Scalable Regret for Learning to Control Network-Coupled Subsystems With Unknown Dynamics

Sagar Sudhakara, Aditya Mahajan, *Senior Member, IEEE*,
Ashutosh Nayyar, *Senior Member, IEEE*, and Yi Ouyang

***Abstract*—In this article, we consider the problem of controlling an unknown linear quadratic Gaussian (LQG) system consisting of multiple subsystems connected over a network. Our goal is to minimize and quantify the regret (i.e., loss in performance) of our learning and control strategy with respect to an oracle who knows the system model. Upfront viewing the interconnected subsystems globally and directly using existing LQG learning algorithms for the global system results in a regret that increases superlinearly with the number of subsystems. Instead, we propose a new Thompson sampling-based learning algorithm which exploits the structure of the underlying network. We show that the expected regret of the proposed algorithm is bounded by $\tilde{\mathcal{O}}(n\sqrt{T})$, where n is the number of subsystems and T is the time horizon. Thus, the regret scales linearly with the number of subsystems. We present numerical experiments to illustrate the salient features of the proposed algorithm.**

***Index Terms*—Linear quadratic systems, networked control systems, reinforcement learning, Thompson sampling (TS).**

I. Introduction

LARGE-SCALE systems comprising multiple subsystems connected over a network arise in a number of applications including power systems, traffic networks, communication networks, and some economic systems [1]. A common feature of such systems is the coupling in their subsystems' dynamics and costs, i.e., the state evolution and local costs of one subsystem depend not only on its own state and control action but also on the states and control actions of other subsystems in the network. Analyzing various aspects of the behavior of such systems and designing control strategies for them under a variety of settings have been long-standing problems of interest in the systems and control literature [2]–[6]. However, there are still many unsolved challenges, especially on the interface between learning and control in the context of these large-scale systems.

In this article, we investigate the problem of designing control strategies for large-scale network-coupled subsystems when some parameters of the system model are not known. Due to the unknown parameters, the control problem is also a learning problem. We adopt a reinforcement learning framework for this problem with the goal of minimizing and quantifying the regret (i.e., loss in performance) of our learning and control strategy with respect to the optimal control strategy based on the complete knowledge of the system model.

The networked system we consider follows linear dynamics with quadratic costs and Gaussian noise. Such linear quadratic Gaussian (LQG) systems are one of the most commonly used modeling frameworks in numerous control applications. Part of the appeal of LQG models is the simple structure of the optimal control strategy when the system model is completely known—the optimal control action in this case is a linear or affine function of the state—which makes the optimal strategy easy to identify and easy to implement. If some parameters of the model are not fully known during the design phase or may change during operation, then it is better to design a strategy that learns and adapts online. Historically, both adaptive control [7] and reinforcement learning [8], [9] have been used to design asymptotically optimal learning algorithms for such LQG systems. In recent years, there has been considerable interest in analyzing the transient behavior of such algorithms which can be quantified in terms of the regret of the algorithm as a function of time. This allows one to assess, as a function of time, the performance of a learning algorithm compared to an oracle who knows the system parameters upfront.

Several learning algorithms have been proposed for LQG systems [10]–[21], and, in most cases, the regret is shown to be bounded by $\tilde{\mathcal{O}}(d_x^{0.5}(d_x + d_u)\sqrt{T})$, where d_x is the dimension of the state, d_u is the dimension of the controls, T is the time horizon, and the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic terms in T. Given

Manuscript received 16 February 2022; revised 28 April 2022; accepted 5 June 2022. Date of publication 17 June 2022; date of current version 13 March 2023. The work of Sagar Sudhakara and Ashutosh Nayyar was supported by NSF under Grant ECCS 1750041 and Grant ECCS 2025732. The work of Aditya Mahajan was supported in part by the Innovation for Defence Excellence and Security (IDEaS) Program under Grant CFPMN2-30. Recommended by Associate Editor G. Russo. *(Corresponding author: Aditya Mahajan.)*

Sagar Sudhakara and Ashutosh Nayyar are with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90007 USA (e-mail: sagarsud@usc.edu; ashutosn@usc.edu).

Aditya Mahajan is with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0G4, Canada (e-mail: aditya.mahajan@mcgill.ca).

Yi Ouyang is with the Preferred Networks America, Burlingame, CA 94010 USA (e-mail: yio@usc.edu).

Digital Object Identifier 10.1109/TCNS.2022.3184107

the lower bound of $\tilde{\Omega}(d_x^{0.5} d_u \sqrt{T})$ (where $\tilde{\Omega}(\cdot)$ notation hides logarithmic terms in T) for regret in LQG systems identified in a recent work [18], the regrets of the existing algorithms have near optimal scaling in terms of time and dimension. However, when directly applied to a networked system with n subsystems, these algorithms would incur $\tilde{\mathcal{O}}(n^{1.5} d_x^{0.5}(d_x + d_u)\sqrt{T})$ regret because the effective dimensions of the state and the controls are nd_x and nd_u, where d_x and d_u are the dimensions of each subsystem. This super-linear dependence on n is prohibitive in large-scale networked systems because the regret per subsystem (which is $\tilde{\mathcal{O}}(\sqrt{n})$) grows with the number of subsystems.

The learning algorithms mentioned above are for a general LQG system and do not take into account any knowledge of the underlying network structure. Our main contribution is to show that by exploiting the structure of the network model, it is possible to design learning algorithms for large-scale network-coupled subsystems where the regret does not grow super-linearly in the number of subsystems. In particular, we utilize a spectral decomposition technique, recently proposed in [22], to decompose the large-scale system into L decoupled systems, where L is the rank of the coupling matrix corresponding to the underlying network. Using the decoupled systems, we propose a Thompson sampling (TS)-based algorithm with $\tilde{\mathcal{O}}(nd_x^{0.5}(d_x + d_u)\sqrt{T})$ regret bound.

A. Related Work

Broadly speaking, three classes of low-regret learning algorithms have been proposed for LQG systems: certainty equivalence (CE) based algorithms, optimism in the face of uncertainty (OFU) based algorithms, and TS-based algorithms. CE is a classical adaptive control algorithm [7]. Recent works [15]–[19] have established near optimal high probability bounds on regret for CE-based algorithms. OFU-based algorithms are inspired by the OFU principle for multiarmed bandits [23]. Starting with the work of Campi et al. [10] and Yadkori and Szepesvári [11], most of the works following the OFU approach [12]–[14] also provide similar high probability regret bounds. TS-based algorithms are inspired by TS algorithm for multiarmed bandits [24]. Most of the works following this approach [19]–[21] establish bounds on expected Bayesian regret of similar near-optimal orders. As argued earlier, most of these works show that the regret scales super-linearly with the number of subsystems and are, therefore, of limited value for large-scale systems.

There is an emerging literature on learning algorithms for networked systems both for LQG models [25]–[30] and Markov decision process (MDP) models [31]–[33]. The works on LQG models propose distributed value based or policy-based learning algorithms and analyze their convergence properties, but they do not characterize their regret. Some of the works on MDP models [32], [33] do characterize regret bounds for OFU- and TS-based learning algorithms, but these bounds are not directly applicable to the LQG model considered in this article.

An important special class of network-coupled systems is mean-field coupled subsystems [34], [35]. There has been considerable interest in reinforcement learning for mean-field models [36], [37], but most of the literature does not consider regret. The basic mean-field coupled model can be viewed as a special case of the network-coupled subsystems considered in this article (see Section VI-A). In a preliminary version of this article [38], we proposed a TS-based algorithm for mean-field coupled subsystems which has a $\tilde{\mathcal{O}}((1 + 1/n)\sqrt{T})$ regret per subsystem. The current article extends the TS-based algorithm to general network-coupled subsystems and establishes scalable regret bounds for arbitrarily coupled networks.

B. Organization

The rest of the article is organized as follows. In Section II, we introduce the model of network-coupled subsystems. In Section III, we summarize the spectral decomposition idea and the resulting scalable method for synthesizing optimal control strategy when the model parameters are known. Then, in Section IV, we consider the learning problem for unknown network-coupled subsystems and present a TS-based learning algorithm with scalable regret bound. We subsequently provide regret analysis in Section V and numerical experiments in Section VI. Finally, Section VII concludes this article.

C. Notation

The notation $A = [a^{ij}]$ means that A is the matrix that has a^{ij} as its (i, j)th element. For a matrix A, A^T denotes its transpose. Given matrices (or vectors) A_1, ..., A_n with the same number of rows, $[A_1, \ldots, A_n]$ denotes the matrix formed by horizontal concatenation. For a random vector v, $\operatorname{var}(v)$ denotes its covariance matrix. The notation $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate Gaussian distribution with mean vector μ and covariance matrix Σ.

For stabilizable (A, B) and positive-definite matrices Q, R, $\text{DARE}(A, B, Q, R)$ denotes the unique positive semidefinite solution of the discrete time algebraic Riccati equation (DARE), which is given as

$$S = A^\mathsf{T} SA - (A^\mathsf{T} SB)(R + B^\mathsf{T} SB)^{-1}(B^\mathsf{T} SA) + Q.$$

II. Model of Network-Coupled Subsystems

We start by describing a minor variation of a model of network-coupled subsystems proposed in [22]. The model in [22] was described in continuous time. We translate the model and the results to discrete time.

A. System Model

1) Graph Structure: Consider a network consisting of n subsystems/agents connected over an undirected weighted simple graph denoted by $\mathcal{G}(N, E, \Psi)$, where $N = \{1, \ldots, n\}$ is the set of nodes, $E \subseteq N \times N$ is the set of edges, and $\Psi = [\psi^{ij}] \in \mathbb{R}^{n \times n}$ is the weighted adjacency matrix. Let $M = [m^{ij}] \in \mathbb{R}^{n \times n}$ be a symmetric coupling matrix corresponding to the underlying graph $\mathcal{G}$. For instance, M may represent the underlying adjacency matrix (i.e., $M = \Psi$) or the underlying Laplacian matrix (i.e., $M = \operatorname{diag}(\Psi \mathbb{1}_n) - \Psi$).

2) State and Dynamics: The states and control actions of agents take values in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_u}$, respectively. For agent $i \in$

N, we use $x_t^i \in \mathbb{R}^{d_x}$ and $u_t^i \in \mathbb{R}^{d_u}$ to denote its state and control action at time t.

The system starts at a random initial state $x_1 = (x_1^i)_{i \in N}$, whose components are independent across agents. For agent i, the initial state $x_1^i \sim \mathcal{N}(0, \Xi_1^i)$, and at any time $t \geq 1$, the state evolves according to

$$x_{t+1}^i = Ax_t^i + Bu_t^i + Dx_t^{\mathcal{G},i} + Eu_t^{\mathcal{G},i} + w_t^i \tag{1}$$

where $x_t^{\mathcal{G},i}$ and $u_t^{\mathcal{G},i}$ are the locally perceived influence of the network on the state of agent i and are given by

$$x_t^{\mathcal{G},i} = \sum_{j \in N} m^{ij} x_t^j \quad \text{and} \quad u_t^{\mathcal{G},i} = \sum_{j \in N} m^{ij} u_t^j. \tag{2}$$

A, B, D, and E are matrices of appropriate dimensions, and $\{w_t^i\}_{t \geq 1}, i \in N$, are independent and identically distributed (i.i.d.) zero-mean Gaussian processes which are independent of each other and the initial state. In particular, $w_t^i \in \mathbb{R}^{d_x}$ and $w_t^i \sim \mathcal{N}(0, W)$. We call $x_t^{\mathcal{G},i}$ and $u_t^{\mathcal{G},i}$ the *network-field* of the states and control actions at node i at time t.

Thus, the next state of agent i depends on its current local state and control action, the current network-field of the states and control actions of the system, and the current local noise.

We follow the same atypical representation of the "vectorized" dynamics as used in [22]. Define x_t and u_t as the global state and control actions of the system

$$x_t = [x_t^1, \dots, x_t^n] \quad \text{and} \quad u_t = [u_t^1, \dots, u_t^n].$$

We also define $w_t = [w_t^1, \dots, w_t^n]$. Similarly, define $x_t^{\mathcal{G}}$ and $u_t^{\mathcal{G}}$ as the global network field of states and actions

$$x_t^{\mathcal{G}} = [x_t^{\mathcal{G},1}, \dots, x_t^{\mathcal{G},n}] \quad \text{and} \quad u_t^{\mathcal{G}} = [u_t^{\mathcal{G},1}, \dots, u_t^{\mathcal{G},n}].$$

Note that $x_t, x_t^{\mathcal{G}}, w_t \in \mathbb{R}^{d_x \times n}$ and $u_t, u_t^{\mathcal{G}} \in \mathbb{R}^{d_u \times n}$ are matrices and not vectors. The global system dynamics may be written as

$$x_{t+1} = Ax_t + Bu_t + Dx_t^{\mathcal{G}} + Eu_t^{\mathcal{G}} + w_t. \tag{3}$$

Furthermore, we may write

$$x_t^{\mathcal{G}} = x_t M^{\mathsf{T}} = x_t M \quad \text{and} \quad u_t^{\mathcal{G}} = u_t M^{\mathsf{T}} = u_t M.$$

3) Per-Step Cost: At any time t, the system incurs a per-step cost given by

$$c(x_t, u_t) = \sum_{i \in N} \sum_{j \in N} [h_x^{ij} (x_t^i)^{\mathsf{T}} Q (x_t^j) + h_u^{ij} (u_t^i)^{\mathsf{T}} R (u_t^j)] \tag{4}$$

where Q and R are matrices of appropriate dimensions and h_x^{ij} and h_u^{ij} are real valued weights. Let $H_x = [h_x^{ij}]$ and $H_u = [h_u^{ij}]$. It is assumed that the weight matrices H_x and H_u are polynomials of M, i.e.,

$$H_x = \sum_{k=0}^{K_x} q_k M^k \quad \text{and} \quad H_u = \sum_{k=0}^{K_u} r_k M^k \tag{5}$$

where K_x and K_u denote the degrees of the polynomials and $\{q_k\}_{k=0}^{K_x}$ and $\{r_k\}_{k=0}^{K_u}$ are real-valued coefficients.

The assumption that H_x and H_u are polynomials of M captures the intuition that the per-step cost respects the graph structure. In the special case when $H_x = H_u = I$, the per-step cost is decoupled across agents. When $H_x = H_u = I + M$, the per-step cost captures a cross-coupling between one-hop neighbors. Similarly, when $H_u = I + M + M^2$, the per-step cost captures a cross-coupling between one- and two-hop neighbors. See [22] for more examples of special cases of the per-step cost defined above.

B. Assumptions on the Model

Since M is real and symmetric, it has real eigenvalues. Let L denote the rank of M and $\lambda^{(1)}, \dots, \lambda^{(L)}$ denote the nonzero eigenvalues. For ease of notation, for $\ell \in \{1, \dots, L\}$, define

$$q^{(\ell)} = \sum_{k=0}^{K_x} q_k (\lambda^{(\ell)})^k \quad \text{and} \quad r^{(\ell)} = \sum_{k=0}^{K_u} r_k (\lambda^{(\ell)})^k$$

where $\{q_k\}_{k=0}^{K_x}$ and $\{r_k\}_{k=0}^{K_u}$ are the coefficients in (5). Furthermore, for $\ell \in \{1, \dots, L\}$, define

$$A^{(\ell)} = A + \lambda^{(\ell)} D \quad \text{and} \quad B^{(\ell)} = B + \lambda^{(\ell)} E.$$

We impose the following assumptions.

(A1) The systems (A, B) and $\{(A^{(\ell)}, B^{(\ell)})\}_{\ell=1}^{L}$ are stabilizable.

(A2) The matrices Q and R are symmetric and positive definite.

(A3) The parameters q_0, r_0, $\{q^{(\ell)}\}_{\ell=1}^{L}$, and $\{r^{(\ell)}\}_{\ell=1}^{L}$ are strictly positive.

Assumption (A1) is needed to ensure that the average cost under the optimal policy is bounded. Assumptions (A2) and (A3) ensure that the per-step cost is strictly positive.

C. Admissible Policies and Performance Criterion

There is a system operator who has access to the state and action histories of all agents and who selects the agents' control actions according to a deterministic or randomized (and potentially history-dependent) policy $u_t = \pi_t(x_{1:t}, u_{1:t-1})$.

Let $\theta^{\mathsf{T}} = [A, B, D, E]$ denote the parameters of the system dynamics. The performance of any policy $\pi = (\pi_1, \pi_2, \dots)$ is measured by the long-term average cost given by

$$J(\pi; \theta) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} c(x_t, u_t) \right]. \tag{6}$$

Let $J(\theta)$ denote the minimum of $J(\pi; \theta)$ over all policies.

We are interested in the setup where the graph coupling matrix M, the cost coupling matrices H_x and H_u, and the cost matrices Q and R are known but the system dynamics θ are unknown and there is a prior distribution on θ. The Bayesian *regret* of a policy π operating for a horizon T is defined as

$$R(T; \pi) := \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} c(x_t, u_t) - TJ(\theta) \right] \tag{7}$$

where the expectation is with respect to the prior on θ, the noise processes, the initial conditions, and the potential randomizations done by the policy π.

III. Background on Spectral Decomposition of the System

In this section, we summarize the main results of [22], translated to the discrete-time model used in this article.

The spectral decomposition described in [22] relies on the spectral factorization of the graph coupling matrix M. Since M is a real symmetric matrix with rank L, we can write it as

$$M = \sum_{\ell=1}^{L} \lambda^{(\ell)} v^{(\ell)} (v^{(\ell)})^{\mathsf{T}} \tag{8}$$

where $(\lambda^{(1)}, \dots, \lambda^{(L)})$ are the nonzero eigenvalues of M and $(v^{(1)}, \dots, v^{(L)})$ are the corresponding eigenvectors.

We now present the decomposition of the dynamics and the cost based on (8) as described in [22].

A. Spectral Decomposition of the Dynamics and Per-Step Cost

For $\ell \in \{1, \dots, L\}$, define *eigenstates* and *eigencontrols* as

$$x_t^{(\ell)} = x_t v^{(\ell)} (v^{(\ell)})^{\mathsf{T}} \quad \text{and} \quad u_t^{(\ell)} = u_t v^{(\ell)} (v^{(\ell)})^{\mathsf{T}} \tag{9}$$

respectively. Furthermore, define *auxiliary state* and *auxiliary control* as

$$\breve{x}_t = x_t - \sum_{\ell=1}^{L} x_t^{(\ell)} \quad \text{and} \quad \breve{u}_t = u_t - \sum_{\ell=1}^{L} u_t^{(\ell)} \tag{10}$$

respectively. Similarly, define $w_t^{(\ell)} = w_t v^{(\ell)} (v^{(\ell)})^{\mathsf{T}}$ and $\breve{w}_t = w_t - \sum_{\ell=1}^{L} w_t^{(\ell)}$.

We now obtain the dynamics of the eigenstates and auxiliary states. Multiplying (3) on the right by $v^{(\ell)}(v^{(\ell)})^{\mathsf{T}}$ and observing that $v^{(\ell)}$ is an eigenvector of M, we get

$$x_{t+1}^{(\ell)} = (A + \lambda^{(\ell)} D) x_t^{(\ell)} + (B + \lambda^{(\ell)} E) u_t^{(\ell),i} + w_t^{(\ell)}. \tag{11}$$

Substituting (3) and (11) in (10), we get

$$\breve{x}_{t+1} = A \breve{x}_t + B \breve{u}_t + \breve{w}_t. \tag{12}$$

Let $x_t^{(\ell),i}$ and $u_t^{(\ell),i}$ denote the ith column of $x_t^{(\ell)}$ and $u_t^{(\ell)}$, respectively; thus, we can write

$$x_t^{(\ell)} = [x_t^{(\ell),1}, \dots, x_t^{(\ell),n}] \quad \text{and} \quad u_t^{(\ell)} = [u_t^{(\ell),1}, \dots, u_t^{(\ell),n}].$$

Similar interpretations hold for $w_t^{(\ell),i}$ and $\breve{w}_t^i$.

Looking at a particular column of (10) and rearranging terms, we can decompose the state and control action at each node $i \in N$ as $x_t^i = \breve{x}_t^i + \sum_{\ell=1}^{L} x_t^{(\ell),i}$ and $u_t^i = \breve{u}_t^i + \sum_{\ell=1}^{L} u_t^{(\ell),i}$. Equation (11) implies that the dynamics of eigenstate $x_t^{(\ell),i}$ depend only on $u_t^{(\ell),i}$ and $w_t^{(\ell),i}$ and are given by

$$x_{t+1}^{(\ell),i} = (A + \lambda^{(\ell)} D) x_t^{(\ell),i} + (B + \lambda^{(\ell)} E) u_t^{(\ell),i} + w_t^{(\ell),i}. \tag{13}$$

Similarly, (12) implies that the dynamics of the auxiliary state $\breve{x}_t^i$ depend only on $\breve{u}_t^i$ and $\breve{w}_t^i$ and are given by

$$\breve{x}_{t+1}^i = A \breve{x}_t^i + B \breve{u}_t^i + \breve{w}_t^i. \tag{14}$$

Furthermore, [22, Proposition 2] implies that per-step cost decomposes as follows:

$$c(x_t, u_t) = \sum_{i \in N} \left[q_0 \breve{c}(\breve{x}_t^i, \breve{u}_t^i) + \sum_{\ell=1}^{L} q^{(\ell)} c^{(\ell)}(x_t^{(\ell),i}, u_t^{(\ell),i}) \right] \tag{15}$$

where[1]

$$\breve{c}(\breve{x}_t^i, \breve{u}_t^i) = (\breve{x}_t^i)^{\mathsf{T}} Q \breve{x}_t^i + \frac{r_0}{q_0} (\breve{u}_t^i)^{\mathsf{T}} R \breve{u}_t^i,$$

$$c^{(\ell)}(x_t^{(\ell),i}, u_t^{(\ell),i}) = (x_t^{(\ell),i})^{\mathsf{T}} Q x_t^{(\ell),i} + \frac{r^{(\ell)}}{q^{(\ell)}} (u_t^{(\ell),i})^{\mathsf{T}} R u_t^{(\ell),i}.$$

Following [22, Lemma 2], we can show that for any $i \in N$

$$\operatorname{var}(w_t^{(\ell),i}) = (v^{(\ell),i})^2 W \quad \text{and} \quad \operatorname{var}(\breve{w}_t^i) = (\breve{v}^i)^2 W \tag{16}$$

where $(\breve{v}^i)^2 = 1 - \sum_{\ell=1}^{L} (v^{(\ell),i})^2$. These covariances do not depend on time because the noise processes are i.i.d.

B. Planning Solution for Network-Coupled Subsystems

We now present the main result of [22], which provides a scalable method to synthesize the optimal control policy when the system dynamics are known.

Based on the decomposition presented in the previous section, we can view the overall system as the collection of the following subsystems:

1) *Eigensystem* (ℓ, i), $\ell \in \{1, \dots, L\}$ and $i \in N$ with state $x_t^{(\ell),i}$, controls $u_t^{(\ell),i}$, dynamics (13), and per-step cost $q^{(\ell)} c^{(\ell)}(x^{(\ell),i}, u^{(\ell),i})$;
2) *Auxiliary system* i, $i \in N$, with state $\breve{x}_t^i$, controls $\breve{u}_t^i$, dynamics (14), and per-step cost $q_0 \breve{c}(\breve{x}_t^i, \breve{u}_t^i)$.

Let $(\theta^{(\ell)})^{\mathsf{T}} = [A^{(\ell)}, B^{(\ell)}] := [(A + \lambda^{(\ell)} D), (B + \lambda^{(\ell)} E)]$, $\ell \in \{1, \dots, L\}$, and $\breve{\theta}^{\mathsf{T}} = [A, B]$ denote the parameters of the dynamics of the eigensystems and auxiliary systems, respectively. Then, for any policy $\pi = (\pi_1, \pi_2, \dots)$, the performance of the eigensystem (ℓ, i), $\ell \in \{1, \dots, L\}$ and $i \in N$, is given by $q^{(\ell)} J^{(\ell),i}(\pi; \theta^{(\ell)})$, where

$$J^{(\ell),i}(\pi; \theta^{(\ell)}) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} c(x_t^{(\ell),i}, u_t^{(\ell),i}) \right].$$

Similarly, the performance of the auxiliary system i, $i \in N$, is given by $q_0 \breve{J}^i(\pi; \breve{\theta})$, where

$$\breve{J}^i(\pi; \breve{\theta}) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} c(\breve{x}_t^i, \breve{u}_t^i) \right].$$

Equation (15) implies that the overall performance of policy π can be decomposed as

$$J(\pi; \theta) = \sum_{i \in N} q_0 \breve{J}^i(\pi; \breve{\theta}) + \sum_{i \in N} \sum_{\ell=1}^{L} q^{(\ell)} J^{(\ell),i}(\pi; \theta^{(\ell)}). \tag{17}$$

The key intuition behind the result of [22] is as follows. By the CE principle for LQ systems, we know that (when the system dynamics are known) the optimal control policy

[1] Recall that (A3) ensures that q_0 and $\{q^{(\ell)}\}_{\ell=1}^{L}$ are strictly positive.

of a stochastic LQ system is the same as the optimal control policy of the corresponding deterministic LQ system where the noises $\{w_t^i\}_{t\geq 1}$ are assumed to be zero. Note that when noises $\{w_t^i\}_{t\geq 1}$ are zero, then the noises $\{w_t^{(\ell),i}\}_{t\geq 1}$ and $\{\breve w_t^i\}_{t\geq 1}$ of the eigensystems and auxiliary systems are also zero. This, in turn, implies that the dynamics of all the eigensystems and auxiliary systems are decoupled. These decoupled dynamics along with the cost decoupling in (17) imply that we can choose the controls $\{u_t^{(\ell),i}\}_{t\geq 1}$ for the eigensystem $((\ell), i)$, $\ell \in \{1,\dots,L\}$ and $i\in N$, to minimize[2] $J^{(\ell),i}(\pi;\theta^{(\ell)})$ and choose the controls $\{\breve u_t^i\}_{t\geq 1}$ for the auxiliary system i, $i\in N$, to minimize3 $\breve J^i(\pi;\breve\theta)$. These optimization problems are standard optimal control problems. Therefore, similar to [22, Theorem 3], we obtain the following result.

Theorem 1: Let $\breve S$ and $\{S^{(\ell)}\}_{\ell=1}^L$ be the solution of the following DARE:

$$\breve S(\breve\theta) = \mathrm{DARE}(A, B, Q, \tfrac{r_0}{q_0}R) \tag{18a}$$

$$\text{and for } \ell \in \{1,\dots,\mathrm{L}\}\mathrm{S}^{(\ell)}(\theta^{(\ell)}) = \mathrm{DARE}(\mathrm{A}^{(\ell)}, \mathrm{B}^{(\ell)}, Q, \tfrac{r^{(\ell)}}{q^{(\ell)}}R). \tag{18b}$$

Define the gains

$$\breve G(\breve\theta) = -\big((B)^\intercal \breve S(\breve\theta)B + \tfrac{r_0}{q_0}R\big)^{-1}(B)^\intercal \breve S(\breve\theta)A \tag{19a}$$

and for $\ell\in\{1,\dots,L\}$

$$G^{(\ell)}(\theta^{(\ell)}) = -\big((B^{(\ell)})^\intercal S^{(\ell)}(\theta^{(\ell)})B^{(\ell)} + \tfrac{r^{(\ell)}}{q^{(\ell)}}R\big)^{-1}(B^{(\ell)})^\intercal S^{(\ell)}(\theta^{(\ell)})A^{(\ell)}. \tag{19b}$$

Then, under assumptions (A1)–(A3), the policy

$$u_t^i = \breve G(\breve\theta)\breve x_t^i + \sum_{\ell=1}^L G^{(\ell)}(\theta^{(\ell)})x_t^{(\ell),i} \tag{20}$$

minimizes the long-term average cost in (6) over all admissible policies. Furthermore, the optimal performance is given by

$$J(\theta) = \sum_{i\in N} q_0 \breve J^i(\breve\theta) + \sum_{i\in N}\sum_{\ell=1}^L q^{(\ell)} J^{(\ell),i}(\theta^{(\ell)}) \tag{21}$$

where $\breve J^i(\breve\theta) = (\breve v^i)^2\,\mathrm{Tr}(W\breve S)$ and for $\ell\in\{1,\dots,L\}$

$$J^{(\ell),i}(\theta^{(\ell)}) = (v^{(\ell),i})^2\,\mathrm{Tr}(WS^{(\ell)}). \tag{22}$$

IV. Learning for Network-Coupled Subsystems

For the ease of notation, we define $z_t^{(\ell),i} = \mathrm{vec}(x_t^{(\ell),i}, u_t^{(\ell),i})$ and $\breve z_t^i = \mathrm{vec}(\breve x_t^i, \breve u_t^i)$. Then, we can write the dynamics (13) and (14) of the eigensystems and the auxiliary systems as

$$x_{t+1}^{(\ell),i} = (\theta^{(\ell)})^\intercal z_t^{(\ell),i} + w_t^{(\ell),i} \qquad \forall i\in N\ \forall \ell\in\{1,\dots,L\}, \tag{23a}$$

$$\breve x_{t+1}^i = (\breve\theta)^\intercal \breve z_t^i + \breve w_t^i \qquad \forall i\in N. \tag{23b}$$

A. Simplifying Assumptions

We impose the following assumptions to simplify the description of the algorithm and the regret analysis.

(A4) The noise covariance W is a scaled identity matrix given by $\sigma_w^2 I$.

(A5) For each $i\in N$, $\breve v^i \neq 0$.

Assumption (A4) is commonly made in most of the literature on regret analysis of LQG systems. An implication of (A4) is that $\mathrm{var}(\breve w_t^i) = (\breve\sigma^i)^2 I$ and $\mathrm{var}(w_t^{(\ell),i}) = (\sigma^{(\ell),i})^2 I$, where

$$(\breve\sigma^i)^2 = (\breve v^i)^2\sigma_w^2 \quad \text{and} \quad (\sigma^{(\ell),i})^2 = (v^{(\ell),i})^2\sigma_w^2. \tag{24}$$

Assumption (A5) is made to rule out the case where the dynamics of some of the auxiliary systems are deterministic.

B. Prior and Posterior Beliefs

We assume that the unknown parameters $\breve\theta$ and $\{\theta^{(\ell)}\}_{\ell=1}^L$ lie in compact subsets $\breve\Theta$ and $\{\Theta^{(\ell)}\}_{\ell=1}^L$ of $\mathbb R^{(d_x+d_u)\times d_x}$. Let $\breve\theta^k$ denote the kth column of $\breve\theta$. Thus, $\breve\theta = [\breve\theta^1,\dots,\breve\theta^{d_x}]$. Similarly, let $\theta^{(\ell),k}$ denote the kth column of $\theta^{(\ell)}$. Thus, $\theta^{(\ell)} = [\theta^{(\ell),1},\dots,\theta^{(\ell),d_x}]$. We use $p\big|_\Theta$ to denote the restriction of probability distribution p on the set Θ.

We assume that $\breve\theta$ and $\{\theta^{(\ell)}\}_{\ell=1}^L$ are random variables that are independent of the initial states and the noise processes. Furthermore, we assume that the priors $\breve p_1$ and $\{p_1^{(\ell)}\}_{\ell=1}^L$ on $\breve\theta$ and $\{\theta^{(\ell)}\}_{\ell=1}^L$, respectively, satisfy the following.

(A6) $\breve p_1$ is given as: $\breve p_1(\breve\theta) = \Big[\prod_{k=1}^{d_x}\breve\xi_1^k(\breve\theta^k)\Big]\Big|_{\breve\Theta}$ where, for $k\in\{1,\dots,d_x\}$, $\breve\xi_1^k = \mathcal N(\breve\mu_1^k, \breve\Sigma_1)$ with mean $\breve\mu_1^k\in\mathbb R^{d_x+d_u}$ and positive-definite covariance $\breve\Sigma_1\in\mathbb R^{(d_x+d_u)\times(d_x+d_u)}$.

(A7) For each $\ell\in\{1,\dots,L\}$, $p_1^{(\ell)}$ is given as:

$$p_1^{(\ell)}(\theta^{(\ell)}) = \left[\prod_{k=1}^{d_x}\xi_1^{(\ell),k}(\theta^{(\ell),k})\right]\Bigg|_{\Theta^{(\ell)}}$$

where for $k\in\{1,\dots,d_x\}$, $\xi_1^{(\ell),k} = \mathcal N(\mu_1^{(\ell),k}, \Sigma_1^{(\ell)})$ with mean $\mu_1^{(\ell),k}\in\mathbb R^{(d_x+d_u)}$ and positive-definite covariance $\Sigma_1^{(\ell)}\in\mathbb R^{(d_x+d_u)\times(d_x+d_u)}$.

These assumptions are similar to the assumptions on the prior in the recent literature on TS for LQ systems [20].

Our learning algorithm (and TS-based algorithms in general) keeps track of a posterior distribution on the unknown parameters based on observed data. Motivated by the nature of the planning solution (see Theorem 1), we maintain separate posterior distributions on $\breve\theta$ and $\{\theta^{(\ell)}\}_{\ell=1}^L$. For each ℓ, we select a subsystem $i_*^{(\ell)}$ such that the $i_*^{(\ell)}$th component of the eigenvector $v^{(\ell)}$ is nonzero (i.e., $v^{(\ell),i_*^{(\ell)}}\neq 0$). At time t, we maintain a posterior distribution $p_t^{(\ell)}$ on $\theta^{(\ell)}$ based on the corresponding eigenstate and action history of the $i_*^{(\ell)}$th subsystem. In other words, for any Borel subset B of $\mathbb R^{(d_x+d_u)\times d_x}$, $p_t^{(\ell)}(B)$ gives the following conditional probability:

$$p_t^{(\ell)}(B) = \mathbb P(\theta^{(\ell)}\in B \mid x_{1:t}^{(\ell),i_*^{(\ell)}}, u_{1:t-1}^{(\ell),i_*^{(\ell)}}). \tag{25}$$

We maintain a separate posterior distribution $\breve p_t$ on $\breve\theta$ as follows. At each time $t>1$, we select a subsystem $j_{t-1} = \arg\max_{i\in N} \breve z_{t-1}^{i\intercal}\breve\Sigma_{t-1}\breve z_{t-1}^i/(\breve\sigma_t^i)^2$, where $\breve\Sigma_{t-1}$ is a covariance

[2] The cost of the eigensystem $((\ell), i)$ is $q^{(\ell)}J^{(\ell),i}(\pi;\theta^{(\ell)})$. From (A3), we know that $q^{(\ell)}$ is positive. Therefore, minimizing $q^{(\ell)}J^{(\ell),i}(\pi;\theta^{(\ell)})$ is the same as minimizing $J^{(\ell),i}(\pi;\theta^{(\ell)})$.

matrix defined recursively in Lemma 1. Then, for any Borel subset B of $\mathbb{R}^{(d_x+d_u)\times d_x}$

$$\breve{p}_t(B) = \mathbb{P}(\breve{\theta} \in B \mid \{\breve{x}_s^{j_s}, \breve{u}_s^{j_s}, \breve{x}_{s+1}^{j_s}\}_{1\le s<t}). \tag{26}$$

See [38] for a discussion on the rule to select j_{t-1}.

Lemma 1: The posterior distributions $p_t^{(\ell)}$, $\ell \in \{1, 2, \dots, L\}$, and $\breve{p}_t$ are given as follows.

1) $p_1^{(\ell)}$ is given by Assumption (A7) and for any $t \ge 1$

$$p_{t+1}^{(\ell)}(\theta^{(\ell)}) = \left[\prod_{k=1}^{d_x} \xi_{t+1}^{(\ell),k}(\theta^{(\ell),k})\right]\Bigg|_{\Theta^{(\ell)}}$$

where for $k \in \{1, \dots, d_x\}$, $\xi_{t+1}^{(\ell),k} = \mathcal{N}(\mu_{t+1}^{(\ell),k}, \Sigma_{t+1}^{(\ell)})$ and

$$\mu_{t+1}^{(\ell)} = \mu_t^{(\ell)} + \frac{\Sigma_t^{(\ell)} z_t^{(\ell),i_*^{(\ell)}} \left(x_{t+1}^{(\ell),i_*^{(\ell)}} - (\mu_t^{(\ell)})^\intercal z_t^{(\ell),i_*^{(\ell)}}\right)^\intercal}{(\sigma^{(\ell),i_*^{(\ell)}})^2 + (z_t^{(\ell),i_*^{(\ell)}})^\intercal \Sigma_t^{(\ell)} z_t^{(\ell),i_*^{(\ell)}}} \tag{27a}$$

$$(\Sigma_{t+1}^{(\ell)})^{-1} = (\Sigma_t^{(\ell)})^{-1} + \frac{1}{(\sigma^{(\ell),i_*^{(\ell)}})^2} z_t^{(\ell),i_*^{(\ell)}} (z_t^{(\ell),i_*^{(\ell)}})^\intercal \tag{27b}$$

where, for each t, μ_t^ℓ denotes the matrix $[\mu_t^{(\ell),1}, \dots, \mu_t^{(\ell),d_x}]$.

2) $\breve{p}_1$ is given by Assumption (A6) and for any $t \ge 1$

$$\breve{p}_{t+1}(\breve{\theta}) = \left[\prod_{k=1}^{d_x} \breve{\xi}_{t+1}^k(\breve{\theta}^k)\right]\Bigg|_{\breve{\Theta}}$$

where for $k \in \{1, \dots, d_x\}$, $\breve{\xi}_{t+1}^k = \mathcal{N}(\breve{\mu}_{t+1}^k, \breve{\Sigma}_{t+1})$, and

$$\breve{\mu}_{t+1} = \breve{\mu}_t + \frac{\breve{\Sigma}_t \breve{z}_t^{j_t} \left(\breve{x}_{t+1}^{j_t} - (\breve{\mu}_t)^\intercal \breve{z}_t^{j_t}\right)^\intercal}{(\breve{\sigma}^{j_t})^2 + (\breve{z}_t^{j_t})^\intercal \breve{\Sigma}_t \breve{z}_t^{j_t}} \tag{28a}$$

$$(\breve{\Sigma}_{t+1})^{-1} = (\breve{\Sigma}_t)^{-1} + \frac{1}{(\breve{\sigma}^{j_t})^2} \breve{z}_t^{j_t} (\breve{z}_t^{j_t})^\intercal \tag{28b}$$

where, for each t, $\breve{\mu}_t$ denotes the matrix $[\breve{\mu}_t^1, \dots, \breve{\mu}_t^{d_x}]$.

Proof: Note that the dynamics of $x_t^{(\ell),i_*^{(\ell)}}$ and $\breve{x}_t^i$ in (23) are linear and the noises $w_t^{(\ell),i_*^{(\ell)}}$ and $\breve{w}_t^i$ are Gaussian. Therefore, the result follows from standard results in Gaussian linear regression [39, Theorem 3]. ■

C. Thompson Sampling Algorithm

We propose a TS-based algorithm called `Net-TSDE` which is inspired by the `TSDE` (Thompson sampling with dynamic episodes) algorithm proposed in [20] and the structure of the optimal planning solution described in Section III-B. The TS part of our algorithm is modeled after the modification of `TSDE` presented in [40].

The `Net-TSDE` algorithm consists of a coordinator $\mathcal{C}$ and $|L|+1$ *actors:* an auxiliary actor $\breve{\mathcal{A}}$ and an eigen actor $\mathcal{A}^\ell$ for each $\ell \in \{1, 2, \dots, L\}$. These actors are described below and the whole algorithm is presented in Algorithm 1.

1) At each time, the coordinator $\mathcal{C}$ observes the current global state x_t, computes the eigenstates $\{x_t^{(\ell)}\}_{\ell=1}^L$ and the auxiliary states $\breve{x}_t$, and sends the eigenstate $x_t^{(\ell)}$ to the eigen actor $\mathcal{A}^{(\ell)}$, $\ell \in \{1, \dots, L\}$, and sends the auxiliary state $\breve{x}_t$ to the auxiliary actor $\breve{\mathcal{A}}$. The eigen actor $\mathcal{A}^{(\ell)}$, $\ell \in \{1, \dots, L\}$, computes the eigencontrol $u_t^{(\ell)}$ and the auxiliary actor $\breve{\mathcal{A}}$ computes the auxiliary control $\breve{u}_t$ (as per the details presented below) and both send their computed controls back to the coordinator $\mathcal{C}$. The coordinator then computes and executes the control action $u_t^i = \sum_{\ell=1}^L u_t^{(\ell),i} + \breve{u}_t^i$ for each subsystem $i \in N$.

Algorithm 1: `Net-TSDE`.

```
initialize eigen actor: $\Theta^{(\ell)}$, $(\mu_1^\ell, \Sigma_1^\ell)$, $t_0^\ell = -T_{\min}$,
    $T_{-1}^\ell = T_{\min}$, $k = 0$, $\theta_k^\ell = 0$
initialize auxiliary actor: $\breve{\Theta}$, $(\breve{\mu}_1, \breve{\Sigma}_1)$, $\breve{t}_0 = -T_{\min}$,
    $\breve{T}_{-1} = T_{\min}$, $k = 0$, $\breve{\theta}_k = 0$.
for $t = 1, 2, \dots$ do
    observe $x_t$
    compute $\{x_t^{(\ell)}\}_{\ell=1}^L$ and $\breve{x}_t$ using (9) and (10).
    for $\ell = 1, 2, \dots, L$ do
        $u_t^{(\ell)} \leftarrow$ EIGEN-ACTOR$(x_t^{(\ell)})$
    $\breve{u}_t \leftarrow$ AUXILIARY-ACTOR$(\breve{x}_t)$
    for $i \in N$ do
        Subsystem $i$ applies control $u_t^i = u_t^{(\ell),i} + \breve{u}_t^i$

function eigen-actor $x_t^{(\ell)}$
    global var $t$
    Update $p_t^{(\ell)}$ according to (27)
    if $(t - t_k^{(\ell)} > T_{\min})$ and
        $((t - t_k^{(\ell)} > T_{k-1}^{(\ell)})$ or $(\det \Sigma_t^{(\ell)} < \frac{1}{2} \det \Sigma_{t_k^{(\ell)}}))$
    then
        $T_k^{(\ell)} \leftarrow t - t_k^{(\ell)}$, $k \leftarrow k + 1$, $t_k^{(\ell)} \leftarrow t$
        sample $\theta_k^{(\ell)} \sim p_t^{(\ell)}$
    return $G^{(\ell)}(\theta_k^{(\ell)}) x_t^{(\ell)}$

function auxiliary-actor $\breve{x}_t$
    global var $t$
    Update $\breve{p}_t$ according to (28)
    if $(t - \breve{t}_k > T_{\min})$ and
        $((t - \breve{t}_k > \breve{T}_{k-1})$ or $(\det \breve{\Sigma}_t < \frac{1}{2} \det \breve{\Sigma}_{t_k^{(\ell)}}))$
    then
        $\breve{T}_k \leftarrow t - \breve{t}_k$, $k \leftarrow k + 1$, $\breve{t}_k \leftarrow t$
        sample $\breve{\theta}_k \sim \breve{p}_t$
    return $\breve{G}(\breve{\theta}_k) \breve{x}_t$
```

2) The eigen actor $\mathcal{A}^{(\ell)}$, $\ell \in \{1, \dots, L\}$, maintains the posterior $p_t^{(\ell)}$ on $\theta^{(\ell)}$ according to (27). The actor works in episodes of dynamic length. Let $t_k^{(\ell)}$ and $T_k^{(\ell)}$ denote the starting time and the length of episode k, respectively. Each episode is of a minimum length $T_{\min}^{(\ell)} + 1$, where $T_{\min}^{(\ell)}$ is chosen as described in [40]. Episode k ends if the determinant of covariance $\Sigma_t^{(\ell)}$ falls below half of its value at time $t_k^{(\ell)}$ (i.e., $\det(\Sigma_t^{(\ell)}) < \frac{1}{2} \det \Sigma_{t_k^{(\ell)}}$) or if the

length of the episode is one more than the length of the previous episode (i.e., $t - t_k^{(\ell)} > T_{k-1}^{(\ell)}$). Thus

$$t_{k+1}^{(\ell)} = \min\Big\{ t > t_k^{(\ell)} + T_{\min}^{(\ell)} \,\Big|\, t - t_k^{(\ell)} > T_{k-1}^{(\ell)} \text{ or } \det \Sigma_t^{(\ell)} < \tfrac{1}{2} \det \Sigma_{t_k^{(\ell)}} \Big\}.$$

At the beginning of episode k, the eigen actor $\mathcal{A}^{(\ell)}$ samples a parameter $\theta_k^{(\ell)}$ according to the posterior distribution $p_{t_k^{(\ell)}}^{(\ell)}$. During episode k, the eigen actor $\mathcal{A}^{(\ell)}$ generates the eigencontrols using the sampled parameter $\theta_k^{(\ell)}$, i.e., $u_t^{(\ell)} = G^{(\ell)}(\theta_k^{(\ell)}) x_t^{(\ell)}$.

3) The auxiliary actor $\breve{\mathcal{A}}$ is similar to the eigen actor. Actor $\breve{\mathcal{A}}$ maintains the posterior $\breve{p}_t$ on $\breve{\theta}$ according to (28). The actor works in episodes of dynamic length. The episodes of the auxiliary actor $\breve{\mathcal{A}}$ and the eigen actors $\mathcal{A}^{(\ell)}$, $\ell \in \{1, 2, \ldots, L\}$, are separate from each other.[3] Let $\breve{t}_k$ and $\breve{T}_k$ denote the starting time and the length of episode k, respectively. Each episode is of a minimum length $\breve{T}_{\min} + 1$, where $\breve{T}_{\min}$ is chosen as described in [40]. The termination condition for each episode is similar to that of the eigen actor $\mathcal{A}^{(\ell)}$. In particular

$$\breve{t}_{k+1} = \min\Big\{ t > \breve{t}_k + \breve{T}_{\min} \,\Big|\, t - \breve{t}_k > \breve{T}_{k-1} \text{ or } \det \breve{\Sigma}_t < \tfrac{1}{2} \det \breve{\Sigma}_{\breve{t}_k} \Big\}.$$

At the beginning of episode k, the auxiliary actor $\breve{\mathcal{A}}$ samples a parameter $\breve{\theta}_k$ from the posterior distribution $\breve{p}_{\breve{t}_k}$. During episode k, the auxiliary actor $\breve{\mathcal{A}}$ generates the auxiliary controls using the the sampled parameter $\breve{\theta}_k$, i.e., $\breve{u}_t = \breve{G}(\breve{\theta}_k)\breve{x}_t$.

Note that the algorithm does not depend on the horizon T.

D. Regret Bounds

We impose the following assumption to ensure that the closed-loop dynamics of the eigenstates and the auxiliary states of each subsystem are stable.

(A8) There exists a positive number $\delta \in (0, 1)$ such that

1) for any $\ell \in \{1, 2, \ldots, L\}$ and $\theta^{(\ell)}, \phi^{(\ell)} \in \Theta^{(\ell)}$ where $(\theta^{(\ell)})^{\mathsf{T}} = [A_{\theta^{(\ell)}}^{(\ell)}, B_{\theta^{(\ell)}}^{(\ell)}]$, we have

$$\rho(A_{\theta^{(\ell)}}^{(\ell)} + B_{\theta^{(\ell)}}^{\ell} G^{(\ell)}(\phi^{(\ell)})) \leq \delta.$$

2) for any $\breve{\theta}, \breve{\phi} \in \breve{\Theta}$, where $(\breve{\theta})^{\mathsf{T}} = [A_{\breve{\theta}}, B_{\breve{\theta}}]$, we have

$$\rho(A_{\breve{\theta}} + B_{\breve{\theta}} \breve{G}(\breve{\phi})) \leq \delta.$$

This assumption is similar to an assumption made in [40] for TS for LQG systems. According to [41, Lemma 1] (also see [18, Theorem 11]), (A8) is satisfied if

$$\Theta^{(\ell)} = \{\theta^{(\ell)} \in \mathbb{R}^{(d_x+d_u)\times d_x} : \|\theta^{(\ell)} - \theta_\circ^{\ell}\| \leq \varepsilon^{(\ell)}\},$$
$$\breve{\Theta} = \{\breve{\theta} \in \mathbb{R}^{(d_x+d_u)\times d_x} : \|\breve{\theta} - \breve{\theta}_\circ\| \leq \breve{\varepsilon}\}$$

where $\theta^{(\ell)}$ and $\breve{\theta}$ are stabilizable and $\varepsilon^{(\ell)}$ and $\breve{\varepsilon}$ are sufficiently small. In other words, the assumption holds when the true system is in a small neighborhood of a known nominal system. Such a small neighborhood can be learned with high probability by running appropriate stabilizing procedures for finite time [18], [41].

The following result provides an upper bound on the regret of the proposed algorithm.

Theorem 2: Under (A1)–(A8), the regret of `Net-TSDE` is upper bounded as follows:

$$R(T; \texttt{Net-TSDE}) \leq \tilde{\mathcal{O}}\big(\alpha^{\mathcal{G}} \sigma_w^2 d_x^{0.5} (d_x + d_u)\sqrt{T}\big)$$

where $\alpha^{\mathcal{G}} = \sum_{\ell=1}^{L} q^{(\ell)} + q_0(n - L)$.

See Section V for proof.

Remark 1: The term $\alpha^{\mathcal{G}}$ in the regret bound partially captures the impact of the network on the regret. The coefficients r_0 and $\{r^{(\ell)}\}_{\ell=1}^{L}$ depend on the network and also affect the regret, but their dependence is hidden inside the $\tilde{\mathcal{O}}(\cdot)$ notation. It is possible to explicitly characterize this dependence, but doing so does not provide any additional insights. We discuss the impact of the network coupling on the regret in Section VI via some examples.

Remark 2: The regret per subsystem is given by $R(T; \texttt{Net-TSDE})/n$, which is proportional to

$$\alpha^{\mathcal{G}}/n = \mathcal{O}\Big(\frac{L}{n}\Big) + \mathcal{O}\Big(\frac{n-1}{n}\Big) = \mathcal{O}\Big(1 + \frac{L}{n}\Big).$$

Thus, the regret per-subsystem scales as $\mathcal{O}(1 + L/n)$. In contrast, for the standard `TSDE` algorithm [20], [40], the regret per subsystem is proportional to $\alpha^{\mathcal{G}}(\texttt{TSDE})/n = \mathcal{O}(n^{0.5})$. This clearly illustrates the benefit of the proposed learning algorithm.

V. Regret Analysis

For the ease of notation, we simply use $R(T)$ instead of $R(T; \texttt{Net-TSDE})$ in this section. Based on (15) and Theorem 1, the regret may be decomposed as

$$R(T) = \sum_{i \in N} q_0 \breve{R}^i(T) + \sum_{i \in N} \sum_{\ell=1}^{L} q^{(\ell)} R^{(\ell),i}(T) \tag{29}$$

where

$$\breve{R}^i(T) := \mathbb{E}\Big[\sum_{t=1}^{T} \breve{c}(\breve{x}_t^i, \breve{u}_t^i) - T\breve{J}^i(\breve{\theta})\Big]$$

and, for $\ell \in \{1, \ldots, L\}$

$$R^{(\ell),i}(T) := \mathbb{E}\Big[\sum_{t=1}^{T} c^{(\ell)}(x_t^{(\ell),i}, u_t^{(\ell),i}) - TJ^{(\ell),i}(\theta^{(\ell)})\Big].$$

Based on the discussion at the beginning of Section III-B, $q_0\breve{R}^i(T)$, $i \in N$, is the regret associated with auxiliary system i and $q^{(\ell)} R^{(\ell),i}(T)$, $\ell \in \{1, \ldots, L\}$ and $i \in N$, is the regret associated with eigensystem (ℓ, i). We now bound $\breve{R}^i(T)$ and $R^{(\ell),i}(T)$ separately.

[3] The episode count k is used as a local variable for each actor.

A. Bound on $R^{(\ell),i}(T)$

Fix $\ell \in \{1,\dots,L\}$. For the component $i_*^{(\ell)}$, the `Net-TSDE` algorithm is exactly the same as the variation of the `TSDE` algorithm of [20] presented in [40]. Therefore, from [40, Theorem 1], it follows that

$$R^{(\ell),i_*^{(\ell)}}(T) \leq \tilde{\mathcal{O}}\big((\sigma^{(\ell),i_*^{(\ell)}})^2 d_x^{0.5}(d_x+d_u)\sqrt{T}\big). \tag{30}$$

We now show that the regret of other eigensystems (ℓ,i) with $i \neq i_*^{(\ell)}$ also satisfies a similar bound.

Lemma 2: The regret of eigensystem (ℓ,i), $\ell \in \{1,\dots,L\}$ and $i \in N$, is bounded as follows:

$$R^{(\ell),i}(T) \leq \tilde{\mathcal{O}}\big((\sigma^{(\ell),i})^2 d_x^{0.5}(d_x+d_u)\sqrt{T}\big). \tag{31}$$

Proof: Fix $\ell \in \{1,\dots,L\}$. Recall from (9) that $x_t^{(\ell)} = x_t v^{(\ell)}(v^{(\ell)})^\mathsf{T}$. Therefore, for any $i \in N$

$$x_t^{(\ell),i} = x_t v^{(\ell)} v^{(\ell),i} = v^{(\ell),i} x_t v^{(\ell)}$$

where the last equality follows because $v^{(\ell),i}$ is a scalar. Since we are using the same gain $G^{(\ell)}(\theta_k^{(\ell)})$ for all agents $i \in N$, we have

$$u_t^{(\ell),i} = G^{(\ell)}(\theta_k^{(\ell)}) x_t^{(\ell),i} = v^{(\ell),i} G^{(\ell)}(\theta_k^{(\ell)}) x_t v^{(\ell)}.$$

Thus, we can write (recall that $i_*^{(\ell)}$ is chosen such that $v^{(\ell),i_*^{(\ell)}} \neq 0$), for all $i \in N$

$$x_t^{(\ell),i} = \left(\frac{v^{(\ell),i}}{v^{(\ell),i_*^{(\ell)}}}\right) x_t^{(\ell),i_*^{(\ell)}} \text{ and } u_t^{(\ell),i} = \left(\frac{v^{(\ell),i}}{v^{(\ell),i_*^{(\ell)}}}\right) u_t^{(\ell),i_*^{(\ell)}}.$$

Thus, for any $i \in N$

$$c^{(\ell)}\left(x_t^{(\ell),i}, u_t^{(\ell),i}\right) = \left(\frac{v^{(\ell),i}}{v^{(\ell),i_*^{(\ell)}}}\right)^2 c^{(\ell)}\left(x_t^{(\ell),i_*^{(\ell)}}, u_t^{(\ell),i_*^{(\ell)}}\right). \tag{32}$$

Moreover, from (22), we have

$$J^{(\ell),i}(\theta^{(\ell)}) = \left(\frac{v^{(\ell),i}}{v^{(\ell),i_*^{(\ell)}}}\right)^2 J^{(\ell),i_*^{(\ell)}}(\theta^{(\ell)}). \tag{33}$$

By combining (32) and (33), we get

$$R^{(\ell),i}(T) = \left(\frac{v^{(\ell),i}}{v^{(\ell),i_*^{(\ell)}}}\right)^2 R^{(\ell),i_*^{(\ell)}}(T).$$

Substitute the bound for $R^{(\ell),i_*^{(\ell)}}(T)$ from (30) and observe that $(v^{(\ell),i}/v^{(\ell),i_*^{(\ell)}})^2 = (\sigma^{(\ell),i}/\sigma^{(\ell),i_*^{(\ell)}})^2$ gives the result. ■

B. Bound on $\breve{R}^i(T)$

The update of the posterior $\breve{p}_t$ on $\breve{\theta}$ does not depend on the history of states and actions of any fixed agent i. Therefore, we cannot directly use the argument presented in [40] to bound the regret $\breve{R}^i(T)$. We present a bound from first principles below.

For the ease of notation, for any episode k, we use $\breve{G}_k$ and $\breve{S}_k$ to denote $\breve{G}(\breve{\theta}_k)$ and $\breve{S}(\breve{\theta}_k)$, respectively. From LQ optimal control theory [42], we know that the average cost $\breve{J}^i(\breve{\theta}_k)$ and the optimal policy $\breve{u}_t^i = \breve{G}_k \breve{x}_t^i$ for the model parameter $\breve{\theta}_k$ satisfy the following Bellman equation:

$$\begin{aligned}\breve{J}^i(\breve{\theta}_k) + (\breve{x}_t^i)^\mathsf{T} \breve{S}_k \breve{x}_t^i &= \breve{c}(\breve{x}_t^i, \breve{u}_t^i) \\ &+ \mathbb{E}\big[\big(\breve{\theta}_k^\mathsf{T} \breve{z}_t^i + \breve{w}_t^i\big)^\mathsf{T} \breve{S}_k \big(\breve{\theta}_k^\mathsf{T} \breve{z}_t^i + \breve{w}_t^i\big)\big].\end{aligned}$$

Adding and subtracting $\mathbb{E}[(\breve{x}_{t+1}^i)^\mathsf{T} \breve{S}_k \breve{x}_{t+1}^i \mid \breve{z}_t^i]$ and noting that $\breve{x}_{t+1}^i = \breve{\theta}^\mathsf{T} \breve{z}_t^i + \breve{w}_t^i$, we get that

$$\begin{aligned}\breve{c}(\breve{x}_t^i, \breve{u}_t^i) &= \breve{J}^i(\breve{\theta}_k) + (\breve{x}_t^i)^\mathsf{T} \breve{S}_k \breve{x}_t^i - \mathbb{E}[(\breve{x}_{t+1}^i)^\mathsf{T} \breve{S}_k \breve{x}_{t+1}^i | \breve{z}_t^i] \\ &+ (\breve{\theta}^\mathsf{T} \breve{z}_t^i)^\mathsf{T} \breve{S}_k ((\breve{\theta})^\mathsf{T} \breve{z}_t^i) - (\breve{\theta}_k^\mathsf{T} \breve{z}_t^i)^\mathsf{T} \breve{S}_k ((\breve{\theta}_k)^\mathsf{T} \breve{z}_t^i).\end{aligned} \tag{34}$$

Let $\breve{K}_T$ denote the number of episodes of the auxiliary actor until horizon T. For each $k > \breve{K}_T$, we define $\breve{t}_k$ to be $T+1$. Then, using (34), we have that for any agent i

$$\begin{aligned}\breve{R}^i(T) &= \underbrace{\mathbb{E}\left[\sum_{k=1}^{\breve{K}_T} \breve{T}_k \breve{J}^i(\breve{\theta}_k) - T\breve{J}^i(\breve{\theta})\right]}_{\text{regret due to sampling error} =: \breve{R}_0^i(T)} \\ &+ \underbrace{\mathbb{E}\left[\sum_{k=1}^{\breve{K}_T} \sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1} \left[(\breve{x}_t^i)^\mathsf{T} \breve{S}_k \breve{x}_t^i - (\breve{x}_{t+1}^i)^\mathsf{T} \breve{S}_k \breve{x}_{t+1}^i\right]\right]}_{\text{regret due to time-varying controller } =: \breve{R}_1^i(T)} \\ &+ \underbrace{\begin{aligned}[t]\mathbb{E}\Bigg[\sum_{k=1}^{\breve{K}_T} \sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1} &\left[(\breve{\theta}^\mathsf{T} \breve{z}_t^i)^\mathsf{T} \breve{S}_k ((\breve{\theta})^\mathsf{T} \breve{z}_t^i)\right. \\ &\left.-(\breve{\theta}_k^\mathsf{T} \breve{z}_t^i)^\mathsf{T} \breve{S}_k ((\breve{\theta}_k)^\mathsf{T} \breve{z}_t^i)\right]\Bigg].\end{aligned}}_{\text{regret due to model mismatch } =: \breve{R}_2^i(T)}\end{aligned} \tag{35}$$

Lemma 3: The terms in (35) are bounded as follows.

1) $\breve{R}_0^i(T) \leq \tilde{\mathcal{O}}((\breve{\sigma}^i)^2 (d_x+d_u)^{0.5}\sqrt{T})$.
2) $\breve{R}_1^i(T) \leq \tilde{\mathcal{O}}((\breve{\sigma}^i)^2 (d_x+d_u)^{0.5}\sqrt{T})$.
3) $\breve{R}_2^i(T) \leq \tilde{\mathcal{O}}((\breve{\sigma}^i)^2 d_x^{0.5}(d_x+d_u)\sqrt{T})$.

Combining these three, we get that

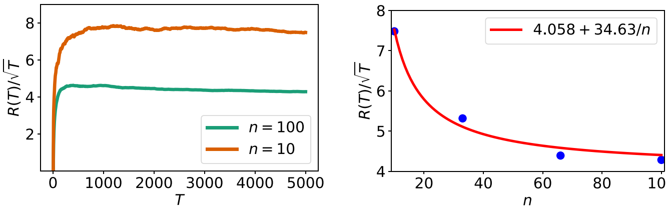
$$\breve{R}^i(T) \leq \tilde{\mathcal{O}}((\breve{\sigma}^i)^2 d_x^{0.5}(d_x+d_u)\sqrt{T}). \tag{36}$$

See Appendix for the proof.

C. Proof of Theorem 2

For ease of notation, let $R^* = \tilde{\mathcal{O}}(d_x^{0.5}(d_x+d_u)\sqrt{T})$. Then, by substituting the result of Lemmas 2 and 3 in (29), we get that

$$\begin{aligned}R(T) &\leq \sum_{i \in N} q_0 (\breve{\sigma}^i)^2 R^* + \sum_{i \in N} \sum_{\ell=1}^{L} q^{(\ell)} (\sigma^{(\ell),i})^2 R^* \\ &\overset{(a)}{=} \sum_{i \in N} q_0 (\breve{v}^i)^2 \sigma_w^2 R^* + \sum_{i \in N} \sum_{\ell=1}^{L} q^{(\ell)} (v^{(\ell),i})^2 \sigma_w^2 R^* \\ &\overset{(b)}{=} \Big(q_0(n-L) + \sum_{\ell=1}^{L} q^{(\ell)}\Big) \sigma_w^2 R^*\end{aligned} \tag{37}$$



(a) (b)

Fig. 1. Regret for mean-field system. (a) $R(T)/\sqrt{T}$ vs T. (b) $R(T)/\sqrt{T}$ vs n.

where (a) follows from (24) and (b) follows from observing that $\sum_{i \in N}(v^{(\ell),i})^2 = 1$ and therefore $\sum_{i \in N}(\breve{v}^i)^2 = n - L$. Equation (37) establishes the result of Theorem 2.

VI. Some Examples

A. Mean-Field System

Consider a complete graph $\mathcal{G}$ where the edge weights are equal to $1/n$. Let M be equal to the adjacency matrix of the graph, i.e., $M = \frac{1}{n}\mathbb{1}_{n \times n}$. Thus, the system dynamics are given by

$$x_{t+1}^i = Ax_t^i + Bu_t^i + D\bar{x}_t + E\bar{u}_t + w_t^i$$

where $\bar{x}_t = \frac{1}{n}\sum_{i \in N} x_t^i$ and $\bar{u}_t = \frac{1}{n}\sum_{i \in N} u_t^i$. Suppose $K_x = K_u = 1$ and $q_0 = r_0 = 1/n$ and $q_1 = r_1 = \kappa/n$, where κ is a positive constant.

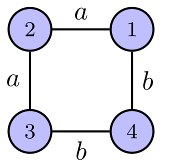
In this case, M has rank $L = 1$, the nonzero eigenvalue of M is $\lambda^{(1)} = 1$, the corresponding normalized eigenvector is $\frac{1}{\sqrt{n}}\mathbb{1}_{n \times 1}$, and $q^{(1)} = r^{(1)} = q_0 + q_1 = (1+\kappa)/n$. The eigenstate is given by $x_t^1 = [\bar{x}_t, \dots, \bar{x}_t]$ and a similar structure holds for the eigencontrol u_t^1. The per-step cost can be written as [see (15)]

$$c(x_t, u_t) = (1+\kappa)\big[\bar{x}_t^{\mathsf{T}} Q \bar{x}_t + \bar{u}_t^{\mathsf{T}} R \bar{u}_t\big]. \\ + \frac{1}{n}\sum_{i \in N}\big[(x_t^i - \bar{x}_t)^{\mathsf{T}} Q (x_t^i - \bar{x}_t) + (u_t^i - \bar{u}_t)^{\mathsf{T}} R (u_t^i - \bar{u}_t)\big].$$

Thus, the system is similar to the mean-field team system investigated in [6].

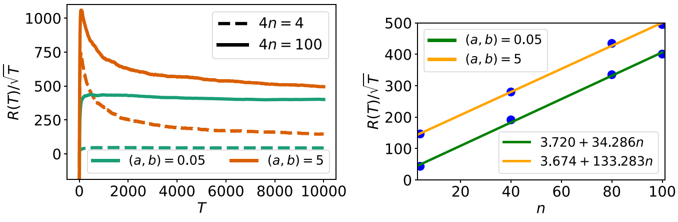
For this model, the network-dependent constant $\alpha^{\mathcal{G}}$ in the regret bound of Theorem 2 is given by $\alpha^{\mathcal{G}} = \big(1 + \frac{\kappa}{n}\big) = \mathcal{O}(1 + \frac{1}{n})$. Thus, for the mean-field system, the regret of Net-TSDE scales as $\mathcal{O}(1 + \frac{1}{n})$ with the number of agents. This is consistent with the discussion following Theorem 2.

We test these conclusions via numerical simulations of a scalar mean-field model with $d_x = d_u = 1$, $\sigma_w^2 = 1$, $A = 1$, $B = 0.3$, $D = 0.5$, $E = 0.2$, $Q = 1$, $R = 1$, and $\kappa = 0.5$. The uncertain sets are chosen as $\Theta^{(1)} = \{\theta^{(1)} \in \mathbb{R}^2 : A + D + (B + E)G^{(1)}(\theta^{(1)}) < \delta\}$ and $\breve{\Theta} = \{\breve{\theta} \in \mathbb{R}^2 : A + B\breve{G}(\breve{\theta}) < \delta\}$ where $\delta = 0.99$. The prior over these uncertain sets is chosen according to (A6)–(A7) where $\breve{\mu}_1 = \mu_1^{(1)} = [1, 1]^{\mathsf{T}}$ and $\breve{\Sigma}_1 = \Sigma_1^{(1)} = I$. We set $T_{\min} = 0$ in Net-TSDE. The system is simulated for a horizon of $T = 5000$ and the expected regret $R(T)$ averaged over 500 sample trajectories is shown in Fig. 1. As expected, the regret scales as $\tilde{\mathcal{O}}(\sqrt{T})$ with time and $\mathcal{O}\big(1 + \frac{1}{n}\big)$ with the number of agents.



$$\begin{bmatrix} 0 & a & 0 & b \\ a & 0 & a & 0 \\ 0 & a & 0 & b \\ b & 0 & b & 0 \end{bmatrix}$$

Fig. 2. Graph $\mathcal{G}^\circ$ with $n = 4$ nodes and its adjacency matrix.



(a) (b)

Fig. 3. Regret for general low-rank network. (a) $R(T)/\sqrt{T}$ vs T. (b) $R(T)/\sqrt{T}$ vs n.

B. General Low-Rank Network

We consider a network with $4n$ nodes given by the graph $\mathcal{G} = \mathcal{G}^\circ \otimes \mathcal{C}_n$, where $\mathcal{G}^\circ$ is a four-node graph shown in Fig. 2 and $\mathcal{C}_n$ is the complete graph with n nodes and each edge weight equal to $\frac{1}{n}$. Let M be the adjacency matrix of $\mathcal{G}$ which is given as $M = M^\circ \otimes \frac{1}{n}\mathbb{1}_{n \times n}$, where M° is the adjacency matrix of $\mathcal{G}^\circ$ shown in Fig. 2. Moreover, suppose $K_x = 2$ with $q_0 = 1$, $q_1 = -2$, and $q_2 = 1$ and $K_u = 0$ with $r_0 = 1$. Note that the cost is not normalized per-agent.

In this case, the rank of M° is 2 with eigenvalues $\pm\rho$, where $\rho = \sqrt{2(a^2 + b^2)}$ and the rank of $\frac{1}{n}\mathbb{1}_{n \times n}$ is 1 with eigenvalue 1. Thus, $M = M^\circ \otimes \frac{1}{n}\mathbb{1}_{n \times n}$ has the same nonzero eigenvalues as M° given by $\lambda^{(1)} = \rho$ and $\lambda^{(2)} = -\rho$. Further, $q^{(\ell)} = (1 - \lambda^{(\ell)})^2$ and $r^{(\ell)} = 1$, for $\ell \in \{1, 2\}$. We assume that $a^2 + b^2 \neq 0.5$, so that the model satisfies (A3).

For this model, the scaling parameter $\alpha^{\mathcal{G}}$ in the regret bound in Theorem 2 is given by

$$\alpha^{\mathcal{G}} = (1 - \rho)^2 + (1 + \rho)^2 + (4n - 2) = 4n + 2\rho^2.$$

Recall that $\rho^2 = (\lambda^{(1)})^2 = (\lambda^{(2)})^2$. Thus, $\alpha^{\mathcal{G}}$ has an explicit dependence on the square of the eigenvalues and the number of nodes.

We verify this relationship via numerical simulations. We consider the graph above with two choices of parameters (a, b): 1) $a = b = 0.05$ and 2) $a = b = 5$. For both cases, we consider a scalar system with parameters the same as the mean-field system considered in Section VI-A. The regret for both cases with different choices of number of agents $4n \in \{4, 40, 80, 100\}$ is shown in Fig. 3. As expected, the regret scales as $\tilde{\mathcal{O}}(\sqrt{T})$ with time and $\mathcal{O}\big(4n + 2\rho^2\big)$ with the number of agents.

VII. Conclusion

In this article, consider the problem of controlling an unknown LQG system consisting of multiple subsystems connected over a network. By utilizing a spectral decomposition technique, we decompose the coupled subsystems into eigensystems and auxiliary systems. We propose a TS-based learning algorithm `Net-TSDE` which maintains separate posterior distributions on the unknown parameters $\theta^{(\ell)}, \ell \in \{1, \ldots, L\}$, and $\breve{\theta}$ associated with the eigensystems and auxiliary systems, respectively. For each eigensystem, `Net-TSDE` learns the unknown parameter $\theta^{(\ell)}$ and controls the system in a manner similar to the `TSDE` algorithm for single agent LQG systems proposed in [20] and [40]. Consequently, the regret for each eigensystem can be bounded using the results of [20] and [40]. However, the part of the `Net-TSDE` algorithm that performs learning and control for the auxiliary system has an agent selection step and thus requires additional analysis to bound its regret. Combining the regret bounds for the eigensystems and auxiliary systems shows that the total expected regret of `Net-TSDE` is upper bounded by $\tilde{\mathcal{O}}(nd_x^{0.5}(d_x + d_u)\sqrt{T})$. The empirically observed scaling of regret with respect to the time horizon T and the number of subsystems n in our numerical experiments agrees with the theoretical upper bound.

The results presented in this article rely on the spectral decomposition developed in [22]. A limitation of this decomposition is that the local dynamics (i.e., the (A, B) matrices) are assumed to be identical for all subsystems and the coupling matrix M is symmetric. Interesting generalizations overcoming these limitations include settings where 1) there are multiple types of subsystems and the (A, B) matrices are the same for subsystems of the same type but different across types, 2) the coupling matrix M is not symmetric, and 3) the subsystems are not identical but approximately identical, i.e., there are nominal dynamics (A°, B°) and the local dynamics (A^i, B^i) of subsystem i are in a small neighborhood of (A°, B°). It may be possible to extend the decomposition in [22] and the learning algorithm of this article to handle cases 1) and 2). For case 3), it may be possible to approximate the nonidentical subsystems by identical subsystems. However, such an approximation may lead to a regret which is linear in time due to the approximation error.

The decomposition in [22] exploits the fact that the dynamics and the cost couplings have the same spectrum (i.e., the same orthonormal eigenvectors). It is also possible to consider learning algorithms which exploit other features of the network such as sparsity in the case of networked MDPs [32], [33].

Appendix
Regret Analysis

A. Preliminary Results

Since $\breve{S}(\cdot)$ and $\breve{G}(\cdot)$ are continuous functions on a compact set $\breve{\Theta}$, there exist finite constants $\breve{M}_J, \breve{M}_{\breve{\theta}}, \breve{M}_S, \breve{M}_G$ such that $\mathrm{Tr}(\breve{S}(\breve{\theta})) \leq \breve{M}_J, \|\breve{\theta}\| \leq \breve{M}_{\breve{\theta}}, \|\breve{S}(\breve{\theta})\| \leq \breve{M}_S$ and $\|[I, \breve{G}(\breve{\theta})^{\mathsf{T}}]^{\mathsf{T}}\| \leq \breve{M}_G$ for all $\breve{\theta} \in \breve{\Theta}$ where $\|\cdot\|$ is the induced matrix norm.

Let $\breve{X}_T^i = \breve{\sigma}^i + \max_{1 \leq t \leq T} \|\breve{x}_t^i\|$. The next two bounds follow from [40, Lemma 4] and [40, Lemma 5].

Lemma 4: For each node $i \in N$, any $q \geq 1$ and any $T > 1$

$$\mathbb{E}\left[\frac{(\breve{X}_T^i)^q}{(\breve{\sigma}^i)^q}\right] \leq \mathcal{O}\left(\log T\right).$$

Lemma 5: For any $q \geq 1$, we have

$$\mathbb{E}\left[\frac{(\breve{X}_T^i)^q}{(\breve{\sigma}^i)^q} \log\left(\sum_{t=1}^{T} \frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\right)\right]$$
$$\leq \mathbb{E}\left[\frac{(\breve{X}_T^i)^q}{(\breve{\sigma}^i)^q} \log\left(\sum_{t=1}^{T}\sum_{i \in N} \frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\right)\right] \leq \tilde{\mathcal{O}}(1). \quad (38)$$

The next lemma gives an upper bound on the number of episodes $\breve{K}_T$.

Lemma 6: The number of episodes $\breve{K}_T$ is bounded as follows:

$$\breve{K}_T \leq \mathcal{O}\left(\sqrt{(d_x + d_u)T \log\left(\sum_{t=1}^{T-1} \frac{(\breve{X}_T^{j_t})^2}{(\breve{\sigma}^{j_t})^2}\right)}\right).$$

Proof: We can follow the same argument as in the proof of [40, Lemma 5]. Let $\breve{\eta} - 1$ be the number of times the second stopping criterion is triggered for $\breve{p}_t$. Using the analysis in the proof of [40, Lemma 5], we can get the following inequalities:

$$\breve{K}_T \leq \sqrt{2\breve{\eta}T}, \quad (39)$$

$$\det(\breve{\Sigma}_T^{-1}) \geq 2^{\breve{\eta}-1} \det(\breve{\Sigma}_1^{-1}) \geq 2^{\breve{\eta}-1}\breve{\lambda}_{\min}^d \quad (40)$$

where $d = d_x + d_u$ and $\breve{\lambda}_{\min}$ is the minimum eigenvalue of $\breve{\Sigma}_1^{-1}$.

Combining (40) with $\mathrm{Tr}(\breve{\Sigma}_T^{-1})/d \geq \det(\breve{\Sigma}_T^{-1})^{1/d}$, we get $\mathrm{Tr}(\breve{\Sigma}_T^{-1}) \geq d\breve{\lambda}_{\min}2^{(\breve{\eta}-1)/d}$. Thus

$$\breve{\eta} \leq 1 + \frac{d}{\log 2} \log\left(\frac{\mathrm{Tr}(\breve{\Sigma}_T^{-1})}{d\breve{\lambda}_{\min}}\right). \quad (41)$$

Now, we bound $\mathrm{Tr}(\breve{\Sigma}_T^{-1})$. From (28b), we have

$$\mathrm{Tr}(\breve{\Sigma}_T^{-1}) = \mathrm{Tr}(\breve{\Sigma}_1^{-1}) + \sum_{t=1}^{T-1} \frac{1}{(\breve{\sigma}^{j_t})^2} \underbrace{\mathrm{Tr}(\breve{z}_t^{j_t}(\breve{z}_t^{j_t})^{\mathsf{T}})}_{=\|\breve{z}_t^{j_t}\|^2}. \quad (42)$$

Note that $\|\breve{z}_t^{j_t}\| = \|[I, \breve{G}(\breve{\theta})^{\mathsf{T}}]^{\mathsf{T}}\breve{x}_t^{j_t}\| \leq \breve{M}_G\|\breve{x}_t^{j_t}\| \leq \breve{M}_G\breve{X}_T^{j_t}$. Using $\|\breve{z}_t^{j_t}\|^2 \leq \breve{M}_G^2(\breve{X}_T^{j_t})^2$ in (42) and substituting the resulting bound on $\mathrm{Tr}(\breve{\Sigma}_T^{-1})$ in (41) and then combining it with the bound on η in (39) give the result of the lemma. ■

Lemma 7: . The expected value of $\breve{K}_T$ is bounded as follows:

$$\mathbb{E}[\breve{K}_T] \leq \tilde{\mathcal{O}}\left(\sqrt{(d_x + d_u)T}\right).$$

Proof: From Lemma 6, we get

$$\mathbb{E}[\breve{K}_T] \leq \mathcal{O}\left(\mathbb{E}\left[\sqrt{(d_x + d_u)T \log\left(\sum_{t=1}^{T-1} \frac{(\breve{X}_T^{j_t})^2}{(\breve{\sigma}^{j_t})^2}\right)}\right]\right)$$

$$\overset{(a)}{\leq} \mathcal{O}\left(\sqrt{(d_x+d_u)T\log\left(\mathbb{E}\left[\sum_{t=1}^{T-1}\frac{(\breve{X}_T^{j_t})^2}{(\breve{\sigma}^{j_t})^2}\right]\right)}\right)$$

$$\leq \mathcal{O}\left(\sqrt{(d_x+d_u)T\log\left(\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in N}\frac{(\breve{X}_T^{i})^2}{(\breve{\sigma}^{i})^2}\right]\right)}\right)$$

$$\overset{(b)}{\leq} \tilde{\mathcal{O}}(\sqrt{(d_x+d_u)T})$$

where (a) follows from Jensen's inequality and (b) follows from Lemma 4. ■

B. Proof of Lemma 3

Proof: We will prove each part separately.

1) Bounding $\breve{R}_0^i(T)$: From an argument similar to the proof of [20, Lemma 5], we get that $\breve{R}_0^i(T) \leq (\breve{\sigma}^i)^2\breve{M}_J\mathbb{E}[\breve{K}_T]$. The result then follows from substituting the bound on $\mathbb{E}[\breve{K}_T]$ from Lemma 7.

2) Bounding $\breve{R}_1^i(T)$:

$$\breve{R}_1^i(T) = \mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}\left[(\breve{x}_t^i)^\top\breve{S}_k\breve{x}_t^i - (\breve{x}_{t+1}^i)^\top\breve{S}_k\breve{x}_{t+1}^i\right]\right]$$

$$= \mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}\left[(\breve{x}_{\breve{t}_k}^i)^\top\breve{S}_k\breve{x}_{\breve{t}_k}^i - (\breve{x}_{\breve{t}_{k+1}}^i)^\top\breve{S}_k\breve{x}_{\breve{t}_{k+1}}^i\right]\right]$$

$$\leq \mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}(\breve{x}_{\breve{t}_k}^i)^\top\breve{S}_k\breve{x}_{\breve{t}_k}^i\right] \leq \mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}\|\breve{S}_k\|\|\breve{x}_{t_k}^i\|^2\right]$$

$$\leq \breve{M}_S\mathbb{E}[\breve{K}_T(\breve{X}_T^i)^2] \tag{43}$$

where the last inequality follows from $\|\breve{S}_k\| \leq \breve{M}_S$. Using the bound for $\breve{K}_T$ in Lemma 6, we get

$$\breve{R}_1^i(T) \leq \mathcal{O}\left(\sqrt{(d_x+d_u)T}\mathbb{E}\left[(\breve{X}_T^i)^2\sqrt{\log\left(\sum_{t=1}^{T-1}\frac{(\breve{X}_T^{j_t})^2}{(\breve{\sigma}^{j_t})^2}\right)}\right]\right). \tag{44}$$

Now, consider the term

$$\mathbb{E}\left[(\breve{X}_T^i)^2\sqrt{\log\left(\sum_{t=1}^{T-1}\frac{(\breve{X}_T^{j_t})^2}{(\breve{\sigma}^{j_t})^2}\right)}\right])$$

$$\overset{(a)}{\leq} \sqrt{\mathbb{E}[(\breve{X}_T^i)^4]\ \mathbb{E}\left[\log\left(\sum_{t=1}^{T-1}\frac{(\breve{X}_T^{j_t})^2}{(\breve{\sigma}^{j_t})^2}\right)\right]}$$

$$\overset{(b)}{\leq} \sqrt{\mathbb{E}[(\breve{X}_T^i)^4]\ \log\left(\mathbb{E}\left[\sum_{t=1}^{T-1}\sum_{i\in N}\frac{(\breve{X}_T^{i})^2}{(\breve{\sigma}^{j_t})^2}\right]\right)}$$

$$\overset{(c)}{\leq} \tilde{\mathcal{O}}((\breve{\sigma}^i)^2) \tag{45}$$

where (a) follows from Cauchy–Schwarz, (b) follows from Jensen's inequality, and (c) follows from Lemma 4. The result then follows from substituting (45) in (43).

3) Bounding $\breve{R}_2^i(T)$: As in [20], we can bound the inner summand in $\breve{R}_2^i(T)$ as

$$\left[(\breve{\theta}^\top\breve{z}_t^i)^\top\breve{S}_k(\breve{\theta}^\top\breve{z}_t^i) - (\breve{\theta}_k^\top\breve{z}_t^i)^\top\breve{S}_k((\breve{\theta}_k)^\top\breve{z}_t^i)\right]$$

$$\leq \mathcal{O}(\breve{X}_T^i\|(\breve{\theta}-\breve{\theta}_k)^\top\breve{z}_t^i\|).$$

Therefore

$$\breve{R}_2^i(T) \leq \mathcal{O}\left(\mathbb{E}\left[\breve{X}_T^i\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}\|(\breve{\theta}-\breve{\theta}_k)^\top\breve{z}_t^i\|\right]\right). \tag{46}$$

The term inside $\mathcal{O}(\cdot)$ can be written as

$$\mathbb{E}\left[\breve{X}_T^i\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}\|(\breve{\theta}-\breve{\theta}_k)^\top\breve{z}_t^i\|\right]$$

$$= \mathbb{E}\left[\breve{X}_T^i\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}\|(\breve{\Sigma}_{t_k}^{-0.5}(\breve{\theta}-\breve{\theta}_k))^\top\breve{\Sigma}_{t_k}^{0.5}\breve{z}_t^i\|\right]$$

$$\leq \mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}\|\breve{\Sigma}_{t_k}^{-0.5}(\breve{\theta}-\breve{\theta}_k)\| \times \breve{X}_T^i\|\breve{\Sigma}_{t_k}^{0.5}\breve{z}_t^i\|\right]$$

$$\leq \sqrt{\mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}\|\breve{\Sigma}_{t_k}^{-0.5}(\breve{\theta}-\breve{\theta}_k)\|^2\right]}$$

$$\times \sqrt{\mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}(\breve{X}_T^i)^2\|\breve{\Sigma}_{t_k}^{0.5}\breve{z}_t^i\|^2\right]} \tag{47}$$

where the last inequality follows from Cauchy–Schwarz inequality.

Following the same argument as [40, Lemma 7], the first part of (47) is bounded by

$$\mathbb{E}\left[\sum_{k=1}^{\breve{K}_T}\sum_{t=\breve{t}_k}^{\breve{t}_{k+1}-1}\|\breve{\Sigma}_{t_k}^{-0.5}(\breve{\theta}-\breve{\theta}_k)\|^2\right] \leq \mathcal{O}(d_x(d_x+d_u)T). \tag{48}$$

For the second part of the bound in (47), we follow the same argument as [40, Lemma 8]. Recall that $\breve{\lambda}_{\min}$ is the smallest eigenvalue of $\breve{\Sigma}_1^{-1}$. Therefore, by (28b), all eigenvalues of $\breve{\Sigma}_t^{-1}$ are no smaller than $\breve{\lambda}_{\min}$. Or, equivalently, all eigenvalues of $\breve{\Sigma}_t$ are no larger than $1/\breve{\lambda}_{\min}$.

Using [11, Lemma 11], we can show that for any $t \in \{t_k, \ldots, t_{k+1}-1\}$,

$$\|\breve{\Sigma}_{t_k}^{0.5}\breve{z}_t^i\|^2 = (\breve{z}_t^i)^\top\breve{\Sigma}_{t_k}\breve{z}_t^i \leq \frac{\det\breve{\Sigma}_t^{-1}}{\det\breve{\Sigma}_{t_k}^{-1}}(\breve{z}_t^i)^\top\breve{\Sigma}_t\breve{z}_t^i$$

$$\leq F_1(\breve{X}_T^i)\,(\breve{z}_t^i)^\top\breve{\Sigma}_t\breve{z}_t^i \tag{49}$$

where $F_1(\breve{X}_T^i) = \left(1+(\breve{M}_G^2(\breve{X}_T^i)^2/\breve{\lambda}_{\min}\breve{\sigma}_w^2)\right)^{\breve{T}_{\min}\vee 1}$ and the last inequality follows from [40, Lemma 10].

Moreover, since all eigenvalues of $\breve{\Sigma}_t$ are no larger than $1/\breve{\lambda}_{\min}$, we have $(\breve{z}_t^i)^\top\breve{\Sigma}_t\breve{z}_t^i \leq \|\breve{z}_t^i\|^2/\breve{\lambda}_{\min} \leq \breve{M}_G^2(\breve{X}_T^i)^2/\breve{\lambda}_{\min}$.

Therefore

$$\begin{aligned}(\breve{z}_t^i)^\mathsf{T}\breve{\Sigma}_t\breve{z}_t^i &\le \left((\breve{\sigma}^i)^2 \vee \frac{\breve{M}_G^2(\breve{X}_T^i)^2}{\breve{\lambda}_{\min}}\right)\left(1 \wedge \frac{(\breve{z}_t^i)^\mathsf{T}\breve{\Sigma}_t\breve{z}_t^i}{(\breve{\sigma}^i)^2}\right) \\ &\le \left((\breve{\sigma}^i)^2 + \frac{\breve{M}_G^2(\breve{X}_T^i)^2}{\breve{\lambda}_{\min}}\right)\left(1 \wedge \frac{(\breve{z}_t^{j_t})^\mathsf{T}\breve{\Sigma}_t\breve{z}_t^{j_t}}{(\breve{\sigma}^{j_t})^2}\right)\end{aligned} \tag{50}$$

where the last inequality follows from the definition of j_t. Let $F_2(\breve{X}_T^i) = \big((\breve{\sigma}^i)^2 + (\breve{\lambda}_{\min}/\breve{M}_G^2(\breve{X}_T^i)^2)\big)$. Then

$$\begin{aligned}\sum_{t=1}^T(\breve{z}_t^i)^\mathsf{T}\breve{\Sigma}_t\breve{z}_t^i &\le F_2(\breve{X}_T^i)\sum_{t=1}^T\left(1 \wedge \frac{(\breve{z}_t^{j_t})^\mathsf{T}\breve{\Sigma}_t\breve{z}_t^{j_t}}{(\breve{\sigma}^{j_t})^2}\right) \\ &= F_2(\breve{X}_T^i)\sum_{t=1}^T\left(1 \wedge \left\|\frac{\Sigma_t^{0.5}\,\breve{z}_t^{j_t}(\breve{z}_t^{j_t})^\mathsf{T}\Sigma_t^{0.5}}{(\breve{\sigma}^{j_t})^2}\right\|\right) \\ &\overset{(a)}{\le} F_2(\breve{X}_T^i)\left[2d\log\left(\frac{\mathrm{Tr}(\breve{\Sigma}_{T+1}^{-1})}{d}\right) - \log\det\Sigma_1^{-1}\right] \\ &\overset{(b)}{\le} F_2(\breve{X}_T^i)\Bigg[2d\log\left(\frac{1}{d}\left(\mathrm{Tr}(\breve{\Sigma}_1^{-1}) + \breve{M}_G\sum_{t=1}^T\frac{(\breve{X}_T^{j_t})^2}{(\breve{\sigma}^{j_t})^2}\right)\right) \\ &\quad - \log\det\Sigma_1^{-1}\Bigg]\end{aligned} \tag{51}$$

where $d = d_x + d_u$ and (a) follows from (28b) and the intermediate step in the proof of [43, Lemma 6]. and (b) follows from (42) and the subsequent discussion.

Using (49) and (51), we can bound the second term of (47) as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T(\breve{X}_T^i)^2\|\breve{\Sigma}_{t_k}^{0.5}\,\breve{z}_t^i\|^2\right] &\le \mathcal{O}\Bigg(d\,\mathbb{E}\Bigg[F_1(\breve{X}_t^i)F_2(\breve{X}_T^i)(\breve{X}_T^i)^2 \\ &\quad \times \log\left(\sum_{t=1}^T\frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\right)\Bigg]\Bigg) \\ &\le \mathcal{O}\Bigg(d(\breve{\sigma}^i)^4\mathbb{E}\Bigg[F_1(\breve{X}_T^i)\frac{F_2(\breve{X}_T^i)}{(\breve{\sigma}^i)^2}\frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\log\left(\sum_{t=1}^T\frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\right)\Bigg]\Bigg) \\ &\le \tilde{\mathcal{O}}(d(\breve{\sigma}^i)^4)\end{aligned} \tag{52}$$

where the last inequality follows by observing that $F_1(\breve{X}_T^i)\frac{F_2(\breve{X}_T^i)}{(\breve{\sigma}^i)^2}\frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\log\left(\sum_{t=1}^T\frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\right)$ is a polynomial in $\breve{X}_T^i/\breve{\sigma}^i$ multiplied by $\log\left(\sum_{t=1}^T\frac{(\breve{X}_T^i)^2}{(\breve{\sigma}^i)^2}\right)$ and, using Lemma 5.

The result then follows by substituting (48) and (52) in (47).■

References

[1] N. Sandell, P. Varaiya, M. Athans, and M. Safonov, “Survey of decentralized control methods for large scale systems,” *IEEE Trans. Autom. Control*, vol. AC-23, no. 2, pp. 108–128, Apr. 1978.

[2] J. Lunze, “Dynamics of strongly coupled symmetric composite systems,” *Int. J. Control*, vol. 44, no. 6, pp. 1617–1640, 1986.

[3] M. K. Sundareshan and R. M. Elbanna, “Qualitative analysis and decentralized controller synthesis for a class of large-scale systems with symmetrically interconnected subsystems,” *Automatica*, vol. 27, no. 2, pp. 383–388, 1991.

[4] G.-H. Yang and S.-Y. Zhang, “Structural properties of large-scale systems possessing similar structures,” *Automatica*, vol. 31, no. 7, pp. 1011–1017, 1995.

[5] S. C. Hamilton and M. E. Broucke, “Patterned linear systems,” *Automatica*, vol. 48, no. 2, pp. 263–272, 2012.

[6] J. Arabneydi and A. Mahajan, “Team-optimal solution of finite number of mean-field coupled LQG subsystems,” in *Proc. Conf. Decis. Control*, Kyoto, Japan, 2015, pp. 5308–5313.

[7] K. J. Astrom and B. Wittenmark, *Adaptive Control*. Boston, MA, USA: Addison-Wesley Longman Publishing, 1994.

[8] S. J. Bradtke, “Reinforcement learning applied to linear quadratic regulation,” in *Proc. Neural Inf. Process. Syst.*, 1993, pp. 295–302.

[9] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, “Adaptive linear quadratic control using policy iteration,” in *Proc. Amer. Control Conf.*, vol. 3, 1994, pp. 3475–3479.

[10] M. C. Campi and P. Kumar, “Adaptive linear quadratic gaussian control: The cost-biased approach revisited,” *SIAM J. Control Optim.*, vol. 36, no. 6, pp. 1890–1907, 1998.

[11] Y. Abbasi-Yadkori and C. Szepesvári, “Regret bounds for the adaptive control of linear quadratic systems,” in *Proc. Annu. Conf. Learn. Theory*, 2011, pp. 1–26.

[12] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, “Optimism-based adaptive regulation of linear-quadratic systems,” *IEEE Trans. Autom. Control*, vol. 66, no. 4, pp. 1802–1808, Apr. 2021.

[13] A. Cohen, T. Koren, and Y. Mansour, “Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1300–1309.

[14] M. Abeille and A. Lazaric, “Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 23–31.

[15] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, “Regret bounds for robust adaptive control of the linear quadratic regulator,” in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 4192–4201.

[16] H. Mania, S. Tu, and B. Recht, “Certainty equivalent control of LQR is efficient,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Dec. 2019, pp. 10154–10164.

[17] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, “Input perturbations for adaptive control and learning,” *Automatica*, vol. 117, 2020, Art. no. 108950.

[18] M. Simchowitz and D. Foster, “Naive exploration is optimal for online LQR,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8937–8948.

[19] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, “On adaptive linear–quadratic regulators,” *Automatica*, vol. 117, Jul. 2020, Art. no. 108982.

[20] Y. Ouyang, M. Gagrani, and R. Jain, “Posterior sampling-based reinforcement learning for control of unknown linear systems,” *IEEE Trans. Autom. Control*, vol. 65, no. 8, pp. 3600–3607, Aug. 2020.

[21] M. Abeille and A. Lazaric, “Improved regret bounds for thompson sampling in linear quadratic control problems,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–9.

[22] S. Gao and A. Mahajan, “Optimal control of network-coupled subsystems: Spectral decomposition and low-dimensional solutions,” *IEEE Trans. Control Netw. Syst.*, vol. 9, no. 2, pp. 657–669, Jun. 2022, doi: 10.1109/TCNS.2021.3124259.

[23] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Mach. Learn.*, vol. 47, no. 2–3, pp. 235–256, 2002.

[24] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in *Proc. Conf. Learn. Theory*, 2012, pp. 39.1–39.26.

[25] H. Wang, S. Lin, H. Jafarkhani, and J. Zhang, “Distributed q-learning with state tracking for multi-agent networked control,” in *Proc. AAMAS*, 2021, pp. 1692–1694.

[26] G. Jing, H. Bai, J. George, A. Chakrabortty, and P. K. Sharma, “Learning distributed stabilizing controllers for multi-agent systems,” *IEEE Control Syst. Lett.*, vol. 6, pp. 301–306, 2022.

[27] Y. Li, Y. Tang, R. Zhang, and N. Li, “Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach,” in *Proc. Conf. Learn. Dyn. Control*, 2020, pp. 814–814.

[28] S. Alemzadeh and M. Mesbahi, “Distributed q-learning for dynamically decoupled systems,” in *Proc. Amer. Control Conf.*, 2019, pp. 772–777.

[29] J. Bu, A. Mesbahi, and M. Mesbahi, “Policy gradient-based algorithms for continuous-time linear quadratic control,” 2020, *arXiv:2006.09178*.

[30] H. Mohammadi, M. R. Jovanovic, and M. Soltanolkotabi, “Learning the model-free linear quadratic regulator via random search,” in *Proc. Learning Dyn. Control*, 2020, pp. 531–539.

[31] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5872–5881.

[32] I. Osband and B. Van Roy, "Near-optimal reinforcement learning in factored MDPs," in *Advances in Neural Information Processing Systems*, vol. 27. Red Hook, NY, USA: Curran Associates, 2014.

[33] X. Chen, J. Hu, L. Li, and L. Wang, "Efficient reinforcement learning in factored MDPs with application to constrained RL," in *Proc. Intl. Conf. Learn. Representations*, 2021.

[34] M. Huang, P. E. Caines, and R. P. Malhamé, "Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized epsilon-Nash equilibria," *IEEE Trans. Autom. Control*, vol. 52, no. 9, pp. 1560–1571, Sep. 2007.

[35] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese J. Math.*, vol. 2, no. 1, pp. 229–260, 2007.

[36] S. G. Subramanian, P. Poupart, M. E. Taylor, and N. Hegde, "Multi type mean field reinforcement learning," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2020, pp. 411–419.

[37] M. A. uz Zaman, K. Zhang, E. Miehling, and T. Başar, "Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games," in *Proc. Conf. Decis. Control*, 2020, pp. 2278–2284.

[38] M. Gagrani, S. Sudhakara, A. Mahajan, A. Nayyar, and Y. Ouyang, "Thompson sampling for linear quadratic mean-field teams." in *Proc. Conf. Decis. Control*, Austin, TX, USA, Dec. 2021.

[39] J. Sternby, "On consistency for the method of least squares using martingale theory," *IEEE Trans. Autom. Control*, vol. AC-22, no. 3, pp. 346–352, Jun. 1977.

[40] M. Gagrani, S. Sudhakara, A. Mahajan, A. Nayyar, and Y. Ouyang, "A relaxed technical assumption for posterior sampling-based reinforcement learning for control of unknown linear systems," in *Proc. Conf. Decis. Control*, Cancun, Mexico, Dec. 2022.

[41] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite-time adaptive stabilization of linear systems," *IEEE Trans. Autom. Control*, vol. 64, no. 8, pp. 3498–3505, Aug. 2019.

[42] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Philadelphia, PA, USA: SIAM Classics, 2015.

[43] Y. Abbasi-Yadkori and C. Szepesvari, "Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm," 2014, *arXiv:1406.3926*.

Sagar Sudhakara received the M.Tech degree in communication and signal processing from the Indian Institute of Technology Bombay, Mumbai, India, in 2016. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Southern California, Los Angeles, CA, USA.

His research interests include reinforcement learning and decentralized stochastic control.

Mr. Sudhakara is a recipient of the USC Annenberg fellowship.

Aditya Mahajan (Senior Member, IEEE) received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Kanpur, India and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of Michigan, Ann Arbor, USA.

He is currently an Associate Professor with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. His principal research interests include learning and control of centralized and decentralized stochastic systems.

Dr. Mahajan currently serves as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *IEEE Control System Letters*, and *Mathematics of Control, Signal, and Systems*. He is the recipient of the 2015 George Axelby Outstanding Paper Award, 2014 CDC Best Student Paper Award (as supervisor), and the 2016 NecSys Best Student Paper Award (as supervisor).

Ashutosh Nayyar (Senior Member, IEEE) received the M.S. degree in electrical engineering and computer science, the M.S. degree in applied mathematics, and the Ph.D. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, MI, USA, in 2008, 2011, and 2011, respectively.

He was a Postdoctoral Researcher with the University of Illinois at Urbana-Champaign, Champaign, IL, USA, and with the University of California, Berkeley, Berkeley, CA, USA, before joining the University of Southern California, Los Angeles, CA, USA, in 2014. His research interests are in decentralized stochastic control, decentralized decision-making in sensing and communication systems, reinforcement learning, game theory, mechanism design, and electric energy systems.

Yi Ouyang received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 2009, and the M.Sc. and Ph.D. degrees in electrical engineering and computer science from the University of Michigan, Ann Arbor, MI, USA, in 2012 and 2015, respectively.

He is currently a Researcher with Preferred Networks, Burlingame, CA, USA. His research interests include reinforcement learning, stochastic control, and stochastic dynamic games.