α is chosen such that LQG-NAIVE and LQG-IF2E have similar behavior for the regret growth, which is evident from Fig. 1. An intuitive way to understand this similarity in regret growth is by plotting the evolution of the minimum eigenvalue of the FIM.

Fig. 2 shows how the minimum eigenvalue of the FIM varies over the time steps. The bold blue line represents the mean growth of $\lambda_{\min}\left(\hat{I}_{H,t}\right)$ of LQG-IF2E, whereas the bold red line represents the mean growth of $\lambda_{\min}\left(\hat{I}_{H,t}\right)$ of LQG-NAIVE. Based on Fig. 2, the behaviors of the FIMs are also similar between the two algorithms. Since the FIM captures the informativity of the data, which influences the parameter learning rate and hence the growth of the regret, one can expect two algorithms to have similar regret growth if their corresponding FIMs behaves similarly. The 'bumps' that are observed in Fig. 2 correspond to the time steps where the model parameter estimate $\hat{\Theta}_t$ was updated. The length of the 'bumps' corresponds approximately to the length of the episodes.

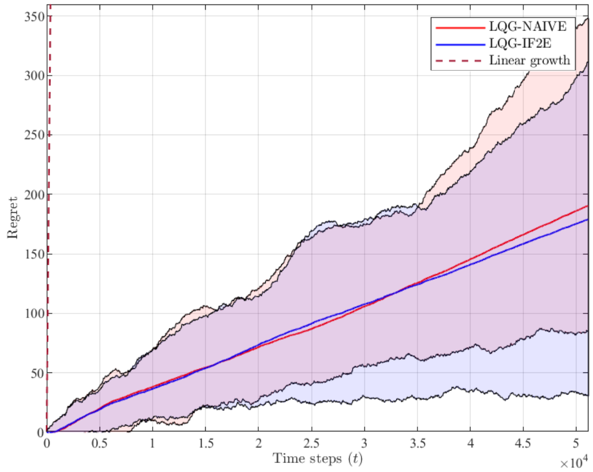

Fig. 1. Regret growth of LQG-NAIVE and LQG-IF2E.

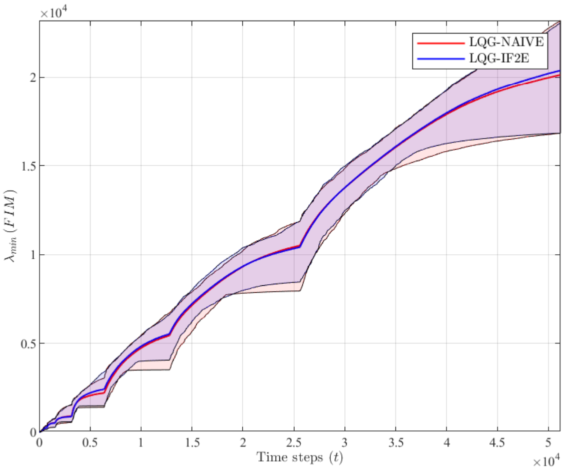

Fig. 2. Growth of the minimum eigenvalue of the estimated FIM.

VI. Conclusions

We have focused on control of unknown partially observable LTI systems in an LQG setting. We have developed two computationally efficient adaptive control algorithms: LQG-NAIVE and LQG-IF2E. The LQG-NAIVE algorithm, based on naive exploration, is more computationally efficient than optimism-in-the-face-of-uncertainty-based exploration. It also has a guaranteed regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$. However, in the regret upper bound, determining how the system constants scale with the dimensions is a topic for future work.

On the other hand, LQG-IF2E extends the 'open-loop' additive excitation signal in LQG-NAIVE to a 'closed-loop' additive excitation by incorporating FIM in designing the covariance of the exploration signal. However, providing finite-time regret guarantees for LQG-IF2E is significantly more challenging, as the additive excitation signal is not i.i.d. This is because the FIM depends on all the previous observations. Both algorithms have been validated in numerical simulations and show competitive performance. Deriving finite-time regret guarantees for FIM-based adaptive control strategies such as LQG-IF2E is also a topic for future work.

References

[1] N. Matni, A. Proutiere, A. Rantzer, and S. Tu, "From self-tuning regulators to reinforcement learning and back again," in *IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 3724–3740.

[2] D. Bertsekas, *Dynamic Programming and Optimal Control: Volume I*. Athena Scientific, 2012, vol. 4.

[3] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Adaptive control and regret minimization in LQG setting," in *American Control Conference (ACC)*, 2021, pp. 2517–2522.

[4] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of LQ systems," in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 1–26.

[5] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for LQ control," *Conference on Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems," *Conference on Advances in Neural Information Processing Systems*, vol. 33, pp. 20 876–20 888, 2020.

[7] T. Kargin, S. Lale, K. Azizzadenesheli, A. Anandkumar, and B. Hassibi, "Thompson sampling for partially observable LQ control," in *American Control Conference (ACC)*, 2023, pp. 4561–4568.

[8] I. Ziemann and H. Sandberg, "Regret lower bounds for learning LQG systems," *arXiv preprint arXiv:2201.01680*, 2022.

[9] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite-sample perspective," *IEEE Control Systems Magazine*, vol. 43, no. 6, pp. 67–97, 2023.

[10] M. Simchowitz and D. Foster, "Naive exploration is optimal for online LQR," in *International Conference on Machine Learning*, 2020, pp. 8937–8948.

[11] K. Colin, M. Ferizbegovic, and H. Hjalmarsson, "Regret minimization for LQ adaptive controllers using Fisher feedback exploration," *IEEE Control Systems Letters*, vol. 6, pp. 2870–2875, 2022.

[12] A. Athrey, O. Mazhar, M. Guo, B. De Schutter, and S. Shi, "Regret analysis of learning-based linear quadratic gaussian control with additive exploration," *arXiv preprint arXiv:2311.02679*, 2023.

[13] M. Verhaegen and V. Verdult, *Filtering and System Identification: A Least Squares Approach*. Cambridge University Press, 2007.

[14] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Regret minimization in partially observable LQ control," *arXiv preprint arXiv:2002.00082*, 2020.

[15] S. Oymak and N. Ozay, "Revisiting Ho–Kalman-based system identification: Robustness and finite-sample analysis," *IEEE Transactions on Automatic Control*, vol. 67, no. 4, pp. 1914–1928, 2021.

[16] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Nonparametric finite time LTI system identification," *arXiv preprint arXiv:1902.01848*, 2019.

[17] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," in *IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 3648–3654.

[18] B. Ho and R. E. Kálmán, "Effective construction of linear state-variable models from input/output functions," *at-Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.

[19] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Regret bound of adaptive control in LQG systems," *arXiv preprint 2003.05999*, 2020.

[20] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2021.