Regularized switched system identification: A statistical learning perspective
Louis Massucci, Fabien Lauer, Marion Gilson

To cite this version:
Louis Massucci, Fabien Lauer, Marion Gilson. Regularized switched system identification: A statistical learning perspective. 7th IFAC International Conference on Analysis and Design of Hybrid Systems, ADHS’2021, Jul 2021, Brussel (Virtual Event), Belgium. hal-03324120

HAL Id: hal-03324120
https://hal.science/hal-03324120
Submitted on 23 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Regularized Switched System Identification: a Statistical Learning Perspective

Louis Massucci *,** Fabien Lauer * Marion Gilson **

* Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
** Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France

Abstract: Switched system identification is a challenging problem, for which many methods were proposed over the last twenty years. Despite this effort, estimating the number of modes of switched systems from input–output data remains a nontrivial and critical issue for most of these methods. This paper discusses a recently proposed statistical learning approach to deal with this issue and proposes to go one step further by considering new results dedicated to regularized models. Optimization algorithms devised to tackle the estimation of such models from data are also proposed and illustrated in a few numerical experiments.

Keywords: Switched systems, System identification, Statistical learning, Model selection.

1. INTRODUCTION

There are a variety of formalisms to describe hybrid dynamical systems of various classes that switch between different subsystems with one mechanism or another. For the purpose of system identification, i.e., the estimation of a model from input–output data, two classes are most often considered: arbitrarily switched systems and piecewise smooth systems. This paper focuses solely on the class of discrete-time switched linear systems and their identification with an autoregressive with exogenous input (ARX) model form. Despite this focus on a rather restricted subclass of hybrid systems, the identification problem becomes highly complex as soon as the switchings are not observed. In this case, not only one has to estimate a submodel for each subsystem, but also to simultaneously assign the data points to the different subsystems, i.e., estimate the active mode at each time step. Formally, this leads to an optimization problem of a combinatorial nature, recognized by the presence of either integer variables or nonconvex and nondifferentiable terms.

While a number of approaches were proposed in the literature to deal with these difficulties (see Lauer and Bloch (2019) for an overview), the additional issue of estimating the number of modes or subsystems remains a critical one, for which very few solutions exist. Many methods estimate the number of modes by relying on a predefined and difficult to tune error tolerance for the model, adding submodels until this tolerance is satisfied (Bemporad et al., 2005; Bako, 2011; Ozay et al., 2012). The seminal work of Vidal et al. (2003) offers an appealing approach based on algebraic arguments, but is applicable often in noiseless (or almost noiseless) conditions, while its extension to noisy data (Ozay et al., 2015) assumes a bounded (typically uniform) noise. A novel approach recently proposed by Massucci et al. (2020) makes use of tools from statistical learning theory (Vapnik, 1998). It amounts to searching for the number of modes that minimizes a probabilistic upper bound on the prediction error of the model, hence offering the best statistical guarantees in terms of prediction accuracy.

In this paper, we propose to further investigate the approach based on statistical learning by considering the benefits of regularization, which is a standard scheme to control the model complexity when learning from data. In particular, we aim at leveraging the new results of Lauer (2020a), which tighten the statistical guarantees for switching models by considering a finer analysis of their complexity. Doing so, we develop new regularization algorithms in Sect. 3 to learn switching models whose complexity is controlled in accordance with the most recent theoretical advances. Then, in Section 4, we derive the statistical guarantees and use them in a model selection framework to estimate the number of modes. The resulting methods are illustrated with a few experiments in Sect. 5 and open issues are discussed in Sect. 6.

Notation. Vectors are written in bold lowercase letters, whereas matrices are written in bold uppercase and upright letters. Random variables are written in italic uppercase letters. Thus, \( X \) is a random vector whose values are denoted by \( x \), whereas \( X \) is a matrix. For any integer \( N > 0 \), \( [N] = \{1, \ldots, N\} \) refers to the set of integers from 1 to \( N \). For \( p \in \{1, 2, \infty\} \), the \( \ell_p \)-norm of a vector \( a \in \mathbb{R}^n \) is denoted by \( \|a\|_p \) and defined as \( \|a\|_p = (\sum_{i=1}^n |a_i|^p)^{1/p} \), except for \( p = \infty \) : \( \|a\|_{\infty} = \max_{i \in [n] } |a_i| \).

2. PROBLEM STATEMENT

Switched linear system identification amounts to estimating the parameter vectors \( \mathbf{w}_j \in \mathbb{R}^d \) of \( C \) linear submodels in ARX form,

\[
f_j(x) = \mathbf{w}_j^T x
\]

with

\[
x_j = [-y_{j-1}, \ldots, -y_{j-n_a}, u_{j-1}, \ldots, u_{j-n_a}]^T,
\]

where \( j = 1, \ldots, N \).

The purpose of system identification, i.e., the estimation of a model from input–output data, two classes are most often considered: arbitrarily switched systems and piecewise smooth systems. This paper focuses solely on the class of discrete-time switched linear systems and their identification with an autoregressive with exogenous input (ARX) model form. Despite this focus on a rather restricted subclass of hybrid systems, the identification problem becomes highly complex as soon as the switchings are not observed. In this case, not only one has to estimate a submodel for each subsystem, but also to simultaneously assign the data points to the different subsystems, i.e., estimate the active mode at each time step. Formally, this leads to an optimization problem of a combinatorial nature, recognized by the presence of either integer variables or nonconvex and nondifferentiable terms.

While a number of approaches were proposed in the literature to deal with these difficulties (see Lauer and Bloch (2019) for an overview), the additional issue of estimating the number of modes or subsystems remains a critical one, for which very few solutions exist. Many methods estimate the number of modes by relying on a predefined and difficult to tune error tolerance for the model, adding submodels until this tolerance is satisfied (Bemporad et al., 2005; Bako, 2011; Ozay et al., 2012). The seminal work of Vidal et al. (2003) offers an appealing approach based on algebraic arguments, but is applicable often in noiseless (or almost noiseless) conditions, while its extension to noisy data (Ozay et al., 2015) assumes a bounded (typically uniform) noise. A novel approach recently proposed by Massucci et al. (2020) makes use of tools from statistical learning theory (Vapnik, 1998). It amounts to searching for the number of modes that minimizes a probabilistic upper bound on the prediction error of the model, hence offering the best statistical guarantees in terms of prediction accuracy.

In this paper, we propose to further investigate the approach based on statistical learning by considering the benefits of regularization, which is a standard scheme to control the model complexity when learning from data. In particular, we aim at leveraging the new results of Lauer (2020a), which tighten the statistical guarantees for switching models by considering a finer analysis of their complexity. Doing so, we develop new regularization algorithms in Sect. 3 to learn switching models whose complexity is controlled in accordance with the most recent theoretical advances. Then, in Section 4, we derive the statistical guarantees and use them in a model selection framework to estimate the number of modes. The resulting methods are illustrated with a few experiments in Sect. 5 and open issues are discussed in Sect. 6.

Notation. Vectors are written in bold lowercase letters, whereas matrices are written in bold uppercase and upright letters. Random variables are written in italic uppercase letters. Thus, \( X \) is a random vector whose values are denoted by \( x \), whereas \( X \) is a matrix. For any integer \( N > 0 \), \( [N] = \{1, \ldots, N\} \) refers to the set of integers from 1 to \( N \). For \( p \in \{1, 2, \infty\} \), the \( \ell_p \)-norm of a vector \( a \in \mathbb{R}^n \) is denoted by \( \|a\|_p \) and defined as \( \|a\|_p = (\sum_{i=1}^n |a_i|^p)^{1/p} \), except for \( p = \infty \) : \( \|a\|_{\infty} = \max_{i \in [n] } |a_i| \).
from $n$ input-output data gathered in

$$X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d}, \quad y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n.$$ 

Note that the available data does not include the knowledge of the active mode for each data point $(x_i, y_i)$, which must also be estimated. Furthermore, we here focus on the additional issue of estimating the number of modes $C$ together with the submodels themselves.

To solve this problem, we will first develop regularized algorithms that can estimate a switched model with a fixed number of modes. Then, we will use these within a global procedure aiming at the estimation of this number.

Throughout the paper, we use $w_j$ to refer to the $j$th subset of $d$ entries in the vector $w \in \mathbb{R}^{C \times d}$ concatenating the $C$ submodel parameter vectors:

$$w = [w_1^T, \ldots, w_C^T]^T.$$

3. REGULARIZED ALGORITHMS FOR A FIXED NUMBER OF MODES

Regularization is a standard technique to control the model complexity while learning from data. It amounts to learning by minimizing a trade-off between a data-fitting term that accounts for the model error, $\mathcal{E}(w, X, y)$, and a regularization term, $\Gamma(w)$, that penalizes highly complex models:

$$\min_{w \in \mathbb{R}^D} \mathcal{E}(w, X, y) + \lambda \Gamma(w), \quad (2)$$

where we assume that the model can be represented by a vector of parameters $w \in \mathbb{R}^D$ and $\lambda > 0$ is a hyperparameter that controls the trade-off between the two terms. Then, various regularization schemes can be devised by choosing different means to compute the error and the model complexity.

The error is typically computed as the mean (or sum) of the pointwise loss over all training data:

$$\mathcal{E}(w, X, y) = \sum_{i=1}^n \ell(w, x_i, y_i)$$

with a so-called loss function $\ell : \mathbb{R}^D \times \mathbb{R}^d \times \mathbb{R} \rightarrow [0, \infty)$. In switched system identification, this function is of the general form

$$\ell(w, x_i, y_i) = \min_{j \in [C]} |y_i - w_j^T x_i|^p, \quad (3)$$

in which $p \in \{1, 2\}$ selects between the absolute and the squared loss. The minimum operation embedded in the loss accounts for the fact that the only error that matters for a data points $(x_i, y_i)$ is the one computed with respect to the submodel $j_i$ that best approximates $y_i$. Therefore, if $\mathcal{E}(w, X, y)$ can be sufficiently minimized, this ensures that there is at least one submodel that accurately estimates $y_i$ for all points.

Regarding the regularization, in this paper, we focus on regularization terms of the form

$$\Gamma(w) = \|\Omega(w)\|_q, \quad (4)$$

where

$$\Omega(w) = [\|w_1\|_2, \ldots, \|w_C\|_2]^T$$

and $q \in \{1, 2, \infty\}$ selects a particular regularization scheme depending on the desired properties of the model. In general terms, regularization with (4) penalizes complex hybrid models, whose complexity is measured at a global level by the $\ell_q$-norm of a vector $\Omega(w)$, in which each component measures the complexity of one of the submodels via the $\ell_2$-norm of its parameter vector. Hence, the global complexity $\Gamma(w)$ is a function of all submodels.

Overall, and throughout the paper, a particular type of regularized switched system identification problem (2) is characterized by the pair $(p, q)$, in which $p$ and $q$ define the loss and the regularization, respectively. For any such pair, (2) is a highly difficult problem, due to the min operation in the loss (3) that expresses the specificity of switched systems on the one hand, but also makes the objective function nonconvex and even nonsmooth and nondifferentiable on the other hand.

In the following, we present extensions of the $k$-LinReg algorithm (Lauer, 2013) that include the various forms of regularization obtained for different values of $q$. The original $k$-LinReg algorithm is a computationally efficient method to obtain approximate solutions to the unregularized switched system identification problem

$$\min_{w \in \mathbb{R}^{C \times d}} \mathcal{E}(w, X, y)$$

for $p = 2$. Though it is not guaranteed to find a global optimum, $k$-LinReg has been shown to often yield satisfactory solutions in practice. Let us first give an overview of its inner workings, which basically consist in an alternation between a classification step and a regression one:

1. Randomly initialize the parameters $w \in \mathbb{R}^{C \times d}$.
2. Classify the data into modes using the current $w_j$’s:

   $$\forall i \in [n], \hat{r}_i = \arg\min_{j \in [C]} (y_i - w_j^T x_i)^2$$

   and set $I_j = \{i \in [n] : \hat{r}_i = j\}$.

3. Update the parameter vectors $w_j$ by minimizing

   $$\sum_{i=1}^n (y_i - w_j^T x_i)^2 = \sum_{j=1}^C \sum_{i \in I_j} (y_i - w_j^T x_i)^2,$$

   which can be achieved by solving $C$ independent linear regression subproblems:

   $$\min_{w_j \in \mathbb{R}^{d \times d}} \sum_{i \in I_j} (y_i - w_j^T x_i)^2, \quad j = 1, \ldots, C.$$

(4) Repeat from Step 2 until convergence.

Though this algorithm can be shown to monotonically decrease the objective in (5), it is not guaranteed to find a global optimum. But its accuracy can be easily improved by restarting the algorithm multiple times from different random initializations (see Lauer (2013) for an analysis of the required number of restarts in practice).

In order to adapt the $k$-LinReg algorithm to the regularized forms of the problem, first note that regularization only affects Step 3 in the above procedure. However, it affects it in a nontrivial manner and in particular the submodels cannot always be independently estimated anymore. Specifically, Step 3 now has to solve (2) for the classification fixed by the $I_j$’s:

$$\min_{w \in \mathbb{R}^{C \times d}} \sum_{j=1}^C \sum_{i \in I_j} |y_i - w_j^T x_i|^p + \lambda \|\Omega(w)\|_q. \quad (6)$$
The different approaches proposed to perform this optimization are detailed below for the different choices of $q \in \{1, 2, \infty\}$.

3.1 Max-norm regularization ($q = \infty$)

For $q = \infty$, Step 3 amounts to solving

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{C} \sum_{j \in I_i} |y_i - w_j^T x_i|^p + \lambda \max_{j \in [C]} \|w_j\|_2,$$

which is a convex optimization problem. More precisely, it can be written as a Second-Order Cone Program (SOCP).

For instance, if we let $X_j$ and $y_j$ be the subsets of rows in $X$ and $y$ with index in $I_j$, and define $\xi_j$ from the error vector $\xi \in \mathbb{R}^n$ in the same manner, this yields, for $p = 1$,

$$\min_{w \in \mathbb{R}^{\mathcal{C} \times \mathcal{R}}, \tau \in \mathbb{R}} \|\mathbf{X}_j^T \mathbf{X}_j + \lambda \tau \|_F,$$

subject to

$$\|w_j\|_2 \leq \tau, \quad j = 1, \ldots, C.$$

Here, for $i \in I_j$, $\xi_i$ is the $i$th entry of $\xi$ which also belongs to $\xi_i$ and corresponds to, at the optimum, to the absolute loss for the $i$th data point assigned to the $j$th mode: $\xi_i^* = |y_i - x_i^T w_j|_2$.

For $p = 2$, a similar formulation can be obtained:

$$\min_{w \in \mathbb{R}^{\mathcal{C} \times \mathcal{R}}, \tau \in \mathbb{R}} \|\mathbf{X}_j^T \mathbf{X}_j + \lambda \tau \|_F,$$

subject to

$$\|w_j\|_2 \leq \tau, \quad j = 1, \ldots, C$$

where $\xi \in \mathbb{R}^C$ now contains a single component for each mode, which, at the optimum, equals the sum of the squared errors for the $j$th mode, i.e., $\xi_j^* = |y_j - X_j w_j|_2^2$.

3.2 $\ell_1$-norm regularization: inducing sparsity

Regularization with $q = 1$ can be used to favor sparse models in terms of the number of submodels: the minimization of the $\ell_1$-norm of $\Omega(w)$ is likely to yield a solution with many zeros in $\Omega(w)$, which corresponds to parameter vectors with $\|w_j\|_2 = 0$ that can probably be discarded from the hybrid model.

Formally, this leads to the optimization problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{C} \sum_{j \in I_i} |y_i - w_j^T x_i|^p + \lambda \sum_{j=1}^{C} \|w_j\|_2,$$

which can be cast into $C$ independent and convex subproblems of the form

$$\min_{w_j \in \mathbb{R}^d} \sum_{i \in I_j} |y_i - w_j^T x_i|^p + \lambda \|w_j\|_2.$$

Though convex, this problem might be difficult to solve due to the presence of the non-convex norm of $w_j$. In practice, it can be set as a SOCP, similar to (7) or (8) with the constraints $\|w_j\|_2 \leq \tau$ replaced by $\|w_j\|_2 \leq \tau_j$, $j = 1, \ldots, C$, and the term $\lambda \sum_{j=1}^{C} \tau_j$ substituted for $\lambda \tau$ in the objective. Yet, these SOCP might take a significant amount of time to solve.

Iterative reweighting algorithm. The solution to problems like (10) can be approached by an iterative scheme, in which one solves

$$\tilde{w}_j = \arg \min_{w_j \in \mathbb{R}^d} \sum_{i \in I_j} |y_i - w_j^T x_i|^p + \lambda \eta_j \|w_j\|_2^2$$

on the basis of the current solution and for a small value of $\varepsilon$ preventing a division by zero. The rationale is that, as the iterations go, the term $\eta_j \|w_j\|_2^2$ tends to the desired value of $\|w_j\|_2^2$. The advantage over (10) is purely computational. For instance, for $p = 2$, (11) is a standard ridge regression problem whose solution can be easily computed by

$$\tilde{w}_j = (X_j^T X_j + \eta \lambda \mathbf{I})^{-1} X_j^T y_j,$$

where $X_j$ and $y_j$ correspond to the subsets of rows in $X$ and $y$ with index in $I_j$. In addition, for large data sets (large $n$), most of the computation in (12) lies in the products $X_j^T X_j$ and $X_j^T y_j$, which can be precomputed to yield very simple iterations that boil down to inverting a small $d \times d$ matrix.

3.3 $\ell_2$-norm regularization ($q = 2$)

For $q = 2$, problem (6) becomes

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{C} \sum_{j \in I_i} |y_i - w_j^T x_i|^p + \lambda \left( \sum_{j=1}^{C} \|w_j\|_2 \right)^{1/2}.$$

Again, this is a convex problem, but more computationally demanding than in the case $q = 1$, since it cannot be decoupled into $C$ independent subproblems. SOCP formulations similar to (7) or (8) are obtained by replacing the constraints $\|w_j\|_2 \leq \tau_j$ by $\|w_j\|_2 \leq \tau_j$, $j = 1, \ldots, C$, together with $\|\tau_1, \ldots, \tau_C\|_2 \leq \tau$.

By using an iterative reweighting approach, a more efficient alternative is to iterate over

$$\tilde{w} = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{C} \sum_{j \in I_i} |y_i - w_j^T x_i|^p + \lambda \eta \sum_{j=1}^{C} \|w_j\|_2^2$$

with $\eta$ initialized to 1 and updated at each iteration by

$$\eta = \frac{1}{\varepsilon + \sqrt{\sum_{j=1}^{C} \|w_j\|_2^2}} = \frac{1}{\varepsilon + \|w\|_2}.$$
number might be unknown and should be estimated from the data along with the model itself. This is a model selection problem, in which one has to choose the “best” structure for the model that is being learned. To tackle this problem, we consider the structural risk minimization (SRM) principle from statistical learning theory. The idea is to derive statistical guarantees on the prediction error, and then to select the model with the best guarantees. In the context of switched system identification, this amounts to searching for the number of modes $C$ that minimizes an upper bound on the prediction error.

4.1 Errors Bounds for Regularized Switched Models

Let us first introduce the statistical framework within which the error bounds used in the SRM approach are derived. In statistical learning, data points $z_i = (x_i, y_i)$ are assumed to come from a realization of the stationary sequence of random variables $Z_i = (X_i, Y_i)$ taking values in $\mathbb{R}^d \times [-M, M]$ (outputs are assumed to be bounded by $M$) and distributed as $(X, Y)$. Then, the prediction error of a switched model $f$ composed of $C$ submodels $f_j$ is computed as the expected value of the pointwise loss defined as the smallest error with respect to all submodels:

$$L(\mathcal{F}) = \mathbb{E} \min_{j \in [C]} |Y - \mathcal{T}_j(X)|^p,$$  

where $\mathcal{F}$ and $\mathcal{T}_j$ are saturated versions of the models yielding outputs in $[-M, M]$. Here, $p \in \{1, 2\}$ provides a choice between the absolute and the squared loss. The prediction error (16) is known as the risk and measures the ability of the model to make accurate predictions on new data points, whereas the empirical risk,

$$\hat{L}(\mathcal{F}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [C]} |Y_i - \mathcal{T}_j(X_i)|^p,$$  

measures the ability of the model to fit the available data sample, of size $n$. With respect to the notations of Section 3, the error term $\mathcal{E}(\omega, X, Y)$ will thus be seen as a realization of the random variable $n\hat{L}(\mathcal{F})$.

So-called risk bounds are upper bounds, that hold with high probability, on the risk (16) in terms of the empirical risk (17) and a confidence interval involving a measure of the complexity of the model class $\mathcal{F}$, i.e., the set of all models possibly returned by the algorithm. The following theorem provides a general risk bound based on the empirical Rademacher complexity of the loss class $\mathcal{L} = \{\ell \in [0, (2M)^p]^\mathbb{Z} : \ell(z) = \min_{j \in [C]} |y - \mathcal{T}_j(x)|^p, \ f \in \mathcal{F}\}$ induced by the model class $\mathcal{F}$.

**Definition 1.** (Empirical Rademacher complexity). Given a sequence $\mathbf{Z}_n = (Z_i)_{1 \leq i \leq n}$ 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

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

where $\sigma_n = (\sigma_i)_{1 \leq i \leq n}$ is a sequence of Rademacher variables, i.e., random variables uniformly distributed in $\{-1, +1\}$.

**Theorem 1.** Let $\mathbf{Z}_n = (Z_i)_{1 \leq i \leq n}$ be a sequence drawn from a stationary $\beta$-mixing distribution. For any $\mu, \alpha > 0$ with $2\mu \alpha = n$ and $\delta > 4(\mu - 1)\beta(a)$, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad L(\mathcal{F}) \leq \hat{L}(\mathcal{F}) + 2\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}) + 3(2M)^p \sqrt{\log \frac{4}{2\mu} \cdot \frac{\delta}{\delta' - 4(\mu - 1)\beta(a)}},$$  

where $\delta' = \delta - 4(\mu - 1)\beta(a)$ and $\mathbf{Z}_n = (Z_0(i-1))_{1 \leq i \leq \mu}$ is a sample of length $\mu$ obtained by taking the first $Z_i$ out of every odd block of $\mathbf{Z}_n$ split into $2\mu$ blocks of length $a$.

Theorem 1 is obtained by application of Theorem 2 in Mohri and Rostamizadeh (2009), which offers a general bound that hold in a system identification context where the data cannot be assumed to be independent but are instead drawn from a mixing process (see Bradley (2005) for more details on the statistical assumption and the definition of the mixing coefficient $\beta(a)$).

Combining Theorem 1 with results from Lauer (2020b), a bound was developed for switched linear systems in Massucci et al. (2020) and used to estimate the number of modes. Its precise formulation involves the model class $\mathcal{F} = \mathcal{F}_0(\Lambda)\mathcal{C}$, built as the product of $C$ independent classes of the form

$$\mathcal{F}_0(\Lambda) = \{ f \in \mathbb{R}^X : f(x) = w^T \mathcal{X}, \|w\|_2 \leq \Lambda \},$$  

and for which the following characterization of the Rademacher complexity can be obtained:

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}_0) \leq p(2M)^{p-1}C \hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}_0) \leq p(2M)^{p-1}C \frac{\Lambda}{\mu} \sqrt{\sum_{i=1}^{\mu} \|X_{2\alpha(i-1)+1}\|^2}.$$

The linear growth of (20) with the number of modes $C$ is a consequence of the choice of $\mathcal{F}$ as a mere product of independent component classes, for which the complexities of the submodels $f_j$ do not interact with each other. However, when using a regularization scheme such as in (4), the complexities of each component submodel $f_j$ influences the overall complexity measured by $\|\mathcal{O}(w)\|_q$. In other words, $\|\mathcal{O}(w)\|_q$ offers a more fine-grained measure of the complexity by looking not only at the number of submodels, but also at the complexity of each submodel. In particular, we have

$$\forall q \in (0, \infty], \quad \|\mathcal{O}(w)\|_q \leq C \max_{j \in [C]} \|w_j\|_2 = \|\mathcal{O}(w)\|_\infty,$$

in which we observe that (20) is especially suitable when regularizing with $q = \infty$, but might fail to accurately measure the complexity of models regularized in a different manner.

To better take into account the regularization with $q \in \{1, 2\}$, we instead make use of the analysis of Lauer (2020a). For models of the regularized class $\mathcal{F}(\Lambda) = \{ f \in \mathcal{F}_0(\Lambda)\mathcal{C} : \|\mathcal{O}(w)\|_q \leq \Lambda \}$, (21) this leads to a refined bound on the Rademacher complexity of the form
Given a maximum number of modes similar to Massucci et al. (2020), to derive the final form of the bound using a proof scheme a discretized value from the estimated \( \Lambda \) as

\[
\tilde{\Lambda}(w) = \min \{ \Lambda \in \{ 1, \ldots, K \} : \Lambda \geq \| \Omega(w) \|_q \}
\]  

(25)

to derive the final form of the bound using a proof scheme similar to Massucci et al. (2020).

**Theorem 2.** Given a maximum number of modes \( C \) and a grid of \( K \) values as in (25), for a sample \( Z_n \) of size \( n \) drawn from a stationary \( \beta \)-mixing distribution, and for any \( \mu, a > 0 \) with \( 2\mu a = n \), and \( \delta > 4C^2K(\mu - 1)\beta(a) \), with probability at least \( 1 - \delta \), for all \( C \in [C] \) and all \( f \in F(\Lambda) \),

\[
L(\mathcal{F}) \leq L(\mathcal{J})
\]

\[
\leq 2p(2M)^{p-1} \alpha(C,q) \Lambda(\mathcal{F}) \sqrt{\sum_{i=1}^{n} \| X_{2a(i-1)+1} \|^2} + 3(2M)^{p} \sqrt{\frac{\log(C^K) + \log \frac{4}{\delta}}{2\mu}}
\]

(26)

where \( \delta' = \delta - 4CK(\mu - 1)\beta(a) \) and \( \tilde{\Lambda}(w) \) is as in (25).

Based on Theorem 2, we can now state the overall model selection procedure, which selects the “best” number of modes \( C \) that minimizes the upper bound in (26). More precisely, we can spare the computation of the last term, which does not depend on \( C \), and thus bypass the issue of estimating the mixing coefficient \( \beta(a) \) influencing \( \delta' \). This leads to Algorithm 1.

**Algorithm 1** SRM with \((p,q)\)-regularization

**Require:** The data \( X \in \mathbb{R}^{n \times d} \), \( y \in \mathbb{R}^n \) and a maximum number of modes \( C \)

**for** \( C = 1 \) to \( C \) **do**

Compute

\[
J(C) = \hat{L}(\mathcal{F}) + 2p(2M)^{p-1} \alpha(C,q) \hat{\lambda}(w)
\]

\[
\times \sqrt{\sum_{i=1}^{n} \| X_{2a(i-1)+1} \|^2}
\]

end for

**return** \( \hat{C} \) and the corresponding model.

**Algorithm 1** SRM with \((p,q)\)-regularization

1. **Input:** Data \( X \in \mathbb{R}^{n \times d} \), target \( y \in \mathbb{R}^n \), maximum number of modes \( C \), regularization parameter \( \mu \), threshold \( \delta \), number of iterations \( n_{\text{iter}} \).

2. **Set** \( C = 1 \).

3. **Repeat**

   a. **Compute** the \( C \) modes using a method that satisfies the \( \delta \)-confidence and \( \mu \)-level.

   b. **Estimate** the objective function \( J(C) \). If \( J(C) \leq \delta \) and \( \mu \), stop.

   c. **Increase** \( C \) by 1.

4. **Return** the model with the minimum \( J(C) \).

**Fig. 1.** Ratio of the computing time of the direct optimization of the SOCPs and the IR schemes versus the number of data \( n \).

**5.1 Optimization Issues and Computing Time**

Here, we compare the two optimization approaches relying either on a straightforward formulation as a SOCP or on an iterative reweighting (IR) scheme. Experiments are run with \( n \) uniformly drawn data points \( x_i \in [-5,5]^d \), \( y_i = x_i^T \theta_i + e_i \), a random switching sequence \( r \in [C]^n \), random parameters \( \theta_j \) \( j \in [-1,1]^d \) and a zero-mean Gaussian noise \( e_i \) of unit variance, for different problem sizes \( n \).

Figure 1 shows the ratio between the computing time of the SOCP solved using MOSEK and that of the IR scheme for \( p = 2 \) and \( q \in \{1,2\} \), plotted with respect to \( n \) for \( C = 3 \) and \( d = 3 \). In all cases, a single initialization of the \( k \)-LinReg procedure is considered with \( \lambda = 1 \) and the IR scheme is stopped after 5 iterations. These results show that, for \( p = 2 \), the IR schemes can offer significant speed-ups up to factors of a few hundreds when compared to a straightforward optimization. Similar speedups were observed for various values of \( C \) and \( d \). However, for the case \( p = 1 \), problem (11) remains a quadratic program, with linear constraints implementing the absolute loss, and for which no explicit formula such as (12) is available.

In this case, the IR scheme iterating over (11), though removing the conic constraints, proved to be slower than the straightforward optimization of (10). Regarding the relative optimality gap induced by the IR approximations, \( (\text{obj}_{\text{IR}} - \text{obj}_{\text{SOCP}})/\text{obj}_{\text{SOCP}} \), where \( \text{obj}_{\text{method}} \) refers to the value of the objective function of (6) computed with \( w \) set to its value estimated by the method, it was satisfactorily below 1% in all our tests. Therefore, in all experiments below, we always use the IR schemes for \( p = 2 \).
Table 1. Average of the FIT criterion (27) computed on the training and test sets with and without regularization over 100 trials.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training FIT</th>
<th>Test FIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>no regularization (Lauer, 2013)</td>
<td>95.6</td>
<td>86.0</td>
</tr>
<tr>
<td>ℓ1-norm regularization (9)</td>
<td>95.7</td>
<td>88.7</td>
</tr>
<tr>
<td>ℓ2-norm regularization (13)</td>
<td>96.3</td>
<td>89.2</td>
</tr>
</tbody>
</table>

Fig. 2. Histogram of $\hat{C}$ using Algorithm 1 for a switched dynamical system with $C = 6$ modes.

5.2 Benefits of Regularization

Regularization is known to be particularly useful to prevent overfitting, i.e., to avoid too complex models that can fit very well the training data but do not generalize well to other data. This phenomenon is illustrated here with experiments similar to Sect. 5.1, but in higher dimension with $n = 500$, $d = 70$, $C = 2$, and with 10 restarts from random initializations of the k-LinReg procedure. The ability of the models to generalize to new test data points is evaluated. Specifically, Table 1 reports the

$$\text{FIT} = \frac{\sum_{i=1}^{n} \min_{j \in [C]} (y_i - w_j^T x_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \times 100\%,$$

(27)

computed either on the training set or on an independent test set of the same size and where $\bar{y}$ is the mean value of the $y_i$'s. While regularization has almost no effect for the training data, it yields an increase of accuracy on the test data.

5.3 Switched System Identification and the Estimation of the Number of Modes

We now consider the estimation of a switched dynamical systems with $C = 6$ modes of the form

$$y_i = \theta_i^T x_i + e_i$$

with orders $n_a = n_b = 2$ and $x_i$ as in (1). Data sets of $n = 300,000$ points are generated with this system, a zero-mean Gaussian input $u_i$ of unit variance and a zero-mean Gaussian noise $e_i$ with a Signal-to-Noise Ratio of 30 dB.

Algorithm 1 is applied with $\overline{C} = 10$, $(p,q) = (1,2)$ and the iterative scheme (14)–(15) to estimate the models. Over 100 trials, Figure 2 shows the histogram of the estimated number of modes $\hat{C}$, which coincides with the true $C = 6$ in 95% of the cases.

6. CONCLUSION

This paper showed how to include regularization in the identification of switched linear systems, how to efficiently solve the resulting optimization problems, and how to take full advantage of this regularization in a model selection framework based on statistical learning.

Future work will try to improve the statistical results to make the model selection method effective with smaller data sets. Another research direction concerns the investigation of the sparsity induced by the ℓ1-norm regularization, and how this could help to estimate the number of modes in practice. Finally, the effect of regularization in a system identification context should be better understood. In particular, the relationship with the instrumental variable method will be investigated.

REFERENCES


