2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Data-Driven Control of Markov Jump Systems: Sample Complexity and Regret Bounds

Zhe Du*[1] Yahya Sattar*[2] Davoud Ataee Tarzanagh[1]
Laura Balzano[1] Necmiye Ozay[1] Samet Oymak[2]

***Abstract*— Learning how to effectively control unknown dynamical systems from data is crucial for intelligent autonomous systems. This task becomes a significant challenge when the underlying dynamics are changing with time. Motivated by this challenge, this paper considers the problem of controlling an unknown Markov jump linear system (MJS) to optimize a quadratic objective in a data-driven way. By taking a model-based perspective, we consider identification-based adaptive control for MJS. We first provide a system identification algorithm for MJS to learn the dynamics in each mode as well as the Markov transition matrix, underlying the evolution of the mode switches, from a single trajectory of the system states, inputs, and modes. Through mixing-time arguments, sample complexity of this algorithm is shown to be $\mathcal{O}(1/\sqrt{T})$. We then propose an adaptive control scheme that performs system identification together with certainty equivalent control to adapt the controllers in an episodic fashion. Combining our sample complexity results with recent perturbation results for certainty equivalent control, we prove that when the episode lengths are appropriately chosen, the proposed adaptive control scheme achieves $\mathcal{O}(\sqrt{T})$ regret. Our proof strategy introduces innovations to handle Markovian jumps and a weaker notion of stability common in MJSs. Our analysis provides insights into system theoretic quantities that affect learning accuracy and control performance. Numerical simulations are presented to further reinforce these insights.**

I. Introduction

A canonical problem at the intersection of control and machine learning is that of adaptive control of an unknown dynamical system. An intelligent autonomous system is likely to encounter such a task; from an observation of the inputs and outputs, it needs to both learn and effectively control the dynamics. A commonly used control paradigm is the Linear Quadratic Regulator (LQR), which is theoretically well understood when system dynamics are linear and known. LQR also provides an interesting benchmark, when system dynamics are unknown, for reinforcement learning (RL) with continuous state and action spaces and for adaptive control [3], [4], [5], [6], [7], [8], [9].

LQR is also generalized to Markov jump linear systems (MJSs) and well understood when system dynamics are known [10], [11]. However, in practice, it is not always possible to have a perfect knowledge of the system dynamics and the Markov transition matrix. For instance, a Mars rover optimally exploring an unknown heterogeneous terrain, optimal solar power generation on a cloudy day, or controlling investments in financial markets may be modeled as MJS-LQR problems with unknown system dynamics [12], [13], [14], [15], [16]. Earlier works have aimed at analyzing the asymptotic properties (i.e., stability) of adaptive controllers for unknown MJSs both in continuous-time [17] and discrete-time [18] settings. For MJSs with independent mode switching, a.k.a. stochastic jump systems, when only the mode transition distribution is unknown, recent work studied data-driven stability verification [19] and stabilization [20] with non-asymptotic guarantees. However, it is difficult to generalize their work to more general MJSs or cases when the mode dynamics is also unknown. There is also recent work on understanding the optimization landscape of MJS-LQR problem [21], which is crucial for model-free learning algorithms. However, despite the practical importance of MJSs, non-asymptotic sample complexity results and regret analysis for MJSs are lacking. The high-level challenge here is the hybrid nature of the problem that requires consideration of both the system dynamics and the underlying Markov transition matrix. A related challenge is that, typically, the stability of MJS is understood only in the *mean-square sense*. This more relaxed notion of *mean-square stability* presents major challenges in learning, controlling, and statistical analysis, which makes statistical tools developed in recent works for linear time-invariant (LTI) systems (e.g., [6], [7], [8], [9]) insufficient due to potentially heavy tailed state distributions.

Contributions: In this paper, we provide the first statistical system identification and regret guarantees for jointly learning and controlling Markov jump linear systems using finitely many samples from a single trajectory while assuming only mean-square stability. Importantly, our guarantees are optimal in the trajectory length T. Specifically, our contributions are as follows[1]:

- **System identification:** For an MJS with s modes, the system dynamics involve a Markov transition matrix $\mathbf{T} \in \mathbb{R}^{s\times s}$ and s state-input matrix pairs $(\mathbf{A}_i, \mathbf{B}_i)_{i=1}^s$. We provide an algorithm (Alg. 1) to estimate these dynamics with the error rate $\mathcal{O}((n+p)\log(T)\sqrt{s/T})$, where n and p are the state and input dimensions respectively, and the $\mathcal{O}(1/\sqrt{T})$ dependence on the trajectory length T is optimal. This constitutes a non-

* Equal contribution.

[1] Department of Electrical Engineering and Computer Science, University of Michigan. Email: {zhedu,tarzanaq,girasole,necmiye}@umich.edu.

[2] Department of Electrical and Computer Engineering, University of California, Riverside. Email: {ysatt001,soymak}@ucr.edu.

♯ A full version of this paper with extended proofs is in [1]. A preliminary version [2] of this paper was presented as a poster at the ICML Workshop on Reinforcement Learning Theory, which is not a formal publication.

[1] orders of magnitude here are up to polylogarithmic factors

TABLE I

COMPARISON WITH PRIOR WORKS IN THE LQR SETTING.

Model	Reference	Regret	Comp. Complexity.	Cost	Sys. Requirement.
LTI	[3]	$\sqrt{T}$	Exponential	Strongly Convex	Controllable
	[22]	$\sqrt{T}$	Exponential	Convex	Controllable
	[23] (one dim. systems)	$\sqrt{T}$	Polynomial	Strongly Convex	Stabilizable
	[24]	$T^{2/3}$	Polynomial	Convex	Stabilizable
	[9]	$\sqrt{T}$	Polynomial	Strongly Convex	Controllable
	[25]	$\sqrt{T}$	Polynomial	Strongly Convex	Strongly Stabilizable
	[7], [26]	$\sqrt{T}$	Polynomial	Strongly Convex	Stabilizable
MJS	**Ours**	$s^2\sqrt{T}$	Polynomial	Strongly Convex	MSS

trivial extension of the recent sample complexity results for identification of LTI systems (see, e.g., [6], [27], [28]) to the MJS setting by introducing a subsampling and truncation procedure to control the independence and tail distribution of the states, required due to the weaker notion of stability and the interaction between the mixing time of the Markov chain and that of the underlying state dynamics.

- $\mathcal{O}(\sqrt{T})$**-regret bound:** We employ our system identification guarantees for the MJS-LQR. When the system dynamics are unknown, we show that the certainty-equivalent adaptive MJS-LQR Algorithm (Alg. 2) achieves a regret bound of $\mathcal{O}(\sqrt{T})$. Importantly, this coincides with the optimal regret bounds for the LTI LQR problem (see Table I).

II. PRELIMINARIES AND PROBLEM SETUP

Notations: We use boldface uppercase (lowercase) letters to denote matrices (vectors). For a matrix $\mathbf{V}$, $\rho(\mathbf{V}), \underline{\sigma}(\mathbf{V}), \lambda_{\min}(\mathbf{V})$ denote its spectral radius, smallest singular value and smallest eigenvalue respectively. We use $\|\cdot\|$ to denote the Euclidean norm of vectors as well as the spectral norm of matrices. Similarly, we use $\|\cdot\|_1$ to denote the ℓ_1-norm of a matrix/vector. The Kronecker product of two matrices $\mathbf{M}$ and $\mathbf{N}$ is denoted as $\mathbf{M}\otimes\mathbf{N}$. $\mathbf{V}_{1:s}$ denotes a set of s matrices $\{\mathbf{V}_i\}_{i=1}^s$ of same dimensions. We define $[s] := \{1, 2, \ldots, s\}$ and $\|\mathbf{V}_{1:s}\| := \max_{i\in[s]} \|\mathbf{V}_i\|$. The i-th row or column of a matrix $\mathbf{M}$ is denoted by $[\mathbf{M}]_{i,:}$ or $[\mathbf{M}]_{:,i}$ respectively. Orders of magnitude notation $\tilde{\mathcal{O}}(\cdot)$ hides $\log(\frac{1}{\delta})$ or $\log^2(\frac{1}{\delta})$ terms.

A. Markov Jump Linear Systems

In this paper we consider the data-driven control of MJSs which are governed by the following state equation,

$$\begin{aligned} &\mathbf{x}_{t+1} = \mathbf{A}_{\omega(t)}\mathbf{x}_t + \mathbf{B}_{\omega(t)}\mathbf{u}_t + \mathbf{w}_t \quad \text{s.t.} \\ &\omega(t) \sim \text{Markov Chain}(\mathbf{T}), \end{aligned} \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^n$, $\mathbf{u}_t \in \mathbb{R}^p$ and $\mathbf{w}_t \in \mathbb{R}^n$ are the state, input, and process noise of the MJS at time t with $\{\mathbf{w}_t\}_{t=0}^\infty \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{w}}^2\mathbf{I}_n)$. There are s modes in total, and the dynamics of mode i is given by the state matrix $\mathbf{A}_i$ and input matrix $\mathbf{B}_i$. The active mode at time t is indexed by $\omega(t) \in [s]$. Throughout, we assume the states $\mathbf{x}_t$ and the modes $\omega(t)$ can be observed at time t. The mode switching sequence $\{\omega(t)\}_{t=0}^\infty$ follows a Markov chain with transition matrix $\mathbf{T} \in \mathbb{R}_+^{s\times s}$ such that for all $t \geq 0$, the ij-th element of $\mathbf{T}$ denotes the conditional probability $[\mathbf{T}]_{ij} := \mathbb{P}\big(\omega(t+1) = j \mid \omega(t) = i\big),\ \forall\ i, j \in [s]$. Throughout, we assume the initial state $\mathbf{x}_0$, the mode switching sequence $\{\omega(t)\}_{t=0}^\infty$, and the noise $\{\mathbf{w}_t\}_{t=0}^\infty$ are mutually independent. We use MJS$(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T})$ to refer to an MJS with state equation (1), parameterized by $(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T})$.

For mode-dependent state-feedback controller $\mathbf{K}_{1:s}$ that yields the input $\mathbf{u}_t{=}\mathbf{K}_{\omega(t)}\mathbf{x}_t$, we use $\mathbf{L}_i{:=}\mathbf{A}_i + \mathbf{B}_i\mathbf{K}_i$ to denote the closed-loop state matrix for mode i. We use $\mathbf{x}_{t+1}{=}\mathbf{L}_{\omega(t)}\mathbf{x}_t$ to denote the noise-free autonomous MJS, either open-loop ($\mathbf{L}_i{=}\mathbf{A}_i$) or closed-loop ($\mathbf{L}_i{=}\mathbf{A}_i + \mathbf{B}_i\mathbf{K}_i$). Due to the randomness in $\{\omega(t)\}_{t=0}^\infty$, it is common to consider the stability of MJS in the mean-square sense which is defined as follows.

Definition 1 (Mean-square stability [11])**.** *We say MJS in* (1) *with* $\mathbf{u}_t = 0$ *is mean-square stable (MSS) if there exists* $\mathbf{x}_\infty, \mathbf{\Sigma}_\infty$ *such that for any initial state* $\mathbf{x}_0$ *and mode* $\omega(0)$*, as* $t \to \infty$*, we have*

$$\|\mathbb{E}[\mathbf{x}_t] - \mathbf{x}_\infty\| \to 0, \quad \|\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\intercal] - \mathbf{\Sigma}_\infty\| \to 0, \tag{2}$$

where the expectation is over the Markovian mode switching sequence $\{\omega(t)\}_{t=0}^\infty$*, the noise* $\{\mathbf{w}_t\}_{t=0}^\infty$ *and the initial state* $\mathbf{x}_0$*. In the noise-free case (*$\mathbf{w}_t = 0$*), we have* $\mathbf{x}_\infty = 0$*,* $\mathbf{\Sigma}_\infty = 0$*. We say MJS in* (1) *with* $\mathbf{w}_t{=}0$ *is (mean-square) stabilizable if there exists mode-dependent controller* $\mathbf{K}_{1:s}$ *such that the closed-loop MJS* $\mathbf{x}_{t+1} = (\mathbf{A}_{\omega(t)} + \mathbf{B}_{\omega(t)}\mathbf{K}_{\omega(t)})\mathbf{x}_t$ *is MSS. We call such* $\mathbf{K}_{1:s}$ *a stabilizing controller.*

The MSS of a noise-free autonomous MJS is related to the spectral radius of an augmented state matrix $\tilde{\mathbf{L}} \in \mathbb{R}^{sn^2\times sn^2}$ with ij-th $n^2\times n^2$ block given by $[\tilde{\mathbf{L}}]_{ij} := [\mathbf{T}]_{ji}\mathbf{L}_j \otimes \mathbf{L}_j$. As discussed in [11, Theorem 3.9], $\tilde{\mathbf{L}}$ can be viewed as the mapping from $\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\intercal]$ to $\mathbb{E}[\mathbf{x}_{t+1}\mathbf{x}_{t+1}^\intercal]$, thus a noise-free autonomous MJS is MSS if and only if $\rho(\tilde{\mathbf{L}}) < 1$.

The analysis of this work highly depends on certain "mixing" of the MJS – the distributions of both state $\mathbf{x}_t$ and mode $\omega(t)$ can converge close to some stationary distributions within finite time, which is guaranteed by the following assumption.

Assumption A1. *The MJS in* (1) *has an ergodic Markov chain and is mean-square stabilizable.*

Ergodicity guarantees that the distribution of $\omega(t)$ converges to a unique strictly positive stationary distribution [29, Theorem 4.3.5]. Throughout, we let $\boldsymbol{\pi}_\infty$ denote the stationary distribution of $\mathbf{T}$ and $\pi_{\min} := \min_i \boldsymbol{\pi}_\infty(i)$. We further define the mixing time [30] of $\mathbf{T}$ as $t_{\mathrm{MC}} := \inf\left\{t \in \mathbb{N} : \max_{i\in[s]} \|([\mathbf{T}^t]_{i,:})^\intercal - \boldsymbol{\pi}_\infty\|_1 \le 0.5\right\}$, to quantify its convergence rate. In our analysis, ergodicity and t_{MC} ensures that the MJS trajectory has enough "visits" to every mode i, thus providing us enough data to learn $\mathbf{A}_i$, $\mathbf{B}_i$ and $[\mathbf{T}]_{i,:}$. On the other hand, stability (or stabilizability) characterized by the spectral radius of $\tilde{\mathbf{L}}$ guarantees the convergence/mixing of $\mathbf{x}_t$, which allows us to obtain weakly dependent samples by properly sub-sampling the MJS trajectory, upon which the sample complexity of learning the matrices $\mathbf{A}_{1:s}$ and $\mathbf{B}_{1:s}$ can be established.

B. Problem Formulation

In this work we consider two major problems under the MJS setting: Data-driven system identification and adaptive control, with the former being the core part of the latter.

(A) System Identification. This problem seeks to estimate unknown system dynamics from data, i.e. from input-state trajectory, when one has the flexibility to design the inputs so that the collected data has nice statistical properties. In the MJS setting, one needs to estimate both the state/input matrices $\mathbf{A}_{1:s}, \mathbf{B}_{1:s}$ for every mode as well as the Markov transition matrix $\mathbf{T}$. In this work, we seek to estimate the MJS dynamics using a single trajectory of states, inputs and mode observations $\{\mathbf{x}_t, \mathbf{u}_t, \omega(t)\}_{t=0}^T$ and provide finite sample guarantees. As mentioned earlier, MJS presents unique statistical analysis challenges due to Markovian jumps and weaker notion of stability. Section III presents our system identification guarantees overcoming these challenges. These guarantees are further integrated into model-based control for MJS-LQR in Section IV.

(B) Online Linear Quadratic Regulator. In this paper, we consider the following finite-horizon Markov jump system linear quadratic regulator (MJS-LQR) problem:

$$\begin{aligned} \inf_{\mathbf{u}_{0:T}} \quad & J(\mathbf{u}_{0:T}) := \sum_{t=0}^{T} \mathbb{E}\left[\mathbf{x}_t^\intercal \mathbf{Q}_{\omega(t)} \mathbf{x}_t + \mathbf{u}_t^\intercal \mathbf{R}_{\omega(t)} \mathbf{u}_t\right], \\ \text{s.t.} \quad & \mathbf{x}_t, \omega(t) \sim \mathrm{MJS}(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T}). \end{aligned} \tag{3}$$

Here, the goal is to design control inputs for the MJS dynamics (1) to minimize the expected quadratic cost function composed of cost matrices $\mathbf{Q}_{1:s}$ and $\mathbf{R}_{1:s}$ that satisfy the following assumption.

Assumption A2. *For all* $i \in [s]$, $\mathbf{R}_i \succ 0, \mathbf{Q}_i \succ 0$.

Assumptions A1 and A2 together guarantee the solvability of MJS-LQR when the dynamics are known [11, Corollary A.21]. In the remaining of the paper, we use MJS-LQR$(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T}, \mathbf{Q}_{1:s}, \mathbf{R}_{1:s})$ to denote MJS-LQR problem (3) composed of MJS$(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T})$ and cost matrices $\mathbf{Q}_{1:s}, \mathbf{R}_{1:s}$.

Recall that, we assume the states $\mathbf{x}_t$ and the modes $\omega(t)$ can be observed for all $t \ge 0$. With these observations, instead of a fixed and open-loop input sequence, one can design closed-loop policies that generate real-time control inputs based on the current observations, e.g. mode-dependent state-feedback controllers. When the dynamics $\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T}$ of the MJS are known, one can solve for the optimal controllers recursively via coupled discrete-time algebraic Riccati equations [11]. In this work, we assume the dynamics are unknown, and only the design parameters $\mathbf{Q}_{1:s}$ and $\mathbf{R}_{1:s}$ are known. Control schemes in this scenario are typically referred to as adaptive control, which usually involves procedures of learning, either the dynamics or directly the controllers. Adaptive control suffers additional costs as (i) the lack of the exact knowledge of the system and (ii) the exploration-exploitation trade-off – the necessity to sacrifice short-term input optimality to boost learning, so that overall long-term optimality can be improved.

Because of this, to evaluate the performance of an adaptive scheme, one is interested in the notion of regret – how much more cost it will incur if one could have applied the optimal controllers? In our setting, we compare the resulting cost against the optimal cost $T \cdot J^\star$, where $J^\star$ is the optimal infinite-horizon average cost

$$J^\star := \limsup_{T\to\infty} \frac{1}{T} \inf_{\mathbf{u}_{0:T}} J(\mathbf{u}_{0:T}), \tag{4}$$

i.e., if one applies the optimal controller for infinitely long, how much cost one would get on average for each single time step. Compared to the regret analysis of standard adaptive LQR problem [24], in MJS-LQR setting, the cost analysis requires additional consideration of Markov chain mixing, which is addressed in this paper.

Algorithm 1: MJS-SYSID

Input: Stabilizing controller $\mathbf{K}_{1:s}$; process and exploration noise variances $\sigma_{\mathbf{w}}^2$ and $\sigma_{\mathbf{z}}^2$; trajectory $\{\mathbf{x}_t, \mathbf{z}_t, \omega(t)\}_{t=0}^T$; control input $\mathbf{u}_t = \mathbf{K}_{\omega(t)}\mathbf{x}_t + \mathbf{z}_t$ with $\mathbf{z}_t \sim \mathcal{N}(0, \sigma_{\mathbf{z}}^2 \mathbf{I}_p)$; data clipping thresholds $c_{\mathbf{x}}$, $c_{\mathbf{z}}$; sub-sampling factor C_{sub}.

Set sub-sampling period and indices

$$L = C_{\mathrm{sub}} \log(T),\ \tau_k = kL,\ \forall k = 1, 2, \ldots, \lfloor T/L \rfloor$$

Estimate $\mathbf{A}_{1:s}, \mathbf{B}_{1:s}$: **for all** modes $i \in [s]$ **do**:

$$S_i = \{\tau_k \mid \omega(\tau_k) = i, \ \|\mathbf{x}_{\tau_k}\| \le c_{\mathbf{x}} \sigma_{\mathbf{w}} \sqrt{\log(T)},\ \|\mathbf{z}_{\tau_k}\| \le c_{\mathbf{z}} \sigma_{\mathbf{z}}\}$$

$$\hat{\boldsymbol{\Theta}}_{1,i}, \hat{\boldsymbol{\Theta}}_{2,i} = \arg\min_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \textstyle\sum_{k\in S_i} \|\mathbf{x}_{k+1} - \boldsymbol{\Theta}_1 \mathbf{x}_k/\sigma_{\mathbf{w}} - \boldsymbol{\Theta}_2 \mathbf{z}_k/\sigma_{\mathbf{z}}\|^2$$

$$\hat{\mathbf{B}}_i = \hat{\boldsymbol{\Theta}}_{2,i}/\sigma_{\mathbf{z}},\ \hat{\mathbf{A}}_i = \hat{\boldsymbol{\Theta}}_{1,i}/\sigma_{\mathbf{w}} - \hat{\mathbf{B}}_i \mathbf{K}_i$$

Estimate $\mathbf{T}$:

$$[\hat{\mathbf{T}}]_{ji} = \frac{\sum_{k=1}^{\lfloor T/L \rfloor} \mathbf{1}_{\{\omega(\tau_k)=i, \omega(\tau_k-1)=j\}}}{\sum_{k=1}^{\lfloor T/L \rfloor} \mathbf{1}_{\{\omega(\tau_k-1)=j\}}}$$

Output: $\hat{\mathbf{A}}_{1:s}, \hat{\mathbf{B}}_{1:s}, \hat{\mathbf{T}}$

III. System Identification for MJS

Our MJS identification procedure is given in Algorithm 1. We assume one has access to a stabilizing controller $\mathbf{K}_{1:s}$,

which is a standard assumption in data-driven control [22], [23], [24], [25], [26] for LTI systems. Note that, if the open-loop MJS is already MSS, then one can simply set $\mathbf{K}_{1:s} = 0$ and carry out MJS identification. Given an MJS trajectory $\{\mathbf{x}_t, \mathbf{z}_t, \omega(t)\}_{t=0}^{T}$ obtained using the input $\mathbf{u}_t = \mathbf{K}_{\omega(t)}\mathbf{x}_t + \mathbf{z}_t$, where $\mathbf{z}_t \sim \mathcal{N}(0, \sigma_{\mathbf{z}}^2 \mathbf{I}_p)$ is the excitation for exploration, we first subsample (index τ_k) it with a sampling period $L = \mathcal{O}(\log(T))$. This will make sure the samples are only weakly dependent, conditioned on the mode observation. We then further subsample the acquired sub-trajectory for bounded states and inputs. This is required because of MSS, which can at most guarantee that the states are bounded in expectation. Two rounds of sampling provides samples with manageable distributional properties and weak temporal dependence. After appropriate scaling, we regress over these samples to obtain the estimates $\hat{\mathbf{A}}_{1:s}, \hat{\mathbf{B}}_{1:s}$. Lastly, using the empirical frequency of observed modes, we estimate $\hat{\mathbf{T}}$.

The following theorem gives our main results on learning the dynamics of an unknown MJS from finite samples obtained from a single trajectory. The complete proof is provided in [1].

Theorem 1 (Identification of MJS)**.** *Suppose we run Algorithm 1 with $c_{\mathbf{x}} = \mathcal{O}(\sqrt{n})$ and $c_{\mathbf{z}} = \mathcal{O}(\sqrt{p})$. Let $\rho = \rho(\tilde{\mathbf{L}})$, where $\tilde{\mathbf{L}}$ is the augmented state matrix of the closed-loop MJS defined in Sec. II-A. Suppose the sub-sampling factor $C_{sub} \geq \mathcal{O}(t_{\mathrm{MC}}/(1-\rho))$ and the trajectory length $T \geq \tilde{\mathcal{O}}(C_{sub}\sqrt{s}(n+p)/\pi_{\min})$. Then, under Assumption A1, with probability at least $1-\delta$, for all $i \in [s]$, we have*

$$\begin{aligned} &\|\hat{\mathbf{A}}_i - \mathbf{A}_i\|, \\ &\|\hat{\mathbf{B}}_i - \mathbf{B}_i\| \end{aligned} \leq \tilde{\mathcal{O}}\Big(\frac{(\sigma_{\mathbf{z}} + \sigma_{\mathbf{w}})}{\sigma_{\mathbf{z}}} \frac{\sqrt{s}(n+p)\log(T)}{\pi_{\min}\sqrt{T}}\Big),$$
$$\|\hat{\mathbf{T}} - \mathbf{T}\|_{\infty} \leq \tilde{\mathcal{O}}\left(\frac{1}{\pi_{\min}}\sqrt{\frac{\log(T)}{T}}\right). \tag{5}$$

The $\log(T)/\sqrt{T}$ dependency on trajectory length T in our system identification result achieves near-optimal statistical error rate compared with, for example, the $1/\sqrt{N}$ rate for LTI identifications using N i.i.d. trajectories [6]. The linear dependency on n of the error bound can potentially be improved to $\sqrt{n}$. Note that, $\pi_{\min}$ dictates the trajectory fraction of the least-frequent mode, thus, in the result $\pi_{\min}^{-1}$ multiplier is unavoidable.

Proof outline for Theorem 1. We omit the discussion on $\hat{\mathbf{T}}$ as its analysis is mainly based on [31]. For $\hat{\mathbf{A}}_i$ and $\hat{\mathbf{B}}_i$, in Algorithm 1, we know $\hat{\mathbf{\Theta}}_{1,i} = \sigma_{\mathbf{w}}(\hat{\mathbf{A}}_i + \hat{\mathbf{B}}_i\mathbf{K}_i)$ and $\hat{\mathbf{\Theta}}_{2,i} = \sigma_{\mathbf{z}}\hat{\mathbf{B}}_i$. Let $\hat{\mathbf{\Theta}}_i := [\hat{\mathbf{\Theta}}_{1,i} \;\; \hat{\mathbf{\Theta}}_{2,i}]$ and $\mathbf{\Theta}_i := [\sigma_{\mathbf{w}}(\mathbf{A}_i + \mathbf{B}_i\mathbf{K}_i) \;\; \sigma_{\mathbf{z}}\mathbf{B}_i]$. Thus, to bound $\|\hat{\mathbf{A}}_i - \mathbf{A}_i\|$ and $\|\hat{\mathbf{B}}_i - \mathbf{B}_i\|$, it suffices to bound $\|\hat{\mathbf{\Theta}}_i - \mathbf{\Theta}_i\|$, which is outlined below.

For mode $i \in [s]$ and all $k \in S_i$. Let $\mathbf{h}_k := [\frac{1}{\sigma_{\mathbf{w}}}\mathbf{x}_k^{\intercal} \;\; \frac{1}{\sigma_{\mathbf{z}}}\mathbf{z}_k^{\intercal}]^{\intercal}$ denote the the regressor used for computing $\hat{\mathbf{A}}_i$ and $\hat{\mathbf{B}}_i$. Then, Alg. 1 solves the following least-squares problem: $\hat{\mathbf{\Theta}}_i^{\intercal} = \arg\min_{\mathbf{\Theta}_i} \frac{1}{2|S_i|}\|\mathbf{Y}_i - \mathbf{H}_i\mathbf{\Theta}_i^{\intercal}\|_F^2$, where $\mathbf{Y}_i$ has $\{\mathbf{x}_{k+1}^{\intercal}\}_{k\in S_i}$ in its rows and $\mathbf{H}_i$ has $\{\mathbf{h}_k^{\intercal}\}_{k\in S_i}$ in its rows. Similarly, let $\mathbf{W}_i$ has $\{\mathbf{w}_k^{\intercal}\}_{k\in S_i}$ in its rows. Then the estimation error of the least-squares estimator $\hat{\mathbf{\Theta}}_i^{\intercal} = (\mathbf{H}_i^{\intercal}\mathbf{H}_i)^{-1}\mathbf{H}_i^{\intercal}\mathbf{Y}_i$ is given by: $\|\hat{\mathbf{\Theta}}_i - \mathbf{\Theta}_i\| = \|\mathbf{H}_i^{\intercal}\mathbf{W}_i\|/\lambda_{\min}(\mathbf{H}_i^{\intercal}\mathbf{H}_i)$. Since $\mathbf{H}_i$ has non-i.i.d. rows, it is therefore not straightforward to upper/lower bound the terms $\|\mathbf{H}_i^{\intercal}\mathbf{W}_i\|$ and $\lambda_{\min}(\mathbf{H}_i^{\intercal}\mathbf{H}_i)$ respectively. To resolve this issue, we rely on MSS and use perturbation-based techniques to indirectly bound these terms. To proceed, for each state $\mathbf{x}_t$ ($t \geq L$), we define its fictional proxy $\bar{\mathbf{x}}_t$, by resetting $\mathbf{x}_{t-L} = 0$ but preserving the mode switching sequence $\omega(\tau)$, the excitation $\mathbf{z}_\tau$ and the noise $\mathbf{w}_\tau$ from $t-L$ to $t-1$. We refer to this as unrolling $\mathbf{x}_t$ until time $t-L$, and we call the obtained $\bar{\mathbf{x}}_t$ as the L-truncated state at time t. Note that one can view $\bar{\mathbf{x}}_t$ as the zero-state response starting from time $t-L$, and $\mathbf{x}_t - \bar{\mathbf{x}}_t$ as the zero-input response. For the latter, we can show that $\|\mathbf{x}_t - \bar{\mathbf{x}}_t\| \leq \mathcal{O}(\sigma_{\mathbf{w}}\sqrt{ns\log(T)})$. Our analysis involves truncating the bounded states $\{\mathbf{x}_k\}_{k\in S_i}$ to obtain the truncated states $\{\bar{\mathbf{x}}_k\}_{k\in S_i}$. Let $\bar{\mathbf{h}}_k := [\frac{1}{\sigma_{\mathbf{w}}}\bar{\mathbf{x}}_k^{\intercal} \;\; \frac{1}{\sigma_{\mathbf{z}}}\mathbf{z}_k^{\intercal}]^{\intercal}$ and $\bar{\mathbf{H}}_i$ has $\{\bar{\mathbf{h}}_k\}_{k\in S_i}$ in its rows. Then, we have $\|\mathbf{H}_i^{\intercal}\mathbf{W}_i\| \leq \|\bar{\mathbf{H}}_i^{\intercal}\mathbf{W}_i\| + \|\mathbf{H}_i^{\intercal}\mathbf{W}_i - \bar{\mathbf{H}}_i^{\intercal}\mathbf{W}_i\|$ and $\lambda_{\min}(\mathbf{H}_i^{\intercal}\mathbf{H}_i) \geq \lambda_{\min}(\bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i) - \|\mathbf{H}_i^{\intercal}\mathbf{H}_i - \bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i\|$. In order to bound $\|\mathbf{H}_i^{\intercal}\mathbf{W}_i\|$ and $\lambda_{\min}(\mathbf{H}_i^{\intercal}\mathbf{H}_i)$, it thus suffices to upper bound $\|\bar{\mathbf{H}}_i^{\intercal}\mathbf{W}_i\|$, $\|\mathbf{H}_i^{\intercal}\mathbf{W}_i - \bar{\mathbf{H}}_i^{\intercal}\mathbf{W}_i\|$ and $\|\mathbf{H}_i^{\intercal}\mathbf{H}_i - \bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i\|$ and lower bound $\lambda_{\min}(\bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i)$.

Upper bounding $\|\bar{\mathbf{H}}_i^T\mathbf{W}_i\|$: By definition of S_i, for all $k \in S_i$, we have $\|\mathbf{x}_k\| \leq \mathcal{O}(\sigma_{\mathbf{w}}\sqrt{ns\log(T)})$, which further implies $\|\bar{\mathbf{x}}_k\| \leq \|\mathbf{x}_k\| + \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \leq \mathcal{O}(\sigma_{\mathbf{w}}\sqrt{ns\log(T)})$. Combining this with $\|\mathbf{z}_k\| \leq \mathcal{O}(\sigma_{\mathbf{z}}\sqrt{p})$, we have that $\|\bar{\mathbf{h}}_k\| \leq \mathcal{O}(\sqrt{s(n+p)\log(T)})$. This implies $\|\bar{\mathbf{H}}_i\| \leq \|\bar{\mathbf{H}}_i\|_F \leq \mathcal{O}(\sqrt{|S_i|s(n+p)\log(T)}) \leq \mathcal{O}(\sqrt{Ts(n+p)})$. To proceed, let $\bar{\mathbf{H}}_i$ has singular value decomposition $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ with $\|\mathbf{\Sigma}\| \leq \mathcal{O}(\sqrt{Ts(n+p)})$. Since $\mathbf{W}_i$ has i.i.d. Gaussian entries, $\mathbf{U}^T\mathbf{W}_i \in \mathbb{R}^{(n+p)\times n}$ has i.i.d. subGaussian columns. As a result, applying [32, Theorem 5.39], we have $\|\mathbf{U}^T\mathbf{W}_i\| \leq \tilde{O}(\sigma_{\mathbf{w}}\sqrt{n+p})$. Therefore, $\|\bar{\mathbf{H}}_i^T\mathbf{W}_i\| \leq \|\mathbf{\Sigma}\|\|\mathbf{U}^T\mathbf{W}_i\| \leq \tilde{O}(\sigma_{\mathbf{w}}(n+p)\sqrt{sT})$.

Lower bounding $\lambda_{\min}(\bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i)$: Recall $\bar{\mathbf{H}}_i$ is composed of all regressor $\bar{\mathbf{h}}_k$ indexed by $k \in S_i$. To proceed, we first construct the set $S_i' := \{\tau_k \;\big|\; \omega(\tau_k) = i,\ \|\mathbf{z}_{\tau_k}\| \leq c_{\mathbf{z}}\sigma_{\mathbf{z}},\ \|\mathbf{x}_{\tau_k}\| \leq c_{\mathbf{x}}\sigma_{\mathbf{w}}\sqrt{\log(T)},\ \|\bar{\mathbf{x}}_{\tau_k}\| \leq c_{\mathbf{x}}\sigma_{\mathbf{w}}\sqrt{\log(T)}/2\}$. Then, picking $L = C_{sub}\cdot\log(T) \geq \tilde{O}(t_{\mathrm{MC}}\log(T)/(1-\rho))$, we can show (a) with high probability $S_i' \subseteq S_i$ and (b) conditioning on the modes, $\{\bar{\mathbf{x}}_k\}_{k\in S_i'}$, $\{\mathbf{z}_k\}_{k\in S_i'}$, and $\{\mathbf{w}_k\}_{k\in S_i'}$ are all independent of each other. This independence allows us to establish the following covariance bound: for $\bar{\mathbf{h}}_k$: $\frac{1}{4}\mathbf{I}_{n+p} \preceq \mathbf{\Sigma}[\bar{\mathbf{h}}_k \mid k \in S_i'] \preceq \mathcal{O}(s(n+p)\log(T))\mathbf{I}_{n+p}$. Let $\bar{\mathbf{H}}_i'$ has $\{\bar{\mathbf{h}}_k\}_{k\in S_i'}$ in its rows. Then, using [32, Theorem 5.41] (by specializing it to non-isotropic rows), we have $\underline{\sigma}(\bar{\mathbf{H}}_i') \geq \sqrt{|S_i'|}/2 - \tilde{O}(\sqrt{s(n+p)\log(T)})$. Then, the fact $S_i' \subseteq S_i$ gives $\lambda_{\min}(\bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i) \geq \lambda_{\min}(\bar{\mathbf{H}}_i'^{\intercal}\bar{\mathbf{H}}_i') \geq \big(\sqrt{|S_i'|}/2 - \tilde{O}(\sqrt{s(n+p)\log(T)})\big)^2$.

Upper bounding $\|\mathbf{H}_i^{\intercal}\mathbf{W}_i - \bar{\mathbf{H}}_i^{\intercal}\mathbf{W}_i\|$: Using simple algebra and MSS, we have $\|\mathbf{H}_i^{\intercal}\mathbf{W}_i - \bar{\mathbf{H}}_i^{\intercal}\mathbf{W}_i\| = \|\sum_{k\in S_i}(\mathbf{h}_k\mathbf{w}_k^{\intercal} - \bar{\mathbf{h}}_k\mathbf{w}_k^{\intercal})\| \leq |S_i|\max_{k\in S_i}\|\mathbf{h}_k - \bar{\mathbf{h}}_k\|\|\mathbf{w}_k\| \leq \tilde{O}(\sigma_{\mathbf{w}}(n+p)\sqrt{sT})$.

Upper bounding $\|\mathbf{H}_i^{\intercal}\mathbf{H}_i - \bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i\|$: Similarly, we have $\|\mathbf{H}_i^{\intercal}\mathbf{H}_i - \bar{\mathbf{H}}_i^{\intercal}\bar{\mathbf{H}}_i\| = \|\sum_{k\in S_i}(\mathbf{h}_k\mathbf{h}_k^{\intercal} - \bar{\mathbf{h}}_k\bar{\mathbf{h}}_k^{\intercal})\| \leq |S_i|\max_{k\in S_i}\|\mathbf{h}_k - \bar{\mathbf{h}}_k\|(\|\mathbf{h}_k\| + \|\bar{\mathbf{h}}_k\|) \leq \mathcal{O}(|S_i'|)$.

Lower bounding $|S_i'|$: To complete the proof, we need to show that the set S_i' has enough samples so that $\underline{\sigma}(\bar{\mathbf{H}}_i') \geq \sqrt{|S_i'|}/2 - \tilde{O}(\sqrt{s(n+p)\log(T)}) \geq \sqrt{|S_i'|}/4$. For this purpose, we use martingale-based techniques to show the following result: when $T \geq T_0 := \tilde{\mathcal{O}}(C_{sub}\sqrt{s}(n+p)/\pi_{\min})$, we have $|S_i'| \geq \mathcal{O}(\pi_{\min}T/(C_{sub}\log(T)))$.

Finalizing the proof: Finally, putting all the above results together, we have $\|\mathbf{H}_i^T\mathbf{W}_i\| \leq \tilde{O}(\sigma_{\mathbf{w}}(n+p)\sqrt{sT})$ and $\lambda_{\min}(\mathbf{H}_i^\intercal\mathbf{H}_i) \geq \mathcal{O}(\pi_{\min}T/(C_{sub}\log(T)))$. Combining these results, we upper bound the estimation error as follows: $\|\hat{\mathbf{\Theta}}_i - \mathbf{\Theta}_i\| \leq \tilde{O}\big((\sigma_{\mathbf{w}}C_{sub}(n+p)\log(T)/\pi_{\min})\sqrt{s/T}\big)$. Specializing this result to the state/input matrices gives us the statement of the theorem. □

IV. Adaptive Control for MJS-LQR

Our adaptive MJS-LQR control scheme is given in Algorithm 2. It is performed on an epoch-by-epoch basis; a fixed controller is used for each epoch, and from epoch to epoch, the controller is updated using the trajectory generated in the most recent epoch. Note that a new epoch is just a continuation of previous epochs instead of restarting the MJS. Similar to the discussion in Section III, we assume, at the beginning of epoch 0, that one has access to a stabilizing controller $\mathbf{K}_{1:s}^{(0)}$. During epoch i, the controller $\mathbf{K}_{1:s}^{(i)}$ is used together with additive exploration noise $\mathbf{z}_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{z},i}^2\mathbf{I}_p)$ to boost learning. At the end of epoch i, the trajectory during that epoch is used to obtain a new MJS dynamics estimate $\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)}$ using Algorithm 1. Then, we set the controller $\mathbf{K}_{1:s}^{(i+1)}$ for epoch $i+1$ to be the optimal controller for the infinite-horizon MJS-LQR$(\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)}, \mathbf{Q}_{1:s}, \mathbf{R}_{1:s})$, which can be computed as follows: For a generic infinite-horizon MJS-LQR$(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T}, \mathbf{Q}_{1:s}, \mathbf{R}_{1:s})$, its optimal controller is given by $\mathbf{K}_{1:s}$ such that for all $j \in [s]$,

$$\mathbf{K}_j := -\big(\mathbf{R}_j + \mathbf{B}_j^\intercal\varphi_j(\mathbf{P}_{1:s})\mathbf{B}_j\big)^{-1}\mathbf{B}_j^\intercal\varphi_j(\mathbf{P}_{1:s})\mathbf{A}_j, \quad (6)$$

where $\varphi_j(\mathbf{P}_{1:s}) := \sum_{k=1}^s[\mathbf{T}]_{jk}\mathbf{P}_k$ and $\mathbf{P}_{1:s}$ is the solution to the following coupled discrete-time algebraic Riccati equations (cDARE):

$$\begin{aligned}\mathbf{P}_j = &\mathbf{A}_j^\intercal\varphi_j(\mathbf{P}_{1:s})\mathbf{A}_j + \mathbf{Q}_j - \mathbf{A}_j^\intercal\varphi_j(\mathbf{P}_{1:s})\mathbf{B}_j \\ &\cdot\big(\mathbf{R}_j + \mathbf{B}_j^\intercal\varphi_j(\mathbf{P}_{1:s})\mathbf{B}_j\big)^{-1}\mathbf{B}_j^\intercal\varphi_j(\mathbf{P}_{1:s})\mathbf{A}_j \quad (7)\end{aligned}$$

for all $j \in [s]$. In practice, cDARE can be solved efficiently via value iteration or LMIs [11]. Note that cDARE may not be solvable for arbitrary parameters, but our theory guarantees that when epoch lengths are appropriately chosen, cDARE parameterized by $\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)}, \mathbf{Q}_{1:s}, \mathbf{R}_{1:s}$ is solvable for every epoch i. This control design based on the estimated dynamics is also referred to as certainty equivalent control.

To achieve theoretical performance guarantees, i.e., sublinear regret, the key is to have a subtle scheduling of epoch lengths T_i and exploration noise variance $\sigma_{\mathbf{z},i}^2$. We choose T_i to increase exponentially with rate $\gamma > 1$, and set $\sigma_{\mathbf{z},i}^2 = \sigma_{\mathbf{w}}^2/\sqrt{T_i}$, which collectively guarantee $\tilde{\mathcal{O}}(\sqrt{T})$ regret when combined with the system identification result from Theorem 1. Intuitively, this scheduling can be interpreted as follows: (i) the increase of epoch lengths guarantees we have more accurate MJS estimates thus more optimal controllers; (ii) as the controller becomes more optimal we can gradually decrease the exploration noise and deploy (exploit) the controller for a longer time. Note that the scheduling rate γ has a similar role to the discount factor in reinforcement learning: smaller γ aims to reduce short-term cost while larger γ aims to reduce long-term cost.

Algorithm 2: Adaptive MJS-LQR

Input: Initial epoch length T_0; initial stabilizing controller $\mathbf{K}_{1:s}^{(0)}$; epoch incremental ratio $\gamma > 1$; data clipping thresholds $c_{\mathbf{x}}$, $c_{\mathbf{z}}$; sub-sampling factor C_{sub}.

for *epoch* $i = 0, 1, 2, \ldots$ **do**

Set epoch length $T_i = \lfloor T_0\gamma^i \rfloor$.

Set exploration noise variance $\sigma_{\mathbf{z},i}^2 = \frac{\sigma_{\mathbf{w}}^2}{\sqrt{T_i}}$.

Evolve MJS T_i steps with $\mathbf{u}_t^{(i)} = \mathbf{K}_{\omega(t)^{(i)}}^{(i)}\mathbf{x}_t^{(i)} + \mathbf{z}_t^{(i)}$ and $\mathbf{z}_t^{(i)} \sim \mathcal{N}(0, \sigma_{\mathbf{z},i}^2\mathbf{I}_p)$.

Record trajectory $\xi^{(i)} = \{\mathbf{x}_t^{(i)}, \mathbf{z}_t^{(i)}, \omega^{(i)}(t)\}_{t=0}^{T_i}$.

$\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)} \leftarrow$ MJS-SYSID$(\mathbf{K}_{1:s}^{(i)}, \sigma_{\mathbf{w}}^2, \sigma_{\mathbf{z},i}^2, \xi^{(i)}, c_{\mathbf{x}}, c_{\mathbf{z}}, C_{sub})$.

$\mathbf{K}_{1:s}^{(i+1)} \leftarrow$ optimal controller of infinite-horizon MJS-LQR$(\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)}, \mathbf{Q}_{1:s}, \mathbf{R}_{1:s})$.

end

A. Regret Analysis

We define filtration $\mathcal{F}_{-1}, \mathcal{F}_0, \mathcal{F}_1, \ldots$ such that $\mathcal{F}_{-1} := \sigma(\mathbf{x}_0, \omega(0))$ is the sigma-algebra generated by the initial state and initial mode, and $\mathcal{F}_i := \sigma(\mathbf{x}_0, \omega(0), \{\{\omega^{(j)}(t)\}_{t=1}^{T_j}\}_{j=0}^i, \mathbf{w}_0, \{\{\mathbf{w}_t^{(j)}\}_{t=1}^{T_j}\}_{j=0}^i, \mathbf{z}_0, \{\{\mathbf{z}_t^{(j)}\}_{t=1}^{T_j}\}_{j=0}^i)$ is the sigma-algebra generated by the randomness up to epoch i. Note that the initial state $\mathbf{x}_0^{(i)}$ of epoch i is also the final state $\mathbf{x}_{T_{i-1}}^{(i-1)}$ of epoch $i-1$, therefore, $\mathbf{x}_0^{(i)}$ is $\mathcal{F}_{i-1}$-measurable, and so is $\omega^{(i)}(0)$. Suppose time t belongs to epoch i, then we define the conditional expected cost at time t as:

$$c_t = \mathbb{E}[\mathbf{x}_t^\intercal\mathbf{Q}_{\omega(t)}\mathbf{x}_t + \mathbf{u}_t^\intercal\mathbf{R}_{\omega(t)}\mathbf{u}_t \mid \mathcal{F}_{i-1}], \quad (8)$$

and cumulative cost as $J_T = \sum_{t=1}^T c_t$. We define the total regret and epoch-i regret as

$$\begin{aligned}\text{Regret}(T) &= J_T - TJ^\star, \\ \text{Regret}_i &= (\sum_{t=1}^{T_i} c_{T_0+\cdots+T_{i-1}+t}) - T_iJ^\star. \end{aligned} \quad (9)$$

Then we can see $\text{Regret}(T) = \mathcal{O}(\sum_{i=1}^{\mathcal{O}(\log_\gamma(T/T_0))}\text{Regret}_i)$ where the regret of epoch 0 is ignored as it does not scale with time T. Let $\mathbf{K}_{1:s}^\star$ denote the optimal controller for the infinite-horizon MJS-LQR$(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T}, \mathbf{Q}_{1:s}, \mathbf{R}_{1:s})$ problem. $\tilde{\mathbf{L}}^{(0)}$ and $\tilde{\mathbf{L}}^\star$ denote the closed-loop augmented state matrices under the initial controller $\mathbf{K}_{1:s}^{(0)}$ and the optimal controller $\mathbf{K}_{1:s}^\star$ respectively, and we let $\bar{\rho} :=$

$\max\{\rho(\tilde{\mathbf{L}}^{(0)}), \rho(\tilde{\mathbf{L}}^{\star})\}$. With these definitions, we have the following sub-linear regret guarantee.

Theorem 2 (Sub-linear regret)**.** *Assume that the initial state* $\mathbf{x}_0 = 0$*, and Assumptions A1 and A2 hold. In Algorithm 2, suppose hyper-parameters* $c_{\mathbf{x}} = \mathcal{O}(\sqrt{n})$*,* $c_{\mathbf{z}} = \mathcal{O}(\sqrt{p})$*,* $C_{sub} \geq \mathcal{O}(t_{\mathrm{MC}}/(1-\bar{\rho}))$*, and* $T_0 \geq \tilde{\mathcal{O}}(C_{sub}\sqrt{s}(n+p)/\pi_{\min})$*. Then, with probability at least* $1-\delta$*, Algorithm 2 achieves*

$$\mathrm{Regret}(T) \leq \tilde{\mathcal{O}}\Big(\frac{s^2p(n^2+p^2)\sigma_{\mathbf{w}}^2}{\pi_{\min}^2}\log^2(T)\sqrt{T}\Big) + \mathcal{O}\Big(\frac{\sqrt{ns}\log^3(T)}{\delta}\Big). \quad (10)$$

Proof outline for Theorem 2. For simplicity, we only show the dominant $\tilde{\mathcal{O}}(\cdot)$ term here and leave the complete proof in [1]. Define the estimation error after epoch i as $\epsilon_{\mathbf{A},\mathbf{B}}^{(i)} := \max_{j\in[s]} \max\{\|\mathbf{A}_j^{(i)} - \mathbf{A}_j\|, \|\mathbf{B}_j^{(i)} - \mathbf{B}_j\|\}$, $\epsilon_{\mathbf{T}}^{(i)} := \|\mathbf{T}^{(i)} - \mathbf{T}\|_{\infty}$. Analyzing the finite-horizon cost and combining the infinite-horizon perturbation results in [33], we can bound epoch-i regret as $\mathrm{Regret}_i \leq \mathcal{O}\big(T_i\sigma_{\mathbf{z},i}^2 + T_i\sigma_{\mathbf{w}}^2(\epsilon_{\mathbf{A},\mathbf{B}}^{(i-1)} + \epsilon_{\mathbf{T}}^{(i-1)})^2\big)$. Plugging in $\sigma_{\mathbf{z},i}^2 = \frac{\sigma_{\mathbf{w}}^2}{\sqrt{T_i}}$ and the upper bounds on the estimation errors $\epsilon_{\mathbf{A},\mathbf{B}}^{(i)} \leq \tilde{\mathcal{O}}\big(\frac{\sigma_{\mathbf{z},i}+\sigma_{\mathbf{w}}}{\sigma_{\mathbf{z},i}\pi_{\min}}\frac{\sqrt{s}(n+p)\log(T_i)}{\sqrt{T_i}}\big)$, $\epsilon_{\mathbf{T}}^{(i)} \leq \tilde{\mathcal{O}}\big(\sqrt{\frac{\log(T_i)}{T_i}}\big)$ from Theorem 1, we have $\mathrm{Regret}_i \leq \tilde{\mathcal{O}}\big(\frac{s^2p(n^2+p^2)\sigma_{\mathbf{w}}^2}{\pi_{\min}^2}\gamma\sqrt{T_i}\log^2(T_i)\big)$. Finally, since $T_i = \mathcal{O}(T_0\gamma^i)$, we have $\mathrm{Regret}(T) = \sum_{i=1}^{\mathcal{O}(\log_\gamma(\frac{T}{T_0}))} \mathrm{Regret}_i \leq \tilde{\mathcal{O}}\big(\frac{s^2p(n^2+p^2)\sigma_{\mathbf{w}}^2}{\pi_{\min}^2}\sqrt{T}\log(\frac{T}{T_0})(\frac{\sqrt{\gamma}}{\sqrt{\gamma}-1})^3\big(\gamma\log(\frac{T}{T_0}) - \sqrt{\gamma}\log(\gamma)\big)\big) = \tilde{\mathcal{O}}\big(\frac{s^2p(n^2+p^2)\sigma_{\mathbf{w}}^2}{\pi_{\min}^2}\mathrm{polylog}(T)\sqrt{T}\big)$. □

One can see the interplay between T and γ from the term $\Big(\frac{\sqrt{\gamma}}{\sqrt{\gamma}-1}\Big)^3\Big(\gamma\log(\frac{T}{T_0}) - \sqrt{\gamma}\log(\gamma)\Big)$ in the proof sketch. Specifically, when horizon T is smaller, a smaller γ minimizes the upper bound, and vice versa. This further provides a mathematical justification for γ being similar to the discount factor in reinforcement learning in early discussions.

One may note that the regret bound in Theorem 2 has $\frac{1}{\delta}$ dependency on the failure probability δ, in addition to the $\log^2(\frac{1}{\delta})$ dependency hidden in $\tilde{O}(\cdot)$ term. This results from the attempt to upper bound the state $\mathbf{x}_t$ in the analysis. Since MSS only describes the stability of $\|\mathbf{x}_t\|^2$ in the expectation sense, one can thus use the Markov inequality only to bound $\|\mathbf{x}_t\|^2$, which produces $\frac{1}{\delta}$. It is straightforward to construct an example showing that under the MSS assumption only, no dependency better than $\frac{1}{\delta}$ can be established. For similar reasons, we choose the expected regret to analyze rather than the random regret used in [24], [34] for standard LQR problems. However, we believe both the random regret result and dependency tighter than $\frac{1}{\delta}$ can be established under slightly stronger notions of stability, e.g. uniform stability. We leave these potential improvements as future work, and for a more detailed discussion, please refer to the extended version [1].

V. Numerical Experiments

We provide numerical experiments to investigate the efficiency of the proposed algorithms and to verify the underlying theory. Throughout, we show results from a synthetic experiment where entries of the true system matrices $(\mathbf{A}_{1:s}, \mathbf{B}_{1:s})$ are generated randomly from a standard normal distribution. We further scale each $\mathbf{A}_i$ to have $\|\mathbf{A}_i\| \leq 0.5$. Since this guarantees the MJS itself is MSS, as we discussed in Sec III, we set controller $\mathbf{K}_{1:s} = 0$ in system identification Algorithm 1 and initial stabilizing controller $\mathbf{K}_{1:s}^{(0)} = 0$ in adaptive MJS-LQR Algorithm 2. For the cost matrices $(\mathbf{Q}_{1:s}, \mathbf{R}_{1:s})$, we set $\mathbf{Q}_j = \underline{\mathbf{Q}}_j\underline{\mathbf{Q}}_j^{\intercal}$, and $\mathbf{R}_j = \underline{\mathbf{R}}_j\underline{\mathbf{R}}_j^{\intercal}$ where $\underline{\mathbf{Q}}_j \in \mathbb{R}^{n\times n}$ and $\underline{\mathbf{R}}_j \in \mathbb{R}^{p\times p}$ are generated from a standard normal distribution. The Markov transition matrix $\mathbf{T} \in \mathbb{R}_+^{s\times s}$ is sampled from a Dirichlet distribution $\mathtt{Dir}((s-1)\cdot\mathbf{I}_s + 1)$, where $\mathbf{I}_s$ denotes the identity matrix. We assume that we have equal probability of starting in any initial mode and the initial state $\mathbf{x}_0 = 0$.

Since for system identification, our main contribution is estimating $\mathbf{A}_{1:s}$ and $\mathbf{B}_{1:s}$ of the MJS, we omit the plots for estimating $\mathbf{T}$. Let $\hat{\Psi}_j = [\hat{\mathbf{A}}_j, \hat{\mathbf{B}}_j]$ and $\Psi_j = [\mathbf{A}_j, \mathbf{B}_j]$. We use $\|\hat{\Psi} - \Psi\|/\|\Psi\| := \max_{j\in[s]} \|\hat{\Psi}_j - \Psi_j\|/\|\Psi_j\|$ to investigate the convergence behavior of MJS-SYSID Algorithm 1. The clipping constants in this algorithm, i.e., C_{sub}, $c_{\mathbf{x}}$, and $c_{\mathbf{z}}$ are chosen based on their lower bounds provided in Theorem 2. The depicted results are averaged over 10 independent Monte Carlo runs.

A. Performance of MJS-SYSID

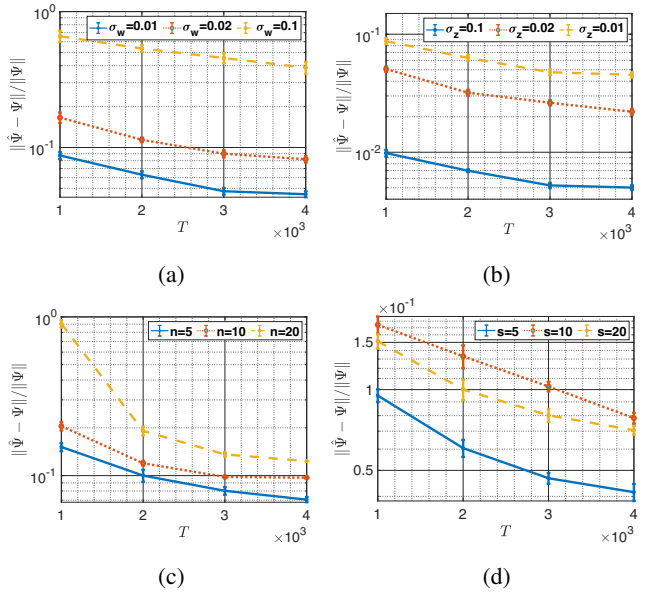
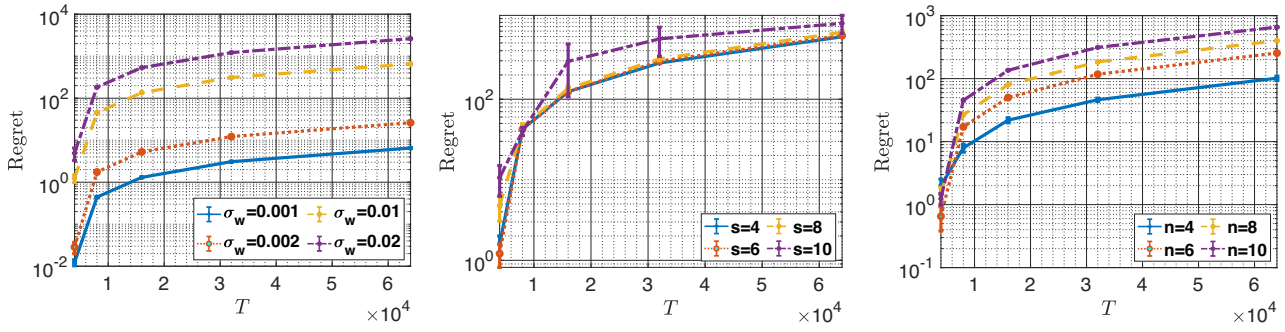


Fig. 1. Performance profiles of MJS-SYSID with varying: (a) process noise $\sigma_{\mathbf{w}}$, (b) exploration noise $\sigma_{\mathbf{z}}$, (c) state dimension n, (d) number of modes s.

In this section, we investigate the performance of our MJS-SYSID method, i.e., Algorithm 1. We first empirically evaluate the effect of the noise/excitation variances $\sigma_{\mathbf{w}}$ and $\sigma_{\mathbf{z}}$. In particular, we study how the estimation errors vary with (i)



(a) (b) (c)

Fig. 2. Performance profiles of Adaptive MJS-LQR with varying: (a) process noise $\sigma_{\mathbf{w}}$, (b) number of modes s, (c) state dimension n.

$\sigma_{\mathbf{w}} = 0.01, \sigma_{\mathbf{z}} \in \{0.01, 0.02, 0.1\}$ and (ii) $\sigma_{\mathbf{z}} = 0.01, \sigma_{\mathbf{w}} \in \{0.01, 0.02, 0.1\}$. The dimension of states, inputs, and modes are set to $n = 5$, $p = 3$, and $s = 5$, respectively. Fig. 1 (a) and (b) demonstrate how the relative estimation error $\|\hat{\Psi} - \Psi\|/\|\Psi\|$ changes as T increases. Each curve on the plot represents a fixed $\sigma_{\mathbf{w}}$ and $\sigma_{\mathbf{z}}$. These empirical results are all consistent with the theoretical bounds of MJS-SYSID given in (5). In particular, the estimation errors degrade with increasing $\sigma_{\mathbf{w}}$ and decreasing $\sigma_{\mathbf{z}}$, respectively.

Next, we fix $\sigma_{\mathbf{w}} = \sigma_{\mathbf{z}} = 0.01$ and investigate the performance of the MJS-SYSID with varying dimension of states, inputs, and modes. Fig. 1 (c) and (d) show how the estimation error $\|\hat{\Psi} - \Psi\|/\|\Psi\|$ changes with (left) $s = 5$, $n \in \{5, 10, 20\}$, $p = n - 2$ and (right) $n = 5$, $p = n - 2$, $s \in \{5, 10, 20\}$. As we can see, the MJS-SYSID has better performance with small n, p and s which is consistent with (5).

B. Performance of Adaptive MJS-LQR

In our next series of experiments, we explore the sensitivity of the regret bounds to the system parameters. In these experiments, we set the initial epoch length $T_0 = 2000$ and incremental ratio $\gamma = 2$. We select five epochs to run Algorithm 2. As an intermediate step for computing controller $\mathbf{K}_{1:s}^{(i+1)}$ in Algorithm 2, the coupled Riccati equations (7) are solved via value iteration, and the iteration stops when the parameter variation between two iterations falls below 10^{-6}, or iteration number reaches 10^4.

Fig. 2 demonstrates how regret bounds vary with (a) $\sigma_{\mathbf{w}} \in \{0.001, 0.002, 0.01, 0.02\}$, $n = 10$, $p = s = 5$; (b) $\sigma_{\mathbf{w}} = 0.01$, $n = 10$, $p = 5$, $s \in \{4, 6, 8, 10\}$, and (c) $\sigma_{\mathbf{w}} = 0.01$, $s = 10$, $p = 5$, $n \in \{4, 6, 8, 10\}$. We see that the regret degrades as $\sigma_{\mathbf{w}}, n$, and s increase. We also see that when $\sigma_{\mathbf{w}}$ is large (T is small), the regret becomes worse quickly as n and s grow larger. These results are consistent with the theoretical bounds in Theorem 2.

VI. Conclusions and Discussion

Markov jump systems are fundamental to a rich class of control problems where the underlying dynamics are changing with time. Despite its importance, statistical understanding (system identification and regret bounds) of MJS have been lacking due to the technicalities such as Markovian transitions and weaker notion of mean-square stability. At a high-level, this work overcomes (much of) these challenges to provide finite sample system identification and model-based adaptive control guarantees for MJS. Notably, resulting estimation error and regret bounds are optimal in the trajectory length and coincide with the standard LQR up to polylogarithmic factors.

While this work leads to some nontrivial progress in statistical understanding of MJS, there is still room for improvement. In identification algorithm only one out of $C_{sub}\log(T)$ data is used, and we seek to devise approaches and corresponding analysis tools that can make use of all the data. Also, as we discussed earlier below Theorem 2, under stronger stability, the $\frac{1}{\delta}$ dependency on failure probability may be improved to $\log(\frac{1}{\delta})$.

As future work, it would be interesting and of practical importance to investigate the case when mode is not observed, which makes both system identification and adaptive quadratic control problems even more non-trivial.

Acknowledgements

Y. Sattar and S. Oymak were supported in part by NSF grant CNS-1932254 and S. Oymak was supported in part by NSF CAREER award CCF-2046816 and ARO MURI grant W911NF-21-1-0312. Z. Du and N. Ozay were supported in part by ONR under grant N00014-18-1-2501 and N. Ozay was supported in part by NSF under grant CNS-1931982 and ONR under grant N00014-21-1-2431. Z. Du, D. Ataee Tarzanagh, and L. Balzano were supported in part by NSF CAREER award CCF-1845076 and AFOSR YIP award FA9550-19-1-0026.

References

[1] 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.

[2] Y. Sattar, Z. Du, D. A. Tarzanagh, N. Ozay, L. Balzano, and S. Oymak, "Identification and adaptive control of markov jump systems: Sample complexity and regret bounds," in *ICML Workshop on Reinforcement Learning Theory*, 2021.

[3] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. of COLT*. JMLR Workshop and Conference Proceedings, 2011, pp. 1–26.

[4] M. Abeille and A. Lazaric, "Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation," in *ICML*. PMLR, 2020, pp. 23–31.

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

[6] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *FOCM*, pp. 1–47, 2019.

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

[8] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Explore more and improve regret in linear quadratic regulators," *arXiv preprint arXiv:2007.12291*, 2020.

[9] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control," in *NeurIPS*, 2019.

[10] H. J. Chizeck, A. S. Willsky, and D. Castanon, "Discrete-time markovian-jump linear quadratic optimal control," *Intl. Journal of Control*, vol. 43, no. 1, pp. 213–231, 1986.

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

[12] L. Blackmore, S. Funiak, and B. C. Williams, "Combining stochastic and greedy search in hybrid estimation," in *AAAI*, 2005, pp. 282–287.

[13] D. Cajueiro, "Stochastic optimal control of jumping markov parameter processes with applications to finance," Ph.D. dissertation, PhD thesis, 2002, Instituto Tecnológico de Aeronáutica-ITA, Brazil, 2002.

[14] K. Loparo and F. Abdel-Malek, "A probabilistic approach to dynamic power system security," *IEEE transactions on circuits and systems*, vol. 37, no. 6, pp. 787–798, 1990.

[15] L. E. Svensson, N. Williams, *et al.*, "Optimal monetary policy under uncertainty: a markov jump-linear-quadratic approach," *Federal Reserve Bank of St. Louis Review*, vol. 90, no. 4, pp. 275–293, 2008.

[16] V. Ugrinovskii* and H. R. Pota, "Decentralized control of power systems via robust control of uncertain markov jump parameter systems," *International Journal of Control*, vol. 78, no. 9, pp. 662–677, 2005.

[17] P. E. Caines and J.-F. Zhang, "On the adaptive control of jump parameter systems via nonlinear filtering," *SIAM J. Control Optim.*, vol. 33, no. 6, pp. 1758–1777, 1995.

[18] F. Xue and L. Guo, "Necessary and sufficient conditions for adaptive stablizability of jump linear systems," *Communications in Information and Systems*, vol. 1, no. 2, pp. 205–224, 2001.

[19] K. Gatsis and G. J. Pappas, "Statistical learning for analysis of networked control systems over unknown channels," *Automatica*, vol. 125, p. 109386, 2021.

[20] M. Schuurmans, P. Sopasakis, and P. Patrinos, "Safe learning-based control of stochastic jump linear systems: a distributionally robust approach," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 6498–6503.

[21] J. P. Jansch-Porto, B. Hu, and G. Dullerud, "Policy learning of mdps with mixed continuous/discrete variables: A case study on model-free control of markovian jump systems," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 947–957.

[22] M. Ibrahimi, A. Javanmard, and B. Van Roy, "Efficient reinforcement learning for high dimensional linear quadratic systems." in *NIPS*, 2012, pp. 2645–2653.

[23] 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.

[24] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Advances in Neural Information Processing Systems*, 2018, pp. 4188–4197.

[25] 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.

[26] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *ICML*. PMLR, 2020, pp. 8937–8948.

[27] Y. Jedra and A. Proutiere, "Finite-time identification of stable linear systems optimality of the least-squares estimator," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 996–1001.

[28] S. Oymak and N. Ozay, "Non-asymptotic identification of lti systems from a single trajectory," *American Control Conference*, 2019.

[29] R. G. Gallager, *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[30] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.

[31] A. Zhang and M. Wang, "Spectral state compression of markov processes," *IEEE transactions on information theory*, vol. 66, no. 5, pp. 3202–3231, 2019.

[32] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012, p. 210–268.

[33] 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.

[34] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems," in *Advances in Neural Information Processing Systems*, 2020.