Piecewise Affine System identification: A least harmonic mean approach

Laurent Bako, Olfa Yahya

Piecewise Affine System identification: A least harmonic mean approach

Laurent Bako[1], Olfa Yahya[2]

Abstract— **In this paper we consider the problem of identifying piecewise affine approximations of nonlinear dynamic systems directly from input-output measurements. This requires estimating a partition of the regression space and an associated set of affine time-invariant models. By relying on a generic Voronoi type of partition, the proposed approach treats these two tasks jointly via the minimization of an appropriate objective function formed as a weighted sum of errors related to all submodels. We minimize this objective function with respect to both the weights and the parameters (describing the partition and the affine submodels) under some heuristic constraint on the weights. It is later observed that the proposed method attempts indeed to optimize a sum (over the available dataset) of harmonic means. The final estimation algorithm consists in solving iteratively a sequence of convex optimization problems. Numerical experiments tend to suggest that the new method exhibits some nice properties: regular convergence and good estimation performance.**

Keywords. system identification, piecewise affine systems, nonsmooth optimization.

I. Introduction

A piecewise affine (PWA) model of a dynamic nonlinear system consists of a partition of the state-input space of the system into local disjoint regions, each of which is associated with an affine (or linear) time-invariant model. Such models are very convenient for representing nonlinear functions (dynamic systems) for example from the perspective of nonlinear control design. In effect, PWA models arise naturally from a basic intuition of control engineering practitioners, the notion of operating point. They are therefore more easily amenable to interpretation and analysis by judiciously using the available knowledge on linear systems. PWA models can be viewed as a means to extend somehow generically linear design and analysis methods to nonlinear dynamic systems [17], [16]. This is probably one of the reason why there has been recently an increase of interest in piecewise affine models identification from input-output data. Beyond being a modeling abstraction of nonlinear systems as argued above, we can conceive the existence of strictly PWA systems, which forms a class of hybrid systems with a state-dependent switching signal (e.g. linear dynamic systems interconnected with static nonlinearities such as deadzones or saturated actuators, ...).

There is indeed a rich literature on PWA system identification. The large diversity of existing methods illustrates a clear surge of interest in these models in the recent years. Examples of methods include multiple local estimation and clustering through K-means [7] or recursive bayesian/incremental learning [9], [2], [6], Minimum partition into feasible sets (MIN PFS) [4]; optimization via mixed integer programming approach [15], sparsity-inducing optimization techniques [11]. The reader is referred to the survey papers [13], [8] for an overview of theory and application of hybrid systems identification.

In this paper we present an estimation framework which applies to both switched and PWA systems. In the PWA case, a distinguishing feature of the new method with respect to the conventional scheme, is that the partition and the corresponding affine submodels are estimated simultaneously. To achieve this, the identification problem is primarily formulated as an optimization problem through the introduction of two main ingredients: (i) a nonsmooth loss function on a generalized error accounting (locally) for the structural properties of the PWA model ; for each submodel, there is one such error associated to it ; (ii) an objective function formed as a sum, over the available samples, of the minimum error on the set of submodels. This is a hard nonconvex problem. Our approach towards a solution is then to consider a weighted objective function involving adjustable weighting functions of the data samples. Finally, an iterative procedure is proposed which alternates between adjusting the weights and updating the parameters. The iteration principle of the final computational algorithm is as follows: Starting from some initial distinct weights values (obtainable for example from a random sampling), the parameters of the partition and those of the affine submodels are jointly computed by solving a convex nonsmooth optimization problem. Given the new estimates of the parameters, the weights are subsequently updated by minimizing again the same objective function but this time, with respect to the weights, under some heuristic constraint. The algorithm keeps iterating this way until some convergence condition is satisfied.
A closer look at the method reveals that it attempts indeed to minimize the sum, over the available samples, of the harmonic means of the generalized errors related to each submodel. This observation turns out to be quite consistent with the fact that the harmonic mean is the closest (among all standard means) to the min-error function.

Outline. The rest of the paper is organized as follows. In Section II, we discuss how a PWA model can be derived naturally from a linearization at multiple operating points. This discussion yields a PWA model which is defined on Voronoi-type of polyhedral partition of the regression space which can be obtained either from the equations of the

[1]L. Bako is with Université de Lyon, Ampère (Ecole Centrale Lyon, INSA Lyon, Université Claude Bernard, CNRS UMR 5005), France, F-69611. E-mail: laurent.bako@ec-lyon.fr

[2]O. Yahya is with Université de Gabès, Tunisia. E-mail: olfayahya@gmail.com

nonlinear system, when available, or directly from input-output data generated by the nonlinear system. In Section III, we formulate the estimation problem for the PWA model. An optimization framework is therefore presented as a possible solution to that problem. In section IV we discuss the algorithmic steps for solving numerically the problem and obtain a PWA model. Section V shows some numerical illustrations concerning the performance of the identification method. Lastly Section VI presents some concluding remarks.

Notations. $\mathbb{R}$ is the set of real numbers while $\mathbb{R}_+$ stands for the set of nonnegative real numbers. $\|\cdot\|_p$ refers to standard vector p-norm. If $v \in \mathbb{R}^n$ is a vector then $\mathrm{diag}(v)$ refers to the diagonal matrix in $\mathbb{R}^{n\times n}$ having the entries of v on its diagonal.

II. PWA APPROXIMATION OF NONLINEAR FUNCTIONS

A. The nonlinear data-generating system

Consider a dynamic system described by a Nonlinear AutoRegressive eXogenous (NARX) model in the form

$$y_t = f(x_t) + e_t \tag{1}$$

where $y_t \in \mathbb{R}$ is the output of the system, e_t is accounting for potential noise and $x_t \in \mathbb{R}^n$ is the regressor vector defined by

$$x_t = \begin{bmatrix} y_{t-1} & \cdots & y_{t-n_a} & u_t^\top & \cdots & u_{t-n_b}^\top \end{bmatrix}^\top \tag{2}$$

with $u_t \in \mathbb{R}^{n_u}$ denoting the input (vector of n_u entries) of the system at time t. n_a and n_b are *known* integers called orders of the model, and $n = n_a + (n_b+1)n_u$. In case $n_a = 0$, the regressor becomes just $x_t = [u_t^\top \quad \cdots \quad u_{t-n_b}^\top]^\top$; the model (1) is then called a Nonlinear Finite Impulse Response (NFIR) model. We will assume that the regressors x_t generated by the system live in a bounded subset $\mathscr{X}$ of $\mathbb{R}^n$, called the regression space; $f : \mathscr{X} \to \mathbb{R}$ is an *unknown* nonlinear function which we will assume, for convenience, to be smooth (at least continuously differentiable).

B. The PWA approximation

As discussed in the introduction, we are interested here in finding a PWA model of the nonlinear system (1) directly from a set $\{(x_t, y_t)\}_{t=1}^N$ of experimental data generated by the nonlinear system (1). For this purpose, we will consider the particular class of PWA functions defined on a Voronoi-type partition of $\mathscr{X}$. Such a class of PWA models is described and motivated in [2]. Indeed if f is continuously differentiable as assumed above, then by selecting appropriately a finite number s of points c_i in $\mathscr{X}$, we can reasonably approximate f by a PWA map $f_{\text{PWA}} : \mathscr{X} \to \mathbb{R}$ defined by

$$f_{\text{PWA}}(x) = \begin{cases} a_1^\top x + b_1 & \text{if} \quad x \in \mathscr{X}_1 \\ \quad\vdots & \quad\vdots \\ a_s^\top x + b_s & \text{if} \quad x \in \mathscr{X}_s \end{cases} \tag{3}$$

where $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$, $i = 1, \ldots, s$, are defined by

$$a_i = \nabla f(c_i), \quad b_i = f(c_i) - c_i^\top \nabla f(c_i) \tag{4}$$

with the notation ∇f referring to the gradient of f with respect to x, and the sets $\mathscr{X}_i$ are given by

$$\mathscr{X}_i = \left\{ x \in \mathscr{X} : \|x - c_i\|_2 \leq \|x - c_j\|_2, \forall j = 1, ..., s \right\}. \tag{5}$$

For some additional arguments on this particular approximation strategy, we refer to [2]. The sets $\{\mathscr{X}_i\}_{i=1}^s$ defined in (5) form a partition of the regression space $\mathscr{X}$, that is, (i) they form a complete cover of the entire domain, $\mathscr{X} = \cup_{i=1}^s \mathscr{X}_i$ and (ii) their interiors are pairwise disjoint, $\mathrm{int}(\mathscr{X}_i) \cap \mathrm{int}(\mathscr{X}_j) = \emptyset$ for all $i \neq j$ with $\mathrm{int}(\cdot)$ referring to the interior. Such a partition which, by Eq. (5), is completely determined by the points $c_1, \ldots, c_s$, is known as Voronoi partition. The regions $\mathscr{X}_i$ are then called the Voronoi cells while the points c_i's are termed the seeds or the generators of the partition. Note that the definition (5) can be generalized to some other distances than the Euclidean one (see e.g., [12] for further details on this matter).

It is not hard to show that each set $\mathscr{X}_i$ is a convex polyhedron, that is, a set resulting from the intersection of a finite number of half-spaces. More precisely, by defining

$$\begin{aligned} H_i &= \begin{bmatrix} c_1 - c_i & \cdots & c_{i-1} - c_i & c_{i+1} - c_i & \cdots & c_s - c_i \end{bmatrix}^\top \\ h_i &= \begin{bmatrix} \beta_{1,i} & \cdots & \beta_{i-1,i} & \beta_{i+1,i} & \cdots & \beta_{s,i} \end{bmatrix}^\top, \end{aligned} \tag{6}$$

we can write $\mathscr{X}_i = \{x \in \mathscr{X} : H_i x \leq h_i\}$ where $\beta_{j,i} = (c_j^\top c_j - c_i^\top c_i)/2$. For illustration purpose, an example of Voronoi partition in $\mathbb{R}^2$ is depicted in Figure 1.

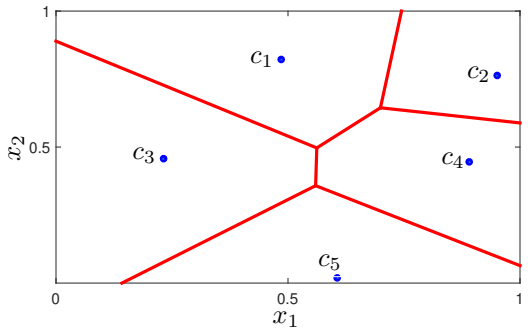

Fig. 1: Example of Voronoi-type partition of $[0\ 1] \times [0\ 1] \subset \mathbb{R}^2$ in 5 polyhedral regions defined by the points c_i, $i = 1, \ldots, 5$, as in Eq. (5).

Finally, we note that the PWA representation on Voronoi-type partition is less general than the usual polyhedral partition which is used for this purpose [6], [7]. But it does enjoy some convenient properties: (a) the seeds c_i can be given a quite intuitive interpretation as *operating points* of the system; (b) the affine functions associated with the local regions can be viewed as resulting from linearization of the nonlinear systems about the operating points; (c) the estimation task consists simply in identifying the parameters $\{(a_i, b_i, c_i)\}_{i=1}^s$ from observations. An interesting point is that, adopting the representation (5) for the regions simplifies slightly the identification of the switching mechanism which

can then be done simultaneously as the parameters of the affine submodels. Furthermore, a posteriori data classification is no longer necessary.

For convenience, we may write in the sequel

$$f_{\text{PWA}}(x) = \tilde{x}^\top \theta_i \quad \text{if} \quad x \in \mathscr{X}_i \tag{7}$$

with $\tilde{x} = [x^\top \quad 1]^\top$ and $\theta_i = [a_i^\top \quad b_i]^\top$.

C. A reference PWA model

Let $\mathscr{F}^s_{\text{PWA}}$ denote the set of all PWA functions f_{PWA} of the form (3)-(5) defined on a partition of $\mathscr{X}$ into s cells. Clearly, $\mathscr{F}^s_{\text{PWA}}$ is a subset of the set of all piecewise continuous functions from the compact set $\mathscr{X}$ to $\mathbb{R}$, which we denote by $\mathscr{F}(\mathscr{X}, \mathbb{R})$. Define a particular PWA function f^o_{PWA} with Voronoi-type partition by

$$f^o_{\text{PWA}} \in \underset{f_{\text{PWA}} \in \mathscr{F}^s_{\text{PWA}}}{\arg\min} \ \|f - f_{\text{PWA}}\|_{\mathscr{F}} \tag{8}$$

where $\|\cdot\|_{\mathscr{F}}$ is some norm on $\mathscr{F}(\mathscr{X}, \mathbb{R})$, for example the L_2 norm. That is, f^o_{PWA} is the function in $\mathscr{F}^s_{\text{PWA}}$ which is closest to f in the sense of the distance induced by the norm $\|\cdot\|_{\mathscr{F}}$. With reference to (7), finding f^o_{PWA} is equivalent to finding the model parameters $\{(\theta^o_i, c^o_i)\}_{i=1}^s$. Intuitively, the approximation error $\|f - f^o_{\text{PWA}}\|_{\mathscr{F}}$ is a nonincreasing function of the number s of affine subsystems.

To begin with, let us observe that the system equation (1) can be conveniently rewritten as

$$y_t = f^o_{\text{PWA}}(x_t) + v_t \tag{9}$$

with $v_t = e_t + \tilde{e}_t$ and $\tilde{e}_t = f(x_t) - f^o_{\text{PWA}}(x_t)$ denoting the approximation (mismatch) error. For easier reference in the following parts of the paper we can put this last equation in the following equivalent form

$$y_t = \tilde{x}_t^\top \theta^o_{\sigma(t)} + v_t, \tag{10}$$

where $\tilde{x}_t = [x_t^\top \quad 1]^\top$ and the subindex $\sigma(t)$ is called the discrete state defined by $\sigma(t) = i \in \{1, \ldots, s\}$ if $x_t \in \mathscr{X}^o_i$ with $\mathscr{X}^o_i$ being the Voronoi cell defined similarly as in (5) from the parameters $c^o_1, \ldots, c^o_s$. Of course, model (10) can account for a class of nonlinear systems as discussed above and also for strictly piecewise affine systems. Because v_t contains both measurement noise and model mismatch we do not associate any stochastic properties, such as independence on the input, to it. It is only required to be bounded.

Problem. From now on we view (10) as the PWA data-generating system. The goal of the estimation task is therefore to infer estimates of the parameters $\{\theta^o_i\}_{i=1}^s$ and $\{c^o_i\}_{i=1}^s$ from a finite collection of data $\{(x_t, y_t)\}_{t=1}^N$.

III. Estimation of PWA models from data

A reasonable approach to the estimation problem stated above would be to solve the optimization problem

$$\min_{\substack{\theta_1, \ldots, \theta_s \\ c_1, \ldots, c_s}} \sum_{i=1}^{s} \sum_{x_t \in \mathscr{X}_i(c_1, \ldots, c_s)} \left(y_t - \theta_i^\top \tilde{x}_t\right)^2 \tag{11}$$

for $\{\theta_i\}_{i=1}^s$ and $\{c_i\}_{i=1}^s$, where $\mathscr{X}_i(c_1, \ldots, c_s)$ is defined as in (5). The purpose of the notation $\mathscr{X}_i(c_1, \ldots, c_s)$ is just to make explicit the dependence of the sets $\mathscr{X}_i$ on the seeds c_i's for more clarity concerning the decision variables.

Estimating simultaneously the partition and the affine subfunctions as in (11) is a problem which is known to be very challenging. It is even hopeless to try to find the exact solution of problem (11) at an affordable computational cost. Therefore, as in virtually all existing works [6], [7], [9], we shall be content with a suboptimal solution.

We start by noting that an immediate and simple algorithm for solving (11) would typically consist in alternating between assigning the data (x_t, y_t) to the regions $\mathscr{X}_i(c_1, \ldots, c_s)$ based on the distances to the c_i's and computing the associated hyperplane parameters $\tilde{\theta}_i$ (analogously to the principle of the K-means algorithm). This can be done by sampling some initial estimates of $(c_1, \ldots, c_s)$ at random from $\mathbb{R}^n$. However the performance of such a basic algorithm is known to be poor due to the nonconvexity of the underlying optimization problem.

Hence we need to design a more efficient estimation procedure. The next discussion is articulated along the following points:

- Formulate the joint estimation problem (11) for the partition and the corresponding affine submodels as a regression-like problem. This is done by means of an appropriate nonsmooth loss function applying to an extended error which accounts for the requirement on the structure of PWA model (see Eq. (12)). We form the ideal cost function as a sum over time of some min-functions as in (14).
- Encapsulate the latter cost in a weighted sum (15) of the sample-losses which is to be minimized with respect to both the weights and the parameters under some multiplicative constraints on the weights (see Eq. (16)).
- Derive an algorithm that tries to find the optimum (see Section IV).

A. The proposed estimation method

In this section, we introduce gradually the idea of our estimation method. As described above, the first step is to reformulate the estimation problem in a way that removes the implicit partition problem in (11). For this purpose, pose $p_i = [\theta_i^\top \quad c_i^\top]^\top$ and $P = [p_1 \quad \cdots \quad p_s]$. Introduce also the notations $\mathbb{S} = \{1, \ldots, s\}$ and $\mathbb{T} = \{1, \ldots, N\}$ for the index sets of the submodels and the data samples respectively. For any sample (x_t, y_t), $t \in \mathbb{T}$, define a loss function

$$\xi_t(p_i, \varepsilon) = \left\| \Lambda \begin{bmatrix} y_t - \theta_i^\top \tilde{x}_t \\ x_t - c_i \end{bmatrix} \right\|_2 + \varepsilon \tag{12}$$

with $\varepsilon \geq 0$ being a small number which will, sometimes, be required to be strictly positive. The goal of ε is essentially to avoid division by zero whenever calculations may involve the inverse of $\xi_t(p_i, \varepsilon)$. $\Lambda \in \mathbb{R}^{(n+1)\times(n+1)}$ is some positive-definite weighting matrix balancing the contribution of the two terms involved in $\xi_t(p_i, \varepsilon)$. A typical choice for Λ can be of the form

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 I_n \end{bmatrix} \tag{13}$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$ and I_n being the identity matrix of dimension n. In that case, one would select λ_1 and λ_2 to balance the contribution of the two terms composing the generalized error $[(y_t - \theta_i^\top \tilde{x}_t) \quad (x_t - c_i)^\top]^\top$. The loss $\xi_t(p_i, \varepsilon)$ is a convex function of p_i which attempts to merge together both the requirement related to submodel fitting and the Voronoi type of partition. Different other choices of loss functions are possible with respect to this purpose. An important feature however is that a sum-of-norms cost is likely to produce better results thanks to the robustness-inducing properties of such costs (see e.g. [3] for some discusions on this point). Then, ideally, we would like to search for the parameter matrix P which minimizes

$$\mathcal{J}(P) = \sum_{t=1}^{N} \min_{i \in \mathbb{S}} \xi_t(p_i, 0). \tag{14}$$

The rationale behind the choice of $\xi_t(p_i, \varepsilon)$ as in the form (12) and the objective function (14) is to drive, for each i, the fitting errors $y_t - \theta_i^\top \tilde{x}_t$ to zero whenever the distance from x_t to c_i is small. But at a given time t, we do not want to minimize all the errors $\xi_t(p_i, 0)$ but only one of them (namely the smallest), the difficulty being that we do not know which one in advance. Note that (14) is still hard to solve as such. Therefore our strategy here is to embed the objective in (14) into the following one

$$F(P, W, \varepsilon) = \sum_{t=1}^{N} \sum_{i=1}^{s} w_{it} \xi_t(p_i, \varepsilon) \tag{15}$$

where $W \triangleq [w_{it}]_{i,t} \in \mathbb{R}^{s \times N}$ is a weighting matrix with positive entries w_{it}, $(i, t) \in \mathbb{S} \times \mathbb{T}$.

Clearly, if we could select the weights such that $w_{it} = 1$ when $i \in \arg\min_{j \in \mathbb{S}} \xi_t(p_j, 0)$ and $w_{it} = 0$ otherwise, then we would recover $F(P, W, \varepsilon) = \mathcal{J}(P)$ for $\varepsilon = 0$. Hence $\mathcal{J}(P)$ appears to be a particular instance of $F(P, W, \varepsilon)$ provided that the weights are appropriately selected. But of course the difficulty is that we do not know $\arg\min_{j \in \mathbb{S}} \xi_t(p_j, 0)$ prior to computing the vectors p_j. Therefore, we will be searching simultaneously for both the weighting matrix W and the parameter matrix P. More explicitly, the following optimization problem is considered:

$$\min_{P,W} F(P, W, \varepsilon) \tag{16}$$

$$\text{subject to} \quad \prod_{i=1}^{s} w_{it} = 1 \; \forall t \in \mathbb{T} \tag{17}$$

$$w_{it} \geq 0 \; \forall (i, t) \in \mathbb{S} \times \mathbb{T} \tag{18}$$

The rationale behind the constraints (17)-(18) on the weights is to promote the property that for all t, whenever a weight $w_{i_o t}$ is relatively large for some i_o, the other weights w_{it}, $i \neq i_o$, should be made somewhat small. The underlying goal which is sought through this property is to favor the desirable scenario that only one subsystem is activated at each time t as in (14).

The next proposition states that problem (16)-(18) is equivalent to minimizing the sum over t of the geometric means of the $\xi_t(p_i, \varepsilon)$ taken along i.

Proposition 1: The optimal value of problem (16) is equal to

$$\min_{P} F^*(P, \varepsilon) \tag{19}$$

where

$$F^*(P, \varepsilon) = s \sum_{t=1}^{N} \Big(\prod_{j=1}^{s} \xi_t(p_j, \varepsilon) \Big)^{1/s}. \tag{20}$$

Proof: According to a general result in optimization theory [14, p. 28], minimizing jointly with respect to (P, W) is equivalent to minimizing first with respect to W and then with respect to P. Hence we can write

$$\inf_{P,W} F(P, W, \varepsilon) = \inf_{P} H(P, \varepsilon)$$

where

$$H(P, \varepsilon) = \inf_{W} \Big\{ F(P, W, \varepsilon) : \text{ subj. to (17)} - \text{(18)} \Big\}. \tag{21}$$

We just need now to show that $H = F^*$. For this purpose, consider a weighting matrix W^* defined by

$$w_{it}^* = \frac{\big(\prod_{j=1}^{s} \xi_t(p_j, \varepsilon) \big)^{1/s}}{\xi_t(p_i, \varepsilon)} \tag{22}$$

Then W^* satisfies the constraints (17)-(18). By definition of $H(P, \varepsilon)$ in (21), it therefore follows that $H(P, \varepsilon) \leq F(P, W^*, \varepsilon)$. Moreover, it can be checked that $F(P, W^*, \varepsilon) = F^*(P, \varepsilon)$ and so, $H(P, \varepsilon) \leq F^*(P, \varepsilon)$. On the other hand, by applying the arithmetic-geometric mean inequality [5] and invoking the constraints on the weighting matrix, we can write

$$\begin{aligned} \Big(\prod_{i=1}^{s} \xi_t(p_i, \varepsilon) \Big)^{1/s} &= \Big(\prod_{i=1}^{s} w_{it} \xi_t(p_i, \varepsilon) \Big)^{1/s} \\ &\leq \frac{1}{s} \sum_{i=1}^{s} w_{it} \xi_t(p_i, \varepsilon). \end{aligned}$$

It follows, by summing over $\mathbb{T}$, that for any feasible W with respect to the constraints (17)-(18), it holds that

$$F^*(P, \varepsilon) \leq F(P, W, \varepsilon),$$

which implies that $F^*(P, \varepsilon) \leq \inf_W F(P, W, \varepsilon) = H(P, \varepsilon)$. In conclusion $F^* = H$ and the result is proved. ∎

What we have shown is that minimizing $F(P, W, \varepsilon)$ for (P, W) under the constraints (17)-(18) is equivalent to minimizing $F^*(P, \varepsilon) = F(P, W^*, \varepsilon)$ for P (see Eq. (20)). For numerical reasons however, we will minimize instead a modified version $F(P, \widetilde{W}^*, \varepsilon)$, where $\widetilde{W}^*$ is a normalized version of W^* defined by $\widetilde{w}_{it}^* = w_{it}^* / \sum_{j=1}^{s} w_{jt}^*$. We will formally denote with ρ the normalization operation, $\widetilde{W}^* = \rho(W^*)$. Indeed, simple algebraic calculations show that

$$F(P, \widetilde{W}^*, \varepsilon) = \sum_{t=1}^{N} \Big(\frac{1}{s} \sum_{i=1}^{s} \frac{1}{\xi_t(p_i, \varepsilon)} \Big)^{-1} \tag{23}$$

that is, $F(P, \widetilde{W}^*, \varepsilon)$ is the sum over $\mathbb{T}$ of the *harmonic means* of the $\{\xi_t(p_i, \varepsilon),\ i = 1, \ldots, s\}$.

As it turns out, the proposed estimation method attempts to minimize the sum of harmonic means of the errors

calculated over all the set of submodels. This makes sense since one can view the harmonic mean of the set of points $\{\xi_t(p_i,\varepsilon),\ i=1,\ldots,s\}$ as an approximation of $\min_{i\in\mathbb{S}}\xi_t(p_i,\varepsilon)$. By the Harmonic-Arithmetic-Geometric inequality it is even the best approximation of the minimum among the standard means.

Remark 1: The problem of identifying switched linear systems can be viewed as a special case of the current framework. In this case, the parameters c_i do not exist so that the vector $\xi_t(p_i,0)$ in (12) reduces to $y_t-\theta_i^\top\tilde{x}_t$. More precisely, the last entries of the θ_i's and of the $\tilde{x}_t$ can be dropped hence leading to $\xi_t(p_i,\varepsilon)=\left|y_t-\theta_i^\top x_t\right|+\varepsilon$. Applying the current framework to switched linear systems with this cost function stands as an interesting alternative to the method developed in [1] where the submodels were extracted one after another by exploiting the sparsity-inducing properties of the ℓ_1 norm. The framework of this paper exploits the same robustness properties more efficiently by estimating all submodels simultaneously.

IV. The estimation algorithm

By summarizing the discussion of Section III-A, we can derive a concrete estimation algorithm for the PWA system parameters. For this purpose, we introduce the map $r:\mathbb{R}^{n\times s}\to\mathbb{R}_+$

$$r(P)=\max_{t\in\mathbb{T}}\min_{i\in\mathbb{S}}\xi_t(p_i,0)$$

with $\xi_t(p_i,0)$ defined as in (12) with $\varepsilon=0$.

Algorithm. Let $\varepsilon^0=1$ and let the weights w_{it}^0 be some distinct positive numbers and set $\widetilde{W}^0=\rho(W^0)$. For all iteration counter $k\geq 0$ generate the sequence of estimates as follows:
If $\varepsilon^k>0$, then

$$P^{k+1}\in\operatorname*{arg\,min}_{P\in\mathbb{R}^{n\times s}}F(P,\widetilde{W}^k,\varepsilon^k) \tag{24}$$

$$\varepsilon^{k+1}=\min\left(\varepsilon^k,r(P^{k+1})\right) \tag{25}$$

$$W^{k+1}\in\operatorname*{arg\,min}_{W\in\mathbb{R}^{s\times N}}\left\{F(P^{k+1},W,\varepsilon^{k+1}):\prod_{i=1}^{s}w_{it}=1\ \forall t\in\mathbb{T}\right\} \tag{26}$$

$$\widetilde{W}^{k+1}=\rho(W^{k+1})=W^{k+1}\left[\operatorname{diag}((W^{k+1})^\top\mathbf{1}_s)\right]^{-1} \tag{27}$$

else, if $\varepsilon^k=0$, then stop the algorithm, i.e., set

$$P^j=P^k\quad\forall\, j\geq k \tag{28}$$

In (27), the notation $\mathbf{1}_s$ refers to a s-dimensional vector filled with ones. As already discussed earlier in this paper, the entries of the matrix W^{k+1} in (26) have the expression given in (22) with P replaced by P^{k+1}. It follows, through the normalization ρ, that the entries of $\widetilde{W}^{k+1}$ take the expression

$$\widetilde{W}_{it}^{k+1}=\frac{1}{\xi_t(p_i^{k+1},\varepsilon)}\Big(\sum_{j=1}^{s}\frac{1}{\xi_t(p_j^{k+1},\varepsilon)}\Big)^{-1}$$

where the p_j^{k+1} refer to the columns of P^{k+1}. Note that (25)-(27) admit solutions which either come in closed-form or are easy to compute. Hence the only step of the algorithm that requires numerical solving is the subproblem (24). It can however be observed that this last optimization problem is convex and can therefore be solved efficiently. Overall the proposed algorithm has a polynomial complexity in the dimensions of the variables.

V. Numerical example

To illustrate numerically the developments of this paper, we consider a nonlinear system defined by

$$y_t=\sin(u_t)+\cos(u_{t-1})+e_t \tag{29}$$

where $u_t\in[-\pi,\pi]$. Hence $x_t=\begin{bmatrix}u_t & u_{t-1}\end{bmatrix}^\top$ is defined on the compact set $\mathscr{X}=[-\pi,\pi]\times[-\pi,\pi]$. $\{e_t\}$ represents a noise sequence. To evaluate the fit between the measured output and that of the model, we use the measure [10]

$$\text{FIT}=\left[1-\frac{\|\hat{y}-y\|_2}{\|y-\bar{y}1_N\|_2}\right]\times 100\%$$

where y is a vector containing the output measurements, $\hat{y}$ is a vector containing the model output, $\bar{y}$ is the average of the measurements and 1_N is an N-dimensional vector of ones.

We apply the estimation method described in the paper to the identification of a piecewise affine model for the nonlinear system (29).

A. Approximation capacity of the PWA model

The goal of our first experiment is to test the theoretical approximation capacity of the PWA model (3) (with Voronoi type of partition) for nonlinear systems. For this purpose, we consider a noise-free scenario and apply the algorithm of Section IV on $N=4000$ data points which result from a random sampling of the input signal. The result depicted in Figure 2 illustrates graphically the approximation power of PWA models with Voronoi-type partition. Moreover, as shown by Figure 3 (noisy case), the approximation error gets logically smaller when the number of affine submodels increases. It can be observed that the approximation error tends to decrease very rapidly when s is small but beyond a certain threshold, the improvement obtained by adding more submodels becomes smaller and smaller.

B. Identification from noisy data

Now we illustrate the sensitivity of the estimation method to measurement noise. The sequence $\{e_t\}$ in (29) is taken to be a zero-mean white Gaussian noise whose relative power (with respect to the noise-free output) is expressed in terms of Signal to Noise Ratio (SNR). Five levels of noise are considered (see Table I). For each level of noise, a Monte-Carlo simulation of size 100 is carried out. For each of the 100 realizations of input-output data, we estimate a PWA model with $s=10$ submodels. The resulting average FITs are collected in Table I in function of the relative power of the noise. What the results reveal is that the proposed identification method is robust to noise since the quality of the estimation decreases smoothly in proportion of the amount of noise. Note however that even in the case where

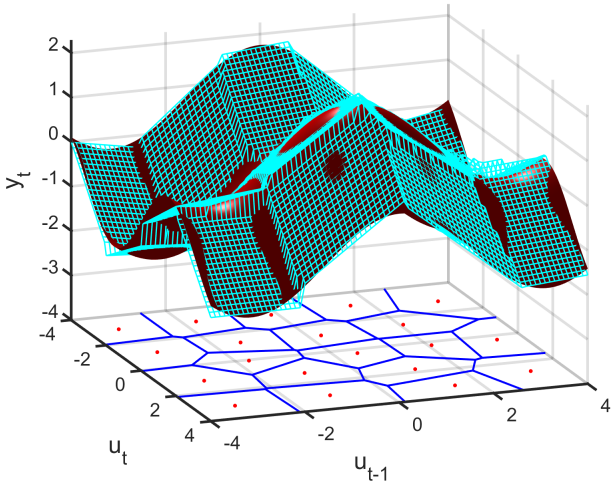

Fig. 2: Graphical comparison of the true nonlinear function defined in (29) and its PWA approximation with 20 affine submodels in a noise-free scenario. The operating points c_i's are materialized by red dots. The Voronoi-type partition of the regression space is drawn in blue line on x-y plane. The true nonlinear system is represented by the red surface while the output of the PWA model is plotted in blue. Design parameters (see Eqs (12)-(13)): $\lambda_1 = 1$, $\lambda_2 = 0.5$.

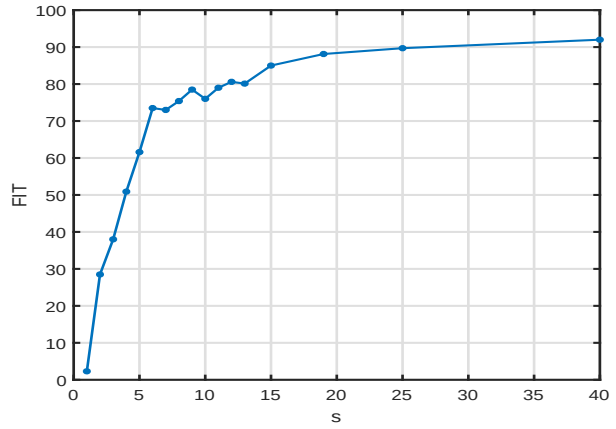

Fig. 3: Evolution of the FIT as a function of the number of affine submodels. Data characteristics: $N = 4000$, SNR= 30 dB. Design parameters of the algorithm: $\lambda_1 = 1$, $\lambda_2 = 0.5$.

there is almost no noise, the FIT does not go higher than 85%. This is due to the fundamental bias induced by the approximation of the *nonlinear system* (29) with a *PWA model* containing 10 submodels. To reduce that bias, one needs to increase the number of submodels (i.e., to enhance the modeling capacity of the model) as shown in Figure 3 where 40 submodels were used to obtain a FIT of about 92% in the presence of moderate noise.

VI. Conclusion

In this paper we have discussed an estimation framework for computing PWA models of nonlinear systems. Although the discussion mostly focuses on PWA models, it applies straightforwardly to switched systems as well. The proposed framework relies on the optimization of a sum-of-norms based objective function which is parameterized by adjustable weighting functions. Optimizing this cost under some constraints on the weights yields in fact the sum of the harmonic means of the errors related to each region of the partition and its associated affine submodel. The simulation results show a good capacity of the algorithm to estimate efficiently PWA models with a large number of submodels and good robustness properties to noise. Future work will include further evaluation of the performance of the proposed method through e.g., convergence analysis and a thorough comparison with other methods.

SNR	5 dB	15 dB	20 dB	30 dB	100 dB
FIT (%)	45.38	77.34	82.41	84.99	85.28

TABLE I: Average FITs obtained for different levels of the noise on a Monte-Carlo simulation of size 100. For each experiment, $N = 2000$ input-output data are generated: estimation is done with 60% of the data and the validation FIT is computed on the whole dataset. Design parameters of the algorithm: $\lambda_1 = 1$, $\lambda_2 = 0.5$.

References

[1] L. Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47:668–677, 2011.

[2] L. Bako, K. Boukharouba, E. Duviella, and S. Lecoeuche. A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis: Hybrid Systems*, 5:242–253, 2011.

[3] L. Bako and H. Ohlsson. Analysis of a nonsmooth optimization approach to robust estimation. *Automatica*, 66:132–145, 2016.

[4] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control*, 50:1567–1580, 2005.

[5] D. S. Bernstein. *Matrix mathematics: theory, facts and formulas*. Princeton University Press, 2009.

[6] V. Breschi, D. Piga, and A. Bemporad. Piecewise affine regression via recursive multiple least squares and multicategory discrimination. *Automatica*, 73:155–162, 2016.

[7] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39:205–217, 2003.

[8] A. Garulli, S. Paoletti, and A. Vicino. A survey on switched and piecewise affine system identification. In *IFAC Symposium on System Identification, Brussels, Belgium*, 2012.

[9] A. L. Juloski, S. Weiland, and W. Heemels. A bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control*, 50:1520–1533, 2005.

[10] L. Ljung. *System Identification Toolbox*, 2018.

[11] H. Ohlsson and L. Ljung. Identification of switched linear regression models using sum-of-norms regularization. *Automatica*, 49:1045–1050, 2013.

[12] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley; 2nd Edition, 2000.

[13] S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: A tutorial. *European Journal of Control*, 13:242–260, 2007.

[14] R. T. Rockafellar and R. J.-B.Wets. *Variational analysis*. Springer Verlag, 1997.

[15] J. Roll, A. Bemporad, and L. Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40:37–50, 2004.

[16] S. Waitman, L. Bako, P. Massioni, G. Scorletti, and V. Fromion. Incremental stability of lur'e systems through piecewise-affine approximations. In *IFAC World Congress, Toulouse, France.*, 2017.

[17] S. Waitman, P. Massioni, L. Bako, G. Scorletti, and V. Fromion. Incremental l2-gain analysis of piecewise-affine systems using piecewise quadratic storage functions. In *IEEE Conference on Decision and Control, Las Vegas, USA*, 2016.