PAC LEARNABILITY OF STABLE SELECTIVE STATE-SPACE MODELS FROM A SINGLE TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective State-Space Models (SSMs) are dynamical systems where the state-transition, input and readout maps may depend on the current input. This mechanism is central to the celebrated Mamba architecture for long-range tasks, but it is also present in more classical model classes used in control engineering and time-series predictions, such as Linear Parameter-Varying (LPV) systems. In this paper we study learnability from a single time-series for a certain class of selective SSMs of which the state-transition, input and readout maps are affine functions of suitable features of the current inputs. We propose a concrete learning algorithm and a finite-sample probabilistic $\widetilde{O}(1/N^{\zeta})$, $\zeta \leq 1/2$ PAC bound on the parameter-estimation (and hence prediction error) for this learning algorithm. The algorithm itself is an extension of the Ho–Kalman realization-based learning algorithm for linear systems, and it relies on computing the so-called Markov parameters of the SSMs. Our results show that the proposed class of selective SSMs satisfies the single-trajectory analogue of PAC learnability, i.e., it is possible to learn a model with high-enough probability and high-enough accuracy, given a sufficient number of data points.

1 INTRODUCTION

State-space models have been used in learning predictive models of time-series for a long time (Brockwell & Davis, 2006; Ljung, 1998), both in control theory and time-series forecasting. In machine learning, state-space models under the form of recurrent neural networks (RNNs) were used for sequences to sequence modeling. A particular class of state-space models are the so called *selective State-Space Models (SSMs)* (Gu et al., 2021; Gu & Dao, 2024; Orvieto et al., 2023; Honarpisheh et al., 2025) which are defined by equations that are affine in the hidden state. The term SSM appeared recently to denote various state-space models (e.g., Mamba) which were used as a recurrent alternative to attention for long sequences (Gu et al., 2021; Gu & Dao, 2024). Later, SSMs have also been used in time-series forecasting, for example in S-Mamba, TimeMachine, and Mamba4Cast (Wang et al., 2025; Ahamed & Cheng, 2024; Bhethanabhotla et al., 2024). While the term SSM appeared recently, the same model class has been widely used in control theory under the name of *state-affine systems* (Sontag, 1979; Isidori, 1985) for several decades. The reason for their popularity is that they are computationally effective while capable of capturing nonlinear behavior. In particular, state-affine/selective SSMs are known to be universal approximators of nonlinear state-space models (Krener, 1974; Boyd & Chua, 1985; Grigoryeva & Ortega, 2018; Cirone et al., 2024).
SSMs, that are used for time-series forecasting or learning models suitable for designing feedback control law, have to be learned from data which is represented as a single long time-series, in contrast to learning from many i.i.d. data samples. In this paper we give theoretical learning guarantees for a concrete learning algorithm that takes a long, noisy time-series as an input and outputs an SSM belonging to a particular subclass of selective models.

Contribution. We assume that the training data is generated by an unknown noisy stable SSM, and we establish a PAC bound on the *parameter estimation error*, i.e., on the difference between the parameter of the data generator and the estimated SSM. Moreover, we consider a class of SSMs which are known as *Linear-Parameter Varying (LPV) systems with affine dependence on scheduling LPV-SAA* (Tóth, 2010; Scherer, 2009; Apkarian et al., 1995). The latter algorithm is based on the extension of the so called Ho-Kalman realization algorithm (Cox et al., 2018), and it can be viewed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

as the extension of the algorithms for which finite-sample bounds were derived for linear state-space models (Oymak & Ozay, 2022; Tsiamis & Pappas, 2019; Hazan et al., 2017; Simchowitz et al., 2019). The PAC bound is $\widetilde{O}(N^{-\zeta}) = O((\ln(N))^r/N^{\zeta})$ for suitable constants $r > 0, \zeta \in (0, \frac{1}{2}]$ which depend on the data generator and the distribution of the input. For suitable white noise inputs we cant take $r = 0$ and $\zeta = \frac{1}{2}$ leading to a $O(1/\sqrt{N})$ bound. The constants of the bound decrease with the Lyapunov exponent of the SSM which generated the data, with the noise variance, and with the richness of the input. In particular, i.i.d. inputs, which are known to be sufficiently rich (persistently exciting) (Ljung, 1998) result in the best rates.

Motivation parameter estimation instead generalization error. First, establishing PAC bounds on the parameter estimation error instead of the generalization error allows to characterize the predictive capacity of the model under distribution shift. While it requires assuming realizability, i.e., that the data is generated by a noisy SSM, the latter is relatively mild due to the universal approximation property of SSMs. Second, SSMs are used not only for forecasting but also for designing feedback control and Reinforcement Learning. For the latter, parameter estimation error is a more suitable metric for the fitness of the model, as it captures predictive power under distribution shift.

Motivation for LPV-SSAs. LPV-SSAs represent a good tradeoff between simplicity, expressivnes and availability of necessary technical results: **(1)** they have a simple structure yet they are widely used for modeling nonlinear phenomena (Tóth, 2010; Scherer, 2009), **(2)** they contain several relevant subclasses of SSMs (e.g., bilinear models, certain subclasses of Mamba), **(3)** there is a rich literature on learning algorithms for them (Cox et al., 2018; Cox & Tóth, 2021), and their structure theory and identifiability are well-understood (Petreczky et al., 2016; Alkhoury et al., 2017); results which are necessary for PAC bounds. In addition, LPV-SSAs might be a suitable stepping stone towards deriving regret bounds for RL algorithms for partially observed non-linear systems. Many regret bounds for partially observed linear state-space systems are derived for RL algorithms which consist of iterating two steps: 1) during the exploration step a model is learned, 2) the control strategy is computed using the model learned during exploration. Regret bounds are then obtained by applying parameter estimation bounds for the exploration step and combining it with bounds on rewards achieved by controllers designed for different parameters. This approach thus relies on PAC bounds for parameter estimation error and availability of analytic formulas for feedback controllers as functions of the model parameters. The class of LPV-SSAs appears to be a natural candidate for extending this approach, as it can be used to model a wide range of nonlinear processes and there are well-established formulas for controllers as functions of model parameters (Scherer, 2009; Rugh & Shamma, 2000; Apkarian et al., 1995).

Related work. PAC bounds for learning RNNs from i.i.d data were investigated in Koiran & Sontag (1998); Chen et al. (2020); Joukovsky et al. (2021) and for switched state-space models in Sarkar et al. (2019). The latter class belongs to the class of SSMs. PAC bounds for SSMs learned from i.i.d. data were investigated in Liu & Li (2024), for Mamba-specific selective-SSM analysis in Honarpisheh et al. (2025), for LPV systems in Rácz et al. (2026) and a length independent PAC bound for linear SSMs composed with time independent nonlinear layers in Rácz et al. (2025b).
There is a rich literature on PAC bounds for learning autoregressive models from a single time-series (Alquier & Wintenberger, 2012; Alquier et al., 2013; Ziemann et al., 2022; Sattar & Oymak, 2022; Massucci et al., 2022; Metakalard et al., 2026). PAC bounds for parameter estimation error for partially observed linear state-space models were explored in Oymak & Ozay (2022); Jedra & Proutiere (2023); Tsiamis & Pappas (2019); Tsiamis et al. (2023); Hazan et al. (2017); Simchowitz et al. (2019; 2018); Sarkar et al. (2021). PAC bounds for the generalization error of learning linear and non-linear state-space models from a single trajectory were explored in Eringis et al. (2024a;b), however they do not address parameter estimation error bounds specifically for SSMs.
In Rácz et al. (2025a) and Sattar et al. (2025) PAC bounds for parameter estimation error for linear switched system and bilinear systems, respectively, with i.i.d. switching and input were presented. The results of the paper at hand cover a more general system class, which includes both switched and bilinear systems, and it does not assume that the switching/input is i.i.d. Moreover, unlike Sattar et al. (2025) we do not assume that the data generator was started at some specific deterministic state which is not always realistic as it implicitly assumes that at the beginning of the training process the system is in rest, with no prior noise/input. Instead, we assume that the data generator is in its stationary regime, generating a stationary process.

Novelty of the paper. Unlike previous literature we **(1)** bound the parameter estimation error of a more general class of SSMs, **(2)** learned from a single trajectory, **(3)** with inputs which are not necessary i.i.d, **(4)** without assuming that, at the moment when the training data was gathered, the system was in rest, i.e., that there was no prior noise.

2 FROM SELECTIVE SSMS TO AFFINE LPV SYSTEMS

Notation. Boldface characters denote probabilistic quantities, deterministic objects are plain type. Scalar, vector and matrix nature of a term is always clear from the context. For words over Q, Q^* denotes the Kleene star, $Q^+ = Q^* \setminus \{\epsilon\}$, and $|w|$ is word length and vw is the concatenation of words v, w. For a vector a, $a^{(j)}$ is its j-th coordinate and a_i the i-th member of a family of vectors, so $a_i^{(j)}[k]$ is the j-th coordinate of the i-th vector at time k; $M^{(i,j)}$ is a matrix entry; superscripts without parentheses are labels. Discrete time is written $[k]$, continuous time (t). We write $\|x\|_P = \sqrt{x^T P x}$; for a matrix F, $\|F\|_P := \sup_{x\neq 0} \|Fx\|_P / \|x\|_P$ is the induced norm, and $\|\cdot\|_{2\to P}$, $\|\cdot\|_{P\to 2}$ denote the induced norms from the Euclidean norm to $\|\cdot\|_P$ and back. $\|\cdot\|_F$ is the Frobenius norm and $\succeq$ is the Loewner-ordering. $\lambda_{\max\,/\,\min}(P)$ and $\sigma_{\max\,/\,\min}(P)$ are the largest / smallest eigen- / singular values of P. $O(\cdot)$ is the standard Big-O notation, while $\widetilde{O}(\cdot)$ ignores polylogarithmic factors. For a summary of constants used in the paper, see Appendix A.

The learning problem considered in this paper is that of *system identification*: consider a pair of stochastic processes $(\mathbf{y}, \tilde{\mathbf{u}})$, where $\tilde{\mathbf{u}}[k]$ corresponds to inputs from $\mathbb{R}^{\tilde{n}_u}$ and $\mathbf{y}$ corresponds to outputs (labels) from $\mathbb{R}^{n_y}$ such that $\mathbf{y}$ is the output response of an unknown noisy, dynamical system

$$\mathbf{x}[k+1] = f_\theta(\mathbf{x}[k], \tilde{\mathbf{u}}[k]) + \mathbf{w}[k], \quad \mathbf{y}[k] = h_\theta(\mathbf{x}[k], \tilde{\mathbf{u}}[k]) + \mathbf{v}[k] \tag{1}$$

to the input $\tilde{\mathbf{u}}$, where $\mathbf{w}, \mathbf{v}$ are i.i.d. noises processes independent of $\tilde{\mathbf{u}}$, and $\mathbf{x}$ is the state process taking values in $\mathbb{R}^{n_x}$. The dynamical system is a parametrization of dynamical systems corresponding to the parameter value $\theta = \theta_{\text{true}}$. The task is to find an estimate $\hat{\theta}_N$ of θ_{true} from a finite sample path $\{y[k] = \mathbf{y}[k](\omega), \tilde{u}[k] = \tilde{\mathbf{u}}[k](\omega)\}_{k=0}^N$, where $\omega \in \Omega$ denotes an outcome in the underlying probability space. The motivation is twofold: prediction and control/reinforcement learning (RL). For prediction, the noiseless version of (1) (when $\mathbf{w} = 0$,$\mathbf{v} = 0$)

$$\hat{\mathbf{x}}[k+1] = f_\theta(\hat{\mathbf{x}}[k], \tilde{\mathbf{u}}[k]), \hat{\mathbf{y}}[k] = h_\theta(\hat{\mathbf{x}}[k], \tilde{\mathbf{u}}[k]) \tag{2}$$

can be considered as a predictor of $\mathbf{y}[k]$ based on past inputs $\mathbf{u}[s]_{s\leq k}$, for which $\theta = \theta_{true}$ is often (e.g., linear RNNs Ljung (1998)) the best (minimal variance) predictor, even for out-of-distribution inputs. Hence, by extension, the the estimate of θ_{true} is expected to give rise to a good predictor. Moreover, for control/RL estimates of θ_{true} are useful for finding the best strategy for choosing the input $\tilde{\mathbf{u}}$. We assume that data generator is an SSM/LPV-SSA, as defined below.

2.1 SELECTIVE STATE-SPACE MODELS

A *selective State-Space Model* (selective SSM) is a discrete-time dynamical system whose coefficient matrices depend on the current input through a fixed feature map

$$f_\theta(x, \tilde{u}) = A_\theta(\phi(\tilde{u}))x + B_\theta(\phi(\tilde{u}))\tilde{u}, \quad h_\theta(x, \tilde{u}) = C_\theta(\phi(\tilde{u}))x + D(\phi(\tilde{u}))\tilde{u} \tag{3}$$

The map $\phi : \mathbb{R}^{\tilde{n}_u} \to \mathbb{R}^{n_p}$ is fixed and contains no learnable parameter; all learnable quantities are collected in θ, and $A_\theta : \mathbb{R}^{n_p} \to \mathbb{R}^{n_x \times n_x}$, $B_\theta : \mathbb{R}^{n_p} \to \mathbb{R}^{n_x \times \tilde{n}_u}$, $C_\theta : \mathbb{R}^{n_p} \to \mathbb{R}^{n_y \times n_x}$ and $D_\theta : \mathbb{R}^{n_p} \to \mathbb{R}^{n_y \times \tilde{n}_u}$ are matrix-valued functions of the features. Selection refers to the fact that the input influences or selects the coefficient matrices at time k via the feature map ϕ. For example, **(1)** when $\phi(\tilde{u}) = \tilde{u}$, a linear selector and A_θ is affine, (3) is a bilinear system (Isidori, 1985); **(2)** when $\phi(\tilde{u}) = (\mathbf{1}_{U_1}(\tilde{u}), \ldots, \mathbf{1}_{U_{n_p}}(\tilde{u}))$ are indicator features for a partition $U_1, \ldots, U_{n_p}$ of the input set, which select one of finitely many coefficient matrices, (3) is a linear switched system (Petreczky, 2015) whose mode is determined by the input;
(3) ϕ can represent the the features of the Mamba core, built from the step $\Delta[k] = \text{softplus}(a + q^T \tilde{\mathbf{u}}[k])$ and the input coordinates (see Appendix D).
Note from (3) that a selective SSM depends on its input trajectory in two separate ways. First, it depends linearly on the input through the terms $B_\theta\big(\phi(\tilde{\mathbf{u}}[k])\big)\tilde{\mathbf{u}}[k]$ and $D_\theta\big(\phi(\tilde{\mathbf{u}}[k])\big)\tilde{\mathbf{u}}[k]$, while the function ϕ or the matrices A_θ, B_θ, C_θ and D_θ may induce a highly nonlinear dependence on u.

While SSMs mostly appear as predictors, i.e., without noise as in (2), it is not unreasonable to consider them as data generators and study the induced system identification problem, as it also provides insight to learnability in terms of prediction error. In the control theory literature SSMs have been known as state-affine systems (Sontag, 1979).

2.2 Linear parameter-varying systems

While SSMs are attractive, for technical reasons we prefer to work with a related class of systems, namely that of Linear Parameter-Varying (LPV) systems (Tóth, 2010). Note that in case of SSMs, the input u enters the equation both linearly and non-linearly, by the virtue of nonlinear dependence of the matrices $A_\theta, B_\theta, C_\theta, D_\theta$ on u. In LPV systems, inputs are divided into two categories: *scheduling variables* which enter the model equations non-linearly, and *control inputs* which enter the model equations linearly.

Formally, we assume that $\tilde{\mathbf{u}}[k] = \begin{bmatrix}\mathbf{p}^T[k], & \mathbf{u}^T[k]\end{bmatrix}^T$ where $\mathbf{u}$ takes values in $\mathbb{R}^{n_u}$ and $\mathbf{p}$ takes values in $\mathcal{P} \subseteq \mathbb{R}^{n_p}$. The process $\mathbf{p}$ corresponds to scheduling variables, and $\mathbf{u}$ to control inputs (referred to as input in the rest of the paper). A *noisy LPV system* is a special case of system (1),

$$f_\theta(x, p, u) = A_\theta(p)x + B_\theta(p)u, \quad h_\theta(x, p, u) = C_\theta(p)x + D_\theta(p)u \tag{4}$$

where $A_\theta : \mathcal{P} \to \mathbb{R}^{n_x \times n_x}$, $B_\theta : \mathcal{P} \to \mathbb{R}^{n_x \times n_u}$, $C_\theta : \mathcal{P} \to \mathbb{R}^{1 \times n_x}$ and $D_\theta : \mathcal{P} \to \mathbb{R}^{1 \times n_u}$ are the coefficient maps. That is, LPV systems are a specific subclass of SSMs, where the components of the inputs which enter the equations nonlinearly are regrouped into scheduling variables.

LPV systems are popular in control theory (Rugh & Shamma, 2000; Tóth, 2010; Scherer, 2009) due to their ability to represent nonlinear systems, while still offering computationally effective methods for synthesizing feedback controllers, using methods from robust control. Applications of LPV systems range from aerospace engineering to robotics.

Setting $\mathbf{u}[k] = \tilde{\mathbf{u}}[k]$ and $\mathbf{p}[k] := \phi(\mathbf{u}[k])$ immediately turns the selective SSM (3) into the LPV system (4) with $A_\theta(\cdot) = A_\theta(\cdot)$, and likewise for B_θ, C_θ, D_θ. The scheduling signal is observed, because ϕ is known and $\mathbf{u}$ is measured, and it depends on the input.

In general, A, B, C, D can be arbitrary functions of the scheduling. However, in this paper we focus on *LPV systems with affine dependence on the scheduling (sometimes called affine LPV state-space representation, or LPV-SSA)* (Tóth, 2010; Petreczky et al., 2016), i.e.

$$\begin{aligned} A_\theta(\mathbf{p}[k]) &= \sum_{i=0}^{n_p} p_i[k] A_{i,\theta}, \quad B_\theta(\mathbf{p}[k]) = \sum_{i=0}^{n_p} p_i[k] B_{i,\theta}, \\ C_\theta(\mathbf{p}[k]) &= \sum_{i=0}^{n_p} p_i[k] C_{i,\theta}, \quad D_\theta(\mathbf{p}[k]) = \sum_{i=0}^{n_p} p_i[k] D_{i,\theta} \end{aligned} \tag{5}$$

for some constant matrices $\{A_{i,\theta}, B_{i,\theta}, C_{i,\theta}, D_{i,\theta}\}_{i=0}^{n_p}$, where $\mathbf{p}[k] = [p_1[k] \ \cdots \ p_{n_p}[k]]^T$ denote the coordinate functions of the scheduling vector and we have $p_0[k] \equiv 1$. We refer to the set $Q = \{0, 1, \ldots, n_p\}$ as scheduling indices. Without loss of generality, we can assume that $|p_i[k]| \leq 1$ as long as $\mathbf{p}$ is bounded, by applying to $\mathbf{p}$ an affine transformation (Tóth, 2010), i.e., $\mathcal{P} \subseteq \{p \in \mathbb{R}^{n_p} \mid \|p\|_\infty \leq 1\}$.

We further assume that $\operatorname{span}\{[1 \ p^T]^T : p \in \mathcal{P}\} = \mathbb{R}^{n_p+1}$, so that the coefficient matrices are determined by the maps $A_\theta(\cdot), \ldots, D_\theta(\cdot)$ on $\mathcal{P}$; and that the output is scalar, i.e. $n_y = 1$.

LPV-SSAs are often used as they are especially suitable for control design and many LPV systems can be transformed to LPV-SSAs by changing the scheduling variable. For instance, if $X_\theta(p)$, $X \in \{A, B, C, D\}$ are polynomials, then we can take as new scheduling parameters as monomials of p. See Tóth (2010) for further examples.

Since SSMs can be viewed as LPV systems and vice versa, many SSMs of interest can be represented by LPV-SSAs. For example, polynomial dependence on the step $\Delta[k]$ from the third example in Section 2.1 or on the input coordinates falls into this category with the monomials of inputs as scheduling variables. Discrete input case in a finite horizon also translates to LPV-SSA representation. For other examples (e.g., Mamba), see Appendix D.

Dropping θ from $X_\theta, X_{i,\theta}$, $X \in \{A, B, C, D\}$. For the sake of simplicity, for LPV-SSA we will identify the parameter set Θ used to parametrize models with the tuple of matrices $\theta = (A_i, B_i, C_i, D_i)_{i \in Q}$ and with the dependence $(A_{i,\theta}, B_{i,\theta}, C_{i,\theta}, D_{i,\theta}) = (A_i, B_i, C_i, D_i)$. Therefor we may omit θ from the notation, and use $A, B, C, D, A_i, B_i, C_i, D_i$ instead of $A_\theta, B_\theta, C_\theta, D_\theta, A_{i,\theta}, B_{i,\theta}, C_{i,\theta}, D_{i,\theta}$.

2.3 PAC-LEARNABILITY: INFORMAL STATEMENT

We show that LPV-SSA satisfy the suitable analogue of PAC-learnability for single long time-series: we present a concrete learning algorithm, which for each sample of $\{y[k], p[k], u[k]\}_{k=0}^{N}$ returns matrices $\{\hat{A}_q, \hat{B}_q, \hat{C}_q, \hat{D}_q\}$, such that the following learning guarantees hold.

Theorem 2.1 (Informal theorem)**.** *Assume* $\mathbf{y}$ *is the response of a LPV-SSA* (1) *with inputs* $\tilde{\mathbf{u}} = \begin{bmatrix}\mathbf{p}^T, & \mathbf{u}^T\end{bmatrix}$ *and with matrices defined by* $\{A_i, B_i, C_i, D_i\}_{i=0}^{n_p}$*. Under suitable assumptions on* $\mathbf{u}, \mathbf{p}$*, there exist constants* $c, r > 0$ *and* $\zeta \in (0, \frac{1}{2}]$ *such that for any* $\delta \in (0, 1)$*, there exists* N_0 *such that* $\forall N \geq N_0$

$$\mathrm{EstErr} := \max_{q \in Q, X \in \{A,B,C,D\}} \|\widehat{X}_q - \bar{X}_q\|_F \leq \frac{c\left(1 + \sqrt{\ln\left(\frac{1}{\delta}\right)}\right)(\ln(N))^r}{N^{\zeta}}$$

holds with probability at least $1 - \delta$ *over all the samples from* $\{\mathbf{y}[k], \mathbf{p}[k], \mathbf{u}[k]\}_{k=0}^{N}$*. In particular, for certain* $\mathbf{u}, \mathbf{p}$*,* $\zeta = \frac{1}{2}$*.*

The learning algorithm, which is based on a Ho-Kalman-like realization algorithm, is presented in Section 3. The precise assumptions are presented in Section 4. The formal statement of the result is presented in Section 5.

3 REDUCED-BASIS HO–KALMAN REALIZATION AND LEARNING ALGORITHM FOR LPV SYSTEMS

In this section we briefly summarize a variant of the Ho-Kalman algorithm that is applicable for LPV systems (Cox et al., 2018). This algorithm estimates the system matrices of an unknown LPV system from a single input-output-scheduling trajectory of this unknown system.

We start by defining the Markov parameters of an LPV-SSA (Petreczky et al., 2016), that is a natural extension of Markov parameters of LTI systems. Let us interpret $Q = \{0, 1, \ldots, n_p\}$ as a finite alphabet. For a word $w = w_0 w_1 \cdots w_{m-1} \in Q^m$ let $A_w := A_{w_0} A_{w_1} \cdots A_{w_{m-1}}$ and $A_\epsilon := I_{n_x}$ where ϵ denotes the empty word. The *Markov parameters* $\{M_w\}_{w \in Q^+}$ are defined as

$$M_{w_0} := D_{w_0}, \qquad M_w := C_{w_0} A_{w_1} \cdots A_{w_{m-2}} B_{w_{m-1}} \in \mathbb{R}^{1 \times n_u} \quad (m \geq 2). \tag{6}$$

In other words, $M_{\alpha\beta\gamma} = C_\alpha A_\beta B_\gamma$ for $\alpha, \gamma \in Q$ and $\beta \in Q^*$. We will later see that the first letter of a word refers to the most recent time step and M_w is the coefficient of $\mathbf{u}[k - m + 1]$'s contribution to the output of the LPV system at time k, weighted by $p_{w_0}[k]\, p_{w_1}[k-1] \cdots p_{w_{m-1}}[k - m + 1]$.

For a word $w = w_0 \cdots w_{m-1} \in Q^m$, define the scheduling-weighted input regressor as

$$\mathbf{z}_w^{\mathbf{u}}[k] = \mathbf{u}[k - m + 1] \prod_{j=0}^{m-1} p_{w_j}[k - j]. \tag{7}$$

We call these regressors because (34) in Lemma B.6 shows that the output of the LPV system can be written in a linear regression-like form with the above products as the observed quantities.

For $L \geq 1$, let $\mathcal{V}_L = \bigcup_{m=1}^{L} Q^m$ be the union of nonempty words of length at most L and let

$$d_L = n_u |\mathcal{V}_L| = n_u \sum_{m=1}^{L} (n_p + 1)^m, \quad \mathbf{Z}_L[k] = \mathrm{col}\big(\mathbf{z}_w^{\mathbf{u}}[k]\big)_{w \in \mathcal{V}_L}, \tag{8}$$

$$G_L = \mathbb{E}[\mathbf{Z}_L[k]\mathbf{Z}_L[k]^T], \qquad \Theta_L = [M_w]_{w \in \mathcal{V}_L} \in \mathbb{R}^{1 \times d_L}, \tag{9}$$

where $\mathbf{Z}_L[k]$ is a d_L-dimensional column vector obtained by stacking vertically the scheduling-weighted input regressors $\mathbf{z}_w^{\mathbf{u}}[k]$ for all $w \in \mathcal{V}_L$. Note that G_L is the Gram matrix of the scheduling-weighted input regressor coordinates collected in $\mathbf{Z}_L[k]$, and the expectation is taken over time. Under mild assumptions (the product of past noises with future scheduling is uncorrelated with $\mathbf{z}_w^{\mathbf{u}}$, Assumption 4.6), $\Theta_L = E[\mathbf{y}[k]\mathbf{Z}_L[k]^T] G_L^{-1}$ i.e., the Markov parameters can be computed from the covariance of outputs with the regressors $\mathbf{Z}_L$.

Next we define the so-called *row and column selections* α and β as follows,

$$\alpha = \{(\nu_i, \alpha_i)\}_{i=1}^{n_x}, \qquad \beta = \{(\mu_j, q_j, l_j)\}_{j=1}^{n_x},$$

where $\nu_i, \mu_j \in Q^*$ have length at most $n_x - 1$ and $l_j \in \{1, \ldots, n_u\}$. They are called selections because they select specific rows and columns of an infinite matrix of Markov parameters. While there exist versions of the current algorithm which use this infinite matrix, the presented variant only uses a finite number of Markov parameters defined by the selections α and β. For a given $q \in Q$ and selections α and β let the matrices $H_{\alpha,\beta}, H_{\alpha,q,\beta}, H_{\alpha,q}, H_{\beta,q}$ be defined elementwise as follows,

$$H^{(i,j)}_{\alpha,\beta} = M^{(l_j)}_{\alpha_i \nu_i \mu_j q_j}, \quad H^{(i,j)}_{\alpha,q,\beta} = M^{(l_j)}_{\alpha_i \nu_i q \mu_j q_j}, \quad H^{(i,j)}_{\alpha,q} = M^{(j)}_{\alpha_i \nu_i q}, \quad H^{(1,j)}_{\beta,q} = M^{(l_j)}_{q \mu_j q_j}. \tag{10}$$

We call these matrices Hankel matrices and under suitable assumptions $\operatorname{rank}(H_{\alpha,\beta}) = n_x$ (Cox et al., 2018). From the Hankel matrices we obtain the following estimates on the system parameters

$$\bar{A}_q = H^{-1}_{\alpha,\beta} H_{\alpha,q,\beta}, \quad \bar{B}_q = H^{-1}_{\alpha,\beta} H_{\alpha,q}, \quad \bar{C}_q = H_{\beta,q}, \quad \bar{D}_q = M_q. \tag{11}$$

It is folklore in control theory that the estimates above are isomorphic to the true parameters, i.e. the system matrices of the LPV system that generated the input-output-scheduling trajectory that was the input of the algorithm (Petreczky & Peeters, 2010). Note that the words needed in (10) have length at most $2n_x + 1$, and only $O((n_p + 1)n_x^2)$ Markov parameters are used.
In reality, we do not have access to the true Markov parameters of a system and must estimate them from the observed data. Given $S_N = \{(y[k], u[k], p[k])\}_{k=0}^N$ and $L \leq N + 1$, let

$$\widehat{\Lambda}_{L,N} = \frac{1}{N - L + 2} \sum_{k=L-1}^{N} \mathbf{y}[k] \mathbf{Z}_L[k]^T, \qquad \widehat{\Theta}_{L,N} = \widehat{\Lambda}_{L,N} G_L^{-1}. \tag{12}$$

The w-indexed block of $\widehat{\Theta}_{L,N}$, denoted by $\widehat{M}_w$, estimates M_w by Lemma B.10, and $\widehat{\Theta}$ estimates Θ. The term G_L^{-1} accounts for correlations between the coordinates of $\mathbf{Z}_L[k]$ (see Lemma B.8). Replacing each M_w in (10) by $\widehat{M}_w$ gives the empirical Hankel blocks and

$$\widehat{A}_q = \widehat{H}^{-1}_{\alpha,\beta} \widehat{H}_{\alpha,q,\beta}, \quad \widehat{B}_q = \widehat{H}^{-1}_{\alpha,\beta} \widehat{H}_{\alpha,q}, \quad \widehat{C}_q = \widehat{H}_{\beta,q}, \quad \widehat{D}_q = \widehat{M}_q. \tag{13}$$

If $\widehat{H}_{\alpha,\beta}$ is singular, the algorithm returns a fixed default model of the same dimensions, with all coefficient matrices zero. It returns the same default if the sample is too short for the prescribed horizon. This convention defines the learning algorithm on every sample.

The main theorem in Section 5 is a theoretical upper bound on the difference between the output of the above algorithm when applied with the true Markov parameters (resulting in $\bar{A}, \bar{B}, \bar{C}, \bar{D}$) and the output when applied with the estimated Markov parameters (resulting in $\widehat{A}_q, \widehat{B}_q, \widehat{C}_q, \widehat{D}_q$).

Remark 3.1. *The learning algorithm assumes that G_L is supplied (see* (12)*). A data-only least-squares version replaces it by $\widehat{G}_{L,N} = \frac{1}{N-L+2} \sum_{k=L-1}^{N} \mathbf{Z}_L[k] \mathbf{Z}_L[k]^T$. We conjecture that our result can extended to the latter case, but it remains future work.*

4 Assumptions

Assumptions on the matrices of the data generator. We assume that the data generator is quadratically stable, and that it is a minimal dimensional representation of its input-output behavior in the absence of noise. The former is necessary to ensure that the generated data is stationary and mixing, and hence learning can take place. The latter is necessary to ensure that the matrices of the generator are unique up to a change of basis, i.e., the parameter estimation problem is well-posed. This assumption is closely related to the well-known concept of *identifiability*, see e.g. Ljung (1998).

Assumption 4.1 (Uniform quadratic stability)**.** *There exist $P \succ 0$ and $\rho \in (0, 1)$ such that*

$$A(p)^T P A(p) \preceq \rho^2 P \qquad \text{for every } p \in \mathcal{P}.$$

Equivalently, $\sup_{p \in \mathcal{P}} \|A(p)\|_P \leq \rho$. *Moreover, we assume that* $K_B = \sum_{i \in Q} \|B_i\|_{2 \to P} < \infty$, $K_C = \sum_{i \in Q} \|C_i\|_{P \to 2} < \infty$ *and* $K_D = \sum_{i \in Q} \|D_i\|_2 < \infty$.

The condition above is not always easy to check. A computationally effective sufficient condition in terms of $\{A_i\}_{i=0}^{n_p}$ is that $A_i^T P A_i \prec \rho^2 P/(1 + n_p)^2$ for all $i \in Q$; in this case $\|A(p)\|_P \leq \sum_{i \in Q} |p_i| \, \|A_i\|_P < \rho$ when $|p_i| \leq 1$. Note that this condition is not necessary. If $\{A_i\}_{i=0}^{n_p}$ the conditions can be checked using standard Linear Matrix Inequalities, see e.g. Boyd et al. (1994).

For LPV-SSAs arising from SSMs, i.e., $\mathbf{p}[k] = \phi(\mathbf{u}[k])$ for the known feature map ϕ and an admissible input set $\mathcal{U} \subseteq \{u : \|u\|_2 \leq K_u\}$, we take $\mathcal{P} = \phi(\mathcal{U})$ and thus the supremum over $\mathcal{P}$ in Assumption 4.1 equals the supremum over $u \in \mathcal{U}$ after substituting $p = \phi(u)$.

Assumption 4.2 (Minimality)**.** *The deterministic LPV-SSA determined by $\{A_i, B_i, C_i, D_i\}_{i=0}^{n_p}$ is minimal in the sense of Petreczky et al. (2016): its extended n_x-step reachability and observability matrices have full row and column rank, respectively.*

The assumption above guarantees that the matrices of the data generator are unique up to a change of basis. In particular, the matrices of the data generator can be taken to be the matrices (11) returned by the Kalman-Ho realization algorithm, since for non-singular matrix T $T\bar{A}_q T^{-1} = A_q$, $T\bar{B}_q = B_q$, $\bar{C}_q T^{-1} = C_q$, $q \in Q$ (Petreczky et al., 2016). In turn, this allows us to relate the estimation error to concentration inequalities for empirical Hankel-matrices.

Assumptions on the input and scheduling. We assume that $\mathbf{u}, \mathbf{p}$ are generated by a special kind of dynamical system called class-$\mathcal{S}$, which, informally is a stable system with a unique steady-state response for any bounded input and it forgets past inputs at an exponential rate. If such systems are driven by bounded i.i.d. input, then their output satisfies certain weak dependence properties (Alquier & Wintenberger, 2012; Eringis, 2025; Dedecker et al., 2007) that we exploit in the proof of Theorem 5.1. Consider a system on a bounded state set $\mathcal{X} \subseteq \mathbb{R}^{n_s}$ and bounded input and output sets, and let

$$s[k+1] = f(s[k], e[k]), \qquad z[k] = h(s[k], e[k]). \tag{14}$$

For initial state s_0 at t_0 and a bi-infinite input path $e : \mathbb{Z} \to E$, let $s(s_0, t_0, e; t)$ denote the state of (14) at $t \geq t_0$.

Definition 4.3 (Class-$\mathcal{S}$ system (Eringis, 2025))**.** *A system of the form* (14) *is of class $\mathcal{S}$ if, for every bounded bi-infinite input $e : \mathbb{Z} \to E$, there exists a unique bounded full state trajectory $s_e : \mathbb{Z} \to \mathcal{X}$ satisfying $s_e(t+1) = f(s_e(t), e[t])$, and constants $C \geq 1$, $\rho \in [0, 1)$, and $L_e, L_{h,s}, L_{h,e} > 0$, independent of the input path, exist such that*

$$\|s(s_0, t_0, e; t) - s_e(t)\| \leq C\rho^{t-t_0}\|s_0 - s_e(t_0)\|, \tag{15}$$

$$\|s_{e_1}(t) - s_{e_2}(t)\| \leq L_e \sum_{j=1}^{\infty} \rho^{j-1}\|e_1(t-j) - e_2(t-j)\|, \tag{16}$$

$$\|h(s_1, e_1) - h(s_2, e_2)\| \leq L_{h,s}\|s_1 - s_2\| + L_{h,e}\|e_1 - e_2\|. \tag{17}$$

Existence of s_e is required by the definition; (15) states exponential convergence to that trajectory.

Assumption 4.4 (Joint bounded data generator)**.** *There exists a class-$\mathcal{S}$ system* (14) *and a bounded i.i.d. process $\boldsymbol{\xi}$, such that for the input $\boldsymbol{\xi}$* (14) *generates as unique stationary output*

$$\mathbf{e}[k] = [\mathbf{u}[k]^T, \mathbf{p}[k]^T, \mathbf{w}[k]^T, \mathbf{v}[k]^T]^T.$$

Furthermore, we assume $\|\mathbf{u}[k]\|_2 \leq K_u$, $\|\mathbf{w}[k]\|_2 \leq K_w$, $|\mathbf{v}[k]| \leq K_v$, and $|\mathbf{p}_i[k]| \leq 1$.

In Lemma B.6 there exists a unique stationary bounded process which is the output response of (14) to the input $\boldsymbol{\xi}$. Hence, the assumption is well-posed and $\mathbf{e}$ is stationary and bounded. The assumption that $\mathbf{u}$ and $\mathbf{p}$ are generated by a dynamical system (14) is widespread system identification (Ljung, 1998), it means that the inputs $\mathbf{u}, \mathbf{p}$ are white noises filtered by stable filters; details and examples are given in Appendix B.2. If $\mathbf{u}, \mathbf{p}$ are i.i.d, then this assumption is satisfied.

Assumption 4.5 (Growing joint persistence of excitation)**.** *For every $L \geq 1$, G_L is positive definite. Moreover, there exist constants $K_G \geq 1$ and $b \geq 0$ such that, for every $L \geq 1$,*

$$\|G_L^{-1}\|_1 \leq K_G e^{bL}, \quad b < \log(1/\rho) \tag{18}$$

where ρ si as in Assumption 4.1.

The assumption above is a *persistency of excitation assumption*, it ensures that the input and scheduling are rich enough. If $\mathbf{u}, \mathbf{p}$ are both zero mean i.i.d. process and $\{\mathbf{p}[k]\}_{k=-\infty}^{\infty}$ and $\{\mathbf{u}[k]\}_{k=-\infty}^{\infty}$ are independent, then G_L is the identity matrix and b can be taken to be zero and hence the assumption above holds for any $\rho \in (0, 1)$. Finally, we want the inputs and noises to be uncorrelated.

Assumption 4.6 (Noise–regressor orthogonality)**.** *We assume that for every $L \geq 1$, $s, k \in \mathbb{Z}$, $s \leq k$, $v_s, \ldots, v_k \in Q$, $\mathbb{E}[\begin{bmatrix}\mathbf{w}^T[s], & \mathbf{v}^T[s]\end{bmatrix}^T \left(\Pi_{j=s}^{k} \mathbf{p}_{v_j}[j]\right) \mathbf{Z}_L[k]^T] = 0$.*

This assumption requires the inputs/scheduling parameters and noises to be uncorrelated. If the scheduling and input are independent of the noises and the noises are zero mean, which is a standard assumption in system identification (Ljung, 1998), the condition above holds. This assumption allows us to eliminate the effect of the noise on the output when estimating Markov-parameters.

5 MAIN RESULT

The main theoretical contribution of the paper is the following theorem on the estimation error of the described Ho–Kalman algorithm for the identification of stable LPV systems.

Theorem 5.1 (Finite-sample bound for the estimated system matrices)**.** *Suppose Assumptions 4.1, 4.2, 4.4, 4.5, and 4.6 hold. Fix $\delta \in (0,1)$ and Ho–Kalman selections α, β for which $H_{\alpha,\beta}$ is nonsingular. There exist a positive constant $K_{\mathrm{CBS}} = O(1/(1-\rho)^3)$ [1] independent of L, N, δ so that the followings hold. For $1 \le L \le N+1$, let d_L as in* (8) *and*

$$
\begin{aligned}
K_y &= \frac{K_C\big(K_B K_u + \sqrt{\lambda_{\max}(P)}K_w\big)}{1-\rho} + K_D K_u + K_v \\
K_1(L,\delta,N) &= \left[\big(K_y K_u + 2B_\xi K_{\mathrm{CBS}}(1+L)^2\big)\sqrt{\frac{2\log(2d_L/\delta)}{N-L+2}} + \frac{K_C K_B K_u^2 \rho^{L-1}}{1-\rho}\right] \\
K_{\mathrm{JG}}(L,\delta,N) &= \sqrt{n_u}\, K_G e^{bL} K_1(L,\delta,N).
\end{aligned}
\tag{19}
$$

Furthermore, let

$$
\lambda_\rho = \log(1/\rho), \qquad \zeta = \frac{1}{2} - \frac{b}{2\lambda_\rho} > 0, \tag{20}
$$

where positivity follows from Assumption 4.5, and choose

$$
L_N = \max\left\{2n_x + 1, \left\lfloor \frac{\log N}{2\lambda_\rho} \right\rfloor\right\}. \tag{21}
$$

Define

$$
\sigma_H = \sigma_{\min}(H_{\alpha,\beta}) > 0, \qquad c_H = \sqrt{n_x \max\{n_x, n_u\}}, \quad K_H = \max_{q\in Q}\left\{1, \|H_{\alpha,q,\beta}\|_2, \|H_{\alpha,q}\|_2\right\},
$$

$$
K_{\mathrm{HK}} = c_H \max\left\{1, \frac{2}{\sigma_H} + \frac{2K_H}{\sigma_H^2}\right\}.
$$

Suppose $N \ge 1$ is large enough that $L_N \le N+1$ and

$$
c_H K_{\mathrm{JG}}(L_N, \delta, N) \le \frac{\sigma_H}{2}. \tag{22}
$$

Then the Ho-Kalman algorithm using G_{L_N} satisfies the following with probability at least $1-\delta$,

$$
\mathrm{EstErr} := \max_{\substack{q\in Q \\ X\in\{A,B,C,D\}}} \|\widehat{X}_q - \bar{X}_q\|_F \le K_{\mathrm{HK}} K_{\mathrm{JG}}(L_N, \delta, N), \tag{23}
$$

where $\bar{X}_q$ and $\widehat{X}_q$ are as in Section 3.

The full proof is in Appendix B.2.2, here we only give a brief overview. The proof has four steps. Class-$\mathcal{S}$ properties and stability convert the joint generator and the LPV recursion into bounded functionals of i.i.d. innovation processes, which can be represented by Causal Bernoulli Shifts (CBS) (Alquier & Wintenberger, 2012; Alquier et al., 2013) (B.7). The stability assumption gives bounded stationary solutions and geometric decay of certain partial finite sums contributing to the output of the system, over words of equal length (Lemma B.6). Summing these groups gives the output expansion; bounding each coordinate of the tail moment gives the truncation bias (Lemma B.8).

[1] K_{CBS} defined in (44) in Appendix B.2 and can be chosen to satisfy $K_{\mathrm{CBS}} = O((1+K_u)[K_D K_u + K_v + K_C(K_B K_u + \sqrt{\lambda_{\max}(P)}K_w)(1-\rho)^{-1}] + K_u + K_u K_C[(1+K_u)K_B + \sqrt{\lambda_{\max}(P)}](1-\rho)^{-2} + K_u K_C(K_B K_u + \sqrt{\lambda_{\max}(P)}K_w)(1-\rho)^{-3})$ and some other implicit constants such as $\max_{1\le i\le n_p}\|A_i\|_P$, and the class-$\mathcal{S}$ constants of the joint data generator.

Applying concentration inequalities for CBSs controls the joint empirical moment, after which G_L^{-1} converts the empirical moment error and truncation bias into the estimation error of Markov parameters (Lemma B.10). Finally, standard Hankel and inverse perturbation bounds pass this error through the reduced-basis Ho–Kalman formulas see the proof of Theorem 5.1.

Discussion of Theorem 5.1. The bound has three separate elements related to sampling, to the truncation and to the properties of the reduced base realization. The first element is the first part of $K_1(L, \delta, N)$ which captures the sampling error related to the joint output-regression moments raising from the data single trajectory. This term is governed by the logarithm of the regressor dimension, d_L. The constants are related to boundedness and temporal dependencies.

The second element is the bias term in $K_1(L, \delta, N)$, it is induced by the truncation of Markov parameters with a higher length than L, and it decays with a ρ^L rate due to uniform stability. As a result, increasing L reduces the truncation bias but it increases $K_G e^{bL}$, the quantitative joint-excitation contribution in K_{JG}. The chosen L_N and λ_ρ act as a balancer between these terms, therefore for a fixed $\delta \in (0, 1)$ the term $K_{\mathrm{JG}}(L_N, \delta, N)$, and consequently the rate of the error, is $\widetilde{O}(1/N^\zeta)$ with probability at least $1 - \delta$, where $\zeta = \frac{1}{2} - \frac{b}{2\lambda_\rho} > 0$. The exponent ζ captures the relation between the decay rate λ_ρ of the system memory and the growth rate b of the inverse joint Gram matrix. In some cases b can be chosen to be zero, for example in case of white noise inputs which is common in control, thus the rate becomes $\widetilde{O}(N^{-1/2})$. Additionally, the regressor dimension enters the bound logarithmically, $\log(d_L) = O(\log(N))$, as a direct result of the chosen L_N.

The third element, K_{HK} measures the conditioning of the selected Ho–Kalman realization, see the end of the proof in Appendix B.2.2, via the term σ_H.

See Appendix C for the extension of Theorem 5.1 to multi-output systems.

6 Conclusion and future work

In this paper we consider learning stable selective SSMs from a single time series through their representation as LPV systems with affine dependence on scheduling. Under the stated assumptions, the input may be temporally correlated and the scheduling signal may depend on the input. We have considered a Ho-Kalman based learning algorithm and have presented a finite-sample probabilistic bound on the parameter estimation error. Consequently the error bound has a rate $\widetilde{O}(1/N^\zeta)$ with $0 < \zeta \leq 1/2$. Among others, the bounds is influenced by the uniform stability condition and the joint persistence of excitation. The resulting theorem provides a PAC-style identification guarantee for the considered family of affine LPVs including certain selective SSMs.

One of the limitation of our work is that the Ho-Kalman learning algorithm assumes that the population Gram matrix, G_L, is known. A fully data-driven implementation requires estimating the population Gram matrix and selecting a suitable Hankel basis. We also assume that the data generator has an LPV representation with affine dependence on the scheduling. Applying the result to a broader class of selective SSMs requires satisfying the stated assumptions on stability, minimality, and persistence of excitation. An additional limitation is the truncated horizon. Even though an increased horizon decreases the truncation bias, it may significantly enlarge the upper bound of the norm of the inverse Gram matrix resulting in balancing between truncation and stability.

Future work includes empirical estimation of G_L, broader coverage of Mamba, and relaxation of the assumptions and especially the connection between stability and the inverse of G_L.

Reproducibility statement

All mathematical objects used by main text are defined either in main text or in the Appendix. The assumptions and the important constants are discussed in the main text. All mathematical claims are either rigorously proven or properly cited.

Statement on the use of AI

AI tools were used to assist with drafting, language editing, literature research and finding relevant related work. The authors checked the correctness of mathematical claims and proofs, citations, and the final text.

References

Md Atik Ahamed and Qiang Cheng. Timemachine: A time series is worth 4 mambas for long-term forecasting, 2024.

Z. Alkhoury, M. Petreczky, and G. Mercère. Identifiability of affine linear parameter-varying models. *Automatica*, 80:62 – 74, 2017. ISSN 0005-1098. doi: http://dx.doi.org/10.1016/j.automatica.2017.01.029. URL http://www.sciencedirect.com/science/article/pii/S0005109817300390.

P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883 – 913, 2012. doi: 10.3150/11-BEJ359. URL https://doi.org/10.3150/11-BEJ359.

P. Alquier, X. Li, and O. Wintenberger. Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 1(2013):65–93, 2013.

Pierre Apkarian, Pascal Gahinet, and Greg Becker. Self-scheduled H_∞ control of linear parameter-varying systems: A design example. *Automatica*, 31(9):1251–1261, 1995. doi: 10.1016/0005-1098(95)00038-X. URL https://doi.org/10.1016/0005-1098(95)00038-X.

Sathya Kamesh Bhethanabhotla, Omar Swelam, Julien Siems, David Salinas, and Frank Hutter. Mamba4cast: Efficient zero-shot time series forecasting with state space models. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Stephen Boyd and Leon O. Chua. Fading memory and the problem of approximating nonlinear operators with volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, November 1985. doi: 10.1109/TCS.1985.1085649.

Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *SIAM Studies in Applied Mathematics*. SIAM, 1994. doi: 10.1137/1.9781611970777. URL https://web.stanford.edu/~boyd/lmibook/.

Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer, New York, 2nd edition, 2006. ISBN 978-0-387-95351-9. doi: 10.1007/978-1-4757-2526-1.

M. Chen, X. Li, and T. Zhao. On generalization bounds of a family of recurrent neural networks. In *AISTATS*, 2020.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3SzrqwupUx.

P. B. Cox and R. Tóth. Linear parameter-varying subspace identification: A unified framework. *Automatica*, 123:109296, 2021.

P. B. Cox, M. Petreczky, and R. Tóth. Towards efficient maximum likelihood estimation of LPV-SS models. *Automatica*, 97(9):392–403, 2018.

Jérôme Dedecker, Paul Doukhan, Gabriel Lang, José Rafael León, Sana Louhichi, and Clémentine Prieur. *Weak Dependence: With Examples and Applications*, volume 190 of *Lecture Notes in Statistics*. Springer Science & Business Media, New York, NY, 2007. ISBN 978-0-387-69951-6. doi: 10.1007/978-0-387-69952-3.

D. Eringis. *Learning with guarantees: A PAC-Bayesian approach to learning time-series.* PhD thesis, Aalborg University, 2025.

D. Eringis, J. Leth, Z-H. Tan, R. Wisniewski, and M. Petreczky. Pac-bayes generalisation bounds for dynamical systems including stable rnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11901–11909, 2024a.

D. Eringis, J. Leth, Zh.-H. Tan, R. Wisniewski, and M. Petreczky. Pac-bayesian error bound, via rényi divergence, for a class of linear time-invariant state-space models. In *ICML*, pp. 12560–12587. PMLR, 2024b.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems. *Journal of Machine Learning Research*, 19(24):1–40, 2018. URL http://jmlr.org/papers/v19/18-020.html.

A. Gu, I. Johnson, K. Goel, K. K. Saab, T. Dao, A. Rudra, and C. Re. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *NeurIPS*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1711.00946.

Arya Honarpisheh, Mustafa Bozdag, Octavia Camps, and Mario Sznaier. Generalization error analysis for selective state-space models through the lens of attention. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f96839fc751b67492e17e70f5c9730e4-Abstract-Conference.html.

A. Isidori. *Nonlinear control systems: an introduction*. Springer, 1985.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of linear systems: Fundamental limits and optimal algorithms. *IEEE Transactions on Automatic Control*, 68(5):2805–2820, 2023.

B. Joukovsky, T. Mukherjee, H. Van Luong, and N. Deligiannis. Generalization error bounds for deep unfolding rnns. In *UAI*, 2021.

P. Koiran and E. D. Sontag. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

A. Krener. Bilinear and nonlinear realization of input-output maps. *SIAM Journal on Control*, 13 (4), 1974.

Fusheng Liu and Qianxiao Li. From generalization analysis to optimization designs for state space models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31383–31405, 2024.

L. Ljung. System identification. In *Signal analysis and prediction*, pp. 163–173. Springer, 1998.

L. Massucci, F. Lauer, and M. Gilson. A statistical learning perspective on switched linear system identification. *Automatica*, 145:110532, 2022. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2022.110532. URL https://www.sciencedirect.com/science/article/pii/S0005109822003934.

A. Metakalard, F. Lauer, K. Colin, and M. Gilson. Uniform error bounds for quantized dynamical models. *IFAC Journal of Systems and Control*, 35:100373, 2026. ISSN 2468-6018. doi: https://doi.org/10.1016/j.ifacsc.2026.100373. URL https://www.sciencedirect.com/science/article/pii/S2468601826000131.

A. Orvieto, S. L. Smith, A. Gu, A. Fernando, C. Gulcehre, R. Pascanu, and S. De. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023.

S. Oymak and N. Ozay. Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis. *IEEE Trans. Autom. Control*, 67(4):1914–1928, 2022.

A. Pavlov and N. Van De Wouw. Steady-state analysis and regulation of discrete-time nonlinear systems. *IEEE transactions on automatic control*, 57(7):1793–1798, 2011.

M. Petreczky. Realization theory of linear hybrid systems. In *Hybrid Dynamical Systems: Observation and control*, pp. 59–101. Cham, Switzerland: Springer International, 2015.

M. Petreczky and R. Peeters. Spaces of nonlinear and hybrid systems representable by recognizable formal power series. In *MTNS*, volume 5, 2010.

M. Petreczky, R. Tóth, and G. Mercère. Realization theory for lpv state-space representations with affine dependence. *IEEE Transactions on Automatic Control*, 62(9):4667–4674, 2016.

D. Rácz, M. Petreczky, and B. Daróczy. A finite-sample bound for identifying partially observed linear switched systems from a single trajectory. *IEEE Control Systems Letters*, 2025a.

Dániel Rácz, Mihály Petreczky, and Bálint Zoltán Daróczy. Length independent generalization bounds for deep SSM architectures via Rademacher contraction and stability constraints. *Transactions on Machine Learning Research*, 2025b. URL https://openreview.net/forum?id=Vo6wHBv07k.

W. Rugh and J. S. Shamma. Research on gain scheduling. *Automatica*, 36(10):1401–1425, 2000.

Dániel Rácz, Martin Gonzalez, Mihály Petreczky, András Benczúr, and Bálint Daróczy. A finite-sample generalization bound for stable LPV systems. *Mathematics of Control, Signals, and Systems*, 2026. doi: 10.1007/s00498-025-00427-7.

T. Sarkar, A. Rakhlin, and M. Dahleh. Nonparametric system identification of stochastic switched linear systems. In *IEEE CDC*, pp. 3623–3628, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther A. Dahleh. Finite time LTI system identification. *Journal of Machine Learning Research*, 22(26):1–61, 2021. URL https://jmlr.org/papers/v22/19-725.html.

Y. Sattar, Y. Jedra, M. Fazel, and S. Dean. Finite sample identification of partially observed bilinear dynamical systems, 2025. URL https://arxiv.org/abs/2501.07652. arxiv:2501.07652.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23(140):1–49, 2022. URL https://jmlr.org/papers/v23/20-617.html.

C. W. Scherer. Robust controller synthesis is convex for systems without control channel uncertainties. In *Model-Based Control: Bridging Rigorous Theory and Advanced Technology*, pp. 13–31. Springer, 2009.

M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *COLT*, pp. 439–473. PMLR, 2018.

M. Simchowitz, R. Boczar, and B. Recht. Learning linear dynamical systems with semi-parametric least squares. In *COLT*, 2019.

E. D. Sontag. Realization theory of discrete-time nonlinear systems: Part I – the bounded case. *IEEE Trans. on Circuits and Systems*, CAS-26(4), 1979.

R. Tóth. *Modeling and identification of linear parameter-varying systems*, volume 403. Springer, 2010.

A. Tsiamis and G. J. Pappas. Finite sample analysis of stochastic system identification. In *CDC*, 2019.

A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023. doi: 10.1109/MCS.2023.3310345.

Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Xiaocui Yang, Han Zhao, Daling Wang, and Yifei Zhang. Is mamba effective for time series forecasting? *Neurocomputing*, 619:129178, 2025.

I. M. Ziemann, H. Sandberg, and N. Matni. Single trajectory nonparametric learning of nonlinear dynamics. In *COLT*, pp. 3333–3364. PMLR, 2022.

A TABLE OF CONSTANTS USED IN THE PAPER

Constant	Defined in
A_0^e	Section B, Assumption B.1
A_0^y	Section B.2.1, proof of Lemma B.7
A_1^e	Section B, Assumption B.1
A_1^y	Section B.2.1, proof of Lemma B.7
α (informal exponent)	Section 2.3, Theorem 2.1 (informal)
$a(H)$	Section B
$\widetilde{a}(H)$	Section B
$\underline{a}_i$	Section D.5, before Eq. (80)
$\overline{a}_i$	Section D.5, before Eq. (80)
a_j^e	Section B.2.1
$a_j(H)$	Section B
a_j^y	Section B.2.1, proof of Lemma B.7
$a_r^{z,w}$	Section B.2.1, Eq. (41)
b	Section 4, Eqs. (18)
B	Section B.2.2, Lemma B.9
B_ξ	Section B, Assumption B.1
C	Section 4, Eq. (15)
c (rate bound)	Section 2.3, Theorem 2.1 (informal)
c (Euler example)	Section D.5, before Eq. (83)
c_H	Section 5, Theorem 5.1
c_H^{mo}	Section C.2, Eq. (61)
c_i	Section D.5, Eq. (80)
δ	Section 5, Theorem 5.1
Δ_0	Section D.5, before Eq. (81)
Δ_H	Section B.2.2, proof of Theorem 5.1
$\Delta_{\max}$	Section D.4
$\Delta_{\min}$	Section D.4
ε	Section B.2.2, Lemma B.9
η	Section D.5
η_A	Section D.5, Eq. (82)

Continued on the next page

Constant	Where defined
ε_C	Section B.2.2, Eq. (52)
$\varepsilon_C^{\mathrm{mo}}$	Section C.1, Eq. (56)
$F_\star$	Section B.1, Proposition B.3
h (Euler example)	Section D.5, before Eq. (83)
h (polynomial example)	Section D.5, before Eq. (81)
h_i	Section D.5, Eq. (80)
K_1	Section 5, Eq. (19)
K_B	Section 4, Assumption 4.1
K_C	Section 4, Assumption 4.1
$K_{\overline{C}}$	Section D.5
K_{CBS}	Section B.2.1, Eq. (44)
K_D	Section 4, Assumption 4.1
K_G	Section 4, Eq. (18)
K_H	Section 5, Theorem 5.1
K_{HK}	Section 5, Theorem 5.1
$K_{\mathrm{HK}}^{\mathrm{mo}}$	Section C.2, Eq. (61)
K_{JG}	Section 5, Eq. (19)
$K_{\mathrm{JG}}^{\mathrm{mo}}$	Section C.1, Eq. (57)
K_u	Section 4, Assumption 4.4
K_v	Section 4, Assumption 4.4
K_w	Section 4, Assumption 4.4
K_w^P	Section B.2.1, Eq. (35)
K_x	Section B.2.1, Eq. (35)
K_y	Section 5, Eq. (19)
K_z	Section B.2.1, Eq. (43)
L_e	Section 4, Eq. (16)
L_e^y	Section B.2.1, proof of Lemma B.7
$L_{f,e}$	Section B.1, Eq. (24)
$L_{f,e}^{\mathrm{LPV}}$	Section B.2.1, proof of Lemma B.6
$L_{h,e}$	Section 4, Eq. (17)
$L_{h,e}^{\mathrm{LPV}}$	Section B.2.1, proof of Lemma B.6
$L_{h,e}^y$	Section B.2.1, proof of Lemma B.7
$L_{h,s}$	Section 4, Eq. (17)
$L_{h,s}^y$	Section B.2.1, proof of Lemma B.7
λ_ρ	Section 5, Eq. (20)
L_θ	Section B.2.2, Lemma B.9
$L_\theta(L)$	Section B.2.1, Eq. (40)
μ	Section D.4
N_0	Section 2.3, Theorem 2.1 (informal)
P	Section 4, Assumption 4.1
r	Section 2.3, Theorem 2.1 (informal)

Continued on the next page

Constant	Where defined
R	Section D.5, Eq. (81)
ρ (LPV system)	Section 4, Assumption 4.1
ρ (class $\mathcal{S}$)	Section 4, Eqs. (15)–(16)
ρ_1, ρ_2	Section B.1, Lemma B.5
$\rho_{\max}$	Section B.1, proof of Lemma B.5
ρ_*	Section D.5, Eq. (79)
ρ_y	Section B.2.1, proof of Lemma B.7
σ_H	Section 5, Theorem 5.1
τ_L	Section B.2.2, Eq. (47)
$\tau_L^{(0)}$	Section B.2.1, Eq. (38)
v_1, v_2	Section D.5, before Eq. (83)
ζ	Section 5, Eq. (20)

B TECHNICAL RESULTS AND PROOFS

Proof overview. The proof has four steps. Class-$\mathcal{S}$ stability converts the joint generator and the LPV recursion into bounded CBS functionals. Uniform quadratic stability gives bounded stationary solutions and geometric decay of the finite sums over words of equal length. Summing these groups gives the output expansion; bounding each coordinate of the tail moment gives the truncation bias. CBS concentration controls the joint empirical moment, after which G_L^{-1} converts moment error into Markov-parameter error. Finally, standard Hankel and inverse perturbation bounds pass this error through the reduced-basis Ho–Kalman formulas.

Throughout the proof, we employ the following additional notations and assumptions which left out from the main text for presentation purposes. These assumptions are all technical and neither of them mean any real restriction compared to the assumptions in the main text.

When the input of a class-$\mathcal{S}$ system is an i.i.d. stochastic process, $\boldsymbol{\xi}$ with $\|\boldsymbol{\xi}[k]\| \le B_\xi$, referred to as innovation, its stationary output $\mathbf{z}$ is a Causal Bernoulli Shift (Alquier & Wintenberger, 2012), $\mathbf{z}[k] = H(\boldsymbol{\xi}[k], \boldsymbol{\xi}[k-1], \ldots)$, and its temporal dependence is measured by the Lipschitz coefficients of H. A scalar CBS map H has coefficients $a_j(H) \ge 0$ if

$$|H(v_0, v_1, \ldots) - H(v_0', v_1', \ldots)| \le \sum_{j\ge 0} a_j(H)\|v_j - v_j'\|.$$

Let $a(H) = \sum_{j\ge 0} a_j(H)$ and $\widetilde{a}(H) = \sum_{j\ge 0} j\, a_j(H)$. Proposition 4.2 of Alquier & Wintenberger (2012) implies that, for bounded innovation, $2B_\xi \widetilde{a}(H)$ bounds the weak-dependence coefficient used in its concentration theorem in the mentioned Proposition. A sufficient contraction condition, examples of admissible generators, and the stationarity and cascade properties used in the proofs are collected in Appendix B.1.

Assumption B.1 (Extension of Assumption 4.4)**.** *In addition to Assumption 4.4, we require the CBS coefficients $\{a_j^e\}_{j\ge 0}$ satisfy*

$$A_0^e := \sum_{j\ge 0} a_j^e < \infty, \qquad A_1^e := \sum_{j\ge 0} j a_j^e < \infty.$$

Let $B_\xi := \|\boldsymbol{\xi}[0]\|_\infty < \infty$.

B.1 CLASS-$\mathcal{S}$ AND CBS PRELIMINARIES

Remark B.2 (Generator examples)**.** *The class-$\mathcal{S}$ part of Assumption 4.4 includes: (i) stable linear filters $s[k+1] = Fs[k] + G\boldsymbol{\xi}[k]$ with an induced norm in which $\|F\| < 1$; (ii) bounded nonlinear*

filters satisfying Proposition B.3; (iii) finite cascades of such filters; and (iv) a temporally correlated input followed by an observed bounded Lipschitz selector $\mathbf{p}[k] = g(\mathbf{u}[k])$. *These constructions address only the joint-generator condition. The joint-Gram and noise–regressor assumptions must still be checked for the resulting process.*

Proposition B.3 (A sufficient condition for class $\mathcal{S}$)**.** *Suppose that* $\mathcal{X}$ *is nonempty and compact,* $f(\mathcal{X}, e) \subseteq \mathcal{X}$ *for every admissible input* e*, and* h *is Lipschitz. Suppose there are a norm* $\|\cdot\|_P$*,* $\rho \in (0,1)$*, and* $L_{f,e} < \infty$ *such that*

$$\|f(s,e) - f(s',e')\|_P \le \rho\|s - s'\|_P + L_{f,e}\|e - e'\| \tag{24}$$

for all admissible s, s', e, e'*. Assume also that, for one fixed* $s_\star \in \mathcal{X}$*,* $F_\star := \sup_e \|f(s_\star, e) - s_\star\|_P < \infty$*. Then* (14) *is of class* $\mathcal{S}$ *with* $C = 1$*, rate* ρ*, and* $L_e = L_{f,e}$ *when all state norms in Definition 4.3 are taken to be* $\|\cdot\|_P$*.*

Proof. For a bounded bi-infinite input path e, define the finite pullback state

$$s_t^{(N)} := s(s_\star, t - N, e; t). \tag{25}$$

Here $s_\star$ is an auxiliary reference state used only at the artificial starting time $t - N$; it is not the initial state of the observed trajectory. Applying (24) for the N subsequent common inputs gives

$$\|s_t^{(N+1)} - s_t^{(N)}\|_P \le F_\star \rho^N, \qquad \|s_e(t) - s_t^{(N)}\|_P \le \frac{F_\star \rho^N}{1-\rho}. \tag{26}$$

The first bound makes $(s_t^{(N)})_{N \ge 0}$ a Cauchy sequence in $\mathcal{X}$. Compactness implies that its limit $s_e(t)$ belongs to $\mathcal{X}$. The second bound follows by summing the first one. Since $s_{t+1}^{(N+1)} = f(s_t^{(N)}, e[t])$, continuity of f gives $s_e(t+1) = f(s_e(t), e[t])$. Boundedness of $\mathcal{X}$ makes this a bounded full state trajectory. Both inequalities are uniform over t, N, and every admissible input path; they are also uniform over reference states in any set for which $F_\star$ has a common finite upper bound. The same contraction proves uniqueness and (15). Iterating (24) for two inputs and taking the pullback limit proves (16). The output condition is assumed. This is the contractive sufficient case of (Pavlov & Van De Wouw, 2011, Theorem 1) and (Eringis, 2025, Appendix to Paper E, Lemma 1.6). □

Lemma B.4 (Stationary solution and fading memory)**.** *For a class-*$\mathcal{S}$ *system, a bounded stationary input process* $\mathbf{e}$ *has bounded stationary steady-state state and output processes. If* $\mathbf{e}$ *is i.i.d., then the steady-state output is a CBS with admissible coefficients*

$$a_0(H_z) = L_{h,e}, \qquad a_j(H_z) = L_{h,s} L_e \rho^{j-1}, \quad j \ge 1. \tag{27}$$

In particular,

$$a(H_z) = L_{h,e} + \frac{L_{h,s} L_e}{1-\rho}, \qquad \widetilde{a}(H_z) = \frac{L_{h,s} L_e}{(1-\rho)^2}. \tag{28}$$

Proof. For $r \in \mathbb{Z}$, define the shift $(\mathsf{T}^r e)[k] = e[k + r]$. Uniqueness of the steady-state trajectory gives

$$s_{\mathsf{T}^r e}(k) = s_e(k + r), \qquad z_{\mathsf{T}^r e}(k) = z_e(k + r). \tag{29}$$

The steady-state maps are measurable because they are pointwise limits of finite-horizon measurable state recursions; hence stationarity of $\mathbf{e}$ and (29) imply stationarity of the state and output processes.

For an input history $(e_0, e_1, \ldots) = (\mathbf{e}[t], \mathbf{e}[t-1], \ldots)$, define $H_s(e_1, e_2, \ldots) = s_e(t)$ and $H_z(e_0, e_1, \ldots) = h(s_e(t), e_0)$. The robustness inequality (16) and the output Lipschitz inequality (17) give (27). Summing the geometric and arithmetic–geometric series gives (28). This is the argument of (Eringis, 2025, Appendix to Paper E, Lemmas 1.4–1.5). □

Lemma B.5 (Cascade closure)**.** *Let* $\mathfrak{S}_1$ *and* $\mathfrak{S}_2$ *be class-*$\mathcal{S}$ *systems, and suppose the output of* $\mathfrak{S}_1$ *is an admissible input of* $\mathfrak{S}_2$*. Their series interconnection is a class-*$\mathcal{S}$ *system. If their exponential rates are* ρ_1 *and* ρ_2*, the cascade can be assigned any rate* $\rho \in (\max\{\rho_1, \rho_2\}, 1)$*; its remaining constants depend only on the constants of the two systems and on* ρ*.*

Proof. The steady-state state of the cascade is $(s_{1,e}(t), s_{2,z_{1,e}}(t))$. Existence and uniqueness follow successively from those of $\mathfrak{S}_1$ and $\mathfrak{S}_2$. In both the initial-state and input-robustness estimates, the only new term is a convolution of the form

$$\sum_{j=0}^{k-1} \rho_2^{k-1-j} \rho_1^j \le k \rho_{\max}^{k-1} \le \frac{\rho^k}{\rho(1-\rho_{\max}}/\rho)^2,$$

where $\rho_{\max} = \max\{\rho_1, \rho_2\}$. Thus these estimates are exponential with rate ρ. The output Lipschitz bound follows by substituting the output bound of $\mathfrak{S}_1$ into that of $\mathfrak{S}_2$. This specializes (Eringis, 2025, Appendix to Paper E, Lemma 1.8). □

B.2 Proofs for the LPV-SSA

B.2.1 Stable expansion and growing-word CBS bounds

For $L \ge 1$, let $\Theta_L = [M_w]_{w \in \mathcal{V}_L} \in \mathbb{R}^{1 \times d_L}$ where the Markov parameters are concatenated in the same order as the blocks of $\mathbf{Z}_L$.

Lemma B.6 (Stationary solution)**.** *Under Assumptions 4.4 and 4.1, the following statements hold.*

1. *Along the observed scheduling path, the LPV recursion has a unique bounded stationary state process* $\mathbf{x}$ *and a bounded stationary output process* $\mathbf{y}$*. The recursions*

$$\begin{aligned} \mathbf{x}^{\mathbf{u}}[k+1] &= A(\mathbf{p}[k])\mathbf{x}^{\mathbf{u}}[k] + B(\mathbf{p}[k])\mathbf{u}[k], \\ \mathbf{x}_\eta[k+1] &= A(\mathbf{p}[k])\mathbf{x}_\eta[k] + \mathbf{w}[k], \\ \boldsymbol{\eta}[k] &= C(\mathbf{p}[k])\mathbf{x}_\eta[k] + \mathbf{v}[k] \end{aligned} \tag{30}$$

   have unique bounded stationary state processes, and

$$\mathbf{x} = \mathbf{x}^{\mathbf{u}} + \mathbf{x}_\eta, \qquad \mathbf{y}[k] = C(\mathbf{p}[k])\mathbf{x}^{\mathbf{u}}[k] + D(\mathbf{p}[k])\mathbf{u}[k] + \boldsymbol{\eta}[k]. \tag{31}$$

   Here $\boldsymbol{\eta}$ *is the output produced by the noises along the same scheduling path.*

2. *The stationary output has the expansion*

$$\mathbf{y}[k] = \sum_{m=1}^{\infty} \left(\sum_{w \in Q^m} M_w \mathbf{z}_w^{\mathbf{u}}[k] \right) + \boldsymbol{\eta}[k]. \tag{32}$$

   The series over m *converges absolutely and uniformly over bounded admissible trajectories, and in mean square. For every* $L \ge 1$*, define*

$$\mathbf{R}_L[k] = \sum_{m=L+1}^{\infty} \left(\sum_{w \in Q^m} M_w \mathbf{z}_w^{\mathbf{u}}[k] \right). \tag{33}$$

   Then

$$\mathbf{y}[k] = \Theta_L \mathbf{Z}_L[k] + \mathbf{R}_L[k] + \boldsymbol{\eta}[k]. \tag{34}$$

   The remainder tends to zero geometrically in L, uniformly over these trajectories.

Proof of Lemma B.6. Constants and contraction. With the constants from Assumptions 4.4 and 4.1, set

$$K_w^P = \sqrt{\lambda_{\max}(P)}\, K_w, \qquad K_x = \frac{K_B K_u + K_w^P}{1-\rho}, \qquad K_y = K_C K_x + K_D K_u + K_v. \tag{35}$$

By Assumption 4.1, for every $p \in \mathcal{P}$ and all x, x',

$$\|A(p)(x - x')\|_P \le \rho \|x - x'\|_P.$$

Writing $\Pi_0[k] = I$ and $\Pi_j[k] = A(\mathbf{p}[k-1]) \cdots A(\mathbf{p}[k-j])$ for $j \ge 1$, we therefore have $\|\Pi_j[k]\|_P \le \rho^j$. Moreover, $\|B(p)\|_{2 \to P} \le K_B$, $\|C(p)\|_{P \to 2} \le K_C$ and $\|D(p)\|_2 \le K_D$ on $\mathcal{P}$.

Class $\mathcal{S}$ and stationarity. The ball $\{x : \|x\|_P \le K_x\}$ is nonempty, compact and forward invariant, since for every admissible (u, p, w) and every x in this ball,

$$\|A(p)x + B(p)u + w\|_P \le \rho K_x + K_B K_u + K_w^P = K_x.$$

On this ball the affine maps in (4) are Lipschitz in (u, p, w, v) for the component-sum norm

$$\|(u, p, w, v)\|_\oplus = \|u\|_2 + \|p\|_1 + \|w\|_P + |v|.$$

Indeed, $\|A(p) - A(p')\|_P \le \max_{i\ge1} \|A_i\|_P \, \|p - p'\|_1$, and similarly for B, C and D. Thus the state map has input Lipschitz constant

$$L_{f,e}^{\mathrm{LPV}} = \max\left\{K_B, 1,\ K_x \max_{i\ge1} \|A_i\|_P + K_u \max_{i\ge1} \|B_i\|_{2\to P}\right\},$$

and the output map has state Lipschitz constant K_C and input Lipschitz constant

$$L_{h,e}^{\mathrm{LPV}} = \max\left\{K_D, 1,\ K_x \max_{i\ge1} \|C_i\|_{P\to 2} + K_u \max_{i\ge1} \|D_i\|_2\right\}.$$

The maxima over $i \ge 1$ run over $i = 1, \dots, n_p$ and are zero when $n_p = 0$. These constants are finite for the fixed coefficient matrices. At $x = 0$, the forcing is uniformly bounded by $K_B K_u + K_w^P$. Hence Proposition B.3 applies with rate ρ.

Cascading the LPV system with the generator of Assumption 4.4 and applying Lemmas B.5 and B.4 gives bounded stationary state and output processes $\mathbf{x}, \mathbf{y}$, with

$$\|\mathbf{x}[k]\|_P \le K_x, \qquad |\mathbf{y}[k]| \le K_C K_x + K_D K_u + K_v = K_y.$$

The same argument after omitting the forcing $\mathbf{w}$, respectively $B(\mathbf{p})\mathbf{u}$, while keeping the original scheduling path fixed, gives bounded stationary solutions $\mathbf{x}^{\mathbf{u}}$ and $\mathbf{x}_\eta$ of the recursions in (30), with

$$\|\mathbf{x}^{\mathbf{u}}[k]\|_P \le \frac{K_B K_u}{1-\rho}, \qquad \|\mathbf{x}_\eta[k]\|_P \le \frac{K_w^P}{1-\rho}.$$

These solutions are unique among bounded full trajectories: the difference d of any two solutions with the same forcing and scheduling path satisfies

$$\|d[k]\|_P \le \rho^j \|d[k-j]\|_P \le \rho^j \sup_{t\in\mathbb{Z}} \|d[t]\|_P \longrightarrow 0 \qquad (j \to \infty).$$

The sum $\mathbf{x}^{\mathbf{u}} + \mathbf{x}_\eta$ solves the full recursion, so uniqueness gives

$$\mathbf{x} = \mathbf{x}^{\mathbf{u}} + \mathbf{x}_\eta, \qquad \mathbf{y}[k] = C(\mathbf{p}[k])\mathbf{x}^{\mathbf{u}}[k] + D(\mathbf{p}[k])\mathbf{u}[k] + \boldsymbol{\eta}[k],$$

where $\boldsymbol{\eta}[k] = C(\mathbf{p}[k])\mathbf{x}_\eta[k] + \mathbf{v}[k]$ and

$$|\boldsymbol{\eta}[k]| \le \frac{K_C K_w^P}{1-\rho} + K_v.$$

All bounds hold almost surely for every k. This proves item 1.

Finite sums over words of equal length. For $w = w_0 \cdots w_{m-1}$, define

$$\mathbf{p}_w[k] = \prod_{r=0}^{m-1} p_{w_r}[k-r], \qquad \mathbf{F}_m[k] = \sum_{w\in Q^m} M_w \mathbf{p}_w[k],$$

and

$$\mathbf{S}_m[k] = \mathbf{F}_m[k]\mathbf{u}[k-m+1] = \sum_{w\in Q^m} M_w \mathbf{z}_w^{\mathbf{u}}[k].$$

With the first letter corresponding to the most recent time, the distributive law gives

$$\begin{aligned}
\mathbf{F}_1[k] &= \sum_{i\in Q} D_i p_i[k] = D(\mathbf{p}[k]),\\
\mathbf{F}_m[k] &= \sum_{w_0,\dots,w_{m-1}\in Q} C_{w_0} A_{w_1} \cdots A_{w_{m-2}} B_{w_{m-1}} \prod_{r=0}^{m-1} p_{w_r}[k-r]\\
&= C(\mathbf{p}[k])\Pi_{m-2}[k]B(\mathbf{p}[k-m+1]) \qquad (m \ge 2).
\end{aligned} \tag{36}$$

The product of A matrices is the identity for $m = 2$. Each inner sum is finite; no infinite series has been rearranged. The realized-product bound therefore yields

$$\|\mathbf{F}_1[k]\|_2 \le K_D, \qquad \|\mathbf{F}_m[k]\|_2 \le K_C \|\Pi_{m-2}[k]\|_P K_B \le K_C K_B \rho^{m-2} \quad (m \ge 2). \tag{37}$$

Together with $\|\mathbf{u}[k-m+1]\|_2 \le K_u$, this gives

$$|\mathbf{S}_1[k]| \le K_D K_u, \qquad |\mathbf{S}_m[k]| \le K_C K_B K_u \rho^{m-2} \quad (m \ge 2).$$

Stationary expansion and grouped tail. The series

$$\sum_{j=1}^{\infty} \Pi_{j-1}[k] B(\mathbf{p}[k-j]) \mathbf{u}[k-j]$$

converges absolutely and uniformly in the P-norm, since its j-th term is bounded by $K_B K_u \rho^{j-1}$. Shifting the series by one time step shows that its sum solves the noise-free recursion. By uniqueness, its sum equals $\mathbf{x}^{\mathbf{u}}[k]$. Multiplying by $C(\mathbf{p}[k])$, adding $D(\mathbf{p}[k])\mathbf{u}[k] + \boldsymbol{\eta}[k]$, and using (36) gives

$$\mathbf{y}[k] = \sum_{m=1}^{\infty} \mathbf{S}_m[k] + \boldsymbol{\eta}[k] = \sum_{m=1}^{\infty} \left(\sum_{w \in Q^m} M_w \mathbf{z}_w^{\mathbf{u}}[k] \right) + \boldsymbol{\eta}[k].$$

In particular,

$$\sum_{m=1}^{\infty} |\mathbf{S}_m[k]| \le K_D K_u + \frac{K_C K_B K_u}{1-\rho}.$$

This is absolute convergence of the series of length groups, not absolute summability of the individual word contributions. The geometric majorant is independent of time and of the bounded admissible trajectory, so convergence is uniform over these trajectories.

For fixed $L \ge 1$, the finite sum $\sum_{m=1}^{L} \mathbf{S}_m[k]$ is exactly $\Theta_L \mathbf{Z}_L[k]$. Thus (34) holds with the tail defined in (33), and

$$\begin{aligned} |\mathbf{R}_L[k]| &\le \sum_{m=L+1}^{\infty} |\mathbf{S}_m[k]| \\ &\le K_C K_B K_u \sum_{m=L+1}^{\infty} \rho^{m-2} = \frac{K_C K_B K_u \rho^{L-1}}{1-\rho} =: \tau_L^{(0)}. \end{aligned} \tag{38}$$

The triangle inequality is applied only to the outer sum over lengths. This proves geometric decay of the remainder uniformly over time and bounded admissible trajectories. It also gives

$$\mathbb{E}|\mathbf{R}_L[k]|^2 \le \left(\tau_L^{(0)}\right)^2 \longrightarrow 0,$$

which proves mean-square convergence and completes item 2. □

Moment processes. For $w \in Q^m$ and $\ell \in \{1, \dots, n_u\}$, define

$$\mathbf{X}_{w,\ell}[k] = \mathbf{y}[k] \left(\mathbf{z}_w^{\mathbf{u}}[k]\right)^{(\ell)}.$$

Let a_r^e be common coordinate-wise CBS coefficients for the joint data generator, and set $a_s^e = 0$ for $s < 0$. Lemma B.4, applied to the class-$\mathcal{S}$ generator of Assumption 4.4, allows coefficients with

$$A_0^e = \sum_{r \ge 0} a_r^e < \infty, \qquad A_1^e = \sum_{r \ge 0} r\, a_r^e < \infty.$$

Use the same innovation norm throughout and write $B_\xi = \operatorname{ess\,sup} \|\boldsymbol{\xi}[0]\| < \infty$.

Lemma B.7 (Growing-word CBS bound)**.** *Under Assumptions 4.4 and 4.1, there is a finite constant K_{CBS}, independent of L, N, δ, such that, for every $L \ge 1$ and every $\mathbf{X}_{w,\ell}$ with $|w| \le L$, $|\mathbf{X}_{w,\ell}[k]| \le K_y K_u$ and*

$$\widetilde{a}(H_{w,\ell}) \le K_{\mathrm{CBS}} (1+L)^2. \tag{39}$$

Consequently its weak-dependence coefficient is bounded by

$$L_\theta(L) = 2 B_\xi K_{\mathrm{CBS}} (1+L)^2. \tag{40}$$

Proof. Fix $w = w_0 \cdots w_{m-1}$. Telescoping the product in (7), using $|p_i| \le 1$ and $\|u\|_2 \le K_u$, gives the following admissible coefficient for every scalar coordinate of $\mathbf{z}_w^{\mathbf{u}}[k]$:

$$a_r^{z,w} := a_{r-m+1}^e + K_u \sum_{j=0}^{m-1} a_{r-j}^e. \tag{41}$$

Therefore

$$a(H_{z,w}) \le A_0^e + K_u m A_0^e, \tag{42}$$

$$\begin{aligned}\widetilde{a}(H_{z,w}) &\le A_1^e + (m-1)A_0^e + K_u \left(m A_1^e + \frac{m(m-1)}{2} A_0^e \right) \\ &\le K_z (1+m)^2, \qquad K_z := (1+K_u)(A_0^e + A_1^e).\end{aligned} \tag{43}$$

The proof of Lemma B.6 establishes that the stable LPV recursion, on its invariant bounded state set, is of class $\mathcal{S}$. The joint data generator followed by that recursion is therefore a class-$\mathcal{S}$ cascade by Lemma B.5. Lemma B.4 gives CBS coefficients a_r^y for $\mathbf{y}$ with

$$A_0^y := \sum_{r \ge 0} a_r^y < \infty, \qquad A_1^y := \sum_{r \ge 0} r\, a_r^y < \infty.$$

More precisely, if $(\rho_y, L_e^y, L_{h,s}^y, L_{h,e}^y)$ are class-$\mathcal{S}$ constants of the cascade, one may take

$$a_0^y = L_{h,e}^y, \qquad a_r^y = L_{h,s}^y L_e^y \rho_y^{r-1} \quad (r \ge 1), \qquad A_1^y = \frac{L_{h,s}^y L_e^y}{(1-\rho_y)^2}.$$

For bounded CBS maps H_1, H_2,

$$|H_1 H_2 - H_1' H_2'| \le |H_2||H_1 - H_1'| + |H_1'||H_2 - H_2'|.$$

Apply this inequality with $H_1 = H_y$, $H_2 = H_{z,w}$, $|H_y| \le K_y$ and $|H_{z,w}| \le K_u$. An admissible product coefficient is $K_u a_r^y + K_y a_r^{z,w}$, so

$$\widetilde{a}(H_{w,\ell}) \le K_u A_1^y + K_y K_z (1+m)^2 \le K_{\mathrm{CBS}} (1+m)^2$$

with the choice

$$K_{\mathrm{CBS}} := K_u A_1^y + K_y K_z, \qquad K_z = (1+K_u)(A_0^e + A_1^e). \tag{44}$$

This proves (39) for $m \le L$, while

$$|\mathbf{X}_{w,\ell}[k]| \le |\mathbf{y}[k]|\, \|\mathbf{z}_w^{\mathbf{u}}[k]\|_2 \le K_y K_u$$

gives the magnitude bound. Proposition 4.2 of Alquier & Wintenberger (2012) yields the weak-dependence bound (40). □

B.2.2 Moment inversion and Ho–Kalman perturbation

Define the output–regressor moment and the tail moment by

$$\Lambda_L = \mathbb{E}[\mathbf{y}[k]\mathbf{Z}_L[k]^T], \qquad B_L = \mathbb{E}[\mathbf{R}_L[k]\mathbf{Z}_L[k]^T]. \tag{45}$$

The next lemma relates these moments to Θ_L.

Lemma B.8 (Truncated joint moment equation)**.** *Under Assumptions 4.4, 4.5, 4.6, and 4.1, for every $L \ge 1$,*

$$\Lambda_L = \Theta_L G_L + B_L, \tag{46}$$

and

$$\|B_L\|_{\max} \le \tau_L := \frac{K_C K_B K_u^2\, \rho^{L-1}}{1-\rho}. \tag{47}$$

Consequently,

$$\Lambda_L G_L^{-1} - \Theta_L = B_L G_L^{-1}. \tag{48}$$

Proof of Lemma B.8. Lemma B.6 gives, for every $L \geq 1$, the exact identity

$$\mathbf{y}[k] = \Theta_L \mathbf{Z}_L[k] + \mathbf{R}_L[k] + \boldsymbol{\eta}[k]. \tag{49}$$

All three terms are bounded for fixed L. Multiply by $\mathbf{Z}_L[k]^T$ and take expectations. The last term vanishes by Assumption 4.6, giving $\Lambda_L = \Theta_L G_L + B_L$. Fix a retained word $v \in \mathcal{V}_L$ and input coordinate $\ell \in \{1, \ldots, n_u\}$. Write $(B_L)_{v,\ell}$ for the corresponding entry of this row vector. Since $|(\mathbf{z}_v^{\mathbf{u}}[k])^{(\ell)}| \leq K_u$, the deterministic bound in (37) makes the following series integrably dominated. Therefore

$$\begin{aligned} |(B_L)_{v,\ell}| &= \left| \mathbb{E}\left[\sum_{m=L+1}^{\infty} \left(\sum_{w \in Q^m} M_w \mathbf{z}_w^{\mathbf{u}}[k] \right) (\mathbf{z}_v^{\mathbf{u}}[k])^{(\ell)} \right] \right| \\ &\leq \sum_{m=L+1}^{\infty} \mathbb{E}\left[\left| \sum_{w \in Q^m} M_w \mathbf{z}_w^{\mathbf{u}}[k] \right| \left| (\mathbf{z}_v^{\mathbf{u}}[k])^{(\ell)} \right| \right] \\ &\leq K_C K_B K_u^2 \sum_{m=L+1}^{\infty} \rho^{m-2} = \frac{K_C K_B K_u^2 \rho^{L-1}}{1-\rho} = \tau_L. \end{aligned} \tag{50}$$

The retained factor $(\mathbf{z}_v^{\mathbf{u}}[k])^{(\ell)}$ is common to every term of the finite inner sum. Extracting it uses no independence, and no cross-length moment is set to zero. Taking the maximum over (v, ℓ) proves (47), without a factor d_L. Assumption 4.5 gives $G_L \succ 0$, so right multiplication by G_L^{-1} proves (48). □

Lemma B.9 (CBS concentration consequence)**.** *If $\mathbf{X}[k]$ is a stationary scalar CBS, $|\mathbf{X}[k]| \leq B$, and its weak-dependence coefficient is at most L_θ, then*

$$\mathbb{P}\left[\left| n^{-1} \sum_{k=1}^{n} \mathbf{X}[k] - \mathbb{E}\mathbf{X}[0] \right| > \varepsilon \right] \leq 2 \exp\left[-\frac{n\varepsilon^2}{2(B + L_\theta)^2} \right].$$

Proof. Proposition 4.2 of Alquier & Wintenberger (2012) converts the CBS coefficients to its weak-dependence coefficient. Apply Theorem 6.6 of that paper to the one-Lipschitz maps $\sum_{k=1}^{n} x_k$ and $-\sum_{k=1}^{n} x_k$, followed by the Chernoff bound and optimization of the exponential parameter. □

Empirical-moment concentration. Lemmas B.7 and B.9 give, uniformly over $|w| \leq L$ and input coordinates ℓ,

$$\mathbb{P}\left[\left| \frac{1}{n} \sum_{k=1}^{n} \mathbf{X}_{w,\ell}[k] - \mathbb{E}\mathbf{X}_{w,\ell}[0] \right| > \varepsilon \right] \leq 2 \exp\left[-\frac{n\varepsilon^2}{2(K_y K_u + L_\theta(L))^2} \right]. \tag{51}$$

For $1 \leq L \leq N+1$, define $n_{N,L} = N - L + 2$ and

$$\varepsilon_C(L, \delta, N) = \big(K_y K_u + L_\theta(L)\big) \sqrt{\frac{2\log(2d_L/\delta)}{n_{N,L}}}. \tag{52}$$

Stationarity allows (51) to be applied to the $n_{N,L}$ observations indexed by $k = L-1, \ldots, N$. A union bound over the d_L scalar moment coordinates gives

$$\mathbb{P}\left[\|\widehat{\Lambda}_{L,N} - \Lambda_L\|_{\max} \leq \varepsilon_C(L, \delta, N) \right] \geq 1 - \delta.$$

No independence between the moment coordinates is needed.

Lemma B.10 (Markov-parameter estimation error)**.** *Under Assumptions 4.4–4.1, for every $1 \leq L \leq N+1$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left[\max_{w \in \mathcal{V}_L} \|\widehat{M}_w - M_w\|_F \leq K_{\mathrm{JG}}(L, \delta, N) \right] \geq 1 - \delta, \tag{53}$$

where K_{JG} is defined in (19)*.*

Proof. On the preceding empirical-moment event, Lemma B.8 gives

$$\widehat{\Theta}_{L,N} - \Theta_L = \big(\widehat{\Lambda}_{L,N} - \Lambda_L + B_L\big)G_L^{-1}.$$

For a row vector a, $\|aG_L^{-1}\|_\infty \le \|a\|_\infty \|G_L^{-1}\|_1$; therefore

$$\|\widehat{\Theta}_{L,N} - \Theta_L\|_{\max} \le K_G e^{bL}\big(\varepsilon_C(L,\delta,N) + \tau_L\big).$$

Each Markov parameter has n_u scalar entries. Conversion to the Frobenius norm multiplies this bound by $\sqrt{n_u}$, and

$$\sqrt{n_u}\, K_G e^{bL}\big(\varepsilon_C(L,\delta,N) + \tau_L\big) = K_{\mathrm{JG}}(L,\delta,N).$$

This proves (53). □

Corollary B.11 (Rate for the required Markov parameters)**.** *Under the preceding assumptions, with L_N and ζ from* (21) *and* (20)*, for each fixed $\delta \in (0,1)$,*

$$K_{\mathrm{JG}}(L_N,\delta,N) = \widetilde{O}(N^{-\zeta}) \qquad (N \to \infty).$$

The suppressed factor is polynomial in $\log N$*. Its constants may depend on the fixed dimensions, on $K_y, K_u, B_\xi, K_{\mathrm{CBS}}$, on $K_B K_C, K_G, b, \rho$, and on* $\log(1/\delta)$*, but not on N.*

Proof. For all sufficiently large N,

$$L_N = \left\lfloor \frac{\log N}{2\lambda_\rho} \right\rfloor, \qquad n_{N,L_N} = N - L_N + 2 \ge N/2.$$

Since $d_L = n_u \sum_{m=1}^{L}(n_p+1)^m$, we have $\log d_{L_N} = O(\log N)$; this also holds when $n_p = 0$, in which case $d_L = n_u L$. Furthermore,

$$L_\theta(L_N) = O((1+\log N)^2).$$

For each fixed δ, the sampling term consequently satisfies

$$e^{bL_N}\varepsilon_C(L_N,\delta,N) = O\left((1+\log N)^{5/2} N^{-1/2+b/(2\lambda_\rho)}\right) = \widetilde{O}(N^{-\zeta}).$$

For the bias term, $b < \lambda_\rho$ and the floor in L_N give

$$e^{bL_N}\rho^{L_N-1} = \rho^{-1}e^{-(\lambda_\rho - b)L_N} \le \rho^{-1}e^{\lambda_\rho - b}N^{-(\lambda_\rho-b)/(2\lambda_\rho)} = O(N^{-\zeta}).$$

Substitution in (19) proves the claim. □

Proof of Theorem 5.1. The preceding results provide the empirical-moment concentration, moment inversion and rate calculation. It remains to pass the Markov-parameter error through the Ho–Kalman reconstruction. This step follows the argument of (Rácz et al., 2025a, Lemma 6 and Theorem 2).

Apply Lemma B.10 at $L = L_N$. Since $L_N \ge 2n_x + 1$, every Markov parameter required by the selected Hankel blocks is covered. On the event in that lemma, which has probability at least $1-\delta$, each such parameter has Frobenius error at most $K_{\mathrm{JG}}(L_N,\delta,N)$. Each scalar entry error is therefore at most the same quantity. The largest selected block has at most $n_x \max\{n_x, n_u\} = c_H^2$ entries, so, simultaneously for $H_{\alpha,\beta}$, $H_{\alpha,q,\beta}$, $H_{\alpha,q}$ and $H_{\beta,q}$,

$$\|\widehat{H} - H\|_F \le \Delta_H := c_H K_{\mathrm{JG}}(L_N,\delta,N).$$

By Weyl's inequality and (22),

$$\sigma_{\min}(\widehat{H}_{\alpha,\beta}) \ge \sigma_H - \Delta_H \ge \sigma_H/2.$$

Thus the algorithm uses the inverse formulas rather than returning its default model, and

$$\|\widehat{H}_{\alpha,\beta}^{-1}\|_2 \le \frac{2}{\sigma_H}.$$

The inverse identity

$$\widehat{H}_{\alpha,\beta}^{-1} - H_{\alpha,\beta}^{-1} = \widehat{H}_{\alpha,\beta}^{-1}(H_{\alpha,\beta} - \widehat{H}_{\alpha,\beta})H_{\alpha,\beta}^{-1}$$

therefore gives

$$\|\widehat{H}_{\alpha,\beta}^{-1} - H_{\alpha,\beta}^{-1}\|_F \le \frac{2\Delta_H}{\sigma_H^2}.$$

For example,

$$\begin{aligned}\widehat{A}_q - \bar{A}_q = \widehat{H}_{\alpha,\beta}^{-1}(\widehat{H}_{\alpha,q,\beta} - H_{\alpha,q,\beta}) \\ + (\widehat{H}_{\alpha,\beta}^{-1} - H_{\alpha,\beta}^{-1})H_{\alpha,q,\beta}.\end{aligned}$$

Using the definition of K_H and the corresponding identity for B_q yields, for every $q \in Q$,

$$\begin{aligned}\|\widehat{A}_q - \bar{A}_q\|_F,\ \|\widehat{B}_q - \bar{B}_q\|_F \le \left(\frac{2}{\sigma_H} + \frac{2K_H}{\sigma_H^2}\right)\Delta_H, \\ \|\widehat{C}_q - \bar{C}_q\|_F \le \Delta_H, \\ \|\widehat{D}_q - \bar{D}_q\|_F \le K_{\mathrm{JG}}(L_N, \delta, N) \le \Delta_H.\end{aligned}$$

The last inequality uses $c_H \ge 1$. Taking the maximum proves (23). Corollary B.11 gives the claimed rate. For fixed δ, $K_{\mathrm{JG}}(L_N, \delta, N) \to 0$, so the smallness condition is satisfied for all sufficiently large N. □

C MULTI-OUTPUT LPV SYSTEMS

We give the details of the multi-output extension stated after Theorem 5.1. As in the switched-system argument, concentration is first applied for a fixed word, input index and output index, followed by a union bound. The Ho–Kalman construction then uses output-indexed row selections.

C.1 DIMENSIONS, ASSUMPTIONS AND MOMENT BOUNDS

Consider (4) with

$$\mathbf{y}[k], \mathbf{v}[k] \in \mathbb{R}^{n_y}, \qquad C_q \in \mathbb{R}^{n_y \times n_x}, \quad D_q \in \mathbb{R}^{n_y \times n_u}, \qquad n_y \ge 1.$$

Use Assumptions 4.4–4.2 with these dimensions: $\|\mathbf{v}[k]\|_2 \le K_v$, the noise response is vector valued, and the orthogonality condition is the matrix equality $\mathbb{E}[\boldsymbol{\eta}[k]\mathbf{Z}_L[k]^T] = 0$. Minimality concerns reachability and observability with all outputs. No independence between output coordinates is assumed.

The definitions of $\mathcal{V}_L, d_L, \mathbf{Z}_L, G_L, K_B, K_C, K_D$ are unchanged, with $\|D_q\|_2$ now the matrix spectral norm and $\|C_q\|_{P\to 2}$ having codomain $\mathbb{R}^{n_y}$. The state bound K_x of Lemma B.6 is unchanged and

$$\|\mathbf{y}[k]\|_2 \le K_y := K_C K_x + K_D K_u + K_v.$$

The Markov parameters in (6) now belong to $\mathbb{R}^{n_y \times n_u}$, while $\Theta_L, \Lambda_L, B_L, \widehat{\Lambda}_{L,N}$ belong to $\mathbb{R}^{n_y \times d_L}$. The grouped expansion and moment identity hold with the same formulas. Indeed, the proof of Lemma B.6 bounds the Euclidean norm of each vector length group by $K_C K_B K_u \rho^{m-2}$ for $m \ge 2$, giving

$$\|\mathbf{R}_L[k]\|_2 \le \frac{K_C K_B K_u \rho^{L-1}}{1-\rho}, \qquad \Lambda_L = \Theta_L G_L + B_L, \qquad \|B_L\|_{\max} \le \tau_L = \frac{K_C K_B K_u^2 \rho^{L-1}}{1-\rho}. \tag{54}$$

Thus the state, tail and population moment arguments require only the dimension changes above.

Choose common CBS coefficients a_j^y that bound every coordinate of $\mathbf{y}$; coefficients for the vector output in the Euclidean norm suffice. Write $A_1^y = \sum_{j\ge 0} j a_j^y$, and use

$$K_z = (1+K_u)(A_0^e + A_1^e), \qquad K_{\mathrm{CBS}} = K_u A_1^y + K_y K_z, \qquad L_\theta(L) = 2B_\xi K_{\mathrm{CBS}}(1+L)^2. \tag{55}$$

The cascade argument and Lemma B.7 give these bounds for every scalar process

$$\mathbf{X}_{w,l,r}[k] = \mathbf{y}^{(r)}[k](\mathbf{z}_w^{\mathbf{u}}[k])^{(l)}, \qquad w \in \mathcal{V}_L, \quad 1 \le l \le n_u, \quad 1 \le r \le n_y.$$

There are $n_y d_L = n_y n_u |\mathcal{V}_L|$ such processes. Define

$$\varepsilon_C^{\mathrm{mo}}(L,\delta,N) = (K_y K_u + L_\theta(L))\sqrt{\frac{2\log(2n_y d_L/\delta)}{n_{N,L}}}, \tag{56}$$

$$K_{\mathrm{JG}}^{\mathrm{mo}}(L,\delta,N) = \sqrt{n_y n_u}\, K_G e^{bL}\big(\varepsilon_C^{\mathrm{mo}}(L,\delta,N) + \tau_L\big). \tag{57}$$

The superscript mo distinguishes these multi-output constants from the scalar-output constants of the main text.

Lemma C.1 (Multi-output Markov-parameter bound)**.** *Under the vector-output assumptions above, for $1 \le L \le N+1$ and $\delta \in (0,1)$, the population-Gram estimator $\widehat{\Theta}_{L,N} = \widehat{\Lambda}_{L,N} G_L^{-1}$ satisfies*

$$\mathbb{P}\left[\max_{w\in\mathcal{V}_L} \|\widehat{M}_w - M_w\|_F \le K_{\mathrm{JG}}^{\mathrm{mo}}(L,\delta,N)\right] \ge 1-\delta. \tag{58}$$

Proof. Lemma B.9 gives, for each (w,l,r),

$$\mathbb{P}\left[\left|\frac{1}{n_{N,L}}\sum_{k=L-1}^{N} \mathbf{X}_{w,l,r}[k] - \mathbb{E}\mathbf{X}_{w,l,r}[0]\right| > \varepsilon_C^{\mathrm{mo}}\right] \le \frac{\delta}{n_y d_L}.$$

A union bound gives $\|\widehat{\Lambda}_{L,N} - \Lambda_L\|_{\max} \le \varepsilon_C^{\mathrm{mo}}$ with probability at least $1-\delta$. On this event,

$$\widehat{\Theta}_{L,N} - \Theta_L = (\widehat{\Lambda}_{L,N} - \Lambda_L + B_L)G_L^{-1}, \qquad \|\widehat{\Theta}_{L,N} - \Theta_L\|_{\max} \le K_G e^{bL}(\varepsilon_C^{\mathrm{mo}} + \tau_L).$$

The matrix inequality follows by applying $\|aF\|_\infty \le \|a\|_\infty \|F\|_1$ to each row a. Each Markov block has $n_y n_u$ entries, so conversion to Frobenius norm proves (58). □

C.2 Joint Ho–Kalman realization

Extend the row selection by an output index:

$$\alpha = \{(\nu_i, \alpha_i, r_i)\}_{i=1}^{n_x}, \qquad \beta = \{(\mu_j, q_j, l_j)\}_{j=1}^{n_x},$$

where $|\nu_i|, |\mu_j| \le n_x - 1$, $\alpha_i, q_j \in Q$, $1 \le r_i \le n_y$, and $1 \le l_j \le n_u$. For $1 \le i,j \le n_x$, $1 \le l \le n_u$, $1 \le r \le n_y$, put

$$\begin{aligned} H_{\alpha,\beta}^{(i,j)} &= M_{\alpha_i \nu_i \mu_j q_j}^{(r_i,l_j)}, & H_{\alpha,q,\beta}^{(i,j)} &= M_{\alpha_i \nu_i q \mu_j q_j}^{(r_i,l_j)}, \\ H_{\alpha,q}^{(i,l)} &= M_{\alpha_i \nu_i q}^{(r_i,l)}, & H_{\beta,q}^{(r,j)} &= M_{q\mu_j q_j}^{(r,l_j)}. \end{aligned} \tag{59}$$

Their dimensions are $n_x \times n_x$, $n_x \times n_x$, $n_x \times n_u$ and $n_y \times n_x$, respectively. All required words still have length at most $L_{\min} = 2n_x + 1$.

Let O_α have row i equal to row r_i of $C_{\alpha_i} A_{\nu_i}$, and let R_β have column j equal to column l_j of $A_{\mu_j} B_{q_j}$. Then

$$\begin{aligned} H_{\alpha,\beta} &= O_\alpha R_\beta, \quad H_{\alpha,q,\beta} = O_\alpha A_q R_\beta, \\ H_{\alpha,q} &= O_\alpha B_q, \qquad H_{\beta,q} = C_q R_\beta. \end{aligned} \tag{60}$$

Joint minimality gives selections for which both square factors are invertible. The population formulas (11) therefore give

$$\bar{A}_q = R_\beta^{-1} A_q R_\beta, \quad \bar{B}_q = R_\beta^{-1} B_q, \quad \bar{C}_q = C_q R_\beta, \quad \bar{D}_q = D_q.$$

The empirical formulas and default-output convention are those of Section 3, with the block dimensions in (59). All outputs share this realization and its state coordinates. No individual scalar-output subsystem is required to be observable or minimal.

Set

$$\begin{aligned} c_H^{\mathrm{mo}} &= \sqrt{n_x \max\{n_x, n_u, n_y\}}, \qquad \sigma_H = \sigma_{\min}(H_{\alpha,\beta}), \\ K_H &= \max_{q\in Q}\{1, \|H_{\alpha,q,\beta}\|_2, \|H_{\alpha,q}\|_2\}, \\ K_{\mathrm{HK}}^{\mathrm{mo}} &= c_H^{\mathrm{mo}} \max\left\{1, \frac{2}{\sigma_H} + \frac{2K_H}{\sigma_H^2}\right\}. \end{aligned} \tag{61}$$

Theorem C.2 (Multi-output finite-sample realization bound)**.** *Under the vector-output versions of Assumptions 4.4–4.2, fix $\delta \in (0,1)$, valid supplied selections (α,β) and a deterministic horizon $L_{\min} \le L \le N+1$. If*

$$c_H^{\mathrm{mo}} K_{\mathrm{JG}}^{\mathrm{mo}}(L,\delta,N) \le \sigma_H/2,$$

then the population-Gram learner satisfies

$$\mathbb{P}\left[\mathrm{EstErr} := \max_{\substack{q\in Q \\ X\in\{A,B,C,D\}}} \|\widehat{X}_q - \bar{X}_q\|_F \le K_{\mathrm{HK}}^{\mathrm{mo}} K_{\mathrm{JG}}^{\mathrm{mo}}(L,\delta,N)\right] \ge 1-\delta. \tag{62}$$

Proof. On the event of Lemma C.1, each scalar Markov entry has error at most $K_{\mathrm{JG}}^{\mathrm{mo}}$. Every Hankel block in (59) has at most $(c_H^{\mathrm{mo}})^2$ entries, hence its Frobenius error is at most $\Delta_H = c_H^{\mathrm{mo}} K_{\mathrm{JG}}^{\mathrm{mo}}$. Weyl's inequality and the inverse identity give

$$\sigma_{\min}(\widehat{H}_{\alpha,\beta}) \ge \sigma_H/2, \qquad \|\widehat{H}_{\alpha,\beta}^{-1} - H_{\alpha,\beta}^{-1}\|_F \le 2\Delta_H/\sigma_H^2.$$

The inverse formulas are therefore used. Subtracting the population products from the empirical products yields

$$\begin{aligned} \|\widehat{A}_q - \bar{A}_q\|_F, \quad \|\widehat{B}_q - \bar{B}_q\|_F &\le (2/\sigma_H + 2K_H/\sigma_H^2)\Delta_H, \\ \|\widehat{C}_q - \bar{C}_q\|_F \le \Delta_H, \qquad \|\widehat{D}_q - \bar{D}_q\|_F &\le K_{\mathrm{JG}}^{\mathrm{mo}}. \end{aligned}$$

These bounds give (62) because $c_H^{\mathrm{mo}} \ge 1$. □

D MAMBA VARIANTS AS LPV SYSTEMS

We give an exact representation of the selective core and two approximations. The constructions specify the affine features and verify stability; applying the identification theorem also requires excitation for the resulting regressors.

D.1 THE SELECTIVE MAMBA CORE

Fix a channel j. The vector $\mathbf{u}[k] \in \mathbb{R}^{n_u}$ is the observed input to the core, after any preceding processing; $\mathbf{x}[k] \in \mathbb{R}^{n_x}$ and the scalar $\mathbf{y}[k]$ denote this channel's state and output. Starting from the continuous-time system

$$\dot{x}(t) = \widetilde{A}x(t) + \widetilde{B}u^{(j)}(t), \qquad y(t) = \widetilde{C}x(t), \tag{63}$$

the Mamba selection mechanism (Gu & Dao, 2024; Honarpisheh et al., 2025) uses $W_B, W_C \in \mathbb{R}^{n_x \times n_u}$ and

$$\Delta[k] = \mathrm{softplus}(a + q^T\mathbf{u}[k]), \qquad \overline{C}[k] = \mathbf{u}[k]^T W_C^T, \tag{64}$$

$$(\overline{A}[k], \overline{B}[k]) = \mathcal{T}(\widetilde{A}, W_B\mathbf{u}[k], \Delta[k]). \tag{65}$$

The fixed channel label is suppressed on these quantities. The two discretization rules considered here are

$$\begin{aligned} \mathcal{T}_{\mathrm{ZOH}}(F,G,h) &= \left(e^{hF}, \int_0^h e^{sF}G\,ds\right), \\ \mathcal{T}_{\mathrm{simp}}(F,G,h) &= (e^{hF}, hG). \end{aligned} \tag{66}$$

For the real diagonal Mamba-1 core, $\widetilde{A} = \mathrm{diag}(a_1,\dots,a_{n_x})$, $a_i < 0$. The implementation uses $\mathcal{T}_{\mathrm{simp}}$, evaluates the diagonal transition entries by scalar exponentials, and reads the updated state:[2]

$$\begin{aligned} \mathbf{x}[k+1] &= \overline{A}[k]\mathbf{x}[k] + \overline{B}[k]\mathbf{u}^{(j)}[k], \\ \mathbf{y}[k] &= \overline{C}[k]\mathbf{x}[k+1] + D_j\mathbf{u}^{(j)}[k]. \end{aligned} \tag{67}$$

[2]The reference scan `selective_scan_ref` and `mamba_simple.py`, commit `e9594ce1`, checked 23 September 2026.

Here $\overline{B}[k]$ is a column and D_j is the scalar skip coefficient. We study the mathematical recurrence with smooth softplus; floating-point errors and the surrounding projections and SiLU branch are outside this core. The polynomial examples below are explicit model approximations.

For identification, use the scalar drive $\mathbf{u}^{(j)}$ as the LPV input; the full observed vector supplies the scheduling features. The pre-update LPV convention is recovered by substitution:

$$\begin{aligned} A(\mathbf{p}[k]) &= \overline{A}[k], & B(\mathbf{p}[k]) &= \overline{B}[k], \\ C(\mathbf{p}[k]) &= \overline{C}[k]\overline{A}[k], & D(\mathbf{p}[k]) &= D_j + \overline{C}[k]\overline{B}[k]. \end{aligned} \tag{68}$$

This also specifies the readout and direct-term features needed below. Working with a scalar drive avoids the forced symmetry $p_\ell \mathbf{u}^{(\ell')} = p_{\ell'} \mathbf{u}^{(\ell)}$ of a vector-input lift with $p_\ell = \mathbf{u}^{(\ell)}$. For several channels, apply the theorem to each channel satisfying its assumptions with confidence δ/n_u.

D.2 Finite features and the Cayley–Hamilton construction

Section D.3 gives the construction for fixed $\widetilde{B}, \widetilde{C}$; here we include the selected input and readout. Let $W_{B,\ell}$ be column ℓ of W_B and $W_{C,m}$ row m of W_C^T. For a general fixed $\widetilde{A}$, Cayley–Hamilton gives

$$\begin{aligned} e^{h\widetilde{A}} &= \sum_{r=0}^{n_x-1} \varphi_r(h;\widetilde{A})\widetilde{A}^r, \\ \Phi(h\widetilde{A}) &= \sum_{r=0}^{n_x-1} \psi_r(h;\widetilde{A})\widetilde{A}^r, \qquad \Phi(Z) = \sum_{m=0}^{\infty} \frac{Z^m}{(m+1)!}. \end{aligned} \tag{69}$$

For ZOH, define the intermediate features

$$p_r^A = \varphi_r(\Delta;\widetilde{A}), \qquad p_{r,\ell}^B = \psi_r(\Delta;\widetilde{A})\Delta u^{(\ell)}, \qquad p_m^C = u^{(m)}. \tag{70}$$

Thus $\overline{A} = \sum_r p_r^A \widetilde{A}^r$, $\overline{B} = \sum_{r,\ell} p_{r,\ell}^B \widetilde{A}^r W_{B,\ell}$, and $\overline{C} = \sum_m p_m^C W_{C,m}$. Together with (68), a scalar basis of

$$\operatorname{span}\{1,\ p_r^A,\ p_{r,\ell}^B,\ p_m^C p_r^A,\ p_m^C p_{r,\ell}^B\} \tag{71}$$

therefore makes all four LPV maps affine. The sums run over $0 \le r < n_x$ and $1 \le \ell, m \le n_u$. Retain $p_0 = 1$, remove linearly redundant functions and normalize the remaining bounded features as in Section 2.2; Section D.4 shows why the unreduced ZOH features can give a singular Gram matrix even with scalar input. The chosen basis and its values must be available to the learning algorithm.

For $\mathcal{T}_{\text{simp}}$, replace $p_{r,\ell}^B$ by $p_\ell^B = \Delta u^{(\ell)}$, with coefficient $W_{B,\ell}$. The same construction works for any finite expansion $\overline{A}(\Delta) = \sum_r A_r p_r^A(\Delta)$; the readout products have coefficients $W_{C,m}A_r$, and the direct products have coefficients $W_{C,m}W_{B,\ell}$. These are expansion coefficients before the final feature enumeration. Cayley–Hamilton establishes representability; its features generally depend on $\widetilde{A}$ and must be known or observed to be used for identification.

D.3 ZOH with fixed input and readout matrices

Here $\widetilde{B}$ and $\widetilde{C}$ are constant, unlike the selected matrices in Section D.2. Consider

$$\dot{x}(t) = \widetilde{A}x(t) + \widetilde{B}u(t), \qquad y(t) = \widetilde{C}x(t), \tag{72}$$

where $\widetilde{A} \in \mathbb{R}^{n_x \times n_x}$ is Hurwitz, $\widetilde{B} \in \mathbb{R}^{n_x \times n_u}$ and $\widetilde{C} \in \mathbb{R}^{1 \times n_x}$. Hold the input constant over each interval of length $\Delta[k] > 0$ and read the output at its start. Exact ZOH gives

$$\begin{aligned} \overline{A}[k] &= e^{\Delta[k]\widetilde{A}}, \\ \overline{B}[k] &= \int_0^{\Delta[k]} e^{s\widetilde{A}}\widetilde{B}\, ds = \Phi(\Delta[k]\widetilde{A})\Delta[k]\widetilde{B}, \end{aligned} \tag{73}$$

with Φ defined in (69). The step may depend on the input, but all three continuous-time matrices remain fixed.

Using the Cayley–Hamilton functions of Section D.2,

$$e^{\Delta\widetilde{A}} = \sum_{r=0}^{n_x-1} \varphi_r(\Delta;\widetilde{A})\widetilde{A}^r,$$
$$\Phi(\Delta\widetilde{A})\Delta\widetilde{B} = \sum_{r=0}^{n_x-1} \psi_r(\Delta;\widetilde{A})\Delta\widetilde{A}^r\widetilde{B}. \tag{74}$$

Thus the intermediate scalar features $p_r^A(\Delta) = \varphi_r(\Delta;\widetilde{A})$ and $p_r^B(\Delta) = \psi_r(\Delta;\widetilde{A})\Delta$ give

$$A(p) = \sum_{r=0}^{n_x-1} \widetilde{A}^r p_r^A, \quad B(p) = \sum_{r=0}^{n_x-1} \widetilde{A}^r\widetilde{B}p_r^B,$$
$$C(p) = \widetilde{C}, \qquad D(p) = 0. \tag{75}$$

After choosing a scalar basis containing $p_0 = 1$ and normalizing its nonconstant features as in Section 2.2, this yields

$$\mathbf{x}[k+1] = A(\mathbf{p}[k])\mathbf{x}[k] + B(\mathbf{p}[k])\mathbf{u}[k], \qquad \mathbf{y}[k] = \widetilde{C}\mathbf{x}[k]. \tag{76}$$

This is an exact affine representation, not yet an identification guarantee. In particular, the intermediate features need not form a basis.

D.4 Feature redundancy and theorem applicability

The ZOH input and transition maps are linked. Since $\widetilde{A}$ is invertible,

$$\overline{B}(\Delta) = \widetilde{A}^{-1}(e^{\Delta\widetilde{A}} - I)\widetilde{B} = -\widetilde{A}^{-1}\widetilde{B} + \sum_{r=0}^{n_x-1} p_r^A(\Delta)\widetilde{A}^{-1}\widetilde{A}^r\widetilde{B}. \tag{77}$$

Consequently, a basis of $\mathrm{span}\{1, p_0^A, \ldots, p_{n_x-1}^A\}$ already suffices for both maps; separate p_r^B features are unnecessary.

For example, when $\widetilde{A} = -a$ with $a > 0$, the unreduced features are

$$p^A = e^{-a\Delta}, \qquad p^B = \frac{1 - e^{-a\Delta}}{a}, \qquad p^A + ap^B = 1. \tag{78}$$

If q_A, q_B denote their scheduling letters, then $\mathbf{z}^{\mathbf{u}}_{q_A}[k] + a\mathbf{z}^{\mathbf{u}}_{q_B}[k] - \mathbf{z}^{\mathbf{u}}_0[k] = 0$. Hence the unreduced G_1 is singular, even for scalar input. Removing redundant features is necessary; an invertible affine reparameterization or a switch to one input coordinate cannot remove this rank defect. A reduced basis does not by itself prove the joint-Gram condition.

Stability, in contrast, follows from Hurwitz stability and $0 < \Delta_{\min} \le \Delta[k] \le \Delta_{\max} < \infty$. Choose $P \succ 0$ and $\mu > 0$ with $\widetilde{A}^T P + P\widetilde{A} \preceq -2\mu P$. Then

$$\|e^{\Delta\widetilde{A}}\|_P \le e^{-\mu\Delta} \le e^{-\mu\Delta_{\min}} < 1,$$

which verifies Assumption 4.1 on the scheduling range. The Cayley–Hamilton functions depend on $\widetilde{A}$; the final basis and its values must be known or observed. Assumptions 4.4, 4.5, 4.6 and 4.2 must still hold for the final representation, with the side information required by Theorem 5.1 supplied.

For the selected input/readout of the Mamba core, use Section D.2, including the product features required by (68); its conditional theorem application is given in Section D.5.

D.5 Examples and applicability of the theorem

The examples keep the selected input, selected readout and skip term in (67). On a bounded input domain, a fixed softplus selector admits bounds $0 < \Delta_{\min} < \Delta_{\max} < \infty$. For the diagonal transition put

$$\rho_* = \max_i e^{a_i\Delta_{\min}} < 1, \qquad \sup_{\Delta\in[\Delta_{\min},\Delta_{\max}]} \|e^{\Delta\widetilde{A}}\|_2 \le \rho_*. \tag{79}$$

Example 1: the exact diagonal core. No transition approximation is needed when its entries are available. Choose known bounds $0 < \underline{a}_i \le e^{a_i\Delta} \le \overline{a}_i < 1$ and set

$$\begin{aligned} c_i &= (\overline{a}_i + \underline{a}_i)/2, \qquad h_i = (\overline{a}_i - \underline{a}_i)/2, \\ p_i^A &= (e^{a_i\Delta} - c_i)/h_i, \quad A(p) = \operatorname{diag}(c_i) + \sum_i h_i E_{ii} p_i^A. \end{aligned} \tag{80}$$

Here E_{ii} has its only nonzero entry, equal to one, at (i,i). Constant entries are absorbed into the constant matrix; dependent features are removed. This is an exact change of basis with $|p_i^A| \le 1$, and (79) proves Assumption 4.1 with $P = I$. It avoids the unscaled coefficients E_{ii} with unit eigenvalue. The feature values require known a_i and Δ, or observed transition entries and step sizes; W_B, W_C, D_j may be unknown. The input/readout products in (71) complete the affine representation.

This construction provides an exact affine representation with available features and a stability certificate. The main theorem will be applicable given the remaining assumptions hold. Note, the price we pay is that the diagonal state matrix has to be known in advance.

Nevertheless, identification of the input and readout matrices only is still a highly nontrivial problem.

Example 2: a centered polynomial approximation. To allow unknown continuous-time transition entries, use features that depend only on the observed step. Let $\Delta_0 = (\Delta_{\min} + \Delta_{\max})/2$, $h = (\Delta_{\max} - \Delta_{\min})/2$, and $p_\Delta = (\Delta - \Delta_0)/h$. For a fixed degree $R \ge 1$, replace $e^{\Delta\widetilde{A}}$ by

$$A_{\mathrm{poly}}(\Delta) = \sum_{r=0}^{R} A_r p_\Delta^r, \qquad A_r = e^{\Delta_0\widetilde{A}} \frac{(h\widetilde{A})^r}{r!}. \tag{81}$$

The powers p_Δ^r are separate scheduling features; their values do not require the unknown a_i. Centering gives $A_0 = e^{\Delta_0\widetilde{A}}$ instead of I. The scalar Taylor remainder yields

$$\sup_\Delta \|e^{\Delta\widetilde{A}} - A_{\mathrm{poly}}(\Delta)\|_2 \le \eta_A := \max_i e^{a_i\Delta_0 + |a_i|h} \frac{(|a_i|h)^{R+1}}{(R+1)!}. \tag{82}$$

If $\rho_* + \eta_A < 1$, Assumption 4.1 holds with $P = I$ and $\rho = \rho_* + \eta_A$. Only the transition is approximated: softplus and both linear selection maps are retained. The complete feature span is generated by $\{1, p_\Delta^r, \Delta u^{(\ell)}, u^{(m)} p_\Delta^r, \Delta u^{(m)} u^{(\ell)}\}$. With known step-selection parameters, or observed steps, these features are available while $\widetilde{A}, W_B, W_C, D_j$ may be unknown. The theorem estimates the resulting affine coefficients up to isomorphism; it does not impose their Taylor factorization.

First-order special case: a recentered Euler transition. For scalar input $u \in [v_1, v_2]$ and a linear positive step $\Delta(u) = qu$, the Euler approximation is $A(u) = I + (qu)\widetilde{A}$. Put $c = (v_1 + v_2)/2$, $h = (v_2 - v_1)/2$ and $p_1 = (c - u)/h$. Then

$$A(u) = A_0 + A_1 p_1, \qquad A_0 = I + qc\widetilde{A}, \quad A_1 = -qh\widetilde{A}. \tag{83}$$

The feature p_1 is known without q or $\widetilde{A}$. If $V_i = I + qv_i\widetilde{A}$ satisfy $\|V_i\|_P \le \rho < 1$ for $i = 1, 2$ with a common P, then $A(u) = \frac{1+p_1}{2}V_1 + \frac{1-p_1}{2}V_2$ verifies Assumption 4.1. This first-order variant uses a linear step selector; Example 2 retains the softplus selector of Mamba.