This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

Real-time Switched System Identification with Online Deterministic Annealing

Christos N. Mavridis*, and Karl Henrik Johansson*

Abstract—**We introduce a real-time identification method for discrete-time state-dependent switching systems in both the input–output and state-space domains. In particular, we design a system of adaptive algorithms running in two timescales; a stochastic approximation algorithm implements an online deterministic annealing scheme at a slow timescale and estimates the mode-switching signal, and a recursive identification algorithm runs at a faster timescale and updates the parameters of the local models based on the estimate of the switching signal. We first focus on piece-wise affine systems and discuss identifiability conditions and convergence properties based on the theory of two-timescale stochastic approximation. In contrast to standard identification algorithms for switched systems, the proposed approach gradually estimates the number of modes and is appropriate for real-time system identification using sequential data acquisition. The progressive nature of the algorithm improves computational efficiency and provides real-time control over the performance-complexity trade-off. Finally, we address specific challenges that arise in the application of the proposed methodology in identification of more general switching systems. Simulation results validate the efficacy of the proposed methodology.**

Index Terms—**Switched System Identification, Piecewise Affine System Identification, Online Deterministic Annealing.**

I. Introduction

Switched systems, described by interacting continuous and discrete dynamics, are a powerful modeling tool in the analysis of systems where logic and continuous processes are interlaced, as in most complex cyber-physical systems. In addition to being able to describe switching dynamics, switched systems can be used as a tool to approximate highly non-linear dynamics by a collection of simpler models, and boost model explainability and robustness, by decomposing the behavior of a complex system into sub-systems where first principles and domain knowledge can be used for precise model tuning [1], [2]. As a result, switched systems have attracted significant attention in the control community.

However, first principles modelling is often too complicated and sub-optimal, and a switched model needs to be identified on the basis of observations. The majority of the work in this area is based on piece-wise affine (PWA) systems, a class of state-dependent switched systems with important applications in identification, verification, and control synthesis of switched and nonlinear systems [2]–[4]. PWA systems are a collection of affine dynamical systems, indexed by a discrete-valued switching variable (mode) that depends on a partitioning of the state-input domain into a finite number of polyhedral regions [2], [3]. The input–output representation of PWA systems is the class of piece-wise affine auto-regressive exogenous (PWARX) systems with the switching signal depending on a partitioning of the domain of a vector containing the recent history of input–output pairs. As the problem of identifying a PWA system can be challenging [5], [6], most existing approaches focus on offline identification methods [7], [8].

A. Contribution and Outline

In this work, we propose a two-timescale stochastic optimization approach for real-time state-dependent switched system identification in both input–output and state-space representations. We first focus on the well-studied case of PWA and PWARX systems. In Section II we present the realization and identifiability conditions for PWA systems, and in Theorem 1 of Section II-B we provide the identifiability conditions for state space PWA systems in the form of a persistence of excitation (PE) criterion. In Section III, we formulate the state-dependent switching system identification problem as a combined identification and prototype-based learning problem and in Sections IV and V we develop a two-timescale stochastic approximation algorithm to solve it in real-time.

In particular, in Section IV we build upon the online deterministic annealing approach [9] to construct a stochastic approximation algorithm that estimates the mode-switching signal, as well as the number of modes, through a bifurcation phenomenon, studied in Section IV-B. In Section V a second stochastic approximation algorithm based on standard adaptive filtering, running at a faster timescale, is developed to update the parameters of the local models based on the estimate of the switching signal. The convergence properties of this system of recursive algorithms are studied in Theorem 4 of Section V-B, and the applicability of the proposed approach in more general state-dependent switching systems is discussed in Section VI. Finally, in Section VII, simulation results validate the efficacy of the proposed approach in PWA systems.

B. Related Work

Most existing switched system identification methods can be categorized by the problem formulation used as optimization-based [7], [10], [11], likelihood-based [12], algebraic [13], or clustering-based [8], [14]–[17], and by the method used as offline [8], [10], [18], [19] or online [20]–[22]. For an extensive review of existing work the readers are referred to [1]–[4], [23], [24] and the references therein.

Algebraic methods are mainly based on transforming a Switched AutoRegressive eXogenous (SARX) model to a "lifted" ARX model that does not depend on the switching sequence [13]. Optimization-based methods rely on solving a

*Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm. emails:{mavridis,kallej}@kth.se.

Research partially supported by the Swedish Foundation for Strategic Research (SSF) grant IPD23-0019.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

large mixed-integer program, which is an NP hard problem that remains tractable only for simple models and small data sets [7], [11]. Therefore, many works focus on relaxation techniques over the same problem [19], [25], that include convexification and expectation-maximization approaches. Finally, clustering-based methods are optimization-based methods that make use of unsupervised learning to estimate the partition of the domain that is needed for the switching signal [8], [10], [14], [16], [17], [26].

Most hybrid identification approaches are offline methods that first classify each observation and estimate the local model parameters (either simultaneously or iteratively), and then reconstruct the partition of the switching signal [8], [10], [14], [16], [25]. In our recent work, we have proposed the use of the online deterministic annealing approach as a clustering method to estimate the partition of the switching signal in real-time [21], [22]. In this work, we extend these methods to provide a complete study of a real-time prototype-based learning algorithm that (i) provides an inherent mechanism to adaptively estimate a set of modes with minimal cardinality, (ii) constitutes a unified switched system identification method for both input–output and state-space representations, and (iii) investigates extensions to more general switching systems.

C. Notation

The sets $\mathbb{R}$ and $\mathbb{Z}$ represent the sets of real and integer numbers, respectively, while $\mathbb{Z}_+$ represents the set of non-negative integers. For a real matrix $A \in \mathbb{R}^{n\times m}$, $A^{\mathrm{T}} \in \mathbb{R}^{m\times n}$ denotes its transpose and $\mathrm{vec}(A) \in \mathbb{R}^{mn}$ the vectorization of A. The $n \times n$ identity matrix is denoted I_n. $A \succeq 0$ is a positive semi-definite matrix, and the condition $A \succeq B$ is understood as $A - B \succeq 0$. Unless otherwise specified, random variables $\mathcal{X} : \Omega \to \mathbb{R}^d$ are defined in a probability space $(\Omega, \mathbb{F}, \mathbb{P})$. The probability of an event is denoted $\mathbb{P}[\mathcal{X} \in S] := \mathbb{P}[\omega \in \Omega : \mathcal{X}(\omega) \in S]$, and the expectation operator $\mathbb{E}[\mathcal{X}] = \int_\Omega \mathcal{X} \mathrm{d}\mathbb{P}$. In case of multiple random variables $(\mathcal{X}, \mathcal{Y})$ and a deterministic function f, the expectation operator $\mathbb{E}[f(\mathcal{X}, \mathcal{Y})]$ is understood with respect to the joint probability measure, while $\mathbb{E}[\mathcal{X}|\mathcal{Y}] := \mathbb{E}[\mathcal{X}|\sigma(\mathcal{Y})]$ denotes the expectation of $\mathcal{X}$ conditioned to the σ-field of $\mathcal{Y}$. Stochastic processes $\{\mathcal{X}(k)\}_k$, $k \in \mathbb{Z}_+$, are defined in the filtered probability space $(\Omega, \mathbb{F}, \{\mathcal{F}_n\}_n, \mathbb{P})$, where $\mathcal{F}_n = \sigma(\mathcal{X}(k)|k \leq n)$, $k \in \mathbb{Z}_+$, is the natural filtration. The normal distribution with mean value μ, and covariance matrix Σ is denoted $\mathcal{N}(\mu, \Sigma)$. The indicator function of the event $[\mathcal{X} \in S]$ is denoted $\mathbb{1}_{[\mathcal{X}\in S]}$ and $\otimes$ denotes the Kronecker product. Finally, "min" (resp. "max") defines the minimization (resp. maximization) operator while "minimize" (resp. "max.") defines a minimization (resp. maximization) problem.

II. Switched and Piecewise Affine Systems

A general discrete-time switched system is described by:

$$\begin{aligned} x_{t+1} &= f_{\sigma_t}(x_t, u_t) + w_t \\ y_t &= g_{\sigma_t}(x_t, u_t) + v_t, \quad t \in \mathbb{Z}_+ \end{aligned} \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the state vector of the system, $u_t \in \mathbb{R}^p$ the input, $y_t \in \mathbb{R}^q$ the output, and $w_t \in \mathbb{R}^n$ and $v_t \in \mathbb{R}^q$ are noise terms. The signal $\sigma_t \in \{1, \dots, s\}$ defines the mode which is active at time t. System (1) is a switched affine system when it can be expressed as:

$$\begin{aligned} x_{t+1} &= A_{\sigma_t} x_t + B_{\sigma_t} u_t + \bar{f}_{\sigma_t} + w_t \\ y_t &= C_{\sigma_t} x_t + D_{\sigma_t} u_t + \bar{g}_{\sigma_t} + v_t, \quad t \in \mathbb{Z}_+. \end{aligned} \tag{2}$$

The matrices $A_i \in \mathbb{R}^{n\times n}$, $B_i \in \mathbb{R}^{n\times p}$, $C_i \in \mathbb{R}^{q\times n}$, $D_i \in \mathbb{R}^{q\times p}$, $\bar{f}_i \in \mathbb{R}^n$, and $\bar{g}_i \in \mathbb{R}^q$ define the affine dynamics for each mode $i \in \{1, \dots, s\}$. System (2) is PWA when σ_t is defined according to a polyhedral partition of the state and input space, i.e., when

$$\sigma_t = i \iff \begin{bmatrix} x_t \\ u_t \end{bmatrix} \in R_i \subset R, \tag{3}$$

where R_i, $i = 1, \dots, s$, are convex polyhedra defining a partition of the state-input domain $R \subseteq \mathbb{R}^{n+p}$, that is when $R_i \cap R_j = \emptyset$ for $i \neq j$, and $\bigcup_i R_i = R$.

Switched affine systems can be expressed in input–output form as (SARX) systems of fixed orders n_a, n_b, such that for every component $y_t^{(i)} \in \mathbb{R}$ of the output vector $y_t \in \mathbb{R}^q$ it holds:

$$y_t^{(i)} = \bar{\theta}_{\sigma_t}^{(i)\mathrm{T}} \begin{bmatrix} r_t \\ 1 \end{bmatrix} + \bar{e}_t^{(i)}, \ i = 1, \dots, q, \tag{4}$$

where the regressor vector $r_t \in \mathbb{R}^{\bar{d}}$, $\bar{d} = qn_a + p(n_b + 1)$, is defined by

$$r_t = [y_{t-1}^{\mathrm{T}} \dots y_{t-n_a}^{\mathrm{T}} u_t^{\mathrm{T}} u_{t-1}^{\mathrm{T}} \dots u_{t-n_b}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{\bar{d}}. \tag{5}$$

The parameter vectors $\bar{\theta}_j^{(i)} \in \mathbb{R}^{\bar{d}+1}$, $j \in \{1, \dots, s\}$, define each ARX mode, and $\bar{e}_t \in \mathbb{R}^q$ is a noise term. Similarly, (4) is PWARX if

$$\sigma_t = i \iff r_t \in P_i \subset P \subseteq \mathbb{R}^{\bar{d}}, \tag{6}$$

and $\{P_i\}_{i=1}^s$ define a polyhedral partition of $P \subseteq \mathbb{R}^{\bar{d}}$.

A. Realization and Identification of PWARX Models

Every observable switched affine system admits a SARX representation. Necessary and sufficient conditions for input–output realization of SARX and PWARX systems are given in [27], and [28], respectively. It is worth mentioning, however, that the number of modes and parameters can grow considerably when a PWA state-space system is converted into a minimum-order equivalent PWARX representation [28]. The general identification problem for a PWARX system of the form (4)-(6) can be formulated as a stochastic optimization problem over the parameters $\{n_a, n_b, s, \{\theta_i\}_{i=1}^s, \{P_i\}_{i=1}^s\}$. We make the following assumption that will allow us to concentrate on the properties of PWARX identification, assuming known $(\tilde{n}_a, \tilde{n}_b)$ subject to potential computational bounds.

Assumption 1. *Upper bounds $(\tilde{n}_a, \tilde{n}_b)$ on the orders of the model (n_a, n_b) are known.*

B. Realization and Identification of PWA State-Space Models

The problem of identifying a state-space representation of a switched affine system can be challenging. Identifiability issues arise regarding the characterization of minimality of discrete-time switched linear systems [5], [6], [22]. In this

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted,
but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

work, to ensure uniqueness of the realizations, given that all subsystems $i \in \{1,\dots,s\}$ share the same state space, we make the following assumptions.

Assumption 2. *$C_i = C$, $\forall i \in \{1,\dots,s\}$ in system* (2)*.*

Assumption 3. *We assume no affine dynamics, i.e., $\bar{f}_{\sigma_t} = 0$, $\bar{g}_{\sigma_t} = 0$, no feed-forward terms, i.e., $D_{\sigma_t} = 0$, full observability, i.e., $C = I_n$, and same zero-mean statistics for the error terms w_t and v_t for every mode of the system.*

Assumption 2 implies that the order n is known (observed) and enforces that the set of observations is acquired using the same observation mechanism, which leads to the realization of (2) being unique. Assumptions 3 simplify the presentation of the proposed methodology without loss of generality. Together, Assumptions 2 and 3 allow for the joint modeling of PWARX and state-space PWA systems, as defined in Section III.

In addition to the realizations of the local systems being non-unique, minimality and identifiability of the switched system does not necessarily imply that of the local subsystems [29]. In Theorem 1, we describe the conditions under which the local linear models of (2) (under Assumptions 2–3) can be identified, even when a subset of them is not controllable (minimal) in isolation.

Theorem 1. *Consider a bounded-input bounded-output linear discrete-time system of the form:*

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \quad t \in \mathbb{Z}_+ \\ y_t &= x_t, \end{aligned} \tag{7}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^p$, $A \in \mathbb{R}^{n\times n}$, and $B \in \mathbb{R}^{n\times p}$. Denote $r_t = [x_t^{\mathrm{T}} u_t^{\mathrm{T}}]^{\mathrm{T}}$. Then, if there exist some $\alpha, \beta, T > 0$ such that

$$\alpha I_{n+p} \preceq \sum_{\tau=t}^{t+T} r_t r_t^{\mathrm{T}} \preceq \beta I_{n+p}, \quad \forall t \geq 0, \tag{8}$$

the augmented parameter matrix $\hat{\Theta}_t = [\hat{A}_t | \hat{B}_t]$ updated by the recursion

$$\hat{\Theta}_{t+1} = \hat{\Theta}_t - \gamma \left(\hat{\Theta}_t r_t - x_{t+1} \right) r_t^{\mathrm{T}}, \quad t \geq 0, \tag{9}$$

for some $\gamma > 0$, asymptotically converges to $\Theta = [A|B]$.

Proof. See Appendix A. □

As a result of Theorem 1, throughout this paper, we make the following assumption to ensure identifiability of (2) under Assumptions 2–3:

Assumption 4. *All linear subsystems $i \in \{1,\dots,s\}$ of* (2) *are asymptotically bounded, and the bounded control input u_t is designed such that for every mode $i \in \{1,\dots,s\}$ of* (2)*, there exist some $\alpha_i, \beta_i, T_i > 0$ for which the following persistence of excitation condition holds:*

$$\alpha_i I_{n+p} \preceq \sum_{\tau=t}^{t+T_i} \begin{bmatrix} x_\tau x_\tau^{\mathrm{T}} & x_\tau u_\tau^{\mathrm{T}} \\ u_\tau x_\tau^{\mathrm{T}} & u_\tau u_\tau^{\mathrm{T}} \end{bmatrix} \preceq \beta_i I_{n+p}, \ \forall t \geq 0. \tag{10}$$

Remark 1. *Informally, condition* (10) *states that not every subsystem in* (2) *needs to be controllable (minimal), as long as the boundaries of each mode (region R_i in the state-input system) are visited often enough.*

Remark 2. *The assumption of asymptotic boundedness and controllability (thus, minimality) for all subsystems of* (2) *would simplify the condition* (10) *to a persistence of excitation criterion for the input u_t for each subsystem separately. However, it is a limiting assumption in a practical sense.*

III. Switched System Identification as an Optimization Problem

Consider a switched linear system of the form:

$$\begin{aligned} \psi_t &= \Theta_i \phi_t + e_t, \\ &= [\phi_t^{\mathrm{T}} \otimes I_m]\theta_i + e_t, \ \text{if } \phi_t \in S_i, \ t \in \mathbb{Z}_+, \end{aligned} \tag{11}$$

where $\psi_t \in \mathbb{R}^m$, $\phi_t \in \mathbb{R}^d$, $\sigma_t \in \{1,\dots,s\}$, $\Theta_i \in \mathbb{R}^{m\times d}$, for all $i = 1,\dots,s$, $\theta_i = \mathrm{vec}(\Theta_i) \in \mathbb{R}^{md}$, $e_t \in \mathbb{R}^m$ is a zero-mean noise signal, and $\{S_i\}_{i=1}^s$ define a polyhedral partition of $S \subseteq \mathbb{R}^d$. System (4) can be written in the form (11) with $\psi_t = y_t \in \mathbb{R}^q$, $\phi_t = [r_t^{\mathrm{T}} 1]^{\mathrm{T}} \in \mathbb{R}^{\bar{d}+1}$, and $\Theta_i = [\bar{\theta}_i^{(1)} \dots \bar{\theta}_i^{(q)}]^{\mathrm{T}}$, where $m = q$, and $d = \bar{d} + 1$. In addition, system (2) under Assumptions 2, 3 can be written in the form (11) with $\psi_t = x_{t+1} \in \mathbb{R}^n$, $\phi_t = [x_t^{\mathrm{T}} u_t^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n+p}$, and $\Theta_i = [A_i | B_i]$, where $m = n$, and $d = n + p$. Notice that, in this case, (11) holds for $(t-1) \in \mathbb{Z}_+$, i.e., $t \in \mathbb{N}$.

Under the identifiability conditions discussed in Section II, the general identification problem for a switching system of the form (11) can be formulated as a stochastic optimization problem over the parameters $\{s, \{\theta_i\}_{i=1}^s, \{S_i\}_{i=1}^s\}$, as follows:

$$\underset{s,\{\theta_i\},\{S_i\}}{\text{minimize}} \ \mathbb{E}\left[\sum_{i=1}^{s} \mathbb{1}_{[\Phi \in S_i]} d_\rho \left(\Psi, [\Phi^{\mathrm{T}} \otimes I_m]\theta_i \right) \right], \tag{12}$$

where $\Psi \in \mathbb{R}^m$ and $\Phi \in \mathbb{R}^d$ represent random variables, realizations of which constitute the system observations, the nonnegative measure d_ρ is an appropriately defined dissimilarity measure, and the expectation is taken with respect to the joint distribution of $(\Psi, \Phi) \in \mathbb{R}^{m+d}$ that depends on the system dynamics, the control input, and the noise term in (11).

It is clear that the optimization problem (12) is computationally hard and becomes intractable as the number of modes and states increases. In particular, the number of modes s is unknown and completely alters the cardinality and the domain of the set of parameter vectors $\{\theta_i\}_{i=1}^s$ that represent the dynamics of the system. In addition, a parametric representation for the polyhedral regions $\{S_i\}$ should be defined.

To represent the regions $\{S_i\}$, we will follow a Voronoi tessellation approach based on prototypes. We introduce a set of parameters $\hat{\phi} := \left\{ \hat{\phi}_i \right\}_{i=1}^K$, $\hat{\phi}_i \in S$ and define the regions:

$$\Sigma_i = \left\{ \phi \in S : i = \arg\min_j d_\rho(\phi, \hat{\phi}_j) \right\}, \ i = 1,\dots,K. \tag{13}$$

The measure d_ρ can be designed such that the Voronoi regions Σ_i are polyhedral, e.g., when d_ρ is a squared Euclidean distance or any Bregman divergence, as will be explained in Section IV-A. In this sense, each S_i can be mapped to a region Σ_j (for $K = s$) or the union of a subset of $\{\Sigma_j\}$ (for $K > s$),

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

according to a predefined rule, as will be explained in Section IV-C. An illustration of this partition is given in Fig. 1.

In addition to the prototype parameters $\left\{\hat{\phi}_i\right\}_{i=1}^K$, we also introduce a set of parameters $\hat{\theta} := \left\{\hat{\theta}_i\right\}_{i=1}^K$, $\hat{\theta}_i \in \mathbb{R}^{md}$, with each $\hat{\theta}_i$ associated with the region Σ_i according to (13). Representing the augmented random vector

$$X = \begin{bmatrix}\Psi \\ \Phi\end{bmatrix} \in \Pi \subseteq \mathbb{R}^{m+d}, \tag{14}$$

we can define a set of augmented codevectors $\mu := \{\mu_i\}_{i=1}^K$ as

$$\mu_i = \begin{bmatrix} z(\phi, \hat{\theta}_i) \\ \hat{\phi}_i \end{bmatrix} \in \Pi,\ i = 1, \ldots, K, \tag{15}$$

where the first component of each μ_i[1] is a mapping $z(\phi, \hat{\theta}_i) = [\phi^{\mathrm{T}} \otimes I_m]\hat{\theta}_i$ that simulates the local model dynamics in (11) with unknown parameters θ_i, and the second component is a set of unknown codevectors $\hat{\phi}_i$ that define the partition in (13).

Problem (12) can then be decomposed into two interconnected stochastic optimization problems. Assuming $\left\{\hat{\theta}_i\right\}_{i=1}^K$ are known, the optimization problem

$$\underset{\hat{\phi}}{\text{minimize}}\ \mathbb{E}\left[\sum_{i=1}^K \mathbb{1}_{\left[\Phi \in \Sigma_i(\hat{\phi})\right]} d_\rho\left(X(\Psi, \Phi), \mu_i(\hat{\theta}_i, \hat{\phi}_i)\right)\right] \tag{16}$$

finds the optimal parameters $\left\{\hat{\phi}_i\right\}_{i=1}^K$ that define the partition $\{\Sigma_i\}_{i=1}^K$ subject to the joint distribution of (Ψ, Φ), and is, therefore, a mode switching signal identification problem.

On the other hand, assuming the partition $\{\Sigma_i\}_{i=1}^K$ (and, therefore, $\{S_i\}_{i=1}^s$) is known, the optimization problem

$$\underset{\hat{\theta}}{\text{minimize}}\ \mathbb{E}\left[\sum_{i=1}^K \mathbb{1}_{[\Phi \in \Sigma_i]} d_\rho\left(\Psi, [\Phi^{\mathrm{T}} \otimes I_m]\hat{\theta}_i\right)\right] \tag{17}$$

is a system identification problem for each mode of the system.

In Section IV we address the question of finding the optimal number K according to a performance-complexity trade-off, as well as finding a mapping between $\{\Sigma_i\}_{i=1}^K$ and $\{S_i\}_{i=1}^{\hat{s}}$ for the lowest possible number $\hat{s} \geq s$. In Section V we tackle the problem of estimating both $\hat{\phi}$ and $\hat{\theta}$ by solving (16) and (17) as a system of interconnected stochastic optimization problems in real-time using principles from two-timescale stochastic approximation theory.

IV. Mode Identification with Online Deterministic Annealing

We aim to construct a recursive stochastic optimization algorithm to solve problem (16) while progressively estimating the number K of the augmented codevectors $\{\mu_i\}_{i=1}^K$, an estimate $\hat{s}$ of the actual number of modes, and a mapping between $\{\Sigma_i\}_{i=1}^K$ and $\{S_i\}_{i=1}^{\hat{s}}$. Recall that the observed data are represented by the random variable $X \in \Pi$ in (14), and the augmented codevectors $\{\mu_i\}_{i=1}^K$ are normally treated as constant parameters to be estimated. To progressively estimate K and $\hat{s}$, we will adopt the online deterministic annealing approach [9], [30], and define a probability space over an arbitrary number of codevectors, while constraining their distribution using a maximum-entropy principle at different levels. First we define a quantizer $Q : \Pi \rightarrow \Pi$ as a stochastic mapping of the form:

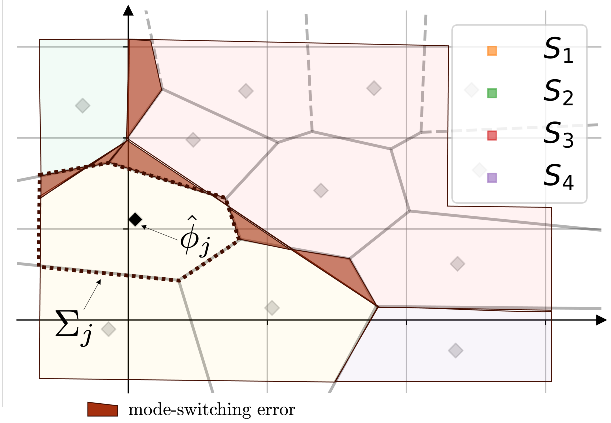

Fig. 1: Illustration of the partition $\{S_i\}_{i=1}^s$ $(s = 4)$ of the state-input space S and its connection to the artificial partition $\{\Sigma_j\}_{j=1}^K$ $(K > s)$. The optimal parameters $\left\{\hat{\phi}_j\right\}$ induce a partition $\{\Sigma_j\}$ that minimizes the mode switching error.

$$Q(x) = \mu_i \ \text{ with probability } p(\mu_i|x). \tag{18}$$

Then we formulate the multi-objective optimization

$$\underset{\hat{\phi}}{\text{minimize}}\ F_\lambda(\mu) = (1-\lambda)D(\mu) - \lambda H(\mu),\ \lambda \in [0, 1), \tag{19}$$

where the dependence on $\hat{\phi}$ comes through $\mu(\hat{\phi})$, the term

$$D(\mu) = \mathbb{E}\left[d_\rho\left(X, Q\right)\right] = \int p(x) \sum_i p(\mu_i|x) d_\rho(x, \mu_i)\ \mathrm{d}x \tag{20}$$

is a generalization of the objective in (16), and

$$\begin{aligned} H(\mu) &= \mathbb{E}\left[-\log P(X, Q)\right] \\ &= H(X) - \int p(x) \sum_i p(\mu_i|x) \log p(\mu_i|x)\ \mathrm{d}x \end{aligned} \tag{21}$$

is the Shannon entropy. This is now a problem of finding the locations $\left\{\hat{\phi}_i\right\}$ and the corresponding probabilities $\{p(\mu_i|x) = \mathbb{P}[Q = \mu_i|X = x]\}$.

Notice that, for $p(\mu_i|x) = \mathbb{1}_{\left[\phi \in \Sigma_i(\hat{\phi})\right]}$ and $\lambda = 0$, (19) is equivalent to (16). In that sense, (19) introduces extra optimization parameters in the probabilities $\{p(\mu_i|x)\}$, and the parameter λ that defines a homotopy F_λ. However, the advantages of this approach are notable, and, perhaps counter-intuitively, lead to numerical optimization solutions with several computational benefits. On the one hand, the Lagrange multiplier $\lambda \in [0, 1)$ controls the trade-off between D and H, which, as will be shown, is a trade-off between performance and complexity. On the other hand, the use of the conditional probabilities $\{p(\mu_i|x)\}$ allows for the definition of the entropy

[1] Throughout this paper we will use the notation μ_i, $\mu_i(\hat{\phi}_i)$, $\mu_i(\hat{\theta}_i)$, $\mu_i(\hat{\theta}_i, \hat{\phi}_i)$, $\mu_i(\phi, \hat{\theta}_i, \hat{\phi}_i)$ interchangeably, to showcase the dependence on the variables of interest in each case.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

term H, which introduces several useful properties [9], [30]–[33]. In particular, as we will show in Section IV-B, reducing the values of λ defines a direction that resembles an annealing process [30], [34] and induces a bifurcation phenomenon, with respect to which, the number of unique codevectors K_λ depends on λ and is finite for any given value of $\lambda > 0$. This process also introduces robustness with respect to initial conditions [30], [35].

A. Solving the Optimization Problem

To solve (19) for a given value of λ, we successively minimize F_λ first with respect to the association probabilities $\{p(\mu_i|x)\}$, and then with respect to the codevector locations μ. The solution of the optimization problem

$$\begin{aligned} F^*_\lambda(\mu) &= \min_{\{p(\mu_i|x)\}} F_\lambda(\mu), \\ \text{s.t.} &\sum_i p(\mu_i|x) = 1, \end{aligned} \tag{22}$$

is given by the Gibbs distributions [36]:

$$p^*(\mu_i|x) = \frac{e^{-\frac{1-\lambda}{\lambda} d_\rho(x,\mu_i)}}{\sum_j e^{-\frac{1-\lambda}{\lambda} d_\rho(x,\mu_j)}}, \ \forall x \in \Pi. \tag{23}$$

In order to minimize $F^*(\mu)$ with respect to $\hat{\phi}$ we set the gradients to zero

$$\frac{\mathrm{d}}{\mathrm{d}\hat{\phi}} F^*_\lambda(\mu) = \frac{\mathrm{d}}{\mathrm{d}\mu} F^*_\lambda(\mu) \frac{\mathrm{d}\mu}{\mathrm{d}\hat{\phi}} = 0 \tag{24}$$

where $\frac{\mathrm{d}\mu}{\mathrm{d}\hat{\phi}} = \begin{bmatrix} 0_{m\times d} \\ I_d \end{bmatrix}$, and

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mu} F^*_\lambda(\mu) &= \frac{\mathrm{d}}{\mathrm{d}\mu} \left((1-\lambda)D(\mu) - \lambda H(\mu)\right) \\ &= \sum_i \int p(x) p^*(\mu_i|x) \frac{\mathrm{d}}{\mathrm{d}\mu_i} d_\rho(x,\mu_i) \ \mathrm{d}x = 0, \end{aligned} \tag{25}$$

where we have used (23) and direct differentiation with similar arguments as in [36]. It follows that $\frac{\mathrm{d}}{\mathrm{d}\hat{\phi}} F^*_\lambda(\mu) = 0$ which implies that

$$\int p(x) p^*(\mu_i|x) \frac{\mathrm{d}}{\mathrm{d}\mu_i} d_\rho(x,\mu_i) \ \mathrm{d}x \begin{bmatrix} 0_{m\times d} \\ I_d \end{bmatrix} = 0, \ \forall i. \tag{26}$$

Equation (26) has a closed-form solution if the dissimilarity measure d_ρ belongs to the family of Bregman divergences [30], [37],information-theoretic dissimilarity measures that include the squared Euclidean distance and the Kullback-Leibler divergence, and are defined as follows:

Definition 1 (Bregman Divergence)**.** *Let $\rho : S \to \mathbb{R}$, be a strictly convex function defined on a vector space $S \subseteq \mathbb{R}^d$ such that ϕ is twice F-differentiable on S. The Bregman divergence $d_\rho : H \times S \to [0, \infty)$ is defined as:*

$$d_\rho(x,\mu) = \rho(x) - \rho(\mu) - \frac{\partial \rho}{\partial \mu}(\mu)(x-\mu),$$

where $x, \mu \in S$, and the continuous linear map $\frac{\partial \rho}{\partial \mu}(\mu) : S \to \mathbb{R}$ is the Fréchet derivative of ρ at μ.

Throughout this manuscript, we will assume that the dissimilarity measure d_ρ in (13) is a Bregman divergence, and, in particular, the squared Euclidean distance. Then the solution to the optimization problem

$$\underset{\hat{\phi}}{\text{minimize}} \ F^*_\lambda\left(\mu(\hat{\phi})\right), \tag{27}$$

where $F^*_\lambda(\mu)$ is the solution of (22) for a given $\lambda \in [0, 1)$ and $p^*(\mu_i|x)$ is given by (23), is given by Theorem 2.

Theorem 2. *If $d_\rho : \Pi \times \Pi \to \mathbb{R}_+$ is a Bregman divergence, then*

$$\hat{\phi}^*_i = \frac{\int \phi p(x) p^*(\mu_i|x) \ \mathrm{d}x}{p^*(\mu_i)} \tag{28}$$

is a solution to the optimization problem (27)*.*

Proof. By definition, for a Bregman divergence $d_\rho : \Pi \times \Pi \to \mathbb{R}_+$ based on a strictly convex function $\rho : \Pi \to \mathbb{R}$, it holds that $\frac{\partial d_\rho}{\partial \mu}(x,\mu) = -\left\langle \nabla^2 \rho(\mu), (x-\mu) \right\rangle$. Similar to [9], with standard algebraic manipulations, (26) then becomes

$$\int (\phi - \hat{\phi}^*_i) p(x) p^*(\mu_i|x) \ \mathrm{d}x = 0, \ \forall i, \tag{29}$$

where $p^*(\mu_i|x)$ is given by (23) and the integral is defined over the domain Π. Eq. (29) is equivalent to (28) since $\int p(x) p^*(\mu_i|x) \ \mathrm{d}x = p^*(\mu_i)$. □

Remark 3. *The partition $\{\Sigma_i\}$ induced by* (13) *and a dissimilarity measure d_ρ that belongs to the family of Bregman divergences, is separated by hyperplanes [37]. As a result, each Σ_i is a polyhedral region for a bounded domain S.*

Based on Theorem 2, Theorem 3 below constructs a gradient-free stochastic approximation algorithm that recursively estimates (28).

Theorem 3. *The sequence $\hat{\phi}_i(t)$ constructed by the recursive updates*

$$\begin{cases} \hat{\rho}_i(t+1) &= \hat{\rho}_i(t) + \beta(t)\left[\hat{p}_i(t) - \hat{\rho}_i(t)\right] \\ \sigma_i(t+1) &= \sigma_i(t) + \beta(t)\left[\phi_t \hat{p}_i(t) - \sigma_i(t)\right], \end{cases} \tag{30}$$

where $x_t = [\psi_t^{\mathrm{T}} \phi_t^T]^{\mathrm{T}}$ represents external input with $\psi_t \sim \Psi$, $\phi_t \sim \Phi$, $\sum_t \beta(t) = \infty$, $\sum_t \beta^2(t) < \infty$, and the quantities $\hat{p}_i(t)$ and $\hat{\phi}_i(t)$ are recursively updated as follows:

$$\hat{\phi}_i(t) = \frac{\sigma_i(t)}{\hat{\rho}_i(t)}, \quad \hat{p}_i(t) = \frac{\hat{\rho}_i(t) e^{-\frac{1-\lambda}{\lambda} d_\rho(x_t,\mu_i(t))}}{\sum_i \hat{\rho}_i(t) e^{-\frac{1-\lambda}{\lambda} d_\rho}(x_t, \mu_i(t))}, \tag{31}$$

with $\mu_i(t) = [z_i^{\mathrm{T}}(\phi_t, \hat{\theta}_i), \hat{\phi}_i(t)^{\mathrm{T}}]^{\mathrm{T}}$, converges almost surely to $\hat{\phi}^*_i$ given in (28)*.*

Proof. The proof follows similar arguments as Theorem 5 of [9]. □

Remark 4. *Intuitively, the quantity $\hat{\rho}_i$ in* (30) *is an estimate of the probability $p(\mu_i)$. In that sense, σ_i becomes an estimate of $\mathbb{E}\left[\mathbb{1}_{\{\mu\}}\Phi\right]$, and $\hat{\phi}_i$ becomes an estimate of $\mathbb{E}\left[\Phi|\mu\right]$, which is a generalization of the centroid form found in clustering algorithms [30].*

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

Remark 5. *Notice that the dynamics of* (30) *can be expressed as:*

$$\hat{\phi}_i(t+1) = \frac{\beta(t)}{\hat{\rho}_i(t)}\left[\frac{\sigma_i(t+1)}{\hat{\rho}_i(t+1)}\left(\hat{\rho}_i(t) - \hat{p}_i(t)\right) + \phi_t\hat{p}_i(t) - \sigma_i(t)\right], \tag{32}$$

where the recursive updates take place for every codevector $\hat{\phi}_i$ sequentially. This is a discrete-time dynamical system that presents bifurcation phenomena with respect to the parameter λ, i.e., the number of equilibria of this system changes with respect to the value λ which is hidden inside the term $\hat{p}_i(t)$ in (31)*. According to this phenomenon, the number of distinct values of $\hat{\phi}_i$ is finite, and the updates need only be taken with respect to these values that we call "effective codevectors". This is discussed in Section IV-B.*

B. Bifurcation Phenomena

In Section IV-A we described how to solve the optimization problem for a given value of the parameter λ. The main idea of the proposed approach is to solve a sequence of optimization problems of the form (19) with decreasing values of λ. This process then becomes a homotopy optimization method. In particular, the usage of the entropy term resembles annealing optimization methods and grants λ the name of a 'temperature' parameter. Notice that, so far, we have assumed an arbitrary number of codevectors K. We will show that the unique values of the set $\left\{\hat{\phi}_i\right\}$ that solves (19), form a finite set of K_λ values that we will refer to as "effective codevectors".

Notice that at high temperature ($\lambda \to 1$), (23) yields uniform association probabilities $p(\mu_i|x) = p(\mu_j|x),\ \forall i, j, \forall x$, and as a result of (28), all pseudo-inputs are located at the same point $\hat{\phi}_i = \mathbb{E}_X[\phi]\,,\ \forall i$, which means that there is one unique "effective" codevector given by $\mathbb{E}_X[\phi]$. As λ is lowered below a critical value, a bifurcation phenomenon occurs, when the number of "effective" codevectors increases, which describes an annealing process [30], [34]. Mathematically, this occurs when the existing solution $\hat{\phi}^*$ given by (28) is no longer the minimum of the free energy F^*, as the temperature λ crosses a critical value. Following principles from variational calculus, we can track the bifurcation by the condition:

$$\left.\frac{d^2}{d\epsilon^2}F^*(\left\{\hat{\phi} + \epsilon\hat{\psi}\right\})\right|_{\epsilon=0} \geq 0, \tag{33}$$

for all choices of finite perturbations $\left\{\hat{\psi}\right\}$. Using (33) and direct differentiation, one can show that bifurcation depends on the temperature coefficient λ (and the choice of the Bregman divergence, through the function ρ) [9], [36]. In other words, the number of codevectors increases countably many times as the value of λ decreases, resulting, at the limit $\lambda \to 0$ ($K \to \infty$), in a consistent density estimator [36]. However, an algorithmic implementation needs to keep in memory only as many codevectors as the number of "effective" codevectors.

In practice. we can detect the bifurcation points by introducing perturbing pairs of codevectors at each temperature level λ. In this way, the codevectors $\hat{\phi}$ are doubled by inserting a perturbation of each $\hat{\phi}_i$ in the set of effective codevectors. The newly inserted codevectors will merge with their pair if a critical temperature has not been reached and separate otherwise. The merging criterion takes the form:

$$\frac{1-\lambda}{\lambda}d_\rho(\hat{\phi}_i, \hat{\phi}_j) \leq \epsilon_n,\ \forall i, j, \tag{34}$$

for a given threshold ϵ_n. The pseudocode for this algorithm is presented in Alg. 1. A detailed discussion on the implementation of the original online deterministic annealing algorithm, its complexity, and the effect of its parameters, can be found in [9], [30], [36].

C. Estimating the number of modes

As illustrated in Fig. 1, the problem formulation developed in Section III defines a possibly imperfect surjective mapping from $\left\{\Sigma_j\right\}_{j=1}^K$ to $\left\{S_i\right\}_{i=1}^s$ such that each S_i is defined as a union of a subset of $\left\{\Sigma_j\right\}_{j=1}^K$. In this section, we define a recursive method to automatically construct this mapping, a critical addition to the methods proposed in [21], [22].

It is worth noting that the construction of $\left\{\Sigma_j\right\}_{j=1}^K$ defines a consistent density estimator of the mode swithcing behavior on S in the limit $\lambda \to 0$ (which induces $K \to \infty$) [36]. However, according to Remark 3, it is possible for this mapping to be perfect even with bounded K if P is a polyhedral partition and the reconstruction is ideal. Then each S_i is perfectly represented, inducing zero mode switching error. In addition, notice that the design of an appropriate termination criterion for Alg. 1 is an open question and is subject to the trade-off between the number K and the minimization of the identification error. In this work, we make use of the condition $K \leq K_{\max}$ as a termination criterion, where $K_{\max}$ represents the computational capacity of the identification device.

Recall that each Σ_j is associated with a parameter vector $\hat{\theta}_j,\ j = 1, \ldots, K$. Assuming a set $\bar{\theta} = \left\{\bar{\theta}_i\right\}_{i=1}^{\hat{s}}$, we define each $\hat{\theta}_j$ as the mapping:

$$\hat{\theta}_j(\bar{\theta}) = \bar{\theta}_i,\ \text{if}\ i = \arg\min_k d_\rho(\hat{\theta}_j, \bar{\theta}_k). \tag{35}$$

In this way $\Sigma_j \in S_i$ if $\hat{\theta}_j(\bar{\theta}) = \bar{\theta}_i$. Therefore, given (35), the goal now is to find $\hat{s}$ and $\bar{\theta}$ such that $\hat{s} = s$, and $\bar{\theta}_i = \theta_i,\ \forall i \in \{1, \ldots, s\}$. We follow a similar approach to the bifurcation mechanism described in Section IV-B. Starting with one codevector $\hat{\phi}_0$, we define $\bar{\theta}_0 = \hat{\theta}_0$. Every time a codevector $\hat{\phi}_j$ is split into a pair of perturbed codevectors, a new $\hat{\theta}_{j'}$ is introduced. After convergence for a given λ, merging of the codevectors is detected by (34). For the insertion of a new $\bar{\theta}_i$ we check the condition:

$$d_\rho(\hat{\theta}_j, \bar{\theta}_i) > \epsilon_s,\ \forall j, \tag{36}$$

with respect to a given threshold ϵ_s. Notice that in contrast to (34), (36) does not depend on λ. If (36) is satisfied, a new $\bar{\theta}_i$ is introduced and $\hat{s} \leftarrow \hat{s} + 1$. This process is integrated in the mode identification algorithm and its pseudocode is presented in Alg. 1.

Remark 6. *Note that $\left\{\hat{\theta}_j\right\}$ are only used as functions of $\bar{\theta}$, and the parameters $\left\{\bar{\theta}_i\right\}$ are the ones that are being updated by the local system identification algorithm that will be presented in Section V.*

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

V. Piecewise Affine System Identification

In this section we review standard recursive system identification for estimating the parameters $\left\{\bar{\theta}_i\right\}$ of the local models given knowledge of the partition $\{S_i\}$.

We show that this kind of recursive identification can be formulated as a stochastic approximation algorithm, and that it can be combined using the theory of two-timescale stochastic approximation with the stochastic approximation method of Section IV to estimate both $\{S_i\}$ and $\left\{\bar{\theta}_i\right\}$ at the same time.

A. Recursive Identification of Local Models

Recall that, given knowledge of the partition $\{S_i\}_{i=1}^{s}$, each local linear model of the PWA system in (11) is completely defined by the parameters $\{\theta_i\}$. In the following, we develop a stochastic approximation recursion to estimate $\left\{\bar{\theta}_i\right\}$. First we define the error:

$$\epsilon(t) = \sum_i \mathbb{1}_{[\phi_t \in S_i]} [\phi_t^{\mathrm{T}} \otimes I_m] \bar{\theta}_i - \psi_t \tag{37}$$

A stochastic gradient descent approach aims to minimize the error:

$$\underset{\bar{\theta}_i}{\operatorname{minimize}} \;\; \frac{1}{2} \mathbb{E}\left[\|\epsilon(t)\|^2\right], \tag{38}$$

using the recursive updates:

$$\begin{aligned} \bar{\theta}_i(t+1) &= \bar{\theta}_i(t) - \alpha(t) \left(\nabla_{\bar{\theta}_i} \epsilon(t)\right) \epsilon(t) \\ &= \bar{\theta}_i(t) - \alpha(t) [\phi_t^{\mathrm{T}} \otimes I_m]^{\mathrm{T}} \epsilon(t) \end{aligned} \tag{39}$$

where $\sum_n \alpha(n) = \infty$, $\sum_n \alpha^2(n) < \infty$. Here the expectation is taken with respect to the joint distribution of (ψ_y, ϕ_t) as explained in Section III. This is a standard recursive identification method and constitutes a stochastic approximation sequence of the form:

$$\bar{\theta}_i(t+1) = \bar{\theta}_i(t) + \alpha(t) \left[h_\theta(\bar{\theta}_i(t)) + M_\theta(t+1)\right], \;\; t \geq 0, \tag{40}$$

where $h_\theta(\bar{\theta}_i) = -\nabla \mathbb{E}\left[\|\epsilon(t)\|^2\right]$ is Lipschitz, and $M(t+1) = \nabla \mathbb{E}\left[\|\epsilon(t)\|^2\right] - \nabla \|\epsilon(t)\|^2$ is a Martingale difference sequence. This sequence converges almost surely to the equillibrium of the differential equation

$$\dot{\bar{\theta}}_i = h_\theta(\bar{\theta}_i), \;\; t \geq 0. \tag{41}$$

which can be shown to be a solution of (38) with standard Lyapunov arguments. For more details the reader is referred to [9], [38]. Moreover, notice that (39) is a vectorized representation of (9), for $\gamma = \alpha(t) > 0$. Therefore, under the PE condition (10) of Assumption 4, and under the zero-mean noise assumption, it follows that $\bar{\theta}_i$ converges asymptotically to θ_i for all $i = 1, \ldots, s$, i.e., the minimum of (38) is achieved.

Algorithm 1 Switched System Identification

```
Set parameters and initialize $\hat{\phi} = \left\{\hat{\phi}_0\right\}, \bar{\theta} = \left\{\bar{\theta}_0\right\}$
while $K < K_{\max}$ and $\lambda > \lambda_{\min}$ do
  Perturb $\hat{\phi}_i \leftarrow \left\{\hat{\phi}_i + \delta, \hat{\phi}_i - \delta\right\}, \forall i$
  Set $t \leftarrow 0$
  repeat
    Observe $x = (\psi, \phi)$ according to (11)
    Update $\bar{\theta}_w$, $w = \arg\min_j d_\rho(\phi, \hat{\phi}_j)$, using (39)
    for $i = 1, \ldots, K$ do
      Update $\hat{\phi}$ using (30), (31)
    end for
    $t \leftarrow t + 1$
  until Convergence
  Discard $\hat{\phi}_i$ if $\frac{1-\lambda}{\lambda} d_\rho(\hat{\phi}_j, \hat{\phi}_i) < \epsilon_n, \forall i, j, i \neq j$
  Insert $\hat{\theta}_i$ in $\bar{\theta}$ if $d_\rho(\hat{\theta}_j, \hat{\theta}_i) > \epsilon_s, \;\forall j$
  Lower temperature $\lambda \leftarrow \gamma\lambda$, $0 < \gamma < 1$
end while
Define $\{\Sigma_i\}_{i=1}^{K}$ using (13)
Define $\hat{s} = \mathrm{card}(\bar{\theta})$
Define $\{S_i\}_{i=1}^{\hat{s}}$ by $\Sigma_j \in S_i$ if $\hat{\theta}_j(\bar{\theta}) = \bar{\theta}_i$
Estimated Model Parameters: $\hat{s}$, $\{S_i\}_{i=1}^{\hat{s}}$, $\left\{\bar{\theta}_i\right\}_{i=1}^{\hat{s}}$
```

B. Combined Mode and Dynamics Identification

Recall that the mode identification method is based on the stochastic approximation updates (30) that can be written with respect to the vectors $\xi_i(t) = [\hat{\rho}_i^{\mathrm{T}}(t) \sigma_i^{\mathrm{T}}(t)]^{\mathrm{T}}$ and a stepsize schedule $\beta(t)$ in the form:

$$\xi_i(t+1) = \xi_i(t) + \beta(t) \left[h_\phi\left(\xi(t), \bar{\theta}(t)\right) + M_\phi(t+1)\right], \;\; t \geq 0, \tag{42}$$

where h_ϕ is Lipschitz, $M_\phi(t)$ is a Martingale difference sequence and the dependence on $\bar{\theta}$ comes from the quantity $\hat{p}_i(t)$ in (31) given (35). At the same time, the recursive system identification technique to estimate $\bar{\theta}$ is a stochastic approximation sequence with a stepsize schedule $\alpha(t)$ of the form:

$$\bar{\theta}_i(t+1) = \bar{\theta}_i(t) + \alpha(t) \left[h_\theta\left(\xi(t), \bar{\theta}(t)\right) + M_\theta(t+1)\right], \;\; t \geq 0, \tag{43}$$

as given in (40). The dependence on ξ, comes through (37), since ξ defines $\hat{\phi}$, which defines $\{\Sigma_i\}_{i=1}^{K}$ through (13), which defines $\{S_i\}_{i=1}^{\hat{s}}$ through the rule $\Sigma_j \in S_i$ if $\hat{\theta}_j(\bar{\theta}) = \bar{\theta}_i$.

Theorem 4 shows how the two recursive algorithms (42) and (43) can be combined using the theory of two-timescale stochastic approximation if ${}^{\beta(t)}\!/_{\alpha(t)} \to 0$, i.e., when the estimation of the partition $\{\Sigma_i\}_{i=1}^{K}$ is updated at a slower rate than the updates of the parameters $\left\{\bar{\theta}_i\right\}_{i=1}^{\hat{s}}$.

Theorem 4. *Consider the sequence* $\{\xi(t)\}_{t \in \mathbb{Z}_+}$ *generated using the updates* (42)*, where* $\xi_i(t) = [\hat{\rho}_i^{\mathrm{T}}(t) \sigma_i^{\mathrm{T}}(t)]^{\mathrm{T}}$*, and* $(\hat{\rho}_i, \sigma_i)$ *are defined in* (30). *Consider the sequence* $\left\{\bar{\theta}(t)\right\}_{t \in \mathbb{Z}_+}$ *generated by the updates* (43). *Let the stepsizes* $(\alpha(t), \beta(t))$ *of* (43) *and* (42)*, respectively, satisfy the conditions* $\sum_n \alpha(n) = \sum_n \beta(n) = \infty$, $\sum_n (\alpha^2(n) + \beta^2(n)) < \infty$, *and* ${}^{\beta(n)}\!/_{\alpha(n)} \to 0$, *with the last condition implying that the iterations for* $\{\xi(t)\}$ *run on a slower timescale than those for* $\left\{\bar{\theta}(t)\right\}$. *If the equation*

$$\dot{\bar{\theta}}(t) = h_\theta(\xi, \bar{\theta}(t)), \;\; \bar{\theta}(0) = \bar{\theta}_0, \tag{44}$$

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

has an asymptotically stable equilibrium $\lambda(\xi)$ for fixed ξ and some Lipschitz mapping λ, and the equation

$$\dot{\xi}(t) = h_\phi(\xi(t), \lambda(\xi(t))),\ \xi(0) = \xi_0, \tag{45}$$

has an asymptotically stable equilibrium ξ^*, then, almost surely, the sequence $(\xi(t), \bar{\theta}(t))$ generated by (42), (43)*, converges to $(\xi^*, \lambda(\xi^*))$.*

Proof. It follows directly from Theorem 2, Ch. 6 of [38]. □

Corollary 4.1. *Condition* (44) *of Theorem 4 is satisfied by the definition of h_θ in* (41)*. Therefore,* (45) *implies the convergence of $\hat{\phi}$ through* (31)*, and of the partition $\{\Sigma_i\}$ through* (13)*.*

Notice that the condition $^{\beta(t)}/_{\alpha(t)} \rightarrow 0$ is of great importance. Intuitively, the stochastic approximation algorithm (42), (43) consists of two components running in different timescales, where the slow component is viewed as quasi-static when analyzing the behavior of the fast transient. In practice, the condition $^{\beta(t)}/_{\alpha(t)} \rightarrow 0$ is satisfied by stepsizes of the form $(\alpha(t), \beta(t)) = (^1/_t, ^1/_{(1+t\log t)})$, or $(\alpha(t), \beta(t)) = (^1/_{t^{2/3}}, ^1/_t)$. Another way of achieving the two-timescale effect is to run the iterations for the slow component with stepsizes $\{\alpha_{t(k)}\}$, where $t(k)$ is a subsequence of t that becomes increasingly rare (i.e. $t(k+1) - t(k) \rightarrow \infty$), while keeping its values constant between these instants. A good policy is to combine both approaches and update the slow component with slower stepsize schedule $\beta(t)$ along a subsequence keeping its values constant in between (e.g., [9], [38]).

VI. General Switched System Identification

In Sections III, IV, and V we have developed a real-time idenitification method for PWA systems. However, neither the proposed methodology, nor the algorithmic implementation of Alg. 1 are constrained to PWA systems. Thus the proposed approach can, in principle, be applied to more general switching and hybrid systems. However, issues may arise with respect to the identifiability conditions, the mode-switching estimation error, and the possibly non-linear local system identification error. In this section, we discuss the applicability of the proposed approach in different cases often encountered in switching control systems, namely switched linear systems with non-polyhedral partition, and switched non-linear systems with polyhedral partition.

A. Switched linear systems with non-polyhedral partition.

In the case of linear local dynamics, the recursive identification method discussed in Section V-A remains unchanged, and the same convergence results hold. However, if the regions S_i of the mode switching partition $\{S_i\}_{i=1}^s$ are non-polyhedral, they cannot be perfectly approximated by a finite union of polyhedral regions $\{\Sigma_i\}_{i=1}^K$. It is worth pointing out that from the convergence results of the online deterministic annealing algorithm [36], it follows that the partition error can be arbitrarily small in the limit $K \rightarrow \infty$ (which is the case when $\lambda \rightarrow 0$). Albeit a nice analytical result, in practice there will always be non-zero error in the estimation of the partition $\{S_i\}_{i=1}^s$. We hereby discuss two ways to deal with this problem. The first is to assume the existence of a non-linear transformation that maps each S_i to a polyhedral region $\bar{S}_i$, and proceed with Alg. 1. General-purpose learning machines, such as artificial neural networks can be incorporated in this process. Further assumptions and analysis is required for this method, which is beyond the scope of this paper. The second refers to mitigating the jumping effect of the identified system to decrease the closed-loop error that naturally occurs due to imperfect mode switching. To this end, recall that, given an observation ϕ_t the dynamics of the identified model are given according to (11) by:

$$\hat{\psi}_t = [\phi_t^{\mathrm{T}} \otimes I_m]\bar{\theta}_i,\ \text{if}\ \phi_t \in \Sigma_j\ \text{and}\ \hat{\theta}_j(\bar{\theta}) = \bar{\theta}_i. \tag{46}$$

To mitigate the jumping behavior one can make use of the association probabilities

$$p(\phi_i|\phi_t) = \frac{e^{-\frac{1-\lambda}{\lambda} d_\rho(\phi_t,\phi_i)}}{\sum_j e^{-\frac{1-\lambda}{\lambda} d_\rho(\phi_t,\phi_i)}}, \tag{47}$$

to instead construct the weighted dynamics:

$$\hat{\psi}_t = \sum_{i=1}^K p^*(\phi_i|\phi_t)[\phi_t^{\mathrm{T}} \otimes I_m]\hat{\theta}_i. \tag{48}$$

This jump-mitigation method has been used in the literature to preserve smoothness of the closed-loop dynamics and is particularly useful when hybrid system identification is used for non-linear function approximation, i.e., when the original system is not hybrid but is to be approximated by a hybrid system with simpler local dynamics.

B. Switched non-linear systems with polyhedral partition.

In this case, often referred to as piece-wise non-linear hybrid systems [39], the mode switching partition $\{S_i\}_{i=1}^s$ is polyhedral, and can be perfectly approximated by a finite union of polyhedral regions $\{\Sigma_i\}_{i=1}^K$. For the identification of the non-linear local dynamics, the recursive identification method discussed in Section V-A needs to be modified. In particular the recursive updates:

$$\bar{\theta}_i(t+1) = \bar{\theta}_i(t) - \alpha(t)\left(\nabla_{\bar{\theta}_i}\epsilon(t)\right)\epsilon(t), \tag{49}$$

given in (39) of the same stochastic gradient descent structure are used, with the error term in this case given by

$$\epsilon(t) = \sum_i \mathbb{1}_{[\phi_t \in S_i]}\hat{f}(\phi_t, \bar{\theta}_i) - \psi_t, \tag{50}$$

where the functions $\hat{f}(\phi_t, \bar{\theta}_i)$ are local parametric models of known form, differentiable with respect to the parameters θ_i. General-purpose learning machines, such as artificial neural networks can be used. Notice that the identification updates remain stochastic approximation updates of the same form, which means that the convergence results of Theorem 4 continue to hold.

VII. Experimental Results

We illustrate the properties and evaluate the performance of the proposed algorithm in the following simulated systems.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

A. Benchmark PWARX System

A benchmark PWARX system, adopted from [8], is given in the input–output representation of (51):

$$y_t = \begin{cases} \theta_1^{\mathrm{T}} \phi_t + e_t, & \text{if } r_t \in P_1 \\ \theta_2^{\mathrm{T}} \phi_t + e_t, & \text{if } r_t \in P_2 \\ \theta_3^{\mathrm{T}} \phi_t + e_t, & \text{if } r_t \in P_3 \end{cases}, \tag{51}$$

where $y_t \in \mathbb{R}^1$, $r_t \in P = [-4, 4]$, $\phi_t = [r_t \; 1]^{\mathrm{T}}$, $(P_1, P_2, P_3) = ([-4, -1], (-1, 2), [2, 4])$, and $(\theta_1, \theta_2, \theta_3) = ([1, 2]^{\mathrm{T}}, [-1, 0]^{\mathrm{T}}, [1, 2]^{\mathrm{T}})$. The simplicity of this example enables the graphical representation of the mode-switching partition and the convergence of the model parameters. At the same time, (51) presents a jump at $r_t = 2$, and same dynamics for different regions of the input space, i.e., $\theta_1 = \theta_3$ while $P_1 \neq P_3$. It can thus be written in the form:

$$y_t = \begin{cases} \theta_2^{\mathrm{T}} \phi_t + e_t, & \text{if } \phi_t \in S_2 \\ \theta_1^{\mathrm{T}} \phi_t + e_t, & \text{otherwise} \end{cases}, \tag{52}$$

where $S_2 = \left\{ \phi = [r \; 1]^{\mathrm{T}} : r \in P_2 \right\}$. A total of $N = 150$ observations under Gaussian noise ($e_t \sim \mathcal{N}(0, 0.2)$) are accessible sequentially.

Algorithm 1 is applied to the observations for $T = 900$ iterations. The temperature parameters used for the online deterministic annealing algorithm are $(\lambda_{\max}, \lambda_{\min}, \gamma) = (0.99, 0.2, 0.8)$, and the stepsizes $(\alpha(t), \beta(t)) = \left({}^{1}\!/\!_{(1+0.01t)}, {}^{1}\!/\!_{(1+0.9t \log t)} \right)$. In addition, $\delta = 0.1$, $\epsilon_n = 1.0$, and $\epsilon_s = 2.0$. At first ($\lambda = \lambda_{\max}$), the algorithm keeps in memory only one codevector $\hat{\phi}_1$ and one model parameter vector $\bar{\theta}_1$, essentially assuming that the system has constant dynamics in the entire domain, i.e., $\hat{S}_1 = \Sigma_1 = \left\{ \phi = [r \; 1]^{\mathrm{T}} : r \in P \right\}$. As new input–output pairs are observed, the estimated parameter $\bar{\theta}_1$ gets updated by the iterations (39). We have assumed $\bar{\theta}_1(0) = [1, 1]^{\mathrm{T}}$.

At the same time, the estimate of $\bar{\theta}_1$ are used to update the location of the codevector towards the mean of the observation domain as shown in (28). As λ is reduced, the bifurcation phenomenon described in Section IV-B takes place, and, after reaching a critical value, the single codevector splits into two duplicates. Now the algorithm assumes that there are two modes in the system and estimates the optimal model parameters $\left\{ \bar{\theta}_1, \bar{\theta}_2 \right\}$ and partition $\left\{ \Sigma_1, \Sigma_2 \right\}$ (through the location of the codevectors $\left\{ \hat{\phi}_1, \hat{\phi}_2 \right\}$). This process continues until a desired termination criterion is reached. In this case it is the minimum temperature parameter $\lambda_{\min}$ that reflects to a potential time and computational constraint of the system. The bifurcation phenomenon is illustrated in Fig. 2 where the locations of the codevectors $\left\{ \hat{\phi}_i \right\}$, $\hat{\phi}_i \in P = [-4, 4]$ generated by Alg. 1 are shown. The algorithm progressively constructs a total of $K = 5$ effective codevectors. The number of modes is estimated with the process explained in Section IV-C. Two modes are estimated with $\bar{\theta} = \left\{ \bar{\theta}_1, \bar{\theta}_2 \right\}$. The association of each effective codevector with each identified mode according to the rule (35) is shown in Fig. 2.

The final estimated partition, the output of the estimated model, and its error with respect to the true model without noise are shown in Fig. 3. A single misclassification instance of the mode at the boundary of the true partition of the input–output domain is observed. This mode switching error can be avoided by allowing λ to go lower, which results in a larger number K of effective codevectors and is indicative of the performance/complexity trade-off of the algorithm. Finally, regarding the effect of the noise variance on the identification error, for $e_t \sim \mathcal{N}(0, 0.2)$, $e_t \sim \mathcal{N}(0, 0.5)$ and $e_t \sim \mathcal{N}(0, 0.7)$, the root-mean-square deviation across the observed samples was computed as $e_{RMS} = 0.504$, $e_{RMS} = 0.642$, and $e_{RMS} = 0.689$, respectively.

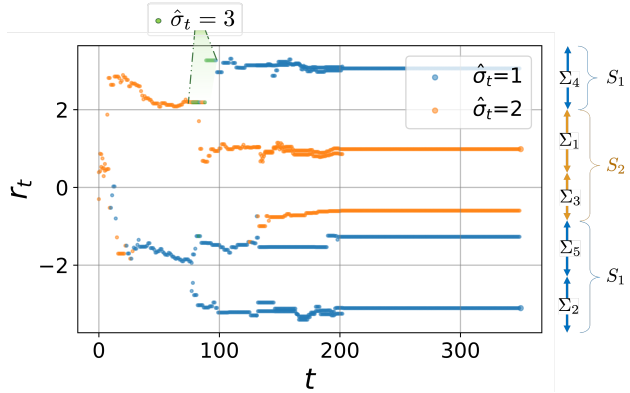

Fig. 2: Evolution of the codevectors $\left\{ \hat{\phi}_i \right\}$ generated by Alg. 1 for system (52) illustrating the bifurcation phenomenon described in Section IV-B. The association of each effective codevector with each identified mode according to the rule (35) is also shown. Notice that a third mode is detected and quickly discarded as explained in Section IV-C

.

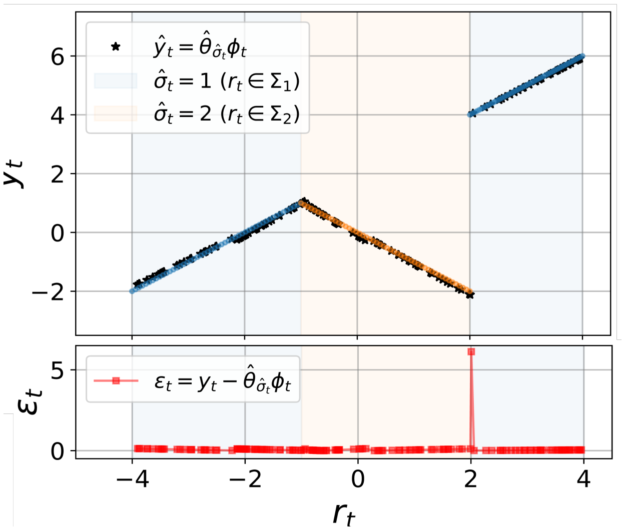

Fig. 3: Identified modes, predicted output and identification error with respect to the true model (52).

B. Comparison with existing methods

Compared to the clustering-based method in [8], the proposed algorithm applied to system (52) shows similar performance while maintaining several advantages. First, the number of modes is not assumed to be known a priori. Second, the proposed identification method can operate in real-time, i.e.,

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

using one forward pass of online observations as opposed to iterating multiple times through a dataset acquired offline. Finally, the progressive nature of the algorithm allows for the exploitation of the performance/complexity trade-off in applications where communication or computational resources are limited.

The same advantages can be observed against more recent methods as well. Consider the following system:

$$y_t = \begin{cases} \theta_1^{\mathrm{T}} \phi_t + e_t, & \text{if } \phi_t \in S_1 \\ \theta_2^{\mathrm{T}} \phi_t + e_t, & \text{otherwise} \end{cases}, \tag{53}$$

where $u_t, y_t \in \mathbb{R}$, $\phi_t = [y_{t-1}, y_{t-2}, u_{t-1}, u_{t-2}, 1] \in \mathbb{R}^5$, $\theta_1 = [0.1, 0.5, -0.4, 0.3, 0]^{\mathrm{T}}$, $\theta_2 = [0.2, 0.4, 0.1, 0.4, 0]^{\mathrm{T}}$, and $S_1 = \left\{\phi \in \mathbb{R}^5 : [1, 0.5, -0.3, 2, 0.2]^{\mathrm{T}} \phi \geq 0\right\}$. Also define a "best fit rate" objective b_f as: $b_f = 1 - \sqrt{\frac{\sum_t \|y_t - \hat{y}_t\|^2}{\sum_t \|y_t - \bar{y}\|^2}}$, where $\bar{y}$ represents the numerical mean value of $\{y_t\}_{t \geq 0}$. In this system, simulated for $t \in [0, T]$, $T = 10000$, the method proposed in [15] achieves $b_f^1 = 0.6568$ in $\tau_1 = 52.28$ seconds [15]. The proposed method achieves $b_f^2 = 0.7792$, constructing $K = 6$ codevectors $\hat{s} = 2$ modes with parameter vectors $\bar{\theta}_1 = [0.07, 0.48, -0.40, 0.29, 0.00]^{\mathrm{T}}$ and $\bar{\theta}_2 = [0.13, 0.43, -0.03, 0.58, 0.01]^{\mathrm{T}}$. In the same desktop machine, the forward loop of the system including the identification computation overhead of the proposed algorithm, lasted a total of $\tau_2 = 16.13$ seconds. This allows real-time operation for systems of the form (53) sampled at frequency $f_s = 620$ Hz or lower, i.e., with sampling period $T_s = 0.0016$ sec or higher. Actual and predicted trajectories for system (53) for the first $T = 100$ timesteps are depicted in Fig. 4. Mode changes are depicted as background color.

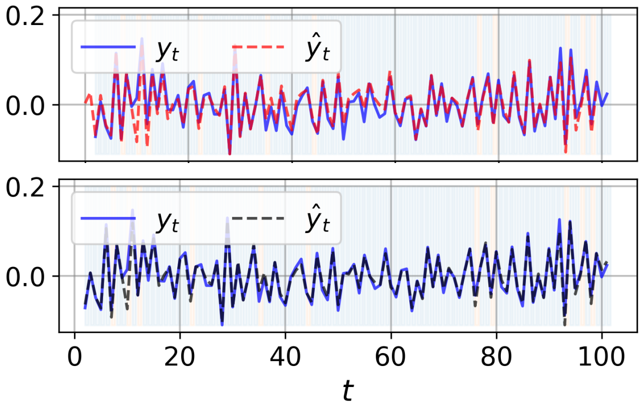

Fig. 4: Actual and predicted trajectories for system (53) using the proposed method (black) and the method in [15] (red).

C. State-Space PWA System

The properties of the proposed algorithm applied to state-space PWA systems are illustrated in this section. A higher-dimensional example can be found in [24]. Consider the following linearized PWA dynamics in the state-space domain:

$$\begin{cases} x_{t+1} &= (I_2 + \mathrm{dt} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}) x_t + \mathrm{dt} \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_t + e_t, \ |u_t| > 1 \\ x_{t+1} &= (I_2 + \mathrm{dt} \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}) x_t + \mathrm{dt} \begin{bmatrix} 0 \\ 0 \end{bmatrix} u_t + e_t, \ |u_t| \leq 1 \end{cases}, \tag{54}$$

where $x_t \in \mathbb{R}^2$, $u_t \in \mathbb{R}$, and $e_t \sim \mathcal{N}(0, 0.5)$. System (54) has two modes ($s = 2$) and the switching signal is defined by the polyhedral regions $R_1 = \left\{[x^{\mathrm{T}} | u^T]^{\mathrm{T}} \in \mathbb{R}^3 : u < -1\right\}$, $R_2 = \left\{[x^{\mathrm{T}} | u^T]^{\mathrm{T}} \in \mathbb{R}^3 : -1 < u < 1\right\}$, and $R_3 = \left\{[x^{\mathrm{T}} | u^T]^{\mathrm{T}} \in \mathbb{R}^3 : 1 < u\right\}$ with $S_1 = R_1 \bigcup R_3$ and $S_2 = R_2$. The dynamics of (54) consist of a controllable double integrator when the input is of sufficient magnitude, and a stable autonomous system, otherwise. In this example, the linear system of the second mode ($s = 2$) is not minimal, and its identification relies on the mode switching behavior of the system, as explained in Section II-B. To preserve the PE conditions of Assumption 4, the input signal is chosen as $u_t = 2\cos(2\pi t * \mathrm{dt})$, $t \in \mathbb{Z}_+$, and the noise term w_t is a zero-mean Gaussian random variable with $\sigma^2 = 0.1$.

The system is allowed to run for $T = 3s$ (seconds), with $\mathrm{dt} = 0.01$, i.e., a total of $N = 300$ observations are acquired online, based on which, the proposed method identifies the switched system in real time. The temperature parameters used for the online deterministic annealing algorithm are $(\lambda_{\max}, \lambda_{\min}, \gamma) = (0.99, 0.1, 0.8)$, $\delta = 0.1$, $\epsilon_n = 0.5$, and $\epsilon_s = 0.01$, and $(\alpha(t), \beta(t)) = ({}^1/_{1+0.01t}, {}^1/_{1+0.9t \log t})$. The estimated parameter $\hat{\theta}_1$ gets updated by the iterations (39). We have assumed $\hat{\theta}_1(0) = [0, 1, 1, 0, 1, 1]^{\mathrm{T}}$. A total of $K = 4$ effective codevectors and $\hat{s} = 2$ modes are estimated.

The identification error and the estimated mode switching error are shown in Fig. 5 in comparison with the true mode switching behavior of the system. More specifically, the algorithm identifies a total of $\hat{s} = 2$ modes with $S_1 = \Sigma_3 \bigcup \Sigma_4$ and $S_2 = \Sigma_1 \bigcup \Sigma_2$. In Figure 6, the convergence of the parameters $\left\{\bar{\theta}_i\right\}$ of each of the $\hat{s} = 2$ local models detected to the actual $\{\theta_i\}_{i=1}^2$ observed are shown. Parameter values that do not appear at $t = 0$ indicate that they belong to modes identified through the bifurcation phenomenon after a certain critical temperature value.

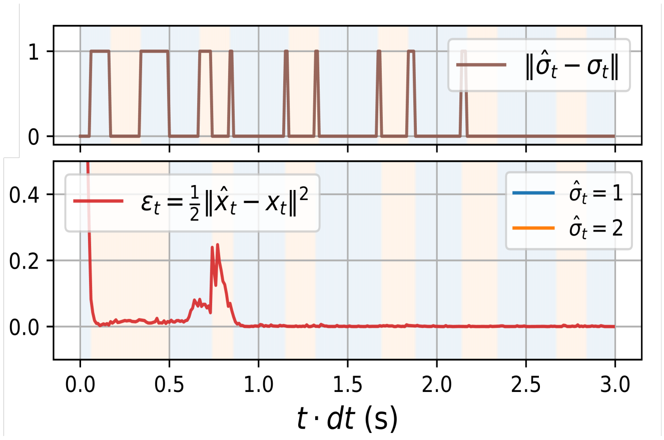

Fig. 5: Identification error over time for system (54). The estimated modes are also compared against the original modes.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

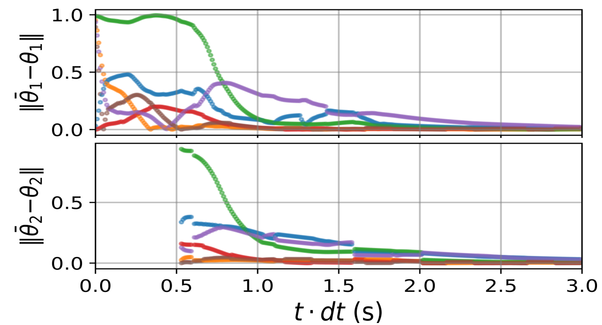

Fig. 6: Convergence of the parameters $\{\bar{\theta}_i\}_{i=1}^2$ to the true values of (54). Parameter values that do not appear at $t = 0$ indicate that belong to modes identified through the bifurcation phenomenon described in Section IV-B.

VIII. Conclusion

A real-time system identification scheme is proposed, appropriate for online identification of both the modes and the subsystems of a discrete-time switched system. The proposed method is computationally efficient compared to standard algebraic, mixed-integer programming, and offline clustering-based methods, and provides real-time control over the performance-complexity trade-off. Future directions will focus on the development of an adaptive annealing schedule with respect to time-dependent changes and extensions to identification of both discrete- and continuous-time partially observable piece-wise affine models in the state-space domain using real-time observations.

References

[1] A. Garulli, S. Paoletti, and A. Vicino, "A survey on switched and piecewise affine system identification," *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 344–355, 2012.

[2] D. Liberzon, *Switching in Systems and Control*. Springer, 2003, vol. 190.

[3] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal, "Identification of hybrid systems a tutorial," *European Journal of Control*, vol. 13, no. 2-3, pp. 242–260, 2007.

[4] A. Moradvandi, R. E. Lindeboom, E. Abraham, and B. De Schutter, "Models and methods for hybrid system identification: a systematic survey," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 95–107, 2023.

[5] R. Vidal, A. Chiuso, and S. Soatto, "Observability and identifiability of jump linear systems," in *IEEE Conference on Decision and Control*, vol. 4, 2002, pp. 3614–3619.

[6] M. Petreczky, L. Bako, and J. H. Van Schuppen, "Realization theory of discrete-time linear switched systems," *Automatica*, vol. 49, no. 11, pp. 3337–3344, 2013.

[7] J. Roll, A. Bemporad, and L. Ljung, "Identification of piecewise affine systems via mixed-integer programming," *Automatica*, vol. 40, no. 1, pp. 37–50, 2004.

[8] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari, "A clustering technique for the identification of piecewise affine systems," *Automatica*, vol. 39, no. 2, pp. 205–217, 2003.

[9] C. Mavridis and J. S. Baras, "Annealing optimization for progressive learning with stochastic approximation," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 2862–2874, 2023.

[10] F. Bianchi, A. Falsone, L. Piroddi, and M. Prandini, "An alternating optimization method for switched linear systems identification," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1071–1076, 2020.

[11] L. Bako, "Identification of switched linear systems via sparse optimization," *Automatica*, vol. 47, no. 4, pp. 668–677, 2011.

[12] D. Piga, A. Bemporad, and A. Benavoli, "Rao-Blackwellized sampling for batch and recursive Bayesian inference of piecewise affine models," *Automatica*, vol. 117, p. 109002, 2020.

[13] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems," in *IEEE International Conference on Decision and Control*, vol. 1, 2003, pp. 167–172.

[14] J. Wang, C. Song, J. Zhao, and Z. Xu, "A PWA model identification method for nonlinear systems using hierarchical clustering based on the gap metric," *Computers & Chemical Engineering*, vol. 138, p. 106838, 2020.

[15] A. Bemporad, V. Breschi, D. Piga, and S. P. Boyd, "Fitting jump models," *Automatica*, vol. 96, pp. 11–21, 2018.

[16] M. Gegundez, J. Aroba, and J. M. Bravo, "Identification of piecewise affine systems by means of fuzzy clustering and competitive learning," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 8, pp. 1321–1329, 2008.

[17] H. Nakada, K. Takaba, and T. Katayama, "Identification of piecewise affine systems based on statistical clustering technique," *Automatica*, vol. 41, no. 5, pp. 905–913, 2005.

[18] A. L. Juloski, S. Weiland, and W. M. H. Heemels, "A Bayesian approach to identification of hybrid systems," *IEEE Transactions on Automatic Control*, vol. 50, no. 10, pp. 1520–1533, 2005.

[19] L. Bako, K. Boukharouba, E. Duviella, and S. Lecoeuche, "A recursive identification algorithm for switched linear/affine models," *Nonlinear Analysis: Hybrid Systems*, vol. 5, no. 2, pp. 242–253, 2011.

[20] Y. Du, F. Liu, J. Qiu, and M. Buss, "A novel recursive approach for online identification of continuous-time switched nonlinear systems," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 15, pp. 7546–7565, 2021.

[21] C. N. Mavridis and J. S. Baras, "Identification of piecewise affine systems with online deterministic annealing," in *IEEE Conference on Decision and Control*, 2023, pp. 4885–4890.

[22] C. N. Mavridis, A. Kanellopoulos, J. S. Baras, and K. H. Johansson, "State-space piece-wise affine system identification with online deterministic annealing," in *European Control Conference*, 2024, pp. 3110–3115.

[23] F. Lauer, G. Bloch, F. Lauer, and G. Bloch, "Hybrid system identification," *Hybrid System Identification: Theory and Algorithms for Learning Switching Models*, pp. 77–101, 2019.

[24] C. Mavridis and K. H. Johansson, "Real-time hybrid system identification with online deterministic annealing," *arXiv preprint arXiv:2408.01730*, 2024.

[25] X. Tang and Y. Dong, "Expectation maximization based sparse identification of cyberphysical system," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 6, pp. 2044–2060, 2021.

[26] M. Yu, F. Bianchi, and L. Piroddi, "A randomized method for the identification of switched NARX systems," *Nonlinear Analysis: Hybrid Systems*, vol. 49, p. 101364, 2023.

[27] S. Paoletti, A. Garulli, J. Roll, and A. Vicino, "A necessary and sufficient condition for input-output realization of switched affine state space models," in *IEEE Conference on Decision and Control*, 2008, pp. 935–940.

[28] S. Paoletti, J. Roll, A. Garulli, and A. Vicino, "On the input-output representation of piecewise affine state space models," *IEEE Transactions on Automatic Control*, vol. 55, no. 1, pp. 60–73, 2009.

[29] M. Petreczky, L. Bako, and J. H. van Schuppen, "Identifiability of discrete-time linear switched systems," in *ACM International Conference on Hybrid Systems: Computation and Control*, 2010, pp. 141–150.

[30] C. N. Mavridis and J. S. Baras, "Online deterministic annealing for classification and clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7125–7134, 2023.

[31] C. N. Mavridis, N. Suriyarachchi, and J. S. Baras, "Maximum-entropy progressive state aggregation for reinforcement learning," in *IEEE Conference on Decision and Control*, 2021, pp. 5144–5149.

[32] C. N. Mavridis and J. S. Baras, "Progressive graph partitioning based on information diffusion," in *IEEE Conference on Decision and Control*, 2021, pp. 37–42.

[33] C. N. Mavridis, N. Suriyarachchi, and J. S. Baras, "Detection of dynamically changing leaders in complex swarms from observed dynamic data," in *International Conference on Decision and Game Theory for Security*. Springer, 2020, pp. 223–240.

[34] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[35] C. Mavridis, E. Noorani, and J. S. Baras, "Risk sensitivity and entropy regularization in prototype-based learning," in *Mediterranean Conference on Control and Automation*, 2022, pp. 194–199.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3619711

[36] C. Mavridis and J. Baras, “Multi-resolution online deterministic annealing: A hierarchical and progressive learning architecture,” 2023. [Online]. Available: https://arxiv.org/abs/2212.08189

[37] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with Bregman divergences,” *Journal of Machine Learning Research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[38] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2009, vol. 48.

[39] F. Lauer and G. Bloch, “Switched and piecewise nonlinear hybrid system identification,” in *Hybrid Systems: Computation and Control: 11th International Workshop, HSCC 2008, St. Louis, MO, USA, April 22-24, 2008. Proceedings 11*. Springer, 2008, pp. 330–343.

[40] B. M. Jenkins, A. M. Annaswamy, E. Lavretsky, and T. E. Gibson, “Convergence properties of adaptive systems and the definition of exponential stability,” *SIAM Journal on Control and Optimization*, vol. 56, no. 4, pp. 2463–2484, 2018.

[41] B. Anderson and J. Moore, “New results in linear system stability,” *SIAM Journal on Control*, vol. 7, no. 3, pp. 398–414, 1969.

APPENDIX A
PROOF OF THEOREM 1.

We construct the system

$$\hat{x}_{t+1} = \hat{A} x_t + \hat{B} u_t, \quad t \in \mathbb{Z}_+, \tag{55}$$

where $\hat{A} \in \mathbb{R}^{n\times n}$, and $\hat{B} \in \mathbb{R}^{n\times p}$. Subtracting (7) from (55), we get:

$$e_{t+1} = \bar{\Theta} r_t, \quad t \in \mathbb{Z}_+, \tag{56}$$

where $e_t = \hat{x}_t - x_t \in \mathbb{R}^n$ is the observation error, $r_t = [x_t^{\mathrm{T}} | u_t^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n+p}$ is the augmented state-input vector as defined in (5), and $\bar{\Theta} = [(\hat{A} - A)|(\hat{B} - B)]$ is an augmented matrix of the system parameters of size $n \times (n+p)$. Then (9) is equivalent to:

$$\bar{\Theta}_{t+1} = \bar{\Theta}_t - \gamma e_{t+1} r_t^{\mathrm{T}}, \quad t \geq 0. \tag{57}$$

Notice that (57) can be written in the form of a linear time-varying dynamical system:

$$\bar{\Theta}_{t+1} = \bar{\Theta}_t (I_{n+p} - \gamma r_t r_t^{\mathrm{T}}), \ t \geq 0. \tag{58}$$

By vectorizing $\bar{\Theta}_t$ such that $\bar{\theta}_t = \mathrm{vec}(\bar{\Theta}_t)$, (58) becomes:

$$\bar{\theta}_{t+1} = (I_{n(n+p)} - \gamma \psi_t \psi_t^{\mathrm{T}}) \bar{\theta}_t = \Xi_t \bar{\theta}_t, \ t \geq 0, \tag{59}$$

where $\otimes$ denotes the Kronecker product, and $\psi_t = [r_t^{\mathrm{T}} \otimes I_n]^{\mathrm{T}}$ is a $n(n+p)\times n$ matrix. We will show that (59) is exponentially stable in the large (Definition 1, [40]) as long as (8) is satisfied. Consider the Lyapunov function candidate $V(t, \bar{\theta}) = \bar{\theta}_t^{\mathrm{T}} \bar{\theta}_t$. It is obvious that there exist $k_1, k_2 > 0$ such that $k_1 \|\bar{\theta}\|^2 \leq V(t, \bar{\theta}) \leq k_2 \|\bar{\theta}\|^2$. Notice that $V(t+1, \bar{\theta}_{t+1}) - V(t, \bar{\theta}_t) = \bar{\theta}_t^{\mathrm{T}} \Xi_t^{\mathrm{T}} \Xi^t \bar{\theta}_t$. As a result, by summing the differences for T timesteps, we get:

$$\begin{aligned}
V(t{+}T+1, \bar{\theta}_{t+T+1}) - V(t, \bar{\theta}_t) = \\
&= \sum_{\tau=t}^{t+T} V(\tau+1, \bar{\theta}_{\tau+1}) - V(\tau, \bar{\theta}_\tau) \\
&= \sum_{\tau=t}^{t+T} \bar{\theta}_\tau^{\mathrm{T}} \left(\Xi_\tau^{\mathrm{T}} \Xi_\tau - I_{n(n+p)} \right) \bar{\theta}_\tau \\
&= \bar{\theta}_t^{\mathrm{T}} \left[\sum_{\tau=t}^{t+T} \Phi(\tau;t)^{\mathrm{T}} \left(\Xi_\tau^{\mathrm{T}} \Xi_\tau - I_{n(n+p)} \right) \Phi(\tau;t) \right] \bar{\theta}_\tau \\
&\leq -\alpha_1 \bar{\theta}_t^{\mathrm{T}} I_{n(n+p)} \bar{\theta}_t = -\alpha_1 V(t, \bar{\theta}_t),
\end{aligned} \tag{60}$$

for some $0 < \alpha_1 < 1$. Here $\Phi(\tau;t) = \Xi_t \Xi_{t+1} \ldots \Xi_{\tau-1}$ is the transition matrix of (59), and the inequality follows from condition (8). Notice that the first inequality in (8) is equivalent to $\alpha I_{n+p} \preceq \sum_{\tau=t}^{t+T} r_\tau^{\mathrm{T}} r_\tau$ and directly implies that $\alpha_2 I_{n(n+p)} \preceq \sum_{\tau=t}^{t+T} \psi_\tau^{\mathrm{T}} \psi_\tau$, for some $\alpha_2 > 0$, as well. As a result $\sum_{\tau=t}^{t+T} \Xi_\tau^{\mathrm{T}} \Xi_\tau \preceq \alpha_3 T I_{n(n+p)}$ for some $0 < \alpha_3 < 1$, and, therefore, $\sum_{\tau=t}^{t+T} \left(\Xi_\tau^{\mathrm{T}} \Xi_\tau - I_{n(n+p)} \right) \preceq -\alpha_4 T I_{n(n+p)}$ for some $0 < \alpha_4 < 1$. Finally this implies that $\left[\sum_{\tau=t}^{t+T} \Phi(\tau;t)^{\mathrm{T}} \left(\Xi_\tau^{\mathrm{T}} \Xi_\tau - I_{n(n+p)} \right) \Phi(\tau;t) \right] \leq -\alpha_1 I_{n(n+p)}$ for some $0 < \alpha_1 < 1$ [41]. Notice that the second inequality of (8) is necessary to ensure non-singularity of the transition matrix $\Phi(\tau;t)$ [40]. Finally, as an immediate result of (60), $V(t+T+1, \bar{\theta}_{t+T}+1) \leq (1-\alpha_1) V(t, \bar{\theta}_t)$, $\forall t \geq 0$, which implies uniform asymptotic stability in the large, and, due to linearity, exponential stability in the large.

Christos N. Mavridis received his Diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2017, and the M.S. and Ph.D. degrees in electrical and computer engineering at the University of Maryland, College Park, MD, in 2021. His research interests include stochastic optimization, learning theory, hybrid systems, and control theory,.

He is currently a postdoc at KTH Royal Institute of Technology, Stockholm, and has been affiliated as a research scientist with the Institute for Systems Research (ISR), University of Maryland, MD, the Nokia Bell Labs, NJ, the Xerox Palo Alto Research Center (PARC), CA, and Ericsson AB, Stockholm.

Dr. Mavridis is an IEEE member, and a member of IEEE/CSS Technical Committee on Security and Privacy. He has received the A. James Clark School of Engineering Distinguished Graduate Fellowship and the Ann G. Wylie Dissertation Fellowship in 2017 and 2021, respectively. He has been a finalist in the Qualcomm Innovation Fellowship US, San Diego, CA, 2018, and he has received the Best Student Paper Award in the IEEE International Conference on Intelligent Transportation Systems (ITSC), 2021.

Karl H. Johansson is Swedish Research Council Distinguished Professor in Electrical Engineering and Computer Science at KTH Royal Institute of Technology in Sweden and Founding Director of Digital Futures. He earned his MSc degree in Electrical Engineering and PhD in Automatic Control from Lund University. He has held visiting positions at UC Berkeley, Caltech, NTU and other prestigious institutions. His research interests focus on networked control systems and cyber-physical systems with applications in transportation, energy, and automation networks. For his scientific contributions, he has received numerous best paper awards and various distinctions from IEEE, IFAC, and other organizations. He has been awarded Distinguished Professor by the Swedish Research Council, Wallenberg Scholar by the Knut and Alice Wallenberg Foundation, Future Research Leader by the Swedish Foundation for Strategic Research. He has also received the triennial IFAC Young Author Prize and IEEE CSS Distinguished Lecturer. He is the recipient of the 2024 IEEE CSS Hendrik W. Bode Lecture Prize. His extensive service to the academic community includes being President of the European Control Association, IEEE CSS Vice President Diversity, Outreach & Development, and Member of IEEE CSS Board of Governors and IFAC Council. He has served on the editorial boards of Automatica, IEEE TAC, IEEE TCNS and many other journals. He has also been a member of the Swedish Scientific Council for Natural Sciences and Engineering Sciences. He is Fellow of both the IEEE and the Royal Swedish Academy of Engineering Sciences.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 12,2025 at 21:43:56 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.