2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Identification of Piecewise Affine Systems with Online Deterministic Annealing

Christos N. Mavridis, and John S. Baras

Abstract— **We propose a new online identification scheme for discrete-time piece-wise affine models based on a system of adaptive algorithms. A stochastic approximation algorithm based on online deterministic annealing runs at a slow timescale, estimating the partition of the space that defines the modes of the system. At the same time, a recursive identification algorithm, running at a higher timescale, updates the parameters of local identification models based on the estimate of the modes. Convergence results under mild assumptions are given based on the theory of two timescale stochastic approximation. In contrast to standard identification algorithms for piecewise affine systems, the proposed approach is appropriate for online system identification using sequential data acquisition, and is computationally more efficient compared to standard algebraic, mixed-integer programming, and clustering-based methods. The progressive nature of the algorithm provides real-time control over the performance-complexity trade-off, desired in practical applications. Experimental results validate the efficacy of the proposed methodology.**

I. Introduction

Switched and Piece-Wise Affine (PWA) systems constitute a class of universal approximation models with important applications in identification, verification, and control synthesis of non-linear, interconnected linear, and hybrid systems [1]–[3]. They are modeled as a collection of affine dynamical systems, often called modes, indexed by a discrete-valued switching variable that depends on a partitioning of the state-input domain into a finite number of polyhedral regions [1], [2]. As such, PWA models have universal approximation properties and can be used to describe hybrid and nonlinear phenomena that are frequent in practical situations [1], [3]. For this reason, identification of PWA systems has been widely investigated in recent years.

Most existing identification approaches for SARX (Switched ARX) systems can be categorized by the problem formulation as optimization-based [4], algebraic [5], [6], or clustering-based [7]–[9], and by the the method used as offline [7] or recursive [6], [10]. Algebraic methods are based on transforming the SARX model to a "lifted" ARX model that does not depend on the switching sequence [5], [6]. Offline optimization-based methods often rely on solving a large mixed integer program that can be tractable only for small data sets [4], or relaxation techniques over the same problem [10]. Finally, clustering-based methods are optimization-based methods that make use of unsupervised learning techniques to estimate the partition of the domain that is needed for the switching signal [7]–[9], [11], [12]. However, most such 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.

In this work, we follow an adaptive clustering-based method to identify a PWARX system from online input-output observations. The estimation of the partition defining the switching signal is based on a Voronoi tessellation with respect to a progressively growing set of codevectors that are computed using an online deterministic annealing learning algorithm [13]. The key idea is to solve a sequence of optimization sub-problems using fast, online, and gradient-free stochastic approximation updates that simulate a dynamical system [13]–[15]. This process progressively estimates the optimal Vornoi tessellation and simulates an annealing process that induces a series of bifurcation phenomena (phase transitions), according to which, the number of codevectors is adjusted [13], [16], thus estimating the number of modes in a PWARX system. Adopting the above adaptive partitioning framework, we develop an online identification scheme for discrete-time PWARX models on a system of adaptive algorithms running in two timescales. A stochastic approximation algorithm based on online deterministic annealing runs at a slow timescale estimating the partition of the space that defines the switching signal, as well as the number of modes (Section III). At the same time, a second stochastic approximation algorithm based on standard recursive system identification methods, running at a higher timescale, updates the parameters of the local models based on the estimate of the switching signal (Section IV-A). The convergence properties of this system of recursive algorithms are studied through the theory of two timescale stochastic approximation (Section IV-B). In contrast to standard identification algorithms for piece-wise affine systems, the proposed approach is appropriate for online system identification using sequential data acquisition, and is computationally more efficient compared to standard algebraic, mixed-integer programming, and clustering-based methods. In addition, the progressive nature of the algorithm provides real-time control over the performance-complexity trade-off, desired in practical applications. Simulation results validate the efficacy of the proposed approach.

II. Switched and Piecewise Affine Systems

A switched affine system is a collection of affine systems, indexed by a discrete-valued switching signal, that share the

The authors are with the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park, USA. emails:{mavridis, baras}@umd.edu.

Research partially supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00111990027, by ONR grant N00014-17-1-2622, and by a grant from Northrop Grumman Corporation.

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10383998

same state. A discrete-time switched affine system in state-space form is described by:

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

where $x_t \in \mathbb{R}^n$ is the state vector of the system, $u_t \in \mathbb{R}^p$ is the input, $y_t \in \mathbb{R}^q$ is the output, and $w_t \in \mathbb{R}^n$ and $v_t \in \mathbb{R}^q$ are noise terms. The signal $\sigma_t \in \{1, \dots, s\}$ represents the discrete state of the system and defines the mode (affine dynamics) which is active at time t. The matrices $A_i \in \mathbb{R}^{n \times n}$, $B_i \in \mathbb{R}^{n \times p}$, $C_i \in \mathbb{R}^{q \times n}$, $D_i \in \mathbb{R}^{q \times p}$, $f_i \in \mathbb{R}^n$, and $g_i \in \mathbb{R}^q$ define the affine dynamics for each mode $i \in \{1, \dots, s\}$. The discrete state σ_t can be either an exogenous input, e.g. triggered by some event, or a function of the system state and input. In particular, when σ_t is defined according to a polyhedral partition of the state and input space, i.e., when

$$\sigma_t = i \iff \begin{bmatrix} x_t \\ u_t \end{bmatrix} \in R_i \subset R, \tag{2}$$

where $\{R_i\}_{i=1}^s$ are convex polyhedra defining a complete partition of the state-input domain $R \subseteq \mathbb{R}^{n+p}$, the switched system is called Piece-Wise Affine (PWA).

Switched systems can be expressed in input-output form as Switched AutoRegressive eXogenous (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:

$$y_t^{(i)} = \theta_{\sigma_t}^{(i)\mathrm{T}} \begin{bmatrix} r_t \\ 1 \end{bmatrix} + e_t^{(i)}, \ i = 1, \dots, q \tag{3}$$

where $r_t \in \mathbb{R}^d$, $d = qn_a + p(n_b + 1)$, is a regression vector given by

$$r_t = [y_{t-1}^{\mathrm{T}} \dots y_{t-n_a}^{\mathrm{T}} u_t^{\mathrm{T}} u_{t-1}^{\mathrm{T}} \dots u_{t-n_b}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^d, \tag{4}$$

$\theta_j^{(i)} \in \mathbb{R}^{d+1}$, $j \in \{1, \dots, s\}$, are the parameter vectors that define each ARX mode, and $e_t \in \mathbb{R}^q$ is a noise term. To simplify the notation, and without loss of generality, in the rest of the paper we assume that $q = 1$. The vector

$$\phi_t = \begin{bmatrix} r_t \\ 1 \end{bmatrix} \in \mathbb{R}^{d+1} \tag{5}$$

is referred to as the extended regression vector. Similarly, a Piece-Wise Affine ARX system (PWARX) is defined according to a polyhedral partition of $P \subseteq \mathbb{R}^d$ as the nonlinear (piece-wise linear) model:

$$y_t = \begin{cases} \theta_1^{\mathrm{T}} \phi_t + e_t, & \text{if } r_t \in P_1 \\ \vdots & \vdots \\ \theta_s^{\mathrm{T}} \phi_t + e_t, & \text{if } r_t \in P_s \end{cases} \tag{6}$$

where $P_i \subset P$, is a polyhedron in $\mathbb{R}^d$, $P_i \cap P_j = \emptyset$ for $i \neq j$, and $\bigcup_i P_i = P$.

A. Identification of PWARX models

In this work we will focus on identification of PWARX models in the input-output form (6). Necessary and sufficient conditions for input-output realization of SARX and PWARX systems are given in [17], and [18], respectively. Future extensions will include identification of PWA systems in state-space form along the lines of [19]. Under certain identifiability conditions, the general identification problem for a PWARX system as given in (6) can be formulated as a stochastic optimization problem over the parameters $\{n_a, n_b, s, \{\theta_i\}_{i=1}^s, \{P_i\}_{i=1}^s\}$ as follows:

$$\min_{n_a, n_b, s, \{\theta_i\}, \{P_i\}} \mathbb{E}\left[\sum_{i=1}^{s} \mathbb{1}_{[r \in P_i]} d(y, \theta_i^{\mathrm{T}} \phi)\right] \tag{7}$$

where the nonnegative measure d is an appropriately defined dissimilarity measure, and the expectation is taken with respect to $(y, r) \in \mathbb{R}^{q+d}$, i.e., the input-output pairs. Problem (7) is generally intractable. Notice that the optimization parameters n_a and n_b representing the model order, and s representing the number of modes, completely alter the number and the domain of θ_i, $i \in \{1, \dots, s\}$ that represent the dynamics of the system with $\theta_i \in \mathbb{R}^d$, $d = qn_a + p(n_b + 1)$. In addition, a parametric representation for the polyhedral regions $P_i, i \in \{1, \dots, s\}$, that form a partition of $P \subseteq \mathbb{R}^d$ satisfying $P_i \subset P$, $P_i \cap P_j = \emptyset$ for $i \neq j$, and $\bigcup_i P_i = P$, should be defined. Finally, if the probability distribution of the error e_t is not assumed known, or of a convenient form, the expectation operation cannot be analytically computed. For these reasons we make the following assumptions.

Assumption 1: We assume that upper bounds $(\tilde{n}_a, \tilde{n}_b)$ on the orders of the model (n_a, n_b), are known.

Assumption 2: For each mode, we assume access to a set of independent observations $\{(\hat{y}_t, \hat{u}_t)\}_{t=1}^N$, $N > \max\{n_a, n_b\}$, of the input-output pairs of the system, which represent realizations of the random variables $(y, u) \in \mathbb{R}^{q+p}$.

Knowledge of the bounds $(\tilde{n}_a, \tilde{n}_b)$ will allow us to concentrate on the properties of PWARX model identification in a constant parameter domain where the highest possible orders (n_a, n_b) are chosen relative to potential computational bounds. In Section III we will propose a recursive algorithm to estimate both the number of modes s and the partition $\{P_i\}_{i=1}^s$ given that the parameters $\{\theta_i\}_{i=1}^s$ are known. Then, in Section IV-A we will review recursive system identification techniques to estimate $\{\theta_i\}_{i=1}^s$ given that s and $\{P_i\}_{i=1}^s$ are known. Finally in Section IV-B we will show that the two recursive systems can be combined using the theory of two-timescale stochastic approximation.

III. Adaptive Partitioning with Online Deterministic Annealing

In this section we will adopt a clustering method to solve the problem of finding s and $\{P_i\}_{i=1}^s$ given that $\{\theta_i\}_{i=1}^s$ are known. In Section IV-B we will show how the proposed methodology can be combined with recursive updates on the parameters $\{\theta_i\}_{i=1}^s$. We introduce a set of variables $\{\rho_i\}_{i=1}^K$,

$\rho_i \in P$ each one representing a region

$$\Sigma_i = \left\{ r \in P : i = \arg\min_j d(r, \rho_j) \right\} \tag{8}$$

for a given dissimilarity measure d. The measure d can be designed such that the Voronoi regions Σ_i are polyhedral, e.g., Euclidean distance or any Bregman divergence, as will be explained in Section III-A. In this sense, each P_i can be mapped to a region Σ_j (for $K = s$) or a union of adjacent sets $\{\Sigma_j\}$ (for $K > s$), as will be explained in Section IV-B.

Problem (7) then becomes a clustering problem:

$$\min_{\{\rho_i\}} \mathbb{E}\left[\sum_{i=1}^{K} \mathbb{1}_{[r \in \Sigma_i]} d(X, \mu_i) \right] \tag{9}$$

on the augmented space of the random variable:

$$X = \begin{bmatrix} \theta \\ r \end{bmatrix} \in \mathbb{S} \subseteq R^{2d+1} \tag{10}$$

defined in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where μ_i is the augmented codevector:

$$\mu_i := \begin{bmatrix} \hat{\theta}_i \\ \rho_i \end{bmatrix} \in S, \ i = 1, \ldots, K, \tag{11}$$

with $\hat{\theta}_i$ being an estimate of θ_i (so far we assume $\hat{\theta}_i = \theta_i$). Here the measure $d : S \times S \to [0, \infty)$ is a dissimilarity measure defined on S. Problem (9) is a hard clustering problem with respect to the parameters $\{\rho_i\}_{i=1}^K$. The lowest possible number K should also be computed.

A. Online Deterministic Annealing

To construct a recursive stochastic optimization algorithm to solve problem (9) and progressively estimate the number K of the augmented codevectors $\{\mu_i\}_{i=1}^K$, we adopt the online deterministic annealing approach introduced in [13]. Recall that the observed data are represented by the random variable $X : \Omega \to S \subseteq \mathbb{R}^{2d+1}$ in (10) defined in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and the augmented codevectors $\{\mu_i\}_{i=1}^K$ are treated as constant parameters to be estimated. According to the online deterministic annealing principles [13], [15], we extend this approach and define a probability space over an infinite number of codevectors, while constraining their distribution using a maximum-entropy principle at different levels. First we define a quantizer $Q : S \to \mathrm{ri}(S)$ as a discrete random variable in the same probability space with countably infinite domain $\mu := \{\mu_i\}$. Then we formulate the multi-objective optimization:

$$\min_{\mu} F_\lambda(\mu) := (1 - \lambda) D(\mu) - \lambda H(\mu), \ \lambda \in [0, 1), \tag{12}$$

where the term

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

takes the place of the objective in (9), and

$$\begin{aligned} H(\mu) &:= \mathbb{E}\left[-\log P(X, Q) \right] \\ &= H(X) - \int p(x) \sum_i p(\mu_i | x) \log p(\mu_i | x) \ \mathrm{d}x \end{aligned} \tag{13}$$

is the Shannon entropy. This is now a problem of finding the locations $\{\mu_i\}$ and the corresponding probabilities $\{p(\mu_i|x)\} := \{p(Q = \mu_i | X = x)\}$. The Lagrange multiplier $\lambda \in [0, 1)$ controls the trade-off between D and H. The entropy term, however, introduces several properties to the approach that can be useful in many applications [13], [15], [20]–[23]. First, it introduces robustness with respect to initial conditions [13], [24]. Second, as we will show in Section III-B, reducing the values of λ defines an annealing process [13], [16] and induces a bifurcation phenomenon that affects the number K of the codevectors.

To solve (12) 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

$$\begin{aligned} F_\lambda^*(\mu) := \min_{\{p(\mu_i|x)\}} & F_\lambda(\mu), \\ \text{s.t.} \sum_i & p(\mu_i|x) = 1 \end{aligned} \tag{14}$$

is given by the Gibbs distributions

$$p^*(\mu_i|x) = \frac{e^{-\frac{1-\lambda}{\lambda} d(x, \mu_i)}}{\sum_j e^{-\frac{1-\lambda}{\lambda} d(x, \mu_j)}}, \ \forall x \in S \tag{15}$$

In order to minimize $F^*(\mu)$ with respect to the codevector locations μ we set the gradients to zero:

$$\begin{aligned} & \frac{d}{d\mu} F_\lambda^*(\mu) = 0 \\ \implies & \frac{d}{d\mu} \left((1 - \lambda) D(\mu) - \lambda H(\mu) \right) = 0 \\ \implies & \sum_i \int p(x) p^*(\mu_i|x) \frac{d}{d\mu_i} d(x, \mu_i) \ \mathrm{d}x = 0 \end{aligned} \tag{16}$$

where we have used (15) and direct differentiation. Equation (16) has a closed-form solution if the dissimilarity measure d belongs to the family of Bregman divergences; information-theoretic measures that play an important role in learning applications and include the widely used Euclidean distance and Kullback-Leibler divergence [13], [25].

Remark 1: The partition $\{\Sigma_i\}$ induced by (8) and a dissimilarity measure d that belongs to the family of Bregman divergences, is separated by hyperplanes, such that each Σ_i is a polyhedral region for a bounded domain P [25].

Throughout this paper, we will assume that the dissimilarity measure d in (8) is a Bregman divergence. Then the following result holds:

Theorem 1 ([15]): If d is a Bregman divergence, then

$$\mu_i^* = \mathbb{E}\left[X | \mu_i \right] = \frac{\int x p(x) p^*(\mu_i|x) \ \mathrm{d}x}{p^*(\mu_i)} \tag{17}$$

is a sufficient solution for the the optimization problem

$$\min_{\mu} F^*(\mu) \tag{18}$$

where $F_\lambda^*(\mu)$ is the solution of (14).

Using Theorem 1, the following Lemma constructs a gradient-free stochastic approximation algorithm that recursively estimates the solution to problem (17):

Lemma 1 ([13]): The sequence $\mu_i(n)$ constructed by the recursive updates

$$\begin{cases} \rho_i(t+1) & = \rho_i(t) + \beta(t)\left[\hat{p}(\mu_i|x_t) - \rho_i(t)\right] \\ \sigma_i(t+1) & = \sigma_i(t) + \beta(t)\left[x_t\hat{p}(\mu_i|x_t) - \sigma_i(t)\right] \end{cases} \tag{19}$$

where $x_t \sim X$, $\sum_t \beta(t) = \infty$, $\sum_t \beta^2(t) < \infty$, and the quantities $\hat{p}(\mu_i|x_t)$ and $\mu_i(t)$ are recursively updated as follows:

$$\mu_i(t) = \frac{\sigma_i(t)}{\rho_i(t)}, \quad \hat{p}(\mu_i|x_t) = \frac{\rho_i(t)e^{-\frac{1-\lambda}{\lambda}d(x_t,\mu_i(t))}}{\sum_i \rho_i(t)e^{-\frac{1-\lambda}{\lambda}d(x_t,\mu_i(t))}}, \tag{20}$$

converges almost surely to a solution of (17).

Remark 2: Notice that we can express the dynamics of the codevector parameters $\mu_i(t)$ directly as:

$$\mu_i(t+1) = \frac{\beta(t)}{\rho_i(t)}\left[\frac{\sigma_i(t+1)}{\rho_i(t+1)}(\rho_i(t) - \hat{p}(\mu_i|x_t)) + (x_t\hat{p}(\mu_i|x_t) - \sigma_i(t)\right] \tag{21}$$

where the recursive updates take place for every codevector μ_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 $\hat{p}(\mu_i|x_t)$ in (20). According to this phenomenon, the number of distinct values of μ_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 III-B.

B. Bifurcation and The Number of Modes

In Section III-A we describe how to solve the optimization problem for a given value of the parameter λ. To define an online deterministic annealing approach, we solve a sequence of optimization problems with decreasing values of λ. This process grants λ the name of a 'temperature' parameter. Notice that, so far, we have assumed a countably infinite set of codevectors. We will show that the unique values of the set $\{\mu_i\}$ that solves (12), form a finite set $K(\lambda)$ of values that we will refer to as "effective codevectors" throughout this paper, and will define the estimated number of modes s.

Notice that at high temperature ($\lambda \to 1$), (15) yields uniform association probabilities $p(\mu_i|x) = p(\mu_j|x),\ \forall i,j,\forall x$, and as a result of (17), all pseudo-inputs are located at the same point $\mu_i = \mathbb{E}[X],\ \forall i$ which means that there is one unique "effective" codevector given by $\mathbb{E}[X]$. As λ is lowered below a critical value, a bifurcation phenomenon occurs, when the number of "effective" codevectors increases, which describes an annealing process [13], [16]. Mathematically, this occurs when the existing solution μ^* given by (17) 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 bifurcation by the condition:

$$\frac{d^2}{d\epsilon^2}F^*(\{\mu + \epsilon\psi\})\Big|_{\epsilon=0} \geq 0 \tag{22}$$

for all choices of finite perturbations $\{\psi\}$. Using (22) and direct differentiation, we can show that bifurcation depends on the temperature coefficient λ (and the choice of the Bregman divergence, through the function ϕ) [15], [26]. In other words, the number of codevectors increases countably many times as the value of λ decreases, and an algorithmic implementation needs only as many codevectors in memory as the number of "effective" codevectors. In practice. we can detect the bifurcation points by introducing perturbing pairs of pseudo-inputs at each temperature level λ. The codevectors μ are doubled by inserting a perturbation of each μ_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 pseudocode for the online determinsitic annealing algorithm and a detailed discussion on its implementation, complexity, parameter sensitivity, can be found in [13], [15], [26].

1) Estimating a minimum number of modes: According to Remark 1, the partition $\{S_i\}$ of S defined by the rule $S_i = \left\{x \in S : i = \arg\min_j d(x,\mu_j)\right\}$ is polyhedral for a bounded domain S. It follows that the partition $\{\Sigma_i\}$ of P defined in (8) is also polyhedral, as each Σ_i can be expressed as a low-dimensional projection of S_i. Therefore, each region P_i in (6) can be mapped to a region Σ_j, if the number of effective codevectors is $K = s$, or a union of adjacent sets $\{\Sigma_j\}$ (for $K > s$). The design of an appropriate termination criterion such that $K = s$ and the identification error is minimized, is not straightforward. Instead, it is often expected that $K > s$. In this case, the inverse process of increasing the temperature parameter λ to merge adjacent sets Σ_i, Σ_j if $\frac{1-\lambda}{\lambda}d(\mu^j,\mu^i) < \epsilon_n$, $i \neq j$, for some parameter $\epsilon_n > 0$ can be followed.

IV. Piecewise Affine System Identification

In this section we review recursive system identification techniques for estimating the parameters θ_i of the local models given knowledge of the partition $\{P_i\}$. Furthermore, we formulate these methods as stochastic approximation methods and show that they can be combined with the stochastic approximation method of estimating $\{P_i\}$ by $\{\Sigma_i\}$ as proposed in Section III.

A. Identification of Local Models

Recall that each local model of the PWARX system in (6) is completely defined by the parameters $\{\theta_i\}$. According to Assumption 2, we assume access to a set of observations $\{(y_t,u_t)\}_{t=1}^N$, $N > \max\{n_a,n_b\}$, of the input-output pairs of the system, which represent realizations of the random variables $(y,u) \in \mathbb{R}^{q+p}$. In the following, we study a stochastic gradient descent and a recursive least-mean-squares identification method to estimate $\left\{\hat{\theta}_i\right\}$. First we define the error:

$$\epsilon(t) = y_t - \hat{\theta}_i^{\mathrm{T}}(t)\phi_t \tag{23}$$

1) Stochastic Gradient Descent: A stochastic gradient descent approach aims to minimize the error:

$$\min_{\hat{\theta}_i} \frac{1}{2}\mathbb{E}\left[\|\epsilon(t)\|^2\right] \tag{24}$$

using the recursive updates:

$$\begin{aligned}\hat{\theta}_i(t+1) &= \hat{\theta}_i(t) - \alpha(t)\left(\nabla_{\theta_i}\epsilon(t)\right)\epsilon(t)^{\mathrm{T}} \\ &= \hat{\theta}_i(t) + \alpha(t)\phi_t\epsilon^{\mathrm{T}}(t)\end{aligned} \tag{25}$$

where $\sum_n \alpha(n) = \infty$, $\sum_n \alpha^2(n) < \infty$. This is a stochastic approximation sequence of the form:

$$\hat{\theta}_i(t+1) = \hat{\theta}_i(t) + \alpha(t)\left[h(\hat{\theta}_i(t)) + M(t+1)\right],\ t \geq 0, \tag{26}$$

where $h(\hat{\theta}_i) = -\nabla\mathbb{E}\left[\|\epsilon(t)\|^2\right]$, and $M(t+1) = \nabla\mathbb{E}\left[\|\epsilon(t)\|^2\right] - \nabla\|\epsilon(t)\|^2$ is a Martingale difference sequence according to Assumption 2 (and under mild assumptions on the existence of the expectation and continuity of the error signal). This sequence converges almost surely to the equillibrium of the differential equation:

$$\dot{\hat{\theta}}_i = h(\hat{\theta}_i),\ t \geq 0, \tag{27}$$

which can be shown to be a solution of (24) with standard Lyapunov arguments.

2) Recursive Least-Mean-Squared Estimation: Stochastic gradient descent is a greedy approach that often converges to poor local minima. An alternative approach is to minimize the error:

$$\min_{\hat{\theta}_i} \frac{1}{2}\mathbb{E}\left[\sum_{\tau=1}^{t}\|\epsilon(\tau)\|^2\right], \tag{28}$$

which can be obtained by setting the gradient to zero (and under mild assumptions on the existence of the expectation and continuity of the error signal) as:

$$\theta_i^*(t) = \mathbb{E}\left[\left\{\sum_{\tau=1}^{t}\phi_t\phi_t^{\mathrm{T}}\right\}^{-1}\sum_{\tau=1}^{t}\phi_t y_t^{\mathrm{T}}\right] \tag{29}$$

We can approximate $\psi_t = \left\{\sum_{\tau=1}^{t}\phi_t\phi_t^{\mathrm{T}}\right\}^{-1}\sum_{\tau=1}^{t}\phi_t y_t^{\mathrm{T}}$, using the least-mean-squared recursive updates:

$$\begin{aligned}\psi_{t+1} &= \psi_t + \frac{p_t}{1+\phi_{t+1}^{\mathrm{T}}p_t\phi_{t+1}}\phi_{t+1}\epsilon^{\mathrm{T}}(t+1) \\ p_{t+1} &= p_t - \frac{p_t\phi_{t+1}\phi_{t+1}^{\mathrm{T}}p_t}{1+\phi_{t+1}^{\mathrm{T}}p_t\phi_{t+1}}\end{aligned} \tag{30}$$

Then, the stochastic approximation sequence:

$$\hat{\theta}_i(t+1) = \hat{\theta}_i(t) + \alpha(t)(\psi_t - \hat{\theta}_i(t)) \tag{31}$$

where $\sum_n \alpha(n) = \infty$, $\sum_n \alpha^2(n) < \infty$ converges almost surely to the solution $\theta_i^*(t) = \mathbb{E}[\psi_t]$, as it is a stochastic approximation approach of the form (26) with $h(\hat{\theta}_i) = \mathbb{E}[\psi_t] - \hat{\theta}_i$, and $M(t+1) = \psi_t - \mathbb{E}[\psi_t]$ being a Martingale difference sequence according to Assumption 2. Similarly, the convergence of this sequence can be studied with standard Lyapunov arguments on the differential equation (27).

B. Combined Partitioning and Local Model Identification

Notice that the estimation updates of the number of modes s and the partition $\{\Sigma_i\}_{i=1}^s$ in (21) is a stochastic approximation algorithm with a stepsize schedule $\beta(t)$. At the same time, the recursive system identification techniques to estimate $\{\theta_i\}_{i=1}^s$ given $\{\Sigma_i\}_{i=1}^s$ in (25) and (31) are stochastic approximation sequences with a stepsize schedule $\alpha(t)$. The two recursive systems can be combined using the theory of two-timescale stochastic approximation if ${}^{\beta(t)}/_{\alpha(t)} \to 0$, i.e., the estimation of the partition $\{\Sigma_i\}_{i=1}^s$ is updated at a slower rate than the updates of the parameters $\{\theta_i\}_{i=1}^s$. This follows directly from Theorem 2 in [15]. In practice, the condition ${}^{\beta(t)}/_{\alpha(t)} \to 0$ is satisfied by stepsizes of the form $(\alpha(t), \beta(t)) = ({}^1/_t, {}^1/_{1+t\log t})$, or $(\alpha(t), \beta(t)) = ({}^1/_{t^{2/3}}, {}^1/_t)$.

V. Experimental Results

We illustrate the properties and evaluate the performance of the proposed algorithm in the following PWA system:

$$y_t = \begin{cases}\theta_1^{\mathrm{T}}\phi_t + e_t, & \text{if } r_t \in P_1 \\ \theta_2^{\mathrm{T}}\phi_t + e_t, & \text{if } r_t \in P_2 \\ \theta_3^{\mathrm{T}}\phi_t + e_t, & \text{if } r_t \in P_3\end{cases} \tag{32}$$

where $y_t \in \mathbb{R}^1$, $r_t \in P = [-4, 4]$, ϕ_t is defined by (5), $(P_1, P_2, P_3) = ([-4,-1], (-1,2), [2,4])$, and $(\theta_1, \theta_2, \theta_3) = ([1,2]^{\mathrm{T}}, [-1,0]^{\mathrm{T}}, [1,2]^{\mathrm{T}})$, as in [7]. The simplicity of this example allows graphical representation of the signaling partition and the convergence of the model parameters. At the same time, it is a switching system that presents a jump at $r_t = 2$, and same dynamics for different regions of the input space, i.e., $\theta_1 = \theta_3$ while $P_1 \neq P_3$.

A total of $N = 150$ observations under Gaussian noise ($e_t \sim N(0, 0.2)$) are accessible in a sequential manner.

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 $(\alpha(t), \beta(t)) = ({}^1/_{1+0.01t}, {}^1/_{1+0.9t\log t})$. At first ($\lambda = \lambda_{\max}$), the algorithm keeps in memory only one codevector ρ_1 and one model parameter vector $\hat{\theta}_1$, essentially assuming that the system has constant dynamics in the entire domain, i.e., $\Sigma_1 = P_1 = P$. As new input-output pairs are observed, the estimated parameter $\hat{\theta}_1$ gets updated by the iterations (30), (31). We have assumed $\hat{\theta}_1(0) = [1,1]^{\mathrm{T}}$. At the same time, the estimate of $\hat{\theta}_1$ are used to update the location of the codevector towards the mean of the observation domain as shown in (17). The converged values of the parameters for $\lambda = \lambda_{\max}$ are used as initial conditions for the next value of λ. As λ is reduced, the bifurcation phenomenon described in Section III-B takes place, and, after reaching a critical value, the single codevector splits into two duplicates. This process continues until the minimum temperature parameter $\lambda_{\min}$ is reached, reflecting a potential time and computational constraint of the system. The bifurcation phenomenon is illustrated in Fig. 1 where the locations of the codevectors $\{\rho_i\}$, $\rho_i \in P = [-4, 4]$ are shown, constructing a total of $K = 5$ effective codevectors. The number of modes is accurately estimated with the inverse process explained in

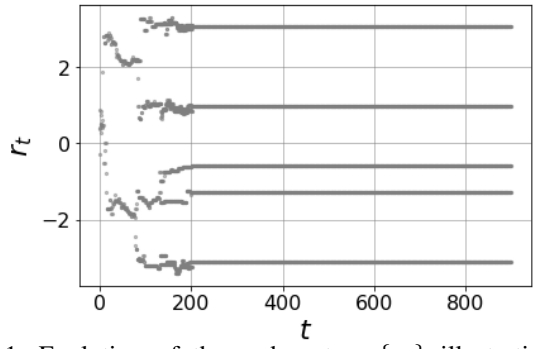

Fig. 1: Evolution of the codevectors $\{\rho_i\}$ illustrating the bifurcation phenomenon described in Section III-B.

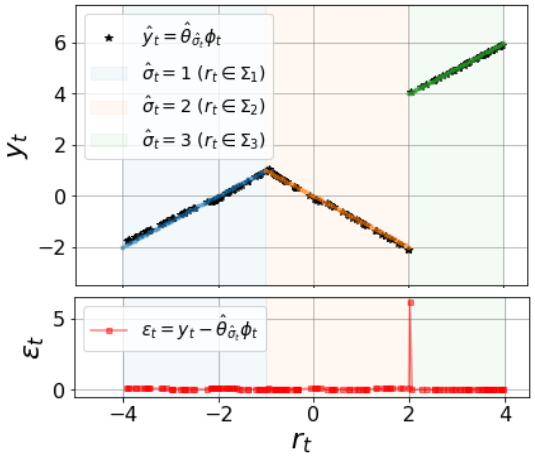

Fig. 2: Estimated partition, output, and error with respect to the true model. A single misclassification instance of the mode appears at the boundary of the true partition.

Section III-B.1. 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. 2.

VI. Conclusion and Future Work

We developed a novel online identification scheme for discrete-time piece-wise affine models based on a system of recursive algorithms. In contrast to standard identification algorithms for piece-wise affine systems, the proposed approach is appropriate for online system identification using sequential data acquisition, and is computationally more efficient compared to existing methods. The progressive nature of the algorithm also provides real-time control over the performance-complexity trade-off. Future directions include extensions of the proposed approach for real-time identification of both discrete- and continuous-time partially observable piece-wise affine models in the state-space domain.

References

[1] D. Liberzon, *Switching in systems and control*. Springer, 2003, vol. 190.

[2] 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.

[3] 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.

[4] 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.

[5] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems," in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 1. IEEE, 2003, pp. 167–172.

[6] R. Vidal, "Recursive identification of switched arx systems," *Automatica*, vol. 44, no. 9, pp. 2274–2287, 2008.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] K. Boukharouba, L. Bako, and S. Lecoeuche, "Identification of piecewise affine systems based on dempster-shafer theory," *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 1662–1667, 2009.

[12] R. Baptista, J. Y. Ishihara, and G. A. Borges, "Split and merge algorithm for identification of piecewise affine systems," in *Proceedings of the 2011 American Control Conference*. IEEE, 2011, pp. 2018–2023.

[13] C. N. Mavridis and J. S. Baras, "Online deterministic annealing for classification and clustering," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[14] ——, "Convergence of stochastic vector quantization and learning vector quantization with bregman divergences," *IFAC-PapersOnLine*, vol. 53, no. 2, 2020.

[15] 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.

[16] 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.

[17] 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 *2008 47th IEEE Conference on Decision and Control*. IEEE, 2008, pp. 935–940.

[18] 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.

[19] C. N. Mavridis, A. Kanellopoulos, K. Vamvoudakis, J. S. Baras, and K. H. Johansson, "Attack identification for cyber-physical security in dynamic games under cognitive hierarchy," *IFAC-PapersOnLine*, 2023.

[20] 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.

[21] 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.

[22] C. N. Mavridis, G. P. Kontoudis, and J. S. Baras, "Sparse gaussian process regression using progressively growing learning representations," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 1454–1459.

[23] 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.

[24] C. Mavridis, E. Noorani, and J. S. Baras, "Risk sensitivity and entropy regularization in prototype-based learning," in *2022 30th Mediterranean Conference on Control and Automation (MED)*. IEEE, 2022, pp. 194–199.

[25] 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.

[26] C. Mavridis and J. Baras, "Multi-resolution online deterministic annealing: A hierarchical and progressive learning architecture," *arXiv preprint arXiv:2212.08189*, 2022.