Proceedings of Machine Learning Research vol 144:1–17, 2022

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

Borna Sayedana BORNA.SAYEDANA@MAIL.MCGILL.CA
Electrical and Computer Engineering, McGill University, Canada

Mohammad Afshari MAFSHARI@UALBERTA.CA
Computer Science, University of Alberta, Canada

Peter E. Caines PETER.CAINES@MCGILL.CA
Electrical and Computer Engineering, McGill University, Canada

Aditya Mahajan ADITYA.MAHAJAN@MCGILL.CA
Electrical and Computer Engineering, McGill University, Canada

Abstract

In this paper, we investigate the problem of system identification for autonomous switched linear systems with complete state observations. We propose switched least squares method for the identification for switched linear systems, show that this method is strongly consistent, and derive data-dependent and data-independent rates of convergence. In particular, our data-dependent rate of convergence shows that, almost surely, the system identification error is $\mathcal{O}\big(\sqrt{\log(T)/T}\big)$ where T is the time horizon. These results show that our method for switched linear systems has the same rate of convergence as least squares method for non-switched linear systems. We compare our results with those in the literature. We present numerical examples to illustrate the performance of the proposed system identification method.

Keywords: Stochastic Switched Linear Systems, System Identification, Least Squares Method, Strong Consistency

1. Introduction

Switched Linear Systems Switched linear systems (SLS) are a good approximation of non-linear time-varying systems arising in various applications including networked control systems (Deaecto et al., 2014) and cyber-physical systems (De Persis and Tesi, 2015; Cetinkaya et al., 2018). There is a rich literature on the stability analysis (e.g., Fang et al. (1994); Fang (1997); Costa et al. (2006)) and optimal control (e.g., Chizeck et al. (1986)) of SLS. However, most of the literature assumes that the system model is known. The problem of system identification, i.e., identifying the dynamics from data, has not received much attention.

System Identification 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 (Goodwin et al., 1980). A commonly used method for system identification is the least squares method. Lai and Wei (1982) provide asymptotic rates of convergence and establish strong consistency of least squares method for regression. These results have been extended to autonomous linear systems by Lai and Wei (1985) and ARMAX systems in Chen and Guo (1986, 1987); Lai and Wei (1986). See Chapter 6 of Caines (2018) for a unified overview.

These classical results provide asymptotic convergence guarantees. In recent years, there has been a significant interest in the machine learning community to establish finite-time convergence guarantees for system identification under a variety of assumptions (Faradonbeh et al., 2020b, 2018; Abbasi-Yadkori and Szepesvári, 2011; Faradonbeh et al., 2020a; Simchowitz et al., 2018; Oymak and Ozay, 2019; Zheng et al., 2021; Lale et al., 2020).

System Identification for Switched Linear Systems System identification of SLS has received less attention in the literature. There is some work on designing asymptotically stable controllers for unknown SLS (Caines and Chen, 1985; Caines and Zhang, 1995; Xue and Guo, 2001) but these papers do not established rates of convergence for system identification. The problem of identification of SLSs using set membership identification has been investigated in Ozay et al. (2015); Hespanhol and Aswani (2020). There are few recent results which establish high probability rates of converge for different models of SLS for subspace methods (Sarkar et al., 2019) and least-square methods (Sattar et al., 2021). We provide a detailed comparison with these papers in Sec. 5.

Contribution We investigate the problem of identifying an unknown (autonomous) SLS with i.i.d. switching. We propose a *switched* least squares method for system identification and provide data-dependent and data-independent rates of convergence for this method. Using these bounds, we establish *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 Lai and Wei (1985). In contrast to the existing high-probability convergence guarantees in the literature, Our results show that the estimation error converges to zero *almost surely*. To the best of our knowledge, this is the first result in the literature which establishes strong consistency and almost sure rates of convergence for SLS.

Organization The rest of the paper is organized as follows. In Sec 2, we state the system model, assumptions, and the main results. In Sec. 3, we sketch the proof of results. We present an illustrative example in Sec. 4 and compare our assumptions and results with the existing literature in Sec. 5. Finally, we conclude in Sec. 6.

Notation Given a matrix A, $A(i,j)$ denotes its (i,j)-th element, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest magnitudes of right eigenvalues, $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\intercal A)}$ denotes the spectral norm. For a square matrix Q, $\operatorname{tr}(Q)$ denotes the trace. When Q is symmetric, $Q \succeq 0$ and $Q \succ 0$ 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 a sequence of positive numbers $\{a_t\}_{t\geq 0}$, $a_T = \mathcal{O}(T)$ denotes $\limsup_{T\to\infty} a_T/T < \infty$, and $a_T = o(T)$ denotes $\limsup_{T\to\infty} a_T/T = 0$. Given a sequence of vectors $\{x_t\}_{t\in\mathcal{T}}$, $\operatorname{vec}(x_t)_{t\in\mathcal{T}}$ 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 $x_{0:t}$.

$\mathbb{R}$ and $\mathbb{N}$ denote the set 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. Convergence in almost sure sense is abbreviated as $a.s.$

2. System model and problem formulation

System model Consider a discrete-time (autonomous) switched linear system. The state of the system has two components: a discrete component $s_t \in \{1, \dots, k\}$ and a continuous component $x_t \in \mathbb{R}^n$. There is a finite set $\mathcal{A} = \{A_1, \dots, A_k\}$ 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 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 is distributed in an independent and identically distributed manner with $\mathbb{P}(s_t = i) = p_i \neq 0$, where $p = (p_1, \dots, p_k)$ is a probability mass function.

Assumptions on the model Let $\mathcal{F}_{t-1} = \sigma(x_{0:t}, s_{0:t})$ denote the sigma-algebra generated by the history of the complete state. Furthermore, let σ_i denote the maximum singular value of A_i, $i \in \{1, \dots, k\}$. It is assumed that the model satisfies the following assumptions:

Assumption 1 *The noise process $\{w_t\}_{t \geq 0}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{\geq 0}$, 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 $\liminf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} w_t w_t^\intercal = C$.*

Assumption 2 *The switching probabilities $p = (p_1, \dots, p_k)$ are such that $\prod_{i=1}^k \sigma_i^{p_i} < 1$.*

Assumption 1 is a standard assumption in the asymptotic analysis of system identification of linear systems (Caines, 2018; Lai and Wei, 1982, 1985; Chen and Guo, 1986, 1987) and allows the noise process to be non-stationary and have heavy tails (as long as moment condition is satisfied). Assumption 2 is a standard assumption for almost sure exponential statbility of *noise-free* switched linear system i.e., when $w_t = 0$ (Fang et al., 1994). Some of the recent results on system identification of Markov jump linear systems impose slightly different assumptions and we compare with those in Sec. 5.

System identification and switched least squares estimates We are interested in the setting where the system dynamics $\mathcal{A}$ and the switching probabilities 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:

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

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 $\mathcal{T}_{i,T} = \{t \leq T \mid s_t = i\}$ 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

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

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

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

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

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

Definition 1 *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 proofs are presented in Sec. 3.

Theorem 2 *Under Assumptions 1 and 2, the switched least squares estimates $\{\hat{A}_{i,T}\}_{i=1}^k$ are strongly consistent, i.e., for each $i \in \{1, \dots, k\}$, we have: $\lim_{T\to\infty} \left\|\hat{A}_{i,T} - A_i\right\|_\infty = 0$, a.s.*

Furthermore, the rate of convergence is upper bounded by the following expression:

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

Remark 3 *Notice that Theorem 2 is not a direct consequence of the decoupling procedure in switched least squares method. The k least squares problems have a common covariate process. Therefore, the convergence of the switched least squares method and the stability of the switched linear systems are interconnected problems. Our proof techniques leverage this connection to establish the consistency of the system identification method.*

We simplify the result in Theorem 2 and characterize the data dependent result found in Theorem 2 in terms of horizon T and the cardinality of the set $\mathcal{T}_{i,T}$.

Corollary 4 *Under Assumptions 1 and 2, for each $i \in \{1, \dots, k\}$, we have:*

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

Remark 5 *The assumption that $p_i \neq 0$ 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 4 still depends on the data. When system identification results are used for adaptive control or reinforcement learning, it is useful have a data-independent characterization of the rate of convergence. We present this characterization in the next theorem.

Theorem 6 *Under Assumptions 1 and 2, the rate of convergence of the switched least squares estimator* $\hat{A}_{i,T}$ *is upper bounded by:*

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

where the constants in the $\mathcal{O}(\cdot)$ *notation do not depend on* p *and* T*. Therefore, the estimation process is strongly consistent, i.e.,* $\lim_{T\to\infty}\left\|\hat{\theta}_T - \theta\right\|_\infty = 0$ *a.s. with the convergence rate given by:*

$$\left\|\hat{\theta}_T - \theta\right\|_\infty \leq \mathcal{O}\sqrt{\log(T)/p^* T}, \quad \textit{a.s.}$$

where $p^* = \min_j p_j$*.*

Theorem 6 shows that Assumptions 1 and 2 guarantee that the switched least squares method for SLS has the same rate of convergence of $\mathcal{O}(\sqrt{\log(T)/T})$ as non-switched case established in Lai and Wei (1985). Moreover, the constants show that the estimation error of i-th least squares problem is proportional to $1/\sqrt{p_i}$; therefore, the rate of convergence of $\hat{\theta}_t$ is proportional to $1/\sqrt{p^*}$, where p^* is the smallest probability of switching in PMF p.

3. Proof of the main results

In this section, we present the proof of Theorems 2 and 6 and Corollary 4. In Section 3.2, we review the background on the rate of convergence for least squares regression. In Section 3.2, we characterize the asymptotic behaviors of continuous state of the system and covariates of the i-th least squares problem. Only proof sketches are presented, see the appendices for complete proofs. The proof the of main theorems are presented in Section 3.3.

3.1. Background on least square estimator

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

$$y_t = \beta^\intercal z_t + w_t, \quad t \geq 0, \tag{5}$$

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:

$$\hat{\beta}_T = \arg\min_{\beta^\intercal} \sum_{\tau=0}^{T} \|y_\tau - \beta^\intercal z_\tau\|^2. \tag{6}$$

The following result by Lai and Wei (1982) characterizes the rate of convergence of $\hat{\beta}_T$ to β in terms of unnormalized covariance matrix of covariates $Z_T := \sum_{\tau=0}^{T} z_\tau z_\tau^\intercal$.

Theorem 7 (Theorem 1 of Lai and Wei (1982)) *Suppose the following conditions are satisfied:* **(C1)** $\lambda_{\min}(Z_T) \to \infty$*, a.s. and* **(C2)** $\log(\lambda_{\max}(Z_T)) = o(\lambda_{\min}(Z_T))$*, a.s. Then the least squares estimate in (6) is strongly consistent with the rate of convergence:*

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

The results of Theorem 7 are valid as long as the covariate process $\{z_t\}_{t\geq 0}$ is $\mathcal{G}_{t-1}$-measurable. For the switched least squares system identification if we take $\mathcal{G}_t$ to be equal to $\mathcal{F}_t$ and verify conditions (C1) and (C2) in Theorem 7, then we can use Theorem 7 to establish the strong consistency and rate of convergence. As mentioned earlier in Remark 3, the empirical covariances are coupled across different components due to the system dynamics. For this reason, establishing (C1) and (C2) is non-trivial. In the next section, we establish properties of the system that enable us to prove (C1) and (C2) for the switched least squares system identification.

3.2. 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 (1) of the continuous component of the state in convolutional form as:

$$x_t = \sum_{\tau=0}^{t-1} \Phi(t-1, \tau+1) w_\tau. \tag{7}$$

In the following lemma, we show that Assumption 2 implies that the sum of norms of the state-transition matrices are uniformly bounded.

Lemma 8 (Uniform Boundedness) *Under Assumption 2, 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.*

Proof (**sketch**) Let $\gamma_t = \sigma_{s_t}$ denote the maximum singular value of the system dynamics A_{s_t} at time t. Basic properties of matrix norm implies that:

$$\|\Phi(t-1, \tau-1)\| \leq \gamma_{\tau-1} \cdots \gamma_{t-1} = \prod_{i=1}^{k} \left[\sigma_i^{m_i(t-1,\tau-1)}\right]^{t-\tau+1}, \tag{8}$$

where $m_i(t-1, \tau-1)$ denotes the average number of times the discrete state equals i in the interval $[\tau-1, t-1]$. By the strong law of large numbers and continuity, we can argue that the term inside the square brackets in the right hand side of Eq. (8) converges almost surely to $\sigma_i^{p_i}$. The result then follows from Assumption 2.[1] ■

Next, we characterize the asymptotic behavior of state of the system x_τ and the matrix $X_{i,\tau}$.

Proposition 9 *Under Assumptions 1 and 2, the following hold a.s. for each $i \in \{1, \cdots, k\}$:* **(P1)** $\sum_{\tau \in \mathcal{T}_{i,T}} \|x_\tau\|^2 = \mathcal{O}(T)$, **(P2)** $\lambda_{\max}(X_{i,T}) = \mathcal{O}(T)$, *and* **(P3)** $\liminf_{T\to\infty} \lambda_{\min}(X_{i,T})/|\mathcal{T}_{i,T}| > 0$.

Proof (**sketch**)

(P1): Starting from Eq. (7), we have

$$\|x_t\|^2 \overset{(a)}{\leq} \left(\sum_{\tau=1}^{T-1} \|\Phi(t-1, \tau+1)\| \|w(\tau)\|\right)^2 \overset{(b)}{\leq} \bar{\Gamma} \sum_{\tau=1}^{T-1} \|\Phi(t-1, \tau+1)\| \|w(\tau)\|^2,$$

1. The actual argument is slightly more nuanced and requires ϵ-δ continuity and convergence arguments.

where (a) follows from property of matrix norms, (b) follows from Cauchy-Schwartz inequality, and Lemma 8. Hence,

$$\sum_{t\in\mathcal{T}_{i,T}} \|x_t\|^2 \le \bar{\Gamma} \sum_{t\in\mathcal{T}_{i,T}} \sum_{\tau=1}^{T-1} \|\Phi(t-1,\tau+1)\| \|w(\tau)\|^2 \le \bar{\Gamma}^2 \sum_{\tau=1}^{T-1} \|w(\tau)\|^2. \tag{9}$$

Now, as argued in Lai and Wei (1985), using Assumption 1 and the strong law of large numbers for Martingale difference sequences, we can show that $\sum_{\tau=1}^{T-1} \|w(\tau)\|^2 = \mathcal{O}(T)$, a.s. Combining this with Eq. (9) completes the proof of (P1).

(P2): Follows from (P1) and the inequality $\lambda_{\max}(X_{i,T}) \le \operatorname{tr}(X_{i,T})$.

(P3): Using the strong law of large numbers for martingale difference sequences, we can show:

$$\left\| \sum_{\tau=0}^{T} A_{s_\tau} x_\tau w_\tau^\intercal + w_\tau x_\tau^\intercal A_{s_\tau}^\intercal \right\| = o(T) \quad \text{a.s.}$$

Using this, we can show that $\sum_{t\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal \succeq \sum_{t\in\mathcal{T}_{i,T}} w_\tau w_\tau^\intercal + o(T)$. Hence,

$$\liminf_{|\mathcal{T}_{i,T}|\to\infty} \frac{\sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal}{|\mathcal{T}_{i,T}|} \succeq \liminf_{|\mathcal{T}_{i,T}|\to\infty} \frac{\sum_{\tau\in\mathcal{T}_{i,T}} w_\tau w_\tau^\intercal}{|\mathcal{T}_{i,T}|} = C \succ 0 \quad \text{a.s.}$$

Therefore, $\lambda_{\min}\Big(\liminf_{|\mathcal{T}_{i,T}|\to\infty} \sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal / |\mathcal{T}_{i,T}|\Big) > 0$, a.s. which proves (P3). ■

Corollary 10 *Proposition 9 implies that the system is stable in the average sense. i.e.*

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{\tau=0}^{T-1} \|x_\tau\|^2 < \infty.$$

3.3. Proof of the Main Results

Using the results established in the previous section, we present proof of the main results

Proof of Theorem 2 To prove this theorem, we check the sufficient conditions in Theorem 7. First notice that $X_{i,T}$ is $\mathcal{F}_{T-1}$ measurable. Also we have:

(C1) By Proposition 9-(P3), we see that $\lambda_{\min}(X_{i,T}) \to \infty$ a.s.; therefore, (C1) in Theorem 7 is satisfied.

(C2) Proposition 9-(P2) and (P3) imply that there exist positive constants C_1, C_2, such that :

$$\limsup_{T\to\infty} \frac{\log(\lambda_{\max}(X_{i,T}))}{\lambda_{\min}(X_{i,T})} \le \limsup_{T\to\infty} \frac{\log(C_1) + \log(T)}{C_2 |\mathcal{T}_{i,T}|} = 0 \quad \text{a.s.} \tag{10}$$

where the last inequality uses the fact that $p_i > 0$ implies $|\mathcal{T}_{i,T}| = \mathcal{O}(T)$, a.s. Therefore, the second condition of Theorem 7 is satisfied.

Therefore, by Theorem 7, for each $i \in \{1, \cdots, k\}$, we have:

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

which proves the claim in Theorem 2.

Proof of Corollary 4 Corollary 4 is the direct consequence of Theorem 2. The right hand side of Eq. (10) implies that for each i, the estimation error $\|\hat{A}_{i,T} - A_i\|_\infty$ is upper-bounded by $\mathcal{O}\big(\sqrt{\log(T)/|\mathcal{T}_{i,T}|}\big)$, $a.s.$

Proof of Theorem 6 We first establish the strong consistency of the parameter $\hat{\theta}_T$. By Theorem 2 and the fact that $k < \infty$, we get:

$$\|\hat{\theta}_T - \theta\|_\infty \leq \max_{i\in\{1,\cdots,k\}} \mathcal{O}\left(\sqrt{\frac{\log\left[\lambda_{\max}(X_{i,T})\right]}{\lambda_{\min}(X_{i,T})}}\right), \quad \text{a.s.}$$

Therefore the result follows by applying Theorem 2 to the argmax of above equation. For the second part notice that by the law of large numbers we have $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = p_i$, a.s. Now, by Corollary 4, we get:

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

which is the claim of Theorem 6.

4. Numerical Simulation

In this section, we illustrate the result of Theorem 2 via an example. Consider a SLS with $n = 2, k = 2$, $A_1 = \left[\begin{smallmatrix}1.5 & 0\\ 0 & 0.2\end{smallmatrix}\right]$, and $A_2 = \left[\begin{smallmatrix}0.01 & 0.1\\ 0.1 & 0.1\end{smallmatrix}\right]$, switching probabilities $p = (0.75, 0.25)$, 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 2, but it is not mean square stable (see the next section). We run the switched least squares for the horizon of $T = 30000$ and repeat the experiment for 20 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.

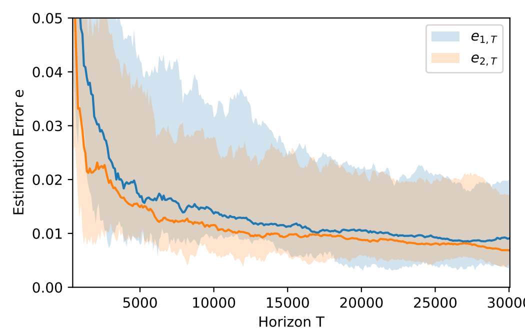

Figure 1: Performance of switched least squares method for the example of Sec. 4. The solid line shows the median across 30 runs and the shaded region shows the 25% to 75% quantile bound.

5. Related work

As mentioned in the introduction, there are two papers which analyze models similar to ours: Sarkar et al. (2019) and Sattar et al. (2021). In this section, we compare our model, assumptions and results from these papers.

Sarkar et al. (2019) investigate the problem of learning the parameters of an unknown SLS of unknown order from input-output data using subspace methods. Under the assumption that the system is mean-square stable, the noise processes are i.i.d. subgaussian, and the system matrices satisfy some technical conditions, they propose an algorithm to estimate an SLS version of the Henkel matrix and obtain parameter estimated by balanced truncation. They show that when the number of samples N_s is sufficiently large, then with high probability the estimation error is $\tilde{\mathcal{O}}(N_s^{-\Delta_s/2})$, where $\Delta_s = \log(1/\rho_{\max})/\log(k/\rho_{\max})$ and $\rho_{\max} = \lambda_{\max}(\sum_{i=1}^k p_i A_i \otimes A_i)$.

The model analyzed in Sarkar et al. (2019) is more general than our model but the proposed algorithms are different. Sarkar et al. (2019) analyze a subspace-based algorithm, while we analyze a switched least squares algorithm. Both of subspace methods and least squares methods are fundamental methods for system identification of linear systems. The results are derived under different assumptions: we impose a slightly weaker assumption on the noise process and our assumption on the stability of the models are different. Moreover, the nature of the results are different: Sarkar et al. (2019) provide high probability rates of convergence while we provide almost sure ones. We disscus the differences between the stability assumptions and the nature of convergence below. Finally, we note that the rate of convergence $\tilde{\mathcal{O}}(N_s^{-\Delta/2})$ depends on the number of subsystems, while our rate of $\tilde{\mathcal{O}}(T^{-1/2})$ does not.

Sattar et al. (2021) consider a SLS where the discrete states evolves in a Markov manner. Such systems are called Markov Jump Linear Systems (MJLS). Under the assumption that the system is mean square stable, the switching distribution is ergodic and the noise is i.i.d. subgaussian, they propose a system identification procedure where random Gaussian noise is injected as control input and system parameters are estimated using least squares. Sattar et al. (2021) show that when T is sufficiently large, then with high probability the estimation error is $\mathcal{O}((\sqrt{k\log T} + \sqrt{\log(1/\delta)})/\sqrt{T})$. Sattar et al. (2021) also propose a certainty equivalent control algorithm and analyze its regret.

The model analyzed in Sattar et al. (2021) is more general than our model and the proposed algorithms are similar. However, the assumptions and the nature of the result differ in a manner similar to those for Sarkar et al. (2019). We impose weaker assumptions on the noise process, our assumption on the stability of the models are different, and we provide almost sure rate of convergence. We discuss the difference between the stability assumptions and the nature of the convergence below.

Discussion on nature of the convergence Both Sarkar et al. (2019) and Sattar et al. (2021) establish high probability rate of convergence. In particular, they show that for any $\delta > 0$ and sufficiently large T, $\|\hat{A}_i - A_i\| \leq \tilde{\mathcal{O}}(f(\delta, T))$ with probability $1-\delta$, where rate of convergence of $f(\delta, T)$ is $o(T)$ but differs in the two papers. In contrast, our results establish an almost sure rate of convergence. Thus our results imply strong consistency of the system identification while the results of Sarkar et al. (2019) and Sattar et al. (2021) do not. This is because strong consistency is defined in terms of almost sure convergence, which is a stronger notion of convergence than convergence in probability implied by the high probability bounds.

On the other hand, the results of Sarkar et al. (2019) and Sattar et al. (2021) are finite-time bounds, i.e., they provide an explicit lower bound on the number of samples needed for the rate of convergence bounds to be valid. In contrast, our result bounds are asymptotic and hold in the limit but do not provide finite time guarantees.

Discussion on Stability Assumption Both Sattar et al. (2021) and Sarkar et al. (2019) assume that the switched system is mean square stable, i.e., there exist 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 $x_0 \in \mathbb{R}$, we have

$$\lim_{\tau\to\infty} \big\|\mathbb{E}[x_\tau] - x_\infty\big\| \to 0 \quad \text{and} \quad \lim_{\tau\to\infty} \big\|\mathbb{E}[x_\tau x_\tau^\intercal] - Q_\infty\big\| \to 0.$$

As shown in Theorem 3.9 of Costa et al. (2006), *mean square stability* is equivalent $\lambda_{\max}(\sum_{i=1}^{k} p_i A_i \otimes A_i) < 1$. Corollary 10 shows that our assumption on stability implies *stability in the average sense* (see Duncan and Pasik-Duncan (1990)), i.e,

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{\tau=0}^{T-1} \|x_\tau\|^2 < \infty.$$

The two notions of stability are different as we illustrate via examples below.

Example 1 *Let* $\theta^\intercal = \{A_1, 0\}$, *and* $p = (p_1, p_2)$, *with* $\lambda_{\max}(p_1 A_1) > 1$ *and* $x_0 \neq 0$. *Then:*

$$\mathbb{E}[x_{\tau+1}] = \mathbb{E}[A_{\sigma_\tau} x_\tau + w_t] = p_1 A_1 \mathbb{E}[x_\tau] = \cdots = (p_1 A_1)^\tau \mathbb{E}(x_0) \implies \lim_{\tau\to\infty} \mathbb{E}(x_\tau) = \infty.$$

Therefore, this system is not mean square stable. However, this system satisfies Assumption 2 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 Assumption 2.*

6. Conclusion and Future Directions

In this paper, we investigated the asymptotic performance of the switched least squares for system identification of (autonomous) switched linear systems. We proposed the switched least squares method and established both data dependent and data independent rates of convergence. We showed this method for system identification is strongly consistent and we derived the almost sure rate of convergence of $\mathcal{O}(\sqrt{\log(T)/T})$. This analysis provide a solid first step toward establishing almost sure regret bounds for adaptive control of SLSs.

The current results are established for autonomous systems with i.i.d. switching when the complete state of the system is observed. Interesting future research directions include relaxing these modeling assumptions and considering non-autonomous (i.e. controlled) systems with more general switching under partial observability.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

Peter E Caines. *Linear stochastic systems*, volume 77. SIAM, 2018.

Peter E Caines and Han-Fu Chen. Optimal adaptive lqg control for systems with finite state process parameters. *IEEE Transactions on Automatic Control*, 30(2):185–189, 1985.

Peter E Caines and Ji-Feng Zhang. On the adaptive control of jump parameter systems via nonlinear filtering. *SIAM journal on control and optimization*, 33(6):1758–1777, 1995.

Ahmet Cetinkaya, Hideaki Ishii, and Tomohisa Hayakawa. Analysis of stochastic switched systems with application to networked control under jamming attacks. *IEEE Transactions on Automatic Control*, 64(5):2013–2028, 2018.

Han-Fu Chen and Lei Guo. Convergence rate of least-squares identification and adaptive control for stochastic systems. *International Journal of Control*, 44(5):1459–1476, 1986.

Han-Fu Chen and Lei Guo. Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost. *SIAM Journal on Control and Optimization*, 25(4):845–867, 1987.

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

Yuan Shih Chow. Local convergence of martingales and the law of large numbers. *The Annals of Mathematical Statistics*, 36(2):552–558, 1965.

Oswaldo Luiz Valle Costa, Marcelo Dutra Fragoso, and Ricardo Paulino Marques. *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.

Claudio De Persis and Pietro Tesi. Input-to-state stabilizing control under denial-of-service. *IEEE Transactions on Automatic Control*, 60(11):2930–2944, 2015.

Grace S Deaecto, Matheus Souza, and José C Geromel. Discrete-time switched linear systems state feedback design with application to networked control. *IEEE Transactions on Automatic Control*, 60(3):877–881, 2014.

Tyrone E Duncan and Bozenna Pasik-Duncan. Adaptive control of continuous-time linear stochastic systems. *Mathematics of Control, signals and systems*, 3(1):45–60, 1990.

Yuguang Fang. A new general sufficient condition for almost sure stability of jump linear systems. *IEEE Transactions on Automatic Control*, 42(3):378–382, 1997.

Yuguang Fang, Kenneth A Loparo, and Xiangbo Feng. Almost sure and δ-moment stability of jump linear systems. *International Journal of Control*, 59(5):1281–1307, 1994.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive control and learning. *Automatica*, 117:108950, 2020a.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *IEEE Transactions on Automatic Control*, 66(4): 1802–1808, 2020b.

Graham Goodwin, Peter Ramadge, and Peter Caines. Discrete-time multivariable adaptive control. *IEEE Transactions on Automatic Control*, 25(3):449–456, 1980.

Pedro Hespanhol and Anil Aswani. Statistical consistency of set-membership estimator for linear systems. *IEEE Control Systems Letters*, 4(3):668–673, 2020.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1): 154–166, 1982.

Tze Leung Lai and Ching Zong Wei. Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems. *Multivariate Analysis VI*, pages 375–393, 1985.

Tze Leung Lai and Ching Zong Wei. Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control*, 31(10):898–906, 1986.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *arXiv preprint arXiv:2003.11227*, 2020.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Necmiye Ozay, Constantino Lagoa, and Mario Sznaier. Set membership identification of switched linear systems with known number of subsystems. *Automatica*, 51:180–191, 2015.

Tuhin Sarkar, Alexander Rakhlin, and Munther Dahleh. Nonparametric system identification of stochastic switched linear systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3623–3628. IEEE, 2019.

Yahya Sattar, Zhe Du, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Identification and adaptive control of Markov jump systems: Sample complexity and regret bounds. *arXiv preprint arXiv:2111.07018*, 2021.

Borna Sayedana, Mohammad Afshari, Aditya Mahajan, and Peter E. Caines. Consistency and Rate of Convergence of Switched Least Squares Estimator for Autonomous Switched Linear Systems, 2021. URL http://www.cim.mcgill.ca/~adityam/projects/reinforcement-learning/preprint/switched-least-squares.pdf.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

William F Stout. *Almost Sure Convergence.* 1974.

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

Yang Zheng, Luca Furieri, Maryam Kamgarpour, and Na Li. Sample complexity of linear quadratic gaussian (LQG) control for output feedback systems. In *Learning for Dynamics and Control*, pages 559–570. PMLR, 2021.

Appendix A. Proof of Lemma 8

Recall that $\sigma_i = \sigma_{\max}(A_i), i \in \{1, \dots, k\}$. Define $\gamma_t = \sigma_{s_t}$. Then, by sub-multiplicative property of the matrix norms, we have:

$$\|\Phi(t-1, \tau+1)\| = \|A_{s_{t-1}} \dots A_{s_{\tau+1}}\| \leq \gamma_{t-1} \cdots \gamma_{\tau+1} =: \Gamma_{t-1,\tau+1}. \tag{11}$$

Given numbers $m_1, \dots, m_k$, define $f(m_1, \dots, m_k) = \sigma_1^{m_1} \cdots \sigma_k^{m_k}$. Let $m_i(t-1, \tau+1) = \sum_{t'=\tau+1}^{\tau-1} \frac{\mathbb{1}\{s_{\tau=i}\}}{t-\tau-1}$ denote the number of times the discrete state equals i in $[\tau+1, t-1]$. Then,

$$\Gamma_{t-1}, \tau+1 = \gamma_{t-1} \cdots \gamma_{\tau+1} = f(m_1(t-1, \tau+1), \dots, m_k(t-1, \tau+1))^{t-\tau-1}.$$

By the strong law of large numbers, we know

$$\lim_{t \to \infty} m_i(t-1, \tau+1) = p_i, \quad \text{a.s.} \quad \forall i \in \{1, \dots, k\}.$$

Furthermore, the rate of convergence of $m_i(t-1, \tau+1)$ to p_i only depends on $\tau+1$ and $t-1$ only through their difference. Thus, for any $\epsilon > 0$, there exists a $N(\epsilon)$ such that for all $t-\tau-1 \geq N(\epsilon)$, $|m_i(t-1, \tau+1) - p_i| < \epsilon \; a.s.$ for all i. By the continuity of $f(\cdot)$, for any $\epsilon' > 0$, there exists a $N'(\epsilon')$ such that for all $t-\tau-1 \geq N'(\epsilon')$, $|f(m_1(t-1, \tau+1), \cdots, m_k(t-1, \tau+1)) - f(p_1, \cdots, p_k)| < \epsilon'$ $a.s.$ Hence,

$$f(m_1(t-1, \tau+1), \dots, m_k(t-1, \tau+1)) < f(p_1, \dots, p_k) + \epsilon' \quad \text{a.s.}$$

By Assumption 2, we know $f(p_1, \dots, p_k) < 1$. Now we can pick ϵ' such that $f(p_1, \dots, p_k) + \epsilon' =: \beta^* < 1$. Then for all $t \geq 1$,

$$\begin{aligned}
&\sum_{\tau=1}^{t-1} f(m_1(t-1, \tau+1), \dots, m_k(t-1, \tau+1))^{t-\tau-1} \\
\leq \; & \sum_{\tau=1}^{t-N(\epsilon')-1} \beta^{*t-\tau-1} + \sum_{\tau=t-N'(\epsilon')}^{t-1} f(m_1(t-1, \tau+1), \dots, m_k(t-1, \tau+1))^{t-\tau-1} \\
< \; & \frac{\beta^{*N'(\epsilon')}}{1-\beta^*} + \sum_{\tau=t-N'(\epsilon')}^{t-1} F_*^{t-\tau-1},
\end{aligned}$$

where $F_* = \max_{p_1, \dots, p_k \in \Delta_k} f(p_1, \dots, p_k)$ (where Δ_k is the k-dimensional simplex), which is clearly bounded. As a result, both terms in the right hand side are bounded which implies the statement in the claim.

Appendix B. Proof of Proposition 9

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

Theorem 11 (Theorem 3.3.1 of Stout (1974)) *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 and for each $\tau \geq 1$ we have $a_\tau \to \infty$ as $\tau \to \infty$, a.s. If for some $p \in (0, 2]$, we have:*

$$\sum_{\tau=0}^{\infty} \frac{\mathbb{E}\big[|X_\tau|^p|\mathcal{F}_{\tau-1}\big]}{a_\tau^p} < \infty,$$

then :

$$\frac{\sum_{\tau=0}^{T} X_\tau}{a_T} \to 0 \quad a.s.$$

B.1. Proof of (P1)

We start by the following Lemma which shows the implication of Assumption 1 on the growth rate of energy of the noise process. This result was presented in Lai and Wei (1985) where a proof sketch was provided. For the sake of completeness, we provide a detailed proof.

Lemma 12 *Assumption 1, implies the following growth rate:*

$$\sum_{\tau=0}^{T} \|w_\tau\|^2 = \mathcal{O}(T), \quad a.s. \tag{12}$$

Proof Let $\zeta_\tau := \|w_\tau\| - \mathbb{E}[\|w_\tau\||\mathcal{F}_{\tau-1}]$. Assumption 1 implies that:

$$\sup_\tau \mathbb{E}[|\zeta_\tau|^2|\mathcal{F}_{\tau-1}] < \infty, \quad \text{a.s.}$$

Hence, by taking $p = 2$ and $a_\tau = \tau$ in Theorem 11, the above expression implies that $\sum_{\tau=0}^{T} \zeta_\tau = o(T)$, $a.s.$ Furthermore, by Assumption 1, $\sum_{\tau=0}^{T} \mathbb{E}[\|w_\tau\|^2|\mathcal{F}_{\tau-1}] = \mathcal{O}(T)$. Therefore, we get:

$$\sum_{\tau=0}^{T} \|w_\tau\|^2 = \sum_{\tau=0}^{T} \zeta_\tau + \sum_{\tau=0}^{T} \mathbb{E}[\|w_\tau\|^2|\mathcal{F}_{\tau-1}] = \sum_{\tau=0}^{T} \zeta_\tau + \mathcal{O}(T) = \mathcal{O}(T), \quad \text{a.s.}$$

■

Using the convolution formula in Eq. (7), we can bound the norm of the state $\|x_t\|^2$ as following:

$$\begin{aligned} \|x_t\|^2 &= \Big(\Big\|\sum_{\tau=1}^{t-1} \Phi(t-1, \tau+1) w(\tau)\Big\|\Big)^2 \overset{(a)}{\leq} \Big(\sum_{\tau=1}^{t-1} \|\Phi(t-1, \tau+1) w(\tau)\|\Big)^2 \\ &\overset{(b)}{\leq} \Big(\sum_{\tau=1}^{t-1} \|\Phi(t-1, \tau+1)\| \|w(\tau)\|\Big)^2 \overset{(c)}{\leq} \Big(\sum_{\tau=1}^{t-1} \Gamma_{t,\tau+1} \|w(\tau)\|\Big)^2 \end{aligned} \tag{13}$$

where (a) follows from triangle inequality and (b) follow from sub-multiplicative property of the matrix norm, and (c) follows from Eq. (11). Now for a fixed i, $i \in \{1, \cdots, k\}$, we have:

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

where (d) follows from Cauchy-Schwarz's inequality, (e) follows from Lemma 8, (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 8.

B.2. Proof of (P2)

First, notice that we have the following lower and upper bounds for maximum eigenvalue of a matrix:

$$\lambda_{\max}\Big(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\intercal\Big) \overset{(a)}{\leq} \operatorname{tr}\Big(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\intercal\Big) = \sum_{t\in\mathcal{T}_{i,T}} \|x_i\|^2$$

where (a) follows from the fact that trace of a matrix is sum its eigenvalues and all eigenvalues of $x_t x_t^\intercal$ are non-negative. Using inequality (a), and Proposition 9-(P1), we get:

$$\lambda_{\max}\Big(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\intercal\Big) = \sum_{t\in\mathcal{T}_{i,T}} \|x_i\|^2 = \mathcal{O}(T) \quad \text{a.s.}$$

which completes the proof.

B.3. Proof of (P3)

B.3.1. PRELIMINARY RESULTS

First we prove the following preliminary lemma:

Lemma 13 *Assumption 1 and 2 imply:*

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

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

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

where (a) follows from Proposition 9-(P1), which implies $S_T = \mathcal{O}(T)$. ■

Lemma 14 *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: $\big[A_{s_\tau}(l,1)x_\tau(1) + \cdots + A_{s_\tau}(l,n)x_\tau(n)\big] w_\tau(p)$. Our goal is to prove:

$$\sum_{\tau=1}^{T} \Big(A_{s_\tau}(l,1)x_\tau(1) + \cdots + A_{s_\tau}(l,n)x_\tau(n) \Big) w_\tau(p) = o(T) \quad a.s.$$

In order to show the above expression, we use Theorem 11 and by setting $a_t = t$ and $p = 2$ we show:

$$\sum_{\tau=1}^{T} \frac{\mathbb{E}\Big[\Big(A_{s_\tau}(l,1)x_\tau(1) + \cdots + A_{s_\tau}(l,n)x_\tau(n)\Big)^2 w_\tau^2(p) \Big| \mathcal{F}_{\tau-1}\Big]}{\tau^2} < \infty \tag{14}$$

We have:

$$\begin{aligned} &\mathbb{E}\Big[\Big(A_{s_\tau}(l,1)x_\tau(1) + \cdots + A_{s_\tau}(l,n)x_\tau(n)\Big)^2 w_\tau^2(p) \Big| \mathcal{F}_{\tau-1}\Big] \\ &\qquad = \sum_{i=1}^{k} p_i \mathbb{E}\Big[\Big(A_i(l,1)x_\tau(1) + \cdots + A_i(l,n)x_\tau(n)\Big)^2 w_\tau^2(p) \Big| \mathcal{F}_{\tau-1}\Big] \end{aligned}$$

Let $A_* = \max_{i\in\{1,\dots,k\}} \|A\|_\infty$. Then, for each fixed i, we have:

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

where (a) is because x_τ is $\mathcal{F}_{\tau-1}$ measurable, and (b) is by Cauchy-Schwarz's inequality. Based on Assumption 1, $\mathbb{E}\big[w_\tau^2(p)\big|\mathcal{F}_{\tau-1}\big]$ is uniformly bounded. Therefore the left hand side of Eq. (14) is bounded by:

$$nA_* \sup_\tau \Big\{ \mathbb{E}[w_\tau^2(p) | \mathcal{F}_{\tau-1}] \Big\} \sum_{\tau=1}^{T} \frac{\|x_\tau\|^2}{\tau^2} \overset{(c)}{\leq} \infty$$

where (c) follows from Lemma 13 . ■

B.3.2. PROOF OF PROPOSITION 9-(P3)

Finally, we prove the statement in the proposition. 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:

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

where (a) follows from Lemma 14. Furthermore, since $|\mathcal{T}_{i,T}| = p_i T$ $a.s.$, we have:

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

(b) holds by Assumption 1. Therefore

$$\liminf_{|\mathcal{T}_{i,T}| \to \infty} \frac{\sum_{\tau \in \mathcal{T}_{i,T}} x_\tau x_\tau^\intercal}{|\mathcal{T}_{i,T}|} \succeq 0 \implies \lambda_{\min}\Big(\liminf_{|\mathcal{T}_{i,T}| \to \infty} \frac{\sum_{\tau \in \mathcal{T}_{i,T}} x_\tau x_\tau^\intercal}{|\mathcal{T}_{i,T}|}\Big) > 0, \quad \text{a.s.}$$

which concludes the proof.

Appendix C. Proof of Corollary 10

Using Eq. (13), we have:

$$\sum_{\tau=1}^{T} \|x_\tau\|^2 = \sum_{i=1}^{k} \sum_{\tau \in \mathcal{T}_{i,T}} \|x_\tau\|^2 \overset{(a)}{=} k\mathcal{O}(T) = \mathcal{O}(T) \quad \text{a.s.}$$

where (a) follows from Prop. 9-(P2).