Online Actuator Selection and Controller Design for Linear Quadratic Regulation with Unknown System Model

Lintao Ye, Ming Chi, Zhi-Wei Liu, and Vijay Gupta

***Abstract*—We study the simultaneous actuator selection and controller design problem for linear quadratic regulation with Gaussian noise over a finite horizon of length T and unknown system model. We consider episodic and non-episodic settings of the problem and propose online algorithms that specify both the sets of actuators to be utilized under a cardinality constraint and the controls corresponding to the sets of selected actuators. In the episodic setting, the interaction with the system breaks into N episodes, each of which restarts from a given initial condition and has length T. In the non-episodic setting, the interaction goes on continuously. Our online algorithms leverage a multiarmed bandit algorithm to select the sets of actuators and a certainty equivalence approach to design the corresponding controls. We show that our online algorithms yield $\sqrt{N}$-regret for the episodic setting and $T^{2/3}$-regret for the non-episodic setting. We extend our algorithm design and analysis to show scalability with respect to both the total number of candidate actuators and the cardinality constraint. We numerically validate our theoretical results.**

I. Introduction

In large-scale control system design, the number of actuators (or sensors) that can be installed is often limited by budget or complexity constraints. The problem of selecting a subset of all the candidate actuators (or sensors), in order to optimize a system objective while satisfying a budget constraint is a classic problem referred to as actuator (or sensor) selection [1]–[8]. However, most of the existing works on this problem assume the knowledge of the system model when designing the actuator (or sensor) selection algorithms. In this work, we are interested in the situation when the system model is unknown [9]. In such a case, the existing algorithms for the actuator selection problem do not apply.

We study the simultaneous actuator selection and controller design problem for Linear Quadratic Regulation (LQR) [10]. The goal is to select a sequence of sets of actuators each with a cardinality constraint, while minimizing the accumulative quadratic cost over a time horizon. We assume that the system model is unknown and the problem needs to be solved in an online manner. We study two settings of the problem: episodic and non-episodic settings. In the episodic setting, the interaction with the system breaks into subsequences, each of which starts from a given initial condition and ends at a terminal time step. In the non-episodic setting, the interaction with the system goes on continuously. Both the episodic and non-episodic settings are widely studied in general reinforcement learning problems, and capture different scenarios in practice [11], [12]. We provide online algorithms to solve the problem, and characterize their regret performance [13]–[15].

Related Work: Actuator (or sensor) selection has been studied in the literature extensively. Since the problem is NP-hard [8], much work in the literature provides approximation algorithms to solve the problem [3], [16], [17]. However, most of the previous work assumes a known system model. Exceptions are [18], [19], where the authors studied an online sensor selection problem for the estimation of a static random variable. Another related work is [20], where the authors considered an unknown continuous-time linear time-invariant system without stochastic noise and studied the problem of selecting a subset of actuators under a cardinality constraint such that a controllability metric of the system is optimized.

The LQR problem with unknown system matrices, also known as the optimal adaptive control problem, has been widely studied [21]–[25]. One standard approach (so-called certainty equivalence) first estimates the system matrices from system trajectories and then uses the estimate of the system matrices to design the control as if the true system matrices are available. Thus, it is crucial to ensure the consistency of the estimate in order to achieve the optimal performance. Based on the consistent estimates returned by least squares as shown by [26], [27] designed a certainty equivalent controller with an additive random perturbation. In [21] and [22], a reward-biased estimate of the system matrices is utilized. In [28], randomly perturbed least squares and Thompson sampling were proposed to obtain the estimates to design the certainty equivalent controller and a regret of nearly square-root growth rate was established. However, the aforementioned works focused on the asymptotic performance of the certainty equivalent controller (as the number of data samples from the system trajectories used for estimating the system matrices goes to infinity). The finite-sample analyses of the certainty equivalence approach have also been studied for learning LQR [29]–[33]. It was shown in [30], [32] that the certainty equivalent controller with a certain additive random perturbation achieves a regret of $\tilde{O}(\sqrt{T})$, where T is the number of time steps in the LQR problem and $\tilde{O}(\cdot)$ hides logarithmic factors in T. Moreover, [34] analyzed the regret of the certainty equivalence approach based on a reward-biased estimate.

Contributions: We now summarize our contributions. First, we formulate the simultaneous actuator selection and controller design problem for LQR with unknown system model. This problem is challenging since it contains both discrete and continuous variables (the sets of actuators and the corresponding controls, respectively). The online algorithms that

The work of the first three authors was supported in part by National Natural Science Foundation of China Grants 62203179, 62222205 and 62373162, and Natural Science Foundation of Hubei Province of China Grants 2022CFB670 and 2022CFA052. The work of Vijay Gupta was supported in part by ARO grant W911NF2310111 and NSF grant 2300355. Lintao Ye, Ming Chi and Zhi-Wei Liu are with the School of Artificial Intelligence and Automation at the Huazhong University of Science and Technology, Wuhan, China. Emails: {yelintao93, zwliu, chiming}@hust.edu.cn. Vijay Gupta is with the Elmore Family School of Electrical and Computer Engineering at Purdue University, IN, USA. Email: gupta869@purdue.edu.

we propose to solve the problem contain two phases. First, the system matrices are estimated based on the data samples from a single system trajectory. Based on the estimated system matrices, the online algorithms then leverage a multiarmed bandit algorithm [35] to select the set of actuators, and leverage the certainty equivalence approach [30] for the controller design. We carefully balance the length of the two phases, when characterizing the regret of the online algorithms.

Second, we consider the actuator selection problem for finite-horizon LQR. We extend the analysis and results for the certainty equivalence approach proposed for learning infinite-horizon LQR (without the actuator selection component) [30], [31] to the finite-horizon setting. The analysis for the finite-horizon setting is more challenging, since the optimal controller for finite-horizon LQR is time-varying in general, while the optimal controller for infinite LQR is time-invariant [36].

Third, we provide a comprehensive study of the problem by considering both the online episodic and non-episodic settings. The non-episodic setting is more challenging than the episodic setting, since the system state cannot be reset to a given initial condition after each episode. However, we show that given a non-episodic instance of the problem, one can first construct a corresponding episodic instance and then apply the proposed online algorithms. We show that our online algorithm for the episodic setting yields a regret of $\tilde{O}(\sqrt{T^2N})$, where N is the number of episodes and T is the number of time steps in each episode. For the non-episodic setting, the online algorithm yields a regret of $\tilde{O}(T^{2/3})$, where T is the horizon length.

Finally, we extend our analysis to efficiently handle instances of the problem when both the total number of candidate actuators and the cardinality constraint scale large. Since the (offline) actuator selection problem for LQR with known system model is NP-hard [8], we leverage a weaker notion of regret, i.e., c-regret, introduced for online algorithms for combinatorial optimization problems [19], [37], [38], and characterize the performance of the online algorithm that we propose for the large-scale problem instances. We show that the c-regret of our online algorithm scales as $\tilde{O}(TN^{2/3})$ (resp., $\tilde{O}(T^{3/4})$) in the episodic (resp., non-episodic) setting, where $c \in (0,1)$ is parameterized by the problem parameters.

An extended version of the paper that contains all the omitted proofs can be found on arXiv as [39].

Notation and terminology: The sets of integers and real numbers are denoted as $\mathbb{Z}$ and $\mathbb{R}$, respectively. For a real number a, let $\lceil a \rceil$ be the smallest integer that is greater than or equal to a. For a matrix $P \in \mathbb{R}^{n\times n}$, let $P^\top$, $\mathrm{Tr}(P)$, and $\{\sigma_i(P) : i \in \{1,\dots,n\}\}$ be its transpose, trace, and set of singular values, respectively. Without loss of generality, the singular values of P are ordered as $\sigma_1(P) \geq \cdots \geq \sigma_n(P)$. Let $\|\cdot\|$ denote the ℓ_2 norm, i.e., $\|P\| = \sigma_1(P)$ for a matrix $P \in \mathbb{R}^{n\times n}$, and $\|x\| = \sqrt{x^\top x}$ for a vector $x \in \mathbb{R}^n$. Let $\|P\|_F = \sqrt{\mathrm{Tr}(PP^\top)}$ be the Frobenius norm of $P \in \mathbb{R}^{n\times m}$. A positive semidefinite matrix P is denoted by $P \succeq 0$, and $P \succeq Q$ if and only if $P - Q \succeq 0$. Let $\mathbb{S}^n_+$ (resp., $\mathbb{S}^n_{++}$) denote the set of $n \times n$ positive semidefinite (resp., positive definite) matrices. Let I be an identity matrix whose dimension can be inferred from the context. For any integer $n \geq 1$, $[n] \triangleq \{1,\dots,n\}$. The cardinality of a finite set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. Let $\mathbb{1}\{\cdot\}$ denote an indicator function.

II. Problem Formulation and Preliminaries

A. Problem Formulation

Consider a discrete-time linear time-invariant system

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is the system dynamics matrix, $x_t \in \mathbb{R}^n$ is the state vector, $B \in \mathbb{R}^{n\times m}$ is the input matrix, $u_t \in \mathbb{R}^m$ is the control, and $\{w_t\}_{t\geq 0}$ are i.i.d. noise terms with zero mean and covariance W for all $t \in \mathbb{Z}_{\geq 0}$. Let $\mathcal{G}$ be the set that contains all the candidate actuators. Denote $B = \begin{bmatrix} B_1 & \cdots & B_{|\mathcal{G}|} \end{bmatrix}$, where $B_i \in \mathbb{R}^{n\times m_i}$ for all $i \in \mathcal{G}$ with $\sum_{i\in\mathcal{G}} m_i = m$. For any $i \in \mathcal{G}$, B_i corresponds to a candidate actuator that can be potentially selected and installed. At each time step $t \in \mathbb{Z}_{\geq 0}$, only a subset of actuators out of all the candidate actuators is selected to provide controls to system (1), due to, e.g., budget constraints. For any $t \in \mathbb{Z}_{\geq 0}$, let $\mathcal{S}_t \subseteq \mathcal{G}$ denote the set of actuators selected for time step t, let $B_{\mathcal{S}_t} \triangleq \begin{bmatrix} B_{i_1} & \cdots & B_{i_{|\mathcal{S}_t|}} \end{bmatrix}$ be the input matrix associated with the actuators in $\mathcal{S}_t$, and let $u_{t,\mathcal{S}_t} \triangleq \begin{bmatrix} u_{t,i_1}^\top & \cdots & u_{t,i_{|\mathcal{S}_t|}}^\top \end{bmatrix}^\top$ be the control provided by the actuators in $\mathcal{S}_t$, where $\mathcal{S}_t = \{i_1,\dots,i_{|\mathcal{S}_t|}\}$. Given a horizon length $T \in \mathbb{Z}_{\geq 1}$ and an actuator selection $\mathcal{S} \triangleq (\mathcal{S}_0,\dots,\mathcal{S}_{T-1})$, we consider the following quadratic cost:

$$J(\mathcal{S}, u_{\mathcal{S}}) \triangleq \Big(\sum_{t=0}^{T-1} x_t^\top Q x_t + u_{t,\mathcal{S}_t}^\top R_{\mathcal{S}_t} u_{t,\mathcal{S}_t} \Big) + x_T^\top Q_f x_T, \tag{2}$$

where $u_{\mathcal{S}} \triangleq (u_{0,\mathcal{S}_0},\dots,u_{T-1,\mathcal{S}_{T-1}})$, $Q, Q_f \in \mathbb{S}^n_+, R \in \mathbb{S}^m_{++}$ are the cost matrices, and $R_{\mathcal{S}_t} \in \mathbb{S}^{m_{\mathcal{S}_t}}_{++}$ (with $m_{\mathcal{S}_t} = \sum_{i\in\mathcal{S}_t} m_i$) is a submatrix of R corresponding to the set $\mathcal{S}_t$.[1]

Given system (1) and $T \in \mathbb{Z}_{\geq 1}$, our goal is to solve the following actuator selection problem for LQR:

$$\begin{aligned} &\min_{\mathcal{S}, u_{\mathcal{S}}} \mathbb{E}\big[J(\mathcal{S}, u_{\mathcal{S}})\big], \\ &s.t.\ \mathcal{S}_t \subseteq \mathcal{G}, |\mathcal{S}_t| = H,\ \forall t \in \{0,\dots,T-1\}, \end{aligned} \tag{3}$$

where $H \in \mathbb{Z}_{\geq 1}$ is a cardinality constraint on the sets of selected actuators, and the expectation is taken with respective to $w_0,\dots,w_{T-1}$. Conditioning on an actuator selection $\mathcal{S} = (\mathcal{S}_0,\dots,\mathcal{S}_{T-1})$, it is well-known that the corresponding optimal control, i.e., $\tilde{u}_{\mathcal{S}} \in \arg\min_{u_{\mathcal{S}}} \mathbb{E}[J(\mathcal{S}, u_{\mathcal{S}})]$, is given by a linear state-feedback controller [36, Chapter 3]

$$\tilde{u}_{t,\mathcal{S}_t} = K_{t,\mathcal{S}} x_t,\ \forall t \in \{0,1,\dots,T-1\}, \tag{4}$$

where the control gain matrix $K_{t,\mathcal{S}} \in \mathbb{R}^{m_{\mathcal{S}_t}\times n}$, with $m_{\mathcal{S}_t} = \sum_{i\in\mathcal{S}_t} m_i$, is given by

$$K_{t,\mathcal{S}} = -(B_{\mathcal{S}}^\top P_{t+1,\mathcal{S}} B_{\mathcal{S}} + R_{\mathcal{S}})^{-1} B_{\mathcal{S}_t}^\top P_{t+1,\mathcal{S}} A, \tag{5}$$

where $P_{t,\mathcal{S}} \in \mathbb{S}^n_+$ is given recursively by the following Discrete Algebraic Riccati Equation (DARE):

$$\begin{aligned} P_{t,\mathcal{S}} = {} & A^\top P_{t+1,\mathcal{S}} A - A^\top P_{t+1,\mathcal{S}} B_{\mathcal{S}_t} \\ & \times (B_{\mathcal{S}_t}^\top P_{t+1,\mathcal{S}} B_{\mathcal{S}_t} + R_{\mathcal{S}_t})^{-1} B_{\mathcal{S}_t}^\top P_{t+1,\mathcal{S}} A + Q \end{aligned} \tag{6}$$

[1] In other words, the matrix $R_{\mathcal{S}_t}$ is obtained by deleting the rows and columns of R indexed by the elements in the set $\mathcal{G} \setminus \mathcal{S}_t$.

initialized with $P_{T,\mathcal{S}} = Q_f$. Moreover, conditioning on an actuator selection $\mathcal{S}$, we know that [36, Chapter 3]

$$J(\mathcal{S}) \triangleq \min_{u_{\mathcal{S}}} \mathbb{E}\big[J(\mathcal{S}, u_{\mathcal{S}})\big] = \mathbb{E}\big[J(\mathcal{S}, \tilde{u}_{\mathcal{S}})\big] = \mathbb{E}[x_0^\top P_{0,\mathcal{S}} x_0] + \sum_{t=0}^{T-1} \mathrm{Tr}(P_{t+1,\mathcal{S}} W). \quad (7)$$

Thus, supposing the system matrices are known, we see that solving Problem (3) is equivalent to solving

$$\begin{aligned} &\min_{\mathcal{S}} J(\mathcal{S}) \\ &s.t.\ \mathcal{S}_t \subseteq \mathcal{G}, |\mathcal{S}_t| = H,\ \forall t \in \{0, \dots, T-1\}. \end{aligned} \quad (8)$$

When the system matrices A and B are unknown, Eqs. (4)-(6) cannot be directly used to design the control $u_{\mathcal{S}}$ conditioning on an actuator selection $\mathcal{S} = (\mathcal{S}_1, \dots, \mathcal{S}_{T-1})$. We now define and solve both the episodic and non-episodic settings of Problem (3). In the sequel, we use superscript k to index an episode and subscript t to index a time step.

B. Online Algorithm for Episodic Setting

In the episodic setting, system (1) starts from an initial condition at the beginning of each episode, and we aim to obtain a solution to Problem (3) by interacting with system (1) for a set of episodes. Specifically, let $N \in \mathbb{Z}_{\geq 1}$ be the total number of episodes and let be the time horizon length of any episode $k \in [N]$. Considering any $k \in [N]$, the dynamics of system (1) in episode k is given by

$$x_{t+1}^k = A x_t^k + B_{\mathcal{S}_t^k} u_{t,\mathcal{S}_t^k}^k + w_t^k, \quad (9)$$

where x_t^k, $u_{t,\mathcal{S}^k}^k$ and w_t^k are the state, control and noise at time step t in episode k, respectively, and $\mathcal{S}^k = (\mathcal{S}_0^k, \dots, \mathcal{S}_{T-1}^k)$ with S_t^k to be the set of actuators selected for time step t, for all $t \in \{0, \dots, T-1\}$. We assume that $\{w_t^k\}_{t=0}^{T-1}$ are i.i.d with $\mathbb{E}[w_t^k] = 0$ and $\mathbb{E}[w_t^k w_t^{k\top}] = W$ for all $t \in \{0, 1, \dots, T-1\}$ and for all $k \in [N]$. We also assume for simplicity that $x_0^k = 0$ for all $k \in [N]$. In this work, we focus on the scenario with $\mathcal{S}_0^k = \cdots = \mathcal{S}_{T-1}^k$ for all $k \in [K]$, i.e., the set of selected actuators in each episode is fixed during that episode. Slightly abusing the notation, we simply denote the set of selected actuators for episode k as $\mathcal{S}^k \subseteq \mathcal{G}$.

Now, similarly to Eq. (2), for any $k \in [N]$ we define the following quadratic cost of episode k when the set of actuators $\mathcal{S}^k \subseteq \mathcal{G}$ is selected to provide $u_{t,\mathcal{S}^k}$ for all $t \in \{0, \dots, T-1\}$:

$$J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k) = \Big(\sum_{t=0}^{T-1} x_t^{k\top} Q^k x_t^k + u_{t,\mathcal{S}^k}^{k\top} R_{\mathcal{S}^k}^k u_{t,\mathcal{S}^k}^k \Big) + x_T^{k\top} Q_f^k x_T^k, \quad (10)$$

where $u_{\mathcal{S}^k}^k = (u_{0,\mathcal{S}^k}^k, \dots, u_{T-1,\mathcal{S}^k}^k)$, $Q^k, Q_f^k \in \mathbb{S}_+^n, R \in \mathbb{S}_{++}^m$ are the cost matrices, and $R_{\mathcal{S}^k}^k \in \mathbb{S}_{++}^{m_{\mathcal{S}^k}}$ (with $m_{\mathcal{S}^k} = \sum_{i \in \mathcal{S}^k} m_i$) is a submatrix of R^k corresponding to the set $\mathcal{S}^k$. Note that we allow different cost matrices across the episodes. We assume that Q^k, Q_f^k and R^k are known for all $k \in [N]$.

At the beginning of each episode $k \in [N]$, an online algorithm for Problem (3) selects a set $S^k \subseteq \mathcal{G}$ (with $|S^k| = H$) of actuators and designs the control $u_{\mathcal{S}^k} = (u_{0,\mathcal{S}^k}, \dots, u_{T-1,\mathcal{S}^k})$ provided by the actuators in $\mathcal{S}^k$. Note that when making the decisions at the beginning of any $k \in [N]$, the following information is available to the online algorithm: (a) the system state trajectories $x^1, \dots, x^{k-1}$, where $x^{k'} \triangleq (x_0^{k'}, \dots, x_{T-1}^{k'})$ for all $k' \in [k-1]$; and (b) previous decisions made by the algorithm, i.e., $\mathcal{S}^1, \dots, \mathcal{S}^{k-1}$ and $u_{\mathcal{S}^1}, \dots, u_{\mathcal{S}^{k-1}}$. Since Q^k, Q_f^k, R^k are assumed to be known, the costs $J_{k'}(\mathcal{S}^{k'}, u_{\mathcal{S}^{k'}}^{k'})$ $\forall k' \in [k-1]$ are also given at the beginning of any episode $k \in [N]$. Thus, the information setting discussed above corresponds to the bandit information setting in online optimization literature (see., e.g., [15]). To characterize the performance of such an online algorithm, denoted as $\mathcal{A}_e$, for Problem (3) in the episodic setting, we aim to minimize the following regret of $\mathcal{A}_e$:

$$R_{\mathcal{A}_e} \triangleq \mathbb{E}_{\mathcal{A}_e}\Big[\sum_{k=1}^{N} J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k) \Big] - \sum_{k=1}^{N} J_k(\mathcal{S}_\star^k), \quad (11)$$

where $\mathbb{E}_{\mathcal{A}_e}[\cdot]$ denotes the expectation with respect to the randomness of the algorithm, $J_k(\mathcal{S}_\star^k)$ is defined as (7) and $\mathcal{S}_\star^k$ is an optimal solution to (8) (with cost matrices Q^k, R^k, Q_f^k and an extra constraint $\mathcal{S}_0 = \cdots = \mathcal{S}_{T-1}$), for all $k \in [N]$.

Remark 1. *Note that $R_{\mathcal{A}_e}$ compares the cost incurred by the online algorithm to the benchmark given by the minimum achievable cost of Problem (3) in the episodic setting. Since $\mathcal{S}_\star^k$ can potentially be different across the episodes in the benchmark in Eq. (11), $R_{\mathcal{A}_e}$ is a dynamic regret [14], [35], [40]. In fact, one can consider any sets $\mathcal{S}_\star^1, \dots, \mathcal{S}_\star^N \subseteq \mathcal{G}$ with $|\mathcal{S}_\star^k| = H$ for all $k \in [N]$ as the benchmark in Eq. (11). For any $\mathcal{S}_\star = (\mathcal{S}_\star^1, \dots, \mathcal{S}_\star^N)$, define*

$$h((\mathcal{S}_\star^1, \dots, \mathcal{S}_\star^N)) = 1 + |\{1 \leq \ell < N-1 : \mathcal{S}_\star^\ell \neq \mathcal{S}_\star^{\ell+1}\}|. \quad (12)$$

Our regret bound for $R_{\mathcal{A}_e}$ holds for a general benchmark $\mathcal{S}_\star = (\mathcal{S}_\star^1, \dots, \mathcal{S}_\star^N)$, where $\mathcal{S}_\star^k$ is any $\mathcal{S}_\star^k \subseteq \mathcal{G}$ with $|\mathcal{S}_\star^k| = H$. If we consider benchmark $\mathcal{S}_\star$ with $h(\mathcal{S}_\star) = 1$, i.e., $\mathcal{S}_\star^1 = \cdots = \mathcal{S}_\star^N$, Eq. (11) reduces to a static regret [13], [35].

C. Online Algorithm for Non-Episodic Setting

In the non-episodic (i.e., continuous) setting, we interact with system (1) over a horizon of length $T \in \mathbb{Z}_{\geq 1}$, where the system is not reset to the initial condition $x_0 = 0$ during the interaction. At the beginning of each time step $t \in \{0, \dots, T-1\}$, the algorithm selects a set $\mathcal{S}_t \subseteq \mathcal{G}$ (with $|\mathcal{S}_t| = H$) of actuators and designs the corresponding control $u_{t,\mathcal{S}_t}$, using the following information available: (a) $x_0, \dots, x_{t-1}$; and (b) $\mathcal{S}_0, \dots, \mathcal{S}_{t-1}$ and $u_{0,\mathcal{S}_0}, \dots, u_{t-1,\mathcal{S}_{t-1}}$. Similar to our arguments above, the information setting corresponds to the bandit setting in the online optimization literature. Denote $\mathcal{S}_{0:t} \triangleq (\mathcal{S}_0, \dots, \mathcal{S}_t)$ and $u_{\mathcal{S}_{0:t}} \triangleq (u_{0,\mathcal{S}_0}, \dots, u_{t,\mathcal{S}_t})$ for all $t \in \{0, \dots, T-1\}$. To characterize the performance of such an online algorithm, denoted as $\mathcal{A}_c$, we minimize the following regret of $\mathcal{A}_c$:

$$R_{\mathcal{A}_c} \triangleq \mathbb{E}_{\mathcal{A}_c}\Big[\sum_{t=0}^{T-1} c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}}) \Big] - J(\mathcal{S}^\star), \quad (13)$$

where $\mathbb{E}_{\mathcal{A}_c}[\cdot]$ denotes the expectation with respect to the randomness of the algorithm, and the benchmark $J(\mathcal{S}^\star)$ is

defined as (7) with $\mathcal{S}^\star = (\mathcal{S}_0^\star, \ldots, \mathcal{S}_{T-1}^\star)$ to be an optimal solution to (8). Note that in Eq. (13) we denote,

$$c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}}) = x_t^\top Q x_t + u_{t,\mathcal{S}_t}^\top R_{\mathcal{S}_t} u_{t,\mathcal{S}_t}, \tag{14}$$

for all $t \in \{0, \ldots, T-2\}$, and

$$c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}}) = x_t^\top Q x_t + u_{t,\mathcal{S}_t}^\top R_{\mathcal{S}_t} u_{t,\mathcal{S}_t} + x_{t+1}^\top Q_f x_{t+1}, \tag{15}$$

for $t = T-1$. Similarly, one can consider a general benchmark $\mathcal{S}^\star$ in Eq. (13) as we described in Remark 1.

III. Algorithm Design for Episodic Setting

We now design an online algorithm for the episodic setting of Problem (3). In our algorithm design, we leverage an algorithm for the multiarmed bandit problem (i.e., the **Exp3.S** algorithm) [35] to select a set $\mathcal{S}^k \subseteq \mathcal{G}$ (with $\mathcal{S}^k = H$) for all k. Given the set $\mathcal{S}^k$ of selected actuators, we then leverage a certainty equivalence approach [30], [31] to design the corresponding control $u_{\mathcal{S}^k}$.

A. Exp3.S Algorithm for Multiarmed Bandit

The MultiArmed Bandit (MAB) problem is specified by a number of episodes N_s, a finite set $\mathcal{Q}$ of possible actions (i.e., arms), and costs of actions $y^1, \ldots, y^{N_s}$ with $y^k = (y_1^k, \ldots, y_{|\mathcal{Q}|}^k)$ for all $k \in [N_s]$, where $\mathcal{Q} = \{1, \ldots, |\mathcal{Q}|\}$ and $y_i^k \in [y_a, y_b]$ (with $y_a, y_b \in \mathbb{R}$) denotes the cost of choosing action i in episode k, for all $k \in [N_s]$ and for all $i \in \mathcal{Q}$. At the beginning of each episode $k \in [N_s]$, one can choose an action from the set $\mathcal{Q}$. Choosing $i_k \in \mathcal{Q}$ for episode $k \in [N_s]$ incurs a cost $y_{i_k}^k$, which is revealed at the end of episode k. To minimize the accumulative cost over the N_s episodes, an online algorithm $\mathcal{A}_M$ chooses action $i_k \in \mathcal{Q}$ for each episode $k \in [N_s]$, where the decision is made based on $i_{k'}$ and $y_{i_{k'}}^{k'}$ for all $k' \in \{1, \ldots, k-1\}$. For any sequence of actions, i.e., $j^{N_s} \triangleq (j_1, \ldots, j_{N_s})$, denote $h(j^{N_s}) = 1 + |\{1 \le k < N_s : j_k \neq j_{k+1}\}|$. We introduce the **Exp3.S** algorithm from [35].

Algorithm 1: Exp3.S

Input: Candidate set $\mathcal{Q}$, total number of episode N_s, parameters $\alpha_1 \in (0,1)$ and $\alpha_2 > 0$.

1 Initialize $\varpi_i^1 = 1$, $\forall i \in [|\mathcal{Q}|]$.

2 **for** $k = 1$ *to* N_s **do**

3 Set $q_i^k = (1-\alpha_1)\frac{\varpi_i^k}{\sum_{j=1}^{|\mathcal{Q}|}\varpi_j^k} + \frac{\alpha_1}{|\mathcal{Q}|}$, $\forall i \in [|\mathcal{Q}|]$.

4 Draw $i_k \in \mathcal{Q}$ according to the probabilities $q_1^k, \ldots, q_{|\mathcal{Q}|}^k$, receive cost $y_{i_k}^k \in [y_a, y_b]$, and normalize $y_{i_k}^k = (y_{i_k}^k - y_a)/(y_b - y_a)$.

5 **for** $j = 1, \ldots, |\mathcal{Q}|$ **do**

6 Set $\hat{y}_j^k = \begin{cases} y_j^k/q_j^k \text{ if} j = i_k, \\ 0 \text{ otherwise,} \end{cases}$

$\varpi_j^{k+1} = \varpi_j^k \exp\left(\frac{\alpha_1 \hat{y}_j^k}{|\mathcal{Q}|}\right) + \frac{e\alpha_2}{|\mathcal{Q}|}\sum_{i=1}^{|\mathcal{Q}|}\varpi_i^k$.

Lemma 1. *[35, Corollary 8.2] Consider any sequence $j^{N_s} = (j_1, \ldots, j_{N_s})$. In Algorithm 1, let $\alpha_2 = 1/N_s$ and*

$$\alpha_1 = \min\left\{1, \sqrt{\frac{|\mathcal{Q}|(h(j^{N_s})\ln(|\mathcal{Q}|N_s) + e)}{(e-1)N_s}}\right\}.$$

Let $\mathbb{E}_M[\cdot]$ denote the expectation with respective to the randomness in the algorithm. Then, we have

$$R_M(j^{N_s}) \triangleq \mathbb{E}_M\left[\sum_{k=1}^{N_s} y_{i_k}^k\right] - \sum_{k=1}^{N_s} y_{j_k}^k$$
$$\le 2(y_b - y_a)\sqrt{e-1}\sqrt{|\mathcal{Q}|N_s(h(j^{N_s})\ln(|\mathcal{Q}|N_s) + e)}. \tag{16}$$

Remark 2. *As argued in [14], [35], the regret bound in (16) holds under the assumption that for any $k \in [N_s]$, $y_{i_k}^k$ does not depend on the previous actions $i_1, , \ldots, i_{k-1}$ chosen by the* **Exp3.S** *algorithm. Other than this assumption, y_i^k can be any real number in $[y_a, y_b]$, and no statistical assumption is made on y_i^k. Also note that the random choices $i_1, \ldots, i_{N_s}$ in line 3 in Algorithm 1 ensure that in each episode $k \in [N_s]$, with some probability, the algorithm explores a new action or commits to the action that gives the lowest cost so far. Lemma 1 shows that such choices of $i_1, \ldots, i_{N_s}$ yield sublinear regret in N_s against an arbitrary benchmark j^{N_s}.*

B. Certainty Equivalence Approach

In this subsection, we assume that a set $\mathcal{S} \subseteq \mathcal{G}$ (with $|\mathcal{S}| = H$) of actuators is selected and fixed for episode $k \in [N]$. We now describe our design of the corresponding control $u_{\mathcal{S}}^k = (u_{0,\mathcal{S}}^k, \ldots, u_{T-1,\mathcal{S}})$, based on the certainty equivalence approach [30], [31]. First, conditioning on the set $\mathcal{S}$ of selected actuators for episode $k \in [N]$, Eq. (4) states that the corresponding optimal control is the linear state-feedback control given by $\tilde{u}_{t,\mathcal{S}}^k = K_{t,\mathcal{S}}^k x_t^k$ for all $t \in \{0, \ldots, T-1\}$, where the control gain matrix $K_{t,\mathcal{S}}^k$ is obtained from Eq. (5) (using the cost matrices Q_k and R_k in episode $k \in [N]$). Since the system matrices A and B are unknown, the certainty equivalence approach leverages estimates of the system matrices, denoted as $\hat{A}$ and $\hat{B}$,[2] in order to compute the control gain matrix [30], [31]. For any $t \in \{0, 1, \ldots, T-1\}$, the certainty equivalent controller for the kth episode of Problem (3) is

$$u_{t,\mathcal{S}}^k = \hat{K}_{t,\mathcal{S}}^k x_t^k, \tag{17}$$
$$\hat{K}_{t,\mathcal{S}}^k = -\big(\hat{B}_{\mathcal{S}}^\top \hat{P}_{t+1,\mathcal{S}} \hat{B}_{\mathcal{S}} + R_{\mathcal{S}}^k\big)^{-1} \hat{B}_{\mathcal{S}}^\top \hat{P}_{k+1,\mathcal{S}} \hat{A} \tag{18}$$
$$\hat{P}_{t,\mathcal{S}}^k = Q^k + \hat{A}^\top \hat{P}_{t+1,\mathcal{S}}^k \hat{A} - \hat{A}^\top \hat{P}_{t+1,\mathcal{S}}^k \hat{B}_{\mathcal{S}} \times \big(\hat{B}_{\mathcal{S}}^\top \hat{P}_{t+1,\mathcal{S}}^k \hat{B}_{\mathcal{S}} + R_{\mathcal{S}}^k\big)^{-1} \hat{B}_{\mathcal{S}}^\top \hat{P}_{t+1,\mathcal{S}}^k \hat{A}, \tag{19}$$

where $\hat{P}_{t,\mathcal{S}}^k \in \mathbb{S}_+^n$ and (19) is initialized with $\hat{P}_{T,\mathcal{S}}^k = Q_f^k$.

Next, we characterize the performance of the resulting certainty equivalent controller, which naturally depends on the estimation errors $\|\hat{A} - A\|$ and $\|\hat{B} - B\|$. Similarly to (7), we denote the expected cost corresponding to $\mathcal{S}$ and $u_{t,\mathcal{S}}^k = \hat{K}_{t,\mathcal{S}}^k x_t^k$ for the kth episode as

$$\hat{J}_k(\mathcal{S}) = \mathbb{E}[J_k(\mathcal{S}, u_{\mathcal{S}}^k)], \tag{20}$$

[2]The estimates $\hat{A}$ and $\hat{B}$ are obtained by some system identification method, using data samples from the system trajectory; we will elaborate more on the system identification part later.

where $J_k(\mathcal{S}, u_{\mathcal{S}}^k)$ is defined in Eq. (10). One can show that the following expression for $\hat{J}_k(\mathcal{S})$ holds [36, Chapter 3]:

$$\hat{J}_k(\mathcal{S}) = \mathbb{E}\big[x_0^{k\top} \tilde{P}_{0,\mathcal{S}}^k x_0^k\big] + \sum_{t=0}^{T-1} \mathrm{Tr}(\tilde{P}_{t+1,\mathcal{S}}^k W), \tag{21}$$

where $\tilde{P}_{t,\mathcal{S}}^k$ satisfies the following recursion with $\tilde{P}_{T,\mathcal{S}}^k = Q_f^k$

$$\begin{aligned}\tilde{P}_{t,\mathcal{S}}^k = Q^k + \hat{K}_{t,\mathcal{S}}^{k\top} R_{\mathcal{S}}^k \hat{K}_{t,\mathcal{S}}^k \\ + (A + B_{\mathcal{S}} \hat{K}_{k,\mathcal{S}})^{k\top} P_{t+1,\mathcal{S}}^k (A + B_{\mathcal{S}} \hat{K}_{t,\mathcal{S}}^k).\end{aligned} \tag{22}$$

We now upper bound $\hat{J}_k(\mathcal{S}) - J_k(\mathcal{S})$ in terms of the estimation error in $\hat{A}$ and $\hat{B}$, where $J_k(\mathcal{S})$ is defined in Eq. (7). Note that both the optimal controller $K_{t,\mathcal{S}}^k$ given by Eq. (4) and the certainty equivalent controller $\hat{K}_{t,\mathcal{S}}^k$ given by Eq. (18) are time-varying for the finite-horizon setting. In contrast, both the optimal controller [36] and the certainty equivalent controller proposed in [30], [31] are time-invariant for the infinite-horizon setting, which are obtained from steady-state solutions to DAREs. Hence, our analysis for the certainty equivalence approach for learning finite-horizon LQR will be more challenging than that in [30], [31] for learning infinite-horizon LQR. To proceed, supposing the estimation error satisfies that $\|A-\hat{A}\| \le \varepsilon$ and $\|B-\hat{B}\| \le \varepsilon$ with $\varepsilon \in \mathbb{R}_{>0}$, we provide upper bounds on $\|K_{t,\mathcal{S}}^k - \hat{K}_{t,\mathcal{S}}^k\|$ and $\|P_{t,\mathcal{S}}^k - \hat{P}_{t,\mathcal{S}}^k\|$, where $P_{t,\mathcal{S}}^k$ (resp., $\hat{P}_{t,\mathcal{S}}^k$) is given by Eq. (6) (resp., Eq. (19)). We need the following mild assumption.

Assumption 1. *We assume that $\sigma_n(Q^k) \ge 1$ and $\sigma_m(R^k) \ge 1$ for all $k \in [N]$.*

In order to simplify the notations in the sequel, we denote

$$\Gamma_{\mathcal{S}} = \max_{k\in[N], t\in[T]} \Gamma_{t,\mathcal{S}}^k, \tag{23}$$

$$\tilde{\Gamma}_{\mathcal{S}} = 1 + \Gamma_{\mathcal{S}}, \tag{24}$$

where $\Gamma_{t,\mathcal{S}}^k = \max\big\{\|A\|, \|B\|, \|P_{t,\mathcal{S}}^k\|, \|K_{t-1,\mathcal{S}}^k\|\big\}$. Moreover, we denote

$$\begin{aligned}\sigma_Q &= \max\Big\{\max_{k\in[N]} \sigma_1(Q^k), \max_{k\in[N]} \sigma_1(Q_f^k)\Big\}, \\ \sigma_R &= \max_{k\in[N]} \sigma_1(R^k).\end{aligned} \tag{25}$$

We have the following result; the proof can be found in [39].

Lemma 2. *Consider any $\mathcal{S} \subseteq \mathcal{G}$, any $k \in [N]$ and any $t \in [T]$. Let $\varepsilon \in \mathbb{R}_{\ge 0}$ and $D \in \mathbb{R}_{\ge 0}$ with $\varepsilon \le 1$ and $D \ge 1$. Suppose that $\|A-\hat{A}\| \le \varepsilon$, $\|B_{\mathcal{S}} - \hat{B}_{\mathcal{S}}\| \le \varepsilon$, and $\|P_{t,\mathcal{S}}^k - \hat{P}_{t,\mathcal{S}}^k\| \le D\varepsilon$, and that Assumption 1 holds. Then,*

$$\|K_{t-1,\mathcal{S}}^k - \hat{K}_{t-1,\mathcal{S}}^k\| \le 3\tilde{\Gamma}_{\mathcal{S}}^3 D\varepsilon, \tag{26}$$

$$\|P_{t-1,\mathcal{S}}^k - \hat{P}_{t-1,\mathcal{S}}^k\| \le 44\tilde{\Gamma}_{\mathcal{S}}^9 \sigma_R D\varepsilon. \tag{27}$$

We make the following assumption on the controllability of the pair (A, B) similar to [30], [41], [42].

Assumption 2. *For any $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$, we assume that the pair $(A, B_{\mathcal{S}})$ in system (1) satisfies that $\sigma_1(\mathcal{C}_{\ell,\mathcal{S}}) \ge \nu$, where $\ell \in [n-1]$, $\nu \in \mathbb{R}_{>0}$ and $\mathcal{C}_{\ell,\mathcal{S}} \triangleq \big[B_{\mathcal{S}} \quad AB_{\mathcal{S}} \quad \cdots \quad A^{\ell-1}B_{\mathcal{S}}\big]$.*

If Assumption 2 is satisfied, we say that $(A, B_{\mathcal{S}})$ is (ℓ, ν)-controllable [30]. Note that if $(A, B_{\mathcal{S}})$ is controllable, $(A, B_{\mathcal{S}})$ can be (ℓ, ν)-controllable for some $\ell \in [n-1]$ that is much smaller than n. One can also check that a sufficient condition for Assumption 2 to hold is that for any actuator $s \in \mathcal{G}$, the pair (A, B_s) is (ℓ, ν)-controllable. Denoting

$$\hat{\mathcal{C}}_{\ell,\mathcal{S}} = \big[\hat{B}_{\mathcal{S}} \quad \hat{A}\hat{B}_{\mathcal{S}} \quad \cdots \quad \hat{A}^{\ell-1}\hat{B}_{\mathcal{S}}\big] \quad \forall \mathcal{S} \subseteq \mathcal{G},$$

we have the following lower bound on $\sigma_n(\hat{\mathcal{C}}_{\ell,\mathcal{S}})$.

Lemma 3. *[30, Lemma 6] Consider any $\mathcal{S} \subseteq \mathcal{G}$. Suppose that $\|A-\hat{A}\| \le \varepsilon$ and $\|B_{\mathcal{S}} - \hat{B}_{\mathcal{S}}\| \le \varepsilon$, where $\varepsilon \in \mathbb{R}_{\ge 0}$. Under Assumption 2, $\sigma_n(\hat{\mathcal{C}}_{\ell,\mathcal{S}}) \ge \nu - \varepsilon \ell^{\frac{3}{2}} \beta^{\ell-1}(\|B_{\mathcal{S}}\| + 1)$, where $\beta \triangleq \max\{1, \varepsilon + \|A\|\}$.*

Lemma 3 states that if ε is small enough, then $\sigma_n(\hat{\mathcal{C}}_{\ell,\mathcal{S}}) > 0$, i.e., $\mathrm{rank}(\hat{\mathcal{C}}_{\ell,\mathcal{S}}) = n$ and the pair $(\hat{A}, \hat{B}_{\mathcal{S}})$ is controllable. We have the following result proved in Appendix A.

Lemma 4. *Consider any $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$ and any $k \in [N]$. If Assumptions 1-2 hold, $\|A-\hat{A}\| \le \varepsilon$ and $\|B_{\mathcal{S}} - \hat{B}_{\mathcal{S}}\| \le \varepsilon$, where $\varepsilon \in \mathbb{R}_{\ge 0}$, then, for any $t \in \{T - \gamma\ell : \gamma \in \mathbb{Z}_{\ge 0}, \gamma\ell \le T\}$, and with $\beta = \max\{1, \varepsilon + \|A\|\}$, it holds that*

$$\|P_{t,\mathcal{S}}^k - \hat{P}_{t,\mathcal{S}}^k\| \le \mu_{t,\mathcal{S}}^k \varepsilon, \tag{28}$$

under the assumption that $\mu_{t,\mathcal{S}}^k \varepsilon \le 1$ with

$$\begin{aligned}\mu_{t,\mathcal{S}}^k \triangleq 32\ell^{\frac{5}{2}} \beta^{2(\ell-1)}(1+\nu^{-1})(1+\|B_{\mathcal{S}}\|)^2 \\ \times \|P_{t,\mathcal{S}}^k\| \max\{\sigma_Q, \sigma_R\}.\end{aligned} \tag{29}$$

Let us further denote

$$\mu_{\mathcal{S}} = 32\ell^{\frac{5}{2}} \tilde{\beta}^{2(\ell-1)}(1+\nu^{-1})\tilde{\Gamma}_{\mathcal{S}}^3 \max\{\sigma_Q, \sigma_R\}, \tag{30}$$

where $\tilde{\beta} = 1 + \|A\|$. Now, combining Lemmas 2 and 4 yields the following result, which upper bounds $\|K_{t,\mathcal{S}}^k - \hat{K}_{t,\mathcal{S}}^k\|$ and $\|P_{t,\mathcal{S}}^k - \hat{P}_{t,\mathcal{S}}^k\|$ for all t; the proof can be found in [39].

Proposition 1. *Consider any $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$. Suppose that Assumptions 1-2 hold, and that $\|A-\hat{A}\| \le \varepsilon$, $\|B_{\mathcal{S}} - \hat{B}_{\mathcal{S}}\| \le \varepsilon$, where $\varepsilon \in \mathbb{R}_{\ge 0}$ and $\mu_{\mathcal{S}}\varepsilon \le 1$. Then, for any $t \in \{0, 1, \ldots, T\}$, it holds that*

$$\|P_{t,\mathcal{S}}^k - \hat{P}_{t,\mathcal{S}}^k\| \le (44\tilde{\Gamma}_{\mathcal{S}}^9 \sigma_R)^{\ell-1} \mu_{\mathcal{S}} \varepsilon, \tag{31}$$

Moreover, for any $t \in \{0, 1, \ldots, T-1\}$, it holds that

$$\|K_{t,\mathcal{S}}^k - \hat{K}_{t,\mathcal{S}}^k\| \le 3\tilde{\Gamma}_{\mathcal{S}}^3 (44\tilde{\Gamma}_{\mathcal{S}}^9 \sigma_R)^{\ell-1} \mu_{\mathcal{S}} \varepsilon. \tag{32}$$

We are now in place to upper bound $\hat{J}_k(\mathcal{S}) - J_k(\mathcal{S})$. We begin with the following result; the proof can be found in [39].

Lemma 5. *Consider any $\mathcal{S} \subseteq \mathcal{G}$ and any $k \in [N]$. Let x_t^k be the state corresponding to the certainty equivalence control $u_{t,\mathcal{S}}^k = \hat{K}_{t,\mathcal{S}}^k x_t^k$, i.e., $x_{t+1}^k = (A + B_{\mathcal{S}}\hat{K}_{t,\mathcal{S}}^k)x_t^k + w_t^k$, where w_t^k is the zero-mean white Gaussian noise process with covariance W for all k. Let $\Delta K_{t,\mathcal{S}}^k \triangleq \hat{K}_{t,\mathcal{S}}^k - K_{t,\mathcal{S}}^k$. Then,*

$$\begin{aligned}\hat{J}_k(\mathcal{S}) - J_k(\mathcal{S}) = \sum_{t=0}^{T-1} \mathbb{E}\Big[x_t^{k\top} \Delta\hat{K}_{t,\mathcal{S}}^{k\top}(R_{\mathcal{S}}^k + B_{\mathcal{S}}^\top P_{t,\mathcal{S}} B_{\mathcal{S}}) \\ \times \Delta\hat{K}_{t,\mathcal{S}}^k x_t^k\Big].\end{aligned} \tag{33}$$

To proceed, consider any $\mathcal{S} \subseteq \mathcal{G}$ and any $k \in [N]$. For any $t_1, t_2 \in \{0, 1, \dots, T\}$ with $t_2 \geq t_1$, we use $\Psi^k_{t_2,t_1}(\mathcal{S})$ to denote the state transition matrix corresponding to $A + B_{\mathcal{S}} K^k_{t,\mathcal{S}}$, i.e.,

$$\Psi^k_{t_2,t_1}(\mathcal{S}) = (A + B_{\mathcal{S}} K^k_{t_2-1,\mathcal{S}})(A + B_{\mathcal{S}} K^k_{t_2-2,\mathcal{S}}) \cdots \times (A + B_{\mathcal{S}} K^k_{t_1,\mathcal{S}}), \quad (34)$$

and $\Psi^k_{t_2,t_1}(\mathcal{S}) \triangleq I$ if $t_1 = t_2$, where $K^k_{t,\mathcal{S}}$ is given by Eq. (5). Similarly, we denote

$$\hat{\Psi}^k_{t_2,t_1}(\mathcal{S}) = (A + B_{\mathcal{S}} \hat{K}^k_{t_2-1,\mathcal{S}}) \times \cdots \times (A + B_{\mathcal{S}} \hat{K}^k_{t_1,\mathcal{S}}), \quad (35)$$

and $\hat{\Psi}^k_{t_2,t_1}(\mathcal{S}) \triangleq I$ if $t_1 = t_2$, where $\hat{K}^k_{t,\mathcal{S}}$ is given by Eq. (18). One can now prove the following result, which shows that the state transition matrix $\Psi^k_{t_2,t_1}(\mathcal{S})$ is exponentially stable; a proof of the result can be found in [43], [44].

Lemma 6. *Consider any $S \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$ and any $k \in [N]$. Supposing Assumptions 1-2 hold. Then, there exist finite constants $\zeta_{\mathcal{S}} \in \mathbb{R}_{\geq 1}$ and $0 < \eta_{\mathcal{S}} < 1$ such that $\|\Psi^k_{t_2,t_1}(\mathcal{S})\| \leq \zeta_{\mathcal{S}} \eta_{\mathcal{S}}^{t_2-t_1}$ for all $t_1, t_2 \in \{0, 1, \dots, T\}$ with $t_2 \geq t_1$.*

The following result characterizes the stability of $\hat{\Psi}^k_{t_2,t_1}(\mathcal{S})$; the proof follows from Lemma 6 and can be found in [39].

Lemma 7. *Consider any $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$ and any $t \in [T]$. Suppose Assumptions 1-2 hold, and $\|K^k_{t,\mathcal{S}} - \hat{K}^k_{t,\mathcal{S}}\| \leq \varepsilon$ for all $t \in \{0, 1, \dots, T-1\}$, where $\varepsilon \in \mathbb{R}_{>0}$. Then, for all $t_1, t_2 \in \{0, 1, \dots, T\}$ with $t_2 \geq t_1$, $\|\hat{\Psi}^k_{t_2,t_1}(\mathcal{S})\| \leq \zeta_{\mathcal{S}} (\frac{1+\eta_{\mathcal{S}}}{2})^{t_2-t_1}$, under the assumption that $\varepsilon \leq \frac{1-\eta_{\mathcal{S}}}{2\|B_{\mathcal{S}}\|\zeta_{\mathcal{S}}}$, where $\zeta_{\mathcal{S}} \geq 1$ and $0 < \eta_{\mathcal{S}} < 1$ are given by Lemma 6.*

Combining Lemmas 5 and 7, and Proposition 1 yields the following result; the proof is included can be found in [39].

Proposition 2. *Consider any $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$ and any $k \in [N]$. Suppose Assumptions 1-2 hold, and $\|A - \hat{A}\| \leq \varepsilon$ and $\|B - \hat{B}\| \leq \varepsilon$, where $\varepsilon \in \mathbb{R}_{>0}$. Then, it holds that*

$$\hat{J}_k(\mathcal{S}) - J_k(\mathcal{S}) \leq \frac{4 \min\{n, m_{\mathcal{S}}\} T \zeta_{\mathcal{S}}^2}{1 - \eta_{\mathcal{S}}^2} \sigma_1(W) \big(\sigma_R + \Gamma_{\mathcal{S}}^3\big) \times \big(3\tilde{\Gamma}_{\mathcal{S}}^3 (20 \tilde{\Gamma}_{\mathcal{S}}^9 \sigma_R)^{\ell-1} \mu_{\mathcal{S}}\big)^2 \varepsilon^2, \quad (36)$$

under the assumption that $\varepsilon \leq \frac{1-\eta_{\mathcal{S}}}{\zeta_{\mathcal{S}} \mu_{\mathcal{S}}}$, where $\zeta_{\mathcal{S}} \geq 1$ and $0 < \eta_{\mathcal{S}} < 1$ are given by Lemma 6, $J_k(\mathcal{S})$ and $\hat{J}_k(\mathcal{S})$ are defined in (7) and (20), respectively, and $m_{\mathcal{S}} = \sum_{i \in \mathcal{S}} m_i$.

Hence, supposing the estimation error of $\hat{A}, \hat{B}$ can be made small enough, Proposition 2 bounds the gap between the (expected) costs incurred by the certainty equivalent controller and the optimal controller that knows the system model A, B.

C. Overall Algorithm Design

We introduce the overall algorithm (Algorithm 2) for the episodic setting of Problem (3), under the assumptions below.

Assumption 3. *We assume that (a) for any $k \in [N]$, $\{w^k_t\}_{t=0}^{T-1}$ are i.i.d Gaussian with $\mathbb{E}[w^k_t] = 0$ and $\mathbb{E}[w^k_t w^{k\top}_t] = \sigma^2 I$, i.e., $w^k_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I)$, where $\sigma \in \mathbb{R}_{\geq 0}$ is known; (b) for any distinct $t_1, t_2 \in \{0, 1 \dots, T-1\}$ and any distinct $k_1, k_2 \in [N]$, the noise terms $w^{k_1}_{t_1}$ and $w^{k_2}_{t_2}$ are independent.*

Assumption 4. *There exist $\mathcal{G}_1, \dots, \mathcal{G}_p$ with $\mathcal{G}_i \subseteq \mathcal{G}$ and $|\mathcal{G}_i| = H$ for all $i \in [p]$ such that $\mathcal{G} = \cup_{i \in [p]} \mathcal{G}_i$ and there is a known stabilizing $K_{\mathcal{G}_i} \in \mathbb{R}^{m_{\mathcal{G}_i} \times n}$ with $\|(A + B_{\mathcal{G}_i} K_{\mathcal{G}_i})^t\| \leq \zeta_0 \eta_0^t$ and $\|K_{\mathcal{G}_i}\| \leq \zeta_0$, $\forall t \in \mathbb{R}_{\geq 0}$ and $\forall i \in [p]$, where $p = \lceil m/H \rceil$, $m_{\mathcal{G}_i} = \sum_{j \in \mathcal{G}_i} m_j$, $\zeta_0 \in \mathbb{R}_{\geq 1}$ and $\eta_0 \in \mathbb{R}_{>0}$ with $0 < \eta_0 < 1$.*

Algorithm 2: Episodic Setting

Input: Parameters $\tau_1, \lambda, N, T, \bar{y}_b$, and $K_{\mathcal{G}_j}$ for all $j \in [p]$ from Assumption 4.

Initialize $N_1 = 1$.
for $j = 1$ *to* p **do**
  Set $N_{j+1} \leftarrow N_j + \tau_1$.
/* System identification phase */
for $j = 1$ *to* p **do**
  for $k = N_j$ *to* $N_{j+1} - 1$ **do**
    Select $\mathcal{S}^k = \mathcal{G}_j$.
    Play $u^k_{\mathcal{G}_j}$ with $u^k_{t,\mathcal{G}_j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(K_{\mathcal{G}_j} x^k_t, 2\sigma^2 \eta_0^2 I)$, $\forall t \in \{0, \dots, T-1\}$.
  Obtain $\hat{\Theta}_{\mathcal{G}_j}$ from (37).
Obtain $\hat{A}$ by extracting the first n columns from $\hat{\Theta}_{\mathcal{G}_1}$; obtain $\hat{B}$ by extracting the last $m_{\mathcal{G}_j}$ columns from $\hat{\Theta}_{\mathcal{G}_j}$ for all $j \in [p]$ and merging them into $\hat{B}$.
/* Control phase */
Initialize an **Exp3.S** subroutine with $N_s = N - N_{p+1} + 1$, $\mathcal{Q} = \{\mathcal{S} \subseteq \mathcal{G} : |\mathcal{S}| = H\}$, and α_1, α_2 according to Lemma 1.[3]
for $k = N_{p+1}$ *to* N **do**
  Enter the $(k - N_{p+1} + 1)$th iteration of the for loop in lines 2-6 in **Exp3.S**; select $\mathcal{S}^k \in \mathcal{Q}$ according to the probabilities $q^k_1, \dots, q^k_{|\mathcal{Q}|}$.
  for $t = 0$ *to* $T - 1$ **do**
    Obtain $\hat{K}^k_{t,\mathcal{S}^k}$ using $\hat{A}, \hat{B}_{\mathcal{S}^k}$ via Eq. (18).
    Play $u^k_{t,\mathcal{S}^k} = \hat{K}^k_{t,\mathcal{S}^k} x^k_t$.
  Receive the cost $y^k_{\mathcal{S}^k} = J_k(\mathcal{S}, u^k_{\mathcal{S}})$; follow lines 4-6 in **Exp3.S** with $y_a = 0$ and $y_b = \bar{y}_b$.
Output: $\mathcal{S}^k, u^k_{\mathcal{S}^k} = (u^k_{0,\mathcal{S}^k}, \dots, u^k_{T-1,\mathcal{S}^k}), \forall k \in [N]$.

Assumption 3(a) ensures that the noise terms from different episodes are independent. Similarly to [30], [31], [42], assuming the noise covariance is $W = \sigma^2 I$ is only made to ease the presentation; our analysis in the remaining of this paper can be extended to w^k_t with general covariance matrix $W \in \mathbb{S}^n_{++}$, where the analysis will then depend on $\sigma_1(W)$ and $\sigma_n(W)$. Note that more general noise models are considered in, e.g., [32], [45], where $\{w^k_t\}_{t=0}^{T-1}$ can be non-stationary and non-Gaussian. We restrict ourselves to the i.i.d. Gaussian noise model of $\{w^k_t\}_{t=0}^{T-1}$ described in Assumption 3, and leave the extension to the more general noise models to future work.

Similar assumptions to Assumption 4 can also be found in [30], [31], [41], [42]. Note that under Assumption 2, the pair $(A, B_{\mathcal{G}_i})$ is controllable for all $i \in [p]$, which guarantees the existence of the $K_{\mathcal{G}_i}$ described in Assumption 4. Moreover,

[3] Note that $h(j^{N_s})$ in Lemma 1 is set to be $h(\mathcal{S}_\star)$ defined in Eq. (12).

the stability of $A+B_{\mathcal{G}_i}K_{\mathcal{G}_i}$ ensures via the Gelfand formula [46] that the finite constants $\zeta_0 \geq 1$ and $0<\eta_0<1$ exist, which may be computed by the LQR cost of the control $u_t = K_{\mathcal{G}_i}x_t$ [31]. Also note that Assumption 4 gives us a set of known stabilizing controllers that we can use in the system identification phase of Algorithm 2 (see our detailed descriptions below). In fact, using the techniques from [37], [45], [47], one can introduce an extra warm-up phase before the system identification phase in Algorithm 2, which learns a stabilizing $K_{\mathcal{G}_i}$ for all $i \in [p]$ from the system trajectory. In particular, as shown in [37], the extra warm-up phase will incur an extra additive factor $2^{O(n)}$ in the regret of Algorithm 2 defined in Eq. (11).

Now, we explain the steps in Algorithm 2.

System identification phase: In lines 4-9, Algorithm 2 computes estimates of A and B, denoted as $\hat{A}$ and $\hat{B}$, respectively. This is achieved by first iteratively selecting the sets $\mathcal{G}_1,\ldots,\mathcal{G}_p$ of actuators and playing the corresponding stabilizing controller given by Assumption 4 for τ_1 episodes. Formally, for any $j \in [p]$, Algorithm 2 selects $\mathcal{S}^k = \mathcal{G}_j$ and plays the control $u^k_{t,\mathcal{G}_j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(K_{\mathcal{G}_j}x^k_t, 2\sigma^2\eta_0^2 I)$ for all time steps $t \in \{0,\ldots,T-1\}$ and all episodes $k \in \{N_j,\ldots,N_{j+1}-1\}$ with $N_{j+1} = N_j + \tau_1$, where $\tau_1 \in \mathbb{Z}_{\geq 1}$ is an input to Algorithm 2 whose value will be specified later. We assume that $u^k_{t,\mathcal{G}_j}$ is independent of the noise $w^{k'}_{t'}$ for all $t' \in \{0,\ldots,T-1\}$ and all $k' \in [N]$. For any $j \in [p]$, the estimate $\hat{\Theta}_{\mathcal{G}_j} \in \mathbb{R}^{n\times(n+m_{\mathcal{G}_j})}$ (with $m_{\mathcal{G}_j} = \sum_{i\in\mathcal{G}_j} m_i$) is obtained by solving the following regularized least squares:[4]

$$\hat{\Theta}_{\mathcal{G}_j} \in \arg\min_{Y} \Big\{\lambda\|Y\|_F^2 + \sum_{k=N_j}^{N_{j+1}-1}\sum_{t=0}^{T-1}\|x^k_{t+1} - Y z^k_{t,\mathcal{G}_j}\|^2\Big\}, \tag{37}$$

where $\lambda \in \mathbb{R}_{>0}$ and

$$z^k_{t,\mathcal{S}} \triangleq \begin{bmatrix} x_t^{k\top} & u_{t,\mathcal{S}}^{k\top}\end{bmatrix}^\top, \tag{38}$$

for all $k \in [N]$, all $t \in \{0,\ldots,T-1\}$ and all $\mathcal{S} \subseteq \mathcal{G}$. For any $j \in [p]$, $\hat{\Theta}_{\mathcal{G}_j}$ can be viewed as an estimate of $\Theta_{\mathcal{G}_j} \triangleq \begin{bmatrix} A & B_{\mathcal{G}_j}\end{bmatrix}$ [31], [49]. Thus, we can obtain estimates of A and B, i.e., $\hat{A}$ and $\hat{B}$, respectively, according to line 9 in Algorithm 2.

Control phase: For any episode $k \in \{N_{p+1},\ldots,N\}$ in lines 11-16 of Algorithm 2, the algorithm calls the **Exp3.S** subroutine to select a set $\mathcal{S}^k$ of actuators, and invokes the certainty equivalence approach described in Section III-B to design $u^k_{t,\mathcal{S}^k} = \hat{K}^k_{t,\mathcal{S}^k}x^k_t$, $\forall t \in \{0,\ldots,T-1\}$, where $\hat{K}^k_{t,\mathcal{S}^k}$ is computed by Eq. (18) using the estimates $\hat{A}, \hat{B}$ obtained from the system identification phase. Here, the **Exp3.S** subroutine is applied to the MAB instance, where the total number of episodes is $N_s = N - N_{p+1} + 1$, the set of all possible actions is $\mathcal{Q} = \{\mathcal{S} \subseteq \mathcal{G} : |\mathcal{S}| = H\}$, and the cost associated with each possible action $\mathcal{S} \in \mathcal{Q}$ in episode k is $y^k_{\mathcal{S}} = J_k(\mathcal{S}, u^k_{\mathcal{S}})$ defined in Eq. (10), where $u^k_{\mathcal{S}} = (u^k_{0,\mathcal{S}},\ldots,u^k_{T-1,\mathcal{S}})$ with $u^k_{t,\mathcal{S}} = \hat{K}_{t,\mathcal{S}}x^k_t$. Thus, each arm in the MAB instance corresponds to a set of actuators with cardinality H.

Finally, one can check that the running time of each episode in Algorithm 2 is $O((n+m)^3T + |\mathcal{Q}|T)$, where the factor $(m+n)^3$ is due to the computation of Eq. (18). Since $|\mathcal{Q}| = \binom{|\mathcal{G}|}{H}$, Algorithm 2 is efficient for instances of Problem (3) with either $|\mathcal{G}|$ (i.e., the total number of candidate actuators) or H (i.e., the cardinality constraint on the set of selected actuators) to be small (or bounded by a constant). Nonetheless, we will later extend our algorithm design to efficiently handle large-scale instances of Problem (3) in Section VI.

Remark 3. *Note from Eq. (10) that the cost of the action chosen by the* **Exp3.S** *subroutine for any episode $k \in \{N_{p+1},\ldots,N\}$ of Algorithm 2, i.e., $J_k(\mathcal{S}^k, u^k_{\mathcal{S}^k})$, does not depend on the previous actions $\mathcal{S}^{N_{p+1}},\ldots,\mathcal{S}^{k-1}$ chosen by the* **Exp3.S** *subroutine. Thus, we know from Remark 2 that the result in Lemma 1 can be applied when we analyze the regret of Algorithm 2 in the next section.*

IV. Regret Analysis for Episodic Setting

In this section, we aim to provide high probability upper bounds on the regret of Algorithm 2 defined in Eq. (11) for the episodic setting of Problem (3). To this end, we first analyze the estimation error of the least squares approach given by (37). For any $j \in [p]$, we denote

$$V_{\mathcal{G}_j} = \lambda I + \sum_{k=N_j}^{N_{j+1}-1}\sum_{t=0}^{T-1} z^k_{t,\mathcal{G}_j} z^{k\top}_{t,\mathcal{G}_j}, \tag{39}$$

where $\mathcal{G}_j$ is given by Assumption 4, $\lambda \in \mathbb{R}_{>0}$, N_j, N_{j+1} are given in Algorithm 2, and $z^k_{t,\mathcal{G}_j}$ is given in Eq. (38). We then have the following result; the proof is similar to that of [41, Lemma 6] and is omitted here for conciseness.

Lemma 8. *Consider any $\mathcal{G}_j$ from Assumption 4, where $j \in [p]$. Let $\Delta_{\mathcal{G}_j} = \Theta_{\mathcal{G}_j} - \hat{\Theta}_{\mathcal{G}_j}$, where $\Theta_{\mathcal{G}_j} = \begin{bmatrix} A & B_{\mathcal{G}_j}\end{bmatrix}$ and $\hat{\Theta}_{\mathcal{G}_j}$ is given by (37). Suppose Assumption 3 holds. Then, for any $\delta \in \mathbb{R}$ with $0<\delta<1$, with probability at least $1-\delta$, it holds*

$$Tr(\Delta^\top_{\mathcal{G}_j} V_{\mathcal{G}_j}\Delta_{\mathcal{G}_j}) \leq 4\sigma^2 n \log\Big(\frac{n\det(V_{\mathcal{G}_j})}{\delta\det(\lambda I)}\Big) + 2\lambda\|\Theta_{\mathcal{G}_j}\|_F^2.$$

For notational simplicity in the sequel, we further denote

$$\begin{aligned}
&\vartheta = \max\{\|A\|,\|B\|\},\ \varepsilon_0 = \min_{\mathcal{S}\subseteq\mathcal{G},|\mathcal{S}|=H}\frac{1-\eta_{\mathcal{S}}}{\zeta_{\mathcal{S}}\mu_{\mathcal{S}}},\\
&\zeta = \max\Big\{\max_{\mathcal{S}\subseteq\mathcal{G},|\mathcal{S}|=H}\zeta_{\mathcal{S}}, \zeta_0\Big\},\ \eta = \max\Big\{\max_{\mathcal{S}\subseteq\mathcal{G},|\mathcal{S}|=H}\eta_{\mathcal{S}}, \eta_0\Big\},\\
&\kappa = \max\Big\{\max_{\mathcal{S}\subseteq\mathcal{G},|\mathcal{S}|=H}\Big(\Gamma_{\mathcal{S}} + \frac{1-\eta_{\mathcal{S}}}{2\|B_{\mathcal{S}}\|\zeta_{\mathcal{S}}}\Big), \eta_0\Big\},\\
&\Gamma = \max_{\mathcal{S}\subseteq\mathcal{G},|\mathcal{S}|=H}\Gamma_{\mathcal{S}},\ \tilde{\Gamma} = \Gamma + 1,
\end{aligned} \tag{40}$$

where $\tilde{\Gamma}_{\mathcal{S}}$ (resp., $\Gamma_{\mathcal{S}}$) is defined in Eq. (24) (resp., (23)), $\zeta_{\mathcal{S}}$ and $\eta_{\mathcal{S}}$ are provided in Lemma 6, and $\mu_{\mathcal{S}}$ is defined in Eq. (30).[5] We then have the following result for the regret of Algorithm 2 defined in Eq. (11), where the results holds for any general benchmark $\mathcal{S}_\star = (\mathcal{S}^1_\star,\ldots,\mathcal{S}^N_\star)$ described in Remark 1.

Theorem 1. *Suppose that Assumptions 1-4 hold. Consider any $\delta \in \mathbb{R}_{>0}$ with $0<\delta<1$. Denote*

$$\tau_0 = \frac{160np\Big(\frac{\lambda\vartheta^2}{\sigma^2} + 2(n+m)\log\big(\frac{8n}{\delta}(p + \frac{TNz_b}{\lambda})\big)\Big)}{T-1}, \tag{41}$$

[4] Note that a solution to (37) can be obtained recursively as a new data sample from the system trajectory becomes available at each time step [48].

[5] Under Assumption 2, one can check via the definition of $\Gamma_{\mathcal{S}}$ in Eq. (23) that Γ is independent of T (see, e.g., [43], [44]).

where

$$z_b = \frac{20\zeta_0^2(1+\eta_0)^2\sigma^2}{(1-\eta_0)^2}\big(2(\vartheta^2+1)\eta_0^2 m + n\big)\log\frac{8TN}{\delta}. \quad (42)$$

In Algorithm 2, let

$$\tau_1 = \Big\lceil \max\Big\{\sqrt{N}, \frac{\tau_0}{\varepsilon_0^2}\Big\}\Big\rceil, \quad (43)$$

$$\bar{y}_b = T(2\sigma_Q + \kappa^2\sigma_R)\frac{4\zeta^2\sigma^2}{(1-\eta)^2}5n\log\frac{8TN}{\delta}. \quad (44)$$

Then, for any $N > \tau_1 p$, with probability at least $1-\delta$,

$$R_{\mathcal{A}_e} = \tilde{O}(n(m+n)^2 p\sqrt{T^2|\mathcal{Q}|h(\mathcal{S}_\star)N}), \quad (45)$$

where $R_{\mathcal{A}_e}$ is defined in Eq. (11), $h(\mathcal{S}_\star)$ is defined in Eq. (12), $\mathcal{Q} = \{\mathcal{S} \subseteq \mathcal{G} : |\mathcal{S}| = H\}$, and $\tilde{O}(\cdot)$ hides polynomial factors in $\log(|\mathcal{Q}|N), \log((m+n)TN/\delta), \sigma_R, \sigma_Q, \sigma, \kappa, \zeta, \tilde{\Gamma}, \ell, \tilde{\beta}, (1-\eta)^{-1}, \nu^{-1}$, where $\tilde{\beta} = 1+\|A\|$, $\nu \in \mathbb{R}_{>0}, \ell \in [n-1]$ are given in Assumption 2, and σ_R, σ_Q are defined in (25).

A. Proof of Theorem 1

Recalling lines 5-7 in Algorithm 2, for any $j \in [p]$ and any $k \in \{N_j, \dots, N_{j+1}-1\}$, one can show that the state of system (9) satisfies that

$$x_{t+1}^k = (A + B_{\mathcal{G}_j}K_{\mathcal{G}_j})x_t^k + B_{\mathcal{G}_j}\tilde{w}_t^k + w_t^k, \quad (46)$$

for all $t \in \{0, \dots, T-1\}$, where $x_0^k = 0$, $K_{\mathcal{G}_j}$ is given by Assumption 4, and $\tilde{w}_t^k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2\sigma^2\eta_0^2 I)$. Also note that w_t^k is independent of w_t^k as we assumed before. For notational simplicity in this proof, denote

$$\tilde{\mathcal{K}} = \{k : N_j \le k \le N_{j+1}-1, j \in [p]\}, \quad (47)$$

$$\mathcal{K} = [N] \setminus \tilde{\mathcal{K}} = \{N_{p+1}, \dots, N\}. \quad (48)$$

Thus, the set $\tilde{\mathcal{K}}$ (resp., $\mathcal{K}$) contains the indices of episodes for the system identification (resp., control) phase in Algorithm 2.

Note that $(\mathcal{S}^1, \dots, \mathcal{S}^N)$ denotes the sequence of the sets of actuators selected by Algorithm 2. From Eq. (11), one can decompose the regret as $R_{\mathcal{A}_e} = R_e^1 + R_e^2 + R_e^3 + R_e^4$ with

$$R_e^1 = \mathbb{E}_{\mathcal{A}_e}\Big[\sum_{k\in\tilde{\mathcal{K}}} J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k)\Big] - \sum_{k\in\tilde{\mathcal{K}}} J_k(\mathcal{S}_\star^k),$$

$$R_e^2 = \mathbb{E}_{\mathcal{A}_e}\Big[\sum_{k\in\mathcal{K}} J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k)\Big] - \sum_{k\in\mathcal{K}} J_k(\mathcal{S}_\star^k, u_{\mathcal{S}_\star^k}^k),$$

$$R_e^3 = \sum_{k\in\mathcal{K}}\big(J_k(\mathcal{S}_\star^k, u_{\mathcal{S}_\star^k}^k) - \hat{J}_k(\mathcal{S}_\star^k)\big),$$

$$R_e^4 = \sum_{k\in\mathcal{K}}\big(\hat{J}_k(\mathcal{S}_\star^k) - J_k(\mathcal{S}_\star^k)\big),$$

where $J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k)$ is defined in Eq. (10) with $u_{\mathcal{S}^k}^k$ given by Algorithm 2, and $J_k(\mathcal{S}_\star^k)$ (resp., $\hat{J}_k(\mathcal{S}_\star^k)$) is given by (7) (resp., (21)). Note that R_e^2, R_e^3 and R_e^4 together correspond to the regret incurred by the exploitation phase in Algorithm 2, and R_e^1 corresponds to the regret incurred by the system identification phase in Algorithm 2. In particular, R_e^2 corresponds to the **Exp3.S** subroutine, and R_e^3, R_e^4 correspond to the certainty equivalent control subroutine.

In order to prove the (high probability) upper bound on $R_{\mathcal{A}_e}$, we will provide upper bounds on R_e^1, R_e^2, R_e^3, and R_e^4 separately in the sequel. First, considering any $0 < \delta < 1$, we define the following probabilistic events:

$$\mathcal{E}_w = \Big\{\|w_{t-1}^k\| \le \sigma\sqrt{5n\log\frac{8TN}{\delta}}, \forall k \in [N], \forall t \in [T]\Big\},$$

$$\mathcal{E}_{\tilde{w}} = \Big\{\|\tilde{w}_{t-1}^k\| \le \eta_0\sigma\sqrt{10m\log\frac{8TN}{\delta}}, \forall k \in \tilde{\mathcal{K}}, \forall t \in [T]\Big\},$$

$$\mathcal{E}_\Theta = \Big\{\mathrm{Tr}(\Delta_{\mathcal{G}_j}^\top V_{\mathcal{G}_j}\Delta_{\mathcal{G}_j}) \le 4\sigma^2 n\log\Big(\frac{8np\det(V_{\mathcal{G}_j})}{\delta\det(\lambda I)}\Big) + 2\lambda\|\Theta_{\mathcal{G}_j}\|_F^2, \forall j \in [p]\Big\},$$

$$\mathcal{E}_z = \Big\{\sum_{k=N_j}^{N_{j+1}-1}\sum_{t=0}^{T-1} z_{t,\mathcal{G}_j}^k z_{t,\mathcal{G}_j}^{k\top} \succeq \frac{(T-1)\tau_1\sigma^2}{80}I, \forall j \in [p]\Big\}.$$

Letting

$$\mathcal{E} = \mathcal{E}_w \cap \mathcal{E}_{\tilde{w}} \cap \mathcal{E}_\Theta \cap \mathcal{E}_z, \quad (49)$$

we have the following result which shows that $\mathcal{E}$ holds with high probability; the proof can be found in [39].

Lemma 9. *For any $0 < \delta < 1$, the event $\mathcal{E}$ defined in Eq. (49) satisfies $\mathbb{P}(\mathcal{E}) \ge 1 - \delta/2$.*

Hence, we will provide upper bounds on R_e^1, R_e^2, R_e^3 and R_e^4, under the event $\mathcal{E}$ defined in Eq. (49). The following result characterizes the estimation error of $\hat{\Theta}_{\mathcal{G}_j}$, for all $j \in [p]$; the proof can be found in [39]. Lemma 10 shows that setting the system identification phase (i.e., $\tau_1 p$) to be sufficiently long (i.e., Eq. (43)) ensures that the estimation error of $\hat{A}, \hat{B}$ is small enough such that the results proved in Section III-B can be applied to bound R_e^3, R_e^4 corresponding to the certainty equivalence subroutine in Algorithm 2.

Lemma 10. *Consider any $0 < \delta < 1$, and suppose that the event $\mathcal{E}$ holds. For any $j \in [p]$, it holds that $\|\hat{\Theta}_{\mathcal{G}_j} - \Theta_{\mathcal{G}_j}\|^2 \le \min\{\frac{\varepsilon_0^2}{p}, \frac{\tau_0}{\sqrt{N}}\}$, where $\Theta_{\mathcal{G}_j} = \begin{bmatrix}A & B_{\mathcal{G}_j}\end{bmatrix}$.*

We then have the following bounds on $R_e^1, R_e^2, R_e^3, R_e^4$; all the proofs are included in Appendix B. In particular, to bound R_e^2, we use $\bar{y}_b$ to normalize (i.e., upper bound) the cost of each episode in the control phase of Algorithm 2 so that the **Exp3.S** subroutine and Lemma 1 can be applied.

Lemma 11. *Under the event $\mathcal{E}$, it holds that*

$$R_e^1 \le \max\{\sigma_Q, \sigma_R\}\frac{\tau_1 p(2\eta_0^2+3)T\zeta_0^2}{(1-\eta_0)^2} \times \big(20\vartheta^2\eta_0^2\sigma^2 m + 10\sigma^2 n\big)\log\frac{8TN}{\delta}, \quad (50)$$

where η_0, ζ_0 are given by Assumption 4.

Lemma 12. *Under the event $\mathcal{E}$, it holds that*

$$R_e^2 \le \bar{y}_b 2\sqrt{e-1}\sqrt{|\mathcal{Q}|N(h(\mathcal{S}_\star)\log(|\mathcal{Q}|N)+e)}. \quad (51)$$

Lemma 13. *Under the event $\mathcal{E}$, the following holds with probability at least $1-\delta/2$:*

$$R_e^3 \le 64\sqrt{TN}\frac{\sigma^2(\sigma_Q+\sigma_R\kappa^2)\vartheta\zeta^3}{(1-\eta^2)(1-\eta)}\sqrt{5n\log\frac{8TN}{\delta}} + 32(\sigma_Q+\sigma_R\kappa^2)\frac{\zeta^2}{1-\eta^2}\sigma^2\sqrt{TN(\log\frac{16TN}{\delta})^3}. \quad (52)$$

Lemma 14. *Under the event $\mathcal{E}$, it holds that*

$$R_e^4 \leq \frac{4\max_{\mathcal{S}\subseteq\mathcal{G},|\mathcal{S}|=H}\{n,m_{\mathcal{S}}\}T\zeta^2\tau_0\sqrt{N}}{(1-\eta^2)}\sigma(\sigma_R+\Gamma^3) \times \left(3\tilde{\Gamma}^6(20\tilde{\Gamma}\sigma_R)^{\ell-1}32\ell^{\frac{5}{2}}\tilde{\beta}^{2(\ell-1)}(1+\nu^{-1})\max\{\sigma_Q,\sigma_R\}\right)^2. \quad (53)$$

Since $\mathbb{P}(\mathcal{E}) \geq 1-\delta/2$ from Lemma 9, we can further apply a union bound and obtain an upper bound on $R_{\mathcal{A}_e}$ that holds with probability at least $1-\delta$. Specifically, one can show using (50)-(53) that $R_e^1 = \tilde{O}(n(m+n)^2p^2T\sqrt{N})$, $R_e^2 = \tilde{O}(nT\sqrt{|Q|Nh(\mathcal{S}_\star)})$, $R_e^3 = \tilde{O}(\sqrt{nTN})$, and $R_e^4 = \tilde{O}(n(n+m)^2T\sqrt{N})$, combining which implies (45) and completes the proof of Theorem 1. ■

B. Discussions about the Results in Theorem 1

Length of the system identification phase: Recall that Eq. (43) specifies the minimum length of the system identification phase in Algorithm 2 (i.e., $\tau_1 p$). To gain more insights on how τ_0, τ_1 depend on other problem parameters, letting the regularization in the least square approach be $\lambda \geq z_b$, and supposing $TN \geq n$ and $TN \geq p$, one can show

$$\frac{\tau_0}{\varepsilon_0^2} = O(1)\frac{\zeta^4\eta_0^4(\vartheta^2+1)^2}{(1-\eta)^4(T-1)}n(m+n)\ell^5\tilde{\Gamma}^6\tilde{\beta}^{4\ell-4} \times \max\{\sigma_R^2,\sigma_Q^2\}(1+\nu^{-1})^2\log\frac{NT}{\delta}, \quad (54)$$

where $O(1)$ is a universal constant. Since $\log(NT/\delta) = o(T)$, we see from Eq. (54) that a larger value of T, i.e., the number of time steps in each episode $k \in [N]$ implies a smaller lower bound on τ_1. Thus, the regret bound in Theorem 1 holds for $N > \tau_1 p$, which can be shown to be equivalent to N being greater than a polynomial in the problem parameters. If $N \geq \tau_0^2/\varepsilon_0^4$, $\tau_1 = \lceil \max\{\sqrt{N}, \tau_0/\varepsilon_0^2\}\rceil$ reduces to $\tau_1 = \lceil\sqrt{N}\rceil$.

Knowledge of the unknown system: One can check that the choices of $\tau_1, \bar{y}_b$ require knowledge of $\sigma_Q, \sigma_R, \zeta_{\mathcal{S}}, \eta_{\mathcal{S}}, \sigma^2, \zeta_0, \eta_0, \vartheta, \ell, \nu, \Gamma_{\mathcal{S}}$ (for all $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$), where $\sigma_Q, \sigma_R, \sigma$ are given by our assumptions on the cost matrices and noise covariance, and ζ_0, η_0 (resp., ℓ, ν) are given by Assumption 4 (resp., Assumption 2). The other parameters may also be computed (or bounded) given some knowledge of the unknown system. First, as shown in [44], for any $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$, $\eta_{\mathcal{S}}$ and $\zeta_{\mathcal{S}}$ can be expressed as $\eta_{\mathcal{S}} = \sqrt{1-1/\max_{t\in[T],k\in[N]}\|P_{t,\mathcal{S}}^k\|}$ and $\zeta_{\mathcal{S}} = \sqrt{\max_{t\in[T],k\in[N]}\|P_{t,\mathcal{S}}^k\|}$. For any $t \in [T]$ and any $k \in [N]$, Eq. (5) yields $\|K_{t-1,\mathcal{S}}^k\| \leq \vartheta^2\|P_{t,\mathcal{S}}^k\|$, and one can further upper bound $\|P_{t,\mathcal{S}}^k\|$ given any stabilizing controller $K_{\mathcal{S}}$ (corresponding to the set of actuators $\mathcal{S}$) (e.g., [36, Chapter 3] and [39]). Thus, by the definition of $\Gamma_{\mathcal{S}}$ in Eq. (23), to compute (or bound) $\eta_{\mathcal{S}}, \zeta_{\mathcal{S}}, \Gamma_{\mathcal{S}}$, we need to know (or upper bound) $\max_{t\in[T],k\in[N]}\|P_{t,\mathcal{S}}^k\|$ and know the upper bound ϑ on $\|A\|$ and $\|B\|$.

Output of Algorithm 2: Since Algorithm 2 uses the **Exp3.S** subroutine to select the sets of actuators in the control phase of Algorithm 2, as we argued in Section III-A, **Exp3.S** produces a (random) sequence of subsets of selected actuators $\mathcal{S}_{N_{p+1}}, \ldots, \mathcal{S}_N$ that can be different across the episodes, which ensures exploring new sets of actuators that have not been chosen before and exploiting the set of actuators that yield the lowest cost up until the current episode. As we showed in Section IV-A, such a sequence of subsets of selected actuators yields a $\sqrt{N}$-regret bound on R_e^2.

Factors in the regret bound: First, the regret bound in Theorem 1 contains the $\sqrt{T^2N}$ factor. Although $\sqrt{T^2N}$ is not sublinear in the total number of time steps in the episodic setting of Problem (3) (i.e., TN), it matches with the optimal regret bound (in terms of the scaling of T, N) that can be achieved by any model-based algorithms for general episodic reinforcement learning problems [50]. If $T = o(\sqrt{N})$ (i.e., the number of time steps in each episode is small relative to the total number of episodes), the factor $\sqrt{T^2N}$ will become sublinear in N. Second, the regret bound contains an exponential factor in ℓ. As we argued before, $\ell << n$ if $\text{rank}(B_{\mathcal{S}})$ is large. In particular, $\ell = 1$ if $\text{rank}(B_{\mathcal{S}}) = n$ (for any $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$). Third, since $R_{\mathcal{A}_e}$ defined in Eq. (11) is a dynamic regret as we argued in Remark 1, the regret bound in (45) contains the factor $\sqrt{h(\mathcal{S}_\star)}$, where $h(\mathcal{S}_\star)$ measures the number of switchings in the benchmark $\mathcal{S}_\star = (\mathcal{S}_\star^1, \ldots, \mathcal{S}_\star^N)$. Such a factor of $h(\mathcal{S}_\star)$ is typical in the bounds on the dynamic regret of online algorithms [35], [40], [51]. If the static regret described in Remark 1 is considered, then $h(\mathcal{S}_\star) = 1$. Finally, the regret bound contains the factor $\sqrt{|\mathcal{Q}|}$ with $|\mathcal{Q}| = \binom{|\mathcal{G}|}{H}$, which will not be a bottleneck if either of $|\mathcal{G}|$ or H is small or bounded by a constant. In fact, a factor of $|\mathcal{Q}| = \binom{|\mathcal{G}|}{H}$ is unavoidable in the regret of any online algorithm defined in Eq. (11) for Problem (3), since Problem (3) is an NP-hard combinatorial optimization problem [8], [19]. In Section VI, we will show how to extend our algorithm design and regret analysis to handle large-scale instances of Problem (3).

V. Algorithm Design for Non-episodic Setting and Regret Analysis

In this section, we consider the non-episodic setting of Problem (3) described in Section II-C. For any $\mathcal{S}_t \subseteq \mathcal{G}$ and any $t \in \{0, \ldots, T-1\}$, we see from Eqs. (14)-(15) that the cost $c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}})$ of time step $t \in \{0, \ldots, T-1\}$ depends on $\mathcal{S}_0, \ldots, \mathcal{S}_{t-1}$ via the state x_t. Hence, Remark 2 implies that the **Exp3.S** algorithm and Lemma 1 cannot be directly applied to solve the non-episodic setting of Problem (3) in the same way as Algorithm 2, which creates the major challenge when we move from the episodic setting to the non-episodic setting.

Nonetheless, given a non-episodic instance of Problem (3) described in Section II-C, one may construct an episodic instance of Problem (3) as follows. First, we group the time steps $0, \ldots, T-1$ in the non-episodic instance of Problem (3) into $N' = \lfloor T/T' \rfloor$ consecutive episodes with length $T' \in \mathbb{Z}_{\geq 1}$, where the kth episode starts at $t = (k-1)T'$ and ends at $t = kT'-1$. In each episode $k \in [N']$, we fix the set of selected actuators, i.e., we let $\mathcal{S}_{(k-1)T'} = \cdots = \mathcal{S}_{kT'-1} = \mathcal{S}^k$, where $\mathcal{S}^k \subseteq \mathcal{G}$ with $|\mathcal{S}^k| = H$.[6] We then follow the notations introduced for the episodic setting in the previous sections.

[6]For simplicity, we assume that $T'N' = T$; otherwise, we can modify the number of time steps in the last episode.

Specifically, we may write the state, control and disturbance at any time step $t \in \{0, \dots, T'-1\}$ in any episode $k \in [N']$ as x_t^k, $u_{t,\mathcal{S}^k}^k$ and w_t^k, respectively, e.g., $x_{(k-1)T'+t} = x_t^k$ and $x_0^k = x_{T'}^{k-1}$ with $x_0^1 = 0$. Note that the initial state x_0^k of any episode $k \in [N']$ is not reset to 0 in the episodic instance of Problem (3) constructed above, and that $x_0^k = x_{T'}^{k-1}$ depends on $\mathcal{S}^1, \dots, \mathcal{S}^{k-1}$. For any episode $k \in [N']$, the cost matrices are set to be $Q^k = Q$, $R^k = R$, $Q_f^k = 0$ if $k < N'$ and $Q_f^k = Q_f$ if $k = N'$. Similarly to Eq. (10), we denote the cost of episode k as $J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k)$, where for notational simplicity we hide the dependency of $J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k)$ on the sets of actuators $\mathcal{S}^1, \dots, \mathcal{S}^{k-1}$ selected before episode k. One can now apply Algorithm 2 to the episodic instance constructed above; the detailed steps are summarized in Algorithm 3. Similarly, the **Exp3.S** subroutine in Algorithm 3 is applied to the MAB instance, where the total number of episodes in **Exp3.S** is $N_s = N - N'_{p+1} + 1$, the set of all possible actions is $\mathcal{Q} = \{\mathcal{S} \subseteq \mathcal{G} : |\mathcal{S}| = H\}$, and the cost associated with each possible action $\mathcal{S} \in \mathcal{Q}$ in episode k is $y_{\mathcal{S}}^k = \frac{1}{T'} J_k(\mathcal{S}^k, u_{\mathcal{S}}^k)$.

Algorithm 3: Non-Episodic Setting

Input: Parameters $\tau'_1, \lambda, N', T', \bar{y}'_b$, and $K_{\mathcal{G}_j}$ for all $j \in [p]$ from Assumption 4.

1 Initialize $N'_1 = 1$.
2 **for** $j = 1$ *to* p **do**
3 Set $N'_{j+1} \leftarrow N'_j + \tau'_1$.
4 Set $T \leftarrow T', N \leftarrow N', N_j \leftarrow N'_j \ \forall j \in [p+1], \bar{y}_b \leftarrow \bar{y}'_b$; follow lines 4-16 in Algorithm 2, where the cost $y_{\mathcal{S}_k}^k$ in line 16 is changed to be $y_{\mathcal{S}^k}^k = \frac{1}{T'} J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k)$.

Output: $\mathcal{S}_t, u_{t,\mathcal{S}_t}, \forall t \in \{0, \dots, T-1\}$.

The intuition behind the above construction is that if we fix a set of actuators $\mathcal{S}^k$ for T' time steps in an episode $k \in [N']$ (and design the corresponding control $u_{\mathcal{S}^k}^k$ based on the certainty equivalence approach), then one can use Lemma 7 to show that the influence of the initial condition x_0^k on the cost $c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}})$ (defined in Eqs. (14)-(15)) at any time step $t \in \{(k-1)T', \dots, kT'-1\}$ in episode k decays exponentially as t increases. This in turn implies that $c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}})$ tends to be independent of $\mathcal{S}^1, \dots, \mathcal{S}^{k-1}$ selected before episode $k \in [N']$, and we can then adopt the analysis developed in Section IV for the episodic setting.

We then prove the following result for the regret $R_{\mathcal{A}_c}$ of Algorithm 3 defined in Eq. (13), where we use the notations introduced in (23)-(25) and (40) (with $N = N'$ and $T = T'$). Note that since we let $\mathcal{S}_{(k-1)T'} = \cdots = \mathcal{S}_{kT'-1} = \mathcal{S}^k$ in any episode $k \in [N']$, we consider the benchmark $\mathcal{S}^\star = (\mathcal{S}_0^\star, \dots, \mathcal{S}_{T-1}^\star)$ in Eq. (13) with $\mathcal{S}_{(k-1)T'}^\star = \cdots = \mathcal{S}_{kT'-1}^\star = \mathcal{S}_\star^k$ for all $k \in [N']$, where $\mathcal{S}$ is any $\mathcal{S}_\star^k \subseteq \mathcal{G}$ with $|\mathcal{S}_\star^k| = H$. The proof of Theorem 2 follows by quantifying the dependency of $c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}})$ on x_0^k as we described above, and carefully adapting the techniques from the proof of Theorem 1. The complete proof of Theorem 2 is included in [39].

Theorem 2. *Suppose Assumptions 1-4 hold. Let* $T' = \lceil (4(e-1)(h(\mathcal{S}_\star)\ln(|\mathcal{Q}|T) + e)|\mathcal{Q}|)^{-1/3} T^{1/3} \rceil$, *and* $N' = \lceil T/T' \rceil$, *where* $\mathcal{Q} = \{\mathcal{S} \subseteq \mathcal{G} : |\mathcal{S}| = H\}$, *and* $h(\mathcal{S}_\star)$ *is defined in Eq. (12). Consider any* $\delta \in \mathbb{R}_{>0}$ *with* $0 < \delta < 1$. *Denote*

$$\tau'_0 = \frac{160np\Big(\frac{\lambda\vartheta^2}{\sigma^2} + 2(n+m)\log\big(\frac{8n}{\delta}(p + \frac{Tz'_b}{\lambda})\big)\Big)}{T'-1},$$

where

$$z'_b = \frac{180\zeta_0^4(1+\eta_0)^2\sigma^2}{(1-\eta_0)^2}\big(2(\vartheta^2+1)\eta_0^2 m + n\big)\log\frac{8T}{\delta}.$$

In Algorithm 3, let $\tau'_1 = \Big\lceil \max\Big\{\sqrt{N'}, \frac{\tau'_0}{\varepsilon_0^2}\Big\}\Big\rceil$ *and*

$$\bar{y}'_b = (2\sigma_Q + \kappa^2\sigma_R)\frac{36\zeta^4\sigma^2}{(1-\eta)^2}(20\vartheta^2\eta_0^2 m + 10n)\log\frac{8T}{\delta}.$$

Then, for any $T > \tau'_1 pT'$ *with* $T' > T_m = \frac{2}{1-\eta}(\frac{1}{3}\log T + \log\zeta) > 0$, *the following holds with probability at least* $1-\delta$*:*

$$R_{\mathcal{A}_c} = \tilde{O}(n(m+n)^2 p^2 \sqrt{|\mathcal{Q}|h(\mathcal{S}_\star)}T^{2/3}). \tag{55}$$

The complete proof of Theorem 2 can be found in [39]. Here, $\tilde{O}(\cdot)$ in Eq. (55) contains similar arguments to those in Theorem 1, and similar arguments to those in Section IV-B can be applied to Theorem 2. In particular, the regret bound in Eq. (13) also contains the factor $\sqrt{h(\mathcal{S}_\star)}$ associated with the benchmark $\mathcal{S}_\star$. Recall that based on the above construction of the episodic instance, we consider the benchmark $\mathcal{S}_\star = (\mathcal{S}_\star^1, \dots, \mathcal{S}_\star^{N'})$ in Theorem 2 with $h(\mathcal{S}_\star) \le 1 + N' = \tilde{O}(T^{2/3})$. In general, for any benchmark $\mathcal{S}_\star$ with $h(\mathcal{S}_\star) = o(T^{2/3})$, the regret bound in Eq. (13) will be sublinear in T.

VI. Handling Large-Scale Problem Instance

We now extend our algorithm design and regret analysis to efficiently handle large-scale instances of Problem (3). We first consider the episodic setting of Problem (3). Leveraging the ideas from [19], we propose to use H statistically independent copies of the **Exp3.S** subroutine in parallel, denoted as $M_1, \dots, M_H$, to choose the H actuators in each episode. Detailed steps are summarized in Algorithm 4, where the system identification phase is the same as Algorithm 2. We now explain how the **Exp3.S** subroutines in Algorithm 4 are used to select the actuators. Consider any $j \in [H]$ and any $k \in \{N_{p+1}, \dots, N\}$. Let $s_j^k \in \mathcal{G}$ be the actuator selected by the **Exp3.S** subroutine M_j and let $\mathcal{S}_j'^k = \{s_1^k, \dots, s_j^k\}$ with $\mathcal{S}_0'^k = \emptyset$. Thus, $\mathcal{S}_H'^k$ is the set of actuators selected by the H **Exp3.S** subroutines. The **Exp3.S** subroutine M_j is applied to the MAB instance, where the total number of episodes is $N_s = N - N_{p+1} + 1$, the set of all possible actions is $\mathcal{Q} = \mathcal{G}$, and the cost associated with each possible action $s \in \mathcal{Q}$ in episode k is

$$y_{j,s}^k = J_k(\mathcal{S}_{j-1}'^k, u_{\mathcal{S}_{j-1}'^k}^k) - J_k(\mathcal{S}_{j-1}'^k \cup \{s\}, u_{\mathcal{S}_{j-1}'^k \cup \{s\}}^k), \tag{56}$$

where $J_k(\cdot,\cdot)$ is defined in Eq. (10). Note that in Algorithm 4, the actual cost that M_j receives by selecting s_j^k is given by

$$\hat{y}_{j,s_j^k}^k \triangleq -J_k(\mathcal{S}^k, u_{\mathcal{S}^k}^k)\mathbb{1}\{s = s_k, j = i, b_k = 1\}, \tag{57}$$

which can be different from the true cost $y_{j,s_j^k}^k$, where b_k is a Bernoulli random variable with parameter ρ, and i and s are sampled from $[H]$ and $\mathcal{G}$ uniformly at random (u.a.r),

respectively. Moreover, the actual set of actuators selected by Algorithm 4 in episode k (i.e., $\mathcal{S}^k$) can also be different from $\mathcal{S}'^k_H$ selected by the **Exp3.S** subroutines, depending on b_k.

Algorithm 4: Large-Scale Problem Instance

Input: Parameters $\tau_1, \lambda, N, T, \bar{y}_b, \rho$, and $K_{\mathcal{G}_j}$ for all $j \in [p]$ from Assumption 4.

Follow lines 1-9 in Algorithm 2 to obtain $\hat{A}, \hat{B}$.

Initialize H independent **Exp3.S** subroutines $M_1, \ldots, M_H$ with $N_s = N - N_{p+1} + 1$, $\mathcal{Q} = \mathcal{G}$, and α_1, α_2 according to Lemma 1.

for $j = 1$ *to* H **do**

  Enter the 1st iteration of the for loop in lines 2-6 in M_j; select $s_j^{N_{p+1}} \in \mathcal{Q}$ according to the probabilities $q_{j,1}^{N_{p+1}}, \ldots, q_{j,|\mathcal{Q}|}^{N_{p+1}}$ computed by line 3 in M_j; construct $\mathcal{S}'^{N_{p+1}}_H = \cup_{j\in[H]} s_j^{N_{p+1}}$.

for $k = N_{p+1}$ *to* N **do**

  Sample $b_k \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\rho)$.

  Sample $i \in [H]$ u.a.r. and $s \in \mathcal{G}$ u.a.r.

  If $b_k = 1$, select $\mathcal{S}^k = \mathcal{S}'^k_{i-1} \cup \{s\}$ for episode k; if $b_k = 0$, select $\mathcal{S}^k = \mathcal{S}'^k_H$ for episode k.

  for $t = 0$ *to* $T-1$ **do**

    Obtain $\hat{K}^k_{t,\mathcal{S}^k}$ using $\hat{A}, \hat{B}_{\mathcal{S}^k}$ via Eq. (18).

    Play $u^k_{t,\mathcal{S}^k} = \hat{K}^k_{t,\mathcal{S}^k} x^k_t$.

  for $j = 1$ *to* H **do**

    Receive the cost $\hat{y}^k_{j,s_j^k}$; follow lines 4-6 in M_j with $y_a = -\bar{y}_b$ and $y_b = \bar{y}_b$; finish the $(k - N_{p+1} + 1)$th iteration of the for loop in lines 2-6 in M_j.

    Enter the $(k - N_{p+1} + 2)$th iteration of the for loop in lines 2-6 in M_j; select s_j^{k+1} according to the probabilities $q_{j,1}^{k+1}, \ldots, q_{j,|\mathcal{Q}|}^{k+1}$; construct $\mathcal{S}'^{k+1}_H = \cup_{j\in[H]} s_j^{k+1}$.

Output: $\mathcal{S}^k, u^k_{\mathcal{S}^k} = (u^k_{0,\mathcal{S}^k}, \ldots, u^k_{T-1,\mathcal{S}^k}), \forall k \in [N]$.

As we argued in Sections III and IV, Problem (3) (i.e., Problem (8)) is NP-hard, and using a single **Exp3.S** subroutine in Algorithm 2 leads to the exponential factor $|\mathcal{Q}| = \binom{|\mathcal{G}|}{H}$ in $|\mathcal{G}|$ in both the running time and the regret bound of Algorithm 2. To overcome the computational bottleneck, Algorithm 4 leverages H **Exp3.S** subroutines each of which selects a single actuator in each episode as we described above. One can check that the running time of each episode in Algorithm 4 is $O(H((n+m)^3 T + |\mathcal{G}|T)) = O(H(n+m)^3 T)$.

To overcome the $|\mathcal{Q}| = \binom{|\mathcal{G}|}{H}$ factor in the regret analysis, we will leverage the notion of c-regret introduced for online algorithms for combinatorial optimization problems (see, e.g., [19], [38]). The c-regret is parameterized by $c \in (0, 1]$ whose value will be specified shortly. For any $k \in [N]$, denote

$$g_k(\mathcal{S}) \triangleq J_k(\emptyset) - J_k(\mathcal{S}), \tag{58}$$

for all $\mathcal{S} \subseteq \mathcal{G}$, where $J_k(\mathcal{S})$ is given by Eq. (7). Now, we augment the elements in the ground set $\mathcal{G}$ (of all the candidate actuators) and define

$$\bar{\mathcal{G}} = \{(s^{N_{p+1}}, \ldots, s^K) : s^k \in \mathcal{G}, k \in \mathcal{K}\} \tag{59}$$

with $\mathcal{K}$ given by Eq. (48). For any $k \in \mathcal{K}$, let $\bar{\mathcal{S}}^k = \{\bar{s}^k \in \bar{s} : \bar{s} \in \bar{\mathcal{S}}\}$ with $\bar{s}^k$ denoting the kth element of the tuple $\bar{s} \in \bar{\mathcal{S}}$. Next, we define $\bar{g}(\bar{\mathcal{S}}) = \sum_{k\in\mathcal{K}} g_k(\bar{\mathcal{S}}^k)$ for all $\bar{\mathcal{S}} \subseteq \bar{\mathcal{G}}$. One can check that Problem (8) (over the episodes in $\mathcal{K}$) can be equivalently written as

$$\max_{\bar{\mathcal{S}}\subseteq\bar{\mathcal{G}}, |\bar{\mathcal{S}}|=H} \bar{g}(\bar{\mathcal{S}}). \tag{60}$$

Since Problem (60) is NP-hard, offline approximation algorithms have been proposed to solve Problem (60) with known system matrices A and B. For example, the (offline) greedy algorithm can be applied to Problem (60) and return a solution $\bar{\mathcal{S}}_g$ such that $\bar{g}(\bar{\mathcal{S}}_g) \geq (1-e^{-c_g}) g(\bar{\mathcal{S}}_\star)$,[7] where $\bar{\mathcal{S}}_\star$ is an optimal solution to Problem (60) and $c_g \in (0, 1]$ is the submodularity ratio of $\bar{g}(\cdot)$ defined to be the largest $c_g \in \mathbb{R}$ such that

$$\sum_{\bar{s}\in\bar{\mathcal{B}}\setminus\bar{\mathcal{A}}} \big(\bar{g}(\bar{\mathcal{A}} \cup \{\bar{s}\}) - \bar{g}(\bar{\mathcal{A}})\big) \geq c_g \big(\bar{g}(\bar{\mathcal{A}} \cup \bar{\mathcal{B}}) - \bar{g}(\bar{\mathcal{A}})\big), \tag{61}$$

for all $\bar{\mathcal{A}}, \bar{\mathcal{B}} \subseteq \bar{\mathcal{G}}$ (see, e.g., [52], [53], for more details).[8] Based on the above arguments, one can view the actuators selected by any M_j from episodes $k = N_{p+1}$ to N as a single action, denoted as $\bar{s}_j = (s_j^{N_{p+1}}, \ldots, s_j^K)$, which corresponds to the element in the jth iteration of the greedy algorithm.

Based on the above arguments, we introduce the following $(1 - e^{-c_g})$-regret to measure the performance of Algorithm 4:

$$R_{\mathcal{A}_l} = (1 - e^{-c_g})\Big(\sum_{k=1}^N J_k(\emptyset) - J_k(\mathcal{S}^k_\star)\Big) - \mathbb{E}_{\mathcal{A}_l}\Big[\Big(\sum_{k=1}^N J_k(\emptyset) - J_k(\mathcal{S}^k, u^k_{\mathcal{S}^k})\Big)\Big], \tag{62}$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the randomness of the algorithm, and $\mathcal{S}^k_\star$ is an optimal solution to (8) (with cost matrices Q^k, R^k, Q^k_f and an extra constraint $\mathcal{S}_0 = \cdots = \mathcal{S}_{T-1}$). Note that the benchmark $\sum_{k=1}^N J_k(\mathcal{S}^k_\star)$ in Eq. (11) is equivalent to the normalized benchmark $\sum_{k=1}^N (J_k(\emptyset) - J_k(\mathcal{S}^k_\star))$ in Eq. (62), since one can replace the objective function $J_k(\mathcal{S})$ in (8) with $J_k(\emptyset) - J_k(\mathcal{S})$, and consider the maximization over all $\mathcal{S}$, which does not change the optimal solution to (8). In words, the benchmark $\sum_{k=1}^N (J_k(\emptyset) - J_k(\mathcal{S}^k_\star))$ in $R_{\mathcal{A}_l}$ is the improvement (i.e., decrease) of the cost of Problem (3) when the sets of actuators $\mathcal{S}^1_\star, \ldots, \mathcal{S}^N_\star$ are selected, over the cost when no actuator is selected for any episode $k \in [N]$. Such a normalization of the benchmark is necessary when analyzing the c-regret of online algorithms for combinatorial optimization problems (see, e.g., [19], for more details). Accordingly, $R_{\mathcal{A}_l}$ compares the optimal improvement in the cost of Problem (3) against the improvement corresponding to Algorithm 4. For our analysis in this section, we make the following assumption.

[7] The greedy algorithm initializes $\bar{\mathcal{S}}_g = \emptyset$ and iteratively adds $\bar{s}_\star \in \arg\max_{\bar{s}\in\bar{\mathcal{G}}} (\bar{g}(\bar{\mathcal{S}}_g \cup \{\bar{s}\}) - \bar{g}(\bar{\mathcal{S}}_g))$ to $\bar{\mathcal{S}}_g$ until $|\bar{\mathcal{S}}_g| = H$.

[8] Note that computing the exact value of c_g from (61) can be intractable. Nonetheless, all of our arguments in Section VI still hold if c_g is replaced with a computable lower bound (e.g., [53]).

Assumption 5. *(a) The matrix $A \in \mathbb{R}^{n\times n}$ in system (1) is stable; (b) the pair (A, B_s) is (ℓ, ν) controllable for all $s \in \mathcal{G}$.*

Under Assumption 5(a), Assumption 4 is naturally satisfied by choosing $\mathcal{G}_j = \emptyset$ and $K_{\mathcal{G}_j} = 0$ for all $j \in [p]$. Recall that Assumption 5(b) is a sufficient condition for Assumption 2 to hold as we argued in Section III-B. Using similar arguments to those in Section IV-A, one show that under Assumption 5 and $\mathcal{E}$ defined in (49), $J_k(\emptyset)$, $J_k(\mathcal{S}^k, u^k_{\mathcal{S}^k})$ and $y^k_{j,s}$ scale linearly with T for all $k \in \{N_{p+1}, \ldots, N\}$. In the sequel, we use the same notations as those defined in (23)-(25) and (40) to denote the parameters of Problem (3), except that we replace $|\mathcal{S}| = H$ in the definitions with $|\mathcal{S}| \leq H$. We then have the following result; the proof is included in Appendix C. The proof extends the analysis in [19] for submodular objective functions (i.e., $c_g = 1$) to approximately submodular functions (i.e., $c_g \in (0, 1]$), and adopts the analyses and results developed in Sections III-IV for Problem (3).

Proposition 3. *Consider any $\delta \in \mathbb{R}_{>0}$ with $0 < \delta < 1$, and the same setting as Theorem 1. Additionally, suppose Assumption 5 holds, and in Algorithm 4 let $\rho = \left(\frac{\log(|\mathcal{G}|N)+e}{N}\right)^{1/3}$. Then, for any $N > \max\{\tau_1 p, (\log(|\mathcal{G}|N) + e)\}$, the following holds with probability at least $1 - \delta$:*

$$R_{\mathcal{A}_l} = \tilde{O}(n(m+n)^2 p^2 T |\mathcal{G}|^{3/2} H^2 h(\mathcal{S}_\star)^{1/2} N^{2/3}), \quad (63)$$

where $h(\mathcal{S}_\star)$ is defined in Eq. (12), and $\tilde{O}(\cdot)$ hides polynomial factors in $\log(|\mathcal{G}|N), \log((m+n)TN/\delta)$ and other parameters of Problem (3).

Next, we consider the non-episodic setting of Problem (3). Following the arguments in Section V, given a non-episodic instance of Problem (3), we can first construct an episodic instance with parameters N', T', and then apply Algorithm 4. Here, the corresponding H **Exp3.S** subroutines in Algorithm 4 are applied to the same MAB instances described above Eq. (56), except that we scale the costs $y^k_{j,s}$ and $\hat{y}^k_{j,s^k_j}$ defined in Eqs. (56) and (57), respectively, by a multiplicative factor $1/T'$. Similarly, following our arguments leading up to Eq. (62) and using the notations in Section II-C, the $(1-e^{-c_g})$-regret of Algorithm 4 in the non-episodic setting is given by

$$R_{\mathcal{A}'_l} = (1 - e^{-c_g})\big(J(\emptyset) - J(\mathcal{S}^\star)\big) - \mathbb{E}_{\mathcal{A}'_l}\Big[J(\emptyset) - \sum_{t=0}^{T-1} c_t(\mathcal{S}_{0:t}, u_{\mathcal{S}_{0:t}})\Big]. \quad (64)$$

where $J(\mathcal{S}^\star)$ is defined as (7), $\mathcal{S}^\star = (\mathcal{S}^\star_0, \ldots, \mathcal{S}^\star_{T-1})$ is an optimal solution to (8) (with an extra constraint $\mathcal{S}^\star_{(k-1)T'} = \cdots = \mathcal{S}^\star_{kT'-1}$ for all $k \in [N']$), $\emptyset$ is a short hand for the T-tuple $(\emptyset, \ldots, \emptyset)$, and $c_t(\cdot, \cdot)$ is defined in Eqs. (14)-(15). The result below is proved in [39].

Proposition 4. *Suppose Assumptions 1-5 hold. Set $T' = \lceil T^{1/4} \rceil$ and set $N', \tau'_1, \bar{y}'_b$ in the same way as Theorem 2. Additionally, in Algorithm 4 let $\rho = \frac{\left(\log(|\mathcal{G}|N')+e\right)^{1/3}}{T^{1/4}}$. Consider any $\delta \in \mathbb{R}_{>0}$ with $0 < \delta < 1$. Then, for any $T > \tau'_1 p T'$ with $T' > T_m = \frac{2}{1-\eta}(\frac{1}{4}\log T + \log\zeta) > 0$, the following holds with probability at least $1 - \delta$:*

$$R_{\mathcal{A}'_l} = \tilde{O}(n(m+n)^2 p^2 |\mathcal{G}|^{3/2} H^2 h(\mathcal{S}_\star)^{1/2} T^{3/4}), \quad (65)$$

where $h(\mathcal{S}_\star)$ is defined in Eq. (12), and $\tilde{O}(\cdot)$ hides polynomial factors in $\log(|\mathcal{G}|N'), \log((m+n)T/\delta)$ and other parameters of Problem (3).

Recalling our arguments in Remark 1, one can check that the regret bound on $R_{\mathcal{A}_l}$ (resp., $R_{\mathcal{A}'_l}$) in Propositions 3 (resp., Proposition 4) also holds for general benchmark $\mathcal{S}_\star = (\mathcal{S}^1_\star, \ldots, \mathcal{S}^N_\star)$ (resp., $\mathcal{S}_\star = (\mathcal{S}^1_\star, \ldots, \mathcal{S}^{N'}_\star)$), where $\mathcal{S}^k_\star$ is any $\mathcal{S}^k_\star \subseteq \mathcal{G}$ with $|\mathcal{S}^k_\star| = H$.

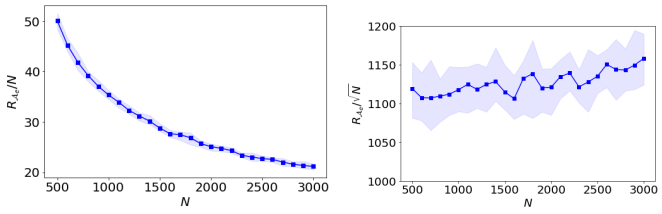
VII. Simulation Results

A. Medium-size Episodic Instances

We validate the results in Theorem 1 for Algorithm 2, using the episodic instances of Problem (3) constructed as follows. We randomly generate the matrices $A \in \mathbb{R}^{5\times 5}$ and $B \in \mathbb{R}^{5\times 10}$ such that Assumption 2 is satisfied and A is unstable. Let each column in $B \in \mathbb{R}^{5\times 10}$ correspond to one candidate actuator, and let the cardinality constraint on the set of selected actuators be $H = 2$. The cost matrices are set to be $Q^k = R^k = I$ and $Q^k_f = 2I$ for all $k \in [N]$. The covariance of the disturbance w^k_t is set to be $W = I$ for all $t \in \{0, \ldots, T-1\}$ and all $k \in [N]$. The number of time steps in any episode $k \in [N]$ is set to be $T = 5$. Given A and B generated above, we construct the known stabilizing $K_{\mathcal{G}_i}$ with $|\mathcal{G}_i| = 2$ for all $i \in [5]$ in Assumption 4. We then apply Algorithm 2 to the instances of Problem (3) constructed above, where the parameters $\tau_1, \bar{y}_b$ are set according to Theorem 1 and $\lambda = 1$ in the least squares (37) for the system identification phase in Algorithm 2. We obtain the regret $R_{\mathcal{A}_e}$ of Algorithm 2 against an optimal static benchmark $\mathcal{S}_\star = (\mathcal{S}^1_\star, \ldots, \mathcal{S}^N_\star)$, i.e., $\mathcal{S}^1_\star = \cdots = \mathcal{S}^N_\star = \mathcal{S}_{\texttt{opt}}$ and $\mathcal{S}_{\texttt{opt}} \in \arg\min_{\mathcal{S}\subseteq\mathcal{G}, |\mathcal{S}|=H} \sum_{k=1}^N J_k(\mathcal{S})$, with $h(\mathcal{S}_\star) = 1$. In Fig. 1, we plot $R_{\mathcal{A}_e}/N$ and $R_{\mathcal{A}_e}/\sqrt{N}$ for different values of the total number of episodes N.[9] From Fig. 1(a), we see that $R_{\mathcal{A}_e}/N$ decreases as N increases. From Fig. 1(b), we see that $R_{\mathcal{A}_e}/\sqrt{N}$ (slightly) increases as N increases. Hence, the results in Fig. 1(a) and (b) match with the regret bound given by Eq. (45). Specifically, the regret bound in Eq. (45) scales as $\sqrt{N}\log N$, which implies that $R_{\mathcal{A}_e}/N = O(\log N/\sqrt{N})$ and $R_{\mathcal{A}_e}/\sqrt{N} = O(\log N)$. Moreover, $R_{\mathcal{A}_e}/N$ is around 20 when $N = 3000$, since the regret bound in Eq. (45) also contains other parameters of Problem (3).

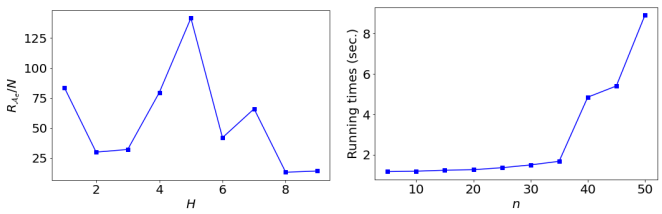
Now, we investigate how the other parameters of Problem (3) influence the performance and running times of Algorithm 2, using the instances of Problem (3) constructed above (with different values of H and n). In Fig. 2, we plot $R_{\mathcal{A}_e}/N$ for different values of the cardinality constraint H, which shows that as H increases, $R_{\mathcal{A}_e}/N$ first increases and then decreases. The result in Fig. 2(a) matches with the regret bound in Eq. (45), since the regret bound contains the factor $\sqrt{|\mathcal{Q}|}$ with $|\mathcal{Q}| = \binom{10}{H}$ in the instances of Problem (3) that we constructed. Note that $R_{\mathcal{A}_e}/N$ in Fig. 2(a) decreases

[9] All the numerical results in Section VII are averaged over 20 experiments and shaded regions display quartiles.



(a) $R_{\mathcal{A}_e}/N$ vs. N (b) $R_{\mathcal{A}_e}/\sqrt{N}$ vs. N

Fig. 1: $R_{\mathcal{A}_e}/N$ and $R_{\mathcal{A}_e}/\sqrt{N}$ against N.



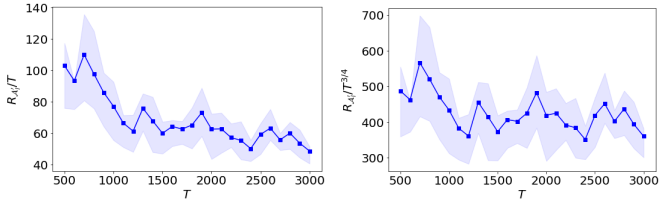
(a) $R_{\mathcal{A}_e}/N$ vs. H (b) Running times vs. n

Fig. 2: The influence of the problem parameter H (resp., n) on the performance (resp., running times) of Algorithm 2.

as H increases from 1 to 2, which is potentially due to the fact that the regret bound in Eq. (45) also contains the factor $p = \lceil 10/H \rceil$. In Fig. 2, we plot the running times of Algorithm 2 for different values of n (i.e., the dimension of the system matrix A). Similarly, we generate $A \in \mathbb{R}^{n\times n}$ for all $n = 5, 10, \ldots, 50$ and $B \in \mathbb{R}^{n\times 15}$ randomly. Fig. 2(b) shows that the running time of Algorithm 2 increases as n increases, which aligns with the time complexity $O((n+m)^3 T + |\mathcal{Q}|T)$ of each episode in Algorithm 2.

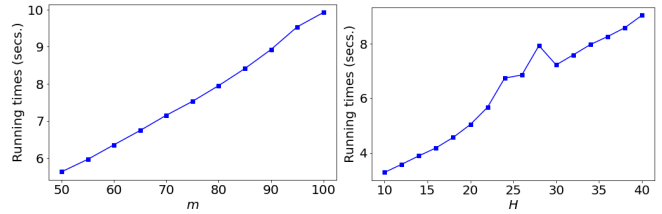
B. Large-Scale Non-Episodic Instances

We next validate the results in Proposition 4 for Algorithm 4 in the non-episodic setting. First, we randomly generate the matrices $A, B \in \mathbb{R}^{50\times 50}$ such that A is stable. Let each column in $B \in \mathbb{R}^{50\times 50}$ correspond to one candidate actuator, and let the cardinality constraint on the set of selected actuators be $H = 20$. The cost matrices are set to be $R = 10^{-3} I, Q = Q_f = 2 \cdot 10^{-3} I$. The covariance of the disturbance w_t is set to be $W = I$ for all $t \in \{0, \ldots, T-1\}$. Since A is stable, we choose the stabilizing $K_{\mathcal{G}_i} = 0$ for all $i \in [p]$ in Assumption 4. As argued in Sections V-VI, we can first construct an episodic instance of Problem (3) given the non-episodic instance generated above, and then apply Algorithm 4, where the parameters $T', N', \tau_1', \bar{y}_b', \rho$ are set according to Proposition 4 and $\lambda = 1$ in the system identification phase in Algorithm 4. Since $\binom{50}{20} \approx 5\times 10^{13}$ and Problem (3) is NP-hard, both Algorithm 3 and obtaining an optimal solution $\mathcal{S}^\star = (\mathcal{S}_1^\star, \ldots, \mathcal{S}_{T-1}^\star)$ to Problem (8) become intractable. Thus, we obtain the regret $R_{\mathcal{A}_l'}$ of Algorithm 4 (in the non-episodic setting) against a random static benchmark $\mathcal{S}^\star = (\mathcal{S}_1^\star, \ldots, \mathcal{S}_{T-1}^\star)$, where $\mathcal{S}_0^\star = \cdots = \mathcal{S}_{T-1}^\star = \mathcal{S}_{\texttt{rand}}$ and $\mathcal{S}_{\texttt{rand}}$ is chosen from $\mathcal{G}$ randomly with $|\mathcal{S}_{\texttt{rand}}| = H$. Moreover, we replace $1 - e^{-c_g}$ with 1 in Eq. (64) so that $R_{\mathcal{A}_l'}$ is lifted to the 1-regret of Algorithm 4. In Fig. 3, we plot $R_{\mathcal{A}_l'}/T$ and $R_{\mathcal{A}_l'}/T^{3/4}$ for different values of the total time steps T. Fig. 3(a) shows that $R_{\mathcal{A}_l'}/T$ decreases as T increases, which aligns with the $T^{3/4}$-regret bound in Eq. (64). Fig. 3(b) shows that $R_{\mathcal{A}_l'}/T^{3/4}$ also tends to decrease as T increases, which potentially implies that the regret bound may not be tight in terms of T. Fig. 3 also shows that Algorithm 4 yields good regret performance in terms of the stronger notion of 1-regret.



(a) $R_{\mathcal{A}_l'}/T$ vs. T (b) $R_{\mathcal{A}_l'}/T^{3/4}$ vs. T

Fig. 3: $R_{\mathcal{A}_l'}/T$ and $R_{\mathcal{A}_l'}/T^{3/4}$ against T.



(a) Running times vs. m (b) Running times vs. H

Fig. 4: The running times of Algorithm 4 against m and H.

As for the running times of Algorithm 4, we plot the running times of Algorithm 4 when applied to the non-episodic instances constructed above with different values of m and H. Fig. 4 aligns with the time complexity $O(H(n+m)^3 T)$ of Algorithm 4 and shows that Algorithm 4 is suitable for large-scale (non-episodic) instances of Problem (3).

VIII. Conclusion

We studied the online actuator selection and controller design problem for LQR with unknown system matrices, under episodic and non-episodic settings. We proposed algorithms to solve the problem and showed that our online algorithms yield sublinear regrets with respect to the horizon length of the problem. We extended our algorithm design and analysis to efficiently handle instances of the problem when both the total number of candidate actuators and the cardinality constraint scale large. We numerically validated our theoretical results.

References

[1] M. Van De Wal and B. De Jager, "A review of methods for input/output selection," *Automatica*, vol. 37, no. 4, pp. 487–510, 2001.

[2] V. Gupta, T. H. Chung, B. Hassibi, and R. M. Murray, "On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage," *Automatica*, vol. 42, no. 2, pp. 251–260, 2006.

[3] A. Olshevsky, "Minimal controllability problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 3, pp. 249–258, 2014.

[4] T. H. Summers, F. L. Cortesi, and J. Lygeros, "On submodularity and controllability in complex dynamical networks," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 1, pp. 91–101, 2016.

[5] L. Ye, S. Roy, and S. Sundaram, "Resilient sensor placement for Kalman filtering in networked systems: Complexity and algorithms," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 4, pp. 1870–1881, 2020.

[6] M. Siami and A. Jadbabaie, "A separation theorem for joint sensor and actuator scheduling with guaranteed performance bounds," *Automatica*, vol. 119, p. 109054, 2020.

[7] V. Tzoumas, L. Carlone, G. J. Pappas, and A. Jadbabaie, "LQG control and sensing co-design," *IEEE Transactions on Automatic Control*, vol. 66, no. 4, pp. 1468–1483, 2020.

[8] L. Ye, N. Woodford, S. Roy, and S. Sundaram, "On the complexity and approximability of optimal sensor selection and attack for Kalman filtering," *IEEE Transactions on Automatic Control*, vol. 66, no. 5, pp. 2146–2161, 2020.

[9] Z.-S. Hou and Z. Wang, "From model-based control to data-driven control: Survey, classification and perspective," *Information Sciences*, vol. 235, pp. 3–35, 2013.

[10] B. D. Anderson and J. B. Moore, *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 2018.

[12] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proc. Conference on Learning Theory*, 2020, pp. 2137–2143.

[13] S. Bubeck, "Introduction to online optimization," *Lecture Notes*, 2011.

[14] R. Arora, O. Dekel, and A. Tewari, "Online bandit learning against an adaptive adversary: from regret to policy regret," in *Proc. International Conference on Machine Learning*, 2012, pp. 1747–1754.

[15] E. Hazan, "Introduction to online convex optimization," *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[16] V. Tzoumas, M. A. Rahimian, G. J. Pappas, and A. Jadbabaie, "Minimal actuator placement with bounds on control effort," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 1, pp. 67–78, 2015.

[17] B. Guo, O. Karaca, T. Summers, and M. Kamgarpour, "Actuator placement under structural controllability using forward and reverse greedy algorithms," *IEEE Transactions on Automatic Control*, vol. 66, no. 12, pp. 5845–5860, 2021.

[18] D. Golovin, M. Faulkner, and A. Krause, "Online distributed sensor selection," in *Proc. ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2010, pp. 220–231.

[19] M. Streeter and D. Golovin, "An online algorithm for maximizing submodular functions," in *Proc. International Conference on Neural Information Processing Systems*, 2008, pp. 1577–1584.

[20] F. Fotiadis and K. G. Vamvoudakis, "Learning-based actuator placement for uncertain systems," in *Proc. IEEE Conference on Decision and Control*, 2021, pp. 90–95.

[21] P. Kumar, "Optimal adaptive control of linear-quadratic-gaussian systems," *SIAM Journal on Control and Optimization*, vol. 21, no. 2, pp. 163–178, 1983.

[22] M. C. Campi and P. Kumar, "Adaptive linear quadratic gaussian control: the cost-biased approach revisited," *SIAM Journal on Control and Optimization*, vol. 36, no. 6, pp. 1890–1907, 1998.

[23] H.-F. Chen and L. Guo, *Identification and stochastic adaptive control*. Springer Science & Business Media, 1991, vol. 5.

[24] K. J. Åström and B. Wittenmark, *Adaptive control*. Courier Corporation, 2013.

[25] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Optimism-based adaptive regulation of linear-quadratic systems," *IEEE Transactions on Automatic Control*, vol. 66, no. 4, pp. 1802–1808, 2020.

[26] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, vol. 10, no. 1, pp. 154–166, 1982.

[27] H.-F. Chen and L. Guo, "Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost," *SIAM Journal on Control and Optimization*, vol. 25, no. 4, pp. 845–867, 1987.

[28] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On adaptive linear–quadratic regulators," *Automatica*, vol. 117, p. 108982, 2020.

[29] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[30] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control," *Advances in Neural Information Processing Systems*, vol. 32, pp. 10 154–10 164, 2019.

[31] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," in *Proc. International Conference on Machine Learning*, 2020, pp. 1328–1337.

[32] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.

[33] I. Ziemann and H. Sandberg, "Regret lower bounds for learning linear quadratic gaussian systems," *arXiv preprint arXiv:2201.01680*, 2022.

[34] A. Mete, R. Singh, and P. Kumar, "Augmented rbmle-ucb approach for adaptive control of linear quadratic systems," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9302–9314, 2022.

[35] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[36] D. P. Bertsekas, *Dynamic programming and optimal control: Vol. 1 4th Edition*. Athena Scientific, 2017.

[37] L. Chen, M. Zhang, H. Hassani, and A. Karbasi, "Black box submodular maximization: Discrete and continuous settings," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1058–1070.

[38] S. P. Salazar and R. Cummings, "Differentially private online submodular maximization," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2021, pp. 1279–1287.

[39] L. Ye, M. Chi, Z.-W. Liu, and V. Gupta, "Online actuator selection and controller design for linear quadratic regulation with unknown system model," *arXiv preprint:2201.10197*, 2022.

[40] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. of International Conference on Machine Learning*, 2003, pp. 928–936.

[41] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *Proc. International Conference on Machine Learning*, 2019, pp. 1300–1309.

[42] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[43] B. D. Anderson and J. B. Moore, "Detectability and stabilizability of time-varying discrete-time linear systems," *SIAM Journal on Control and Optimization*, vol. 19, no. 1, pp. 20–32, 1981.

[44] R. Zhang, Y. Li, and N. Li, "On the regret analysis of online lqr control with predictions," in *Proc. American Control Conference*, 2021, pp. 697–703.

[45] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Reinforcement learning with fast stabilization in linear dynamical systems," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2022, pp. 5354–5390.

[46] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 2012.

[47] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite-time adaptive stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3498–3505, 2018.

[48] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[49] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. Conference on Learning Theory*, 2011, pp. 1–26.

[50] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *Proc. International Conference on Machine Learning*, 2017, pp. 263–272.

[51] L. Zhang, T. Yang, Z.-H. Zhou, *et al.*, "Dynamic regret of strongly adaptive methods," in *Proc. International conference on machine learning*, 2018, pp. 5882–5891.

[52] A. A. Bian, J. M. Buhmann, A. Krause, and S. Tschiatschek, "Guarantees for greedy maximization of non-submodular functions with applications," in *Proc. International Conference on Machine Learning*, 2017, pp. 498–507.

[53] L. F. Chamon, A. Amice, and A. Ribeiro, "Approximately supermodular scheduling subject to matroid constraints," *IEEE Transactions on Automatic Control*, vol. 67, no. 3, pp. 1384–1396, 2021.

APPENDIX A: PROOFS PERTAINING TO THE CERTAINTY EQUIVALENCE APPROACH

Proof of Lemma 4

Our proof is based on a similar idea to that for the proof of [30, Proposition 3]. To simplify the notations in the proof,

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2024.3422105

we assume that $T = \varphi\ell$ for some $\varphi \in \mathbb{Z}_{\geq 1}$; otherwise we only need to focus on the time steps from $T - \tilde{\varphi}\ell$ to T of Problem (3), where $\tilde{\varphi}$ is the maximum positive integer such that $T - \tilde{\varphi}\ell \geq 0$. Under the assumption that $T = \varphi\ell$, we need to show that (28) holds for $t \in \{0, \ell, \ldots, \varphi\ell\}$. Note that (28) holds for $t = T$, since $P^k_{T,\mathcal{S}} = \hat{P}^k_{T,\mathcal{S}} = Q^k_f$. In the rest of this proof, we drop the dependency of various terms on $\mathcal{S}$ and k for notational simplicity, while the proof works for any $\mathcal{S} \subseteq \mathcal{G}$ (with $|\mathcal{S}| = H$) and any $k \in [N]$. First, for any $\gamma \in \mathbb{Z}_{\geq 1}$ (with $\gamma\ell \leq T$), let us consider the noiseless LQR problem for system (1), i.e., $x_{t+1} = Ax_t + Bu_t$, from time step $\gamma\ell$ to T. Let the initial state $x_{\gamma\ell}$ be any vector in $\mathbb{R}^n$ with $\|x_{\gamma\ell}\| \leq 1$. Similarly to Eq. (10), we define the cost

$$\tilde{J}(A, B, u_{\gamma\ell:T-1}) \triangleq \Big(\sum_{j=\gamma}^{\varphi-1} \sum_{t=0}^{\ell-1} x_{j\ell+t}^\top Q x_{j\ell+t} + u_{j\ell+t}^\top R u_{j\ell+t} \Big) + x_T^\top Q_f x_T,$$

where $u_{\gamma\ell:T-1} = (u_{\gamma\ell}, \ldots, u_{T-1})$. Again, we know from [36] that the minimum value of $\tilde{J}(A, B, u_{\gamma\ell:T-1})$ (over all control policies $u_{\gamma\ell:T-1}$) is achieved by $\tilde{u}_t = K_{t,\mathcal{S}} x_t$ for all $t \in \{\gamma\ell, \gamma\ell+1, \ldots, T-1\}$, where $K_{t,\mathcal{S}}$ is given by Eq. (5). Moreover, we know that $\tilde{J}(A, B, \tilde{u}_{\gamma\ell:T-1}) = x_{\gamma\ell}^\top P_{\gamma\ell} x_{\gamma\ell}$, where $P_{\gamma\ell}$ can be obtained from Eq. (6) with $P_T = Q_f$.

Next, consider another LTI system given by $\hat{x}_{t+1} = \hat{A}\hat{x}_t + \hat{B}\hat{u}_t$ over the same time horizon and starting from the same initial state $\hat{x}_{\gamma\ell} = x_{\gamma\ell}$ as we described above. Similarly, define the corresponding cost as $J(\hat{A}, \hat{B}, \hat{u}_{\gamma\ell:T-1})$, where $\hat{u}_{\gamma\ell:T-1} = (\hat{u}_{\gamma\ell}, \ldots, \hat{u}_{T-1})$. Similarly, the minimum value of $J(\hat{A}, \hat{B}, \hat{u}_{\gamma\ell:T-1})$ (over all control policies $\hat{u}_{\gamma\ell:T-1}$) is achieved by $u'_t = \hat{K}^k_{t,\mathcal{S}} \hat{x}_t$ for all $t \in \{\gamma\ell, \gamma\ell+1, \ldots, T-1\}$, where $\hat{K}_{t,\mathcal{S}}$ is given in Eq. (18). The minimum cost is given by $J(\hat{A}, \hat{B}, u'_{\gamma\ell:T-1}) = x_{\gamma\ell}^\top \hat{P}_{\gamma\ell} x_{\ell\gamma}$, where $\hat{P}_{\gamma\ell}$ can be obtained from Eq. (19) with $\hat{P}_T = Q_f$. Moreover, note that $J(\hat{A}, \hat{B}, u'_{\gamma\ell:T-1}) \leq J(\hat{A}, \hat{B}, \hat{u}_{\gamma\ell:T-1})$, where $\hat{u}_{\gamma\ell:T-1}$ is an arbitrary control policy and the inequality follows from the optimality of $u'_{\gamma\ell:T-1}$. Recalling that ε is assumed to be small enough such that the right-hand side of (28) is smaller than or equal to 1, one can obtain from Lemma 3 that $\sigma_n(\hat{\mathcal{C}}_{\ell,\mathcal{S}}) \geq \frac{\nu}{2} > 0$, which implies that the pair $(\hat{A}, \hat{B})$ is controllable. Now, one can follow similar arguments to those for the proof of [30, Proposition 3] and show that $\hat{u}_{\gamma\ell:\varphi\ell-1}$ can be chosen such that $\hat{x}_{\varphi'\ell} = x_{\varphi'\ell}$ for all $\varphi' \in \{\gamma, \gamma+1, \ldots, \varphi\}$. It then follows from the above arguments that

$$x_{\gamma\ell}^\top \hat{P}_{\gamma\ell} x_{\gamma\ell} - x_{\gamma\ell}^\top P_{\gamma\ell} x_{\gamma\ell} \leq \Big(\sum_{j=\gamma}^{\varphi-1} \sum_{t=0}^{\ell-1} \hat{x}_{j\ell+t}^\top Q \hat{x}_{j\ell+t} + \hat{u}_{j\ell+t}^\top R \hat{u}_{j\ell+t} - x_{j\ell+t}^\top Q x_{j\ell+t} - u_{j\ell+t}^\top R u_{j\ell+t} \Big). \quad (66)$$

One can further follow similar arguments to those for the proof of [30, Proposition 3] and show that $\hat{u}_{\gamma\ell:T-1}$ in Eq. (66) can be chosen such that the following holds:

$$x_{\gamma\ell}^\top \hat{P}_{\gamma\ell} x_{\gamma\ell} - x_{\gamma\ell}^\top P_{\gamma\ell} x_{\gamma\ell} \leq \frac{1}{2} \mu_{\gamma\ell} \varepsilon, \quad (67)$$

under the assumption that $\frac{1}{2}\mu_{\gamma\ell}\varepsilon \leq 1$, where $\mu_{\gamma\ell}$ (i.e., $\mu^k_{\gamma\ell,\mathcal{S}}$) is defined in Eq. (29). Now, reversing the roles of (A, B) and $(\hat{A}, \hat{B})$ in the arguments above, one can also obtain that

$$x_{\gamma\ell}^\top P_{\gamma\ell} x_{\gamma\ell} - x_{\gamma\ell}^\top \hat{P}_{\gamma\ell} x_{\gamma\ell} \leq \frac{1}{2} \mu_{\gamma\ell} \frac{\|\hat{P}_{\gamma\ell}\|}{\|P_{\gamma\ell}\|} \varepsilon, \quad (68)$$

under the assumption that $\frac{1}{2}\mu_{\gamma\ell}\frac{\|\hat{P}_{\gamma\ell}\|}{\|P_{\gamma\ell}\|}\varepsilon \leq 1$.[10] Note from Eq. (6) and Assumption 1 that $P_{\gamma\ell} \succeq Q \succeq I$, and note that (67) and (68) hold for any $x_{\gamma\ell} \in \mathbb{R}^n$ with $\|x_{\gamma\ell}\| \leq 1$ as we discussed above. It then follows from (67) that $\lambda_1(\hat{P}_{\gamma\ell}) \leq \lambda_1(P_{\gamma\ell}) + 1$, i.e., $\|\hat{P}_{\gamma\ell}\| \leq \|P_{\gamma\ell}\| + 1 \leq 2\|P_{\gamma\ell}\|$. Hence, we have from (67) and (68) that $\lambda_1(\hat{P}_{\gamma\ell} - P_{\gamma\ell}) \leq \mu_{\gamma\ell}\varepsilon$ and $\lambda_1(P_{\gamma\ell} - \hat{P}_{\gamma\ell}) \leq \mu_{\gamma\ell}\varepsilon$, which further implies (28). ■

Appendix B: Proofs pertaining to Theorem 1

Proof Sketch of Lemma 11: The lemma can be proved by upper bounding $\|x^k_t\|$ and $\|u^k_t\|$ for all $t\{0, \ldots, T-1\}$ and all $k \in \tilde{\mathcal{K}}$ under the event $\mathcal{E}$. Details are included in [39]. ■

Proof of Lemma 12: Consider any episode $k \in \mathcal{K}$ in Algorithm 2. Noting that $x^k_0 = 0$, one can show that the state of system (9) corresponding to $\mathcal{S}^k$ selected in line 12 of Algorithm 2 satisfies $x^k_{t+1} = \sum_{i=0}^t \hat{\Psi}^k_{t,i}(\mathcal{S}^k) w^k_i$, where $\hat{\Psi}^k_{t,i}(\mathcal{S}^k)$ is defined in Eq. (35). Moreover, supposing that the event $\mathcal{E}$ holds, we know from Lemma 10 that $\|\hat{\Theta}_{\mathcal{G}_j} - \Theta_{\mathcal{G}_j}\| \leq \varepsilon_0/\sqrt{p}$ for all $j \in [p]$. It follows that $\hat{A}$ and $\hat{B}$ obtained in line 9 of Algorithm 2 satisfy that $\|\hat{A} - A\| \leq \varepsilon_0$ and $\|\hat{B} - B\| \leq \varepsilon_0$, which also implies that $\|\hat{B}_{\mathcal{S}^k} - B_{\mathcal{S}^k}\| \leq \varepsilon_0$, where $\hat{B}_{\mathcal{S}^k}$ contains the columns of $\hat{B}$ that correspond to $\mathcal{S}^k$. Now, one can obtain from the choice of ε_0 in (40) and Proposition 1 that $\|\hat{K}^k_{t,\mathcal{S}^k} - K^k_{t,\mathcal{S}^k}\| \leq \frac{1-\eta_{\mathcal{S}^k}}{2\|B_{\mathcal{S}^k}\|\zeta_{\mathcal{S}^k}}$, $\forall t \in \{0, \ldots, T-1\}$, which also implies that $\|\hat{K}^k_{t,\mathcal{S}^k}\| \leq \kappa$, $\forall t \in \{0, \ldots, T-1\}$, where $\hat{K}^k_{t,\mathcal{S}^k}$ and $K^k_{t,\mathcal{S}^k}$ are given by Eqs. (18) and (5), respectively. We have from Lemma 7 that $\|\hat{\Psi}^k_{t_2,t_1}(\mathcal{S}^k)\| \leq \zeta_{\mathcal{S}^k}(\frac{1+\eta_{\mathcal{S}^k}}{2})^{t_2-t_1}$, for all $t_1, t_2 \in \{0, \ldots, T-1\}$ with $t_2 \geq t_1$, where we know from Lemma 6 that $0 < (1+\eta_{\mathcal{S}^k})/2 < 1$. One can now use similar arguments to those for [31, Lemma 38] and show that

$$\|x^k_t\| \leq \frac{2\zeta_{\mathcal{S}^k}}{1-\eta_{\mathcal{S}^k}} \max_{k' \in \mathcal{K}, t' \in \{0,\ldots,T-1\}} \|w^{k'}_{t'}\|.$$

Thus, under the event $\mathcal{E}$ defined in Eq. (49), we have that $\|x^k_t\| \leq \frac{2\zeta\sigma}{1-\eta}\sqrt{5n \log \frac{8TN}{\delta}}$, for all $k \in \mathcal{K}$ and all $t \in \{0, \ldots, T\}$. Furthermore, from (10) and the fact that $u^k_{t,\mathcal{S}^k} = \hat{K}^k_{t,\mathcal{S}^k} x^k_t$, one can similarly show that under $\mathcal{E}$,

$$J_k(\mathcal{S}^k, u^k_{\mathcal{S}^k}) \leq T(2\sigma_Q + \kappa^2\sigma_R) \frac{4\zeta^2\sigma^2}{(1-\eta)^2} 5n \log \frac{8TN}{\delta} = \bar{y}_b.$$

To proceed, recall that we use the **Exp3.S** algorithm in Algorithm 2 to select $\mathcal{S}^k$ for all $k \in \mathcal{K}$. As we argued in Section III, each action in the **Exp3.S** algorithm corresponds to a set $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{G}| = H$, i.e., the set of all possible actions $\mathcal{Q}$ in the **Exp3.S** algorithm is given by $\mathcal{Q} = \{\mathcal{S} \subseteq \mathcal{G} : |\mathcal{S}| = H\}$. Moreover, the cost of the action corresponding to $\mathcal{S}^k$ in

[10] Note that the proof technique in [30] is for the infinite-horizon (noiseless) LQR problem, which can be adapted to the finite-horizon setting studied here. The details of such an adaption are omitted for conciseness.

episode $k \in [N]$ is given by $J_k(\mathcal{S}^k, u^k_{\mathcal{S}^k})$. Thus, we can replace y_b (resp., y_a) in (16) with $\bar{y}_b$ (resp., 0), and obtain that (51) holds under the event $\mathcal{E}$. ■

Proof Sketch of Lemma 13: We provide a proof sketch here; the detailed proof can be found in [39]. Consider any $k \in \mathcal{K}$. As in the proof of Lemma 12, and applying Eqs. (9) and (22), one can show that $J_k(\mathcal{S}^k_\star, u^k_{\mathcal{S}^k_\star}) = \Big(\sum_{t=0}^{T-1} 2w_t^{k\top} \tilde{P}^k_{t+1,\mathcal{S}^k_\star}(A + B_{\mathcal{S}^k_\star})x^k_t + w_t^{k\top} \tilde{P}^k_{t+1,\mathcal{S}^k_\star} w^k_t \Big)$, where we note that $x^k_0 = 0$ and $\tilde{P}^k_{T,\mathcal{S}^k_\star} = Q^k_f$. From the definition of R^3_e, one can show that $R^3_e = \sum_{k\in\mathcal{K}} \big(\sum_{t=0}^{T-1} 2w_t^{k\top} \tilde{P}^k_{t+1,\mathcal{S}^k_\star}(A + B_{\mathcal{S}^k_\star})x^k_t + w_t^{k\top} \tilde{P}^k_{t+1,\mathcal{S}^k_\star} w^k_t - \sigma^2 \mathrm{Tr}(\tilde{P}^k_{t+1,\mathcal{S}^k_\star}) \big)$. The proof follows by upper bounding the terms in this summation via adapting [41, Lemmas 31&32]. ■

Proof of Lemma 14: As in the proof of Lemma 12, under the event $\mathcal{E}$ defined in Eq. (49), $\hat{A}$ and $\hat{B}$ obtained in line 9 of Algorithm 2 satisfy $\|\hat{A} - A\| \le \sqrt{\frac{\tau_0}{\sqrt{N}}}$ and $\|\hat{B} - B\| \le \sqrt{\frac{\tau_0}{\sqrt{N}}}$, which also implies that $\|\hat{B}_{\mathcal{S}} - B_{\mathcal{S}}\| \le \sqrt{\frac{\tau_0}{\sqrt{N}}}$ for all $\mathcal{S} \subseteq \mathcal{G}$ with $|\mathcal{S}| = H$. Under the event $\mathcal{E}$, one can then show via the choice of ε_0 in (40) and Proposition 2 that (14) holds. ■

Appendix C: Proof Sketch of Proposition 3

We provide a proof sketch here; the detailed proof can be found in [39]. Similarly to the proof of Theorem 1 provided in Section IV-A, the regret $R_{\mathcal{A}_l}$ of Algorithm 4 can be decomposed as $R_{\mathcal{A}_l} = R^1_{\mathcal{A}_l} + R^2_{\mathcal{A}_l}$, where $R^1_{\mathcal{A}_l}$ corresponds to the system identification phase and the certainty equivalence subroutine, and $R^2_{\mathcal{A}_l}$ corresponds to the **Exp3.S** subroutines $M_1, \ldots, M_H$. Suppose the event $\mathcal{E}$ defined in (49) holds. Following similar arguments to those for Lemmas 11, 13 and 14 in the proof of Theorem 1, one can show that $R^1_{\mathcal{A}_l} = \tilde{O}(n(m+n)^2 p^2 T\sqrt{N})$. We then focus on upper bounding $R^2_{\mathcal{A}_l}$. To proceed, for any $\bar{\mathcal{S}} \subseteq \bar{\mathcal{G}}$ with $\bar{\mathcal{G}}$ defined in Eq. (59) and any $k \in \mathcal{K}$, define $f_k(\bar{\mathcal{S}}) = J_k(\emptyset) - J_k(\bar{\mathcal{S}}^k, u^k_{\bar{\mathcal{S}}^k})$, where $\bar{\mathcal{S}}^k = \{\bar{s}^k \in \bar{s} : \bar{s} \in \bar{\mathcal{S}}\}$ with $\bar{s}^k$ denoting the kth element of the tuple $\bar{s} \in \bar{\mathcal{S}}$, and $u^k_{t,\bar{\mathcal{S}}^k} = \hat{K}^k_{t,\bar{\mathcal{S}}^k} x^k_t$ for all $t \in \{0, \ldots, T-1\}$ with $\hat{K}^k_{t,\bar{\mathcal{S}}^k}$ obtained via Eq. (18) using $\hat{A}, \hat{B}$ from Algorithm 4. For any $\bar{\mathcal{S}} \subseteq \bar{\mathcal{G}}$, we define $\bar{f}(\bar{\mathcal{S}}) = \sum_{k\in\mathcal{K}} f_k(\bar{\mathcal{S}})$. For any $j \in [H]$, we further denote $\bar{\mathcal{S}}'_j = \{\bar{s}_1, \ldots, \bar{s}_j\}$ with $\bar{\mathcal{S}}_0 = \emptyset$, where $\bar{s}_j = (s_j^{N_{p+1}}, \ldots, s_j^K)$ contains the actuators selected by M_j from episodes $k = N_{p+1}$ to N. Eq. (56) implies that the regret of any M_j in Algorithm 4 can be written as

$$r_j = \max_{\bar{s}\in\bar{\mathcal{G}}} \big\{ \bar{f}(\bar{\mathcal{S}}'_{j-1} \cup \bar{s}) - \bar{f}(\bar{\mathcal{S}}'_{j-1}) \big\} - \big(\bar{f}(\bar{\mathcal{S}}'_{j-1} \cup \bar{s}_j) - \bar{f}(\bar{\mathcal{S}}'_{j-1}) \big).$$

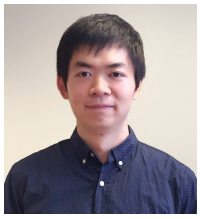
Following similar arguments to those in the proof of Lemmas 13-14, one can show via the definition of Algorithm 4:

$$\begin{aligned} R^2_{\mathcal{A}_l} &= (1 - e^{-c'}) \sum_{k\in\mathcal{K}} \big(J_k(\emptyset) - J_k(\mathcal{S}^k_\star, u^k_{\mathcal{S}^k_\star}) \big) \\ &\quad - \mathbb{E}_{\mathcal{A}_l} \Big[\sum_{k\in\mathcal{K}} J_k(\emptyset) - f_k(\bar{\mathcal{S}}'_H) - J_k(\mathcal{S}^k, u^k_{\mathcal{S}^k}) + f_k(\bar{\mathcal{S}}'_H) \Big] \\ &\le \tilde{O}(n(n+m)^2 T\sqrt{N}) + \mathbb{E}_{\mathcal{A}_l} \Big[\sum_{j=1}^{H} r_j \Big] + \mathbb{E}_{\mathcal{A}_l}[N_e] \tilde{O}(nT). \end{aligned}$$

Thus, it remains to bound $\mathbb{E}_{\mathcal{A}_l}[r_j]$ for all $j \in [H]$. Eq. (57) and the definition of Algorithm 4 yield that for any $j \in [H]$, any $s \in \mathcal{G}$ and any $k \in \mathcal{K}$, $\mathbb{E}_{\mathcal{A}_l}[\hat{y}^k_{j,s}] = \frac{\rho}{|\mathcal{G}|H} y^k_{j,s} + \frac{\rho}{|\mathcal{G}|H} J_k(\mathcal{S}'^k_{j-1}, u^k_{\mathcal{S}'^k_{j-1}})$. One can also show that $y^k_{j,s} \in [-\bar{y}_b, \bar{y}_b]$ for all $j \in [H]$, all $s \in \mathcal{G}$ and all $k \in \mathcal{K}$. Following similar arguments to those for [19, Lemma 5&Theorem 13], one can now show via Lemma 1 that for any $j \in [H]$,

$$\begin{aligned} \mathbb{E}_{\mathcal{A}_l}[r_j] &\le \frac{|\mathcal{G}|H}{\rho} \tilde{O}(nT) \mathbb{E}_{\mathcal{A}_l} \Big[\sqrt{|\mathcal{G}| N_e (h(\mathcal{S}_\star) \ln(|\mathcal{G}|N_e) + e)} \Big] \\ &\le |\mathcal{G}| H \tilde{O}(nT) \sqrt{\frac{|\mathcal{G}|}{\rho} N (h(\mathcal{S}_\star) \ln(|\mathcal{G}|N) + e)}. \end{aligned}$$

Combining these arguments with the choice of ρ, we obtain $R^2_{\mathcal{A}_l} = \tilde{O}(n(m+n)^2 T |\mathcal{G}|^{3/2} H^2 h(\mathcal{S}_\star)^{1/2} N^{2/3})$, which together with the upper bound on $R^1_{\mathcal{A}_l}$ complete the proof. ■



Lintao Ye is a Lecturer in the School of Artificial Intelligence and Automation at the Huazhong University of Science and Technology, Wuhan, China. He received his M.S. degree in Mechanical Engineering in 2017, and his Ph.D. degree in Electrical and Computer Engineering in 2020, both from Purdue University, IN, USA. He was a Postdoctoral Researcher at the University of Notre Dame, IN, USA. His research interests are in the areas of optimization algorithms, control theory, estimation theory, and network science.

Ming Chi is a Professor in the School of Artificial Intelligence and Automation at the Huazhong University of Science and Technology, Wuhan, China. He received the Ph.D. degree in Control Science and Engineering from the Huazhong University of Science and Technology in 2013. His research interests include networked control systems, multi-agent systems, complex networks, and hybrid control systems.

Zhi-Wei Liu is a Professor with the School of Artificial Intelligence and Automation at the Huazhong University of Science and Technology, Wuhan, China. He received the B.S.degree in Information Management and Information System from Southwest Jiaotong University, Chengdu, China, in 2004, and the Ph.D.degree in Control Science and Engineering from the Huazhong University of Science and Technology in 2011. His current research interests include cooperative control and optimization of distributed network systems.

Vijay Gupta is the Elmore Professor of Electrical and Computer Engineering at Purdue University. He received his B. Tech degree at Indian Institute of Technology, Delhi, and his M.S. and Ph.D. at California Institute of Technology, all in Electrical Engineering. He received the 2018 Antonio J Rubert Award from the IEEE Control Systems Society, the 2013 Donald P. Eckman Award from the American Automatic Control Council and a 2009 National Science Foundation (NSF) CAREER Award. His research interests are broadly in the interface of communication, control, distributed computation, and human decision making.