Identification of hybrid systems: a tutorial

Simone Paoletti, Aleksandar Lj. Juloski, Giancarlo Ferrari-Trecate and René Vidal

Abstract— **This tutorial paper is concerned with the identification of hybrid models, i.e. dynamical models whose behavior is determined by interacting continuous and discrete dynamics. Methods specifically aimed at the identification of models with a hybrid structure are of very recent date. After discussing the main issues and difficulties connected with hybrid system identification, and giving an overview of the related literature, this paper focuses on four different approaches for the identification of switched affine and piecewise affine models, namely an algebraic procedure, a Bayesian procedure, a clustering-based procedure, and a bounded-error procedure. The main features of the selected procedures are presented, and possible interactions to still enhance their effectiveness are suggested.**

I. Introduction

Hybrid systems are heterogeneous dynamical systems whose behavior is determined by interacting continuous and discrete dynamics. The continuous dynamics is described by variables taking values from a continuous set, while the discrete dynamics is described by variables taking values from a discrete, typically finite, set. The continuous or discrete-valued variables may depend on independent variables such as time, which in turn may be continuous or discrete. Some of the variables can also be discrete-event driven in an asynchronous manner.

Hybrid systems arise not only from the interaction of logic devices and continuous processes. They can be used to describe real phenomena that exhibit discontinuous behaviors. For instance, the trajectory of a bouncing ball results from the alternation between free fall and elastic contact. Moreover, hybrid models can be used to approximate continuous phenomena by concatenating different models from a simple class. For instance, a nonlinear dynamical system can be approximated by switching among various linear models.

Due to their many potential applications, hybrid systems have attracted increasing attention in the control community during the last decade. Numerous results on analysis, verification, computation, stability and control of hybrid systems have appeared in the literature. However, most of the theoretical developments hinge on the assumption that a hybrid model of the process at hand is available. In some situations it is possible to obtain such a model starting from first principles. On the other hand, first principles modelling is too complicated or even impossible to apply in most practical situations, and the model needs to be identified on the basis of experimental data.

S. Paoletti is with Dipartimento di Ingegneria dell'Informazione, Università di Siena, Via Roma 56, 53100 Siena, Italy. E-mail: paoletti@dii.unisi.it.

A. Lj. Juloski is with Department of Electrical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. E-mail: A.Juloski@tue.nl.

G. Ferrari-Trecate is with Dipartimento di Informatica e Sistemistica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy. E-mail: giancarlo.ferrari@unipv.it.

R. Vidal is with Center for Imaging Science, Department of Biomedical Engineering, Johns Hopkins University, 3400 N. Charles St., Baltimore MD 21218, USA. E-mail: rvidal@cis.jhu.edu.

A. Paper contribution

In the first part, this paper introduces the topic of hybrid system identification by focusing in particular on the identification of switched affine and PieceWise Affine (PWA) models. PWA systems are a class of hybrid systems obtained by partitioning the state-input domain into a finite number of non-overlapping convex polyhedral regions, and by considering linear/affine subsystems in each region [66]. Since PWA models are equivalent to several classes of hybrid models [4], [34], [67], PWA system identification techniques are suitable to obtain hybrid models from data. Moreover, the universal approximation properties of PWA maps [14], [49] make PWA models attractive also for nonlinear system identification [64].

Identification of PWA models is a challenging problem that involves the estimation of both the parameters of the affine submodels, and the coefficients of the hyperplanes defining the partition of the state-input domain (or the regressors domain, for models in input-output form). The main difficulty lies in the fact that the identification problem includes a classification problem where each data point must be associated to the most suitable submodel. Concerning the partitioning, two alternative approaches can be distinguished:

1) the partition is fixed a priori;
2) the partition is estimated along with the submodels.

In the first case, data classification is very simple, and estimation of the submodels can be carried out by resorting to standard linear identification techniques. In the second case, the regions must be shaped to the clusters of data, and the strict relation among data classification, parameter estimation and region estimation makes the identification problem very hard to cope with. The problem is even harder when also the number of submodels must be estimated.

Different techniques leading to PWA models of smooth dynamical systems can be found in the extensive literature on nonlinear black-box identification. A nice overview is presented in [61]. However, most of these approaches assume that the system dynamics is continuous. Recently, novel contributions allowing for discontinuities have been proposed in both the hybrid systems and the nonlinear identification communities. An iterative algorithm that sequentially estimates the parameters of the model and classifies the data through the use of adapted weights is described in [60]. A

method based on statistical clustering of measured data via a Gaussian mixture model and support vector classifiers is presented in [56]. Several optimization problem formulations of the identification problem are proposed in [54], [55]. In [62] the identification problem is formulated for two subclasses of PWA models, namely Hinging Hyperplane ARX (HHARX) and Wiener PWARX (W-PWARX) models, and solved via mixed-integer linear or quadratic programs. Subspace identification of piecewise linear systems is addressed in [10], [71], while recursive identification of switched hybrid systems is addressed in [32], [75].

Among the proposed approaches, contributions of the authors of this paper are represented by four different procedures for the identification of switched affine and piecewise affine models, namely the algebraic procedure [78], the clustering-based procedure [27], the Bayesian procedure [47], and the bounded-error procedure [5]. These techniques have been successfully applied in several real problems, such as the identification of the electronic component placement process in pick-and-place machines [5], [43], [47], the modelling of a current transformer [27], traction control [11], and motion segmentation in computer vision [76], [77]. The main features of the selected techniques are summarized in the second part of the paper. Possible interactions to still enhance their effectiveness are also suggested.

B. Paper outline

This paper is organized as follows. Section II introduces the classes of switched affine and piecewise affine models, both in state space and input-output form. Section III reports several formulations of the identification problem for these model classes, and presents an overview of the related literature. Different identification approaches are classified along the lines proposed in [61]. The problems of data classification and region estimation are addressed in Section IV for those approaches that firstly classify the data, then estimate the affine dynamics, and finally reconstruct the polyhedral partition. Most recent contributions for the identification of models with hybrid and discontinuous characteristics belong to this category. Four procedures falling into the category analyzed in Section IV, are finally described and discussed in Section V. Section VI draws the conclusions, and foreshadows interesting topics for future research.

II. Switched affine and piecewise affine models

Switched affine models are defined as collections of linear/affine models, connected by switches that are indexed by a discrete-valued additional variable, called the discrete state. Models for which the discrete state is determined by a polyhedral partition of the state-input domain, are called piecewise affine models. They can be used to model a large number of physical processes (see, e.g. [3], [17], [18], [48], [69]), and are suitable to approximate virtually any nonlinear dynamics, e.g., via multiple linearizations at different operating points. Moreover, piecewise affine models are equivalent to several classes of hybrid models, and can therefore be used to describe systems exhibiting hybrid structure.

A. Models in state space form

A discrete-time switched affine model in *state space* form is described by the equations

$$\begin{aligned} \boldsymbol{x}_{k+1} &= A_{\sigma(k)}\,\boldsymbol{x}_k + B_{\sigma(k)}\,\boldsymbol{u}_k + f_{\sigma(k)} + \boldsymbol{w}_k \\ \boldsymbol{y}_k &= C_{\sigma(k)}\,\boldsymbol{x}_k + D_{\sigma(k)}\,\boldsymbol{u}_k + g_{\sigma(k)} + \boldsymbol{v}_k, \end{aligned} \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^n$, $\boldsymbol{u}_k \in \mathbb{R}^p$ and $\boldsymbol{y}_k \in \mathbb{R}^q$ are, respectively, the (*continuous*) state, the input and the output of the system at time $k \in \mathbb{Z}$, and $\boldsymbol{w}_k \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^q$ are noise/error terms. The *discrete* state $\sigma(k)$, describing in what affine dynamics the system is at time k, is assumed to take only a finite number of values, i.e. $\sigma(k) \in \{1, \dots, s\}$, where s is the number of affine submodels. In general, $\sigma(k)$ can be a function of k, $\boldsymbol{x}_k$, $\boldsymbol{u}_k$, or some other external input. The real matrices/vectors A_i, B_i, f_i, C_i, D_i and g_i, $i = 1, \dots, s$, having appropriate dimensions, describe each affine dynamics. Hence, model (1) can be seen as a collection of affine models with continuous state $\boldsymbol{x}_k$, connected by switches that are indexed by the discrete state $\sigma(k)$.

The evolution of the discrete state can be described in a variety of ways. In *Jump Linear* (JL) models, $\sigma(k)$ is an unknown, deterministic and finite-valued input. In *Jump-Markov Linear* (JML) models, the dynamics of $\sigma(k)$ is modelled as an irreducible Markov chain governed by the transition probabilities $\pi(i,j) \triangleq P\big(\sigma(k+1) = j \,\big|\, \sigma(k) = i\big)$. In *PieceWise Affine* (PWA) models [66], $\sigma(k)$ is given by the rule

$$\sigma(k) = i \quad \text{iff} \quad \left[\begin{smallmatrix} \boldsymbol{x}_k \\ \boldsymbol{u}_k \end{smallmatrix}\right] \in \Omega_i, \quad i = 1, \dots, s, \tag{2}$$

where $\{\Omega_i\}_{i=1}^s$ is a complete partition[1] of the state-input domain $\Omega \subseteq \mathbb{R}^{n+p}$. The regions Ω_i are assumed to be convex polyhedra described by

$$\Omega_i = \Big\{ \left[\begin{smallmatrix} \boldsymbol{x} \\ \boldsymbol{u} \end{smallmatrix}\right] \in \mathbb{R}^{n+p} \,:\, \bar{H}_i \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{u} \\ 1 \end{bmatrix} \preceq_{[i]} \boldsymbol{0} \Big\}, \tag{3}$$

where $\bar{H}_i \in \mathbb{R}^{\bar{\mu}_i \times (n+p+1)}$, $i = 1, \dots, s$, and $\bar{\mu}_i$ is the number of linear inequalities defining the ith polyhedral region Ω_i. With abuse of notation, in (3) the symbol $\preceq_{[i]}$ denotes a μ_i-dimensional vector whose elements can be the symbols $\leq$ and $<$ in order to avoid that the regions Ω_i overlap over common boundaries.

Remark 2.1: PWA models form a special class of hybrid models. Other descriptions for hybrid systems include *Mixed Logical Dynamical* (MLD) models [6], *Linear Complementarity* (LC) models [33], [70], *Extended Linear Complementarity* (ELC) models [20], and *Max-Min-Plus-Scaling* (MMPS) models [21]. Equivalences among these five classes of systems are proven in [4], [34]. Such results are very important for transferring theoretical properties and tools (e.g., control and identification techniques) from one class to another, as they imply that one can choose the most convenient hybrid modelling framework for the study of a particular hybrid system.

[1] A collection $\{\mathcal{A}_i\}_{i=1}^s$ is said to be a (complete) partition of $\mathcal{A} \subseteq \mathbb{R}^m$ if $\cup_{i=1}^s \mathcal{A}_i = \mathcal{A}$ and $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$, $\forall\, i \neq j$.

B. Models in input-output form

For fixed model orders n_a and n_b, a *Switched affine AutoRegressive eXogenous* (SARX) model is defined by introducing the regression vector

$$\boldsymbol{r}_k = [\ \boldsymbol{y}_{k-1}^\top\ \cdots\ \boldsymbol{y}_{k-n_a}^\top\ \boldsymbol{u}_k^\top\ \boldsymbol{u}_{k-1}^\top\ \cdots\ \boldsymbol{u}_{k-n_b}^\top\ ]^\top, \quad (4)$$

and then by expressing the output $\boldsymbol{y}_k$ as a piecewise affine function of $\boldsymbol{r}_k$, namely

$$\boldsymbol{y}_k = \theta_{\sigma(k)}^\top \left[\begin{smallmatrix} \boldsymbol{r}_k \\ 1 \end{smallmatrix}\right] + \boldsymbol{e}_k, \quad (5)$$

where $\sigma(k) \in \{1, \ldots, s\}$ is the discrete state, s is the number of submodels, θ_i, $i = 1, \ldots, s$, are the matrices of parameters defining each submodel, and $\boldsymbol{e}_k \in \mathbb{R}^q$ is a noise/error term. In the following, the vector $\varphi_k = \left[\begin{smallmatrix} \boldsymbol{r}_k \\ 1 \end{smallmatrix}\right]$ will be called the *extended* regression vector.

SARX models represent a subclass of the switched affine models (1), and can be easily transformed into that form by defining the continuous state as

$$\boldsymbol{x}_k = [\ \boldsymbol{y}_{k-1}^\top\ \cdots\ \boldsymbol{y}_{k-n_a}^\top\ \boldsymbol{u}_{k-1}^\top\ \cdots\ \boldsymbol{u}_{k-n_b}^\top\ ]^\top. \quad (6)$$

As for the models in state space form, the evolution of the discrete mode $\sigma(k)$ can be described in a variety of ways. In *PieceWise affine AutoRegressive eXogenous* (PWARX) models the switching mechanism is determined by a polyhedral partition of the regressors domain $\mathcal{R} \subseteq \mathbb{R}^d$, where $d = q \cdot n_a + p \cdot (n_b + 1)$. This means that for these models the discrete state $\sigma(k)$ is given by

$$\sigma(k) = i \quad \text{iff} \quad \boldsymbol{r}_k \in \mathcal{R}_i, \quad i = 1, \ldots, s, \quad (7)$$

where $\{\mathcal{R}_i\}_{i=1}^s$ is a complete partition of $\mathcal{R}$. Each region $\mathcal{R}_i$ is a convex polyhedron described by

$$\mathcal{R}_i = \left\{\boldsymbol{r} \in \mathbb{R}^d\ :\ H_i \left[\begin{smallmatrix} \boldsymbol{r} \\ 1 \end{smallmatrix}\right] \preceq_{[i]} \boldsymbol{0}\right\}, \quad (8)$$

where $H_i \in \mathbb{R}^{\mu_i \times (d+1)}$, $i = 1, \ldots, s$, μ_i is the number of linear inequalities defining the ith polyhedral region $\mathcal{R}_i$ and, as in (3), the symbol $\preceq_{[i]}$ denotes a μ_i-dimensional vector whose elements can be the symbols $\leq$ and $<$. In general, the shape of $\mathcal{R}$ reflects the physical constraints on the inputs and the outputs of the system. For instance, typical constraints on the output can be $\|\boldsymbol{y}_k\|_\infty \leq y_{max}$ or $\|\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\|_\infty \leq \Delta y_{max}$, where $\|\cdot\|_\infty$ is the infinity norm of a vector, and y_{max} and Δy_{max} are fixed bounds.

By introducing the piecewise affine map

$$f(\boldsymbol{r}) = \begin{cases} \theta_1^\top \varphi & \text{if} \quad H_1 \varphi \preceq_{[1]} \boldsymbol{0} \\ \vdots & \vdots \\ \theta_s^\top \varphi & \text{if} \quad H_s \varphi \preceq_{[s]} \boldsymbol{0}, \end{cases} \quad (9)$$

with $\varphi = \left[\begin{smallmatrix} \boldsymbol{r} \\ 1 \end{smallmatrix}\right]$, it will be useful to rewrite the model defined by (5), (7) and (8) as

$$\boldsymbol{y}_k = f(\boldsymbol{r}_k) + \boldsymbol{e}_k. \quad (10)$$

Remark 2.2: The PWA map (9) can be discontinuous along the boundaries defined by the polyhedra (8), as shown in Fig. 1. Though, for the sake of simplicity, in the following the subscript $[i]$ will be removed from the notation $\preceq_{[i]}$, one must always take care of the definition of the regions, to avoid that the PWA map is multiply defined over common boundaries of the regions $\mathcal{R}_i$.

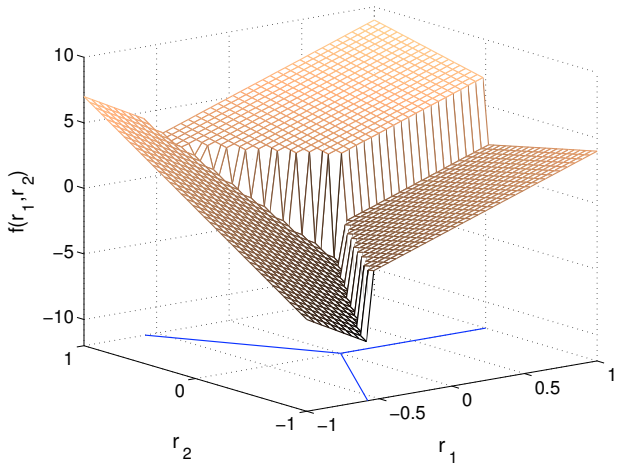

Fig. 1. Discontinuous PWA map of two variables with $s = 3$ regions.

III. Hybrid system identification

In this section, the identification problem will be firstly addressed for input-output models, and then for state space models. An overview of the related literature is finally presented. For the sake of clarity, single input-single output systems (i.e. $p = q = 1$) are considered. To this aim, notations y_k, u_k and e_k will be used instead of $\boldsymbol{y}_k$, $\boldsymbol{u}_k$ and $\boldsymbol{e}_k$. The discussion can be straightforwardly extended to multi input-single output systems (i.e. $p > 1$ and $q = 1$). Multi input-multi output systems (i.e. $p > 1$ and $q > 1$) are also handled by state-space techniques, while in the input-output case one can identify a model for each output by considering the other outputs as additional inputs[2].

A. Identification problem for SARX models

For SARX models (5), the general identification problem reads as follows.

Problem 3.1: Given a collection of N input-output pairs (y_k, u_k), $k = 1, \ldots, N$, estimate the model orders n_a and n_b, the number of submodels s, and the parameter vectors θ_i, $i = 1, \ldots, s$. Moreover, estimate the discrete state $\sigma(k)$ for $k > \max\{n_a, n_b\}$.

If the system generating the data has the structure (5), an exact algebraic solution to Problem 3.1 is presented in [51], [74], [78] for the case of noiseless data (though the approach can be amended to work also with noisy data). The algorithm only requires to fix upper bounds $\bar{n}_a$, $\bar{n}_b$, and $\bar{s}$ on the model orders and the number of submodels, respectively. A description of the algorithm will be given in Section V.

If the model orders are fixed, the problem is to fit the data to s hyperplanes. This problem is addressed in the field of

[2]Though this approach may lead in general to a larger number of regions than necessary, since the overall partition is obtained by intersecting the partitions of the single models.

data analysis, and several approaches are proposed where s is either estimated from data or fixed a priori. One way to estimate s is by solving the following problem.

Problem 3.2: Given $\delta > 0$, find the smallest number s of vectors θ_i, $i = 1, \dots, s$, and a mapping $k \mapsto \sigma(k)$ such that

$$|y_k - \varphi_k^\top \theta_{\sigma(k)}| \leq \delta \tag{11}$$

for all $k = \bar{n}, \dots, N$, where $\bar{n} = \max\{n_a, n_b\} + 1$.

Problem 3.2 consists in finding a *Partition* of the system of inequalities

$$|y_k - \varphi_k^\top \theta| \leq \delta\,, \quad k = \bar{n}, \dots, N, \tag{12}$$

into a *Minimum* number of *Feasible Subsystems* (MIN PFS problem). The bound δ in (12) is not necessarily given a priori (e.g., if the noise is bounded, and the bound is known), rather it can be adjusted in order to find the desired trade off between model accuracy and complexity. In fact, the smaller δ, the larger is typically the number of submodels needed to fit the data[3], while on the other hand, the larger δ, the worse is the fit, since larger errors are allowed. Figure 2 shows two typical plots of the number of submodels and the Mean Squared Error (MSE) as a function of δ when solving Problem 3.2 for a given data set. The choice of a suitable δ is typically made at the knee of the s-curve, where also the MSE is kept low. The MIN PFS problem is NP-hard, and a suboptimal greedy randomized algorithm to tackle its solution is proposed in [1].

If s is fixed, the well-known optimization approach used in linear system identification (i.e. choose the parameters of a linear model such that they minimize some prediction error norm) can be generalized to the identification of SARX models. Given a nonnegative function $\ell(\cdot)$, such as $\ell(\varepsilon) = \varepsilon^2$ or $\ell(\varepsilon) = |\varepsilon|$, the estimation of the parameter vectors θ_i, $i = 1, \dots, s$, and of the discrete state $\sigma(k)$ can be in fact formulated as the following optimization problem:

$$\begin{cases} \min\limits_{\theta_i,\, \chi_{k,i}} & \displaystyle\sum_{k=\bar{n}}^{N} \sum_{i=1}^{s} \ell\big(y_k - \varphi_k^\top \theta_i\big)\chi_{k,i} \\ s.t. & \displaystyle\sum_{i=1}^{s} \chi_{k,i} = 1 \qquad \forall\, k \\ & \chi_{k,i} \in \{0,1\} \quad \forall\, k, i. \end{cases} \tag{13}$$

In (13), each binary variable $\chi_{k,i}$ describes whether the data point $(y_k, \boldsymbol{r}_k)$ is associated to the ith submodel, under the constraint that each data point must be associated to only one submodel. The discrete state $\sigma(k)$ can be finally reconstructed according to the rule:

$$\sigma(k) = i \quad \text{iff} \quad \chi_{k,i} = 1. \tag{14}$$

The optimization problem in (13) is a *mixed integer* program that is computationally intractable, except for small instances. In principle, branch and bound algorithms could be applied, but the search tree increases exponentially with the number of data N and the number of submodels s.

[3]In this case overfit may occur, i.e. the model adjusts to the particular noise realization.

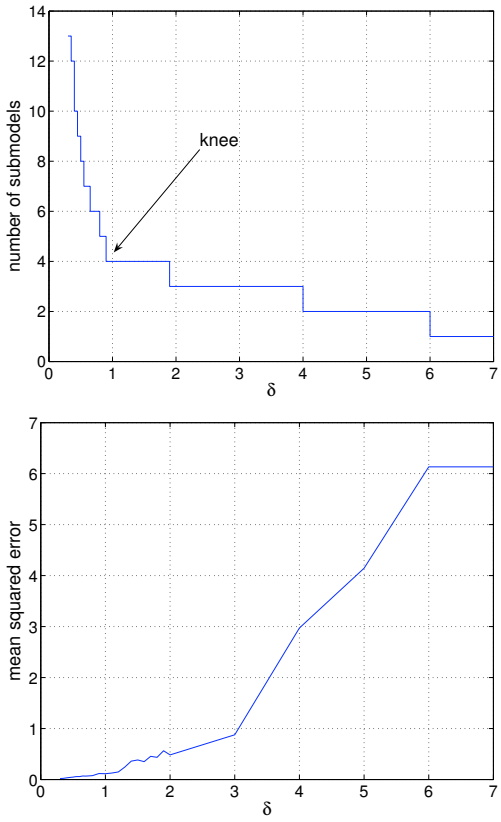

Fig. 2. Number of submodels and mean squared error as a function of δ for a data set generated by a SARX system with four discrete states and Gaussian additive noise with zero mean and variance $\sigma^2 = 0.1$.

It is shown in [55] that (13) can be transformed into a smooth constrained optimization problem by relaxing the integer constraints, i.e. by requiring $\chi_{k,i} \in [0,1]$, $\forall\, k, i$. The global optimum of the relaxed problem coincides with the global optimum of (13). Moreover, an integer solution can be readily obtained from the solution of the relaxed problem. By the same reasoning, it is also shown that (13) can be transformed into the following non-smooth unconstrained optimization problem:

$$\min_{\theta_i} \sum_{k=\bar{n}}^{N} \min_{i=1,\dots,s} \ell\big(y_k - \varphi_k^\top \theta_i\big). \tag{15}$$

In order to not get trapped in a local minimum, suitable optimization techniques must be used to tackle the solution of the equivalent problems. It is reported in [55] that state-of-the-art solvers, such as [38], are able to solve (15) in reasonable time at least for sample problems.

An alternative to the formulation (13) is the clustering algorithm proposed in [12], which groups the given data points into s clusters by generating s planes that represent a local solution to the non-convex problem of minimizing the sum of squares of the 2-norm distances between each point and a nearest plane.

B. Identification problem for PWARX models

For PWARX models defined by (5), (7) and (8), the general identification problem reads as follows.

Problem 3.3: Given a collection of N input-output pairs (y_k, u_k), $k = 1, \dots, N$, estimate the model orders n_a and n_b, the number of submodels s, the parameter vectors θ_i and the regions $\mathcal{R}_i$, $i = 1, \dots, s$.

Note that, in the case of piecewise affine models, the partition of the regressors domain automatically implies the estimation of the discrete state according to (7).

All techniques specifically developed for the identification of PWARX models, assume fixed orders n_a and n_b. The estimation of the model orders can be based on preliminary data analysis, and carried out by algebraic techniques such as [51], [74], or classical model order selection techniques (see [50]). Hence, in the following the orders n_a and n_b are given, and $\bar{n} = \max\{n_a, n_b\} + 1$.

The considered identification problem consists in finding the PWARX model that best matches the given data according to a specified criterion of fit. It involves the estimation of:

- The number of discrete states s.
- The parameters θ_i, $i = 1, \dots, s$, of the affine submodels.
- The coefficients H_i, $i = 1, \dots, s$, of the hyperplanes defining the partition of the regressors set.

This issue also underlies a classification problem such that each data point is associated to one region, and to the corresponding submodel. The simultaneous optimal estimation of all the quantities mentioned above is a very hard, computationally intractable problem. To the best of the authors' knowledge, no satisfactory formulation in the form of a single optimization problem has been even provided for it. One of the main concerns is how to choose s in a sensible way. For instance, perfect fit is obtained by letting $s = N$, i.e. one submodel per each data point, which is clearly an inadequate solution. Penalties on increasing s should be therefore introduced in order to keep the number of submodels reasonably low, and to avoid overfit because the model is given too many degrees of freedom. An additional difficulty is how to express efficiently the constraint that the collection $\{\mathcal{R}_i\}_{i=1}^{s}$ must form a complete partition of the regressors domain $\mathcal{R}$.

The problem becomes easy if the number of discrete states s is fixed, and the regions (8) are either known or fixed *a priori*. In that case each regression vector $\boldsymbol{r}_k$ can be associated to one submodel according to (7). Hence, by introducing the quantities

$$\chi_{k,\,i} = \begin{cases} 1 & \text{if } \boldsymbol{r}_k \in \mathcal{R}_i \\ 0 & \text{otherwise} \end{cases} \qquad \forall\, k, i, \tag{16}$$

the identification problem reduces to the following optimization problem:

$$\min_{\theta_i} \frac{1}{N} \sum_{k=\bar{n}}^{N} \sum_{i=1}^{s} \ell\big(y_k - \varphi_k^\top \theta_i\big)\chi_{k,\,i}, \tag{17}$$

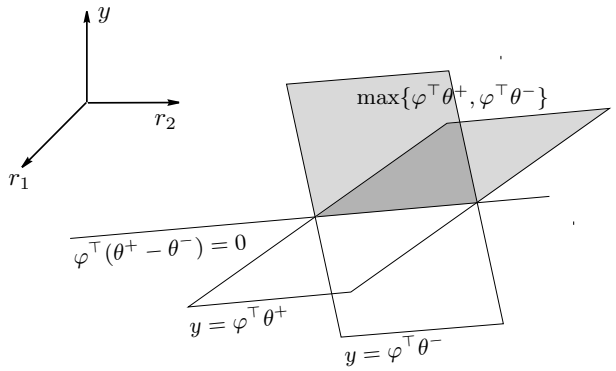

Fig. 3. Two hinging hyperplanes $y = \varphi^\top \theta^-$ and $y = \varphi^\top \theta^+$, and the corresponding hinge function $y = \max\{\varphi^\top \theta^+, \varphi^\top \theta^-\}$, where $\varphi = [\, r_1 \; r_2 \; 1 \,]^\top$.

where $\ell(\cdot)$ is a given nonnegative function. If $\ell(\varepsilon) = \varepsilon^2$, (17) is an ordinary least-squares problem in the unknowns θ_i.

In [61], [62] the identification problem is reformulated for the class of Hinging-Hyperplane ARX (HHARX) models [14], which are described by

$$\begin{aligned} y_k &= f(\boldsymbol{r}_k; \theta) + e_k \\ f(\boldsymbol{r}_k; \theta) &= \varphi_k^\top \theta_0 + \sum_{i=1}^{M} \sigma_i \max\{\varphi_k^\top \theta_i, 0\}, \end{aligned} \tag{18}$$

where $\theta = [\theta_0^\top \;\; \theta_1^\top \dots \theta_M^\top]^\top$, and $\sigma_i \in \{-1, 1\}$ are fixed a priori. It is easy to see that HHARX models are a subclass of PWARX models for which the PWA map (9) is continuous. The number of submodels s is bounded by the quantity $\sum_{j=0}^{d} \binom{M}{j}$, which only depends on the length d of the regression vector, and the number M of hinge functions (see Fig. 3). The identification problem considered in [62] selects the optimal parameter vector θ^* by solving

$$\theta^* = \arg\min_{\theta} \sum_{k=\bar{n}}^{N} |y_k - f(\boldsymbol{r}_k; \theta)|^p, \tag{19}$$

where $p = 1$ or 2. Assuming a priori known bounds on θ (which can be taken arbitrarily large), (19) can be reformulated as a mixed-integer linear or quadratic program (MILP/MIQP) by introducing auxiliary continuous variables $z_i(k) = \max\{\varphi_k^\top \theta_i, 0\}$, and binary variables

$$\delta_i(k) = \begin{cases} 0 & \text{if } \; \varphi_k^\top \theta_i \le 0 \\ 1 & \text{otherwise.} \end{cases} \tag{20}$$

The MILP/MIQP problems can then be solved for the global optimum. The optimality of the described approach comes at the cost of a theoretically very high worst-case computational complexity, which means that it is mainly suitable for small-scale problems (e.g., when it is very costly to obtain data). To be able to handle somewhat larger problems, different suboptimal approximations are proposed in [61]. Various extensions are also possible for handling non-fixed σ_i, discontinuities, general PWARX models, etc., again at the cost of increased computational complexity.

Most of the heuristic and suboptimal approaches that are applicable, or at least related, to the identification of PWARX models, either assume a fixed s, or adjust s iteratively (e.g., by adding one submodel at a time) in order to improve the fit. A few techniques allow for the automatic estimation of s from data. An overview of the related literature is presented in Section III-D.

C. Identification problem for state space models

For switched affine models defined by (1), or piecewise affine models defined by (1), (2) and (3), the general identification problem reads as follows.

Problem 3.4: Given a collection of N input-output pairs (y_k, u_k), $k = 1, \ldots, N$, estimate the model order n, the number of submodels s, and the 6-tuples $(A_i, B_i, f_i, C_i, D_i, g_i)$, $i = 1, \ldots, s$. Moreover, estimate the discrete state $\sigma(k)$, $k = 1, \ldots, N$, and, if the model is piecewise affine, the regions Ω_i, $i = 1, \ldots, s$.

As for the models in input-output form, the difficulty of Problem 3.4 depends on which quantities are assumed to be known. Nevertheless, while for SARX/PWARX models the identification problem is easy if all the quantities (including the switching sequence) are known, and only the parameters of the submodels must be estimated, an additional difficulty arises when dealing with the identification of state space models. If the switching sequence is known, the matrices of each submodel can still be estimated by classical techniques such as subspace identification methods. However, as pointed out in [71], the matrices of the submodels are obtained up to a linear state transformation. This state transformation is different, in general, for each of the submodels. To combine the submodels they need to be transformed into the same state basis. In [71] it is discussed how the transitions between the submodels can be used to this aim. The algorithm requires a sufficiently large number of transitions for which the states at the transition are linearly independent.

Heuristics and suboptimal techniques for the identification of switched and piecewise affine state space models are summarized in the next subsection.

D. Literature overview

In this subsection, an overview of different approaches to the identification of switched affine and piecewise affine models is presented. The description is not intended to be exhaustive, and the interested reader is referred to [61] for additional details. The list of references in [61] is completed here with most recent contributions.

1) Switched affine models: Emphasis on the identification of SARX models is put in the contributions [51], [74], [78], where an algebraic procedure for the estimation of the model orders, the number of discrete state and the model parameters, is proposed. The identification of SARX models is also considered in [58], where it is assumed that switchings occur with a certain probability at each time step, and [72], [73], where identification schemes for multi-mode and Markov models are developed. Switched affine models in state space form are considered in [10], [36], [71]. While in [71] the discrete state is assumed to be known, and the focus is mainly on determining the state transformations to express all the submodels in the same state basis (see Section III-C), in [10] the number of discrete states and the switching times are estimated from data. In both contributions, subspace identification techniques are used to identify the individual submodels. In [36], the estimation of the model orders, the number of submodels and the switching times is carried out by embedding the input-output data into a higher dimensional space, where the problem becomes the one of segmenting the data into distinct subspaces.

2) Piecewise affine models: Work on regression with PWA maps can be found in many fields, such as neural networks, electrical networks, time-series analysis, function approximation. Most of the related approaches assume that the system dynamics is continuous. Indeed, enabling the estimation of discontinuous models is a key feature of algorithms specifically designed for hybrid system identification. This is motivated by the fact that logic conditions can be represented through discontinuities in the state-update and output maps of the identified PWA model.

Remark 3.1: If the PWA map is assumed to be continuous, the model parameters and the partition of the domain are not independent. For instance, consider the PWA map (9) with $s = 2$. If (9) is continuous, at the switching surface between the two modes it must hold that $\theta_1^\top \begin{bmatrix} \boldsymbol{r} \\ 1 \end{bmatrix} = \theta_2^\top \begin{bmatrix} \boldsymbol{r} \\ 1 \end{bmatrix}$, and hence $\boldsymbol{r}$ must satisfy

$$(\theta_1 - \theta_2)^\top \begin{bmatrix} \boldsymbol{r} \\ 1 \end{bmatrix} = \mathbf{0}. \tag{21}$$

Equation (21) defines a hyperplane which divides the domain into two regions. Each mode of the PWA map is valid on one side of the hyperplane. Exploiting constraints of the type of (21) can be helpful to the identification process.

Different categories of approaches to PWA system identification can be distinguished depending on how the partitioning into regions is done. It follows from the discussion in Section III-B that there are mainly two alternative approaches: either the partition is defined a priori, or it is estimated along with the different submodels.

The first approach requires to define a priori the gridding of the domain. For instance, rectangular regions with sides parallel to the coordinate axes are used in [9], while simplices (i.e. polytopes with $d + 1$ corners, where d is the dimension of the domain) are considered in [23] and [40]. This approach drastically simplifies the estimation of the linear/affine submodels, since standard linear identification techniques can be used to estimate the submodels, given enough data points in each region. On the other hand, it has the drawback that the number of regions and the need for experimental data, grow exponentially with d. This approach is therefore impracticable for high-dimensional systems.

The second approach consists in estimating the submodels and the partition of the domain either simultaneously or iteratively. This should allow for the use of fewer regions, since the regions are shaped according to the available data. Depending on how the partition is determined, Roll [61]

further distinguishes among four different categories of approaches.

1) The first category relies on the direct formulation of a suitable criterion function to be minimized, such as (19). The parameters of the affine submodels and the coefficients of the hyperplanes defining the partition of the domain are therefore estimated simultaneously by minimizing the criterion function through numerical methods (e.g., Gauss-Newton search). The algorithms proposed in [3], [15], [29], [41], [59] fall into this category. This way of tackling the identification problem is straightforward, but has the drawback that the optimization algorithm may get trapped in a local minimum. Techniques for reducing the risk of getting stuck in a local minimum can be used, at the cost of increased computational complexity.
2) The second category of approaches is an extension of the first one, and gives more flexibility with respect to the number of submodels. All parameters are identified simultaneously for a model with a very simple partition. If the resulting model is not satisfactory, new submodels/regions are added, in order to improve the value of a criterion function. In other words, instead to be solved at once, the overall identification problem is divided into several steps, each consisting in an easier problem to solve. The algorithms proposed in [14], [22], [35], [37], [39] fall into this category. The algorithm [14] has been analyzed in [59]. The paper [41] also describes an iterative method for introducing new partitions on the domain, when the error obtained is not satisfactory. As for the first category of approaches, there is still a risk to get stuck in a local minimum. When adding new submodels, one should also take into consideration the risk of overfit.
3) The third category contains a variety of approaches, sharing the characteristic that the parameters of the submodels and the partition of the domain are identified iteratively or in different steps, each step considering either the submodels or the regions. The algorithms proposed in [5], [27], [47], [56], [60] start by classifying the data points and estimating the linear/affine submodels simultaneously. Then, region estimation is carried out by resorting to standard linear separation techniques. In [54], the position of rectangular regions is optimized one by one iteratively. Then, each rectangular region is divided into simplices, in which affine submodels are finally identified. In [52], a greedy randomized adaptive search procedure is used to iteratively and heuristically find good partitions of the domain. Other approaches can be found in [30] and [31].
4) The last category of approaches estimates the partition using only information concerning the distribution of the regression vectors, and not the corresponding output values. This means that the domain is partitioned in such a way that each region contains a suitable number of experimental data to estimate an affine submodel. The algorithms proposed in [16], [68] fall into this category. The major drawback of this category of approaches is that, without considering the output values, a set of data which really should be associated to the same submodel might be split arbitrarily.

It is stressed that most of the aforementioned approaches (e.g, [3], [14], [16], [22], [29], [35], [37], [41], [59]) assume that the system dynamics is continuous, while, e.g., [5], [27], [47], [56], [60] allow for discontinuities. Moreover, only few approaches (e.g., those in the second category, [5], [56], and [26], which is an extension of [27]) estimate also the number of submodels from data.

3) Other hybrid model classes: Recently, some contributions have focused on the class of *PieceWise Output Error* (PWOE) models, which are defined by the equations

$$\begin{aligned} y_k &= w_k + e_k \\ w_k &= f(\boldsymbol{r}_k), \end{aligned} \tag{22}$$

where $f(\cdot)$ is the PWA map (9), and the regression vector $\boldsymbol{r}_k$ is built as

$$\boldsymbol{r}_k = [\ w_{k-1}\ \dots\ w_{k-n_a}\ u_k\ u_{k-1}\ \dots\ u_{k-n_b}\ ]^\top. \tag{23}$$

In [63] a prediction-error minimization method for piecewise linear output-error predictors is derived under the assumption that the discrete state is known at each time step. Estimation of the discrete state is made possible in [46], where a Bayesian method for identification of PWOE models is proposed.

4) Recursive identification approaches: All the aforementioned algorithms operate in a batch mode, i.e. the model is identified after all the input-output data have been collected. Since the computational complexity of batch algorithms depends on the number of data points, such algorithms may not be suitable for real time applications. An online algorithm for the identification of SARX/PWARX models is proposed in [65]. It exploits a mixture of recursive identification and pattern recognition techniques in order to identify the current parameter values. A different approach is pursued in the recent contributions [32], [75]. A standard recursive identification algorithm is used to estimate the parameters of a "lifted" ARX model which is independent of the switching sequence, and is built by applying a polynomial embedding to the input-output data. Then, estimates of the ARX submodel parameters are obtained by differentiation. This approach also enables for the estimation of the model orders and the number of submodels.

IV. Data classification and region estimation

As pointed out in Section III-D, identification methods allowing for discontinuities in the PWA map (9) are best suited in the context of hybrid systems, since they allow logic conditions to be represented by abrupt changes in the system dynamics. Most recent contributions, such as [5], [27], [47], [56], [60], have thus focused on regression with discontinuous PWA maps. It is interesting to note that all the above mentioned approaches share the idea to tackle

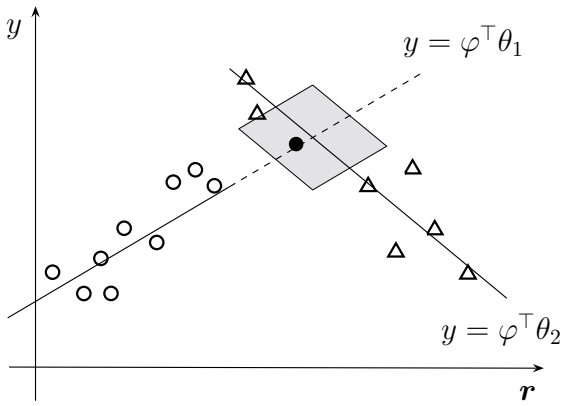

Fig. 4. Example showing the problem of intersecting submodels. The data point denoted by the black circle could be in principle attributed to both submodels. Wrong attribution yields two non-linearly separable clusters of points.

the identification problem by firstly classifying the data and estimating the affine submodels, and then estimating the partition of the regressors domain. In this section, the data classification step is discussed in view of the subsequent step of region estimation. Moreover, a brief overview of linear separation techniques is given, and issues related to the estimation of the partition from a finite number of points are highlighted.

A. Data classification

Methods for the identification of PWARX models that firstly classify the data points and estimate the affine submodels, and then estimate the partition of the regressors domain, split in practice the identification problem into the identification of a SARX model, followed by the shaping of the regions to the clusters of data. In this respect, such methods can be also considered as methods for the identification of SARX models, if the final region estimation step is not addressed. Vice versa, methods developed for the identification of SARX models, such as [51], [74], [78], can be used to initialize the procedures for the identification of PWARX models.

However, in view of the subsequent step of region estimation, data classification for the identification of PWARX models needs to be carefully addressed. The main problem to deal with is represented by data points that are consistent with more than one submodel, namely data points lying in the proximity of the intersection of two or more submodels. Wrong attribution of these data points may lead to misclassifications when estimating the polyhedral regions.

In order to clarify this point, Fig. 4 shows a data set obtained from a one-dimensional PWA model with $s = 2$ discrete modes. It is assumed that the parameter vectors θ_1 and θ_2 have been previously estimated, no matter which method has been used. If each data point $(y_k, \boldsymbol{r}_k)$ is associated to the submodel i^* such that the prediction error is minimized, i.e. according to the rule

$$i^* = \arg \min_{i=1,\dots,s} |y_k - \varphi_k^\top \theta_i|, \tag{24}$$

the point denoted by the black circle is attributed to the first submodel. This yields two non-linearly separable clusters of points. It is stressed that the issue addressed in this example does not depend on the particular choice of (24) for associating each data point to one submodel. If data classification and parameter estimation are performed by solving Problem 3.2 for a given $\delta > 0$, the point denoted by the black circle is still attributed to the first submodel in this case. The gray area in Fig. 4 represents the region of all data points satisfying

$$|y_k - \varphi_k^\top \theta_i| \leq \delta \tag{25}$$

for both $i = 1$ and $i = 2$. These data points are termed *undecidable*, because they could be in principle attributed to both submodels.

The identification procedures [5], [27], [47], [56], [60] deal with the problem of intersecting submodels in different ways. For instance, an ad-hoc refinement procedure based on the certainly attributed closest neighbors is proposed in [5], weights for misclassification are introduced in [47], and clustering in a feature space is pursued in [27]. These three approaches will be described in Section V.

B. Region estimation

After the data classification step, providing the estimates of the discrete state $\sigma(k) \in \{1, \dots, s\}$, it is possible to form s clusters of regression vectors as

$$\mathcal{A}_i = \left\{ \boldsymbol{r}_k \,:\, \sigma(k) = i \right\}, \quad i = 1, \dots, s. \tag{26}$$

The problem of region estimation consists in finding a complete polyhedral partition $\left\{\mathcal{R}_i\right\}_{i=1}^{s}$ of the regressors domain $\mathcal{R}$ such that $\mathcal{A}_i \subseteq \mathcal{R}_i$ for all $i = 1, \dots, s$. The polyhedral regions (8) are defined by hyperplanes. Hence, the considered problem is equivalent to that of separating s sets of points by means of linear classifiers (hyperplanes). This problem can be tackled in two different ways:

- a) Construct a linear classifier for each pair $(\mathcal{A}_i, \mathcal{A}_j)$, with $i \neq j$.
- b) Construct a piecewise linear classifier which is able to discriminate among s classes.

In the first approach, a separating hyperplane is constructed for each pair $(\mathcal{A}_i, \mathcal{A}_j)$, $i \neq j$. This amounts to solve $s(s-1)/2$ *two-class* linear separation problems. Given two sets $\mathcal{A}_i$ and $\mathcal{A}_j$, $i \neq j$, the linear separation problem is to find $w \in \mathbb{R}^d$ and $\gamma \in \mathbb{R}$ such that

$$\begin{aligned} w^\top \boldsymbol{r}_k + \gamma > 0 \quad \forall\, \boldsymbol{r}_k \in \mathcal{A}_i \\ w^\top \boldsymbol{r}_k + \gamma < 0 \quad \forall\, \boldsymbol{r}_k \in \mathcal{A}_j. \end{aligned} \tag{27}$$

This problem can be easily rewritten as a feasibility problem with linear inequality constraints by introducing the quantities

$$z_k = \begin{cases} 1 & \text{if } \boldsymbol{r}_k \in \mathcal{A}_i \\ -1 & \text{if } \boldsymbol{r}_k \in \mathcal{A}_j. \end{cases} \tag{28}$$