A statistical learning perspective on switched linear system identification

Louis Massucci, Fabien Lauer, Marion Gilson

A Statistical Learning Perspective on Switched Linear System Identification

Louis Massucci[a,b], Fabien Lauer[a], Marion Gilson[b]

[a] *Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France*
[b] *Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France*

Abstract

Hybrid systems form a particularly rich class of dynamical systems based on the combination of multiple continuous subsystems and a discrete mechanism deciding which one of these is active at a given time. Their identification from input-output data involves nontrivial issues that were partly solved over the last twenty years thanks to numerous approaches. However, despite this effort, estimating the number of modes (or subsystems) of hybrid systems remains a critical and open issue. This paper focuses on switched linear systems and proposes an analysis of their identification based on statistical learning theory. This leads to new theoretically sound bounds on the prediction error of switched models on the one hand, and a practical method for the estimation of the number of modes on the other hand. The latter is inspired by the structural risk minimization principle developed in statistical learning for model selection. The proposed analysis is conducted under various assumptions on the model class and regularization schemes for which new algorithms are presented. Numerical experiments are also provided to illustrate the accuracy of the proposed method.

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

1. Introduction

Hybrid systems are dynamical systems that include interacting continuous and discrete dynamical behaviors. This results in systems that switch from an operating mode with certain continuous dynamics to another [12]. Hybrid systems can be used to describe real phenomena that exhibit discontinuous behaviors, or to estimate continuous ones with a class of more simple models. For instance, nonlinear dynamical systems can be approximated by multiple linear subsystems. The switchings between the different operating modes can be either state-dependent for piecewise smooth systems or time-dependent for arbitrarily switched systems.

This paper focuses on the identification of arbitrarily switched linear systems from input-output data. The switching mechanism, which is considered to be externaly generated, and the true number of modes are assumed to be unknown. In this case, the identification problem is highly complex as it requires not only to estimate a submodel for each subsystem but also to classify data points between the different submodels. Over the last decades, a lot of methods have been devised to estimate such systems. As specified in [12], switched system identification is divided into two main approaches. The first one assumes a fixed or known number of modes, as proposed in [27, 13, 9, 15]. The second one assumes the number of subsystems unknown and consists in estimating the minimal number of modes satisfying a predefined threshold on the error. Several papers focus on this approach, as, e.g., [3, 1, 22, 21, 7, 24, 6]. In each of these methods, the threshold on the error is the main hyperparameter that influences the recovery of the correct number of modes. Therefore, for both categories of approaches, the estimation of the number of modes remains a critical issue. Meanwhile, machine learning techiques have been widely used recently to produce new solutions in system identification as in, e.g, [24, 15, 6].

In this paper, we analyze switched system identification in the framework of statistical learning theory [26, 20] and investigate its consequences regarding the estimation of the number of modes. Statistical learning theory aims at describing the mechanisms at play when learning predictive models from data. In particular, it can be used to derive upper bounds on the prediction error of the models. Such bounds enjoy desirable properties such as being distribution-free, non-asymptotic and independent of the algorithm used to optimize the model. This paper develops new bounds tailored for switched system identification and proposes a new method to estimate the number of modes. This method relies on a generalization error bound, i.e., an upper bound on the expected error of the model. These bounds are derived with an independence assumption on the data which does not hold in system identification. This issue has been overcome in [19], where bounds for dependent data are obtained under a mixing assumption. More precisely, error bounds for dependent data are based on measures of dependence called mixing coefficients [4] and on the independent block sequence construction due to [29]. The general framework of [19] is here combined with the ones described in [10] and [11], which present bounds for switched regression models in the static case, to propose new error bounds for switched system identification. Then, the Structural Risk Minimization (SRM) principle is applied on the basis of these bounds in order to estimate the number of modes. This requires to produce another *uniform* error bound suitable for the practical setting of model selection. Preliminary versions of this work [16, 18] that worked with various assumptions on the model class are here unified in a general framework that encompasses and fully

exploits regularization.

Paper Organization: Section 2 presents the switched system identification problem and introduces tools from statistical learning theory and their extension based on mixing arguments. In Section 3, error bounds are derived for switched system identification, and a model selection method to estimate the number of modes is proposed. Section 3.3 also discusses the algorithmic consequences for switched system identification. Section 4 illustrates the application of the new model selection technique we propose. Section 5 concludes the paper and discusses open issues.

Notation: Vectors are written in bold and lowercase letters. For any integer $n \geq 1$, $[n]$ 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. For a vector $\boldsymbol{a} \in \mathbb{R}^n$ and any $q \in (0,\infty)$, $\|\boldsymbol{a}\|_q = (\sum_{k=1}^n |a_k|^q)^{1/q}$ denotes its ℓ_q-norm for $q \geq 1$ or ℓ_q-quasi-norm for $q \in (0,1)$, while $\|\boldsymbol{a}\|_\infty = \max_{k\in[n]} |a_k|$ is its ℓ_∞-norm. Given two sets $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{Y}^\mathcal{X}$ denotes the set of functions from $\mathcal{X}$ to $\mathcal{Y}$.

2. Preliminaries

We start by introducing the switched system identification problem in Sect. 2.1. Then, the basics of learning theory are presented in Sect. 2.2 and their extension to the specific context of system identification is further detailed in Sect. 2.3.

2.1. Switched linear system identification

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

$$\begin{aligned} y_i &= f_{r_i}(\boldsymbol{x}_i) + e_i, \\ f_j(\boldsymbol{x}_i) &= \boldsymbol{w}_j^T \boldsymbol{x}_i, \qquad j = 1,\dots,C, \end{aligned} \tag{1}$$

where $y_i \in \mathbb{R}$ is the output, $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ the regression vector, $r_i \in \{1,\dots,C\}$ the discrete state or mode, C the number of submodels, f_j with $j \in \{1,\dots,C\}$ the linear submodel of parameters $\boldsymbol{w}_j \in \mathbb{R}^d$ and $e_i \in \mathbb{R}$ a noise term. The regressor $\boldsymbol{x}_i \in \mathbb{R}^d$, $d = n_a + n_b$, with the model orders n_a and n_b, is given by

$$\boldsymbol{x}_i = \left[-y_{i-1},\ \dots, -y_{i-n_a},\ u_{i-1},\dots,u_{i-n_b}\right]^T, \tag{2}$$

where the u_{i-k}'s denote the delayed inputs.[1] Hybrid system identification is concerned with the case where the mode r_i is unknown, and we here focus on arbitrary switchings that can occur at any time step without dwell time and assume that the orders n_a and n_b are known.

Problem 1. *Given a data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a set $\mathcal{F}$ of possible hybrid models, estimate the number of submodels C, the hybrid model $\boldsymbol{f} = \{f_j\}_{j=1}^C$ made of C linear submodels $f_j(x) = \boldsymbol{w}_j^T \boldsymbol{x}$ and the switching sequence $\mathbf{r} = (r_i)_{1\leq i\leq n} \in [C]^n$.*

[1] MISO systems with $\boldsymbol{u}_i \in \mathbb{R}^{d_u}$, $d_u > 1$, can be considered with $\boldsymbol{x}_i = [y_{i-1},\dots,y_{i-n_a},\boldsymbol{u}_i^T,\boldsymbol{u}_{i-1}^T,\dots,\boldsymbol{u}_{i-n_b}^T]^T$.

More precisely, the accuracy of a switching model is measured in terms of a pointwise switching loss function

$$\ell(\boldsymbol{f},\boldsymbol{x},y) = \min_{j\in[C]} |y - f_j(\boldsymbol{x})|^p. \tag{3}$$

Here, $p \in \{1,2\}$ selects between the absolute and the squared loss. The minimum operation included in the loss accounts for the fact that the only error that matters for a data points $(\boldsymbol{x}_i, y_i)$ is the one computed with respect to the submodel f_j that best estimates y_i. Therefore, in the case the error over all data is sufficiently minimized, this ensures that there is at least one submodel that accurately estimates y_i for all points.

2.2. Statistical learning theory and error bounds

Statistical learning theory aims at obtaining guarantees on models learned from data under minimal assumptions. For instance, knowledge on the kind of noise, or on the structure of the system being modeled is not needed. These guarantees typically take the form of upper bounds on the expected error of the model, called the risk or generalization error. First, for a clear understanding, some preliminaries are recalled.

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 the unknown probability distribution of the 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 $\mathbf{Z}_n = (Z_i)_{1\leq i\leq n} = ((X_i,Y_i))_{1\leq i\leq n}$ 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)

$$L(f) = \mathbb{E}_{X,Y}\ell(f,X,Y), \tag{4}$$

defined as the expected value, $\mathbb{E}_{X,Y}\ell(f,X,Y) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f,x,y)p(x,y)\,dxdy$, of the loss function $\ell(f,X,Y)$. This loss function measures the pointwise prediction error between $f(X)$ and Y and is generally of the form $\ell(f,X,Y) = |Y - f(X)|^p$ with $p \in \{1,2\}$, for standard regression problems without switchings.

In practice, the expected value cannot be computed due to the unknown probabilty distribution of Z. Therefore, many methods minimize instead the empirical risk

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

for a realization of $\mathbf{Z}_n = \mathbf{z}_n$. So-called generalization error bounds are upper bounds on the risk (4) 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:

$$\mathbb{P}\left\{\forall f \in \mathcal{F},\ L(f) \leq \hat{L}(f) + \epsilon(n,\mathcal{F},\delta)\right\} \geq 1-\delta, \tag{6}$$

where the probability $\mathbb{P}$ is over the random draw of $\mathbf{Z}_n$. A large part of learning theory is dedicated to deriving the tightest confidence interval $\epsilon(n,\mathcal{F},\delta)$, which depends on the sample size n, but also on the capacity of the model class $\mathcal{F}$. More precisely,

the capacity of the loss class

$$\mathcal{L} = \left\{ \ell \in \mathbb{R}^{\mathcal{Z}} : \ \ell(z) = \ell(f, x, y), \ f \in \mathcal{F} \right\} \tag{7}$$

must be regarded, and can be measured for instance with the Rademacher complexity, as initiated by [8, 2].

Definition 1 (Empirical 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*

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

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

The Rademacher complexity measures the capacity of the loss class to provide functions whose outputs align well with the random directions $\boldsymbol{\sigma}_n$. A general bound based on the Rademacher complexity is the following:

Theorem 1 (Theorem 1 in [20]). *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 $Z \in \mathcal{Z}$. Then, for any fixed $\delta \in (0, 1)$, with probability at least $1 - \delta$, uniformly over all $\ell \in \mathcal{L}$,*

$$\mathbb{E}_Z \ell(Z) \le \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i) + 2\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}) + 3B \sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \tag{9}$$

To derive such bounds as in Theorem 1, the loss function needs to be bounded. Assuming 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*

$$\bar{t} = \min(M, \max(-M, t)). \tag{10}$$

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, we can state that for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(\overline{f}, x, y) \le \ell(f, x, y)$, and the generalization error (4) of the clipped model, $L(\overline{f})$, is always smaller than $L(f)$. Then, $\overline{f}$ is used instead of f to make predictions and upper bounds on $L(\overline{f})$ are derived.

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

$$\mathcal{F} = \left\{ f : f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}, \ \boldsymbol{w} \in \mathbb{R}^d, \ \|\boldsymbol{w}\|_2 \le \Lambda \right\} \tag{11}$$

and the loss class (induced by the clipped model)

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

The Rademacher complexity of $\mathcal{L}$ can be related to that of $\mathcal{F}$ by using a contraction argument (Theorem 5 in Appendix A). This yields the relation

$$\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}) \le 4M \hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}), \tag{12}$$

which in turn can be bounded using standard computations for Rademacher complexities (see Theorem 6 in Appendix A) as

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) \le \Lambda \sqrt{\sum_{i=1}^{n} \|\boldsymbol{X}_i\|_2^2} / n. \tag{13}$$

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 (11)*,*

$$L(\overline{f}) \le \hat{L}(\overline{f}) + \frac{8M\Lambda \sqrt{\sum_{i=1}^{n} \|\boldsymbol{X}_i\|_2^2}}{n} + 12M^2 \sqrt{\frac{\log \frac{2}{\delta}}{2n}}, \tag{14}$$

where $L(\overline{f})$ stands for the expected mean square error (MSE) and $\hat{L}(\overline{f})$ for the MSE over the training sample.

2.3. Learning theory for system identification

One may point out that the independence assumption in Theorem 1 prevents its application to system identification. This is due to the regressor vector $\boldsymbol{x}_i$ in (2), which depends on lagged outputs y_{i-k}. In the following, the non-i.i.d setting is studied in order to make the general error bounds suitable for dynamical systems. In this paper, we characterize the degree of dependence between data using the β-mixing coefficient. According to [28], β-mixing is a natural assumption on non-i.i.d stochastic processes. The following assumes that the sample $\mathbf{Z}_n$ is taken from a stationary β-mixing process.

Definition 3 (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, $i \le k \le j$. Then, for any positive integer k, the β-mixing coefficient of the stochastic process $\mathbf{Z}$ is defined as*

$$\beta(k) = \mathbb{E}_{B \in \sigma_{-\infty}^{0}} \left\{ \sup_{A \in \sigma_k^{\infty}} |\mathbb{P}(A|B) - \mathbb{P}(A)| \right\}. \tag{15}$$

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 (15) and $\beta(k) = 0$ for all $k \ge 1$. 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 [4].

Rademacher complexity bounds for dependent data were obtained in [19] by applying the independent block sequence construction of [29]. In short, the whole data set is split into a sequence of 2μ blocks of length a, and only a subset of μ blocks instead of the entire set of n dependent data points is considered (see Figure 1). By doing so, for β-mixing processes, concentration inequalities can be applied to a sequence of independent

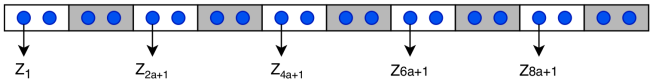

Figure 1: Illustration of the block sequence with $n = 20$, $a = 2$ data points per block and $\mu = 5$ odd blocks (white) and μ even blocks (grey). The sample $\mathbf{Z}_\mu = (Z_1, Z_{2a+1}, Z_{4a+1}, \ldots, Z_{8a+1})$ contains the first point of each odd block. The dependence between the odd blocks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

blocks distributed as those in the odd subset while controlling the error thus introduced. This leads to the following adaptation of Theorem 1.

Theorem 2 (Theorem 2 in [19]). *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 $\ell \in \mathcal{L}$,*

$$\mathbb{E}\ell(Z_1) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(Z_i) + 2\hat{\mathcal{R}}_{\mathbf{Z}_\mu}(\mathcal{L}) + 3B\sqrt{\frac{\log\frac{4}{\delta'}}{2\mu}}, \tag{16}$$

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.

In Theorem 2, the confidence interval decreases with the "effective number of data" $\mu = n/2a$, in comparison with Theorem 1, where it more directly depends on the number of data points n. In addition, in Theorem 2, the confidence interval also slightly increases due to the use of δ' instead of δ, which also implies the constraint $\delta > 4(\mu-1)\beta(a)$. The apparent increase of δ with respect to μ is related to the fact that the error between the analysis on $\mathbf{Z}_n$ and that on $\mathbf{Z}_\mu$ increases with the sample size. However, for β-mixing sequences, a can typically be chosen so that $(\mu-1)\beta(a)$ is small. Also note that this is inherent to the approach of [28] and the mixing arguments. A more recent approach without this weakness was developed in [22] for linear system identification. However, this remains limited to the analysis of the empirical risk minimizer $\hat{f} \in \operatorname{argmin}_{f\in\mathcal{F}} \hat{L}(f)$, which, in the context of switched systems, remains elusive in most cases. Thus, in this paper, we prefer to use mixing arguments to derive *uniform* error bounds that apply more broadly to *any* model $f \in \mathcal{F}$, including those that do not minimize $\hat{L}(f)$.

3. Statistical learning theory for hybrid system identification

We now turn to the adaptation of statistical learning theory to hybrid system identification. More precisely, we focus on linear hybrid systems as in (1), whose complexity is measured by $\|\boldsymbol{w}_j\|_2$ for each linear submodel. Indeed, $\|\boldsymbol{w}_j\|_2$ corresponds to the magnitude of the gradient of the submodel f_j, and thus, the larger $\|\boldsymbol{w}_j\|_2$ is, the more pronounced the variation of f_j's outputs are.

We further concentrate on the hybrid model class

$$\mathcal{F} = \left\{\boldsymbol{f} = (f_j)_{1\leq j\leq C} : f_j(\boldsymbol{x}) = \boldsymbol{w}_j^T\boldsymbol{x},\ j = 1\ldots C,\ \|\Omega(\boldsymbol{f})\|_q \leq \Lambda\right\},$$

where

$$\Omega(\boldsymbol{f}) = [\|\boldsymbol{w}_1\|_2,\ \ldots,\ \|\boldsymbol{w}_C\|_2]^T \tag{17}$$

and q, typically chosen in $\{1, 2, \infty\}$, selects a particular regularization scheme depending on the intended properties of the model. In general terms, the regularization embedded in (17) penalizes complex hybrid models, whose complexity is measured at a global level by the ℓ_q-norm of a vector $\Omega(\boldsymbol{f})$, in which each component measures the complexity of one of the submodels according to the ℓ_2-norm of its parameter vector.

Section 3.1 below discusses the risk bounds that can be derived for such models. Then, Section 3.2 proposes to adapt these bounds to make them hold uniformly over a range of values for C in order to proceed to model selection. Section 3.3 finally discusses the algorithmic consequences of the considered regularization schemes.

3.1. Error bounds for Switched linear system identification

We now have all the ingredients needed to derive a new generalization error bound for switched system identification including regularization in the form of (17)-(17). The following presents the adaptation of the bound of Section 2.3 to switched system identification.

Theorem 3. *Let $\mathcal{F}$ be a vector-valued function class with C components as in (17) and the loss ℓ be as in (3). 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 $\boldsymbol{f} \in \mathcal{F}$:*

$$L(\overline{\boldsymbol{f}}) \leq \hat{L}(\overline{\boldsymbol{f}}) + 2p(2M)^{p-1}\alpha(C,q)\frac{\Lambda\sqrt{\sum_{i=1}^{\mu}\|\boldsymbol{X}_{2a(i-1)+1}\|_2^2}}{\mu} + 3(2M)^p\sqrt{\frac{\log\frac{4}{\delta'}}{2\mu}},$$

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

$$\alpha(C,q) = \begin{cases} C, & \text{if } q = \infty \\ \frac{q}{q-1}C^{1-\frac{1}{q}}, & \text{if } 1 < q < \infty \\ 1+\log C, & \text{if } q = 1, \\ \frac{1}{1-q}, & \text{if } 0 < q < 1. \end{cases} \tag{18}$$

Proof of Theorem 3. See Appendix B.

In the bound of Theorem 3, the effects of the different regularizations are carried by the parameter $\alpha(C,q)$. The case $q = \infty$ corresponds to the case where the submodels are considered as independent. $q = 2$ is the more classic scenario, where the dependence between the submodels is considered through the sum of squares of the submodels complexities. For $q = 1$, sparse models with few nonzero vectors $\boldsymbol{w}_j$ are favored and this translates into a low dependence on C for the bound.

3.2. Uniform bounds for model selection

In switched system identification, a critical issue is the estimation of the number of modes. This is a model selection

problem that we propose to tackle with an approach from statistical learning theory. This method is the structural risk minimization principle, that consists in minimizing a generalization error bound with respect to the hyperparameters. In our case, the hyperparameter is the number of modes, C.

Formally, 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 $\boldsymbol{f} \in \mathcal{F}$ and $\mathbf{r} \in [C]^n$. In this scheme, the generic algorithm is applied for all numbers of modes C within a predefined range. Then, the "best" model is selected on the basis of a criterion $J(C)$ designed below[2] from the error bound.

Algorithm 1 Switched system SRM with (p, q)-regularization

Require: The data $((\boldsymbol{x}_i, y_i))_{1\le i\le n} \in (\mathbb{R}^d \times \mathbb{R})^n$ and a maximum number of modes $\overline{C}$

for $C = 1$ **to** $\overline{C}$ **do**

Estimate $\boldsymbol{f}$ with C modes using Alg. 2 from Sect. 3.3. (or any algorithm working with a fixed C)

Compute

$$J(C) = \hat{L}(\overline{\boldsymbol{f}}) + 2p(2M)^{p-1}\alpha(C,q)\tilde{\Lambda}(\boldsymbol{f})\frac{\sqrt{\sum_{i=1}^{\mu} \|\boldsymbol{x}_{2a(i-1)+1}\|_2^2}}{\mu}.$$

end for

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

return $\hat{C}$ and the corresponding model.

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 Λ for the model class (17), 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 Λ could be computed *a posteriori* as

$$\Lambda(\boldsymbol{f}) = \|\Omega(\boldsymbol{f})\|_q \tag{19}$$

from the estimated $\boldsymbol{f}$. More precisely, we have to consider a discretized value

$$\tilde{\Lambda}(\boldsymbol{f}) = \min\left\{\Lambda \in \{\Lambda_1, \dots, \Lambda_K\} \;:\; \Lambda \ge \|\Omega(\boldsymbol{f})\|_q\right\}. \tag{20}$$

In other words, a grid of K possible values of Λ is considered, and $\tilde{\Lambda}(\boldsymbol{f})$ represents the first point in the grid that is larger than the actual value of $\Lambda(\boldsymbol{f})$.

Theorem 4. *Given a maximum number of modes $\overline{C}$, a model class $\mathcal{F}$ as in* (17)*, and a grid $\{\Lambda_1, \dots, \Lambda_K\}$ of K values as in* (20) *such that $\Lambda_K = \Lambda$, 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 $1-\delta$, for all $C \in [\overline{C}]$ and all $\boldsymbol{f} \in \mathcal{F}$,*

$$L(\overline{\boldsymbol{f}}) \le \hat{L}(\overline{\boldsymbol{f}}) + \frac{2p(2M)^{p-1}\alpha(C,q)\tilde{\Lambda}(\boldsymbol{f})\sqrt{\sum_{i=1}^{\mu} \|\boldsymbol{X}_{2a(i-1)+1}\|_2^2}}{\mu} + 3(2M)^p\sqrt{\frac{\log(\overline{C}K) + \log\frac{4}{\delta'}}{2\mu}}, \tag{21}$$

where $\delta' = \delta - 4\overline{C}K(\mu-1)\beta(a)$ and $\tilde{\Lambda}(\boldsymbol{f})$ is as in (20)*.*

Proof of Theorem 4. See Appendix C.

We can now formulate the criterion $J(C)$ used in the general model selection framework of Algorithm 1 on the basis of this bound. In particular, we leave aside constant terms that do not depend on C nor on the estimated model $\boldsymbol{f}$ in the bound (21) and obtain

$$J(C) = \hat{L}(\overline{\boldsymbol{f}}) + \epsilon(\boldsymbol{f}, C), \tag{22}$$

$$\epsilon(\boldsymbol{f}, C) = \frac{2p(2M)^{p-1}\alpha(C,q)\tilde{\Lambda}(\boldsymbol{f})\sqrt{\sum_{i=1}^{\mu} \|\boldsymbol{x}_{2a(i-1)+1}\|_2^2}}{\mu}$$

with a and μ such that $2\mu a = n$.

Note in particular that the precise value of the mixing coefficient $\beta(a)$ in the bound of Theorem 4 does not influence the model selection procedure and the values of $J(C)$.

Example 2. *Consider a switched system composed of $C = 3$ linear subsystems of orders $n_a = n_b = 2$ with parameter vectors*

$$\begin{aligned}\boldsymbol{w}_1 &= [-0.4 \;\; 0.25 \;\; -0.15 \;\; 0.08]^T,\\ \boldsymbol{w}_2 &= [1.55 \;\; -0.58 \;\; -2.1 \;\; 0.96]^T,\\ \boldsymbol{w}_3 &= [1 \;\; -0.24 \;\; -0.65 \;\; 0.30]^T.\end{aligned} \tag{23}$$

This system is used to generate a data set of $n = 80\,000$ 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 equal to 30 dB with respect to the output signal. The active mode r_i is uniformly distributed in $\{1, 2, 3\}$.*

Applying Algorithm 1 with $p = 2$ and $q = \infty$, a representative curve of the bound $J(C)$ is obtained as in Fig. 2, where we can see a minimum achieved at $C = 3$. The confidence interval is linearly increasing along with the number of modes. On the opposite, the MSE curve is monotonically decreasing. The sum of both curves, that gives $J(C)$, shows a minimum that helps us to find the number of modes.

3.3. Algorithmic consequences

Regularization is a standard technique used in practice to control the model complexity while learning from data. Instead of constraining the model, it often amounts to learning by minimizing a trade-off between a data fitting term, $\hat{L}(\boldsymbol{f})$, and a regularization term, $\Gamma(\boldsymbol{f})$, that penalizes highly complex models:

$$\min_{\boldsymbol{f}\in\mathcal{F}_0} \hat{L}(\boldsymbol{f}) + \lambda\,\Gamma(\boldsymbol{f}), \tag{24}$$

[2] $J(C)$ also depends on the estimated model $\boldsymbol{f}$, but we keep the notation short to emphasize the number of modes C.

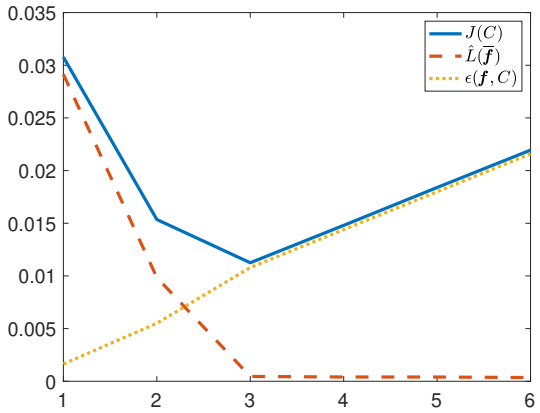

Figure 2: Resulting bound from Example 2 versus the number of modes C.

where $\mathcal{F}_0$ is an unconstrained model class such as

$$\mathcal{F}_0 = \left\{ \boldsymbol{f} = (f_j)_{1 \le j \le C} : f_j(\boldsymbol{x}) = \boldsymbol{w}_j^T \boldsymbol{x},\ \boldsymbol{w}_j \in \mathbb{R}^d,\ j = 1, \dots, C \right\}$$

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 term and the model complexity.

For the regularization, here, we focus on regularization terms of the form $\Gamma(\boldsymbol{f}) = \|\Omega(\boldsymbol{f})\|_q$. Hence, the global complexity $\Gamma(\boldsymbol{f})$ is a function of all submodels complexities. Overall, a specific type of regularized switched system identification problem (24) is denoted by the pair (p, q), in which p and q define the loss and the regularization, respectively. Then, the regularization constant λ will define how much the complex models will be penalized, hence playing an inverse role to Λ in (17).[3] For any pair (p, q), (24) 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 on the other hand.

Here, extensions of the k-LinReg algorithm [9] that include the various forms of regularization obtained for different values of q are presented. While other algorithms could be used and extended in our framework, k-LinReg is a computationally efficient method to obtain approximate solutions to the classic switched system identification problem without regularization 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 by restarting the algorithm multiple times from different random initializations (see [9] for the estimation of its probability of success). It basically consists in an alternation between a classification step and a regression one, as detailed in Algorithm 2.

The main difference between Alg. 2 and the vanilla k-LinReg is that, in most cases, the submodels have to be estimated all together when solving (25), whereas this optimization can be decoupled into C subproblems without regularization. More details on how to efficiently solve this convex problem in practice can be found in [18].

[3] See [5] for more details on the relation between λ and Λ.

Algorithm 2 Switched system identification with (p, q)-regularization

Require: The data $((\boldsymbol{x}_i, y_i))_{1 \le i \le n} \in (\mathbb{R}^d \times \mathbb{R})^n$, the choice of (p, q), and the number of restarts R

Randomly initialize the parameters $\boldsymbol{w}_j \in \mathbb{R}^d$, $j = 1, \dots, C$

repeat

Classify the data into modes using the current $\boldsymbol{w}_j$'s:

$$\forall i \in [n], \quad \hat{r}_i = \operatorname*{argmin}_{j \in [C]} |y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i|^p$$

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

Update the parameter vectors $\boldsymbol{w}_j$ by solving

$$\min_{\{\boldsymbol{w}_j \in \mathbb{R}^d\}_{j=1}^C} \sum_{j=1}^{C} \sum_{i \in I_j} |y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i|^p + \lambda \|\Omega(\boldsymbol{f})\|_q. \tag{25}$$

until Convergence

return the estimated parameter vectors $\boldsymbol{w}_j$, $j = 1, \dots, C$, and mode sequence $\hat{\boldsymbol{r}}$.

4. Numerical experiments

We now turn to an experimental evaluation of the model selection procedure of Algorithm 1, implemented in Matlab, for the estimation of a switched dynamical system of the form (1)–(2) with $C = 6$ modes and orders $n_a = n_b = 2$. Data sets of $n = 300\,000$ points are generated with this system for random parameters $\boldsymbol{w}_j \in [-1, 1]^d$, a zero-mean Gaussian input u_i of unit variance and a uniform noise e_i with an SNR of 30 dB.

Algorithm 1 is applied with $\overline{C} = 10$ and $(p, q) = (1, 2)$. In Algorithm 2, problem (25) is solved using the iterative scheme proposed in [18]. Over 100 trials, the true number of modes $C = 6$ was successfully recovered in all cases.

The procedure has also been confronted to a zero-mean Gaussian noise with an SNR of 30 dB. In this case, Y is not bounded and violates our assumption. However, the estimated number of modes $\hat{C}$ coincides with the true $C = 6$ in 95% of the cases, as shown by the histogram in Figure 3.

5. Conclusion

This paper presented a statistical learning perspective to switched system identification. This provided new error bounds for system identification, that take into account both the non-i.i.d nature of the data inherent to dynamical systems and the various complexities of regularized hybrid models. Building on these error bounds, a novel model selection method for the estimation of the number of modes of switched systems was proposed. The various forms of regularization considered also led to new extensions of the k-LinReg algorithm. Numerical experiments showed promising results regarding the estimation

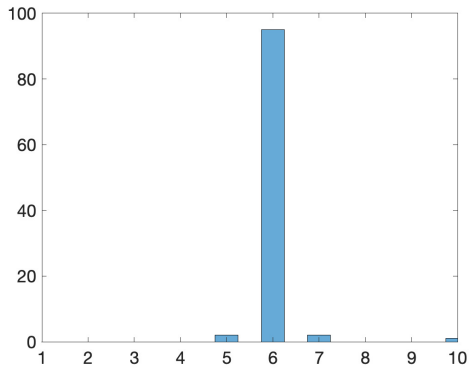

Figure 3: Histogram of $\hat{C}$ using Algorithm 1 for a switched dynamical system with $C = 6$ modes under Gaussian noise.

of the number of modes. Additional experimental results can be found in [17, 18].

Future work will study the mixing condition in the proposed theorems. Indeed, the estimation of the mixing coefficient $\beta(a)$ is needed to fully compute the bound and produce accuracy guarantees. Its value also influences the relevance of the bound, due to the constraint on the confidence index δ. Alternatively, other methods that do not require the mixing condition to derive bounds could be explored, as in [23]. Another research direction concerns the investigation of the sparsity induced by the ℓ_1-norm regularization ($q = 1$), 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 [25] will be investigated .

References

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

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

[3] 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.

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

[5] Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

[6] V. De Iuliis, F. Smarra, C. Manes, and A D'Innocenzo. On the stability of switched ARX models, with an application to learning via regression trees. *Proc. of the 7th IFAC Conference on Analysis and Design of Hybrid Systems ADHS*, 54(5):61–66, 2021.

[7] S. Hojjatinia, C. M. Lagoa, and F. Dabbene. Identification of switched autoregressive exogenous systems from large noisy datasets: Identification of switched autoregressive exogenous systems from large noisy datasets. *International Journal of Robust and Nonlinear Control*, 2020.

[8] 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.

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

[10] F. Lauer. Error bounds for piecewise smooth and switching regression. *IEEE Transactions on Neural Networks and Learning Systems*, 31(4):1183–1195, 2020.

[11] F. Lauer. Risk bounds for learning multiple components with permutation-invariant losses. In *Proc. of the 23th Int. Conf. on Artificial Intelligence and Statistics (AISTATS), Virtual Meeting*, 2020.

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

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

[14] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[15] A. Madary, H. R. Momeni, A. Abate, and K. G. Larsen. A Bayesian framework for large-scale identification of nonlinear hybrid systems. *Proc. of the 7th IFAC Conference on Analysis and Design of Hybrid Systems ADHS*, 54(5):259–264, 2021.

[16] L. Massucci, F. Lauer, and M. Gilson. Structural risk minimization for switched system identification. In *Proc. of the 59th IEEE Conference on Decision and Control*, 2020.

[17] L. Massucci, F. Lauer, and M. Gilson. How statistical learning can help to estimate the number of modes in switched system identification? In *Proc. of the 19th IFAC Symposium on system identification, Virtual Meeting*, 2021.

[18] L. Massucci, F. Lauer, and M. Gilson. Regularized switched system identification: a statistical learning perspective. In *Proc. of the 7th IFAC Conference on Analysis and Design of Hybrid Systems, Virtual Meeting*, 2021.

[19] 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.

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

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

[22] 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.

[23] M. Simchowitz, H. Mania, S. Tu, M.I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Proceedings of the Conference on Computational Learning Theory (COLT), Stockholm, Sweden*, 2018.

[24] F. Smarra, G. D. Di Girolamo, V. De Iuliis, A. Jain, R. Mangharam, and A. D'Innocenzo. Data-driven switching modeling for MPC using regression trees and random forests. *Nonlinear Analysis: Hybrid Systems*, 36:100882, 2020.

[25] T. Söderström and P. G. Stoica. Instrumental variable methods for system identification. *Lecture notes in control and information sciences*, 57, 1983.

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

[27] 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.

[28] M. Vidyasagar. *Learning and generalisation: with applications to neural networks*. Springer, 2013.

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

Appendix A. Basic results from the literature

We here recall the contraction principle for Rademacher complexities and a standard result for linear models.

Theorem 5. *(Contraction principle [14]) If the function* $\phi : \mathbb{R} \to \mathbb{R}$ *is* L_ϕ*-Lipshitz, i.e.,* $\forall (u, v) \in \mathbb{R}^2$, $|\phi(u) - \phi(v)| \leq L_\phi |u - v|$, *then,*

$$\hat{\mathcal{R}}_{\mathbf{Z}_n}(\{\phi \circ f : f \in \mathcal{F}\}) \leq L_\phi \hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{F}).$$

Theorem 6. *(Rademacher complexity of linear models [2]) Let $\mathcal{F}$ be a class of linear functions $\mathcal{F} = \{f : f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x},\ \|\boldsymbol{w}\|_2 \leq \Lambda\}$. Then,*

$$\hat{\mathcal{R}}_{\mathbf{X}_n}(\mathcal{F}) \leq \frac{\Lambda}{n}\sqrt{\sum_{i=1}^{n}\|\mathbf{X}_i\|_2^2}.$$

Appendix B. Proof of Theorem 3

Theorem 3 will result from an analysis of the Rademacher complexity of switching models inspired by [11] and the extension of learning theory results to system identification via Theorem 2. First, note that the loss (3) is permutation-invariant in the sense of [11], which means that we can focus on the ordered class

$$\tilde{\mathcal{F}} = \left\{\tilde{\boldsymbol{f}} : \boldsymbol{f} \in \mathcal{F}\right\},$$

where

$$\tilde{\boldsymbol{f}} = (f_{l(1)}, \ldots, f_{l(C)}),$$

in which $l(j)$ is the jth element of a permutation of $[C]$ that ensures

$$\|\boldsymbol{w}_{l(1)}\|_2 \geq \|\boldsymbol{w}_{l(2)}\|_2 \cdots \geq \|\boldsymbol{w}_{l(C)}\|_2.$$

Then, applying Lemma 2 in [11], we obtain

$$\tilde{\mathcal{F}} \subseteq \Pi_q = \prod_{j=1}^{C}\left\{f_j : f_j(\boldsymbol{x}) = \boldsymbol{w}_j^T\boldsymbol{x}, \|\boldsymbol{w}_j\|_2 \leq j^{-\frac{1}{q}}\Lambda\right\}.$$

Thus $\mathcal{L}_{\tilde{\mathcal{F}}} \subseteq \mathcal{L}_{\Pi_q}$, which further yields

$$\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}_{\tilde{\mathcal{F}}}) \leq \hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}_{\Pi_q}).$$

Since Π_q is a product of independent component classes, the decomposition result in Theorem 3 of [11] then gives

$$\hat{\mathcal{R}}_{\mathbf{Z}_n}(\mathcal{L}_{\tilde{\mathcal{F}}}) \leq 2\sum_{j=1}^{C}\hat{\mathcal{R}}_{\mathbf{X}_n}\left(\left\{f_j : f_j(\boldsymbol{x}) = \boldsymbol{w}_j^T\boldsymbol{x}, \|\boldsymbol{w}_j\|_2 \leq j^{-\frac{1}{q}}\Lambda\right\}\right).$$

Finally, using Theorem 6 in Appendix A and the fact that $\sum_{j=1}^{C} j^{-\frac{1}{q}} \leq \alpha(C,q)$ for all $C \geq 2$ and $q \in (0,\infty]$ completes the proof.

Appendix C. Proof of Theorem 4

Let $\mathcal{F}_k$ be defined as in (17) with Λ_k instead of Λ. For any fixed C, Λ_k and $\delta_0' > 0$, let $\delta_0 = \delta_0' + 4(\mu-1)\beta(a)$ and

$$\epsilon(C,\Lambda_k,\delta_0') = \frac{2p(2M)^{p-1}\alpha(C,q)\Lambda_k\sqrt{\sum_{i=1}^{\mu}\|\boldsymbol{X}_{2a(i-1)+1}\|_2^2}}{\mu} + 3(2M)^p\sqrt{\frac{\log\frac{4}{\delta_0'}}{2\mu}}.$$

Then, Theorem 3 gives

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

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

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

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

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

with

$$\epsilon(C,\Lambda_k,\delta_0') = \frac{2p(2M)^{p-1}\alpha(C,q)\Lambda_k\sqrt{\sum_{i=1}^{\mu}\|\boldsymbol{X}_{2a(i-1)+1}\|_2^2}}{\mu} + 3(2M)^p\sqrt{\frac{\log(\overline{C}K) + \log\frac{4}{\delta'}}{2\mu}}.$$

For any $C \in [\overline{C}]$, we can rewrite $\mathcal{F}$ in (17) with constraint $\|\Omega(\boldsymbol{f})\|_q \leq \Lambda$ as $\mathcal{F} = \bigcup_{k\in[K]}\mathcal{F}_k$; and any $\boldsymbol{f} \in \mathcal{F}$ also belongs to $\mathcal{F}_k$ with, by (20), k such that $\Lambda_k = \tilde{\Lambda}(f)$. Thus, (C.1) is equivalent to (21) and the statement is proved.