A randomized approach to switched nonlinear systems identification

Federico Bianchi * **Maria Prandini** * **Luigi Piroddi** *

* *Dipartimento di Elettronica, Informazione e Bioingegneria,*
Politecnico di Milano, Milano, Italy
(e-mail: {federico.bianchi, maria.prandini, luigi.piroddi}@polimi.it).

Abstract: This paper addresses the identification of Switched Nonlinear AutoRegressive eXogenous (SNARX) systems characterized as a collection of nonlinear dynamical systems (modes), each one described via a discrete time NARX model, indexed by a discrete-valued variable (switching signal). We propose a novel approach which, given a realization of the input/output signals collected from the system, jointly classifies the data attributing them to the different modes, and identifies the model structure and parameters for each mode. The involved optimization problem is partly combinatorial due to the data classification over modes and the model structure selection, and partly continuous due to the parameter estimation required to complete the identification of the dynamical models assigned to the different modes. A probabilistic framework is employed to address the problem, where Categorical and Bernoulli distributions are respectively used for the assignment of modes over time and for the structure selection of the NARX models describing the modes. A randomized procedure is then proposed to solve the problem, based on a sample-and-evaluate strategy that progressively refines the induced SNARX model probability distribution. The approach is tested on a numerical example taken from the literature, where it shows promising results.

Keywords: NARX model identification, switched systems, randomized algorithms.

1. INTRODUCTION

Hybrid systems are dynamical systems characterized by interleaved continuous and discrete dynamics. Most research regarding the identification of hybrid systems has focused on switched affine (SA) and piecewise affine (PWA) models which are both characterized by a finite set of affine continuous dynamics (modes) and a discrete-valued switching signal that dictates the switching between modes. In SA systems, the switching signal is an exogenous input, while in PWA systems it is an endogenous signal generated by the system evolution. Matching input-output formalisms have also been introduced, namely the Switched ARX (SARX) and the PieceWise affine ARX (PWARX) models, respectively. In general terms, the identification task requires to jointly classify the data (assigning samples to modes) and estimating the model parameters (for each mode). This task can be accomplished by solving a mixed-integer optimization problem over the model structure and parameters, and the discrete variables governing the assignment of the samples. The complexity of the optimization problem stems from its underlying combinatorial nature, related to both the sample assignment to modes and the model structure selection.

Many approaches have been proposed over the last two decades for the case of *linear* local models, (Paoletti et al., 2007), (Garulli et al., 2012). Far fewer works have tackled the extension to *nonlinear* local models, which is often required in complex modeling applications. Indeed, if the underlying hybrid system has nonlinear dynamics, using a hybrid model with linear local models generally induces an unnecessary multiplication of the modes (multiple linear local models are required to adequately reproduce the nonlinear dynamics) and consequently of the switching instants, aggravating the combinatorial nature of the identification problem, and impairing the interpretation of the switching signal in terms of actual physical phenomena. A first attempt in this direction is documented in (Lauer and Bloch, 2008), where a method based on kernel regression and Support Vector Machine (SVM) is discussed. This method however suffers from the curse-of-dimensionality problem since it optimizes over a set of variables that grows with the number of data N and the number of modes N_M. This work was extended in (Lauer et al., 2011), leading to a method that can be applied to large datasets thanks to a reduced-size kernel function, (Le et al., 2011). However, the resulting continuous optimization problem is non-convex and thus global convergence is not guaranteed. In (Bako et al., 2010) a method is proposed instead, which ultimately results in a convex minimization problem after a relaxation step. However, a sufficient condition guaranteeing the optimality of the relaxed convex problem solution was provided only under a noiseless assumption. The most recent contribution dealing with nonlinear hybrid systems was proposed in (Le et al., 2013) as an extension to the sparse optimization based method, (Bako et al., 2010). Specifically, a kernel expansion was introduced as in (Le et al., 2011) to deal with nonlinearities and the notion of robust sparsity was introduced to treat noisy data. This approach results in a sequence of relaxed convex optimization programs, each equivalently formulated as a Support Vector Regression (SVR) problem that can be solved efficiently even for large data sets. However, this method still relies on a convex relaxation strategy, which does not guarantee the equivalence with the original problem. Furthermore, it requires the careful setting of several parameters, which appears to be far from trivial (*e.g.*, factor C, which defines the trade-off between model complexity and accuracy, or the weights w_i adopted to improve the sparsity of the solution).

This paper introduces an iterative randomized approach for the identification of Switched Nonlinear ARX (SNARX) models. The optimization problem is reformulated in a probabilistic framework by introducing Categorical and Bernoulli distributions governing respectively the attribution of the samples to the modes and the selection of the model structures of the local models. More in detail, a Categorical distribution is associated to represent the mode assigned to each time instant, and a Bernoulli distribution models the presence of a given term in a local model. The optimization problem can be tackled by means of an iterative sample-and-evaluate approach which allows to progressively refine the overall probability distribution representing the hybrid model. Specifically, at each step one samples the Categorical distributions associated to the time instants to generate instances of the switching signal and (independently) extracts instances of the local models from the Bernoulli distributions. This sampling process is used to gather information to tune the probability distribution by reinforcing the probability of the most promising mode assignments and local model terms. The progressive refinement of the probability distribution ultimately tends to a limit distribution representing a precise hybrid model. Under the assumption that there exists only one optimal switched model, one can prove that the solution of the reformulated problem equals that of the original optimization problem. The method requires few a-priori assumptions and design parameters. Furthermore, it is capable of operating with noisy data and medium-large datasets.

2. PRELIMINARIES

Consider a data-set of N input-output pairs $\{(u(t), y(t)), t = 1, \ldots, N\}$ collected from a single-input single-output system. The objective of the identification problem is to determine a relationship between past observations $\{u(t-1), u(t-2), \ldots, y(t-1), y(t-2), \ldots\}$ and future predicted output $\hat{y}(t)$, in the form of an input-output recursive model:

$$\hat{y}(t) = g(u(t-1), u(t-2), \ldots, y(t-1), y(t-2), \ldots),$$

so as to best fit the observations in the data-set.

2.1 The NARX model class

In this work we assume that the nonlinear dynamics of each mode can be represented by a NARX model of the type:

$$y(t) = g\left(\boldsymbol{x}(t); \boldsymbol{\vartheta}\right) + e(t)$$

where $\boldsymbol{x}(t) = [y(t-1), \ldots, y(t-n_y), u(t-1), \ldots, u(t-n_u)]$ is a finite-dimensional vector of the most recent past observations (n_y and n_u being suitable maximum lags), $e(t)$ is a stochastic process characterized as a sequence of i.i.d. zero mean random variables, and $g(\cdot)$ is an unknown nonlinear function parameterized via a vector $\boldsymbol{\vartheta} = [\vartheta_1, \ldots \vartheta_n]^T$ of coefficients. NARX models are the natural extension of ARX models to the nonlinear case, and have earned widespread interest in the literature in view of their flexibility and representation capabilities, (Billings, 2013). The corresponding prediction form is given by:

$$\hat{y}(t) = g\left(\boldsymbol{x}(t); \boldsymbol{\vartheta}\right). \tag{1}$$

In NARX models, the nonlinear mapping $g(\cdot)$ is often expressed as a linear combination of (nonlinear) basis functions φ_j, $j = 1, \ldots, n$:

$$g\left(\boldsymbol{x}(t); \boldsymbol{\vartheta}\right) = \sum_{j=1}^{n} \vartheta_j \varphi_j(t),$$

so that the predictor can be reduced to the following compact notation:

$$\hat{y}(t) = \boldsymbol{\varphi}(t)^T \boldsymbol{\vartheta}, \tag{2}$$

where all basis functions are collected in the *regression vector* $\boldsymbol{\varphi} = [\varphi_1, \ldots \varphi_n]^T$. Accordingly, the elements of the regression vector are called *regressors*. Parameter estimation is carried out with the least squares (LS) algorithm followed by a statistical Student's t-test to determine the statistical relevance of each regressor, and if some regressor turns out not to be statistically relevant, it is classified as *redundant*.

A popular choice for the representation of $g(\cdot)$ is the polynomial functional expansion, whereby the regressors are monomials of elements in $\boldsymbol{x}(t)$ up to a given order n_d. This functional expansion extends gracefully from linear models and typically allows an easier model interpretation but it suffers from the curse of dimensionality. However, one is not bound to use full polynomial expansions and, indeed, models with few selected terms can provide highly accurate and robust models. A crucial component of NARX model identification methods is therefore the selection of the essential terms of a model.

2.2 The SNARX model class

A SNARX system is a collection of NARX systems (2) indexed by a finite-valued switching signal σ. Given a regression vector $\boldsymbol{\varphi}$ of size n, the output predictor of a SNARX model is defined as

$$\hat{y}(t) = \boldsymbol{\varphi}(t)^T \boldsymbol{\vartheta}^{(\sigma(t))}, \tag{3}$$

where $\sigma(t) \in \{1, \ldots, N_M\}$ is the value taken by the switching signal at time t and defines which mode is active at that time instant, N_M is the number of modes, and $\boldsymbol{\vartheta}^{(i)}$ is the parameter vector defining the dynamics of the i-th NARX mode, including its structure ($\boldsymbol{\vartheta}^{(i)}$ has zero entries for the regressors that are not present in the i-th mode).

The identification problem for a SNARX model (3) consists in estimating from a data-set of N input-output pairs $\{(y(t), u(t)), t = 1, \ldots, N\}$, the number of modes N_M and the corresponding model parameterizations $\boldsymbol{\vartheta}^{(i)}$, $i = 1, \ldots, N_M$, as well the switching signal $\sigma(t)$, $t = 1, \ldots, N$. Consistently with most of the literature on switched system identification, we here address the case where Assumption 1 holds.

Assumption 1. The number of modes N_M is known.

Under Assumption 1, one may tackle the identification problem by adopting a classical prediction error approach and solving the following optimization problem [1]:

$$\min_{\{\{\beta_i(t)\}_{t=1}^N, \boldsymbol{\vartheta}^{(i)}\}_{i=1}^{N_M}} \sum_{t=1}^{N}\sum_{i=1}^{N_M} \beta_i(t)\cdot\left(y(t)-\boldsymbol{\varphi}(t)^T\boldsymbol{\vartheta}^{(i)}\right)^2$$

$$\text{subject to:} \qquad (4)$$

$$\sum_{i=1}^{N_M}\beta_i(t)=1,\ \beta_i(t)\in\{0,1\},\ t=1,\dots N, i=1,\dots,N_M,$$

where $\beta_i(t)$ is a binary variable which encodes the assignment of sample t to mode i, from which the switching signal $\sigma(t)$, $t=1,\dots,N$, can be reconstructed according to:

$$\sigma(t)=i \Longleftrightarrow \beta_i(t)=1. \qquad (5)$$

Unfortunately, the optimization problem (4) is a mixed integer program which is typically computationally intractable due to its combinatorial nature.

Motivated by this observation, and inspired by the Randomized Model Structure Selection (RaMSS) method, (Falsone et al., 2015), we propose in the next section a probabilistic approach to determine a solution to (4).

3. A PROBABILISTIC APPROACH TO SNARX IDENTIFICATION

Let $\mathcal{S}=\{0,1\}^n$ be the set of all possible NARX model structures $\boldsymbol{s}$ compatible with the set of regressors $\{\varphi_1,\cdots,\varphi_n\}$, such that $s(k)=1$ if φ_k belongs to it, and $s(k)=0$ otherwise. If a SNARX model (3) has N_M modes, then its structure can be defined as a collection $s=\left(\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N_M)}\right)$ of N_M NARX structures, with $\boldsymbol{s}^{(i)}\in\mathcal{S}$, $i=1,\dots,N_M$. Let also $\boldsymbol{\sigma}=[\sigma(1),\dots,\sigma(N)]$ denote the switching signal along $[1,N]$, and $\Sigma=\{1,\dots,N_M\}^N$ the set where $\boldsymbol{\sigma}$ takes values.

We can now associate a candidate SNARX model to a pair $\lambda=(\boldsymbol{\sigma},s)$ taking values in $\Lambda=\Sigma\times\mathcal{S}^{N_M}$ and rate it according to the performance index $\mathcal{L}:\Lambda\to[0,1]$ defined as

$$\mathcal{J}(\lambda)=e^{-K_\lambda\mathcal{L}(\lambda)}, \qquad (6)$$

where $K_\lambda>0$ is a design parameter and

$$\mathcal{L}(\lambda)=\min_{\{\boldsymbol{\vartheta}^{(i)}\}_{i=1}^{N_M}}\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{N_M}\beta_i(t)\cdot\varepsilon_i^2(t). \qquad (7)$$

Note that for a given $\lambda=(\boldsymbol{\sigma},s)$, $\beta_i(t)$ in (7) is uniquely determined by $\boldsymbol{\sigma}$ through (5), whereas $\varepsilon_i(t)=y(t)-\hat{y}_{\boldsymbol{s}^{(i)}}(t)$ is the prediction error associated to mode i with structure $\boldsymbol{s}^{(i)}$ and parameters $\boldsymbol{\vartheta}^{(i)}$.

Computing $\mathcal{L}(\lambda)$ in (7) involves segmenting the data-set in multiple data-sets associated with the different NARX modes and then computing the parameters of each NARX mode i with structure $\boldsymbol{s}^{(i)}$ solving a LS problem. Indeed, given a switching signal $\boldsymbol{\sigma}$ (or, equivalently, the corresponding $\beta_i(t)$ variables, see (5)), we can define the number of pairs in the data-set that are associated with mode i as $N_i=\sum_{t=1}^N\beta_i(t)$. Then, $\mathcal{L}(\lambda)$ with $\lambda=(\boldsymbol{\sigma},(\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N_M)})$ can be rewritten as

$$\mathcal{L}(\lambda)=\frac{1}{N}\sum_{i:\,N_i\neq 0}N_i\cdot\mathcal{L}_i(\boldsymbol{\sigma},\boldsymbol{s}^{(i)}), \qquad (8)$$

where $\mathcal{L}_i(\boldsymbol{\sigma},\boldsymbol{s}^{(i)})$ measures the accuracy of the model of the ith mode, when assuming structure $\boldsymbol{s}^{(i)}$ and with the switching signal $\boldsymbol{\sigma}$. Index $\mathcal{L}_i(\boldsymbol{\sigma},\boldsymbol{s}^{(i)})$ is well-defined if $N_i\neq 0$ and is given by:

$$\mathcal{L}_i(\boldsymbol{\sigma},\boldsymbol{s}^{(i)})=\min_{\boldsymbol{\vartheta}^{(i)}}\frac{1}{N_i}\sum_{t=1}^{N}\beta_i(t)\cdot\varepsilon_i^2(t). \qquad (9)$$

According to the introduced notation, the SNARX identification problem can be formalized as that of finding, among all the possible $\lambda=(\boldsymbol{\sigma},s)$ values, the one which maximizes the performance index $\mathcal{J}(\lambda)$ in (6) and does not have redundant terms in s, *i.e.,* the mode parameter $\boldsymbol{\vartheta}^{(i)}$ identified on the dataset given the data classification and SNARX structure specified by λ satisfies $\vartheta_k^{(i)}\neq 0$ for any $k\in\{1,\dots,n\}$ such that $s_k^{(i)}=1$, for all modes $i=1,\dots,N_M$. Under the assumption that there exists only one such λ, this can be written as:

$$\lambda^\star=\arg\max_{\lambda\in\Lambda}\mathcal{J}(\lambda), \qquad (10)$$

where, with a slight abuse of notation, Λ denotes the set of all non redundant λ values. The so-obtained $\lambda^\star=(\boldsymbol{\sigma}^\star,s^\star)$ defines the switching signal $\boldsymbol{\sigma}^\star$ and the structure $s^\star=\left(\boldsymbol{s}^{(1)\star},\dots,\boldsymbol{s}^{(N_M)\star}\right)$ of all the modes in the identified SNARX model. The parameters of each mode i are the solutions of the following LS problems:

$$\boldsymbol{\vartheta}^{(i)\star}=\arg\min_{\boldsymbol{\vartheta}^{(i)}}\sum_{t=1}^{N}\beta_i^\star(t)\cdot\left(y(t)-\hat{y}_{\boldsymbol{s}^{(i)\star}}(t)\right)^2, \qquad (11)$$

where $\beta_i^\star(t)$ is recovered from $\boldsymbol{\sigma}^\star$ based on (5).

Problem (10) is computationally intractable since in principle one should explore exhaustively the set of all possible candidate models Λ whose cardinality grows with the size N of the data-set, the number N_M of modes and the number n of the regressors. In analogy with what done with the RaMSS algorithm (Falsone et al., 2015), we recast the problem in a probabilistic framework by introducing a discrete random variable Φ which takes values in Λ according to some probability distribution $\mathbb{P}_\Phi$. The probability distribution $\mathbb{P}_\Phi$ will be iteratively updated based on the data-set starting from a tentative distribution, until convergence. Ideally, the tuning procedure should make $\mathbb{P}_\Phi$ converge to a point mass probability concentrated on the best SNARX model describing the actual system.

The average performance of Φ with probability distribution $\mathbb{P}_\Phi$ is given by:

$$\mathbb{E}\left[\mathcal{J}(\Phi)\right]=\sum_{\lambda\in\Lambda}\mathbb{P}_\Phi(\lambda)\mathcal{J}(\lambda), \qquad (12)$$

which is a convex combination of the performance of all SNARX models and possible switching sequences in Λ. Then the value $\lambda^\star$ that maximizes $\mathcal{J}(\lambda)$ in (10) can be also obtained as:

$$\lambda^\star=\arg\max_{\lambda\in\Lambda}\mathbb{P}_\Phi^\star(\lambda), \qquad (13)$$

where $\mathbb{P}_\Phi^\star=\arg\max_{\mathbb{P}_\Phi}\mathbb{E}\left[\mathcal{J}(\Phi)\right]$. Indeed, if $\mathbb{P}_\Phi$ varies over all the possible distributions on Λ then the maximum of

[1] Unselected and redundant regressors are set to 0 in $\boldsymbol{\vartheta}^{(i)}$, for all modes $i=1,\dots,N_M$.

the average performance in equation (12) is obtained by making all the probability mass concentrate on the best SNARX model.

A key point is to choose a suitable parametrization for the probability distribution $\mathbb{P}_\Phi$ and an appropriate tuning of such a parametrization based on the data-set processing so as to make the probability mass of $\mathbb{P}_\Phi$ concentrate on $\lambda^\star$ in (10). We here assume independence between the switching signal and the structure of the SNARX model when defining $\mathbb{P}_\Phi$ and express it as:

$$\mathbb{P}_\Phi(\lambda) = \mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\sigma}) \cdot \mathbb{P}_{\boldsymbol{\varsigma}}(s), \tag{14}$$

where $\boldsymbol{\xi}$ is the random vector associated with the switching signal and taking values in Σ according to the probability distribution $\mathbb{P}_{\boldsymbol{\xi}}$ and $\boldsymbol{\varsigma}$ is the random vector associated with the SNARX structure and taking values in $\mathcal{S}^{N_M}$ according to $\mathbb{P}_{\boldsymbol{\varsigma}}$.

3.1 Parametrization of $\mathbb{P}_{\boldsymbol{\xi}}$

We allow for switching to possibly occur at specified time instants in a sequence $\mathcal{T}_s = \{t_k\}_{k=1}^{N_s}$, with $1 = t_1 < t_2 < \cdots < t_{N_s} \leq N$, and encode this a-priori information within $\mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\sigma})$ by attributing zero probability to the values of $\boldsymbol{\sigma}$ such that $\boldsymbol{\sigma} \notin \Sigma_{\mathcal{T}_s}$, where

$$\Sigma_{\mathcal{T}_s} = \{\boldsymbol{\sigma} : \sigma(t) = \sigma(t_k),\, t \in [t_k, t_{k+1}),\, k = 1, \ldots, N_s\}$$

with $t_{N_s+1} = N + 1$. Note that the case when no a-priori information on the switching times is available can be embedded in this framework by setting $t_k = k$ and making k range from 1 to $N_s = N$.

Let $\sigma(t_k)$, $k = 1, \ldots, N_s$, be independent and distributed according to a Categorical distribution

$$\xi(t_k) \sim \texttt{Categorical}\left(\boldsymbol{p}(t_k)\right),$$

where vector $\boldsymbol{p}(t_k) = [\eta_{t_k}^{(1)}, \ldots, \eta_{t_k}^{(N_M)}]$ collects the probabilities $\eta_{t_k}^{(i)}$, $i = 1, \ldots, N_M$, of drawing any of the modes at time instant t_k (denoted Mode Extraction Probabilities (MEPs) in the following). Clearly, $\sum_{i=1}^{N_M} \eta_{t_k}^{(i)} = 1$.

Recalling the relationship (5) between $\sigma(t)$ and $\beta_i(t)$, the probability distribution of $\boldsymbol{\xi}$ with switching times $\{t_k\}_{k=1}^{N_s}$ is given by

$$\mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\sigma}) = \begin{cases} \displaystyle\prod_{k=1}^{N_s} \prod_{i=1}^{N_M} \left(\eta_{t_k}^{(i)}\right)^{\beta_i(t_k)}, & \boldsymbol{\sigma} \in \Sigma_{\mathcal{T}_s} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

3.2 Parametrization of $\mathbb{P}_{\boldsymbol{\varsigma}}$

Consider now the problem of choosing a suitable parametrization for $\mathbb{P}_{\boldsymbol{\varsigma}}(s)$, where $s = \left((\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(N_M)}\right)$, $\boldsymbol{s}^{(i)} \in \mathcal{S}$ describing a possible model structure for mode i. Following the approach introduced in (Falsone et al., 2015) we associate to each mode i a random vector $\boldsymbol{\rho}^{(i)}$ which collects the Bernoulli random variables $\rho_k^{(i)} \sim Be\left(\mu_k^{(i)}\right)$, $k = 1, \cdots, n$, where the success probability $\mu_k^{(i)}$ (denoted Regressor Inclusion Probability (RIP)) represents the belief that the regressor belongs to the true model . This induces a probability distribution:

$$\mathbb{P}_{\boldsymbol{\rho}^{(i)}}(\boldsymbol{s}^{(i)}) = \prod_{j:\varphi_j \in \boldsymbol{s}^{(i)}} \mu_j^{(i)} \prod_{j:\varphi_j \notin \boldsymbol{s}^{(i)}} \left(1 - \mu_j^{(i)}\right).$$

Under the assumption of independence between the mode structures, we then have

$$\mathbb{P}_{\boldsymbol{\varsigma}}(s) = \prod_{i=1}^{N_M} \mathbb{P}_{\boldsymbol{\rho}^{(i)}}(\boldsymbol{s}^{(i)}).$$

3.3 Update of $\mathbb{P}_\Phi$

We parameterized the probability distribution $\mathbb{P}_\Phi$ through some scalar parameters $\eta_{t_k}^{(i)}$ and $\mu_j^{(i)}$, $k = 1, \ldots, N_s$, $j = 1, \ldots, n$, $i = 1, \ldots, N_M$. We next formulate tuning rules for $\eta_{t_k}^{(i)}$ and $\mu_j^{(i)}$ aiming at concentrating the probability distribution $\mathbb{P}_\Phi$ in equation (14) on $\lambda^\star$. The tuning rules are implemented via a randomized procedure that involves extracting sample values $\lambda = (\boldsymbol{\sigma}, s)$ of $\Phi = (\boldsymbol{\xi}, \boldsymbol{\varsigma})$ and evaluating the performance of the extracted SNARX structure s based on the dataset segmentation induced by the extracted switching signal realization $\boldsymbol{\sigma}$.

To this purpose we define

$$\delta_{t_k}^{(i)} = \mathbb{E}_{\mathbb{P}_\Phi}\left[\mathcal{J}(\Phi) | \xi(t_k) = i\right] - \mathbb{E}_{\mathbb{P}_\Phi}\left[\mathcal{J}(\Phi) | \xi(t_k) \neq i\right], \tag{16}$$

which compares the average performance of Φ (switching signal and SNARX structure) where mode i is assigned to time instant t_k with the average performance of the remaining Φ's. Index $\delta_{t_k}^{(i)}$ can be used directly to update $\eta_{t_k}^{(i)}$ at the next iteration, according to:

$$\eta_{t_k}^{(i)} \leftarrow \eta_{t_k}^{(i)} + \chi \delta_{t_k}^{(i)}, \tag{17}$$

where the step size $\chi > 0$ is a design parameter.

A similar formulation applies also to the update of the RIP $\mu_j^{(i)}$ indicating the likelihood of having regressor φ_j in the structure of the ith mode. Specifically, we define:

$$\begin{aligned} \ell_j^{(i)} &= \mathbb{E}_{\mathbb{P}_\Phi}\left[\mathcal{J}_i(\boldsymbol{\xi}, \boldsymbol{\varsigma}^{(i)}) | \varphi_j \in \boldsymbol{\varsigma}^{(i)} \wedge \vartheta_j^{(i)} \neq 0\right] \\ &\quad - \mathbb{E}_{\mathbb{P}_\Phi}\left[\mathcal{J}_i(\boldsymbol{\xi}, \boldsymbol{\varsigma}^{(i)}) | \varphi_j \notin \boldsymbol{\varsigma}^{(i)} \vee \vartheta_j^{(i)} = 0\right], \end{aligned} \tag{18}$$

where we set

$$\mathcal{J}_i(\boldsymbol{\sigma}, \boldsymbol{s}^{(i)}) = \begin{cases} e^{-K_\rho \mathcal{L}_i(\boldsymbol{\sigma}, \boldsymbol{s}^{(i)})}, & N_i \neq 0 \\ 0, & N_i = 0 \end{cases} \tag{19}$$

with $\mathcal{L}_i$ defined in (9). This leads to

$$\mu_j^{(i)} \leftarrow \mu_j^{(i)} + \gamma_i \ell_j^{(i)}, \tag{20}$$

where $\gamma_i > 0$ is the step size for mode i.

The update of $\mu_j^{(i)}$ is based on the difference between the average performance of structures for mode i where φ_j appears and that of the remaining structures. Such an average performance is assessed on the segments of the data-set that are assigned to mode i in the switching signal.

An exact computation of the expected values in (16) and (18) cannot be obtained in practice, since it would require to consider exhaustively all the possible sequences and structures. We then adopt a Monte Carlo approach to estimate such values by drawing a multi-sample of Φ based on the current $\mathbb{P}_\Phi$ distribution. The representation of $\mathbb{P}_\Phi$ as the product of $\mathbb{P}_{\boldsymbol{\xi}}$ and $\mathbb{P}_{\boldsymbol{\varsigma}}$ simplifies the extraction procedure. In practice, we obtain a sample of Φ by extracting separately a sample of the switching signal $\boldsymbol{\xi}$ and a sample of the SNARX structure $\boldsymbol{\varsigma}$. The update rules in (17) and (20) are then implemented based on the estimated expected values, and the updated sequences $\eta_{t_k}^{(i)}$ and $\mu_j^{(i)}$ are renormalized in order to represent valid probabilities.

3.4 Implementation issues

The choice of the step sizes χ and γ_i, $i = 1, \ldots, N_M$, in (17) and (20), is crucial since it influences the convergence of the algorithm. Following (Falsone et al., 2015), we adaptively tune χ and γ_i, $i = 1, \ldots, N_M$, taking into account the dispersion of the performance of the associated SNARX model for χ, and the NARX mode i for γ_i, as induced by $\mathbb{P}_\Phi$. Specifically,

$$\chi = \left(10\left(\mathcal{J}_{\text{best}} - \overline{\mathcal{J}}\right) + 0.1\right)^{-1} \tag{21}$$

and

$$\gamma_i = \left(10\left(\mathcal{J}_{i,\text{best}} - \overline{\mathcal{J}}_i\right) + 0.1\right)^{-1} \tag{22}$$

where $\mathcal{J}_{\text{best}}$ and $\mathcal{J}_{i,\text{best}}$, $\overline{\mathcal{J}}$ and $\overline{\mathcal{J}}_i$ are, respectively, the best and mean values of $\mathcal{J}$ and $\mathcal{J}_i$, evaluated on the extracted samples for Φ. The idea underlying expressions (21–22) is that in its early stages the algorithm has to collect information by freely exploring the solution space and thus the correction terms should have a low weight. As the algorithm proceeds, the parameter corrections will become more reliable and convergence will be sped up (the lower is the dispersion in the performance of the extracted models, the higher the value of the step sizes).

The convergence of the algorithm is also influenced by the choice of K_λ and K_ρ in (6) and (19). Specifically, high values increase the convergence speed but they can trap the algorithm into local minima.

A suitable initialization of the algorithm is obtained by setting equal small probabilities $\mu_j^{(i)}$, thus encouraging the extraction of small models at the early stages of the algorithm. With a similar rationale, the mode extraction probabilities $\eta_{t_k}^{(i)}$ are initially set equal.

4. SIMULATION RESULTS

In the following tests we will consider the SNARX system presented in (Lauer and Bloch, 2008), which switches between mode 1

$$y(t) = -0.905y(t-1) + 0.9u(t-1) + e(t),$$

and mode 2

$$y(t) = -0.4y(t-1)^2 + 0.5u(t-1) + e(t),$$

where $e(t)$ is a zero mean Gaussian noise of variance 0.012 and $u(t)$ is uniformly distributed in the interval $[0, 1]$.

4.1 Example 1: Single switching

The first example illustrates a typical run of the algorithm. An observation window of length $N = 1200$ is considered, which contains a single switching event at $t = 400$ from mode 1 to mode 2. The a-priori information about the switching time is employed to divide the observation window into two time periods, before and after the switch, obtaining $\Sigma = \{(1,1), (1,2), (2,1), (2,2)\}$. Notice that the algorithm is only given the information that $t = 400$ is a *possible* switching instant, and has to determine based on the input–output data if a switching actually took place at that time. Because of this, the sequences $\{(1,1), (2,2)\}$ are also admissible (otherwise, the problem could have been easily separated into two independent NARX identification problems). The design parameters have been set to $n_y = n_u = n_d = 2$ (for a total of 15 possible regressors), $N_p = 200$, $K_\rho = K_\lambda = 10$, and the initial MEPs equally probable, whereas the initial RIPs are all set to 0.05.

Figures 1 and 2 illustrate, respectively, the evolution of the RIPs associated to the two modes. Similarly, Figure 3, shows the evolution of the MEPs for the two time-periods, *i.e.* $[1, 399]$ and $[400, 1200]$. At the onset of the algorithm, the two modes are equally probable and consequently also the respective model structures are similar. The algorithm initially assigns more probability to mode 1 for the first period (this is actually an arbitrary choice), but still remains unresolved regarding the assignment of the second period. After a first exploration phase, lasting 15–20 iterations, the algorithm operates a first selection on the model structures of the two modes. More in detail, the linear regressors $u(t-1)$ and $y(t-1)$ are picked for both modes (their RIPs rapidly rise to 1). Shortly after, the algorithm recognizes a more complex behavior in the data and selects an additional (nonlinear) regressor for the second mode ($y(t-1)^2$). At the same time, it decides that the two time periods actually correspond to different dynamics (*i.e.* there is an actual mode switching at $t = 400$), and gradually assigns the second one with greater probability to mode 2. At iteration 25, the assignment of the time periods to the modes is final, as well as the structure of mode 1 (all RIPs are set to 0, except those pertaining to the previously selected linear terms). It takes the algorithm a few more iterations to finalize also the structure of the 2^{nd} mode (when the RIP of $y(t-1)^2$ reaches 1, the last two redundant regressors are dropped). By iteration 38 the two modes have been assigned the correct structure. The corresponding identified parameters are $[-0.8928,\ 0.9063]$ for mode 1, and $[-0.4233,\ 0.5084]$ for mode 2.

It is worth mentioning that the algorithm was capable of inferring from the data the existence of both linear and nonlinear dynamics without any prior knowledge, while in (Lauer and Bloch, 2008) a linear kernel and a RBF kernel had been chosen specifically for the two modes.

4.2 Example 2: Multiple switchings

In this second example, an observation window of $N = 2000$ is considered, which contains four switchings at times $t = 400$ (from mode 1 to mode 2), $t = 1500$ (from mode 2 to mode 1), $t = 1600$ (from mode 1 to mode 2), and $t = 1700$ (from mode 2 to mode 1). This time, the only *a priori* information available to the identification algorithm is that switchings may possibly occur at times $t_k = 100(k-1)$, $k = 2, 3, \ldots, 20$. To give an idea of the underlying complexity of the combinatorial problem, this amounts to 20 sub-periods, which corresponds to $2^{20} = 1048576$ possible switching signals. The same design parameters of Example 1 have been used.

Table 1 reports the aggregated results obtained from 100 runs of the algorithm on the same data realization.

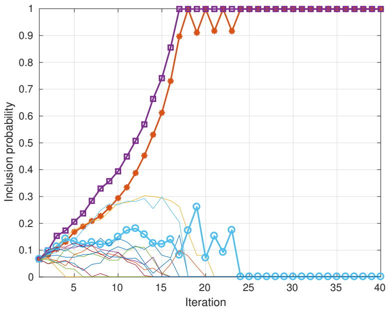

Fig. 1. Example 1: Evolution of the RIPs for the first mode ($u(t-1)$: square, $y(t-1)$: star, $y(t-1)^2$: circle).

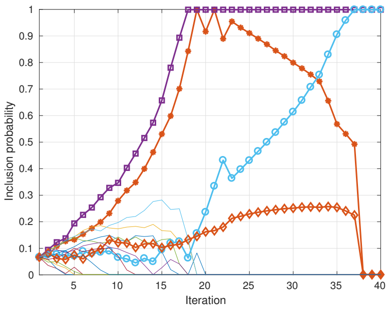

Fig. 2. Example 1: Evolution of the RIPs for the second mode ($u(t-1)$: square, $y(t-1)$: star, $y(t-1)^2$: circle, $y(t-1)u(t-2)$: diamond).

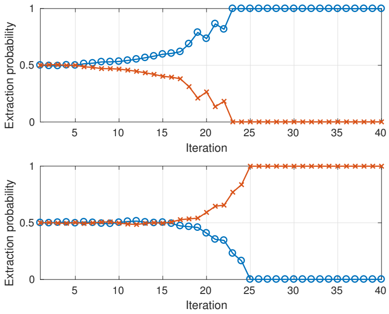

Fig. 3. Example 1: Evolution of the MEPs (circle for mode 1 and cross for mode 2), first (top) and second (bottom) time periods.

The proposed algorithm performs well in both the mode assignment and the model identification by exploring a small fraction of the total number of possible switching signals ($2^{20} = 1048576$) and models ($2^{15} - 1 = 32767$). Both the switching sequence and the model for mode 1 were identified correctly in all runs. As for mode 2, it happened sporadically (2 times) that regressor $y(t-1)$ was selected instead of $y(t-1)^2$, with a slight performance loss. As already seen in Figure 2 for Example 1, nonlinear terms tend to be selected after the linear ones, causing the algorithm to be trapped in a local minimum of the loss function (9). Indeed, $\mathcal{L}_i$ takes the value 0.0120 for the wrong model and 0.0118 for the correct one, leading to an almost negligible difference of 0.0015 between the corresponding $\mathcal{J}_i$ values. This could explain the occasional inability to escape from the local minimum, particularly in view of (22), if there were few correct models in the extracted population.

It is worth noticing that despite this occasional failure, the algorithm has always been able to capture from the data the existence of two different modes, and to assign them correctly to the subperiods.

Table 1. Ex. 2: Monte Carlo simulation results.

Average # of iterations	123.9
Average elapsed time [s]	39.14
Percentage of correct sequence selection	100%
Average # of explored sequences	16795
Percentage of correct model selection (mode 1)	100%
Average # of explored models (mode 1)	687
Percentage of correct model selection (mode 2)	98%
Average # of explored models (mode 2)	684

5. CONCLUSIONS

A randomized batch method has been presented for the identification of switched nonlinear systems, based on the NARX model family. It recasts the optimization problem which characterizes the identification of the system in a probabilistic framework by defining a probability distribution over the space of possible switched models.

The initial results with the proposed algorithm emphasize its capability of assigning correctly the time periods between switchings to the modes and in choosing the correct model structures for the different modes, which is a valuable aspect considering the non trivial interaction between these two tasks. Future work will focus on the study of the robustness of the presented approach to noise, considering also the case when the input signal is not fully exciting. Furthermore, the dependencies between modes could be modeled to improve the sampling procedure (*e.g.*, introducing a minimum dwell time), thereby allowing to address also the case with no *a priori* information on the switching times at an affordable computational cost.

REFERENCES

Bako, L., Boukharouba, K., and Lecoeuche, S. (2010). An l_0–l_1 norm based optimization procedure for the identification of switched nonlinear systems. In 49^{th} *IEEE Conference on Decision and Control*, 4467–4472.

Billings, S.A. (2013). *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains.* Wiley.

Falsone, A., Piroddi, L., and Prandini, M. (2015). A randomized algorithm for nonlinear model structure selection. *Automatica*, 60, 227–238.

Garulli, A., Paoletti, S., and Vicino, A. (2012). A survey on switched and piecewise affine system identification. In 16^{th} *IFAC Symposium on System Identification*, 344–355. Brussels, Belgium.

Lauer, F. and Bloch, G. (2008). Switched and piecewise nonlinear hybrid system identification. In *International Workshop on Hybrid Systems: Computation and Control*, 330–343.

Lauer, F., Bloch, G., and Vidal, R. (2011). A continuous optimization framework for hybrid system identification. *Automatica*, 47(3), 608–613.

Le, V.L., Bloch, G., and Lauer, F. (2011). Reduced-size kernel models for nonlinear hybrid system identification. *IEEE Transactions on Neural Networks*, 22(12), 2398–2405.

Le, V.L., Lauer, F., Bako, L., and Bloch, G. (2013). Learning nonlinear hybrid systems: from sparse optimization to support vector regression. In *Proceedings of the* 16^{th} *International Conference on Hybrid systems: computation and control*, 33–42.

Paoletti, S., Juloski, A.L., Ferrari-Trecate, G., and Vidal, R. (2007). Identification of hybrid systems a tutorial. *European Journal of Control*, 13(2–3), 242–260.