Piecewise affine systems identification: a learning theoretical approach

Maria Prandini

***Abstract*— In this paper we study the problem of the identification of a hybrid model for a nonlinear system, based on input-output data measurements. We consider in particular the identification of piecewise affine models of nonlinear single-input/single-output systems through the prediction error minimization approach. The objective of this work is to analyze the performance of the identified model as the number of data used in the identification procedure grows to infinity. We consider a stochastic setting where the input and output signals are strictly stationary stochastic processes. Under suitable ergodicity assumptions, we show that the identified model is asymptotically optimal. The adopted approach is based on recent developments in statistical learning theory, and appears promising for studying the finite-sample properties of the identified model.**

I. INTRODUCTION

We consider a system identification problem consisting in building a mathematical model of a single-input/single output (SISO) discrete time nonlinear system based on input-output data measurements.

The standard approach to system identification consists in a two-step procedure: i) select a class of candidate models, and ii) choose the "best" model in the candidate models class according to a certain criterion.
In the Prediction Error Minimization (PEM) approach, the goal is determining a law to predict the future values of the system output from previous observations. A predictor is associated to each candidate model and the quality of a model is evaluated in terms of the performance of the associated predictor on the collected input-output data.

Specifically, let u and y be the input and output of the system under study. Denote by $\hat{y}_k$ the prediction of the output y of the system at time k, based on the observed input and output data up to time $k-1$ for some candidate predictor. Suppose that the values of the prediction errors

$$\epsilon_k := y_k - \hat{y}_k,$$

can be computed for already seen system outputs at every time $k = 1, \ldots, t$. Then, a standard criterion used to quantify the accuracy of the predictor under consideration is the quadratic cost:

$$\frac{1}{t}\sum_{k=1}^{t}\epsilon_k^2. \tag{1}$$

Minimizing a cost such as (1) over the set of candidate predictors (i.e. the set of candidate laws mapping the past observations in the output prediction) returns a predictor. If the predictor set was derived from a model set, one can go a step back and determine the model corresponding to that predictor.

To simplify the identification procedure, it is convenient to describe the candidate models class through a finite dimensional parameter vector, say θ, so that a model is selected by picking up a parameter value for θ in some set of admissible values Θ. This parameterization in some cases is artificially introduced, whereas in some other cases has a physical interpretation and naturally derives from the problem description.

If the candidate models set is parameterized by $\theta \in \Theta$, we can denote the prediction of y_k associated to the model with parameter θ as $\hat{y}_k(\theta)$, and define the function $J_t : \Theta \to \Re$

$$J_t(\theta) = \frac{1}{t}\sum_{k=1}^{t}(y_k - \hat{y}_k(\theta))^2, \tag{2}$$

which maps each $\theta \in \Theta$ to a positive real number representing the value taken by the cost function (1) when the model with parameter θ is used for predicting the output of the system.

The problem of identifying a model for the system is then reduced to that of solving the following parameter optimization problem:

$$\text{Determine } \hat{\theta}_t \text{ such that } J_t(\hat{\theta}_t) = \min_{\theta \in \Theta} J_t(\theta). \tag{3}$$

The desired property of the identified model with parameter $\hat{\theta}_t$ is that the accuracy of the associated predictor does not deteriorate when it is used for predicting output data different from those used in the identification process, though generated under the same operating conditions (*generalization property*).

Depending on the excitation properties of the input u, the identified model could be useful not only for prediction, but also for other purposes (control, estimation, simulation, etc.). A discussion on this latter aspect goes beyond the scope of this paper. The interested reader is referred to [1] for experiment design in a linear setting.

Here we consider the problem of evaluating the generalization property of the model identified based on the PEM approach with quadratic cost in a stationary environment.

PEM methods can be given a very meaningful interpretation when used in a stationary environment. If the input and the output of the system are strictly stationary processes, the

Research supported by the European Commission under the project HYBRIDGE IST-2001-32460, and by MIUR under the project "New techniques for the identification and adaptive control of industrial systems".

Maria Prandini is with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy prandini@elet.polimi.it

generalization properties of the model with parameter θ can be measured through

$$\bar{J}(\theta) = E[(y_k - \hat{y}_k(\theta))^2], \tag{4}$$

where the expectation is taken with respect to the joint probability distribution of the random variables y_k and $\hat{y}_k(\theta)$, which are, respectively, the output and its prediction at a generic time instant k.
$\bar{J}(\hat{\theta}_t)$ represents the average prediction performance when the model identified based on the input-output data collected up to time t is used to predict unobserved output data generated by the system in the same stationary conditions. For this reason $\bar{J}(\hat{\theta}_t)$ is also called *generalization error*.

The best achievable performance for a predictor belonging to the candidate set is given by

$$\bar{J}^\star = \inf_{\theta \in \Theta} \bar{J}(\theta). \tag{5}$$

The *empirical cost* $J_t(\theta)$ in (2) represents an estimate of the *theoretical cost* $\bar{J}(\theta)$ in (4), which in fact cannot be computed because the joint probability distribution of y_k and $\hat{y}_k(\theta)$ is not known. According to this interpretation, by minimizing $J_t(\theta)$ one actually aims at minimizing $\bar{J}(\theta)$.

In general, we cannot expect that the predictor with parameter $\hat{\theta}_t$ —obtained by minimizing the empirical cost based on a finite number of observed data— is optimal, i.e., $\bar{J}(\hat{\theta}_t) = \bar{J}^\star$ with $t < \infty$. Our best expectation is that the predictor with parameter $\hat{\theta}_t$ becomes optimal asymptotically, when the number of data used in the identification procedure tends to infinity, i.e., $\bar{J}(\hat{\theta}_t) \to \bar{J}^\star$, as $t \to \infty$.
Since the value taken by $\hat{\theta}_t$, $t \geq 1$, depends on the realization of the u and y processes used for the identification, there might be some "bad" realizations of the u and y processes such that asymptotic optimality is not achieved. For this reason, we relax our requirement on asymptotic optimality by admitting that bad realizations might occur but with zero probability. This can be formalized as follows:

$$\bar{J}(\hat{\theta}_t) \to \bar{J}^\star, \text{ as } t \to \infty, \text{ a.s. (almost surely).} \tag{6}$$

In this paper, we study the asymptotic performance of the PEM method when the model to be identified is a Nonlinear AutoRegressive eXogenous (NARX) model with a piecewise affine structure (PWARX model).

Piecewise affine systems are a class of hybrid systems which has been studied extensively in the literature. This is partly due to their modeling capabilities, since they are an equivalent representation for different hybrid systems (linear complementary systems [2], systems obtained by the combination of linear systems and finite automata [3], and hybrid systems in the mixed logic dynamical form [4]).

In principle, one can exploit the "simple" structure of a piecewise affine system —characterized by affine dynamics pieced together— to develop analysis and control design methodologies inspired by linear systems theory. Results that confirm this intuition can actually be found in [4]-[6], just to name a few. Moreover, some nonlinear functions can be approximated with arbitrary accuracy by piecewise affine functions [7], so that analysis and control design methodologies developed for piecewise affine systems can be applied also to these nonlinear systems [8].

Various contributions on the identification of PWARX models have been presented in the literature, see e.g. [9]-[12]. These papers are concerned with the issue of developing efficient algorithms for solving the optimization problem (3). The present paper addresses the issue of assessing the identified model quality, and as such should be seen as complementary to these contributions. The methodology adopted is inspired by ideas from statistical learning theory and stochastic processes analysis. Statistical learning theory provides useful tools for convergence analysis in system identification, allowing to extend the asymptotic results in [13], [1] to a more general setting. In [14], a converge result for PEM methods is proven in a parameter-free context by using the ϵ-net concept as complexity measure for the models class ([15]). Here, we concentrate on the identification of parameterized PWARX models for nonlinear systems and prove asymptotic optimality based on a different complexity measure known as Pollard-dimension. A general picture on the application of statistical learning theory to system identification is given in [16], [17].

The rest of the paper is organized as follows. In Section II we precisely formulate the identification problem that we are studying, whereas Section III is devoted to the analysis of the asymptotic performance of the identified model. Some conclusions are drawn in Section IV.

II. MATHEMATICAL FRAMEWORK

Suppose that we want to identify a model for describing a SISO system with input u and output y, based on the input-output data collected up to time t.

For this purpose we consider as candidate models class the set of PWARX models described by

$$y_k = f(\phi_{k-1}; \theta) + e_k, \tag{7}$$

where $\phi_{k-1} = [y_{k-1}, \ldots, y_{k-n}, u_{k-1}, \ldots, u_{k-m}]^T \in \Re^{n+m}$ is the regression vector containing past samples of the output y and input u, and function $f(\cdot;\theta) : \Re^{n+m} \to \Re$ has the following piecewise linear structure:

$$f(\phi; \theta) = \gamma_0 + \alpha_0 \phi + \sum_{i=1}^{p} \eta_i \left| \gamma_i + \alpha_i \phi \right|, \tag{8}$$

where $\theta = [\gamma_0\ \alpha_0\ \eta_1\ \gamma_1\ \alpha_1\ \ldots\ \eta_p\ \gamma_p\ \alpha_p]$ belongs to some compact set $\Theta \subset \Re^{(p+1)(n+m+2)-1}$.

As for $\{e_k\}$, it is a scalar white process with zero mean, independent of $\{u_k\}$, and not directly measurable.

Function (8) has been largely studied in the literature on nonlinear function approximation and circuit analysis, and in these two contexts is known under the name of hinging hyperplanes function ([7]) and Chua's canonical representation ([18]), respectively.

Remark 1: Note that there are some redundancies in the parameterization of the candidate models. For instance,

$$\eta_i \left|\gamma_i + \alpha_i \phi\right| = \frac{\eta_i}{a} \left|a\gamma_i + a\alpha_i \phi\right|, \quad \forall a > 0.$$

Also, the value of function f does not change if we change the ordering of the terms in the summation in (8). These redundancies cause the map associating to a parameter θ a model to be not bijective, hence, a *structural identifiability problem*. This problem can be alleviated by considering suitable constraints on the admissible θ values when defining the set Θ ([9]). □

The predictor associated to model (7) with parameter θ is given by:

$$\hat{y}_k(\theta) = f(\phi_{k-1}; \theta).$$

Correspondingly, the empirical cost (2) and the theoretical cost (4) takes respectively the form

$$J_t(\theta) = \frac{1}{t} \sum_{k=1}^{t} (y_k - f(\phi_{k-1}; \theta))^2 \tag{9}$$

$$\bar{J}(\theta) = E_P[(y_k - f(\phi_{k-1}; \theta))^2], \tag{10}$$

where P is the joint probability distribution of the random variables y_k and ϕ_{k-1}, and E_P is used to denote the expected value with respect to P.

In the next section, we study the asymptotic properties of the identified model. We shall prove that the optimality result (6) holds in our setting under some assumptions on the stochastic processes $\{u_k\}$ and $\{y_k\}$ that are described below.

Assumption 1: $\{u_k\}$ and $\{y_k\}$ are strictly stationary processes taking values in some compact set $U \subset \Re$ and $Y \subset \Re$, respectively. □

The assumption that the input and output processes take values in compact sets is quite technical, and is required for ensuring that the prediction error $y_k - f(\phi_{k-1}; \theta)$ is bounded. This allows the use of certain results of the statistical learning theory for analyzing the identified model.

We need also to make some ergodicity assumption on the $\{u_k\}$ and $\{y_k\}$ processes for the asymptotic optimality result (6) to hold. $\hat{\theta}_t$ appearing in (6) is the minimizer of the empirical cost (9). Intuitively, the correlation in time of the $\{u_k\}$ and $\{y_k\}$ processes has to decay at a sufficiently fast rate for the empirical cost (9) to converge to the theoretical cost (10) as t tends to infinity.

In a linear setting, this directly translates into the requirement that the system under study is asymptotically stable. In a nonlinear setting, it is difficult to translate the needed ergodicity assumption in stability-like requirements on the data generation mechanism. This characterization is in fact the subject of ongoing research activities [19].

Here, we limit ourselves to express it directly in terms of correlation properties of the input and output processes through the concept of beta-mixing process, which is briefly explained hereafter (see e.g. [20]).

Let us consider a strictly stationary process $\{s_k\}$ with probability distribution $\bar{P}$. The beta-mixing coefficients of $\{s_k\}$ are defined as:

$$\beta_t = \sup_{A \in \sigma_t} \{|\bar{P}(A) - (\bar{P}^0_{-\infty} \times \bar{P}^{\infty}_1)(A)|\}$$

where $\bar{P}^0_{-\infty}$ and $\bar{P}^{\infty}_1$ are the semi-infinite marginals of $\bar{P}$, and σ_t denotes the σ-algebra generated by the sets of random variables $\{s_k, k \leq 0\}$ and $\{s_k, k \geq t\}$. The sequence $\{\beta_t, t \geq 1\}$ is bounded below by zero and not increasing because $\sigma_{t+1} \subseteq \sigma_t$, $\forall t$. Therefore $\bar{\beta} = \lim_{t \to \infty} \beta_t$ exists and is $\bar{\beta} \geq 0$. If $\bar{\beta} = 0$, then $\{s_k\}$ is a beta-mixing process.

We are now in a position to formulate our assumption. We require that

Assumption 2: The process $\{(y_k, \phi_{k-1})\}$ is geometrically beta-mixing, i.e., its beta-coefficients satisfy: $\beta_t \leq \rho^t$, $\forall t$, for some $\rho < 1$. □

III. ASYMPTOTIC OPTIMALITY OF THE IDENTIFIED MODEL

In this section we prove that the asymptotic result (6) holds in our setting, i.e., when the classical PEM method with quadratic cost function is used to identify a PWARX model based on input-output data collected in a stationary environment, where Assumptions 1 and 2 are satisfied.

We start by making some preliminary observations.

Fix $\theta \in \Theta$. The process $\{v_k(\theta)\}$ defined by

$$v_k(\theta) = (y_k - f(\phi_{k-1}; \theta))^2$$

is strictly stationary with mean $\bar{J}(\theta)$. $J_t(\theta)$ is the empirical estimate of the mean of process $\{v_k(\theta)\}$ based on t samples. Under suitable ergodicity assumptions, by the strong law of large numbers one can prove that

$$J_t(\theta) = \frac{1}{t} \sum_{k=1}^{t} v_k(\theta) \to \bar{J}(\theta) = E[v_k(\theta)], \text{ a.s.}$$

for every $\theta \in \Theta$ (point-wise convergence). However, we need a stronger property than point-wise convergence to prove that the minimum of the empirical cost converges to the minimum of the theoretical cost:

$$\bar{J}(\hat{\theta}_t) \to \bar{J}^\star = \inf_{\theta \in \Theta} \bar{J}(\theta), \text{ as } t \to \infty, \text{ a.s.} \tag{11}$$

The following lemma shows that uniform convergence of $J_t(\theta)$ to $\bar{J}(\theta)$ is a sufficient condition for (11) to hold.

Lemma 1: Suppose that

$$\sup_{\theta \in \Theta} |J_t(\theta) - \bar{J}(\theta)| \to 0, \text{ as } t \to \infty, \text{ a.s.} \tag{12}$$

Then, $\bar{J}(\hat{\theta}_t) \to \inf_{\theta \in \Theta} \bar{J}(\theta)$, as $t \to \infty$, a.s..

Proof. Observe that the following chain of inequalities holds:

$$\begin{aligned}\bar{J}(\hat{\theta}_t) &= J_t(\hat{\theta}_t) + (\bar{J}(\hat{\theta}_t) - J_t(\hat{\theta}_t)) \\ &\leq J_t(\hat{\theta}_t) + \sup_{\theta\in\Theta} \left|\bar{J}(\theta) - J_t(\theta)\right| \\ &= \inf_{\theta\in\Theta} \left[\bar{J}(\theta) + \left(J_t(\theta) - \bar{J}(\theta)\right)\right] + \sup_{\theta\in\Theta} \left|\bar{J}(\theta) - J_t(\theta)\right| \\ &\leq \inf_{\theta\in\Theta} \bar{J}(\theta) + 2\sup_{\theta\in\Theta} \left|\bar{J}(\theta) - J_t(\theta)\right|.\end{aligned}$$

Therefore,

$$\inf_{\theta\in\Theta} \bar{J}(\theta) \leq \bar{J}(\hat{\theta}_t) \leq \inf_{\theta\in\Theta} \bar{J}(\theta) + 2\sup_{\theta\in\Theta} \left|\bar{J}(\theta) - J_t(\theta)\right|.$$

The thesis then immediately follows from the assumption that $\sup_{\theta\in\Theta} |J_t(\theta) - \bar{J}(\theta)| \to 0$, as $t \to \infty$, a.s.. □

Lemma 1 is instrumental to our derivations, and is of general use. It actually proved to be useful in contexts different from identification, such as adaptive and robust control ([21], [22]), where the problem to be solved was replacing the minimization of the expected value of some control cost with the minimization of its empirical mean, which is much easier to compute.

Our objective now is showing that the uniform convergence property (12) is satisfied in our setting.

Remark 2: The problem of uniform convergence of empirical means has been studied in various contexts and, in particular, in the econometric literature. In [23] conditions are given under which point-wise convergence implies uniform convergence. The interesting feature of the approach for proving uniform convergence adopted in the present paper is that, in contrast with the approach in [23], it is based on results and bounds that hold true for an arbitrary finite number of data. These intermediate results could be useful for analyzing finite-sample property of PEM methods. □

Define

$$h(w;\theta) := (z - f(x;\theta))^2,$$

where $w = (z, x) \in W := Y \times X$, with $X := Y^n \times U^m$.

The empirical cost (9) and the theoretical cost (10) can be expressed in terms of function $h(\cdot;\theta)$ as

$$\begin{aligned}J_t(\theta) &= \frac{1}{t}\sum_{k=1}^{t} h((y_k, \phi_{k-1});\theta) \\ \bar{J}(\theta) &= E_P[h((y_k, \phi_{k-1});\theta)].\end{aligned}$$

Property (12) is then the almost sure convergence of empirical means (ASCEM) property for the family of functions $\{h(\cdot;\theta) : W \to \Re_+,\ \theta \in \Theta\}$ (here θ is regarded as a parameter which identifies a single function $h(\cdot,\theta)$ from W to $\Re_+$) with respect to the process $\{w_k = (y_k, \phi_{k-1})\}$ ([24]).

The ASCEM property of a family of function has been studied mainly with reference to function approximation and classification problems, where i) the process $\{w_k\}$ is a sequence of independent and identically distributed (i.i.d.) random variables, and ii) the function to approximate is bounded or takes a finite number of values.

As for condition ii), due to the continuity of function f in (8), the fact that Θ is compact, and Assumption 1, $M := \max_{w\in W, \theta\in\Theta} h(w;\theta)$ exists and is finite, and, therefore, $h(\cdot;\theta) : W \to [0, M]$, $\forall\theta \in \Theta$.

As for the i.i.d. assumption, it is not verified in our setting since we are trying to identify a dynamical system. Nevertheless, we shall first study the i.i.d. case. Based on the results obtained in this case, we shall then prove that the ASCEM property holds in the case of interest, where $\{w_k\}$ is a strictly stationary process with geometric beta-mixing properties.

A. ASCEM property in the i.i.d. case

Suppose that $\{w_k\}$ is a sequence of i.i.d. random variables taking values in the compact set W, each one with the same distribution P (equal to the joint distribution of y_k and ϕ_{k-1}). Define

$$q(t,\epsilon) := P^t\{(w_1, \dots, w_t) : \sup_{\theta\in\Theta} |\tilde{J}_t(\theta) - \bar{J}(\theta)| > \epsilon\}, \tag{13}$$

where the product probability $P^t = P \times P \times \cdots \times P$, t times, represents the joint probability distribution of the i.i.d. random variables $(w_1, \dots, w_t)$, and $\tilde{J}_t(\theta) := \frac{1}{t}\sum_{k=1}^{t} h(w_k;\theta)$ is the empirical cost obtained from samples w_k, $k = 1, \dots, t$, independently extracted from W with the same distribution P.

If $q(t,\epsilon) \to 0$, as $t \to \infty$, $\forall\epsilon > 0$, then, $\sup_{\theta\in\Theta} |\tilde{J}_t(\theta) - \bar{J}(\theta)| \to 0$ in probability, or, equivalently, $\{h(\cdot;\theta),\ \theta \in \Theta\}$ has the uniform convergence of empirical means property in probability (UCEM property) with respect to the i.i.d. process $\{w_k\}$.

In the last two decades, the UCEM property for general classes of functions has been largely studied in the statistical learning literature. General conditions for this property to hold are now available (see e.g. [24]-[28]). For our purposes the main result is that $\{h(\cdot;\theta) : W \to [0, M],\ \theta \in \Theta\}$ has the UCEM property if its Pollard(P)-dimension (cf. [24, pag.74]) is finite. Moreover, letting d be the P-dimension, $q(M,\epsilon)$ defined in (13) is upper bounded as follows ([24, Theorem 7.1]):

$$q(t,\epsilon) \leq 8\left(\frac{16Me}{\epsilon}\ln\frac{16Me}{\epsilon}\right)^d \exp\left(-\frac{t\epsilon^2}{32M^2}\right). \tag{14}$$

Then, if d is finite, $\sum_{t=1}^{\infty} q(t,\epsilon) < \infty$, $\forall\epsilon > 0$, and, by the Borel-Cantelli Lemma, the a.s. convergence of $\sup_{\theta\in\Theta} |\tilde{J}_t(\theta) - \bar{J}(\theta)|$ to zero (or equivalently the ASCEM property for $\{h(\cdot;\theta) : W \to [0, M],\ \theta \in \Theta\}$ with respect to the i.i.d. process $\{w_k\}$) immediately follows from (14).

We next show that the P-dimension of $\{h(\cdot,\theta) : W \to [0, M],\ \theta \in \Theta\}$ is actually finite, thus concluding the proof that $\{h(\cdot,\theta) : W \to [0, M],\ \theta \in \Theta\}$ has the ASCEM property in the i.i.d. case.

Reportedly, the computation of a P-dimension on the basis of its definition is a hard task, and this has been for long an important bottleneck in the application of the theory of uniform convergence of empirical means. In [26] and [27], a powerful technique for the evaluation of the P-dimension of a function class satisfying general conditions has been introduced. This technique is used for proving the following result.

Proposition 1:

$$\text{P-dimension}(\{h(\cdot,\theta) : W \to [0,M],\, \theta \in \Theta\}) < \infty \qquad (15)$$

Proof. Consider the family of functions

$$\mathcal{H} = \{h(\cdot,\theta) : W \to [0,M], \theta \in \Theta\},$$

where $W \subset \Re^{n+m+1}$ and $\Theta \subset \Re^{(p+1)(n+m+2)-1}$.
Given a function $h(\cdot,\theta)$ in $\mathcal{H}$, let

$$g((w,c);\theta) := H(h(w;\theta) - c),$$

where $c \in [0,M]$ is an additional variable and $H(\cdot)$ is the Heaviside function ($H(x) = 1$, if $x \geq 0$, $H(x) = 0$, if $x < 0$). Also, let

$$\mathcal{G} := \{g((\cdot,\cdot);\theta) : W \times [0,M] \to \{0,1\}, \theta \in \Theta\}.$$

Then, by [24, Lemma 10.1]

$$\text{P-dimension}(\mathcal{H}) = \text{VC-dimension}(\mathcal{G})$$

(see e.g. [24, pag. 69] for the definition of the VC-dimension). Thus, the original problem of computing the P-dimension($\mathcal{H}$) is reduced to the one of computing the VC-dimension($\mathcal{G}$). This computation can be carried out by resorting to [24, Corollary 10.2] as explained next.
Given any set S, let I_S be the indicator function of S. We prove below that $g((w,c);\theta)$ can be written as

$$g((w,c);\theta) = I_S((w,c),\theta), \qquad (16)$$

where $S \subset \Re^{n+m+1} \times [0,M] \times \Re^{(p+1)(n+m+2)-1}$ is a set with a particular structure. Precisely,

$$S = \text{Boolean formula applied to } \{S_i\}_{i=1}^{2^p+p},$$

where S_i, $i = 1,\ldots,2^p + p$, are sets given by $S_i = \{\tau_i((w,c),\theta) > 0\}$ with τ_i polynomials in θ whose largest degree is $v = 4$ (a Boolean formula is any set expression containing union, intersection, and complementation).
Before proving (16), we note that the statement (15) can be obtained from (16) by applying the bound on the VC-dimension($\mathcal{G}$) in [24, Corollary 10.2], which, in our notations, can be written as follows

$$\begin{aligned}\text{VC-dimension}(\mathcal{G}) &\leq 2q\log_2(4ev(2^p + p)) \\ &= 2q\log_2(16e(2^p + p)),\end{aligned}$$

where $q := (p+1)(n+m+2) - 1$.

The proof is now completed by showing (16).
Let us recall that

$$h(w;\theta) := (z - f(x;\theta))^2,\ w = (z,x) \in W,$$

where f has the following piecewise linear structure:

$$f(x;\theta) = \gamma_0 + \alpha_0 x + \sum_{i=1}^{p} \eta_i \left|\gamma_i + \alpha_i x\right|,$$

being $\theta = [\gamma_0\ \alpha_0\ \eta_1\ \gamma_1\ \alpha_1\ \ldots\ \eta_p\ \gamma_p\ \alpha_p]$.

Consider the set $\mathcal{I}$ of all subsets of set $K = \{1,2,\ldots,p\}$. Let as denote the elements of $\mathcal{I}$ as $I_1, I_2, \ldots, I_{2^p}$.
For every $i = 1,\ldots,2^p$, define

$$S_i = \{\tau_i((w,c),\theta) > 0\}, \quad i = 1,\ldots,2^p,$$

with

$$\begin{aligned}\tau_i(((z,x),c),\theta) =& - \Big[z - \Big(\gamma_0 + \alpha_0 x + \sum_{k\in I_i}(\eta_k\alpha_k + \eta_k\gamma_k x)\Big) \\ & - \sum_{k\in K\setminus I_i}(\eta_k\alpha_k + \eta_k\gamma_k x)\Big)\Big]^2 + c.\end{aligned}$$

Define p additional sets as follows:

$$S_i = \{\tau_i((w,c),\theta) > 0\}, \quad i = 2^p+1,\ldots,2^p+p,$$

where

$$\tau_i((w,c),\theta) = \alpha_i + \gamma_i x, \quad i = 2^p+1,\ldots,2^p+p.$$

If we set

$$S = \cup_{i=1}^{2^p}\left(\bar{S}_i \cap \left(\left(\cup_{k\in I_i} S_{2^p+k}\right) \cap \left(\cup_{k\in K\setminus I_i} \bar{S}_{2^p+k}\right)\right)\right),$$

then (16) follows. □

B. ASCEM property in the beta-mixing case

Consider the strictly stationary stochastic process $\{w_k = (y_k, \phi_{k-1})\}$ satisfying the beta-mixing Assumption 2. Each random variable w_k takes values over the compact set W and has probability distribution P.

Similarly to the i.i.d. case, we define

$$q_{\text{mix}}(t,\epsilon) := P_t\{(w_1,\ldots,w_t) : \sup_{\theta\in\Theta} |J_t(\theta) - \bar{J}(\theta)| > \epsilon\},$$

where P_t denotes the joint probability distribution of the random variables $(w_1,\ldots,w_t)$, and $J_t(\theta)$ and $\bar{J}(\theta)$ are the empirical and theoretical costs defined in (9) and (10).

By the same line of reasoning as in the i.i.d. case, if we can show that $q_{\text{mix}}(t,\epsilon)$ tends to zero as $t \to \infty$ at a sufficiently fast rate, $\forall\epsilon > 0$, then the ASCEM property for the family of functions $\{h(\cdot;\theta) : W \to [0,M],\ \theta \in \Theta\}$ with respect to the process $\{w_k = (y_k, \phi_{k-1})\}$ follows from the Borel-Cantelli Lemma. The asymptotic optimality result (6) is then an immediate consequence of Lemma 1.

We next show that $q_{\text{mix}}(t,\epsilon)$ tends to zero as $t \to \infty$ at a faster rate than $\frac{1}{t^2}$, $\forall\epsilon > 0$, thus concluding the proof of asymptotic optimality. The following result proven in [20, Theorem 2] is fundamental for this.

Proposition 2 ([20, Theorem 2]): Fix a sequence of integers $\{k_t\}$ such that $k_t \leq t$, $\forall t$. Then,

$$q_{\text{mix}}(t,\epsilon) \leq t\beta_{k_t} + k_t \max\{q(l_t+1,\epsilon), q(l_t,\epsilon)\}, \qquad (17)$$

where $l_t = \lfloor t/k_t \rfloor$ denotes the integer part of t/k_t and $q(k, \epsilon)$ is defined in (13). □

Proposition 3:

$$q_{\text{mix}}(t, \epsilon) = o(1/t^2), \quad \forall \epsilon > 0$$

Proof. By plugging in (17) the bounds on β_t and $q(t, \epsilon)$ respectively given in Assumption 2 and (14), we get

$$q_{\text{mix}}(t, \epsilon) \leq t\rho^{k_t} + k_t c(\epsilon)\bar{\rho}(\epsilon)^{l_t}$$

where we set $c(\epsilon) := 8\left(\frac{16Me}{\epsilon}\ln\frac{16Me}{\epsilon}\right)^d$ and $\bar{\rho}(\epsilon) := \exp\left(-\frac{\epsilon^2}{32M^2}\right)$. If we define

$$\nu(\epsilon) := \max\{\rho, \bar{\rho}(\epsilon)\}(< 1)$$

and choose $k_t = \lfloor \sqrt{t} + 1 \rfloor$, then, it is easily seen that, for all $t \geq 3$,

$$q_{\text{mix}}(t, \epsilon) \leq t\nu(\epsilon)^{\sqrt{t}} + (\sqrt{t} + 1)c(\epsilon)\bar{\nu}(\epsilon)^{\sqrt{t}-2}.$$

Since the right-hand-side of this equation is a $o(1/t^2)$, this concludes the proof. □

IV. Conclusions

In this paper we adopted a statistical learning theory approach for the analysis of the asymptotic learning capability of a PEM identification method, when the model to be identified is piecewise affine and the input-output data are collected from the system operating in stationary conditions. We showed that if the input-output processes are geometrically beta-mixing, then, the model with the best prediction performance is identified asymptotically. This result is proven with reference to single-input/single-output systems affine systems, but it can be generalized to the multiple-input/multiple-output case.

The adopted methodology could be useful for studying the case when the model to be identified belongs to a more general class of hybrid systems. Also, some of the intermediate results obtained in this paper are not asymptotic, hence, they could be a good starting point for assessing the quality of models identified based on a finite number of data. This is an interesting direction of research.

V. Acknowledgments

The author would like to thank Prof. M.C. Campi and Dr. M. Vidyasagar for the interesting and inspiring discussions on system identification and statistical learning theory.

References

[1] L. Ljung, *System Identification: Theory for the User*. Prentice Hall, Englewood Cliffs, NJ, 1999.

[2] W. P. M. H. Heemels, B. De Schutter, and A. Bemporad, “Equivalence of hybrid dynamical models,” *Automatica*, vol. 37, no. 7, pp. 1085–1091, 2001.

[3] E. D. Sontag, “Interconnected automata and linear systems: A theoretical framework in discrete time,” in *Hybrid Systems III: Verification and Control*, ser. Lecture Notes in Computer Science, R. Alur, T. A. Henzinger, and E. D. Sontag, Eds., no. 1066. Springer-Verlag, 1996, pp. 436–448.

[4] A. Bemporad and M. Morari, “Control of systems integrating logic, dynamics, and constraints,” *Automatica*, vol. 35, no. 3, pp. 407–428, 1999.

[5] A. Bemporad, F. Borrelli, and M. Morari, “Piecewise linear optimal controllers for hybrid systems,” in *Proc. Americal Control Conf.*, Chicago, IL, 2000.

[6] A. Bemporad, G. Ferrari-Trecate, and M. Morari, “Observability and controllability of piecewise affine and hybrid systems,” *IEEE Trans. Automat. Contr.*, vol. 45, pp. 1806–1876, 2000.

[7] L. Breiman, “Hinging hyperplanes for regression, classification, and function approximation,” *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 999–1013, 1993.

[8] F. Borrelli, A. Bemporad, M. Fodor, and D. Hrovat, “A hybrid approach to traction control,” in *Hybrid Systems: Computation and Control. 4th International Workshop (HSCC01)*, ser. Lecture Notes in Computer Science, M. D. Di Benedetto and A. Sangiovanni-Vincentelli, Eds., vol. 2034. Springer-Verlag, 2001, pp. 162–174.

[9] A. Bemporad, J. Roll, and L. Ljung, “Identification of hybrid systems via mixed-integer programming,” in *Proc. of the 40th IEEE Conf. on Decision and Control*, 2001, pp. 786–792.

[10] J. Roll, “Robust verification and identification of piecewise affine systems,” Ph.D. dissertation, Linköping University, 2001.

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

[12] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, “A greedy approach to identification of piecewise affine models,” in *Hybrid Systems: Computation and Control*, ser. Lecture Notes in Computer Science, O.Maler and A.Pnueli, Eds., vol. 2623. Springer-Verlag, 2003, pp. 97–112.

[13] L. Ljung, “Convergence analysis of parametric identification methods,” *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 770–783, 1978.

[14] T. Johansen and E. Weyer, “On convergence proofs in system identification - a general principle using ideas from learning theory,” *Syst. & Contr. Lett.*, vol. 34, pp. 85–92, 1998.

[15] V. N. Vapnik, *Estimation of dependences based on empirical data*. Springer-Verlag, 1982.

[16] M. Vidyasagar and R. Karandikar, “A learning theory approach to system identification and stochastic adaptive control,” in *IFAC Symp. on Adaptation and Learning*, August 2001.

[17] ——, “System identification: A learning theory approach,” in *Proc. 40th IEEE Conf. on Decision and Control*, December 2001.

[18] L. O. Chua and A.-C. Deng, “Canonical piecewise-linear representation,” *IEEE Trans. Circuits Syst.*, vol. 35, no. 1, pp. 101–111, 1988.

[19] R. Karandikar and M. Vidyasagar, “Probably approximately correct learning with beta mixing input sequences,” private communication.

[20] ——, “Rates of uniform convergence of empirical means with mixing processes,” *Statistics and Probability Letters*, vol. 58, pp. 297–307, 2002.

[21] M. C. Campi and M. Prandini, “Randomized algorithms for the synthesis of cautious adaptive controllers,” *Syst. & Contr. Letters*, vol. 49, pp. 21–36, 2003.

[22] M. Vidyasagar, “Randomized algorithms for robust controller synthesis using statistical learning theory,” *Automatica*, vol. 37, no. 10, pp. 1515–28, 2001.

[23] D. W. Andrews, “Generic uniform convergence,” *Econometric Theory*, vol. 8, pp. 241–257, 1992.

[24] M. Vidyasagar, *A theory of learning and generalization: with applications to neural networks and control systems*. Springer-Verlag, London, 1997.

[25] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag, New York, Inc., 1995.

[26] M. Karpinski and A. J. Macintyre, “Polynomial bounds for VC dimension of sigmoidal neural networks,” in *Proc. 27th ACM Symp. on Theory of Computing*, 1995, pp. 200–208.

[27] ——, “Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks,” *J. Comput. Sys. Sci.*, vol. 54, pp. 169–176, 1997.

[28] V. N. Vapnik, *Statistical learning theory*. John Wiley & Sons, Inc., 1998.