Relative Almost Sure Regret Bounds for Certainty Equivalence Control of Markov Jump Systems

Borna Sayedana, Mohammad Afshari, Peter E. Caines, Aditya Mahajan

Abstract— **In this paper, we consider learning and control problem in an unknown Markov jump linear system (MJLS) with perfect state observations. We first establish a generic upper bound on regret for any learning based algorithm. We then propose a certainty equivalence-based learning algorithm and show that this algorithm achieves a regret of $\mathcal{O}(\sqrt{T}\log(T))$ relative to a certain subset of the sample space. As part of our analysis, we revisit the switched least squares system identification algorithm of [1], [2] for autonomous MJLS and generalize it to controlled MJLS, establishing strong consistency and almost sure rates of convergence of this method.**

I. Introduction

The main goal of reinforcement learning and adaptive control is simultaneous learning and control of unknown dynamical systems. Due to continuity and unboundedness of the state and action spaces in control setups, classical reinforcement learning algorithms do not achieve good performance. Recently, there has been a surge of interest in designing reinforcement learning algorithms for linear quadratic regulators (LQR) and analyzing the performance of these algorithms [3]–[9]. These results exploit the linearity, time-invariancy, and structure of the cost function in the proposed algorithms and analysis.

Markov jump systems are a mathematical formulation which model time-varying dynamical systems with abrupt and stochastic changes in the dynamics. These systems find application in cyber-physical system [10], networked control systems [11], [12], etc. In this paper, we investigate the problem of simultaneous learning and controlling an unknown Markov jump linear system (MJLS). We use the switched least squares method proposed in [1], [2] in the closed-loop setup for the system identification and use the system estimates in a certainty equivalence controller.

The problem of learning and controlling MJLS systems has recently received some attention in the literature. The sensitivity analysis of certainty equivalence controller to the system parameter is investigated in [13]. Based on the results of [13], a system identification algorithm and a certainty equivalence controller is proposed in [14] where it is shown that the proposed method achieves the regret of $\tilde{\mathcal{O}}(\sqrt{T})$ with high probability, where T denotes the time horizon, and notation $\tilde{O}$ hides logarithmic factors of T. It is shown in [15] that policy gradient method converges to the optimal policy for MJLS systems. The performance of Thompson sampling algorithm in controlling networked control systems as a special case of switched linear systems is investigated in [16]. The problem of system identification of Markov jump linear systems from a single trajectory is investigated in [1], [2] and [14].

A. Contributions

- We characterize the almost sure (relative to a certain subset of the noise process and the algorithm randomization) regret bounds for general class of linear adaptive polices.
- We use switched least squares method for closed-loop system identification of MJLS systems, show that this method is strongly consistent, and establish that its rate of convergence is $\mathcal{O}(\sqrt{\log(T)/T})$.
- We propose a version of certainty equivalence controller based on the switched least squares system identification method, and show that this algorithm achieves a regret of $\mathcal{O}(\sqrt{T}\log(T))$ relative to a certain subset of the sample space.
- We show that there exists a finite identification horizon T_0 for which this algorithm achieves the almost sure regret of $\mathcal{O}(\sqrt{T}\log(T))$ on the entire sample space.

B. Organization

The rest of the paper is organized as follows. In Sec II, we review some standard results about MJLS systems that are useful in our analysis. In Sec. III, we characterize the notion of almost sure regret criteria. In Sec. V, we present our system identification method and reinforcement learning algorithm. The main results are presented in Sec. VI. We concluded our results in Sec. VII.

C. Notation

Given a matrix A, $A(i,j)$ denotes its (i,j)-th element, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest magnitudes of right eigenvalues, $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^{\intercal}A)}$ denotes the spectral norm. For a square matrix Q, $\operatorname{Tr}(Q)$ denotes the trace. When Q is symmetric, $Q \succeq 0$ and $Q \succ 0$ denote that Q is positive semi-definite and positive definite, respectively. For two square matrices, Q_1 and Q_2 of the same dimension, $Q_1 \succeq Q_2$ means $Q_1 - Q_2 \succeq 0$. Given two matrices A and B, $A \otimes B$ denotes the Kronocher product of the two matrices.

The authors are with the Department of Electrical and Computer Engineering, McGill University, 3480 Rue University, Montreal, QC H3A 0E9, Canada. Emails: {borna.sayedana, mohammad.afshari2}@mail.mcgill.ca, {peter.caines, aditya.mahajan}@mcgill.ca.

This work was supported in part by Fonds de Recherche du Québec, Nature et Technologies (FRQNT), Bourses de doctorat en recherche Grant 316558, Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2021-03511 and RGPIN-2019-05336.

Given a sequence of positive numbers $\{a_t\}_{t\geq 0}, a_T = \mathcal{O}(T)$ means that $\limsup_{T\to\infty} a_T/T < \infty$, and $a_T = o(T)$ means that $\limsup_{T\to\infty} a_T/T = 0$. Given a sequence of vectors $\{x_t\}_{t\in\mathcal{T}}$, $\text{vec}(x_t)_{t\in\mathcal{T}}$ denotes the vector formed by vertically stacking $\{x_t\}_{t\in\mathcal{T}}$. Given a sequence of random variables $\{x_t\}_{t\geq 0}$, $x_{0:t}$ is a short hand for $(x_0, \cdots, x_t)$ and $\sigma(x_{0:t})$ denotes the sigma field generated by random variables $x_{0:t}$. When describing values that are taken by consecutive variables, for example s_t and s_{t+1}, we use s to denote a generic value of s_t and s_+ to denote a generic values of s_{t+1}. Given a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$, Ω denotes the sample space, $\omega \in \Omega$ denotes a generic elementary event, $\mathbb{P}(\cdot)$ denotes the probability measure and $\mathbb{E}[\cdot]$ denotes the expectation operator.

$\mathbb{R}$ and $\mathbb{N}$ denote the sets of real and natural numbers. For a set $\mathcal{T}$, $|\mathcal{T}|$ denotes its cardinality. For a vector x, $\|x\|$ denotes the Euclidean norm. For a matrix A, $\|A\|$ denotes the spectral norm and $\|A\|_\infty$ denotes the element with the largest absolute value. $\text{diag}(A_1, A_2, \ldots, A_n)$ denotes the block diagonal matrix, where the blocks are matrices $A_1, A_2, \ldots, A_n$.

II. Background on Markov Jump Linear Systems

We start by a review of stability of autonomous Markov Jump Linear Systems and the basic results for optimal control of Markov Jump Linear Systems.

A. Stability of autonomous Markov Jump Linear Systems

Consider an autonomous discrete-time MJLS with continuous state $x_t \in \mathbb{R}^n$ and the discrete state $s_t \in \mathcal{S} = \{1, 2, \ldots, d\}$. The system starts with a known initial state (x_1, s_1). The continuous state evolves over time according to

$$x_{t+1} = A_{s_t} x_t, \quad t \geq 1, \tag{1}$$

where the set $\{A_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ consists of the system dynamics matrices. The discrete state evolves in a time-homogeneous Markov manner according to a transition matrix H. We will refer to the above system as MJLS system $(\{A_s\}_{s\in\mathcal{S}}, H)$.

We assume that the Markov chain $\{s_t\}_{t\geq 1}$ is irreducible and aperiodic, and therefore, has a stationary distribution $\{\rho_s\}_{s\in\mathcal{S}}$.

Definition 1. *The MJLS system* (1) *is called Mean Square Stable (MSS) if for any initial state* (x_1, s_1), $\lim_{t\to\infty} \|\mathbb{E}[x_t]\| = 0$, *and* $\lim_{t\to\infty} \|\mathbb{E}[x_t x_t^\intercal]\| = 0$.

The following characterizations of MSS follow from [17, Theorem 3.9]:

Proposition 1. *The following conditions are equivalent:*

1) *The MJLS system in* (1) *is MSS.*
2) *Transition probability matrix H and matrices $\{A_s\}_{s\in\mathcal{S}}$ satisfy:*

$$\lambda_{\max}\Big((H^\intercal \otimes I_{n^2})\,\text{diag}(A_1 \otimes A_1, \ldots, A_d \otimes A_d)\Big) < 1.$$

3) *The MJLS system* (1) *is exponentially stochastically stable , i.e., there exists $\beta \geq 1$ and $0 < \zeta < 1$ such that for any initial state (x_1, s_1), we have*

$$\mathbb{E}[\|x_t\|^2] \leq \beta\zeta^t \|x_0\|^2, \quad t \geq 1.$$

4) *The MJLS system* (1) *is stochastically stable (SS), i.e., for all initial state (x_1, s_1), we have*

$$\sum_{t=0}^{\infty} \mathbb{E}[\|x_t\|^2] < \infty.$$

B. Optimal control of Markov Jump Linear Systems

Consider a discrete-time MJLS with continuous state $x_t \in \mathbb{R}^n$, discrete state $s_t \in \mathcal{S}$, control input $u_t \in \mathbb{R}^m$, and disturbance $w_t \in \mathbb{R}^n$. The system starts with a known initial state (x_1, s_1). The continuous state evolves over time according to:

$$x_{t+1} = A_{s_t} x_t + B_{s_t} u_t + w_t, \quad t \geq 1, \tag{2}$$

where $\{A_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ and $\{B_s \in \mathbb{R}^{n\times m}\}_{s\in\mathcal{S}}$ are the system dynamics matrices, and $\{w_t\}_{t\geq 1}$ is an i.i.d. process with $\mathbb{E}[w_t] = 0$ and $\mathbb{E}[w_t w_t^\intercal] = \sigma_w^2 I$. The discrete state evolves in a time-homogeneous Markov manner, independent of $\{w_t\}_{t\geq 1}$, according to a transition matrix H. We assume that the Markov chain $\{s_t\}_{t\geq 1}$ is irreducible and aperiodic, and therefore, has a stationary distribution $\{\rho_s\}_{s\in\mathcal{S}}$.

The system incurs a per-step cost

$$c(x_t, s_t, u_t) := x_t^\intercal Q_{s_t} x_t + u_t^\intercal R_{s_t} u_t, \tag{3}$$

where $\{Q_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ and $\{R_s \in \mathbb{R}^{m\times m}\}_{s\in\mathcal{S}}$ are positive definite matrices. The objective is to design a controller which observes the state of the system and chooses control inputs to minimize the long term average cost given by

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Bigg[\sum_{t=1}^{T} c(x_t, s_t, u_t)\Bigg]. \tag{4}$$

1) Stochastic stabilizability and stochastic detectability: We now define two important properties of MJLS systems:

Definition 2. *The MJLS system* (2) *is* stochastically stabilizable, *if there exists gain matrices $\{F_s \in \mathbb{R}^{m\times n}\}_{s\in\mathcal{S}}$ such that the autonomous MJLS system $(\{A_s - B_s F_s\}_{s\in\mathcal{S}}, H)$ is MSS.*

Definition 3. *The MJLS system* (2) *is* stochastically detectable , *if there exists gain matrices $\{K_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ such that the autonomous MJLS system $(\{A_s - K_s Q_s^{1/2}\}_{s\in\mathcal{S}}, H)$ is MSS.*

Note that one can check stochastic stability and stochastic detectability via Linear Matrix inequalities (LMIs). For instance, a check for stochastic stabilizability is given by [17, Proposition 3.42].

Proposition 2. *The MJLS system* (2) *is stochastically stabilizable if and only if there exist matrices $\{W_s^2 \in \mathbb{R}^{n\times m}\}_{s\in\mathcal{S}}$*

and positive semi-definite matrices $\{W_s^1 \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ and $\{W_s^3 \in \mathbb{R}^{m\times m}\}_{s\in\mathcal{S}}$ such that:

$$\begin{aligned}\sum_{s\in\mathcal{S}} H_{ss'}(A_s W_s^1 A_s^\intercal &+ B_s (W_s^2)^\intercal A_s^\intercal \\ &+ A_s W_s^2 B_s^\intercal + B_s W_s^3 B_s^\intercal) < W_{s'}^1 \qquad \forall s' \in \mathcal{S}, \\ \begin{bmatrix} W_s^1 & W_s^2 \\ (W_s^2)^\intercal & W_s^3 \end{bmatrix} &\geq 0, \qquad \forall s \in \mathcal{S}, \\ W_s^1 &> 0, \qquad \forall s \in \mathcal{S}.\end{aligned}$$

A similar test for stochastic detectability follows by replacing B_s by $(Q_s^{1/2})^\intercal$ in the above proposition.

2) Optimal control of MJLS: We assume that the system satisfies the following:

Assumption 1. *The MJLS system in* (2) *is stochastically stabilizable and stochastically detectable.*

The following result follows from [18, Theorem 45 and Theorem 51].

Theorem 1. *Under Assumption 1, the minimum value of the average cost* (4) *is*

$$\sigma_w^2 \sum_{s\in\mathcal{S}} \sum_{s_+\in\mathcal{S}} \rho_s H_{ss_+} \operatorname{Tr}(P_{s_+}) \tag{5}$$

and is achieved by the feedback policy

$$u_t = -L_{s_t} x_t, \quad t \geq 1, \tag{6}$$

where the gains $\{L_s\}_{s\in\mathcal{S}}$ are given by

$$L_s = (R_s + B_s^\intercal \bar{P}_s B_s)^{-1} B_s^\intercal \bar{P}_s A_s, \quad s \in \mathcal{S} \tag{7}$$

and $\{P_s\}_{s\in\mathcal{S}}$ is the solution of the following set of algebraic Riccati equations:

$$\bar{P}_s = \sum_{s_+\in\mathcal{S}} H_{ss_+} P_{s_+}, \quad s \in \mathcal{S}, \tag{8}$$

$$P_s = Q_s + A_s^\intercal \bar{P}_s A_s \tag{9}$$

$$- A_s^\intercal \bar{P}_s B_s^\intercal (R_s + B_s^\intercal \bar{P}_s B_s)^{-1} B_s^\intercal \bar{P}_s A_s, \quad s \in \mathcal{S}. \tag{10}$$

As established in [18, Theorem 45], the optimal control law is stabilizing in the following sense.

Proposition 3. *The autonomous system MJLS system $(\{A_s - B_s L_s\}_{s\in\mathcal{S}}, H)$ is MSS.*

Remark 1. The result of Proposition 3 in [18, Lemma 45] states that the system $(\{A_s - B_s L_s\}_{s\in\mathcal{S}}, H)$ is stochastically stable. As established in Proposition 1, stochastic stability is equivalent to MSS, so we have stated Prop. 3 in terms of MSS.

III. The learning problem

A. Some remarks on notation

1) Notation for probability spaces: We need a somewhat elaborate notation to describe our notion of regret. The MJLS system described above is a stochastic system with two stochastic inputs: the noise process $\{w_t\}_{t\geq 1}$ and the switching process $\{s_t\}_{t\geq 1}$. In addition, the learning algorithm may randomize while choosing control actions as well. We assume that the noise process and randomization done by the algorithm are defined on a probability space $(\Omega_1, \mathcal{F}_1, \mu_1)$ and the switching process is defined on a separate probability space $(\Omega_2, \mathcal{F}_2, \mu_2)$. Since the processes $\{w_t\}_{t\geq 1}$ and $\{s_t\}_{t\geq 1}$ and the randomization done by the algorithm are independent, we consider the probability space

$$(\Omega, \mathcal{F}, \mu) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \otimes \mu_2),$$

where $\mathcal{F}_1 \otimes \mathcal{F}_2$ is the product sigma algebra given by $\sigma(D_1 \times D_2 : D_1 \in \mathcal{F}_1, D_2 \in \mathcal{F}_2)$, and $\mu_1 \otimes \mu_2$ is the product measure on $\mathcal{F}_1 \otimes \mathcal{F}_2$, i.e., for any $D_1 \in \mathcal{F}_1$, $D_2 \in \mathcal{F}_2$, we have $\mu(D_1 \times D_2) = \mu_1(D_1)\mu_2(D_2)$. We will use the tuple $(\Omega, \mathcal{F}, \mu)$ as the probability space to define all the system variables. We abbreviate almost surely with respect to measure $\mu(\cdot)$ as μ-*a.s.* and almost surely with respect to measure $\mu_1(\cdot)$ as μ_1-*a.s.*.

2) Notation for policy dependent sample paths: To avoid confusion, we also use a slightly elaborate notation to indicate sample paths of state and action corresponding to a specific policy. Let $\theta = \{A_s, B_s\}_{s\in\mathcal{S}}$ denote the parameters of the system dynamics. Suppose the control input u_t is chosen as a function of the history of state and actions $(x_{1:t}, s_{1:t}, u_{1:t-1})$ according to a possibly randomized history-dependent measurable policy π. Then for any $\omega = (\omega_1, \omega_2) \in \Omega$, we use the notation $\{x_t^\pi(\omega), s_t(\omega_2), u_t^\pi(\omega)\}_{t\geq 1}$ to denote the states and the control actions along the sample path ω for the system when the controller is following policy π. Note that the discrete component of state, $s_t(\omega_2)$ only depends on ω_2 and does not depend on the policy π.

When it is clear from the context, we will not explicitly indicate the dependence on θ, π, and ω.

B. Regret definition

We are interested in the setting where the system parameters θ are unknown and the cost parameters $\{(Q_s, R_s)\}_{s\in\mathcal{S}}$ and transition matrix H are known. A learning agent observes the state (x_t, s_t) of the system and chooses the control input u_t according to a possibly history-dependent randomized measurable policy π. For any fixed realization $\omega_1 \in \Omega_1$ of the system noise and possible randomization by the algorithm, let

$$J_T^\pi(\omega_1) = \int_{\Omega_2} \sum_{t=1}^{T} c(x_t^\pi(\omega_1, \omega_2), s_t(\omega_2), u_t^\pi(\omega_1, \omega_2)) \mu_2(d\omega_2)$$

denote the performance of policy π along the sample path ω_1 for the horizon T averaged over the realizations of mode switching.

The (frequentist) **regret** of policy π is given by

$$\mathcal{R}_T^\pi(\omega_1) = J_T^\pi(\omega_1) - J_T^{\pi_\theta^*}(\omega_1)$$

where π_θ^* is the optimal policy corresponding to parameters θ.

Note that the notion of regret can be defined at different degrees of granularity. In particular, regret may be defined as a random variable which depends on the realization of the noise sequences and the randomizations done by the algorithm. Or it may be defined in terms of expectation over noise and algorithm randomization. In this paper, we take an intermediate approach: we define regret as a random variable which depends on the realization of the process noise and the randomizations done by the algorithm, but take the expectation over the discrete switching sequence.

IV. An upper bound on regret for adaptive linear policies with persistence of excitation

Let $\mathcal{F}_t = \sigma(x_{1:t-1}, s_{1:t-1}, u_{1:t-1})$ denote the sigma algebra generated by the observations of the history of states and actions of the learning agent at the beginning of time t. Motivated by the structure of the optimal policy presented in Theorem 1, we restrict attention to adaptive linear policies defined below.

Definition 4 (Adaptive linear policy)**.** *An adaptive linear policy π with persistence of excitation is characterized by a sequence of gains $\{\hat{L}_s(t) \in \mathbb{R}^{m\times n}\}_{s\in\mathcal{S}, t\geq 1}$, where $\{\hat{L}_s(t)\}_{s\in\mathcal{S}}$ is $\mathcal{F}_t$-measurable, and an independent noise process $\{\nu_t\}_{t\geq 1}$, $\nu_t \in \mathbb{R}^n$, where $\nu_t \sim \mathcal{N}(0, \sigma_t^2 I)$. The control input chosen by policy π at time t is given by $u_t = -\hat{L}_{s_t}(t)x_t + \nu_t$.*

Theorem 2. *Consider an adaptive linear policy π with persistence of excitation with gains $\{\hat{L}_s(t)\}_{s\in\mathcal{S},t\geq 1}$ and noise-level $\{\sigma_t^2\}_{t\geq 1}$. The regret of policy π may be decomposed as follows*

$$\mathcal{R}^\pi_T(\omega_1) = \mathcal{O}\big(\mathcal{R}^\pi_{1,T}(\omega_1)\big) + \mathcal{O}\big(\mathcal{R}^\pi_{2,T}(\omega_1)\big) + \mathcal{R}^\pi_{3,T}(\omega_1) \quad (11)$$

where

$$\mathcal{R}^\pi_{1,T}(\omega_1) = \int_{\Omega_2} \sum_{t=1}^{T} r^\pi_{1,t}(x^\pi_t(\omega_1,\omega_2), s_t(\omega_2))\mu_2(d\omega_2)$$

with $r^\pi_{1,t}(x_t, s_t)$ given by

$$x_t^\intercal(\hat{L}_{s_t}(t) - L_{s_t})^\intercal[R_{s_t} + B_{s_t}\bar{P}_{s_t}B_{s_t}](\hat{L}_{s_t}(t) - L_{s_t})x_t^\intercal,$$

and

$$\mathcal{R}^\pi_{2,T}(\omega_1) = \int_{\Omega_2} \sum_{t=1}^{T} r^\pi_{2,t}(\nu_t(\omega_1), s_t(\omega_2))\mu_2(d\omega_2)$$

with $r^\pi_{2,t}(\nu_t, s_t)$ given by $\nu_t^\intercal[R_{s_t} + B_{s_t}\bar{P}_{s_t}B_{s_t}]\nu_t$, and

$$\mathcal{R}^\pi_{3,T}(\omega_1) = \int_{\Omega_2} r^\pi_{3,t}(x^\pi_{T+1}(\omega), x^{\pi^*_\theta}_{T+1}(\omega), s_{T+1}(\omega_2))\mu_2(d\omega_2)$$

with $\omega = (\omega_1, \omega_2)$ and $r^\pi_{3,t}(x_{T+1}, x^{\pi^*_\theta}_{T+1}, s_{T+1})$ given by $(x^{\pi^*_\theta}_{T+1})^\intercal P_{s_{T+1}} x^{\pi^*_\theta}_{T+1} - x^\intercal_{T+1} P_{s_{T+1}} x_{T+1}$, where recall that $x^{\pi^*_\theta}$ denotes the state corresponding to the optimal policy π^*_θ.

The proof is presented in Appendix I.

V. A certainty equivalence based learning algorithm

A. Overview of the learning algorithm

We consider a specific type of certainty equivalence-based learning algorithm and analyze its regret by using Theorem 2. The algorithm consists of two phases: a *system identification phase* which lasts for a fixed time $T^{(0)}$; and an *adaptation phase*, which last for the remainder of the time that the system is running. The adaptation phase runs in episodes, and the length of k-th episode is $\lfloor \alpha^k T^{(0)} \rfloor$, where $\alpha > 1$ is a constant. We use $t^{(k)}$ to denote the start time of episode k and use $T^{(k)}$ to denote the length of episode k.

Before describing the two phases in detail, we need to define the notion of stabilizing gains.

Definition 5. *A set of gain matrices $\{\bar{L}_s \in \mathbb{R}^{m\times n}\}_{s\in\mathcal{S}}$ is said to be* stabilizing *for the MJLS system* (2) *if the autonomous system $(\{A_s - B_s\bar{L}_s\}_{s\in\mathcal{S}}, H)$ is MSS.*

We make the following assumption:

Assumption 2. *The learning agent has access to a set of stabilizing controllers $\{\bar{L}_s\}_{s\in\mathcal{S}}$.*

Assumption 2 is a common assumption in the literature of reinforcement learning for LQR systems [4], [7]–[9], [14]. During the system identification phase, the control input is chosen as $u_t = -\bar{L}_{s_t}x_t + \nu_t$, where ν_t is i.i.d., zero mean Gaussian random noise with covariance $I/\sqrt{T^{(0)}}$. We then use the system identification algorithm used in the next section to generate an initial estimate $\hat{\theta}^{(0)}$.

During episode k of the adaption phase, at time $t^{(k)}$, we pick control gains $\{\hat{L}^{(k)}_s\}_{s\in\mathcal{S}}$ to be the optimal control gains corresponding to the estimate $\hat{\theta}^{(k-1)}$. During the episode, we choose the control input as $u_t = -\hat{L}^{(k)}_{s_t}x_t + \nu_t$, where ν_t is i.i.d., zero mean Gaussian random noise with covariance $I/\sqrt{T^{(k)}}$. At the end of the k-th episode, we use the system identification algorithm described in the next section to generate a new estimate $\hat{\theta}^{(k)}$ based on all the data seen in episode k.

A detailed description of the learning algorithm is presented in Algorithm 1.

B. The system identification algorithm

In this section, we describe the system identification algorithm used in both phases. This algorithm is a variation of the switched least squares system identification algorithm presented in [1], [2] for autonomous system.

For uniformity of notation, we allow $k = 0$ to mean the system identification phase and set $t^{(0)} = 1$ and $\hat{L}^{(0)}_s = \bar{L}_s$ for $s \in \mathcal{S}$. Now consider a generic k-th episode, $k \in \{0, 1, \dots\}$, which is of length $T^{(k)}$. During this episode, the control input is chosen as

$$u_t = -\hat{L}^{(k)}_{s_t}x_t + \nu_t,$$

where ν_t is random noise chosen as $\nu_t \sim \mathcal{N}(0, \sigma^2_{(k)}I)$, where $\sigma^2_{(k)} = 1/\sqrt{T^{(k)}}$. Thus, Eq. (2) may be written as

$$x_{t+1} = A_{s_t}x_t - B_{s_t}\hat{L}^{(k)}_{s_t}x_t + B_{s_t}\nu_t + w_t \quad (12)$$

```
Algorithm 1: Certainty equiv. based learning algorithm
input   : A set of stabilizing controllers $\{\bar{L}_s\}_{s\in\mathcal{S}}$
          Time $T^{(0)}$; Scaling factor $\alpha > 1$.

System ID :
Initialize $\hat{L}_s^{(0)} = \bar{L}_s$, for all $s \in \mathcal{S}$.
Initialize $t^{(0)} = 1$.
for time $t \in \{t^{(0)}, \dots, t^{(0)} + T^{(0)} - 1\}$ do
    Sample $\nu_t \sim \mathcal{N}(0, \sigma^2_{(0)} I)$, where $\sigma^2_{(0)} = 1/\sqrt{T^{(0)}}$.
    Apply control input $u_t = \hat{L}^{(0)}_{s_t} x_t + \nu_t$.
end

Generate estimate $\hat{\theta}^{(0)}$ using (14) and (15).

Adaptation:
for episode $k = 1, 2, \dots$ do
    Initialize $t^{(k)} = t + 1$; $T^{(k)} = \lfloor \alpha^k T^{(0)} \rfloor$.
    Choose $\{\hat{L}^{(k)}_s\}_{s\in\mathcal{S}}$ using (7) for system $\hat{\theta}^{(k-1)}$.
    Set $\sigma^2_{(k)} = 1/\sqrt{T^{(k)}}$.
    for time $t \in \{t^{(k)}, \dots, t^{(k)} + T^{(k)} - 1\}$ do
        Sample $\nu_t \sim \mathcal{N}(0, \sigma^2_{(k)} I)$.
        Apply control input $u_t = \hat{L}^{(k)}_{s_t} x_t + \nu_t$.
    end
    Generate estimate $\hat{\theta}^{(k)}$ using (14) and (15)
end
```

or, equivalently,

$$x_{t+1} = \eta^{(k)}_{s_t} z_t + w_t. \tag{13}$$

where $\{\eta^{(k)}_s \in \mathbb{R}^{n\times(n+m)}\}_{s\in\mathcal{S}}$ is given by

$$\eta^{(k)}_s := [A_s - B_s \hat{L}^{(k)}_s, B_s], \quad s \in \mathcal{S},$$

and $z_t^\intercal := [x_t^\intercal, \nu_t^\intercal] \in \mathbb{R}^{n+m}$.

At the end of the episode, we generate estimates $\hat{\eta}^{(k)} := \{\hat{\eta}^{(k)}_s \in \mathbb{R}^{n\times(n+m)}\}_{s\in\mathcal{S}}$ by solving the following switched least squares problem:

$$\hat{\eta}^{(k)} = \mathop{\arg\min}_{\eta^{(k)} = \{\eta^{(k)}_s : s\in\mathcal{S}\}} \sum_{t=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|x_{t+1} - \eta^{(k)}_{s_t} z_t\|^2. \tag{14}$$

We then compute estimates $\{\hat{B}^{(k)}_s\}_{s\in\mathcal{S}}$ and $\{\hat{A}^{(k)}_s\}_{s\in\mathcal{S}}$ as:

$$\hat{B}^{(k)}_s = \hat{\eta}^{(k)}_s \begin{bmatrix} 0_{n\times n} \\ I_{m\times n} \end{bmatrix}, \quad \hat{A}^{(k)}_s = \hat{\eta}^{(k)}_s \begin{bmatrix} I_{n\times n} \\ \hat{L}^{(k)}_s \end{bmatrix}, \quad s \in \mathcal{S}. \tag{15}$$

We denote the estimated parameters as $\hat{\theta}^{(k)}_s := [\hat{A}^{(k)}_s, \hat{B}^{(k)}_s]$, $s \in \mathcal{S}$ and use $\hat{\theta}^{(k)} := \{\hat{\theta}^{(k)}_s\}_{s\in\mathcal{S}}$ to denote the estimated parameters of the model.

VI. The main results

A. Asymptotic regret of certainty equivalence algorithm

In our analysis, we need to assume that the proposed learning algorithm at all times generates estimates such that the gains corresponding to those estimates stabilize the original system.

Definition 6. *Given the set of stabilizing controllers $\{\bar{L}_s\}_{s\in\mathcal{S}}$, time T_0 and scaling factor α, let $\mathcal{A}_0$ be the set of all sample paths $\omega_1 \in \Omega_1$ such that for almost all $\omega_2 \in \Omega_2$ and $k \geq 1$ the gains $\{\hat{L}^{(k)}_s(\omega_1, \omega_2)\}_{s\in\mathcal{S}}$ are stabilizing for MJLS system* (2).

Assumption 3. *We assume $\mu_1(\mathcal{A}_0) > 0$.*

In our results below, we restrict attention to the sample paths $\omega_1 \in \mathcal{A}_0$. Note that the process $\{s_t\}_{t\geq 0}$ remains Markov on the set $\mathcal{A}_0 \times \Omega_2$ with the same transition probabilities. We assume that $\mu_1(\mathcal{A}_0) > 0$, which is weaker than the stability assumption implicitly imposed in [4] for (non-switching) LQR model, where it was assumed that $\mu(\mathcal{A}_0) = 1$.

By an argument similar to that used in [2] for autonomous systems, we can show that if the controller used in an episode is stable and the episode is asymptotically large, the estimates generated by switched least squares system identification algorithm described in Sec. V-B converge almost surely to the correct parameters. We can also characterize the rate of convergence, as shown below:

Theorem 3. *On the set $\mathcal{A}_0$, the estimate $\hat{\theta}^{(k)}$ is strongly consistent, i.e. $\lim_{k\to\infty} \|\hat{\theta}^{(k)} - \theta\| = 0$, μ_1-a.s. Furthermore, the error of the system identification method is upper bounded by:*

$$\limsup_{k\to\infty} \frac{\|\hat{\theta}^{(k)} - \theta\|}{\sqrt{\log(T^{(k)})/\sigma^{(k)} T^{(k)}}} < \infty, \quad \mu_1\text{-}a.s. \tag{16}$$

The proof is presented in Appendix II.

Following theorem establishes the regret bound for Algorithm 1. This regret matches with the regret of LQR problems established in [3]–[5], [7], [9] and the regret of MJLS-LQR established in [14].

Theorem 4. *On the set $\mathcal{A}_0$, the regret of Algorithm 1 is given by:*

$$\mathcal{R}^{\hat{\pi}}_T \leq \mathcal{O}(\sqrt{T}\log(T)) \quad \mu_1\text{-}a.s.$$

The proof is presented in Appendix III.

B. Sufficient conditions for stability

In characterizing the almost sure regret of adaptive control problems, ensuring the stability of the system is a challenging problem. Our results in Theorems 3 and 4 are derived on the set $\mathcal{A}_0$. In this section, we try to weaken this requirement by characterizing a set which is larger than $\mathcal{A}_0$. For the MJLS system in (2) with parameters θ, let $L^\theta = \{L^\theta_s\}_{s\in\mathcal{S}}$ denote the set of optimal control gains. Define:

$$\mathcal{B}_\epsilon(L^\theta) := \Big\{\{\hat{L}_s\}_{s\in\mathcal{S}} : \|\hat{L}_s - L^\theta_s\| \leq \epsilon, \forall s \in \mathcal{S}\Big\},$$

as a ball in the space of gain matrices with radius ϵ centered at L^θ.

Lemma 1. *[14, Lemma C.1] For the MJLS in* (2), *there exists a radius ϵ_θ such that all the gains $\{\hat{L}_s\}_{s\in\mathcal{S}} \in \mathcal{B}_{\epsilon_\theta}(L^\theta)$ are stabilizing for θ.*

We define $\mathcal{B}_\delta(\theta) \coloneqq \big\{\{\hat{\theta}_s\}_{s\in\mathcal{S}} : \|\hat{\theta}_s - \theta_s\| \le \delta, \forall s \in \mathcal{S}\big\}$. Now let δ_θ be the radius such that if $\hat{\theta} \in \mathcal{B}_{\delta_\theta}(\theta)$ then $L^{\hat{\theta}} \in \mathcal{B}_{\epsilon_\theta}(L_\theta)$.

We now characterize the connection between the assumptions on the stability and length of the identification phase $T^{(0)}$. Consider a system identification setup in which we use adaptive linear policy $\{\mathring{L}_s\}_{s\in\mathcal{S}}$ with persistent of excitation $\nu_t \sim \mathcal{N}(0, \mathring{\sigma}^2 I)$, where $\{\mathring{L}_s\}_{s\in\mathcal{S}}$ is a stabilizing controller. We get $x_{t+1} = \mathring{\eta}_{s_t} z_t + w_t$, where $\mathring{\eta}_s \coloneqq [A_s - B_s\mathring{L}_s, B_s]$. We estimate $\hat{\mathring{\eta}}_T \coloneqq \{\hat{\mathring{\eta}}_{s,T} \in \mathbb{R}^{n\times(n+m)}\}_{s\in\mathcal{S}}$ by solving:

$$\hat{\mathring{\eta}}_T = \operatorname*{arg\,min}_{\mathring{\eta}=\{\mathring{\eta}_s : s\in\mathcal{S}\}} \sum_{t=1}^{T} \|x_{t+1} - \mathring{\eta}_{s_t} z_t\|^2. \tag{17}$$

We generate the estimate $\hat{\theta}_T$ from $\hat{\mathring{\eta}}_T$ similarly to (15). To explicitly emphasize the functional dependence of $\hat{\theta}_T$ on $\mathring{L} = \{\mathring{L}_s\}_{s\in\mathcal{S}}$ and $\omega \in \Omega$, we use the notation $\hat{\theta}_T(\mathring{L}, \omega)$. Similar to Theorem 3, we can establish that if $\{\mathring{L}_s\}_{s\in\mathcal{S}}$ is stabilizing controller, then $\lim_{T\to\infty} \|\hat{\theta}_T(\mathring{L}, \omega) - \theta\| = 0$, μ-$a.s.$ and the error of the system identification is upper bounded by:

$$\limsup_{T\to\infty} \frac{\|\hat{\theta}_T(\mathring{L}, \omega) - \theta\|}{\sqrt{\log(T)/\mathring{\sigma}^2 T}} < \infty, \quad \mu\text{-}a.s. \tag{18}$$

Now for any generic stabilizing gain $\tilde{L} = \{\tilde{L}_s\}_{s\in\mathcal{S}}$, define

$$T_{\delta_\theta}(\tilde{L}, \omega) \coloneqq \inf\{T \in \mathbb{N} : \forall t \ge T, \|\hat{\theta}_T(\tilde{L}, \omega) - \theta\| \le \delta_\theta\},$$
$$\bar{T}_{\delta_\theta}(\omega) \coloneqq \sup_{\tilde{L}\in\mathcal{B}_{\epsilon_\theta}(L^\theta)} T_{\delta_\theta}(\tilde{L}, \omega).$$

A consequence of the result in (18) is that for any stabilizing $\tilde{L}$, $\mathbb{P}(T_{\delta_\theta}(\tilde{L}, \omega) < \infty) = 1$ and consequently $\mathbb{P}(\bar{T}_{\delta_\theta}(\omega) < \infty) = 1$. For any $T > 0$, define

$$\mathcal{A}_{\delta_\theta}(T) \coloneqq \{\omega \in \Omega : \bar{T}_{\delta_\theta}(\omega) \le T - 1\}.$$

Proposition 4. *The set $\mathcal{A}_{\delta_\theta}(T)$ satisfies following properties:*

1) *If $T < T'$, we have $\mathcal{A}_{\delta_\theta}(T) \subseteq \mathcal{A}_{\delta_\theta}(T')$.*
2) *For any $\omega \in \Omega$, there exists a T_0 such that: $\omega \in \mathcal{A}_{\delta_\theta}(T_0) \subseteq \Omega$, μ-$a.s.$*
3) *There exists a $T_0 < \infty$ such that $\Omega = \mathcal{A}_{\delta_\theta}(T_0)$, μ-$a.s.$*

The proof is omitted due to space constraints; however, this proposition is a consequence of the result in Theorem 3.

Theorem 5. ***(Sufficient condition for stability)*** *Suppose the initial stabilizing controller $\{\bar{L}_s\}_{s\in\mathcal{S}} \in \mathcal{B}_{\epsilon_\theta}(L^\theta)$, then the results of Theorem 3 and 4 are valid on the set $\mathcal{A}_{\delta_\theta}(T^{(0)})$.*

The proof is presented in IV.

VII. Conclusion and Future Directions

In this paper, we investigate the problem of simultaneous learning and control of a Markov jump linear system using complete state observation. We derive an almost sure regret decomposition for the general class of adaptive linear policies with persistence of excitation. We propose a version of certainty equivalence controller which uses the switched least squares method for the closed-loop system identification. Our analysis shows that the error of the system identification method is $\mathcal{O}(\sqrt{\log(T)/T})$, and the regret of certainty equivalence controller reaches $\mathcal{O}(\sqrt{T}\log(T))$ almost surely. Our guarantees are stated for specific subset of Ω. We show we can make this subset arbitrary large by increasing $T^{(0)}$. Finding an algorithm with performance guarantees independent of the set $\mathcal{A}_{\delta_\theta}(T^{(0)})$, extending these results to the case of partial observation and analyzing algorithms such as Thompson sampling with the tool developed in Theorem 2 is left for future works.

References

[1] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Strong consistency and rate of convergence of switched least squares system identification for autonomous markov jump linear systems," *arXiv preprint arXiv:2112.10753*, 2021.

[2] ——, "Consistency and rate of convergence of switched least squares system identification for autonomous markov jump linear systems," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 6678–6685.

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

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

[5] ——, "Optimism-based adaptive regulation of linear-quadratic systems," *IEEE Trans. Autom. Control*, vol. 66, no. 4, pp. 1802–1808, 2020.

[6] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems," *arXiv preprint arXiv:2003.11227*, 2020.

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

[8] M. Abeille and A. Lazaric, "Improved regret bounds for thompson sampling in linear quadratic control problems," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1–9.

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

[10] G. S. Deaecto, M. Souza, and J. C. Geromel, "Discrete-time switched linear systems state feedback design with application to networked control," *IEEE Trans. Autom. Control*, vol. 60, no. 3, pp. 877–881, 2014.

[11] C. De Persis and P. Tesi, "Input-to-state stabilizing control under denial-of-service," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2930–2944, 2015.

[12] A. Cetinkaya, H. Ishii, and T. Hayakawa, "Analysis of stochastic switched systems with application to networked control under jamming attacks," *IEEE Trans. Autom. Control*, vol. 64, no. 5, pp. 2013–2028, 2018.

[13] Z. Du, Y. Sattar, D. A. Tarzanagh, L. Balzano, S. Oymak, and N. Ozay, "Certainty equivalent quadratic control for markov jump systems," *arXiv preprint arXiv:2105.12358*, 2021.

[14] Y. Sattar, Z. Du, D. A. Tarzanagh, L. Balzano, N. Ozay, and S. Oymak, "Identification and adaptive control of Markov jump systems: Sample complexity and regret bounds," *arXiv preprint arXiv:2111.07018*, 2021.

[15] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Policy optimization for markovian jump linear quadratic control: Gradient method and global convergence," *IEEE Transactions on Automatic Control*, 2022.

[16] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Thompson-sampling based reinforcement learning for networked control of unknown linear systems," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 723–730.

[17] O. L. V. Costa, M. D. Fragoso, and R. P. Marques, *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.

[18] A. Czornik, *On control problems for jump linear systems*. Wydawn. Politechniki Śląskiej, 2003.

[19] T. L. Lai and C. Z. Wei, "Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems," *Multivariate Analysis VI*, pp. 375–393, 1985.

[20] ——, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *Ann. Statist.*, vol. 10, no. 1, pp. 154–166, 1982.

Appendix I
Proof of Theorem 2

We start with the completion of squares lemma.

Lemma 2. *For $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ and matrices A, B, S, R with appropriate dimensions, we have*

$$\begin{aligned}&u^\intercal Ru + (Ax+Bu)^\intercal P(Ax+Bu) + x^\intercal Qx = \\ &(u + L(P,R,A,B)x)^\intercal [R + B^\intercal PB](u + L(P,R,A,B)x) \\ &\qquad\qquad\qquad\qquad + x^\intercal K(P,A,B,R,Q)x,\end{aligned}$$

where

$$\begin{aligned}L(P,R,A,B) &\coloneqq -(R+B^\intercal PB)^{-1}B^\intercal PA. \\ K(P,A,B,R,Q) &\coloneqq Q + A^\intercal PA \\ &\quad - A^\intercal PB(R+B^\intercal PB)^{-1}B^\intercal PA.\end{aligned}$$

Remark 2. Notice that in (8), we have:

$$L_s = L(\bar{P}_s, R_s, A_s, B_s), \quad P_s = K(\bar{P}_s, A_s, B_s, R_s, Q_s).$$

We assume that π and ω_1 are fixed and do not explicitly include their dependence on the terms. Instead, we will use x_t as a short-hand for $x_t^\pi(\omega)$ and x_t^* as a short-hand for $x_t^{\pi_\theta^*}(\omega)$. We also use $\tilde{s}_t$ instead of $s_t(\omega_2)$, where we use the superscript tilde to highlight the fact that we are not referring to a specific realization of the discrete state at time t rather marginalizing over all possible realizations. By recursively applying completion of squares (Lemma 2), we can show the following:

Lemma 3. *For any policy π we have*

$$\begin{aligned}&\int_{\Omega_2}\Bigg[\sum_{t=1}^T c(x_t, s_t, u_t) + x_{T+1}^\intercal P_{s_{T+1}} x_{T+1}\Bigg]\mu_2(d\omega_2) \\ &\quad = \int_{\Omega_2}\Bigg[x_1^\intercal \bar{P}_{\tilde{s}_1} x_1 \\ &\qquad + \sum_{t=1}^T (u_t + L_{\tilde{s}_t}x_t)^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}](u_t + L_{\tilde{s}_t}x_t) \\ &\qquad + \sum_{t=1}^T \Big[2w_t^\intercal \bar{P}_{\tilde{s}_t}(A_{\tilde{s}_t}x_t + B_{\tilde{s}_t}u_t) + w_t^\intercal \bar{P}_{\tilde{s}_t} w_t\Big]\Bigg]\mu_2(d\omega_2).\end{aligned} \tag{19}$$

Using the decomposition in (19) in the expression for regret, and substituting $u_t = -\hat{L}_{\tilde{s}_t}(t)x_t + \nu_t$ for policy π and substituting $u_t = -L_{\tilde{s}_t}x_t$ for policy π^*, we get the following:

Lemma 4. *For any adaptive linear policy π with persistence of excitation, we have*

$$\begin{aligned}R_T^\pi(\omega_1) = \int_{\Omega_2}\Bigg[&\sum_{t=1}^T x_t^\intercal(\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}] \\ &\qquad\qquad\qquad (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})x_t \\ + &\sum_{t=1}^T \big[\nu_t^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}]\nu_t \\ &\qquad + 2\nu_t^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}](\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})x_t\big] \\ + &\sum_{t=1}^T 2w_t^\intercal \bar{P}_{\tilde{s}_t}\big[(A_{\tilde{s}_t} - B_{\tilde{s}_t}L_{\tilde{s}_t})(x_t - x_t^*) \\ &\qquad - B_{\tilde{s}_t}(\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})x_t + B_{\tilde{s}_t}\nu_t\big] \\ + &[(x_{T+1}^*)^\intercal \bar{P}_{\tilde{s}_{T+1}} x_{T+1}^* - x_{T+1}^\intercal \bar{P}_{\tilde{s}_{T+1}} x_{T+1}]\Bigg]\mu_2(d\omega_2).\end{aligned} \tag{20}$$

We first recall the following result [19, Corollary 10]

Lemma 5. *Given a filtration $\{\mathcal{F}_t\}_{t\geq 1}$, suppose w_t is a martingale difference process adapted to $\{\mathcal{F}_t\}_{t\geq 1}$ and y_{t+1} is $\mathcal{F}_t$-measurable. Then,*

$$\sum_{t=1}^T y_t^\intercal w_t = \mathcal{O}\Big(\sqrt{Y_T \log(Y_T)}\Big), \quad a.s.$$

where $Y_T = \sum_{t=1}^T y_t^\intercal y_t$.

An implication of Lemma 5 is that

$$\begin{aligned}\int_{\Omega_2}\Bigg[\sum_{t=1}^T w_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}(\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})x_t\Bigg]\mu_2(d\omega_2) \\ = \mathcal{O}\Big(\sqrt{\mathcal{R}_{1,T}^\pi(\omega_1)\log \mathcal{R}_{1,T}^\pi(\omega_1)}\Big),\end{aligned} \tag{21}$$

where $\mathcal{R}_{1,T}^\pi(\omega_1)$ is defined in Theorem 2. By the same argument, we also have

$$\begin{aligned}\int_{\Omega_2}\Bigg[\sum_{t=1}^T \nu_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}(\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})x_t\Bigg]\mu_2(d\omega_2) \\ = \mathcal{O}\Big(\sqrt{\mathcal{R}_{1,T}^\pi(\omega_1)\log \mathcal{R}_{1,T}^\pi(\omega_1)}\Big),\end{aligned} \tag{22}$$

and

$$\begin{aligned}\int_{\Omega_2}\Bigg[\sum_{t=1}^T 2w_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}\nu_t\Bigg]\mu_2(d\omega_2) \\ = \mathcal{O}\Big(\sqrt{\mathcal{R}_{2,T}^\pi(\omega_1)\log \mathcal{R}_{2,T}^\pi(\omega_1)}\Big).\end{aligned} \tag{23}$$

Now, by Prop. 3, the autonomous MJLS system $(\{A_s - B_sL_s\}_{s\in\mathcal{S}}, H)$ is MSS. Based on the fact that MSS implies exponential stochastic stability (Prop. 1), we can show that

$$\begin{aligned}&\int_{\Omega_2}\Bigg[\sum_{t=1}^T w_t^\intercal \bar{P}_{\tilde{s}_t}(A_{\tilde{s}_t} - B_{\tilde{s}_t}L_{\tilde{s}_t})(x_t - x_t^*)\Bigg]\mu_2(d\omega_2) \\ &\quad = \mathcal{O}\Bigg(\int_{\Omega_2}\Bigg[\sum_{t=1}^T w_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}(\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})x_t\Bigg]\mu_2(d\omega_2)\Bigg)\end{aligned} \tag{24}$$

which is therefore also upper bounded by the right hand side of (21). The result of Theorem 2 then follows from substituting (21)–(24) in Lemma 4.

Appendix II
Proof of Theorem 3

The proof of this theorem is based on a notion of stability called stability in the average sense in [1].

Definition 7. *Let $\{x_t\}_{t\geq 1}$ denote the state process corresponding to the MJLS system $A_{s_t}x_t + w_t$. We say this system is stable in the average sense, if: $\sum_{t=1}^{T}\|x_t\|^2 = \mathcal{O}(T)\quad \mu\text{-}a.s.$,*

Proposition 5. *[1, Proposition 3] If the MJLS system $(\{A_s\}_{s\in\mathcal{S}}, H)$ is MSS, then the MJLS system: $A_{s_t}x_t + w_t$ is stable in the average sense.*

By the assumptions in Theorem 3, we know $(\{A_s - B_sL_s\}_{s\in\mathcal{S}}, H)$ is MSS; therefore, by Proposition 5, and the fact that σ_t^2 is finite, we get that MJLS $x_{t+1} = (A_{s_t} - B_{s_t}\hat{L}^{(k)}_{s_t})x_t + B_{s_t}\nu_t + w_t$ is stable in the average sense. Recall that $\eta^{(k)}_{s_t} := \left[A_{s_t} - B_{s_t}\hat{L}^{(k)}_{s_t}, B_{s_t}\right]$, and $z_t := \begin{bmatrix} x_t \\ \nu_t\end{bmatrix}$, and we have:

$$x_{t+1} = \eta^{(k)}_{s_t} z_t + w_t. \tag{25}$$

Let $\mathcal{T}^{(k)}_{i,T} = \{t^{(k)} \leq t < t^{(k)} + T : s_t = i\}$ denote the time indices until the time T, when the discrete state of the system equals i at the k-th episode. Note that for $t \in \mathcal{T}^{(k)}_{i,T}$, $\eta_{s_t} = \eta_i$. Therefore, we have:

$$\hat{\eta}^{(k)}_{i,T} := \arg\min_{\eta} \sum_{t\in\mathcal{T}^{(k)}_{i,T}} \|x_{t+1} - \eta_i z_t\|^2, \quad \forall i \in \{1,\dots,d\}.$$

Let $Z_{i,T}$ denote $\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T}$, which we call the unnormalized empirical covariance of the augmented state process when $s_t = i$. Now we look at $\lambda_{\max}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T})$ and $\lambda_{\min}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T})$. We have:

$$\begin{aligned}\lambda_{\max}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T}) &\leq \operatorname{tr}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T}) \\ &= \sum_{t\in\mathcal{T}^{(k)}_{i,T}} \|z_t\|^2 \leq \sum_{t=1}^{T}\|z_t\|^2\end{aligned}$$

By Proposition 5, we know $\sum_{t=1}^{T}\|x_t\|^2 = \mathcal{O}(T)$ μ-$a.s.$ and by [19, Eq. 3.1] we know $\sum_{t=1}^{T}\|\nu_t\|^2 = \mathcal{O}(T)$ μ-$a.s.$, which implies:

$$\lambda_{\max}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T}) = \mathcal{O}(T) \quad \mu\text{-}a.s.$$

On the other hand, we have:

$$z_t z_t^\mathsf{T} = \begin{bmatrix} x_t x_t^\mathsf{T} & x_t\nu_t^\mathsf{T} \\ \nu_t x_t^\mathsf{T} & \nu_t\nu_t^\mathsf{T}\end{bmatrix}$$

Similar to [1, Lemma 3], we can show $\sum_{t\in\mathcal{T}^{(k)}_{i,T}} \|x_t\nu_t^\mathsf{T} + \nu_t x_t^\mathsf{T}\| = o(T)$ μ-$a.s.$; therefore,

$$\begin{aligned}&\lambda_{\min}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T}) \\ &= \mathcal{O}\big(\min\big\{\lambda_{\min}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} x_t x_t^\mathsf{T}), \lambda_{\min}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} \nu_t\nu_t^\mathsf{T})\big\}\big) \quad \mu\text{-}a.s.\end{aligned}$$

By [1, Proposition 1-P2], we know $\liminf_{T\to\infty}\lambda_{\min}(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\mathsf{T})/\mathcal{T}^{(k)}_{i,T} \geq 0$ μ-$a.s.$, and since $\sigma^2_{(k)} > 0$, we get $\liminf_{k\to\infty}\lambda_{\min}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} \nu_t\nu_t^\mathsf{T})/\mathcal{T}^{(k)}_{i,T} \geq \sigma^2_{(k)}$ μ-$a.s.$ Therefore,

$$\lambda_{\min}(\sum_{t\in\mathcal{T}^{(k)}_{i,T}} z_t z_t^\mathsf{T}) > \sigma^2_{(k)}\mathcal{T}_{i,T} \quad \mu\text{-}a.s.$$

Given a filtration $\{\mathcal{G}_t\}_{t\geq 0}$, consider the following regression model:

$$y_t = \beta^\mathsf{T} z_t + w_t, \quad t \geq 0, \tag{26}$$

where $\beta \in \mathbb{R}^n$ is an unknown parameter, $z_t \in \mathbb{R}^n$ is $\mathcal{G}_{t-1}$-measurable covariate process, y_t is the observation process, and $w_t \in \mathbb{R}$ is a noise process. Then the least squares estimate $\hat{\beta}_T$ of β is given by:

$$\hat{\beta}_T = \arg\min_{\beta^\mathsf{T}} \sum_{\tau=0}^{T}\|y_\tau - \beta^\mathsf{T} z_\tau\|^2. \tag{27}$$

The following result by [20] characterizes the rate of convergence of $\hat{\beta}_T$ to β in terms of unnormalized covariance matrix of covariates $Z_T := \sum_{\tau=0}^{T} z_\tau z_\tau^\mathsf{T}$.

Theorem 6 (see [20, Theorem 1]). *Suppose the following conditions are satisfied:* **(S1)** *$\lambda_{\min}(Z_T) \to \infty$, a.s. and* **(S2)** *$\log(\lambda_{\max}(Z_T)) = o(\lambda_{\min}(Z_T))$, a.s. Then the least squares estimate in* (27) *is strongly consistent with the rate of convergence:*

$$\|\hat{\beta}_T - \beta\|_\infty = \mathcal{O}\Bigg(\sqrt{\frac{\log\left[\lambda_{\max}(Z_T)\right]}{\lambda_{\min}(Z_T)}}\Bigg) \quad a.s.$$

Therefore, by Theorem 6, and the fact that $\sigma^2_{(k)}\mathcal{T}_{i,T} = \mathcal{O}(\sigma^2_{(k)}T)$ we get that:

$$\lim_{k\to\infty}\|\hat{\theta}_{(k)} - \theta\| = \mathcal{O}\big(\sqrt{\log(T^{(k)})/\sigma^{(k)}T^{(k)}}\big) \quad \mu_1\text{-}a.s.$$

Appendix III
Proof of Theorem 4

Lemma 6. *The regret in the k-th episode satisfies:*

$$\limsup_{k\to\infty}\frac{\int_{\Omega_2}\left[\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} c(x_\tau, s_\tau, u_\tau)\right]\mu_2(d\omega_2)}{(T^{(k-1)})^{1/2}\log(T^{(k-1)})} < \infty \quad \mu\text{-}a.s.$$

Proof. (*Sketch*) In the k-th episode, $\hat{L}^{(k)}_s$ is computed based on the estimate $\hat{\theta}^{(k-1)}$. Assumption 3 implies that on the set $\mathcal{A}_0$, the set of the gains $\{\hat{L}^{(k)}_s\}_{s\in\mathcal{S}}$ is stabilizing for all $k \geq 0$. Setting $\sigma^2_{(k-1)} = 1/\sqrt{T^{(k-1)}}$ in Theorem 3 implies:

$$\limsup_{k\to\infty}\frac{\|\hat{\theta}^{(k-1)} - \theta\|}{\sqrt{\log(T^{(k-1)})}/(T^{(k-1)})^{1/4}} < \infty, \quad \mu\text{-}a.s. \tag{28}$$

By the continuity of the gains $\hat{L}_s^{(k)}$ in the parameter $\hat{\theta}^{(k-1)}$ we get,

$$\limsup_{k\to\infty} \frac{\|\hat{L}_s^{(k)} - L_s\|}{\sqrt{\log(T^{(k-1)})}/(T^{(k-1)})^{1/4}} < \infty, \quad \mu\text{-}a.s. \quad (29)$$

In the proof of Theorem 3, we established that

$$\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|x_t\|^2 = \mathcal{O}(T^{(k)}) \quad \mu\text{-}a.s.$$

The gain $\hat{L}_s^{(k)}$ is fixed during the episode. Therefore by plugging in $\|\hat{L}_s^{(k)} - L_s\|$, $\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|x_t\|^2$ in the $\mathcal{R}_{1,T}^{\pi}(\omega_1)$, and $\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|\nu_t\|^2 = \mathcal{O}(\sqrt{T^{(k)}})$ in $\mathcal{R}_{2,T}^{\pi}$, we get the desired result. □

Let $\tilde{T}^{(k)} = \sum_{m=0}^{k} T^{(m)} = T^{(0)}(\alpha^{k+1}-1)/(\alpha-1)$, which implies $k = \mathcal{O}(\log_\alpha(\tilde{T}^{(k)}/T^{(0)})$.

Proof of Theorem 4 follows by adding the terms in previous lemma. Due to the limited space, the proof is omitted.

APPENDIX IV
PROOF OF THEOREM 5

We prove this result by induction. We show on the set $\mathcal{A}_{\delta_\theta}(T^{(0)})$ if $\hat{\theta}^{(k)} \in \mathcal{B}_{\delta_\theta}(\theta)$, then $\hat{\theta}^{(k+1)} \in \mathcal{B}_{\delta_\theta}(\theta)$. As the basis of induction, since $\{\bar{L}_s\}_{s\in\mathcal{S}} \in \mathcal{B}_{\epsilon_\theta}(L_\theta)$, then Theorem 3 and the definition of $\mathcal{A}_{\delta_\theta}(T^{(0)})$ imply that $\hat{\theta}^{(0)} \in \mathcal{B}_{\delta_\theta}(\theta)$.

Now assume that $\hat{\theta}^{(k)} \in \mathcal{B}_{\delta_\theta}(\theta)$. Lemma 1 implies that $\{\hat{L}_s^{(k)}\}_{s\in\mathcal{S}}$ is stabilizing. Moreover, since $T^{(k)} \geq T^{(0)}$, Theorem 3 and definition of $\mathcal{A}_{\epsilon_\theta}(T^{(0)})$ imply that $\|\hat{\theta}^{(k+1)} - \theta\| \leq \epsilon_\theta$. Hence $\hat{\theta}^{(k+1)} \in \mathcal{B}_{\delta_\theta}(\theta)$. This completes the proof of the induction step.