Structural risk minimization for switched system identification

Louis Massucci, Fabien Lauer, Marion Gilson

Structural Risk Minimization for Switched System Identification

Louis Massucci, Fabien Lauer, Marion Gilson

***Abstract*— This paper deals with the identification of hybrid dynamical systems that switch arbitrarily between modes. In particular, we focus on the critical issue of estimating the number of modes. A novel method inspired by model selection techniques in statistical learning is proposed. Specifically, the method implements the structural risk minimization principle, which relies on the minimization of an upper bound on the expected prediction error of the model. This so-called generalization error bound is first derived for static switched systems using Rademacher complexities. Then, it is extended to handle non independent observations from a single trajectory of a dynamical system. Finally, it is further tailored to the needs of model selection via a uniformization step. An illustrative example of the behavior of the method and its ability to recover the true number of modes is presented.**

I. INTRODUCTION

Hybrid systems are dynamical systems that include both interacting continuous and discrete dynamical behaviors. This results in systems that switch, according to the value of the discrete variables, between different operating modes with continuous dynamics. This paper focuses on switched linear systems, for which the continuous dynamics are linear and the switching mechanism is arbitrary, i.e., the discrete variable arbitrarily triggering the switch from a mode to an other is an unobserved external input. The identification of such systems, i.e., their estimation from input-output data, is a complex problem whenever the active mode associated to every data point is unknown [1]. A number of efficient methods have been devised over the last 15 years, but a critical issue remains: the estimation of the number of modes. Indeed, as reviewed in [2], switched system identification methods can be classified into two groups: those that work with a fixed number of modes (such as [3], [4], [5]), and those that estimate the minimal number of modes satisfying a predefined threshold on the error (such as [6], [7], [8] or [9] which uses a penalized minimization form of this principle rather than a strict constraint). Therefore, in all cases, the methods require the value of a hyperparameter that directly implies the number of modes for the model.

In this paper, we propose to tackle the model selection problem of estimating the number of modes with a structural risk minimization (SRM) approach inspired from statistical learning [10]. This approach relies on the minimization of an upper bound on the expected error of the model, called the risk or generalization error. Much of learning theory is devoted to the derivation of such generalization error bounds, however most often with an independence assumption on the data, which is not compatible with the system identification context where data points come from a single (or a few) system trajectory. Alternatively, most works related to error bounds for dependent data are based on measures of dependence called mixing coefficients [11] and the independent block sequence construction due to [12]. These ideas were exploited to produce non-asymptotic error bounds for system identification in [13], [14]. In line with more recent work in learning theory, [15] derived error bounds for dependent data based on Rademacher complexities.

In this paper, we combine the general framework of [15] with recent results on Rademacher complexities of static switched models [16] to derive the first generalization error bound for switched system identification. Then, we derive an SRM approach to estimate the number of modes on the basis of this bound. This requires to produce another *uniform* error bound suitable for the practical setting of model selection for most switched system identification algorithms. To the best of our knowledge, this is the first time that such statistical learning techniques are applied in this switched system identification context.

Paper organization: Section II introduces the hybrid system framework and formally exposes the identification problem. Section III gives the necessary background on learning theory and derives the error bound for switched systems. Then, Section IV details the proposed model selection approach, which is illustrated on a numerical example in Section V. Section VI concludes the paper and discusses open issues.

Notation: For any positive integer n, $[n]$ denotes the set of integers from 1 to n. An upright bold letter with a subscript, e.g., $\mathbf{t}_n$, denotes a sequence $(t_i)_{1\leq i\leq n}$ of length n, which should not be confused with the vector $\boldsymbol{t}_i$ of index i. Capital letters are used for random variables. Given two sets $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{Y}^{\mathcal{X}}$ denotes the set of functions from $\mathcal{X}$ to $\mathcal{Y}$.

II. GENERAL FRAMEWORK

In this paper, we focus on discrete-time Single Input Single Output (SISO) Autoregressive with external input (ARX) hybrid systems of the form

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\theta}_{q_i} + e_i \tag{1}$$

with $y_i \in \mathbb{R}$ the output, $\boldsymbol{x}_i \in \mathbb{R}^d$ the regression vector, $q_i \in \{1, \dots, C\}$ the discrete state or mode, C the number of modes, $\boldsymbol{\theta}_j$ with $j = 1, \dots, C$ the parameter vector of the jth mode and $e_i \in \mathbb{R}$ a noise term. The regressor $\boldsymbol{x}_i \in \mathbb{R}^d$,

LM is with the Université de Lorraine, CNRS, LORIA, CRAN, F-54000 Nancy, France louis.massucci@univ-lorraine.fr

FL is with the Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France fabien.lauer@loria.fr

MG is with the Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France marion.gilson@univ-lorraine.fr

$d = n_a + n_b$, with the model orders n_a and n_b, is given by:

$$\boldsymbol{x}_i = [-y_{i-1},\ \ldots, -y_{i-n_a},\ u_{i-1}, \ldots, u_{i-n_b}]^T, \quad (2)$$

where the u_{i-k}'s denote the delayed inputs.

The goal of switched system identification is to find a model $\boldsymbol{f} = \{f_j\}_{j=1}^C$ made of C component submodels f_j that estimate the continuous behaviors of the system in the different modes. The problem can be set as follows.

Problem 1. Given a data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a set of possible submodels $\mathcal{F}$, estimate the number of submodels C, the submodels $\boldsymbol{f} = \{f_j\}_{j=1}^C$, $\boldsymbol{f} \in \mathcal{F}^C$, with $\mathcal{F}^C = \mathcal{F} \times \mathcal{F} \times \cdots \times \mathcal{F}$ and the switching sequence $\mathbf{q} = (q_i)_{1 \le i \le n} \in [C]^n$.

In this paper, we concentrate on the model selection subproblem of determining the number C of submodels. More precisely, we consider the general scheme in Algorithm 1, in which we assume that we have access to a generic switched system identification algorithm working with a fixed C to estimate $\boldsymbol{f} \in \mathcal{F}^C$ and $\mathbf{q} \in [C]^n$. In this scheme, the generic algorithm is applied for all number of modes C within a predefined range. Then, the "best" model is selected on the basis of a criterion $J(C)$.[1] Here, we develop a model selection method based on a criterion $J(C)$ inspired by the SRM principle in statistical learning. This criterion, derived below, will basically take the form of a generalization error bound.

Algorithm 1 General model selection scheme

Require: The data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a maximum number of modes $\overline{C}$
  for $C = 1$ to $\overline{C}$ **do**
    Run the generic algorithm to estimate a model $\boldsymbol{f}$ with C modes
    Compute a criterion $J(C)$
  end for
  Select the "best" number of modes

$$\hat{C} = \operatorname*{argmin}_{C \in [\overline{C}]} J(C)$$

  return the selected model with $\hat{C}$ modes

III. ERROR BOUNDS FOR SWITCHED SYSTEM IDENTIFICATION

A short introduction to learning theory and regression error bounds is presented in Sect. III-A, before the exposition of error bounds dedicated to static switched systems in Sect. III-B. Then, we discuss the non-independent case in Sect. III-C and finally derive a bound applicable to switched dynamical systems in Sect. III-D.

A. Preliminaries

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the output space with $\mathcal{Y} = [-M, M]$ for some $M > 0$. A relationship between inputs and outputs is characterized by an unknown probability distribution of a random pair $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$, of probability density function $p(x, y)$. Given a realization of a sample $\mathbf{Z}_n = (Z_i)_{1 \le i \le n} = ((X_i, Y_i))_{1 \le i \le n}$ of n independent copies of Z, the aim of regression is to learn the model $f : \mathcal{X} \to \mathcal{Y}$ that minimizes, over a certain model class $\mathcal{F}$, the generalization error (or risk)

$$L(f) = \mathbb{E}_{X,Y} \ell(f, X, Y), \quad (3)$$

defined as the expected value

$$\mathbb{E}_{X,Y} \ell(f, X, Y) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(f, x, y) p(x, y)\, dxdy \quad (4)$$

of the loss function $\ell(f, X, Y)$. This loss function measures the pointwise error in predicting $f(x)$ instead of y and is typically of the form $\ell(f, X, Y) = |y - f(x)|^p$ with $p \in \{1, 2\}$.

As the probability distribution of Z is unknown, we cannot compute the expected value and many methods minimize instead the empirical risk

$$\hat{L}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, X_i, Y_i), \quad (5)$$

for a realization of $Z_n = \mathbf{z}_n$. So-called generalization error bounds are upper bounds on the risk (3) that typically involve the empirical risk (5) and hold in the non-asymptotic case with high probability and uniformly over the model class $\mathcal{F}$, i.e., bounds of the form:

$$\mathbb{P}\left\{\forall f \in \mathcal{F},\ L(f) \le \hat{L}(f) + \epsilon(n, \mathcal{F}, \delta)\right\} \ge 1 - \delta. \quad (6)$$

A large part of learning theory is devoted to the characterization of the tightest confidence interval $\epsilon(n, \mathcal{F}, \delta)$ in (6), which typically depends on the capacity of the model class $\mathcal{F}$. More precisely, the capacity of the loss class

$$\mathcal{L} = \{\ell(z) = \ell(f, x, y),\ f \in \mathcal{F}\} \quad (7)$$

must be considered, and can be measured for instance with the Rademacher complexity, as initiated by [17], [18].

Definition 1 (Rademacher complexity)**.** Given a sequence $\mathbf{Z}_n = (Z_i)_{1 \le i \le n}$ of random variables $Z_i \in \mathcal{Z}$, the *empirical Rademacher complexity* of a class $\mathcal{L}$ of functions from $\mathcal{Z}$ to $\mathbb{R}$ is defined as

$$\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{\ell \in \mathcal{L}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(Z_i) \,\middle|\, \mathbf{Z}_n\right], \quad (8)$$

where $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \le i \le n}$ is a sequence of Rademacher variables, i.e., random variables uniformly distributed in $\{-1, +1\}$. The *Rademacher complexity* of $\mathcal{L}$ is

$$\mathcal{R}_n(\mathcal{L}) = \mathbb{E}_{\mathbf{Z}_n} \hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}). \quad (9)$$

A general bound based on the Rademacher complexity is the following:

Theorem 1 (Theorem 1 in [19])**.** *Let $\mathcal{L}$ be a class of functions from $\mathcal{Z}$ into $[0, B]$ and $\mathbf{Z}_n = (Z_i)_{1 \le i \le n}$ be a sequence of independent copies of the random variable*

[1] $J(C)$ also depends on the estimated model $\boldsymbol{f} \in \mathcal{F}^C$, but we keep the notation short to put the emphasis on the number of modes C.

$Z \in \mathcal{Z}$. Then, for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over all $\ell \in \mathcal{L}$,

$$\mathbb{E}_Z \ell(Z) \leq \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i) + 2\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}) + 3B\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \quad (10)$$

To derive risk bounds from Theorem 1, the loss function must be bounded. Given a bounded output $Y \in [-M, M]$, this can be obtained by clipping the model.

Definition 2 (Clipping)**.** For any $M > 0$ and $t \in \mathbb{R}$, we define the clipped version $\bar{t}$ of t as

$$\bar{t} = \begin{cases} -M, & \text{if } t < -M \\ t, & \text{if } t \in [-M, M] \\ M, & \text{if } t > M. \end{cases} \quad (11)$$

The clipped version $\overline{f}$ of a function $f : \mathcal{X} \to \mathbb{R}$ is obtained by clipping its output: $\forall x \in \mathcal{X}$, $\overline{f}(x) = \overline{f(x)}$. And $\overline{\mathcal{F}}$ denotes the clipped function class $\{\overline{f} : \ f \in \mathcal{F}\}$.

Indeed, since for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(f, x, y) \geq \ell(\overline{f}, x, y)$, the risk (3) of the clipped model, $L(\overline{f})$, is always smaller than $L(f)$ and we can consider that the final result of the estimation is $\overline{f}$ instead of f and derive upper bounds on $L(\overline{f})$.

Example 1 (Linear regression)**.** For linear regression in $\mathcal{X} = \mathbb{R}^d$, we can consider the model class

$$\mathcal{F} = \{f : f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}, \ \boldsymbol{w} \in \mathbb{R}^d, \ \|\boldsymbol{w}\| \leq R_w\} \quad (12)$$

and the loss class (induced by the clipped model)

$$\mathcal{L} = \{\ell \in [0, 4M^2]^{\mathcal{Z}} : \ell(z) = (y - \overline{f}(x))^2, \ f \in \mathcal{F}\}.$$

Using a contraction argument (see [19]), the Rademacher complexity of $\mathcal{L}$ can be related to that of $\mathcal{F}$, with $\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}) \leq 4M\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F})$, which in turn can be bounded using standard computations for Rademacher complexities [18] as

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) \leq \frac{R_w\sqrt{\sum_{i=1}^{n} \|\boldsymbol{X}_i\|^2}}{n}. \quad (13)$$

Thus, Theorem 1 leads to the following error bound for linear regression: for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$ as in (12),

$$L(\overline{f}) \leq \hat{L}(\overline{f}) + \frac{8MR_w\sqrt{\sum_{i=1}^{n} \|\boldsymbol{X}_i\|^2}}{n} + 12M^2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \quad (14)$$

B. Switching Regression Bounds

When estimating arbitrarily switching systems (in the static context of switching regression or for hybrid dynamical system identification), the selection of the active submodel among the C submodels f_j is embedded in the loss function:

$$\ell(\boldsymbol{f}, x, y) = \min_{j \in [C]} |y - f_j(x)|^p \quad (15)$$

with $p \in \{1, 2\}$. By minimizing the error thus defined, we ask that at least one of the submodels $f_j \in \mathcal{F}$ accurately approximates the output y.

Error bounds for switching regression in the static case were obtained in [16] through the decomposition of the Rademacher complexity of the loss class based on (15) in terms of the one of $\mathcal{F}$:

$$\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}) \leq p(2M)^{p-1} C\, \hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}). \quad (16)$$

Example 2 (Switching linear regression)**.** For switching linear regression with $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{F}$ as in (12) and the loss (15) with $p = 2$, Theorem 1 and (16) combined with (13) guarantee that, for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $\boldsymbol{f} \in \mathcal{F}^C$,

$$L(\overline{\boldsymbol{f}}) \leq \hat{L}(\overline{\boldsymbol{f}}) + \frac{8MCR_w\sqrt{\sum_{i=1}^{n} \|\boldsymbol{X}_i\|^2}}{n} + 12M^2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Note that, due to the independence assumption in Theorem 1, this bound cannot be used directly for switched system identification, where $\boldsymbol{x}_i$ depends on lagged outputs y_{i-k} due to (2).

C. Bounds for Dependent Data

The following assumes that the sample $\mathbf{Z}_n$ is taken from a stationary β-mixing process.

Definition 3 (Stationarity)**.** A sequence of random variables $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$ is said to be stationary if for any t and non-negative integers m and k, the random vectors $(Z_t, \ldots, Z_{t+m})$ and $(Z_{t+k}, \ldots, Z_{t+k+m})$ have the same distribution.

The time index t does not affect the distribution of a variable Z_t in a stationary sequence.

Definition 4 (β-mixing)**.** Let $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{\infty}$ be a stationary sequence of random variables. For any $i, j \in \mathbb{Z} \cup \{-\infty, \infty\}$, let σ_i^j denote the σ-algebra generated by the random variables Z_k, $i \leq k \leq j$. Then, for any positive integer k, the β-mixing coefficient of the stochastic process $\mathbf{Z}$ is defined as

$$\beta(k) = \mathbb{E}_{B \in \sigma_{-\infty}^{0}} \left\{ \sup_{A \in \sigma_k^{\infty}} |\mathbb{P}[A|B] - \mathbb{P}[A]| \right\}. \quad (17)$$

If $\beta(k) \to 0$ as $k \to \infty$, then $\mathbf{Z}$ is said to be β-mixing.

For a sequence of independent variables, $\mathbb{P}[A|B] = \mathbb{P}[A]$ in (17) and $\beta(k) = 0$ for all $k \geq 1$. For β-mixing processes, the dependence between two events separated by k time steps weakens as a function of k. For more details on mixing processes, refer to [11]. In this paper, we simply rely on Definition 4 to characterize the degree of dependence between data.

Rademacher complexity bounds for dependent data were obtained in [15] by applying the independent block sequence construction of [12]. In short, the idea is to see the whole data set as blocks of length a and by considering only a subset of μ blocks instead of the set of n data points (see Figure 1). By doing so, for β-mixing processes, concentration inequalities can be applied to a sequence of independent blocks distributed as those in the odd subset while controlling the

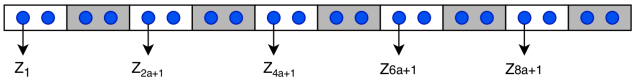

Fig. 1. Illustration of the block sequence with $n = 20$, $a = 2$ data points per block and $\mu = 5$ odd blocks (white) and μ even blocks (grey). The sample $\mathbf{Z}_\mu = (Z_1, Z_{2a+1}, Z_{4a+1}, \ldots, Z_{8a+1})$ contains the first point of each odd block. The dependence between the odd blocks (and the points of $\mathbf{Z}_\mu$) decreases with the length a of the even blocks separating them.

error thus introduced. This leads to the following adaptation of Theorem 1.

Theorem 2 (Theorem 2 in [15])**.** *Let $\mathcal{L}$ be a class of functions from $\mathcal{Z}$ into $[0, B]$ and $\mathbf{Z}_n = (Z_i)_{1\leq i\leq n}$ be a sequence drawn from a stationary β-mixing distribution. For any $\mu, a > 0$ with $2\mu a = n$ and $\delta > 4(\mu-1)\beta(a)$, with probability at least $1-\delta$, uniformly over all $\ell \in \mathcal{L}$,*

$$\mathbb{E}\ell(Z_1) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(Z_i) + 2\hat{\mathcal{R}}_{\mathbf{Z}_\mu}(\mathcal{L}) + 3B\sqrt{\frac{\log\frac{4}{\delta'}}{2\mu}}, \quad (18)$$

where $\delta' = \delta - 4(\mu-1)\beta(a)$ and $\mathbf{Z}_\mu = (Z_{2a(i-1)+1})_{1\leq i\leq\mu}$ is a sample of length μ as in Fig. 1.

Compared with Theorem 1, Theorem 2 shows a confidence interval that decreases with the "effective number of data" $\mu = n/2a$ instead of the original number n. In addition, the confidence interval also slightly increases due to the use of δ' instead of δ.

D. Bounds for Switched System Identification

We now have all the ingredients needed to derive a new generalization error bound for switched system identification.

Theorem 3. *Let $\mathcal{F}^C$ be a vector valued function class with C components from $\mathcal{F}$ as in* (12) *and the loss ℓ be as in* (15)*. Then, for any sample $\mathbf{Z}_n = ((\boldsymbol{X}_i, Y_i))_{1\leq i\leq n} \in (\mathbb{R}^d \times \mathcal{Y})^n$ drawn from a stationary β-mixing distribution, and for any $\mu, a > 0$ with $2\mu a = n$ and $\delta > 4(\mu-1)\beta(a)$, with probability at least $1-\delta$, the following inequality holds for all $\boldsymbol{f} \in \mathcal{F}^C$:*

$$L(\overline{\boldsymbol{f}}) \leq \hat{L}(\overline{\boldsymbol{f}}) + \frac{2p(2M)^{p-1}CR_w\sqrt{\sum_{i=1}^{\mu}\|\boldsymbol{X}_{2a(i-1)+1}\|^2}}{\mu} + 3(2M)^p\sqrt{\frac{\log\frac{4}{\delta'}}{2\mu}},$$

where $\delta' = \delta - 4(\mu-1)\beta(a)$.

Proof: Apply Theorem 2 and bound the Rademacher complexity of the loss class based on (15) with (16) and (13).

IV. MODEL SELECTION

We now turn to the model selection issue of estimating the number of modes, for which we derive an SRM approach. The SRM principle consists in tuning the hyperparameters by minimizing a generalization error bound, which in our case is given by Theorem 3. However, before we can proceed with the practical implementation in Sect. IV-B, this bound will be tailored to the specific needs of model selection in Sect. IV-A.

A. Uniform Bounds

Since the general scheme in Algorithm 1 requires to compute the bound for all $C \in [\overline{C}]$ with the same data set, we need a bound that holds uniformly over all C. In addition, the bound in Theorem 3 holds with a predefined radius R_w for the model class (12), whereas most practical algorithms for switched system identification do not impose constraints on the parameter vectors $\boldsymbol{w}_j$. To fill this gap between theory and practice, we would need a bound in which the radius R_w could be computed *a posteriori* as

$$R_w = \max_{j\in[C]}\|\boldsymbol{w}_j\| \quad (19)$$

from the estimated $\boldsymbol{w}_j$'s. This can usually be obtained via a uniformization over a well chosen discretization, as, e.g., in Appendix F of [16]. However, here, additional terms in Theorem 3 due to the dependence of the data breaks the possibility to apply the same technique and we have to rely on a predefined grid of K discretized values,

$$\mathcal{G} = \{R_1, \ldots, R_K\},$$

with $R_K = R_{max}$. Then, we can derive a bound in which the radius R_w is replaced by its discretized value

$$\tilde{R}_w = \min_{k\in[K]} R_k, \quad \text{s.t } R_k \geq R_w, \quad \text{with } R_w \text{ as in (19).} \quad (20)$$

Theorem 4 (Uniform Bound)**.** *Given a constant R_{max}, let $\mathcal{F}$ be as in* (12) *with R_{max} instead of R_w. Then for a sample $\mathbf{Z}_n$ of size n drawn from a stationary β-mixing distribution, and for any $\mu, a > 0$ with $2\mu a = n$, and $\delta > 4\overline{C}K(\mu-1)\beta(a)$, with probability at least $1-\delta$:*

$$\forall C \in [\overline{C}],\ \forall \boldsymbol{f} \in \mathcal{F}^C,$$
$$L(\overline{\boldsymbol{f}}) \leq \hat{L}(\overline{\boldsymbol{f}}) + \frac{2p(2M)^{p-1}C\tilde{R}_w\sqrt{\sum_{i=1}^{\mu}\|\boldsymbol{X}_{2a(i-1)+1}\|^2}}{\mu} + 3(2M)^p\sqrt{\frac{\log(\overline{C}K) + \log\frac{4}{\delta'}}{2\mu}}, \quad (21)$$

with $\delta' = \delta - 4\overline{C}K(\mu-1)\beta(a)$ and $\tilde{R}_w$ as in (20)*.*

Note that the cost of the uniformization is merely a $\log(\overline{C}K)$ term within the second square root.

Proof: Let $\mathcal{F}_k$ be defined as in (12) with R_k instead of R_w. For any fixed C, R_k and $\delta_0' > 0$, let $\delta_0 = \delta_0' + 4(\mu-1)\beta(a)$ and

$$\epsilon(C, R_k, \delta_0') = \frac{2p(2M)^{p-1}CR_k\sqrt{\sum_{i=1}^{\mu}\|\boldsymbol{X}_{2a(i-1)+1}\|^2}}{\mu} + 3(2M)^p\sqrt{\frac{\log\frac{4}{\delta_0'}}{2\mu}}.$$

Then, Theorem 3 gives

$$\mathbb{P}\left\{\exists \boldsymbol{f} \in \mathcal{F}_k^C,\ L(\overline{\boldsymbol{f}}) > \hat{L}(\overline{\boldsymbol{f}}) + \epsilon(C, R_k, \delta_0')\right\} \leq \delta_0.$$

Thus, by the union bound and $\overline{C}K$ applications of Theorem 3 with confidence δ_0, we have

$$\begin{aligned}
&\mathbb{P}\left\{\exists C \in [\overline{C}], k \in [K], \boldsymbol{f} \in \mathcal{F}_k^C, L(\overline{\boldsymbol{f}}) \geq \hat{L}(\overline{\boldsymbol{f}}) + \epsilon(C, R_k, \delta_0')\right\} \\
&\leq \sum_{C=1}^{\overline{C}} \sum_{k=1}^{K} \mathbb{P}\left\{\exists \boldsymbol{f} \in \mathcal{F}_k^C,\ L(\overline{\boldsymbol{f}}) \geq \hat{L}(\overline{\boldsymbol{f}}) + \epsilon(C, R_k, \delta_0')\right\} \\
&\leq \overline{C}K\delta_0.
\end{aligned}$$

Then, for any $\delta = \delta' + 4\overline{C}K(\mu - 1)\beta(a)$ with $\delta' > 0$, set $\delta_0' = \delta'/\overline{C}K$. This leads to $\delta_0 = \delta/\overline{C}K$ and thus

$$\begin{aligned}
&\mathbb{P}\left\{\forall C \in [\overline{C}], k \in [K], \boldsymbol{f} \in \mathcal{F}_k^C, L(\overline{\boldsymbol{f}}) \leq \hat{L}(\overline{\boldsymbol{f}}) + \epsilon(C, R_k, \delta_0')\right\} \\
&= 1 - \mathbb{P}\left\{\exists C \in [\overline{C}], k \in [K], \boldsymbol{f} \in \mathcal{F}_k^C,\right. \\
&\qquad \left. L(\overline{\boldsymbol{f}}) > \hat{L}(\overline{\boldsymbol{f}}) + \epsilon(C, R_k, \delta_0')\right\} \\
&\geq 1 - \overline{C}K\delta_0 = 1 - \delta
\end{aligned} \tag{22}$$

with

$$\begin{aligned}
\epsilon(C, R_k, \delta_0') = &\frac{2p(2M)^{p-1} C R_k \sqrt{\sum_{i=1}^{\mu} \|\boldsymbol{X}_{2a(i-1)+1}\|^2}}{\mu} \\
&+ 3(2M)^p \sqrt{\frac{\log(\overline{C}K) + \log\frac{4}{\delta'}}{2\mu}}.
\end{aligned}$$

For any $C \in [\overline{C}]$, we can rewrite $\mathcal{F}^C$ with radius $R_{\max}$ as $\mathcal{F}^C = \bigcup_{k \in [K]} \mathcal{F}_k^C$; and any $\boldsymbol{f} \in \mathcal{F}^C$ also belongs to $\mathcal{F}_k^C$ with, by (20), k such that $R_k = \tilde{R}_w$. Thus, (22) is equivalent to (21) and the statement is proved.

B. Proposed Method

The proposed model selection method for switched system identification is now detailed. In Sect. IV-A, we derived a uniform generalization error bound that holds uniformly over any number of modes $C \leq \overline{C}$. Thus, a criterion $J(C)$ can be devised on the basis of this bound to work in the general model selection framework of Algorithm 1. More precisely, the structural risk minimization principle amounts in this case to selecting the "best" number of modes $\hat{C}$ as the one that minimizes the bound (21) on the generalization error. By leaving aside constant terms that do not depend on C nor on the estimated model $\boldsymbol{f}$ in the bound, the model selection procedure can be written as

$$\hat{C} = \underset{C \in [\overline{C}]}{\operatorname{argmin}}\, J(C), \tag{23}$$

where

$$J(C) = \hat{L}(\overline{\boldsymbol{f}}) + \epsilon(\boldsymbol{f}, C), \tag{24}$$

$$\epsilon(\boldsymbol{f}, C) = \frac{2p(2M)^{p-1} C \tilde{R}_w \sqrt{\sum_{i=1}^{\mu} \|\boldsymbol{x}_{2a(i-1)+1}\|^2}}{\mu}$$

with a and μ such that $2\mu a = n$. Note in particular that the precise value of $\beta(a)$ in the bound of Theorem 3 does not influence the model selection procedure and the values of $J(C)$.

V. NUMERICAL EXAMPLE

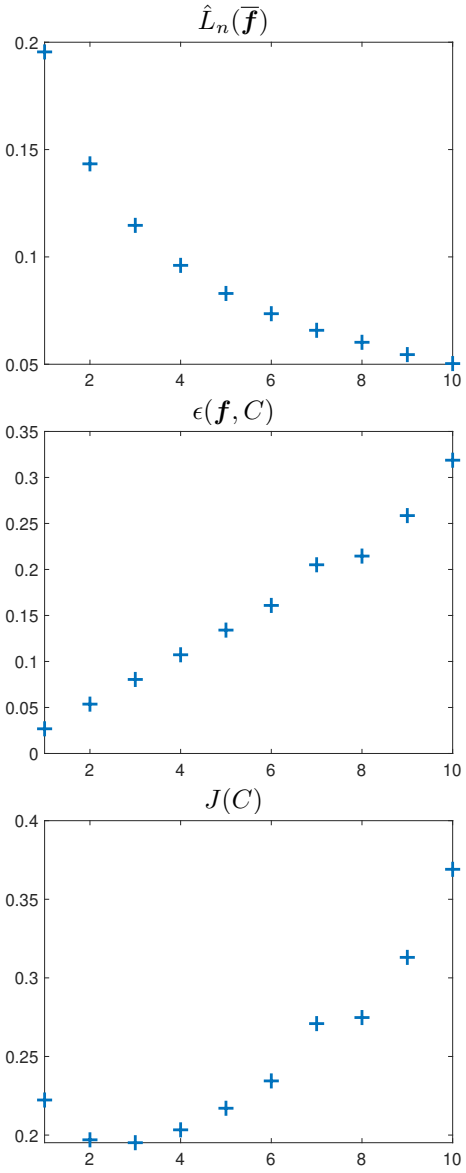

Fig. 2. *Top:* the empirical error $\hat{L}_n(\overline{\boldsymbol{f}})$ decreases as a function of the number C of modes. *Middle:* The second term in $J(C)$ (24), i.e., $\epsilon(\boldsymbol{f}, C)$, increases with C. *Bottom:* the sum of the two terms, i.e., $J(C)$, has a minimum at $\hat{C} = 3$ which coincides with the true number of modes.

In this section, we illustrate the method proposed in Sect. IV on an example taken from [7]: the identification of a switched system composed of $C = 3$ linear subsystems of orders $n_a = n_b = 2$ with parameter vectors

$$\begin{aligned}
\theta_1 &= [-0.4 \quad 0.25 \quad -0.15 \quad 0.08]^T, \\
\theta_2 &= [1.55 \quad -0.58 \quad -2.1 \quad 0.96]^T, \\
\theta_3 &= [1 \quad -0.24 \quad -0.65 \quad 0.30]^T.
\end{aligned} \tag{25}$$

This system is used to generate a data set of $n = 400\,000$ points with (1)–(2) under the following conditions. The excitation input u_i is a zero-mean Gaussian signal of unit variance. The noise e_i is a white Gaussian noise whose magnitude is such that the Signal to Noise Ratio (SNR) is

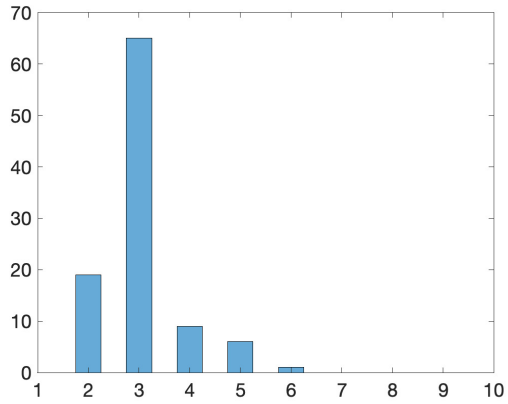

Fig. 3. Histogram of the estimated number of modes $\hat{C}$ over 100 trials with true $C = 3$.

equal to 10 dB with respect to the output signal. The active mode q_i is uniformly distributed in $\{1, 2, 3\}$.

Algorithm 1 is applied with $J(C)$ computed as in (24) for $p = 1$, $a = 2$ and $\overline{C} = 10$. The K-LinReg algorithm [5], which remains particularly efficient on large data sets with a satisfactory accuracy, is used to estimate a model $\boldsymbol{f}$ with C modes at each iteration. Figure 2 illustrates the behavior of Algorithm 1 and $J(C)$ together with its two constituents: the empirical error $\hat{L}_n(\overline{\boldsymbol{f}})$ and the term $\epsilon(\boldsymbol{f}, C)$ inherited from the bound (21). On the one hand, as submodels are added to $\boldsymbol{f}$ and the number C of modes is increased, the error naturally decreases (top plot). On the other hand, the model complexity increases with C, which implies an increase of $\epsilon(\boldsymbol{f}, C)$ (middle plot). Finally, the minimum of $J(C)$ at $\hat{C} = 3$ (bottom plot) offers the optimal trade-off between the two terms and leads to the correct estimation of the number of modes.

To evaluate the robustness of the method over several data sets, another set of experiments is conducted. One hundred new data sets are generated under the same conditions as before, but with colored noise $\tilde{e}_i$ such that $\tilde{e}_i - 0.35\tilde{e}_{i-1} = e_i + 0.5e_{i-1}$ and a signal-to-noise ratio of 5dB. Figure 3 shows the histogram of the estimated number of modes $\hat{C}$. As can be seen from these results, the method correctly estimates the number of modes (with $\hat{C} = 3$) in 65% of the trials and remains close to the true value otherwise, even in this situation with a large and colored noise.

VI. CONCLUSIONS

This paper investigated a model selection method inspired by the structural risk minimization principle from statistical learning to estimate the number of modes in switched system identification. For this method, new generalization error bounds for switched dynamical systems were derived.

Preliminary results showed that the proposed approach can recover the true number of modes in the considered experimental setting. However, the error bounds on which our model selection method is based inherently depend on the number of data. Currently, the method is effective only when this number is sufficiently large in order to make the confidence interval of the bound of the same order of magnitude as the empirical risk $L(\overline{\boldsymbol{f}})$ and correctly balance the two terms. This is in part due to the dependence in the data, which degrades the confidence interval through the use of the number of blocks $\mu = n/2a$ instead of the number of data n. Tightening the bound thus appears as an important issue to make the method applicable to a broader range of problems, especially for the case $p = 2$.

Future work will also concentrate on the characterization of the β-mixing coefficient for switched systems, for instance by extending results available for linear systems [13], [14]. Though the value of this coefficient does not directly influence the proposed model selection strategy, its computation would be needed in practice to apply our generalization bounds as guarantees on the prediction error.

REFERENCES

[1] F. Lauer. On the complexity of switching linear regression. *Automatica*, 74:80–83, 2016.

[2] F. Lauer and G. Bloch. *Hybrid system identification: Theory and algorithms for learning switching models*. Springer, 2019.

[3] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proc. of the 42nd IEEE Conference on Decision and Control (CDC), Maui, HI, USA*, pages 167–172, 2003.

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

[5] F. Lauer. Estimating the probability of success of a simple algorithm for switched linear regression. *Nonlinear Analysis: Hybrid Systems*, 8:31–47, 2013.

[6] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control*, 50(10):1567–1580, 2005.

[7] L. Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47(4):668–677, 2011.

[8] N. Ozay, M. Sznaier, C.M. Lagoa, and O.I. Camps. A sparsification approach to set membership identification of switched affine systems. *IEEE Transactions on Automatic Control*, 57(3):634–648, 2012.

[9] H. Ohlsson and L. Ljung. Identification of switched linear regression models using sum-of-norms regularization. *Automatica*, 49(4):1045–1050, 2013.

[10] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[11] R.C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys*, 2:107–144, 2005.

[12] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

[13] E. Weyer. Finite sample properties of system identification of arx models under mixing conditions. *Automatica*, 36(9):1291–1299, 2000.

[14] M. Vidyasagar and R. L. Karandikar. A learning theory approach to system identification and stochastic adaptive control. *Journal of Process Control*, 18(3):421 – 430, 2008.

[15] M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems 21*, pages 1097–1104, 2009.

[16] Fabien Lauer. Error bounds for piecewise smooth and switching regression. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1183–1195, 2019.

[17] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[18] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[19] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, second edition, 2018.