Strong Consistency and Rate of Convergence of Switched Least Squares System Identification for Autonomous Markov Jump Linear Systems

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

arXiv:2112.10753v2 [cs.LG] 4 Feb 2023

***Abstract*—In this paper, we investigate the problem of system identification for autonomous Markov jump linear systems (MJS) with complete state observations. We propose switched least squares method for identification of MJS, show that this method is strongly consistent, and derive data-dependent and data-independent rates of convergence. In particular, our data-independent rate of convergence shows that, almost surely, the system identification error is $\mathcal{O}(\sqrt{\log(T)/T})$ where T is the time horizon. These results show that switched least squares method for MJS has the same rate of convergence as least squares method for autonomous linear systems. We derive our results by imposing a general stability assumption on the model called stability in the average sense. We show that stability in the average sense is a weaker form of stability compared to the stability assumptions commonly imposed in the literature. We present numerical examples to illustrate the performance of the proposed method.**

I. Introduction

Markov jump linear systems (MJS) are a good approximation of non-linear time-varying systems arising in various applications including networked control systems [2] and cyber-physical systems [3], [4]. There is a rich literature on the stability analysis (e.g., [5]–[7]) and optimal control (e.g., [8]) of MJS. However, most of the literature assumes that the system model is known. The question of system identification, i.e., identifying the dynamics from data, has not received much attention in this setup.

The problem of identifying the system model from data is a key component for control synthesis for both offline control methods and online control methods including adaptive control and reinforcement learning [9], [10]. There are four main approaches for system identification of linear systems: (i) maximum likelihood estimation which maximizes the likelihood function of the unknown parameter given the observation (e.g. see [11]); (ii) minimum prediction error methods which minimize the estimation error (residual process) according to some loss function (e.g. see [12], [13]); (iii) subspace methods, which find a minimum state space realization given the input, output data (e.g. see [14], [15]); (iv) least squares method which estimates the unknown parameter by considering the model as a regression problem (e.g. see [16], [17]).

These methods differ in terms of structural assumptions on the model (e.g. system order), hypotheses on the stochastic process, and convergence properties and guarantees.

Structural assumptions require the system to be stable in some sense (e.g., mean square stable, exponentially stable, etc.), and stochastic hypotheses restrict the noise processes to be of a certain type, (e.g., Gaussian, sub-Gaussian, or Martingale difference sequences).

Convergence properties characterize the asymptotic behavior of system identification methods. The basic requirements for any system identification method is its consistency, asymptotic normality and rates of convergence, that is to establish that estimates converge asymptotically to the true unknown parameter and characterize the rate of convergence. System identification methods can be *weakly* consistent (i.e., estimates converge in probability) or *strongly* consistent (i.e., estimates convergence almost surely). For linear systems, there is a vast literature that establishes the consistency and rates of convergence for a variety of methods (e.g. see [10], [17] for a unified overview). Another characterization of the convergence is finite-time guarantees which provide lower-bounds on the number of samples required so that estimates have a specified degree of accuracy with a specified high probability [18]–[26]. As the number of samples grow to infinity, these results establish weak consistency of the proposed methods.

System identification of MJS and switched linear systems (SLS) has received less attention in the literature. There is some work on designing asymptotically stable controllers for unknown SLS [27]–[29] but these papers do not establish rates of convergence for system identification. There are some recent papers which provide finite time guarantees and rate of convergence for SLS [30], [31] and MJS [32]. System identification of a globally asymptotically stable SLS with controlled switching signal is investigated in [30], while the system identification of an unknown order SLS using subspace methods is investigated in [31]. Both these methods are developed for SLS and are not directly applicable to MJS. The model analyzed in [32] is an MJS system. Under the assumption that the system is mean square stable, the switching distribution is ergodic and the noise is i.i.d. subgaussian, it is established that the convergence rate is $\mathcal{O}(\sqrt{\log T/T})$ with high probability. Then a certainty equivalence control algorithm is proposed and its regret is analyzed. Note that

An earlier version of this paper is presented in 61st Conference on Decision and Control (CDC), Cancun, Mexico, 2022 [1].

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

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

if we let the number of samples go to infinity, these results imply *weak* consistency of the proposed methods for MJS systems. As far as we are aware, there is no existing result which establishes *strong* consistency of a method for system identification of MJS.

A. Contributions

- We propose a *switched* least squares method for system identification of an unknown (autonomous) MJS and provide data-dependent and data-independent rates of convergence for this method.
- We prove *strong* consistency of the switched least squares method and establish a $\mathcal{O}(\sqrt{\log(T)/T})$ rate of convergence, which matches with the rate of convergence of non-switched linear systems established in [16]. In contrast to the existing high-probability convergence guarantees in the literature, our results show that the estimates converge to the true parameters *almost surely*. Therefore, our results provide guarantees which are different in nature compared to parallel works.
- The main challenge in establishing strong consistency for MJS systems is the interplay between the empirical covariance process and stability of the MJS system. We shed light on this connection and show that stability in the average sense is a sufficient condition for strong consistency.
- Our results are derived under weaker assumptions compared to the existing literature (Note that the preliminary version of this paper [1] had assumed a stronger stability condition). Most existing results assume that the MJS system is mean square stable. We prove that mean square stability implies stability in the average sense. Furthermore, we show that a commonly used sufficient condition for almost sure stability of noise-free MJS system also implies stability in the average sense.

B. Organization

The rest of the paper is organized as follows. In Sec II, we present the system model, assumptions, and the main results. In Sec. III, we prove the main results. In Sec. IV, we explain the connection of stability in the average sense with mean square stability and almost sure stability. We present an illustrative example in Sec. V. We conclude in Sec. VI.

C. Notation

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

Given a sequence of positive numbers $\{a_t\}_{t\geq 0}, a_T = \mathcal{O}(T)$ means that $\limsup_{T\to\infty} a_T/T < \infty$, and $a_T = o(T)$ means that $\limsup_{T\to\infty} a_T/T = 0$. Given a sequence of vectors $\{x_t\}_{t\in\mathcal{T}}$, $\mathrm{vec}(x_t)_{t\in\mathcal{T}}$ denotes the vector formed by vertically stacking $\{x_t\}_{t\in\mathcal{T}}$. Given a sequence of random variables $\{x_t\}_{t\geq 0}$, $x_{0:t}$ is a short hand for $(x_0, \cdots, x_t)$ and $\sigma(x_{0:t})$ denotes the sigma field generated by random variables $x_{0:t}$. Given a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$, Ω denotes the sample space, $\omega \in \Omega$ denotes elementary events, $\mathbb{P}(\cdot)$ denotes the probability measure and $\mathbb{E}[\cdot]$ denotes the expectation operator.

$\mathbb{R}$ and $\mathbb{N}$ denote the sets of real and natural numbers. For a set $\mathcal{T}$, $|\mathcal{T}|$ denotes its cardinality. For a vector x, $\|x\|$ denotes the Euclidean norm. For a matrix A, $\|A\|$ denotes the spectral norm and $\|A\|_\infty$ denotes the element with the largest absolute value. $\mathrm{diag}(\cdot)$ is the block diagonal matrix. Convergence in almost sure sense is abbreviated as $a.s.$

II. System model and problem formulation

Consider a discrete-time (autonomous) MJS. The state of the system has two components: a discrete component $s_t \in \mathcal{S} \coloneqq \{1, \ldots, k\}$ and a continuous component $x_t \in \mathbb{R}^n$. There is a finite set $\mathcal{A} = \{A_1, \ldots, A_k\}$ of system matrices, where $A_i \in \mathbb{R}^{n\times n}$. The continuous component x_t of the state starts at a fixed value x_0 and the initial discrete state s_0 starts according to a prior distribution π_0. The continuous state evolves according to:

$$x_{t+1} = A_{s_t} x_t + w_t, \quad t \geq 0, \tag{1}$$

where $\{w_t\}_{t\geq 0}$, $w_t \in \mathbb{R}^n$, is a noise process. The discrete component evolves in a Markovian manner according to a time-homogeneous irreducible and aperiodic transition matrix P, i.e. $\mathbb{P}(s_{t+1} = j | s_t = i) = P_{ij}$.

Let $\pi_t = (\pi_t(1), \ldots, \pi_t(k))$ denote the probability distribution of the discrete state at time t and π_∞ denote the stationary distribution. We assume $\pi_\infty(i) \neq 0$ for all i. Let $\mathcal{F}_{t-1} = \sigma(x_{0:t}, s_{0:t})$ denote the sigma-algebra generated by the history of the complete state.

It is assumed that the noise process satisfies the following:

Assumption 1. *The noise process $\{w_t\}_{t\geq 0}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t\geq 0}$, i.e., $\mathbb{E}[|w_t|] < \infty$ and $\mathbb{E}[w_t \mid \mathcal{F}_{t-1}] = 0$. Furthermore, there exists a constant $\alpha > 2$ such that $\sup_{t\geq 0} \mathbb{E}[\|w_t\|^\alpha \mid \mathcal{F}_{t-1}] < \infty$ a.s. and there exists a symmetric and positive definite matrix $C \in \mathbb{R}^{n\times n}$ such that $\liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} w_t w_t^\intercal = C$ a.s.*

Assumption 1 is a standard assumption in the asymptotic analysis of system identification of linear systems [16], [17], [33]–[35] and allows the noise process to be non-stationary and have heavy tails (as long as moment condition is satisfied). We use the following notion of stability for the MJS system (1).

Definition 1. *The MJS system* (1) *is called stable in the* average sense *if almost surely:*

$$\sum_{t=1}^{T} \|x_t\|^2 = \mathcal{O}(T) \quad \textit{i.e.} \quad \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \|x_t\|^2 < \infty.$$

Assumption 2. *The MJS system* (1) *is stable in the average sense.*

The notion of stability in the average sense has been used in a few papers in the literature of linear systems [36], [37]; however, in the MJS literature, the commonly used notions of stability are mean square stability and almost sure stability of *noise-free* system. We compare stability in the average sense with both of these notions in Sec. IV. Specifically, we show that mean square stability implies stability in the average sense. Moreover, we show a common sufficient condition for almost sure stability of noise-free system implies stability in the average sense for MJS system (1). Therefore, the assumption of stability in the average sense is weaker than the commonly imposed stability assumptions imposed in the literature.

A. System identification and switched least squares estimates

We are interested in the setting where the system dynamics $\mathcal{A}$ and the switching transition matrix P are unknown. Let $\theta^\intercal = [A_1, \dots, A_k] \in \mathbb{R}^{n \times nk}$ denote the unknown parameters of the system dynamics matrices. We consider an agent that observes the complete state (x_t, s_t) of the system at each time and generates an estimate $\hat{\theta}_T$ of θ as a function of the observation history $(x_{0:T}, s_{0:T})$. A commonly used estimate in such settings is the least squares estimate:

$$\hat{\theta}_T^\intercal = \underset{\theta^\intercal = [A_1, \dots, A_k]}{\arg\min} \sum_{t=0}^{T-1} \|x_{t+1} - A_{s_t} x_t\|^2. \tag{2}$$

The components $[\hat{A}_{1,T}, \dots, \hat{A}_{k,T}] = \hat{\theta}_T^\intercal$ of the least squares estimate can be computed in a switched manner. Let $\mathcal{T}_{i,T} = \{t \le T \mid s_t = i\}$ denote the time indices until time T when the discrete state of the system equals i. Note that for each $t \in \mathcal{T}_{i,T}$, $A_{s_t} = A_i$. Therefore, we have

$$\hat{A}_{i,T} \coloneqq \underset{A_i \in \mathbb{R}^{n \times n}}{\arg\min} \sum_{t \in \mathcal{T}_{i,T}} \|x_{t+1} - A_i x_t\|^2, \quad \forall i \in \{1, \cdots, k\}. \tag{3}$$

Let $X_{i,T}$ denote $\sum_{t \in \mathcal{T}_{i,T}} x_t x_t^\intercal$, which we call the unnormalized empirical covariance of the continuous component of the state at time T when the discrete component equals i. Then, $\hat{A}_{i,T}$ can be computed recursively as follows:

$$\hat{A}_{i,T+1} = \hat{A}_{i,T} + \left[\frac{X_{i,T}^{-1} x_T (x_{T+1} - \hat{A}_{i,T} x_T)^\intercal}{1 + x_T^\intercal X_{i,T}^{-1} x_T}\right] \mathbb{1}\{s_{T+1} = i\} \tag{4}$$

where $X_{i,T}$ may be updated as $X_{i,T+1} = X_{i,T} + \left[x_{T+1} x_{T+1}^\intercal\right] \mathbb{1}\{s_{T+1} = i\}$. Due to the switched nature of the least squares estimate, we refer to above estimation procedure as *switched least squares* system identification.

B. The main results

A fundamental property of any sequential parameter estimation method is strong consistency, which we define below.

Definition 2. *An estimator $\hat{\theta}_T$ of parameter θ is called strongly consistent if $\lim_{T \to \infty} \hat{\theta}_T = \theta$, a.s.*

Our main result is to establish that the switched least squares estimator is strongly consistent. We do so by providing two different characterization of the rate of convergence. We first provide a data-dependent rate of convergence which depends on the spectral properties of the unnormalized empirical covariance. We then present a data-independent characterization of rate of convergence which only depends on T. All the proofs are presented in Sec. III.

Theorem 1. *Under Assumptions 1 and 2, the switched least squares estimates $\{\hat{A}_{i,T}\}_{i=1}^k$ are strongly consistent, i.e., for each $i \in \mathcal{S}$, we have: $\lim_{T \to \infty} \|\hat{A}_{i,T} - A_i\|_\infty = 0$, a.s. Furthermore, the rate of convergence is upper bounded by:*

$$\|\hat{A}_{i,T} - A_i\|_\infty \le \mathcal{O}\left(\sqrt{\frac{\log\left[\lambda_{\max}(X_{i,T})\right]}{\lambda_{\min}(X_{i,T})}}\right), \quad \text{a.s.}$$

Remark 1. Theorem 1 is not a direct consequence of the decoupling procedure in the switched least squares method. The k least squares problems have a common covariate process $\{x_t\}_{t \ge 1}$. Therefore, the convergence of the switched least squares method and the stability of the MJS are interconnected problems. Our proof techniques carefully use the stability properties of the system to establish the consistency of the system identification method.

We simplify the result of Theorem 1 and characterize the data dependent result of Theorem 1 in terms of horizon T and the cardinality of the set $\mathcal{T}_{i,T}$.

Corollary 1. *Under Assumptions 1 and 2, for each $i \in \mathcal{S}$, we have:*

$$\|\hat{A}_{i,T} - A_i\|_\infty \le \mathcal{O}\left(\sqrt{\log(T)/|\mathcal{T}_{i,T}|}\right), \quad \text{a.s.}$$

Remark 2. The assumption that $\pi_\infty(i) \neq 0$ implies that for sufficiently large T, $|\mathcal{T}_{i,T}| \neq 0$ almost surely, therefore the expressions in above bounds are well defined.

The result of Corollary 1 still depends on data. When system identification results are used for adaptive control or reinforcement learning, it is useful to have a data-independent characterization of the rate of convergence. We present this characterization in the next theorem.

Theorem 2. *Under Assumptions 1 and 2, the rate of convergence of the switched least squares estimator $\hat{A}_{i,T}$, $i \in \mathcal{S}$ is upper bounded by:*

$$\|\hat{A}_{i,T} - A_i\|_\infty \le \mathcal{O}\big(\sqrt{\log(T)/\pi_\infty(i)T}\big), \quad \text{a.s.}$$

where the constants in the $\mathcal{O}(\cdot)$ notation do not depend on Markov chain $\{s_t\}_{t \ge 0}$ and horizon T. Therefore, the estimation process $\{\hat{\theta}_T\}_{T \ge 1}$ is strongly consistent, i.e., $\lim_{T \to \infty} \|\hat{\theta}_T - \theta\|_\infty = 0$ a.s. Furthermore, the rate of convergence is upper bounded by:

$$\|\hat{\theta}_T - \theta\|_\infty \le \mathcal{O}\big(\sqrt{\log(T)/\pi^* T}\big), \quad \text{a.s.}$$

where $\pi^* = \min_{j \in \mathcal{S}} \pi_\infty(j)$.

Theorem 2 shows that Assumptions 1 and 2 guarantee that the switched least squares estimator for MJS has the same rate of convergence of $\mathcal{O}(\sqrt{\log(T)/T})$ as non-switched case established in [16]. Moreover, the upper bound in Theorem 2

shows that the estimation error of $\hat{A}_{i,T}$ is proportional to $1/\sqrt{\pi_\infty(i)}$; therefore, the rate of convergence of $\hat{\theta}_T$ is proportional to $1/\sqrt{\pi^*}$, where π^* is the smallest probability in the stationary distribution π_∞.

Remark 3. SLS is a special case of MJS in which the discrete state evolves in an i.i.d. manner. The results presented in this section are valid for the SLS after substituting stationary distribution π_∞ with the i.i.d. PMF of switching probabilities defined over discrete state.

III. Proofs of the main results

A. Preliminary results

We first state the Strong Law of Large Numbers (SLLN) for Martingale Difference Sequences (MDS).

Theorem 3. *(see [38, Theorem 3.3.1]) Suppose $\{X_\tau\}_{\tau\geq 1}$is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_\tau\}_{\tau\geq 1}$. Let a_τ be $\mathcal{F}_{\tau-1}$ measurable for each $\tau \geq 1$ and we have $0 < a_\tau \to \infty$ as $\tau \to \infty$, a.s. If for some $p \in (0,2]$, we have: $\sum_{\tau=1}^{\infty} \mathbb{E}[|X_\tau|^p|\mathcal{F}_{\tau-1}]/a_\tau^p < \infty$, then: $\lim_{T\to\infty} \sum_{\tau=1}^{T} X_\tau/a_T = 0$ a.s.*

Lemma 1. *The assumptions on the process $\{s_t\}_{t\geq 0}$ imply that $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = \pi_\infty(i)$, a.s.*

Proof. $\{s_t\}_{t\geq 0}$ is an aperiodic and irreducible Markov chain, hence, by the Ergodic Theorem (Theorem 4.1, [39]), $\{s_t\}_{t\geq 0}$ is ergodic and therefore $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = \pi_\infty(i)$ a.s. □

Lemma 2. *Assumption 1 and 2 imply:*

$$\sum_{\tau=1}^{\infty} \|x_\tau\|^2/\tau^2 < \infty \quad a.s.$$

Proof. The result is a direct consequence of Abel's lemma. Let $S_T \coloneqq \sum_{\tau=1}^{T} \|x_\tau\|^2$, then we have:

$$\begin{aligned}
\sum_{\tau=1}^{T} \frac{\|x_\tau\|^2}{\tau^2} &= \sum_{\tau=1}^{T} \frac{S_\tau - S_{\tau-1}}{\tau^2} \\
&= \frac{S_T}{T^2} - \frac{S_0}{1} + \sum_{\tau=2}^{T} S_{\tau-1}\Big(\frac{1}{(\tau-1)^2} - \frac{1}{\tau^2}\Big) \\
&\overset{(a)}{=} \frac{S_T}{T^2} - \frac{S_0}{1} + \sum_{\tau=2}^{T} \mathcal{O}(\tau-1)\Big(\frac{2\tau-1}{\tau^2(\tau-1)^2}\Big) \\
&= \frac{S_T}{T^2} - \frac{S_0}{1} + \sum_{\tau=2}^{T} \mathcal{O}\Big(\frac{1}{\tau^2}\Big) < \infty
\end{aligned}$$

where (a) follows from Assumption 2. □

Lemma 3. *We have the following:*

$$\Big\| \sum_{\tau=1}^{T} A_{s_\tau} x_\tau w_\tau^\intercal + w_\tau x_\tau^\intercal A_{s_\tau}^\intercal \Big\| = o(T) \quad a.s.$$

Proof. We prove the limit element-wise. The (l,p)-th element of the matrix $A_{s_\tau} x_\tau w_\tau^\intercal$ is $\Big[\sum_{j=1}^{n} A_{s_\tau}(l,j)x_\tau(j)\Big]w_\tau(p)$. We calculate the term:

$$\mathbb{E}\Big[\Big(\sum_{j=1}^{n} A_{s_\tau}(l,j)x_\tau(j)w_\tau(p)\Big)^2 \Big| \mathcal{F}_{\tau-1}\Big]. \tag{5}$$

Let $A_* = \max_{i\in\mathcal{S}} \|A_i\|_\infty$, then

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{j=1}^{n} A_{s_\tau}(l,j)x_\tau(j)\Big)^2 w_\tau^2(p)\Big|\mathcal{F}_{\tau-1}\Big] \\
&\overset{(a)}{\leq} A_*^2 \sup_\tau \mathbb{E}[w_\tau^2(p)\big|\mathcal{F}_{\tau-1}]\Big(\sum_{j=1}^{n} x_\tau(j)\Big)^2 \\
&\overset{(b)}{\leq} nA_*^2 \sup_\tau \mathbb{E}\big[w_\tau^2(p)\big|\mathcal{F}_{\tau-1}\big]\|x_\tau\|^2,
\end{aligned}$$

where (a) uses the fact that s_τ and x_τ are $\mathcal{F}_{\tau-1}$ measurable and that $|A_{s_\tau}(l,j)| \leq A_*$ and (b) is by Cauchy-Schwarz's inequality. Therefore:

$$\begin{aligned}
&\sum_{\tau=1}^{T} \frac{\mathbb{E}\Big[\Big(\big[\sum_{j=1}^{n} A_{s_\tau}(l,j)x_\tau(j)\big]w_\tau(p)\Big)^2\Big|\mathcal{F}_{\tau-1}\Big]}{\tau^2} \\
&\qquad \leq nA_*^2 \sup_\tau \Big\{\mathbb{E}[w_\tau^2(p)|\mathcal{F}_{\tau-1}]\Big\} \sum_{\tau=1}^{T} \frac{\|x_\tau\|^2}{\tau^2} \overset{(c)}{\leq} \infty.
\end{aligned}$$

Since $\alpha > 2$ in Assumption 1, and finiteness of higher order moments imply finiteness of lower order moments, we get $\mathbb{E}\big[w_\tau^2(p)\big|\mathcal{F}_{\tau-1}\big]$ is uniformly bounded. This fact along with Lemma 2 imply (c). The result then follows by applying Theorem 3 by setting $a_t = t$ and $p = 2$. □

We characterize the asymptotic behavior of the matrix $X_{i,T}$.

Proposition 1. *Under Assumptions 1 and 2, the following hold a.s. for each $i \in \mathcal{S}$:*
(P1) $\lambda_{\max}(X_{i,T}) = \mathcal{O}(T)$, *a.s.*
(P2) $\liminf_{T\to\infty} \lambda_{\min}(X_{i,T})/|\mathcal{T}_{i,T}| > 0$, *a.s.*

Remark 4. Property (P1) shows that when the system is stable in the average sense, $\lambda_{\max}(X_{i,T})$ cannot grow faster than linearly with time. Therefore, the stability of the system controls the rate at which $X_{i,T}$ can grow. Property (P2) shows that when the noise has a minimum covariance, $\lambda_{\min}(X_{i,T})$ cannot grow slower than linearly with time.

Proof of (P1). The maximum eigenvalue of a matrix can be upper bounded as follows:

$$\begin{aligned}
\lambda_{\max}\Big(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\intercal\Big) &\overset{(a)}{\leq} \mathrm{tr}\Big(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\intercal\Big) = \sum_{t\in\mathcal{T}_{i,T}} \|x_t\|^2 \\
&\leq \sum_{t=1}^{T} \|x_t\|^2 = \mathcal{O}(T)
\end{aligned}$$

where (a) follows from the fact that trace of a matrix is sum of its eigenvalues and all eigenvalues of $x_t x_t^\intercal$ are non-negative. □

Proof of (P2). For $\tau \geq 1$, we have:

$$\begin{aligned}
x_\tau x_\tau^\intercal =& (A_{s_{\tau-1}}x_{\tau-1} + w_{\tau-1})(A_{s_{\tau-1}}x_{\tau-1} + w_{\tau-1})^\intercal \\
=& A_{s_{\tau-1}}x_{\tau-1}x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal \\
&+ A_{s_{\tau-1}}x_{\tau-1}w_{\tau-1}^\intercal + w_{\tau-1}x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal + w_{\tau-1}w_{\tau-1}^\intercal.
\end{aligned}$$

Since $A_{s_{\tau-1}} x_{\tau-1} x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal$ is positive semi-definite, we have:

$$x_\tau x_\tau^\intercal \succeq A_{s_{\tau-1}} x_{\tau-1} w_{\tau-1}^\intercal + w_{\tau-1} x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal + w_{\tau-1} w_{\tau-1}^\intercal.$$

By summing over $\tau \in \mathcal{T}_{i,T}$, we get:

$$\begin{aligned}\sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal &\succeq \sum_{\tau\in\mathcal{T}_{i,T}} w_{\tau-1} w_{\tau-1}^\intercal + x_0 x_0^\intercal \mathbb{1}\{s_0 = i\} \\ &\quad + \sum_{\tau\in\mathcal{T}_{i,T}} \left[A_{s_{\tau-1}} x_{\tau-1} w_{\tau-1}^\intercal + w_{\tau-1} x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal\right] \\ &\overset{(a)}{\succeq} \sum_{\tau\in\mathcal{T}_{i,T}} w_{\tau-1} w_{\tau-1}^\intercal + o(T) \quad \text{a.s.}\end{aligned}$$

where (a) follows from Lemma 3 and $x_0 x_0^\intercal \mathbb{1}\{s_0 = i\} \succeq 0$. Furthermore, since $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = \pi_\infty(i)$ a.s. by Lemma 1 and $\pi_\infty(i) \neq 0$ by assumptions on $\{s_\tau\}_{\tau\geq 0}$, we have:

$$\begin{aligned}&\liminf_{|\mathcal{T}_{i,T}|\to\infty} \frac{\sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal}{|\mathcal{T}_{i,T}|} \\ &\qquad \succeq \liminf_{|\mathcal{T}_{i,T}|\to\infty} \frac{\sum_{\tau\in\mathcal{T}_{i,T}} w_{\tau-1} w_{\tau-1}^\intercal}{|\mathcal{T}_{i,T}|} \overset{(b)}{=} C \succ 0 \quad \text{a.s.}\end{aligned}$$

where (b) holds by Assumption 1 and independence of $\{w_\tau\}_{\tau\geq 0}$ and $\{s_\tau\}_{\tau\geq 0}$ processes. Therefore

$$\liminf_{|\mathcal{T}_{i,T}|\to\infty} \lambda_{\min}\Big(\frac{\sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal}{|\mathcal{T}_{i,T}|}\Big) \succ 0.$$

□

B. Background on least square estimator

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

$$y_t = \beta^\intercal z_t + w_t, \quad t \geq 0, \tag{6}$$

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

$$\hat{\beta}_T = \arg\min_{\beta^\intercal} \sum_{\tau=0}^{T} \|y_\tau - \beta^\intercal z_\tau\|^2. \tag{7}$$

The following result by [33] characterizes the rate of convergence of $\hat{\beta}_T$ to β in terms of unnormalized covariance matrix of covariates $Z_T \coloneqq \sum_{\tau=0}^{T} z_\tau z_\tau^\intercal$.

Theorem 4 (see [33, Theorem 1])**.** *Suppose the following conditions are satisfied:* **(S1)** $\lambda_{\min}(Z_T) \to \infty$*, a.s. and* **(S2)** $\log(\lambda_{\max}(Z_T)) = o(\lambda_{\min}(Z_T))$*, a.s. Then the least squares estimate in* (7) *is strongly consistent with the rate of convergence:*

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

Theorem 4 is valid for all the $\mathcal{G}_{t-1}$-measurable covariate processes $\{z_t\}_{t\geq 0}$. For the switched least squares system identification if we take $\mathcal{G}_t$ to be equal to $\mathcal{F}_t$ and verify conditions (S1) and (S2) in Theorem 4, then we can use Theorem 4 to establish its strong consistency and rate of convergence. As mentioned earlier in Remark 1, the empirical covariances are coupled across different components due to the system dynamics.

C. Proof of Theorem 1

To prove this theorem, we check the sufficient conditions in Theorem 4. First requirement that $X_{i,T}$ is measurable w.r.t. $\mathcal{F}_{T-1}$, follows by the definition of $X_{i,T}$. Conditions (S1) and (S2) are verified in the following.

(S1) By Proposition 1-(P2), we see that $\lambda_{\min}(X_{i,T}) \to \infty$ a.s.; therefore, (S1) in Theorem 4 is satisfied.

(S2) Proposition 1-(P1) and (P2) imply that there exist positive constants C_1, C_2, such that :

$$\begin{aligned}&\limsup_{T\to\infty} \frac{\log(\lambda_{\max}(X_{i,T}))}{\lambda_{\min}(X_{i,T})} \\ &\qquad \leq \limsup_{T\to\infty} \frac{\log(C_1) + \log(T)}{C_2 |\mathcal{T}_{i,T}|} = 0 \quad \text{a.s.}\end{aligned}$$

where the last equality follows by Lemma 1 (i.e. $|\mathcal{T}_{i,T}| = \mathcal{O}(T)$, a.s.). Therefore, the second condition of Theorem 4 is satisfied.

Therefore, by Theorem 4, for each $i \in \mathcal{S}$, we have:

$$\big\|\hat{A}_{i,T} - A_i\big\|_\infty \leq \mathcal{O}\Bigg(\sqrt{\frac{\log\left[\lambda_{\max}(X_{i,T})\right]}{\lambda_{\min}(X_{i,T})}}\Bigg), \quad \text{a.s.} \tag{8}$$

which proves the claim in Theorem 1.

D. Proof of Corollary 1

Corollary 1 is the direct consequence of Theorem 1 and Proposition 1. Proposition 1-(P1) implies that $\lambda_{\max}(X_{i,T}) = \mathcal{O}(\log(T))$. By substituting $\lambda_{\max}(X_{i,T})$ with $\mathcal{O}(\log(T))$ in the right hand side of Eq. (8), we get that for each $i \in \mathcal{S}$, the estimation error $\|\hat{A}_{i,T} - A_i\|_\infty$ is upper-bounded by $\mathcal{O}\big(\sqrt{\log(T)/|\mathcal{T}_{i,T}|}\big)$, *a.s.*

E. Proof of Theorem 2

We first establish the strong consistency of the parameter $\hat{\theta}_T$. By Theorem 1 and the fact that $k < \infty$, we get:

$$\big\|\hat{\theta}_T - \theta\big\|_\infty \leq \max_{i\in\mathcal{S}} \mathcal{O}\Bigg(\sqrt{\frac{\log\left[\lambda_{\max}(X_{i,T})\right]}{\lambda_{\min}(X_{i,T})}}\Bigg), \quad \text{a.s.}$$

Therefore, the result follows by applying Theorem 1 to the argmax of above equation. For the second part notice that by Lemma 1, we know $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = \pi_\infty(i)$, a.s. Now, by Corollary 1, we get:

$$\big\|\hat{A}_{i,T} - A_i\big\|_\infty \leq \mathcal{O}\Bigg(\sqrt{\frac{\log(T)}{|\mathcal{T}_{i,T}|}}\Bigg) = \mathcal{O}\Bigg(\sqrt{\frac{\log T}{\pi_\infty(i) T}}\Bigg) \quad \text{a.s.}$$

which is the claim of Theorem 2.

IV. Discussion on Stability in the Average Sense

The main results of this paper are derived under Assumption 2 i.e., the MJS system (1) is stable in the average sense. In this section, we discuss the connection between this notion of stability and more common forms of stability, i.e., mean square stability and almost sure stability.

A. Stability on the average sense and mean square stability

A common assumption on the stability of MJS systems (e.g., [31] and [40]) is mean square stability defined as following:

Definition 3. *The MJS system* (1) *is called mean square stable (MSS) if there exists a deterministic vector* $x_\infty \in \mathbb{R}^n$ *and a deterministic positive definite matrix* $Q_\infty \in \mathbb{R}^{n\times n}$ *such that for any deterministic initial state* x_0 *and* s_0 *, we have:* $\lim_{\tau\to\infty}\left\|\mathbb{E}[x_\tau]-x_\infty\right\| \to 0$*, and* $\lim_{\tau\to\infty}\left\|\mathbb{E}[x_\tau x_\tau^\intercal]-Q_\infty\right\| \to 0$.

Proposition 2 (see [7, Theorem 3.9])**.** *The system is MSS, if and only if* $\lambda_{\max}\big((P^\intercal \otimes I_{n^2})\,\mathrm{diag}(A_i \otimes A_i)\big) < 1$.

We now show that stability in the average sense is a weaker notion of stability than MSS.

Proposition 3. *If the MJS system* (1) *is mean square stable, then the system is stable in the average sense.*

The proof if presented in Appendix A.

Remark 5. Proposition 3 shows that MSS implies Assumption 2. Therefore, the results of Theorem 1 and 2 also hold when Assumption 2 is replaced by the assumption that the system is MSS.

B. Stability on the average sense and almost sure stability

Consider the noise free version of the MJS system (1) with the following dynamics:

$$x_{t+1} = A_{s_t} x_t, \quad t \geq 0. \tag{9}$$

Definition 4. *The system* (9) *is called almost surely stable if, for any deterministic initial state* x_0 *and* s_0 *we have:*

$$\lim_{t\to\infty} \|x_t\| = 0, \quad a.s.$$

A common sufficient condition to check the almost sure stability of MJS system (9) *is given below.*

Proposition 4 (see [7, Theorem 3.47])**.** *If the stationary distribution* $\pi_\infty = (\pi_\infty(1), \ldots, \pi_\infty(k))$ *satisfies* **(C1)** $\pi_\infty(i) \neq 0$ *for all* i *and* **(C2)** $\prod_{i=1}^k \sigma_{\max}(A_i)^{\pi_\infty(i)} < 1$*, then the system* (9) *is almost surely stable.*

We now show that (C1) and (C2) are also sufficient conditions for stability in the average sense.

Proposition 5. *If the MJS system* (1) *satisfies (C1) and (C2), then the system is stable in the average sense.*

Proof is presented in Appendix B.

Remark 6. Proposition 5 shows that (C1) and (C2) imply Assumption 2. Therefore, the results of Theorem 1 and 2 also hold when Assumption 2 is replaced by the assumption that the system satisfies (C1) and (C2).

C. Discussion on Non-Comparable Stability Assumption

The following examples illustrate that neither MSS nor conditions (C1) and (C2) in Proposition 5 is stronger than the other.

Example 1. Let $\theta^\intercal = \{A_1, 0\}$, and $p = (p_1, p_2)$ is an i.i.d. probability transition, with $\lambda_{\max}(p_1 A_1) > 1$ and $x_0 \neq 0$. Then $\mathbb{E}[x_{\tau+1}] = \mathbb{E}[A_{\sigma_\tau} x_\tau + w_t] = p_1 A_1 \mathbb{E}[x_\tau] = \cdots = (p_1 A_1)^\tau \mathbb{E}(x_0)$, which implies $\lim_{\tau\to\infty} \mathbb{E}(x_\tau) = \infty$. Therefore, this system is not mean square stable. However, this system satisfies conditions (C1) and (C2) in Prop. 5 and therefore is stable in the average sense.

Example 2. Consider non-switched system with matrix A, with $\lambda_{\max}(A) < 1$ and $\sigma_{\max}(A) > 1$. This system is mean square stable, but it doesn't satisfy the conditions (C1) and (C2) in Proposition 5.

V. Numerical Simulation

In this section, we illustrate the result of Theorem 1 via an example. Consider a MJS with $n = 2, k = 2$, $A_1 = \left[\begin{smallmatrix} 1.5 & 0 \\ 0 & 0.2 \end{smallmatrix}\right]$, and $A_2 = \left[\begin{smallmatrix} 0.01 & 0.1 \\ 0.1 & 0.1 \end{smallmatrix}\right]$, probability transition matrix $P = \left[\begin{smallmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{smallmatrix}\right]$ and i.i.d. $\{w_t\}_{t\geq 0}$ with $w_t \sim \mathcal{N}(0, I)$. Note that the example satisfies Assumptions 1 and conditions (C1) and (C2) of Proposition 5 (and, therefore, Assumption 2), but it is not mean square stable. We run the switched least squares for a horizon of $T = 10^6$ and repeat the experiment for 100 independent runs. We plot the estimation error $e_{i,T} = \|\hat{A}_{i,t} - A_1\|_\infty$ versus time in Fig. 1. The plot shows that the estimation error is converging almost surely even though the system is not mean square stable. In Fig. 2, logarithm of the estimation error versus logarithm of the horizon is plotted. The linearity of the graph along with approximate slope of -0.5 shows that $e_{i,T} = \tilde{O}(1/\sqrt{T})$.

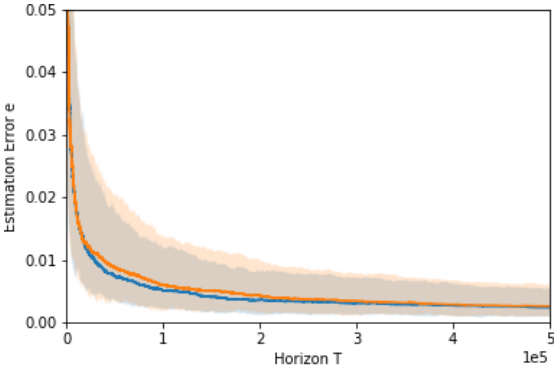


Fig. 1. Performance of switched least squares method for the example of Sec. V. The solid line shows the mean across 100 runs and the shaded region shows the 25% to 75% quantile bound.

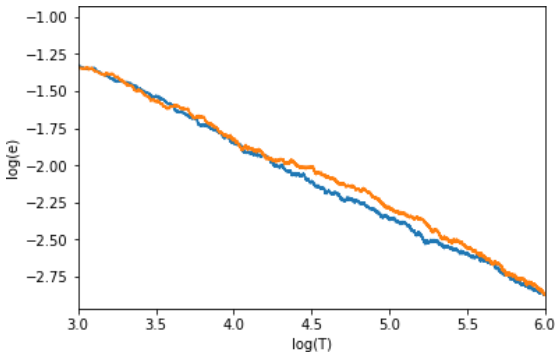


Fig. 2. Logarithm of the estimation error versus logarithm of the horizon is plotted.

VI. Conclusion and Future Directions

In this paper, we investigated system identification of (autonomous) Markov jump linear systems. We proposed the switched least squares method, showed it is strongly consistent and derived the almost sure rate of convergence of $\mathcal{O}(\sqrt{\log(T)/T})$. This analysis provides a solid first step toward establishing almost sure regret bounds for adaptive control of MJS.

We derived our results assuming that system is stable in the average sense and we showed that this is a weaker assumption compared to mean square stability.

The current results are established for autonomous systems with Markov switching when the complete state of the system is observed. Interesting future research directions include relaxing these modeling assumptions and considering controlled systems under partial state observability and unobserved jump times.

References

[1] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Consistency and rate of convergence of switched least squares system identification for autonomous switched linear systems," in *2022 IEEE 61st Decis Contr P (CDC)*. IEEE, 2022.

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

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

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

[5] Y. Fang, K. A. Loparo, and X. Feng, "Almost sure and δ-moment stability of jump linear systems," *Int J Control*, vol. 59, no. 5, pp. 1281–1307, 1994.

[6] Y. Fang, "A new general sufficient condition for almost sure stability of jump linear systems," *IEEE Trans. Autom. Control*, vol. 42, no. 3, pp. 378–382, 1997.

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

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

[9] G. Goodwin, P. Ramadge, and P. Caines, "Discrete-time multivariable adaptive control," *IEEE Trans. Autom. Control*, vol. 25, no. 3, pp. 449–456, 1980.

[10] L. Ljung, *System identification*. Springer, 1998.

[11] J. Rissanen and P. Caines, "The strong consistency of maximum likelihood estimators for ARMA processes," *Ann. Statist.*, pp. 297–315, 1979.

[12] L. Ljung, "On the consistency of prediction error identification methods," in *Math. Sci. Eng.* Elsevier, 1976, vol. 126, pp. 121–164.

[13] P. E. Caines and L. Ljung, "Prediction error estimators: Asymptotic normality and accuracy," in *1976 IEEE Decis Contr P*. IEEE, 1976, pp. 652–658.

[14] B. Ho and R. E. Kálmán, "Effective construction of linear state-variable models from input/output functions," *at-Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.

[15] A. Lindquist and G. Picci, "State space models for Gaussian stochastic processes," in *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J. C. Willems, Eds. Dordrecht: Springer Netherlands, 1981, pp. 169–204.

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

[17] P. E. Caines, *Linear stochastic systems*. SIAM, 2018.

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

[19] ——, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

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

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

[22] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*. PMLR, 2018, pp. 439–473.

[23] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *2019 P Amer Contr Conf (ACC)*. IEEE, 2019, pp. 5655–5661.

[24] Y. Zheng, L. Furieri, M. Kamgarpour, and N. Li, "Sample complexity of linear quadratic Gaussian (LQG) control for output feedback systems," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 559–570.

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

[26] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite sample perspective," *arXiv preprint arXiv:2209.05423*, 2022.

[27] P. E. Caines and H.-F. Chen, "Optimal adaptive LQG control for systems with finite state process parameters," *IEEE Trans. Autom. Control*, vol. 30, no. 2, pp. 185–189, 1985.

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

[29] F. Xue and L. Guo, "Necessary and sufficient conditions for adaptive stablizability of jump linear systems," *Commun. Inf. Syst.*, vol. 1, no. 2, pp. 205–224, 2001.

[30] S. Shi, O. Mazhar, and B. De Schutter, "Finite-sample analysis of identification of switched linear systems with arbitrary or restricted switching," *IEEE Control Systems Letters*, vol. 7, pp. 121–126, 2023.

[31] T. Sarkar, A. Rakhlin, and M. Dahleh, "Nonparametric system identification of stochastic switched linear systems," in *2019 IEEE 58th Decis Contr P (CDC)*. IEEE, 2019, pp. 3623–3628.

[32] 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.

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

[34] H.-F. Chen and L. Guo, "Convergence rate of least-squares identification and adaptive control for stochastic systems," *Int J Control*, vol. 44, no. 5, pp. 1459–1476, 1986.

[35] ——, "Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost," *SIAM J. Control Optim.*, vol. 25, no. 4, pp. 845–867, 1987.

[36] T. E. Duncan and B. Pasik-Duncan, "Adaptive control of continuous-time linear stochastic systems," *Math. Control signals systems*, vol. 3, no. 1, pp. 45–60, 1990.

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

[38] W. F. Stout, *Almost Sure Convergence*. Academic Press, 1974.

[39] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer Science & Business Media, 2013, vol. 31.

[40] Y. Sattar and S. Oymak, "Non-asymptotic and accurate learning of nonlinear dynamical systems," *arXiv preprint arXiv:2002.08538*, 2020.

APPENDIX A
PROOF OF PROPOSITION 3

Proof. Since the system is MSS, there exists a positive definite matrix $Q_\infty \in \mathbb{R}^{n\times n}$ such that $\lim_{\tau\to\infty}\mathbb{E}[x_\tau x_\tau^\intercal] = Q_\infty$, which implies $\lim_{\tau\to\infty}\mathrm{Tr}(\mathbb{E}[x_\tau x_\tau^\intercal]) = \mathrm{Tr}(Q_\infty)$. Since $\mathrm{Tr}(\mathbb{E}[xx^\intercal]) = \mathbb{E}[\mathrm{Tr}(xx^\intercal)] = \mathbb{E}[x^\intercal x]$, MSS implies that sequence of real numbers $\{\mathbb{E}(\|x_\tau\|^2)\}_{\tau\geq 0}$ converges to $\mathrm{Tr}(Q_\infty)$ and therefore:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=1}^{T}\mathbb{E}(\|x_\tau\|^2) = \mathrm{Tr}(Q_\infty) < \infty \tag{10}$$

Define events

$$E_n = \Big\{\omega\in\Omega : \limsup_{T\to\infty}\frac{1}{T}\sum_{\tau=1}^{T}\|x_\tau\|^2 \leq n\Big\}, \quad \forall n\in\mathbb{N}$$

and

$$E = \bigcup_{n=0}^{\infty} E_n = \Big\{\omega\in\Omega : \limsup_{T\to\infty}\frac{1}{T}\sum_{\tau=1}^{T}\|x_\tau\|^2 < \infty\Big\}.$$

Now, by the continuity of probability measure from below, we have:

$$\mathbb{P}(E) = \mathbb{P}(\bigcup_{n=0}^{\infty} E_n) = \lim_{n\to\infty}\mathbb{P}(E_n). \tag{11}$$

Note that

$$\begin{aligned}\mathbb{P}(E_n) &= \mathbb{P}\big(\limsup_{T\to\infty}\frac{1}{T}\sum_{\tau=1}^{T}\|x_\tau\|^2 \leq n\big)\\ &\overset{(a)}{\geq} \limsup_{T\to\infty}\mathbb{P}\big(\frac{1}{T}\sum_{\tau=1}^{T}\|x_\tau\|^2 \leq n\big)\\ &\overset{(b)}{\geq} 1 - \limsup_{T\to\infty}\frac{\big(\sum_{\tau=1}^{T}\mathbb{E}\|x_\tau\|^2\big)}{Tn}\\ &\overset{(c)}{\geq} 1 - \frac{\mathrm{Tr}(Q_\infty)}{n},\end{aligned}$$

where (a) follows from reverse Fatou's lemma, (b) follows from the Markov inequality and (c) follows from Eq. (10). Substituting the above in equation (10), we get

$$\mathbb{P}(E) \geq \lim_{n\to\infty}\Big(1 - \frac{\mathrm{Tr}(Q_\infty)}{n}\Big) = 1.$$

Therefore $\mathbb{P}(E) = 1$, and the system is stable in the average sense. □

APPENDIX B
PROOF OF PROPOSITION 5

A. Asymptotic Behavior of Continuous Component

To simplify the notation, we assume that $x_0 = 0$ which does not entail any loss of generality. Let $\Phi(t-1,\tau+1) = A_{s_{t-1}}\cdots A_{s_{\tau+1}}$ denote the state transition matrix where we follow the convention that $\Phi(t,\tau) = I$, for $t < \tau$. Then we can write the dynamics in Eq. (1) of the continuous component of the state in convolutional form as:

$$x_t = \sum_{\tau=0}^{t-1}\Phi(t-1,\tau+1)w_\tau. \tag{12}$$

where $\|\Phi(t-1,\tau+1)\| = \|A_{s_{t-1}}\dots A_{s_{\tau+1}}\|$, and

$$\|A_{s_{t-1}}\dots A_{s_{\tau+1}}\| \leq \sigma_{s_{t-1}}\cdots\sigma_{s_{\tau+1}} =: \Gamma_{t-1,\tau+1} \tag{13}$$

where $\sigma_{s_t} = \sigma_{\max}(A_{s_t})$. In the following lemma, it is established that the conditions (C1) and (C2) in Prop. 5 imply that the sum of norms of the state-transition matrices are uniformly bounded.

Lemma 4 (see [1, Lemma 1])**.** *Under the conditions (C1) and (C2) in Prop. 5, there exists a constant $\bar\Gamma < \infty$ such that for all $T > 1$, $\sum_{\tau=0}^{T-1}\|\Phi(T-1,\tau+1)\| \leq \bar\Gamma$, a.s.*

The following Lemma shows the implication of Assumption 1 on the growth rate of energy of the noise process.

Lemma 5 ([16, Eq. (3.1)])**.** *Under Assumption 1 $\sum_{\tau=0}^{T}\|w_\tau\|^2 = \mathcal{O}(T)$, a.s.*

Using the convolution formula in Eq. (12), we can bound the norm of the state $\|x_t\|^2$ as following:

$$\begin{aligned}\|x_t\|^2 &= \Big(\big\|\sum_{\tau=0}^{t-1}\Phi(t-1,\tau+1)w(\tau)\big\|\Big)^2\\ &\overset{(a)}{\leq}\Big(\sum_{\tau=0}^{t-1}\|\Phi(t-1,\tau+1)w(\tau)\|\Big)^2\\ &\overset{(b)}{\leq}\Big(\sum_{\tau=0}^{t-1}\|\Phi(t-1,\tau+1)\|\|w(\tau)\|\Big)^2\\ &\overset{(c)}{\leq}\Big(\sum_{\tau=0}^{t-1}\Gamma_{t,\tau+1}\|w(\tau)\|\Big)^2\end{aligned} \tag{14}$$

where (a) follows from triangle inequality and (b) follow from sub-multiplicative property of the matrix norm, and (c) follows from Eq. (13). Now for a fixed i, $i\in\mathcal{S}$, we have:

$$\begin{aligned}\sum_{t=0}^{T}\|x_t\|^2 &\leq \sum_{t=0}^{T}\Big(\sum_{j=0}^{t-1}\Gamma_{j+1,t-1}\|w(j)\|\Big)^2\\ &\overset{(d)}{\leq}\sum_{t=0}^{T}\Big(\sum_{j=0}^{t-1}\Gamma_{j+1,t-1}\Big)\Big(\sum_{j=0}^{t-1}\Gamma_{j+1,t-1}\|w(j)\|^2\Big)\\ &\overset{(e)}{\leq}\bar\Gamma\sum_{t=0}^{T}\Big(\sum_{j=0}^{t-1}\Gamma_{j+1,t-1}\|w(j)\|^2\Big)\\ &\overset{(f)}{\leq}\bar\Gamma\sum_{j=0}^{T-1}\Big(\sum_{t=0}^{T}\Gamma_{j+1,t-1}\Big)\|w(j)\|^2\\ &\overset{(g)}{\leq}\bar\Gamma^2\sum_{j=0}^{T-1}\|w(j)\|^2 = \mathcal{O}(T) \quad \text{a.s.}\end{aligned}$$

where (d) follows from Cauchy-Schwarz's inequality, (e) follows from Lemma 4, (f) follows from changing the order of summation, and (g) follows from boundedness of sub-sums of $\sum_{\tau=0}^{T-1}\Gamma_{\tau+1,T-1}$, and Lemma 4.