# Identification and Adaptive Control of Markov Jump Systems: Sample Complexity and Regret Bounds

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

## Abstract

Learning how to effectively control unknown dynamical systems is crucial for autonomous systems. This task becomes more challenging 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. By taking a model-based perspective, we consider identification-based adaptive control for MJSs. 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 $\tilde{\mathcal{O}}(1/\sqrt{T})$. 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 the proposed adaptive control scheme achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret, which can be improved to $\hat{\mathcal{O}}(\log(T))$ with partial knowledge of the system. Our analysis introduces innovations to handle MJS specific challenges (e.g. Markovian jumps) and provides insights into system theoretic quantities that affect learning accuracy and control performance.

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of California, Riverside, USA [2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA. Correspondence to: Yahya Sattar <ysatt001@ucr.edu>, Zhe Du <zhedu@umich.edu>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

## 1. Introduction

A canonical problem at the intersection of machine learning and control 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 (Campi & Kumar, 1998; Abbasi-Yadkori & Szepesvári, 2011; Dean et al., 2019; Mania et al., 2019; Lale et al., 2020a; Abeille & Lazaric, 2020).

A generalization of linear dynamical systems that can capture dynamics that switch between multiple linear systems, called modes, according to an underlying finite Markov chain is Markov jump linear systems (MJSs). MJS allows for modeling a richer set of problems where the underlying dynamics can abruptly change over time. One can, similarly, generalize the LQR paradigm to MJS by using mode-dependent cost matrices, which allows different control goals under different modes. While the MJS-LQR problem is also well understood when one has perfect knowledge of the system dynamics (Chizeck et al., 1986; Costa et al., 2006), in practice, it is not always possible to know 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. Earlier works have aimed at analyzing the asymptotic properties (i.e., stability) of adaptive controllers for unknown MJS both in continuous-time (Caines & Zhang, 1995) and discrete-time (Xue & Guo, 2001) settings, however, despite the practical importance of MJS, non-asymptotic sample complexity results and regret analysis for MJS 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

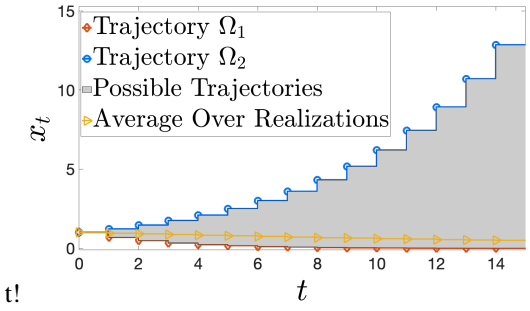


*Figure 1.* State trajectories for a two-mode MJS $\begin{cases} x_{t+1} = 0.7x_t \\ x_{t+1} = 1.2x_t \end{cases}$ with Markov matrix $\begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$ and $\mathbf{x}_0 = 1$. Red and blue curves: mode switching sequence $\Omega_1 = \{1, 1, \ldots\}, \Omega_2 = \{2, 2, \ldots\}$. Yellow curve: average over realizations. Gray area: region for all possible trajectories.

is that, typically, the stability of MJS is understood only in the *mean-square sense*. This is in stark contrast to deterministic stability (e.g., as in LQR), where the system is guaranteed to converge towards an equilibrium point in the absence of noise. On the other hand, the convergence of MJS trajectories towards an equilibrium depends heavily on how the switching between modes occurs.

Figure 1 shows an example (reproduced from (Costa et al., 2006)) of an MJS that is stable in the mean square sense despite having an unstable mode. Clearly, under an unfavorable mode switching sequence, the system trajectory can still blow up. High-probability light tail bounds are therefore not applicable without very strong assumptions on the joint spectral radius of different modes (cf. (Sarkar et al., 2019)). Perhaps more surprisingly, there are examples of MJS with all modes individually stable, however due to switching, the system exhibits an unstable behavior on average, and the MJS is not mean square stable. Therefore, finding controllers to individually stabilize the mode dynamics does not guarantee that overall system will be stable when mode switches over time. This more relaxed notion of *mean-square stability* presents major challenges in learning, controlling, and the statistical analysis.

**Contributions:** In this paper, we provide the first comprehensive system identification and regret guarantees for learning and controlling Markov jump linear systems using a single trajectory. Importantly, our guarantees are optimal in the trajectory length $T$. Specifically, our contributions are:

**(I) System identification:** For an MJS with $s$ modes, the system dynamics involve Markov chain 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 optimal error rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$ [1]. Specifically, the sample complexity grows as $T \gtrsim \text{poly}(s)(n+p)$ where $n$ and $p$ are the state and input dimensions respectively.

**(II) $\tilde{\mathcal{O}}(\sqrt{T})$-regret bound:** We employ the system identification guarantees for the MJS-LQR. When system dynamics are unknown, we show that our certainty-equivalent adaptive MJS-LQR algorithm (Alg. 2) achieves a regret of $\tilde{\mathcal{O}}(\sqrt{T})$. Remarkably, this coincides with the optimal regret bound for the standard LQR problem obtained via certainty equivalence (Mania et al., 2019). Furthermore, we show that when the input matrices are known, the regret bound can be significantly improved to $\mathcal{O}(\text{polylog}(T))$, which coincides with the case in (Cassel et al., 2020) for standard LQR.

## 2. Preliminaries and Problem Setup

We use boldface uppercase (lowercase) letters to denote matrices (vectors). For a matrix $\mathbf{V}$, $\rho(\mathbf{V})$ denotes its spectral radius. 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 $[s] := \{1, 2, \ldots, s\}$. Throughout, $\tilde{\mathcal{O}}(\cdot)$ and $\hat{\mathcal{O}}(\cdot)$ hide $\text{polylog}(\frac{1}{\delta})$ and $\text{poly}(\frac{1}{\delta})$ terms respectively.

### 2.1. Markov Jump Linear Systems

In this paper we consider the identification and control of MJS which are governed by the following state equation,

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

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 $t$. Throughout, we assume $\mathbf{x}_0 \sim \mathcal{D}_x$ and $\{\mathbf{w}_t\}_{t=0}^\infty \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{w}}^2 \mathbf{I}_n)$. 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 $\omega(t) \in [s]$. The MJS 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 $[\mathbf{T}]_{ij} := \mathbb{P}\left(\omega(t+1) = j \mid \omega(t) = i\right)$, $\forall\, i, j \in [s]$. Throughout, we assume that the initial state $\mathbf{x}_0$, Markov chain $\{\omega(t)\}_{t=0}^\infty$, and noise $\{\mathbf{w}_t\}_{t=0}^\infty$ are mutually independent. We use $\text{MJS}(\mathbf{A}_{1:s}, \mathbf{B}_{1:s}, \mathbf{T})$ 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 (Costa et al., 2006))**.**

[1] Here $\tilde{\mathcal{O}}(\cdot)$ hides ploylogarithmic factors in $T$, $1/\delta$ etc.

*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*

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

*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 (mean-square) stability 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 $[\tilde{\mathbf{L}}]_{ij} := [\mathbf{T}]_{ji}\mathbf{L}_j \otimes \mathbf{L}_j$. Specifically, if $\rho(\tilde{\mathbf{L}}) < 1$, a noise-free autonomous MJS can be shown to satisfy MSS (Costa et al., 2006).

**Assumption A1.** *The MJS in* (1) *is stabilizable, and its underlying Markov chain (*$\mathbf{T}$*) is ergodic.*

Stabilizability allows us to use a mixing argument to obtain weakly dependent sub-trajectories by properly subsampling the original trajectory, and ergodicity guarantees that the Markov chain converges to a unique stationary distribution. Throughout, $\boldsymbol{\pi}_\infty$ denotes the stationary distribution of $\mathbf{T}$ with $\pi_{\min} := \min_i \boldsymbol{\pi}_\infty(i)$. We further define the mixing time (Levin & Peres, 2017) of $\mathbf{T}$ as $t_{\mathrm{MC}} := \inf\left\{t \in \mathbb{N} : \max_{i\in[s]} \|([\mathbf{T}^t]_{i,:})^\intercal - \boldsymbol{\pi}_\infty\|_1 \le 0.5\right\}$, where $[\mathbf{T}^t]_{i,:}$ denotes the $i_{th}$ row of $\mathbf{T}^t$. Note that $t_{MC}$ plays a key role in the mixing time of the overall MJS. In the analysis, $\pi_{\min}$ guarantees one could obtain enough data for each mode, while the mixing time $t_{MC}$ of the MJS determines the fraction of the data that provably helps towards learning the system.

### 2.2. Problem Formulation

In this paper, we consider two major problems under the MJS setting: System identification and adaptive control, with identification being the core part of adaptive control. **(A) System Identification.** This problem seeks to estimate unknown system dynamics from data, i.e. from input-output trajectory and the mode observation, when one has the flexibility to design the input 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 matrix $\mathbf{T}$. In this work, we seek to estimate the MJS dynamics using only a single trajectory $\{\mathbf{x}_t, \mathbf{u}_t, \omega(t)\}_{t=0}^T$ and provide finite sample guarantees. Section 3 presents our system identification results. **(B) Online Linear Quadratic Regulator.** In this paper, we consider the following finite-horizon Markov jump system linear quadratic regulator (MJS-LQR) problem:

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

The goal is to design inputs to minimize the expected quadratic cost composed of positive semi-definite matrices $\mathbf{Q}_{1:s}$ and $\mathbf{R}_{1:s}$ under the MJS dynamics. We will 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). We assume the following for cost matrices.

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

We assume the state $\mathbf{x}_t$ and mode $\omega(t)$ can be observed at time $t$. With these observations, instead of a fixed and open-loop input sequence, one can design closed-loop policies that generate real-time input based on 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 Riccati equations (Costa et al., 2006). In our 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 infinite-horizon cost $T \cdot J^\star$ where $J^*$ 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 (Dean et al., 2018), in MJS-LQR setting, the analysis requires additional consideration of Markov chain mixing, which is addressed in this paper.

## 3. 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 (Dean et al., 2018). Note that if the open-loop MJS is already MSS, then one can simply set $\mathbf{K}_{1:s}^{(0)} = 0$. The following theorem gives our main results on learning the dynamics of an unknown MJS from finite samples obtained from a single trajectory.

**Algorithm 1** MJS-SYSID

**Input:** A mean square stabilizing controller $\mathbf{K}_{1:s}$, dynamics noise $\sigma_{\mathbf{w}}^2$, exploration noise $\sigma_{\mathbf{z}}^2$, trajectory $\{\mathbf{x}_t, \mathbf{z}_t, \omega(t)\}_{t=0}^{T}$ generated using input $\mathbf{u}_t = \mathbf{K}_{\omega(t)}\mathbf{x}_t + \mathbf{z}_t$ with $\mathbf{z}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{z}}^2 \mathbf{I}_p)$, data clipping thresholds $c_{\mathbf{x}}$, $c_{\mathbf{z}}$, subsampling factor $C_{\text{sub}}$.

Set subsampling period $L = C_{\text{sub}} \log(T)$

Set subsampling indices $\tau_k = kL$ for $k = 1, 2, \ldots, \lfloor T/L \rfloor$

**Estimate** $\mathbf{A}_{1:s}, \mathbf{B}_{1:s}$: **for all** modes $i \in [s]$ **do**: $S_i = \{\tau_k \mid \omega(\tau_k) = i,\ \|\mathbf{x}_{\tau_k}\| \le c_{\mathbf{x}} \sigma_{\mathbf{w}} \sqrt{\log(T)},\ \|\mathbf{z}_{\tau_k}\| \le c_{\mathbf{z}} \sigma_{\mathbf{z}}\}$,

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

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

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

**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. Suppose $\mathbf{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{w}}^2 \mathbf{I}_n)$. Suppose the trajectory length $T \ge \tilde{\mathcal{O}}(\log^2(T)(n+p)/\pi_{\min})$ and the sampling factor satisfies $C_{sub} \ge t_{\text{MC}} \cdot \mathcal{O}(1/(1-\rho))$. Then, under Assumption A1, with probability at least $1 - \delta$, for all $i \in [s]$, we have*

$$\max \left\{ \begin{matrix} \|\hat{\mathbf{A}}_i - \mathbf{A}_i\|, \\ \|\hat{\mathbf{B}}_i - \mathbf{B}_i\| \end{matrix} \right\} \le \tilde{\mathcal{O}}\left( \frac{\sigma_{\mathbf{z}} + \sigma_{\mathbf{w}}}{\sigma_{\mathbf{z}} \pi_{\min}} \frac{(n+p)\log(T)}{\sqrt{T}} \right),$$

$$\text{and } \|\hat{\mathbf{T}} - \mathbf{T}\|_\infty \le \tilde{\mathcal{O}}\left( \frac{1}{\pi_{\min}} \sqrt{\frac{\log(T)}{T}} \right). \tag{5}$$

**Corollary 1.** *Consider the setting of Algorithm 1. When $\mathbf{B}_{1:s}$ are known, setting $\sigma_{\mathbf{z}} = 0$ and solving only for the state matrices leads to a stronger upper bound $\|\hat{\mathbf{A}}_i - \mathbf{A}_i\| \le \tilde{\mathcal{O}}(\frac{(n+p)\log(T)}{\pi_{\min}\sqrt{T}})$.*

Our system identification result achieves optimal statistical error rate of $\tilde{O}(1/\sqrt{T})$. The sample complexity grows quadratically in state dimension $n$, which can potentially be improved to linear via a more refined control on the state-covariance (see (Simchowitz et al., 2018; Dean et al., 2019) for standard linear systems). It also grows with the inverse of the minimum mode frequency as $\pi_{\min}^{-1}$. Note that, $\pi_{\min}$ dictates the trajectory fraction of the least-frequent mode, thus, $\pi_{\min}^{-1}$ multiplier is not avoidable. In Corollary 1, we show that, when the knowledge of $\mathbf{B}$ is assumed, $\mathbf{A}$ can be estimated regardless of the exploration strength $\sigma_{\mathbf{z}}$. This is because the excitation for the state matrix arises from $\mathbf{w}_t$.

**Proof outline for Theorem 1:** Our proof strategy for Algorithm 1 addresses the challenges introduced by MJS and mean-square stability. We only emphasize the core technical challenges. In Algorithm 1, we subsample the trajectory. At a high-level, this will help us upper/lower bound the empirical covariance matrix formed by the subsampled state-input pairs $(\mathbf{x}_{\tau_k}, \mathbf{z}_{\tau_k})$ for all $\tau_k \in S_i$. Initial subsampling (with spacing $L$) aims to reduce the statistical dependence across the input data $(\mathbf{x}_t, \mathbf{z}_t)_{t \ge 0}$ to obtain a weakly-dependent sub-trajectory with indices $\tau_k$. This dependence is due to the mode sequence $\omega(t)$ – unique to the MJS setting – and the system's memory (contribution of the earlier states on the current state). Thus $L$ is primarily a function of the mixing-time of $\mathbf{T}$ and the spectral radius of the MJS system. Unlike related works on sysid and regret analysis (Simchowitz et al., 2018; Dean et al., 2018; Lale et al., 2020a; Oymak & Ozay, 2019; Lale et al., 2020b), mean-square stability does not lead to strong high-probability bounds, as one can only bound $\|\mathbf{x}_t\|$ or $\mathbf{x}_t \mathbf{x}_t^\intercal$ in the expectation sense. The second subsampling restricts our attention to the *bounded* $(\mathbf{x}_{\tau_k}, \mathbf{z}_{\tau_k})$ pairs on mode $i$. This boundedness enables us to control the covariance matrix despite MSS and potentially heavy-tailed states via non-asymptotic toolset (e.g. Thm 5.44 of (Vershynin, 2010)). However, heavy-tailed empirical covariance lower bounds require independence and our subsampled data are only "approximately independent" (coupled over modes and history). To make matters worse, the fact that we sample bounded states introduces further dependencies. To resolve this, we introduce a novel strategy to construct an independent subset of *processed states* from this larger dependent set. The independence is ensured by conditioning on the mode-sequence and truncating the contribution of earlier states. We then use perturbation-based techniques to deal with actual (non-truncated) states. The fnal igredient is showing that, for each mode $1 \le i \le s$, with high probability, this carefully-crafted subset contains enough samples to ensure a well-conditioned covariance (with excitation provided by $\mathbf{z}_t, \mathbf{w}_t$). With this in place, after two rounds of subsampling, least-squares will accurately estimate $\mathbf{A}$ and $\mathbf{B}$ for all modes with rate $1/\sqrt{T}$.

## 4. Adaptive Control for MJS-LQR

Our adaptive MJS-LQR 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 newly collected trajectory.

Similar to the discussion in Section 3, we assume at the beginning one has access to a stabilizing controller $\mathbf{K}_{1:s}^{(0)}$. During epoch $i$, controller $\mathbf{K}_{1:s}^{(i)}$ is used together with additive exploration noise $\mathbf{z}_t$ to boost learning. At the end of epoch $i$, the trajectory during this epoch is used to obtain a new MJS dynamics estimate $\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)}$ through 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 solved

**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 bound $c_{\mathbf{x}}$, $c_{\mathbf{z}}$, sub-sampling factor $C_{sub}$.
**for** $i = 0, 1, 2, \dots$ **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 for $T_i$ steps with $\mathbf{u}_t^{(i)} = \mathbf{K}_{\omega(t)}^{(i)} \mathbf{x}_t^{(i)} + \mathbf{z}_t^{(i)}$ with $\mathbf{z}_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{z},i}^2 \mathbf{I}_p)$ and record the trajectory $\Phi_i := \{\mathbf{x}_t^{(i)}, \mathbf{z}_t^{(i)}, \omega^{(i)}(t)\}_{t=0}^{T_i}$.
  $\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)}$
    $= \text{MJS-SYSID}(\mathbf{K}_{1:s}^{(i)}, \sigma_{\mathbf{w}}^2, \sigma_{\mathbf{z},i}^2, \Phi_i, c_{\mathbf{x}}, c_{\mathbf{z}}, C_{sub})$
  $\mathbf{K}_{1:s}^{(i+1)} \leftarrow$ optimal controller for the infinite-horizon MJS-LQR$(\mathbf{A}_{1:s}^{(i)}, \mathbf{B}_{1:s}^{(i)}, \mathbf{T}^{(i)}, \mathbf{Q}_{1:s}, \mathbf{R}_{1:s})$.
**end for**

efficiently via value iteration or via LMIs (Costa et al., 2006). Note that this control design based on the estimated dynamics is also referred to as certainty equivalent control.

To have theoretically guaranteed performance, i.e. sub-linear 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 $\mathcal{O}(\log(T)\sqrt{T})$ regret when combined with the system identification result from Theorem 1. Intuitively, this scheduling has interpretations from two folds: (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 exploration noise and deploy (exploit) the controller for a longer time. Note that the scheduling rate $\gamma$ has similar role to the discount factor in reinforcement learning: smaller $\gamma$ aims to reduce short-term cost while larger $\gamma$ aims to reduce long-term cost.

### 4.1. Regret Analysis

We define filtration $\mathcal{F}_{-1}, \mathcal{F}_0, \mathcal{F}_1, \dots$ such that $\mathcal{F}_{-1} := \sigma(\mathbf{x}_0, \omega(0))$ is the sigma algebra generated by the initial state and mode, and $\mathcal{F}_i := \sigma(\mathbf{x}_0, \omega(0), \{\{\omega^{(j)}(t)\}_{t=1}^{T_j}\}_{j=0}^{i}, \mathbf{w}_0, \{\mathbf{w}_{1:T_j}^{(j)}\}_{j=0}^{i}, \mathbf{z}_0, \{\mathbf{z}_{1:T_j}^{(j)}\}_{j=0}^{i})$ is the sigma algebra generated by the randomness up to epoch $i$. Note that since the initial state $\mathbf{x}_0^{(i)}$ of epoch $i$ is 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 $\omega(0)^{(i+1)}$. Suppose time step $t$ belongs to epoch $i$, then we define the following conditional expected cost at time $t$. $c_t = \mathbb{E}[\mathbf{x}_t^\intercal \mathbf{Q}_{\omega(t)} \mathbf{x}_t + \mathbf{u}_t^\intercal \mathbf{R}_{\omega(t)} \mathbf{u}_t \mid \mathcal{F}_{i-1}]$, and cumulative cost as $J_T = \sum_{t=1}^{T} c_t$. We define the total regret and epoch-$i$ regret as

$$\text{Regret}(T) = J_T - T J^\star,$$

$$\text{Regret}_i = (\sum_{t=1}^{T_i} c_{T_0+\cdots+T_{i-1}+t}) - T_i J^\star. \tag{6}$$

One can refer to Appendix C.4 for more discussion on the regret definition. With these definitions, we have the following result.

**Theorem 2** (Sub-linear regret)**.** *If $T_0$, $C_{sub}$, $c_{\mathbf{x}}$, and $c_{\mathbf{z}}$ are large enough, then under Assumption A1 and A2, with probability at least $1-\delta$, Algorithm 2 achieves*

$$\text{Regret}(T) \le \hat{\mathcal{O}}(\log(T)) + \tilde{\mathcal{O}}(\log^2(T)\sqrt{T}). \tag{7}$$

From Corollary 1, we know when $\mathbf{B}_{1:s}$ are known, no further exploration noise is needed to learn $\mathbf{A}_{1:s}$ or $\mathbf{T}$, this applies to the adaptive MJS-LQR setting as well. Getting rid of exploration noise improves the regret as follows.

**Corollary 2** (Poly-log regret)**.** *When $\mathbf{B}_{1:s}$ are known, it suffices to set $\sigma_{\mathbf{z},i} = 0$ for all $i$ in Algorithm 2. Then, Algorithm 2 achieves* $\text{Regret}(T) \le \hat{\mathcal{O}}(\log(T)) + \tilde{\mathcal{O}}(\log^3(T))$.

**Proof outline for Theorem 2:** For simplicity, we only show the dominant $\mathcal{O}(\log^2(T)\sqrt{T})$ term. Define the estimation error after epoch $i$ as $\epsilon_{\mathbf{A},\mathbf{B}}^{(i)} := \max_{j\in[s]} \max\{\|\mathbf{A}_j^{(i)} - \mathbf{A}_j\|, \|\mathbf{B}_j^{(i)} - \mathbf{B}_j\|\}$, $\epsilon_{\mathbf{T}}^{(i)} := \|\mathbf{T}^{(i)} - \mathbf{T}\|_\infty$, Using perturbation result (Du et al., 2021) for infinite-horizon MJS-LQR together with new finite-horizon cost analysis, we can bound epoch-$i$ regret as follows: $\text{Regret}_i \le \mathcal{O}\left(T_i \sigma_{\mathbf{z},i}^2 + T_i \sigma_{\mathbf{w}}^2 \left(\epsilon_{\mathbf{A},\mathbf{B}}^{(i-1)} + \epsilon_{\mathbf{T}}^{(i-1)}\right)^2\right)$. Next, plugging in the upper bounds on the estimation errors $\epsilon_{\mathbf{A},\mathbf{B}}^{(i)} \le \mathcal{O}\left(\frac{\sigma_{\mathbf{z},i}+\sigma_{\mathbf{w}}}{\sigma_{\mathbf{z},i}} \frac{\log(T_i)}{\sqrt{T_i}}\right)$, $\epsilon_{\mathbf{T}}^{(i)} \le \mathcal{O}\left(\sqrt{\frac{\log(T_i)}{T_i}}\right)$ from Theorem 1, and using the exploration variance $\sigma_{\mathbf{z},i}^2 = \frac{\sigma_{\mathbf{w}}^2}{\sqrt{T_i}}$, we have $\text{Regret}_i \le \mathcal{O}(\sigma_{\mathbf{w}}^2 \sqrt{\gamma} \sqrt{T_i} \log^2(T_i))$. Finally, we have: $\text{Regret}(T) = \sum_{i=1}^{\mathcal{O}(\log_\gamma(\frac{T}{T_0}))} \text{Regret}_i \le \mathcal{O}\left(\sigma_{\mathbf{w}}^2 \log(\frac{T}{T_0}) \sqrt{\frac{T}{T_0}} \left(\frac{\sqrt{\gamma}}{\sqrt{\gamma}-1}\right)^3 \left(\sqrt{\gamma}\log(\frac{T}{T_0}) - \log(\sqrt{\gamma})\right)\right) = \mathcal{O}(\log^2(T)\sqrt{T})$.

## 5. 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 poly-logarithmic factors.

## Acknowledgements

Y. Sattar and S. Oymak were supported in part by NSF under grant CNS-1932254. 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. L. Balzano was supported in part by NSF CAREER award CCF-1845076, NSF BIGDATA award IIS1838179, ARO YIP award W911NF1910027, and the Institute for Advanced Study Charles Simonyi Endowment.

## References

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

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

Caines, P. E. and Zhang, J.-F. On the adaptive control of jump parameter systems via nonlinear filtering. *SIAM J. Control Optim.*, 33(6):1758–1777, 1995.

Campi, M. C. and Kumar, P. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM J. Control Optim.*, 36(6):1890–1907, 1998.

Cassel, A., Cohen, A., and Koren, T. Logarithmic regret for learning linear quadratic regulators efficiently. In *International Conference on Machine Learning*, pp. 1328–1337. PMLR, 2020.

Chizeck, H. J., Willsky, A. S., and Castanon, D. Discrete-time markovian-jump linear quadratic optimal control. *International Journal of Control*, 43(1):213–231, 1986.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. Online linear quadratic control. In *International Conference on Machine Learning*, pp. 1029–1038. PMLR, 2018.

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

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

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

Du, Z., Ozay, N., and Balzano, L. Mode clustering for markov jump systems. *arXiv preprint arXiv:1910.02193*, 2019.

Du, Z., Sattar, Y., Tarzanagh, D. A., Balzano, L., Oymak, S., and Ozay, N. Certainty equivalent quadratic control for markov jump systems. *arXiv e-prints*, 2021.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Goel, G. and Hassibi, B. The power of linear controllers in lqr control. *arXiv preprint arXiv:2002.02574*, 2020.

Hsu, D., Kakade, S., Zhang, T., et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

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

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

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Mania, H., Tu, S., and Recht, B. Certainty equivalence is efficient for linear quadratic control. In *NeurIPS*, 2019.

Oymak, S. Stochastic gradient descent learns state equations with nonlinear activations. In *Conference on Learning Theory*, pp. 2551–2579, 2019.

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

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. In *ICML*, pp. 5610–5618. PMLR, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. Nonparametric system identification of stochastic switched linear systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3623–3628. IEEE, 2019.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. *arXiv preprint arXiv:1802.08334*, 2018.

Tu, S., Boczar, R., Packard, A., and Recht, B. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Xue, F. and Guo, L. Necessary and sufficient conditions for adaptive stablizability of jump linear systems. *Communications in Information and Systems*, 1(2):205–224, 2001.

Zhang, A. and Wang, M. State compression of markov processes via empirical low-rank estimation. *arXiv preprint arXiv:1802.02920*, 2018.