HAL
open science

Error Bounds for Piecewise Smooth and Switching Regression

Fabien Lauer

▶ **To cite this version:**

Fabien Lauer. Error Bounds for Piecewise Smooth and Switching Regression. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31 (4), pp.1183-1195. 10.1109/TNNLS.2019.2919444 . hal-01566136v3

HAL Id: hal-01566136

https://hal.science/hal-01566136v3

Submitted on 27 May 2019

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.

Error Bounds for Piecewise Smooth and Switching Regression

Fabien Lauer

***Abstract*—The paper deals with regression problems, in which the nonsmooth target is assumed to switch between different operating modes. Specifically, piecewise smooth (PWS) regression considers target functions switching deterministically via a partition of the input space, while switching regression considers arbitrary switching laws. The paper derives generalization error bounds in these two settings by following the approach based on Rademacher complexities. For PWS regression, our derivation involves a chaining argument and a decomposition of the covering numbers of PWS classes in terms of the ones of their component function classes and the capacity of the classifier partitioning the input space. This yields error bounds with a radical dependency on the number of modes. For switching regression, the decomposition can be performed directly at the level of the Rademacher complexities, which yields bounds with a linear dependency on the number of modes. By using once more chaining and a decomposition at the level of covering numbers, we show how to recover a radical dependency. Examples of applications are given in particular for PWS and swichting regression with linear and kernel-based component functions.**

***Index Terms*—Learning theory, guaranteed risk, regression, Rademacher complexity, covering number, chaining.**

I. Introduction

The paper deals with regression problems, in which the nonsmooth target is assumed to switch between different operating modes. Specifically, we focus on two different (but related) settings: piecewise smooth (PWS) regression and switching regression. In PWS regression, the target function is assumed to switch between modes deterministically via a partition of the input space, while in switching regression the switchings can be arbitrary.

Switching regression was introduced by [1] and early algorithms include the one of [2] and the expectation-maximization methods of [3], [4], [5]. Regression trees [6], and subsequent improvements [7], [8], are well-known early examples of piecewise regression models, together with the mixtures of experts [9], which however usually consider smooth switchings. More recently, most of the work in this field was produced by the control community for hybrid dynamical system identification [10]. Hybrid systems are dynamical systems that switch between different operating modes and their identification corresponds to the estimation of a model from input–output data without knowledge of the mode that generated each data point or the switching law. This raises many practical difficulties related to the intertwining of a regression task (the estimation of the component functions modeling the system behavior in each mode) and a classification one (partitioning the data into modes). Work in this field focuses mostly on the corresponding optimization issues [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] and the analysis of algorithmic complexity [23], [24].

This activity produced a number of practical methods for minimizing the empirical error of switching models with various optimization accuracy and computational efficiency. However, few results are available in terms of statistical guarantees for the obtained models [25], [16], [26], [27], and most of them are established in a parametric estimation framework or under restrictive conditions on the target function. Instead, we here aim at obtaining generalization error bounds for switching models in the agnostic learning framework [28] including the nonparametric setting and without particular assumptions on the target. For this purpose, we will use the tools of statistical learning and follow a standard approach to derive error bounds based on Rademacher complexities [29], [30], [31]. While doing so, we particularly pay attention to the dependency of the obtained bound on the number of modes. In this respect, our work is related to recent discussions on the dependency of error bounds for margin multi-category classifiers on the number of categories, see, e.g., [32], [33], [34]. As in these works, a crucial role will be played by the decomposition of a global capacity measure as a function of capacities of component function classes.

Specifically, we bound the Rademacher complexity of PWS classes using the chaining method [35] and covering numbers. We propose a decomposition scheme to express the covering numbers of a PWS class in terms of those of its component function classes and the capacity of the classifier defining the partition of the input space. For a large set of PWS classes, this results in error bounds with a radical dependency on the number of modes and efficient convergence rates when compared to the results of [33], [34] in multi-category classification. For switching regression, we follow a similar path but also consider a more straightforward approach, in which we apply the decomposition at the level of the Rademacher complexities themselves, without invoking covering numbers. A comparison of the two approaches shows that decomposing at the level of covering numbers is more advantageous with respect to the number of modes, with however a slightly worse rate of convergence with respect to the sample size for kernel-based classes.

Paper organization: Section II formally exposes the two considered settings: PWS regression in Sect. II-A and switching regression in II-B. Then, Section III derives error bounds for the PWS case in subsection III-A and for switching regression in subsection III-B. Section IV concludes the paper. Throughout the paper, a number of technical results are retained in Appendix, more precisely, in App. A for those

F. Lauer is with the Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France, e-mail: (see https://members.loria.fr/FLauer/).

from the literature and in App. B–F for newly derived ones.

Notation: For an integer n, $[n]$ denotes the set of integers from 1 to n. A bold letter with a subscript, e.g., $\boldsymbol{t}_n$, denotes a sequence, $(t_i)_{1\leq i\leq n}$. Given two sets $\mathcal{X}$ and $\mathcal{Y}$, the set of functions from $\mathcal{X}$ into $\mathcal{Y}$ is written as $\mathcal{Y}^{\mathcal{X}}$.

II. Theoretical framework

Let $\mathcal{X}$ denote the input space and let the output space be $\mathcal{Y} = [-M, M]$ for some $M > 0$. We assume that the relationship between inputs and outputs is characterized by the probability distribution of the random pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ and further assume that this distribution is unknown. Given a realization of the sample $((X_i, Y_i))_{1\leq i\leq n}$ of n independent copies of (X, Y), the aim of regression is to learn the model f that minimizes, over a certain function class to be defined below, the (expected) risk. In this paper, we define the risk from loss functions that can be clipped at M.

Definition 1 (Clipping)**.** *For any $M > 0$ and $t \in \mathbb{R}$, we define the clipped version $\bar{t}$ of t as*

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

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

Recall from [36] that a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$ can be clipped at M when, for all $(y, t) \in \mathcal{Y} \times \mathbb{R}$,

$$\ell(y, \bar{t}) \leq \ell(y, t).$$

A. PWS regression

For regression, the most common choice of loss function is the squared loss, $\ell_2(y, t) = (y - t)^2$. However, our results will hold more generally for any ℓ_p-loss, $\ell_p(y, t) = |y - t|^p$, with $p \in [1, \infty)$, and the corresponding ℓ_p-risks.

Definition 2 (ℓ_p-risk and empirical ℓ_p-risk)**.** *For $p \in [1, \infty)$, the ℓ_p-risk of a function f from $\mathcal{X}$ into $\mathbb{R}$ is*

$$L_p(f) = \mathbb{E}_{X,Y}|Y - f(X)|^p$$

and the corresponding empirical ℓ_p-risk evaluated on an n-sample $((X_i, Y_i))_{1\leq i\leq n}$ is

$$\hat{L}_{p,n}(f) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - f(X_i)|^p. \tag{1}$$

Since the ℓ_p-losses can be clipped at M, the ℓ_p-risk of a clipped function, $L_p(\bar{f})$, is always smaller than the one of the unclipped f. The following thus considers that the final result of the learning procedure estimating f is $\bar{f}$ and derives bounds on the risk of $\bar{f}$.

In PWS regression, the empirical risk minimization (ERM) principle amounts to minimizing (1) for a given realization of the training sample $((X_i, Y_i))_{1\leq i\leq n}$ over a predefined class of piecewise smooth functions:

Definition 3 (PWS class)**.** *Given a sequence $(\mathcal{F}_k)_{1\leq k\leq C}$ of classes of component functions from $\mathcal{X}$ into $\mathbb{R}$ and a set of classifiers $\mathcal{G}$ from $\mathcal{X}$ into $[C]$, we define the PWS class of functions*

$$\mathcal{F}_{\mathcal{G}} = \big\{f \in \mathbb{R}^{\mathcal{X}} : f(x) = f_{g(x)}(x), \\ g \in \mathcal{G},\ (f_k)_{1\leq k\leq C} \in (\mathcal{F}_k)_{1\leq k\leq C}\big\}.$$

However, in most cases, such a minimization is an NP-hard problem [23] and practical algorithms can rarely be guaranteed to yield an empirical risk minimizer. Therefore, we will aim at obtaining uniform bounds on the ℓ_p-risk holding (with high probability) for all f in $\mathcal{F}_{\mathcal{G}}$ and not only for global minimizers of (1).

B. Switching regression

Switching regression differs from PWS regression in the assumptions made regarding the switchings in the data generating process. While PWS regression assumes that the switchings are a deterministic function of x, switching regression deals with arbitrary switchings.

In order to allow for such arbitrary switchings, we define classes of switching functions with vector-valued functions and embed the selection of the component function used to predict Y in the definition of the loss functions and risks.

Definition 4 (Switching class)**.** *Given a sequence $(\mathcal{F}_k)_{1\leq k\leq C}$ of classes of component functions from $\mathcal{X}$ into $\mathbb{R}$, we define the switching class of vector-valued functions from $\mathcal{X}$ into $\mathbb{R}^C$ as*

$$\mathcal{F}^S = \Big\{f : f(x) = (f_k(x))_{1\leq k\leq C},\ f_k \in \mathcal{F}_k,\ 1 \leq k \leq C\Big\}.$$

Definition 5 (ℓ_p-switching risks)**.** *For $p \in [1, \infty)$, the switching ℓ_p-risk of a vector-valued function f from $\mathcal{X}$ into $\mathbb{R}^C$ is*

$$L_p^S(f) = \mathbb{E}_{X,Y} \min_{k\in[C]} |Y - f_k(X)|^p$$

and the corresponding switching empirical ℓ_p-risk evaluated on an n-sample $((X_i, Y_i))_{1\leq i\leq n}$ is

$$\hat{L}_{p,n}^S(f) = \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[C]} |Y_i - f_k(X_i)|^p. \tag{2}$$

Again, clipped versions of f, i.e., $\bar{f} = (\bar{f}_k)_{1\leq k\leq C}$, will be used and it is easy to verify that the switching ℓ_p-losses are clippable in the sense that

$$\forall (y, t) \in \mathcal{Y} \times \mathbb{R}^C, \quad \min_{k\in[C]} |y - \bar{t}_k|^p \leq \min_{k\in[C]} |y - t_k|^p.$$

Such switching loss functions formalize the goal of accurately learning a collection of submodels so that, for all inputs, at least one submodel can predict the output well. Such a setting appears for instance in hybrid dynamical system identification [10] and a number of computer vision applications [37]. As for PWS regression, the ERM principle here corresponds to the minimization of (2) over $\mathcal{F}^S$, which is also an NP-hard problem in most cases [24]; hence the need for bounds holding uniformly over $\mathcal{F}^S$.

III. Error bounds

The following is dedicated to establishing upper bounds on the expected risks, $L_p(f)$ and $L_p^S(f)$, in terms of the empirical risks, $\hat{L}_{p,n}(f)$ and $\hat{L}_{p,n}^S(f)$, and a confidence (semi-)interval or control term depending on the function classes over which the bounds should hold uniformly.

We first introduce a general error bound based on the Rademacher complexity of the function class of interest.

Definition 6 (Rademacher complexity)**.** *For $n \in \mathbb{N}^*$, let $\boldsymbol{T}_n = (T_i)_{1\le i\le n}$ be an n-sample of independent copies of the random variable $T \in \mathcal{T}$, let $\boldsymbol{\sigma}_n = (\sigma_i)_{1\le i\le n}$ be a sequence of independent random variables uniformly distributed in $\{-1,+1\}$. Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. The* empirical Rademacher complexity *of $\mathcal{F}$ given $\boldsymbol{T}_n$ is*

$$\hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(T_i) \middle| \boldsymbol{T}_n\right].$$

The Rademacher complexity *of $\mathcal{F}$ is*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{T}_n}\hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{T}_n\boldsymbol{\sigma}_n}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(T_i).$$

Theorem 1 (After, e.g., Theorem 3.1 in [31])**.** *Let $\mathcal{L}$ be a class of functions from $\mathcal{Z}$ into $[0,1]$ and $(Z_i)_{1\le i\le n}$ be a sequence of independent copies of the random variable $Z \in \mathcal{Z}$. Then, for a fixed $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over all $\ell \in \mathcal{L}$,*

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

In the remaining of the paper, we assume without loss of generality that $Y \in [-M, M]$ with $M = \frac{1}{2}$, since otherwise we can recover this setting by rescaling Y. This choice is made in order to guarantee that the ℓ_p-losses remain bounded by 1 and that the corresponding function classes satisfy the assumptions of Theorem 1.

A. Error bounds for PWS classes

Our derivation of error bounds for PWS classes starts with the following consequence of Theorem 1, whose proof is given in Appendix B.

Theorem 2. *Let $\mathcal{F}$ be a real-valued function class. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the ℓ_p-risk of Definition 2 is bounded uniformly $\forall \bar{f} \in \bar{\mathcal{F}}$ as*

$$L_p(\bar{f}) \le \hat{L}_{p,n}(\bar{f}) + 2p\mathcal{R}_n\left(\bar{\mathcal{F}}\right) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Then, it remains to bound the Rademacher complexity of the clipped PWS class $\bar{\mathcal{F}}_{\mathcal{G}}$ which can be expressed from the clipped $\bar{\mathcal{F}}_k$'s as in Definition 3.

For this purpose, we will apply the chaining method [35] and introduce other capacity measures: the covering numbers.

Definition 7 (Pseudo-metric)**.** *Given a sequence $\boldsymbol{t}_n \in \mathcal{T}^n$, $d_{q,\boldsymbol{t}_n}$ is the empirical pseudo-metric defined $\forall (f,f') \in \left(\mathbb{R}^{\mathcal{T}}\right)^2$ and $q \in [1,\infty)$ by*

$$d_{q,\boldsymbol{t}_n}(f,f') = \left(\frac{1}{n}\sum_{i=1}^n |f(t_i) - f'(t_i)|^q\right)^{\frac{1}{q}}$$

and for $q = \infty$ by

$$d_{\infty,\boldsymbol{t}_n}(f,f') = \max_{i\in[n]} |f(t_i) - f'(t_i)|.$$

Definition 8 (Covering numbers)**.** *Given a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{T}}$ and a (pseudo-)metric ρ over $\mathbb{R}^{\mathcal{T}}$, the* covering number *$\mathcal{N}(\epsilon,\mathcal{F},\rho)$ at scale ϵ of $\mathcal{F}$ for the distance ρ is the smallest cardinality of the proper ϵ-net $\mathcal{H} \subseteq \mathcal{F}$ of $\mathcal{F}$ such that $\forall f \in \mathcal{F}$, $\rho(f,\mathcal{H}) < \epsilon$.* Uniform covering numbers *are defined for all pseudo-metrics as in Definition 7 by*

$$\mathcal{N}_q(\epsilon,\mathcal{F},n) = \sup_{\boldsymbol{t}_n\in\mathcal{T}^n} \mathcal{N}(\epsilon,\mathcal{F},d_{q,\boldsymbol{t}_n}).$$

By considering covering numbers at different scales, chaining allows one to bound the Rademacher complexity of $\bar{\mathcal{F}}_{\mathcal{G}}$ whose diameter is $2M = 1$ as follows (see Theorem 4 in Appendix A): for any $N \in \mathbb{N}^*$,

$$\hat{\mathcal{R}}_n(\bar{\mathcal{F}}_{\mathcal{G}}) \le 2^{-N} + 6\sum_{j=1}^N 2^{-j}\sqrt{\frac{\log\mathcal{N}(2^{-j},\bar{\mathcal{F}}_{\mathcal{G}},d_{2,\boldsymbol{x}_n})}{n}}. \quad (3)$$

The task is now to bound the covering numbers of the function class $\bar{\mathcal{F}}_{\mathcal{G}}$. This is done below by decomposing them in terms of the ones of the component function classes $\bar{\mathcal{F}}_k$ on the one hand and of the capacity of the classifier $\mathcal{G}$ on the other hand. In particular, we will measure the capacity of $\mathcal{G}$ with the growth function.

Definition 9 (Trace and growth function)**.** *Let $\mathcal{G}$ be a set of classifiers from $\mathcal{X}$ to $[C]$. The* trace *of $\mathcal{G}$ on a set $\boldsymbol{x}_n \in \mathcal{X}^n$ is the set*

$$\mathcal{G}_{\boldsymbol{x}_n} = \{(g(x_1),\dots,g(x_n)) : g \in \mathcal{G}\} \subseteq [C]^n$$

and the growth function *of $\mathcal{G}$ is defined by*

$$\forall n \in \mathbb{N}, \quad \Pi_{\mathcal{G}}(n) = \sup_{\boldsymbol{x}_n\in\mathcal{X}^n} |\mathcal{G}_{\boldsymbol{x}_n}|.$$

1) Decomposition of the covering numbers: The following gives two results based on two different techniques to optimize the dependency of the decomposition on the number of component functions (or modes), C. Note that these results are stated in terms of the clipped classes, but can be proved similarly for the unclipped ones.

Lemma 1. *Given a PWS class $\mathcal{F}_{\mathcal{G}}$ as in Definition 3, we have, for any $q \in [1,\infty) \cup \{\infty\}$,*

$$\mathcal{N}(\epsilon,\bar{\mathcal{F}}_{\mathcal{G}},d_{q,\boldsymbol{x}_n}) \le \Pi_{\mathcal{G}}(n)\prod_{k=1}^C \mathcal{N}\left(\frac{\epsilon}{C^{1/q}},\bar{\mathcal{F}}_k,d_{q,\boldsymbol{x}_n}\right).$$

Lemma 1, proved in App. C, provides a bound on covering numbers in L_q-norm, which is most advantageous with respect

to the dependency on C for $q = \infty$. Covering numbers in L_∞-norm can be used in chaining thanks to the following easy to verify inequality:

$$\mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{2,\boldsymbol{x}_n}) \leq \mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{\infty,\boldsymbol{x}_n}). \tag{4}$$

However, this bound can be crude and not optimal in terms of the sample size n. Furthermore, Sauer-Shelah lemmas used to bound the covering numbers of the component classes $\bar{\mathcal{F}}_k$ can typically be made independent of n for $q = 2$ but not for $q = \infty$ (see Lemma 8 below). Nonetheless, in some cases as emphasized in [34], the relationship (4) can be sufficient to obtain a good dependency on both C and n in the final chained bound.

For comparison, the following lemma provides a bound with the same dependency on C than the one in Lemma 1 for $q = \infty$ while relying solely on L_2-norm covering numbers. Its proof, detailed in App. D, uses a slightly different technique based on the introduction of a collection of pseudo-metrics, all derived from the L_2-norm but based on different samples. The other ingredient is the non-increasing nature of uniform covering numbers of Glivenko-Cantelli (GC) classes [38] with respect to n, proved in App. E. Note that focusing on uniform GC classes is not very restrictive as this coincides with all learnable classes [39].

Lemma 2. *Given a PWS class $\mathcal{F}_\mathcal{G}$ as in Definition 3 with uniform GC classes $\bar{\mathcal{F}}_k$, $1 \leq k \leq C$, we have*

$$\mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{2,\boldsymbol{x}_n}) \leq \Pi_\mathcal{G}(n) \prod_{k=1}^{C} \mathcal{N}_2(\epsilon, \bar{\mathcal{F}}_k, n).$$

2) Metric entropy bounds: The decomposition results above readily yield general bounds on the metric entropy, $\log \mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{2,\boldsymbol{x}_n})$, of a PWS class $\bar{\mathcal{F}}_\mathcal{G}$ to be used in (3) and expressed in terms of Natarajan and fat-shattering dimensions.

Definition 10 (Natarajan dimension)**.** *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[C]$. A set $\{x_i\}_{i=1}^n \subset \mathcal{X}$ is said to be shattered by $\mathcal{G}$ if there exist two functions a and b from $\mathcal{X}$ into $[C]$ such that for every $i \in [n]$, $a(x_i) \neq b(x_i)$ and for every subset $I \subseteq [n]$, there is a function $g \in \mathcal{G}$ satsifying: $\forall i \in I$, $g(x_i) = a(x_i)$ and $\forall i \in [n] \setminus I$, $g(x_i) = b(x_i)$. The* Natarajan dimension *$d_\mathcal{G}$ of $\mathcal{G}$ is the maximal cardinality of a set $\{x_i\}_{i=1}^n \subset \mathcal{X}$ shattered by $\mathcal{G}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite Natarajan dimension.*

Definition 11 (Fat-shattering dimension [40])**.** *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. For $\epsilon > 0$, a set $\{x_i\}_{i=1}^n \subset \mathcal{X}$ is said to be* ϵ-shattered *by $\mathcal{F}$ if there is a witness $\boldsymbol{b}_n \in \mathbb{R}^n$ such that, for every subset $I \subseteq [n]$, there is a function $f \in \mathcal{F}$ satisfying: $\forall i \in I$, $f(x_i) \geq b_i + \epsilon$ and $\forall i \in [n] \setminus I$, $f(x_i) \leq b_i - \epsilon$. The* fat-shattering dimension with margin *ϵ of the class $\mathcal{F}$, $d_\mathcal{F}(\epsilon)$, is the maximal cardinality of a set $\{x_i\}_{i=1}^n \subset \mathcal{X}$ ϵ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite fat-shattering dimension with margin ϵ.*

In particular, our first decomposition result in Lemma 1 yields the following metric entropy bound.

Proposition 1 (PWS metric entropy bound 1)**.** *Given a PWS class $\mathcal{F}_\mathcal{G}$, let $d_\mathcal{G}$ denote the Natarajan dimension of $\mathcal{G}$ and $d_\mathcal{F}(\epsilon) = \max_{k\in[C]} d_{\bar{\mathcal{F}}_k}(\epsilon)$ denote the pointwise maximum of the fat-shattering dimensions of the $\bar{\mathcal{F}}_k$'s. Then, for any $\epsilon \in (0, 1]$ and $n \in \mathbb{N}^*$,*

$$\log \mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{2,\boldsymbol{x}_n}) \leq d_\mathcal{G} \log \frac{Cen}{2d_\mathcal{G}} + 6Cd_\mathcal{F}\left(\frac{\epsilon}{4}\right) \log^2 \frac{2en}{\epsilon}.$$

Proof. By application of (4) and Lemma 1 with $q = \infty$, we have

$$\log \mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{2,\boldsymbol{x}_n}) \leq \log \Pi_\mathcal{G}(n) + C \max_{k\in[C]} \log \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{\infty,\boldsymbol{x}_n}).$$

Then, we use two generalized Sauer-Shelah lemmas: Lemma 6 in App. A to bound the first term with the Natarajan dimension and the result from [39] (in the form of Lemma 7 with $M = 1/2$) to bound the last one in terms of the fat-shattering dimension. This yields

$$\max_{k\in[C]} \log \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{\infty,\boldsymbol{x}_n}) \leq 2d_\mathcal{F}\left(\frac{\epsilon}{4}\right) \log_2 \frac{2en}{d_\mathcal{F}(\frac{\epsilon}{4})\epsilon} \log \frac{4n}{\epsilon^2},$$

which, after using $d_\mathcal{F}(\epsilon/4) \geq 1$ and the relations $2/\log 2 < 3$ and $\sqrt{n} < en$, gives the result. If $d_\mathcal{F}(\epsilon/4) = 0$, the statement holds trivially since for all $k \in [C]$, $(f, f') \in \bar{\mathcal{F}}_k^2$ and $x \in \mathcal{X}$, $|f(x) - f'(x)|/2 < \epsilon/4$, which implies that $d_{\infty,\boldsymbol{x}_n}(f, f') < \epsilon/2$ and thus that $\mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{\infty,\boldsymbol{x}_n}) \leq \mathcal{N}(\frac{\epsilon}{2}, \bar{\mathcal{F}}_k, d_{\infty,\boldsymbol{x}_n}) = 1$. □

Conversely, our second decomposition result in Lemma 2 yields the following bound.

Proposition 2 (PWS metric entropy bound 2)**.** *Given a PWS class $\mathcal{F}_\mathcal{G}$ with uniform GC classes $\bar{\mathcal{F}}_k$, $1 \leq k \leq C$, let $d_\mathcal{G}$ denote the Natarajan dimension of $\mathcal{G}$ and $d_\mathcal{F}(\epsilon) = \max_{k\in[C]} d_{\bar{\mathcal{F}}_k}(\epsilon)$ denote the pointwise maximum of the fat-shattering dimensions of the $\bar{\mathcal{F}}_k$'s. Then, for any $\epsilon \in (0, 1]$ and $n \in \mathbb{N}^*$,*

$$\log \mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{2,\boldsymbol{x}_n}) \leq d_\mathcal{G} \log \frac{Cen}{2d_\mathcal{G}} + 20Cd_\mathcal{F}\left(\frac{\epsilon}{96}\right) \log \frac{7}{\epsilon}.$$

Proof. By application of Lemma 2, we have

$$\log \mathcal{N}(\epsilon, \bar{\mathcal{F}}_\mathcal{G}, d_{2,\boldsymbol{x}_n}) \leq \log \Pi_\mathcal{G}(n) + C \max_{k\in[C]} \log \mathcal{N}_2(\epsilon, \bar{\mathcal{F}}_k, n).$$

Then, we bound the first term as in Proposition 1 by Lemma 6 and the last one with the dimension-free Sauer-Shelah lemma from [41] (in the form of Lemma 8 with $M = 1/2$). □

These two bounds share the same dependency on C, but the first one in Proposition 1 will lead to a worse dependency on n when used in (3) for chaining. However, as discussed in [34], this is not due to the dimension-free nature of the bound in Proposition 2 (which is independent of n). Indeed, the dependency on n of the final chained bound is mostly impacted by how the metric entropy depends on $\frac{1}{\epsilon}$. Here, Proposition 2 exhibits a $O(\log \frac{1}{\epsilon})$ whereas the bound in Proposition 1 is in $O(\log^2 \frac{1}{\epsilon})$, which will translate into a $\sqrt{\log n}$ gain for the chained bound based on Proposition 2 (note however that due to the constants the true gain would be only visible in practice for very large n).

3) Applications: We now turn to specific examples of PWS classes. In particular, we focus on Euclidean input spaces $\mathcal{X} \subseteq \mathbb{R}^d$ with $d \geq 2$ and PWS classes constructed from a set of linear classifiers, i.e., for which

$$\mathcal{G} = \{g \in [C]^{\mathcal{X}} : g(x) = \operatorname*{argmax}_{k\in[C]} \langle w_k, x\rangle,\ w_k \in \mathbb{R}^d\}. \quad (5)$$

In this case, the Natarajan dimension of $\mathcal{G}$ satisfies $(C-1)(d-1) \leq d_{\mathcal{G}} \leq Cd$ (see, e.g., Corollary 29.8 in [42]) and Lemma 6 yields

$$\log \Pi_{\mathcal{G}}(n) \leq Cd \log\left(\frac{neC}{2(C-1)(d-1)}\right) \leq Cd\log(3n). \quad (6)$$

a) General PWS classes: We start with general PWS classes $\mathcal{F}_{\mathcal{G}}$ based on linear classifiers as in (5) and component function classes $\mathcal{F}_k$ that satisfy a polynomial growth assumption on the fat-shattering dimension, as considered, e.g., in [43], [33], [34]:

$$\forall \epsilon > 0, \quad d_{\mathcal{F}}(\epsilon) = \max_{k\in[C]} d_{\bar{\mathcal{F}}_k}(\epsilon) \leq \alpha\epsilon^{-\beta} \quad (7)$$

for some positive numbers α and β. For example, if the $\mathcal{F}_k$'s are implemented by neural networks with l hidden layers, (7) can be satisfied with $\beta = 2(l+1)$ [44]. Note that (7) implies that the $\bar{\mathcal{F}}_k$'s are uniform GC classes [39].

Using the general metric entropy bound of Proposition 2 with (6) and the assumption (7) in (3) leads to

$$\hat{\mathcal{R}}_n(\bar{\mathcal{F}}_{\mathcal{G}}) \leq 2^{-N} + \frac{6S_N}{\sqrt{n}} \quad (8)$$

with

$$S_N = \sum_{j=1}^{N} 2^{-j}\sqrt{d_{\mathcal{G}}\log\frac{Cen}{2d_{\mathcal{G}}} + 20Cd_{\mathcal{F}}\left(\frac{2^{-j}}{96}\right)\log(7\cdot 2^j)}$$
$$\leq \sqrt{C}\sum_{j=1}^{N} 2^{-j}\sqrt{d\log(3n) + 20\cdot 96^{\beta}\alpha 2^{j\beta}\log(2^{j+3})}.$$

Therefore, the dependency on the number of modes C is radical for all such PWS classes and the degree β of the polynomial growth of the fat-shattering dimensions only influences the convergence rate in n of the Rademacher complexity. This convergence rate can be specified as follows:

$$S_N \leq 2^{\frac{\beta}{2}}\sqrt{C}\sum_{j=1}^{N} 2^{j(\frac{\beta}{2}-1)}\sqrt{\frac{d\log(3n)}{2^{j\beta}} + 20\cdot 96^{\beta}\alpha\log(2^{j+3})}$$
$$\leq 2^{\frac{\beta}{2}}\sqrt{C}\sum_{j=1}^{N} 2^{j(\frac{\beta}{2}-1)}\sqrt{\frac{d\log(3n)}{2^{\beta}} + 14\cdot 96^{\beta}\alpha(j+3)}$$
$$\leq 2^{\frac{\beta}{2}}\sqrt{C}\sum_{j=1}^{N} 2^{j(\frac{\beta}{2}-1)}\sqrt{\frac{d\log(3n)}{2^{\beta}} + 56\cdot 96^{\beta}\alpha N}.$$

By setting $N = \lceil \log_2 n^{\frac{1}{\beta}} \rceil \leq \frac{1}{\beta}\log_2(2^{\beta}n)$, this gives, for all $\beta \geq 1$,

$$S_N \leq \sqrt{C\left[d + \frac{56\cdot 192^{\beta}\alpha}{\beta}\right]\log_2(2^{\beta}n)}\sum_{j=1}^{N} 2^{j(\frac{\beta}{2}-1)},$$

while $2^{-N} \leq n^{-1/\beta}$ for the first term of (8). Thus, for $\beta = 2$, we obtain

$$S_N < \sqrt{C\left[d + 28\cdot 192^2\alpha\right]\log_2(4n)}\frac{1}{2}\log_2(4n)$$

and, with respect to n,

$$\hat{\mathcal{R}}_n(\bar{\mathcal{F}}_{\mathcal{G}}) = O\left(\frac{\log^{\frac{3}{2}} n}{\sqrt{n}}\right). \quad (9)$$

For $\beta > 2$, we have

$$\sum_{j=1}^{N} 2^{j(\frac{\beta}{2}-1)} = \frac{2^{(\frac{\beta}{2}-1)(N+1)} - 2^{(\frac{\beta}{2}-1)}}{2^{(\frac{\beta}{2}-1)} - 1} < \frac{2^{(\frac{\beta}{2}-1)(N+1)}}{2^{(\frac{\beta}{2}-1)} - 1}$$
$$< \frac{4^{(\frac{\beta}{2}-1)}}{2^{(\frac{\beta}{2}-1)} - 1} n^{(\frac{1}{2}-\frac{1}{\beta})}, \quad (10)$$

which gives

$$\hat{\mathcal{R}}_n(\bar{\mathcal{F}}_{\mathcal{G}}) = O\left(\frac{\sqrt{\log n}}{n^{\frac{1}{\beta}}}\right). \quad (11)$$

Overall, the convergence rates obtained are similar to the ones derived in [33] for multi-category classification based on score function classes satisfying (7). However, thanks to Lemma 2, the radical dependency on C is more favorable than the dependency on the number of categories in [33].

b) Kernel-based PWS classes: Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of reproducing kernel K [45] and consider PWS classes with component function classes from this RKHS:

$$\mathcal{F}_k = \{f_k \in \mathcal{H} : \|f_k\|_{\mathcal{H}} \leq R_{\mathcal{H}}\}. \quad (12)$$

Since the covering numbers and the fat-shattering dimensions can only decrease when going from $\mathcal{F}_k$ to $\bar{\mathcal{F}}_k$, the bounds on $d_{\mathcal{F}_k}(\epsilon)$ given by Lemma 9 in App. A with $\phi : x \mapsto K(x,\cdot)$ also apply to the clipped component function classes and we have

$$\forall \epsilon > 0, \quad d_{\mathcal{F}}(\epsilon) \leq R_x^2 R_{\mathcal{H}}^2 \epsilon^{-2},$$

where $R_x = \sup_{x\in\mathcal{X}} \|K(x,\cdot)\|_{\mathcal{H}} = \sup_{x\in\mathcal{X}}\sqrt{K(x,x)}$. Thus, for $\mathcal{G}$ as in (5), the results above are applicable with $\alpha = R_x^2 R_{\mathcal{H}}^2$ and $\beta = 2$. This yields

$$\hat{\mathcal{R}}_n(\bar{\mathcal{F}}_{\mathcal{G}}) \leq \frac{1}{\sqrt{n}} + 3\log_2^{\frac{3}{2}}(4n)\sqrt{\frac{C}{n}\left(d + 28\cdot 192^2 R_x^2 R_{\mathcal{H}}^2\right)}$$

and a radical dependency on C with a convergence rate in $O\left(\log^{3/2}(n)/\sqrt{n}\right)$.

c) PWA classes: Let $\mathcal{F}_{\mathcal{G}}$ be a piecewise affine (PWA) class corresponding to Definition 3 with $\mathcal{G}$ as in (5) and linear function classes

$$\mathcal{F}_k = \{f_k \in \mathbb{R}^{\mathcal{X}} : f_k(x) = \langle w_k, x\rangle,\ \|w_k\|_2 \leq R_w\}. \quad (13)$$

The general results above could be applied similarly to this case (with $\phi : x \mapsto x$ in Lemma 9) in order to yield a bound in the flavor of the one obtained for kernel-based PWS classes. A slightly more efficient approach uses estimates of covering numbers that do not involve the fat-shattering dimension, as those given by Theorem 3 in [46]. Thus, instead of the metric entropy bounds of the previous subsection, one could use

Lemma 2 with such estimates in the chaining formula (3). Yet, the convergence rate would remain in $O\left(\log^{3/2}(n)/\sqrt{n}\right)$.

In fact, the metric entropy bound of [46] is suitable for large dimensional cases as it only involves the logarithm of the input dimension d. Yet, since in our case the dimension already appears outside of log terms when bounding $\log \Pi_{\mathcal{G}}(n)$ by (6), we can use more simple results that depend linearly on d, but enjoy a much better dependence on ϵ. In particular, the following can be easily derived from classical results on the covering of unit balls in $\mathbb{R}^d$ (see, e.g., Exercise 2.2.14 in [35]):

$$\forall \epsilon \leq R_x R_w, \quad \log \mathcal{N}_\infty(\epsilon, \mathcal{F}_k, n) \leq d \log \frac{(2+R_w)R_x}{\epsilon}. \tag{14}$$

Using this in (3) with (4), Lemma 1 and (6), we obtain

$$\begin{aligned}\hat{\mathcal{R}}_n(\bar{\mathcal{F}}_{\mathcal{G}}) &\leq 2^{-N} + 6\sqrt{\frac{Cd}{n}} \sum_{j=1}^{N} 2^{-j} \sqrt{\log\left(3n(2+R_w)R_x 2^j\right)} \\ &< 2^{-N} + 6\sqrt{\frac{Cd}{n} \log\left(3n(2+R_w)R_x 2^N\right)}.\end{aligned}$$

Setting $N = \lceil \log_2 \sqrt{n} \rceil \leq \log_2(2\sqrt{n})$ then yields an improved convergence rate:

$$\begin{aligned}\hat{\mathcal{R}}_n(\bar{\mathcal{F}}_{\mathcal{G}}) &< \frac{1}{\sqrt{n}} + 6\sqrt{\frac{Cd}{n} \log\left(6(2+R_w)R_x n^{3/2}\right)} \\ &= O\left(\sqrt{\frac{\log n}{n}}\right).\end{aligned}$$

B. Error bounds for switching regression

We now come back to the switching regression setting of Sect. II-B. Here, the focus is on the switching loss class of functions from $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$,

$$\mathcal{L}^S_{p,\mathcal{F}^S} = \{\ell \in [0,1]^{\mathcal{Z}} : \ell(x,y) = \min_{k\in[C]} |y - \bar{f}_k(x)|^p,\ \bar{f} \in \bar{\mathcal{F}}^S\}, \tag{15}$$

induced by the vector-valued function class $\mathcal{F}^S$ (Def. 4). Indeed, Theorem 1 applied to $\mathcal{L}^S_{p,\mathcal{F}^S}$ yields, with probability at least $1-\delta$ and uniformly over $\bar{\mathcal{F}}^S$, the following bound on the ℓ_p-switching risk of Definition 5:

$$L^S_p(\bar{f}) \leq \hat{L}^S_{p,n}(\bar{f}) + 2\mathcal{R}_n\left(\mathcal{L}^S_{p,\mathcal{F}^S}\right) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{16}$$

To finalize the bound, it remains to estimate the Rademacher complexity $\mathcal{R}_n\left(\mathcal{L}^S_{p,\mathcal{F}^S}\right)$. For this purpose, we consider two decomposition schemes: one at the level of Rademacher complexities and another one at the level of covering numbers.

1) Decomposition of the Rademacher complexity: In the case of switching regression, we can decompose directly at the level of the Rademacher complexities, without requiring the chaining machinery and covering numbers. In particular, we can derive the following decomposition result relating the Rademacher complexity of $\mathcal{L}^S_{p,\mathcal{F}^S}$ to the ones of the component function classes $\mathcal{F}_k$ with a linear dependency on the number C of modes.

Theorem 3. *Let $\mathcal{F}^S$ be a vector-valued function class as in Definition 4. Then, the Rademacher complexity of $\mathcal{L}^S_{p,\mathcal{F}^S}$ (15) is bounded by*

$$\mathcal{R}_n\left(\mathcal{L}^S_{p,\mathcal{F}^S}\right) \leq p \sum_{k=1}^{C} \mathcal{R}_n\left(\mathcal{F}_k\right).$$

Proof. Let us define the following classes of functions from $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$:

$$\forall k \in [C], \quad \mathcal{E}_k = \{e_k \in \mathbb{R}^{\mathcal{Z}} : e_k(x,y) = y - \bar{f}_k(x),\ \bar{f}_k \in \bar{\mathcal{F}}_k\}. \tag{17}$$

By using the facts that for any $(a_k)_{1\leq k\leq C} \in \mathbb{R}^C$, $\min_{k\in[C]} a_k = -\max_{k\in[C]} -a_k$ and that σ_i and $-\sigma_i$ share the same distribution, we have:

$$\begin{aligned}\mathcal{R}_n\left(\mathcal{L}^S_{p,\mathcal{F}^S}\right) &= \mathbb{E}_{\boldsymbol{X}_n \boldsymbol{Y}_n \boldsymbol{\sigma}_n} \sup_{\bar{f}\in\bar{\mathcal{F}}^S} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \min_{k\in[C]} |Y_i - \bar{f}_k(X_i)|^p \\ &= \mathbb{E}_{\boldsymbol{X}_n \boldsymbol{Y}_n \boldsymbol{\sigma}_n} \sup_{\bar{f}\in\bar{\mathcal{F}}^S} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i \max_{k\in[C]} -|Y_i - \bar{f}_k(X_i)|^p \\ &= \mathbb{E}_{\boldsymbol{X}_n \boldsymbol{Y}_n \boldsymbol{\sigma}_n} \sup_{(e_k\in\mathcal{E}_k)_{k\in[C]}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \max_{k\in[C]} -|e_k(X_i,Y_i)|^p \\ &\leq \sum_{k=1}^{C} \mathcal{R}_n(-|\mathcal{E}_k|^p),\end{aligned}$$

where the inequality is obtained by application of Lemma 5 in Appendix A. By taking into account the range of $|\mathcal{E}_k|$, i.e., $[0,1]$, and the Lipschitz constant of $\phi(u) = u^p$ for u in that interval, we obtain by contraction (Lemma 4) that $\mathcal{R}_n(-|\mathcal{E}_k|^p) \leq p\mathcal{R}_n(\mathcal{E}_k)$. Then, following the last steps of the proof of Theorem 2 (Appendix B) leads to

$$\mathcal{R}_n(\mathcal{E}_k) \leq \mathcal{R}_n(\bar{\mathcal{F}}_k) \leq \mathcal{R}_n(\mathcal{F}_k),$$

which completes the proof. □

For switching linear regression with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{F}_k$ set as in (13), we can combine (16) with Theorem 3 and Lemma 10 in App. A to get the risk bound: With probability at least $1-\delta$, for all $\bar{f} \in \bar{\mathcal{F}}^S$,

$$L^S_p(\bar{f}) \leq \hat{L}^S_{p,n}(\bar{f}) + 2pC\frac{R_x R_w}{\sqrt{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \tag{18}$$

For switching nonlinear regression based on component function classes from an RKHS $\mathcal{H}$ of reproducing kernel K as in (12), a similar result holds with $R_x = \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}$ and $R_w = R_{\mathcal{H}}$.

2) Chaining and decomposition of the covering numbers: In another context, namely, multi-category classification as studied in [32], [33], [34], the decomposition of the Rademacher complexity in terms of those of the component function classes yields a linear dependency on the number of categories, while chaining and decomposition at the level of covering numbers allows one to obtain a radical dependency. We now evaluate the possibility of reducing the linear dependency on the number of modes C of the bound in Theorem 3 with such an approach.

Consider the risk bound (16) based on the Rademacher complexity of the real-valued class $\mathcal{L}^S_{p,\mathcal{F}^S}$ defined in (15). We

can bound the covering numbers of this class thanks to the structural result of Lemma 13 in Appendix E as follows.

Lemma 3. *Let $\mathcal{F}^S$ be a vector-valued function class as in Definition 4. Then, for $\mathcal{L}^S_{p,\mathcal{F}^S}$ as defined in* (15)*, the following holds for any $q \in [1,\infty) \cup \{\infty\}$:*

$$\mathcal{N}(\epsilon, \mathcal{L}^S_{p,\mathcal{F}^S}, d_{q,\boldsymbol{z}_n}) \leq \prod_{k=1}^{C} \mathcal{N}\left(\frac{\epsilon}{pC^{1/q}}, \bar{\mathcal{F}}_k, d_{q,\boldsymbol{x}_n}\right).$$

Proof. Let $\mathcal{E}_k$ be as in (17) and $\mathcal{E}_k^p$ denote the class $\{|e_k|^p : e_k \in \mathcal{E}_k\}$. Then, $\mathcal{L}^S_{p,\mathcal{F}^S}$ is the pointwise minimum class $\{\min_{k\in[C]} e_k : e_k \in \mathcal{E}_k^p\}$ and Lemma 13 gives

$$\mathcal{N}(\epsilon, \mathcal{L}^S_{p,\mathcal{F}^S}, d_{q,\boldsymbol{z}_n}) \leq \prod_{k=1}^{C} \mathcal{N}\left(\frac{\epsilon}{C^{1/q}}, \mathcal{E}_k^p, d_{q,\boldsymbol{z}_n}\right).$$

The contraction principle for covering numbers (see Lemma 27.3 in [42]) implies that

$$\mathcal{N}(\epsilon, \mathcal{E}_k^p, d_{q,\boldsymbol{z}_n}) \leq \mathcal{N}\left(\frac{\epsilon}{p}, \mathcal{E}_k, d_{q,\boldsymbol{z}_n}\right).$$

Since, for all pair of functions $e_k(x,y) = y - \bar{f}_k(x)$ and $e'_k(x,y) = y - \bar{f}'_k(x)$, $\forall (x,y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $|e_k(x,y) - e'_k(x,y)| = |\bar{f}_k(x) - \bar{f}'_k(x)|$, we have $\forall \boldsymbol{z}_n \in \mathcal{Z}^n$, $d_{q,\boldsymbol{z}_n}(e_k, e'_k) = d_{q,\boldsymbol{x}_n}(\bar{f}_k, \bar{f}'_k)$ and

$$\mathcal{N}(\epsilon, \mathcal{E}_k, d_{q,\boldsymbol{z}_n}) \leq \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{q,\boldsymbol{x}_n}).$$

Putting all these inequalities together concludes the proof. □

In order to optimize the dependency on C, we apply chaining (Theorem 4 in App. A) to estimate the Rademacher complexity of $\mathcal{L}^S_{p,\mathcal{F}^S}$ with the relationship (4) and covering numbers in L_∞-norm controlled by Lemma 3. This yields a radical dependency on C:

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{L}^S_{p,\mathcal{F}^S}) &\leq 2^{-N} + 6\sum_{j=1}^{N} 2^{-j}\sqrt{\frac{\log \mathcal{N}(2^{-j}, \mathcal{L}^S_{p,\mathcal{F}^S}, d_{2,\boldsymbol{z}_n})}{n}} \\ &\leq 2^{-N} + 6\sqrt{\frac{C}{n}}\sum_{j=1}^{N} 2^{-j}\sqrt{\max_{k\in[C]} \log \mathcal{N}\left(\frac{2^{-j}}{p}, \bar{\mathcal{F}}_k, d_{\infty,\boldsymbol{x}_n}\right)}\end{aligned} \tag{19}$$

and a convergence rate that depends on the capacity of the $\bar{\mathcal{F}}_k$'s as measured by their covering numbers.

In particular, for classes with fat-shattering dimensions that grow no more than polynomially with ϵ^{-1}, as in (7), Lemma 7 (Appendix A) yields

$$\max_{k\in[C]} \log \mathcal{N}\left(\epsilon, \bar{\mathcal{F}}_k, d_{\infty,\boldsymbol{x}_n}\right) \leq 6 \cdot 4^\beta \alpha \epsilon^{-\beta} \log^2 \frac{2en}{\epsilon}$$

and (19) leads to

$$\begin{aligned}&\hat{\mathcal{R}}_n(\mathcal{L}^S_{p,\mathcal{F}^S}) \\ &\quad\leq 2^{-N} + 6\sqrt{\frac{C}{n}}\sum_{j=1}^{N} 2^{-j}\sqrt{6\cdot 4^\beta \alpha p^\beta 2^{j\beta}\log^2(2enp2^j)} \\ &\quad\leq 2^{-N} + 6\cdot 2^\beta\sqrt{\frac{6\alpha p^\beta C}{n}}\log(2enp2^N)\sum_{j=1}^{N} 2^{j\left(\frac{\beta}{2}-1\right)}.\end{aligned}$$

Setting $N = \lceil \log_2 n^{\frac{1}{\beta}} \rceil \leq \frac{1}{\beta}\log_2(2^\beta n)$ gives, for $\beta = 2$,

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{L}^S_{p,\mathcal{F}^S}) &\leq \frac{1}{\sqrt{n}} + 12p\sqrt{\frac{6\alpha C}{n}}\log(4epn^{\frac{3}{2}})\log_2(4n) \\ &= O\left(\frac{\log^2 n}{\sqrt{n}}\right).\end{aligned}$$

For $\beta > 2$, recalling (10) leads to

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{L}^S_{p,\mathcal{F}^S}) &\leq \frac{1}{n^{\frac{1}{\beta}}} + \frac{3\cdot 2^{2\beta-1}\sqrt{6\alpha p^\beta C}}{2^{\left(\frac{\beta}{2}-1\right)}-1}\frac{\log(4epn^{\frac{1}{\beta}+1})}{n^{\frac{1}{\beta}}} \\ &= O\left(\frac{\log n}{n^{\frac{1}{\beta}}}\right).\end{aligned}$$

Overall, we observe an additional factor in $O(\sqrt{\log n})$ compared to the bounds in (9) and (11) for PWS regression with similar component function classes.

a) Switching kernel regression: For kernel-based classes $\mathcal{F}_k$ as in (12), we could apply the results above with $\beta = 2$ thanks to Lemma 9. However, for such function classes, the covering numbers can be more efficiently bounded without invoking a Sauer-Shelah lemma and the fat-shattering dimension. In particular, for L_∞-norm covering numbers, we can use Lemma 11 in (19) and obtain, with $N = \lceil \log_2 \sqrt{n} \rceil \leq \log_2(2\sqrt{n})$,

$$\begin{aligned}&\hat{\mathcal{R}}_n(\mathcal{L}^S_{p,\mathcal{F}^S}) \\ &\quad\leq \frac{1}{\sqrt{n}} + 36pR_xR_{\mathcal{H}}\log_2(2\sqrt{n})\sqrt{\frac{C}{n}\log(30pR_xR_{\mathcal{H}}n^{3/2})} \\ &\quad= O\left(\frac{\log^{\frac{3}{2}} n}{\sqrt{n}}\right).\end{aligned}$$

Thus, for switching kernel regression, the bound is essentially of the same order as the one for kernel-based PWS regression.

Compared with (18), we gained a $\sqrt{C}$ but also introduced a $\log^{3/2} n$ factor, which is only beneficial when $\log^3 n < C < \sqrt{n}$ (up to constant factors). This limited range over which chaining provides a gain is due to the use of kernel-based classes whose Rademacher complexity can be directly and very efficiently bounded.

b) Switching linear regression: For switching linear regression with $\mathcal{X} \subset \mathbb{R}^d$ and classes $\mathcal{F}_k$ as in (13), the convergence rate is much better and in fact similar to that of (18). To see this, note that with (14) we can apply the integral form of chaining (Theorem 4) to obtain

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{L}^S_{p,\mathcal{F}^S}) &\leq \frac{12}{\sqrt{n}}\int_0^{1/2}\sqrt{\log \mathcal{N}(\epsilon, \mathcal{L}^S_{p,\mathcal{F}^S}, d_{\infty,\boldsymbol{z}_n})}\,d\epsilon \\ &\leq 12\sqrt{\frac{C}{n}}\int_0^{1/2}\sqrt{\max_{k\in[C]}\log \mathcal{N}(\epsilon/p, \bar{\mathcal{F}}_k, d_{\infty,\boldsymbol{x}_n})}\,d\epsilon \\ &\leq 12\sqrt{\frac{Cd}{n}}\int_0^{pR_wR_x}\sqrt{\log\left(\frac{p(2+R_w)R_x}{\epsilon}\right)}\,d\epsilon \\ &\leq 12pR_wR_x\sqrt{\log(2/R_w+1)}\sqrt{\frac{Cd}{n}}.\end{aligned} \tag{20}$$

By comparing with Theorem 3 and (18), we had to pay a $\sqrt{d}$ factor in exchange for a $\sqrt{C}$ one, which is advantageous in low or moderate-dimensional cases, as those that often occur in applications such as hybrid system identification [22].

IV. Conclusions

The paper derived error bounds for piecewise smooth and switching regression. These bounds are based on a decomposition of the capacity measure of the class of interest in terms of those of its component function classes. Different levels of decomposition were explored to optimize the dependency of the bounds on the number of components and a radical dependency was obtained for both PWS and switching regression via chaining and decomposition at the level of covering numbers. We note that this radical dependency is not a final characterization of the optimal growth rate. Indeed, the application of chaining could have been optimized (for instance by replacing n by $n/\sqrt{C}$ when setting the value of N) to yield (only slightly) better growth rates at the cost of more complex expressions for the bounds. In addition, recent work in [47] indicate that better convergence rates (with a $\sqrt{\log n}$ exchanged for a power of $\log C$) could be obtained for switching regression in Sect. III-B2 by using covering numbers in L_q-norm with $2 < q < \infty$ and the Sauer-Shelah lemma of that paper.

From a practical viewpoint, most algorithms consider the unconstrained minimization of a penalized empirical risk functional rather than the minimization of the empirical risk over a constrained function class as in (12) or (13). In such cases, with Hilbert space balls, the risk bounds can be made uniform over the radius of the ball using arguments similar to those of Theorem 4.5 in [31]. For instance, for switching regression with (13), the bound (18) leads to (see Appendix F): For all $\delta \in (0,1)$, with probability at least $1-\delta$, for all $f = (f_k)_{1\leq k\leq C}$ with $f_k(x) = \langle w_k, x\rangle$, $W_f = \max_{k\in[C]} \|w_k\|_2 < \infty$, and $\tilde{W}_f = \max\{1, W_f\}$,

$$L_p^S(\bar{f}) \leq \hat{L}_{p,n}^S(\bar{f}) + 4pC\frac{R_x\tilde{W}_f}{\sqrt{n}} + \sqrt{\frac{\log\frac{3}{\delta}}{2n}} + \sqrt{\frac{\log\log_2 2\tilde{W}_f}{n}}. \tag{21}$$

Open issues include the followings.

Decomposition. While we could also directly decompose the Rademacher complexity of a switching loss class, the efficient decomposition of PWS classes at the level of Rademacher complexities remains an open issue. Even if we can expect a worse dependency on the number of modes, as for the arbitrarily switching case, the convergence rates might be better for specific component function classes such as linear ones or RKHS balls.

Non-independent case. We assumed independence of the sampled data. Extending our results to the non-independent case would be of primary interest for time-series prediction and their application to hybrid dynamical system identification, where the input typically includes lagged values of the output. Works in that direction could follow the bounding schemes developed in [48].

Nonlinear classifiers in PWS regression. Sharpening the analysis of the PWS case to obtain nontrivial bounds with a set of classifiers of infinite Natarajan dimension remains an open issue. This would typically require new decomposition results based on a scale-sensitive measure of capacity for the classifier instead of the growth function.

Model selection. Based on our results, practical procedures implementing structural risk minimization could be envisioned to tune the number of modes C. Indeed, the empirical risk could be minimized for a sequence of PWS or switching classes with increasing C, before selecting the model with lowest value of the error bound.

Appendix A
Technical results from the literature

We recall the contraction principle for Rademacher complexities.

Lemma 4 (After Theorem 4.12 in [49])**.** *If the function $\phi : \mathbb{R} \to \mathbb{R}$ is L_ϕ-Lipschitz, i.e., if $\forall (u,v) \in \mathbb{R}^2,\ |\phi(u) - \phi(v)| \leq L_\phi|u-v|$, then, for the class $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$,*

$$\mathcal{R}_n(\phi \circ \mathcal{F}) \leq L_\phi \mathcal{R}_n(\mathcal{F}).$$

The following chaining technique due to Dudley relates the Rademacher complexity to the covering numbers.

Theorem 4. *Let $\mathcal{F}$ be a real-valued function class over $\mathcal{T}$ and, for any $\boldsymbol{t}_n \in \mathcal{T}^n$, let $D_\mathcal{F} = \sup_{(f,f')\in\mathcal{F}^2} d_{2,\boldsymbol{t}_n}(f,f')$ denote its diameter. Then, for any $N \in \mathbb{N}^*$,*

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{D_\mathcal{F}}{2^N} + 6D_\mathcal{F}\sum_{j=1}^N 2^{-j}\sqrt{\frac{\log\mathcal{N}(D_\mathcal{F}2^{-j}, \mathcal{F}, d_{2,\boldsymbol{t}_n})}{n}}$$

and, if the integral exists,

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq 12\int_0^{D_\mathcal{F}/2}\sqrt{\frac{\log\mathcal{N}(\epsilon, \mathcal{F}, d_{2,\boldsymbol{t}_n})}{n}}d\epsilon.$$

The following result upper bounds the Rademacher complexity of a class of functions defined as the pointwise maximum of a set of functions.

Lemma 5 (After Lemma 8.1 in [31])**.** *Let $(\mathcal{U}_k)_{1\leq k\leq K}$ be a sequence of K classes of functions from $\mathcal{Z}$ into $\mathbb{R}$. Then, for the class $\mathcal{U} = \{u \in \mathbb{R}^\mathcal{Z} : u(z) = \max_{k\in[K]} u_k(z),\ u_k \in \mathcal{U}_k\}$,*

$$\hat{\mathcal{R}}_n(\mathcal{U}) \leq \sum_{k=1}^K \hat{\mathcal{R}}_n(\mathcal{U}_k).$$

We use the following generalized Sauer-Shelah lemmas.

Lemma 6 (After Corollary 5 in [50] and Theorem 9 in [51])**.** *Let $d_\mathcal{G}$ be the Natarajan dimension of $\mathcal{G}$. Then, for any $\boldsymbol{x}_n \in \mathcal{X}^n$,*

$$|\mathcal{G}_{\boldsymbol{x}_n}| \leq \sum_{i=1}^{d_\mathcal{G}}\binom{n}{i}\binom{C}{2}^i \leq \left(\frac{neC}{2d_\mathcal{G}}\right)^{d_\mathcal{G}}.$$

Lemma 7 (After Lemma 3.5 in [39])**.** *For a class $\mathcal{F}$ of functions from $\mathcal{X}$ into $[-M,M]$, let $d_\mathcal{F}(\epsilon)$ denote its fat-shattering dimension. Then, for any $\epsilon \in (0, 2M]$ and $n \in \mathbb{N}^*$,*

$$\mathcal{N}_\infty(\epsilon, \mathcal{F}, n) \leq 2\left(\frac{16M^2n}{\epsilon^2}\right)^{d_\mathcal{F}(\frac{\epsilon}{4})\log_2\left(\frac{4Men}{d_\mathcal{F}(\frac{\epsilon}{4})\epsilon}\right)}.$$

Lemma 8 (After Theorem 1 in [41] and Lemma 3 in [33])**.** *For a class $\mathcal{F}$ of functions from $\mathcal{X}$ into $[-M,M]$, let $d_\mathcal{F}(\epsilon)$*

denote its fat-shattering dimension. Then, for any $\epsilon \in (0, 2M]$ *and* $n \in \mathbb{N}^*$,

$$\mathcal{N}_2(\epsilon, \mathcal{F}, n) \leq \left(\frac{13M}{\epsilon}\right)^{20 d_{\mathcal{F}}(\frac{\epsilon}{96})}.$$

For linear and/or kernel-based classes, the different capacity meaures can be bounded as follows.

Lemma 9 (After Theorem 4.6 in [52])**.** *Given a Hilbert space* $\mathcal{H}$ *and a mapping* $\phi : \mathcal{X} \to \mathcal{H}$*, let* $\mathcal{X} \subseteq \{x \in \mathcal{X} : \|\phi(x)\|_{\mathcal{H}} \leq R_x\}$ *and* $\mathcal{F} = \{f \in \mathbb{R}^{\mathcal{X}} : \ f(x) = \langle w, \phi(x)\rangle_{\mathcal{H}}, \ \|w\|_{\mathcal{H}} \leq R_w\}$*. Then, for any* $\epsilon > 0$,

$$d_{\mathcal{F}}(\epsilon) \leq \left(\frac{R_x R_w}{\epsilon}\right)^2.$$

Lemma 10 (After Theorem 5.5 in [31])**.** *Given a Hilbert space* $\mathcal{H}$ *and a mapping* $\phi : \mathcal{X} \to \mathcal{H}$*, let* $\mathcal{X} \subseteq \{x \in \mathcal{X} : \|\phi(x)\|_{\mathcal{H}} \leq R_x\}$ *and* $\mathcal{F} = \{f \in \mathbb{R}^{\mathcal{X}} : \ f(x) = \langle w, \phi(x)\rangle_{\mathcal{H}}, \ \|w\|_{\mathcal{H}} \leq R_w\}$*. Then, for any* $n \in \mathbb{N}^*$,

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{R_x R_w}{\sqrt{n}}.$$

Lemma 11 (After Theorem 4 in [46])**.** *Given a Hilbert space* $\mathcal{H}$ *and a mapping* $\phi : \mathcal{X} \to \mathcal{H}$*, let* $\mathcal{X} \subseteq \{x \in \mathcal{X} : \|\phi(x)\|_{\mathcal{H}} \leq R_x\}$ *and* $\mathcal{F} = \{f \in \mathbb{R}^{\mathcal{X}} : \ f(x) = \langle w, \phi(x)\rangle_{\mathcal{H}}, \ \|w\|_{\mathcal{H}} \leq R_w\}$*. Then, for any* $\epsilon \in (0, R_x R_w)$,

$$\begin{aligned}\log \mathcal{N}_\infty(\epsilon, \mathcal{F}, n) &\leq 36 \frac{R_x^2 R_w^2}{\epsilon^2} \log\left(2\left\lceil \frac{4 R_x R_w}{\epsilon} + 2\right\rceil n + 1\right) \\ &\leq 36 \frac{R_x^2 R_w^2}{\epsilon^2} \log\left(\frac{15 R_x R_w n}{\epsilon}\right),\end{aligned}$$

and for $\epsilon \geq R_x R_w$, $\log \mathcal{N}_\infty(\epsilon, \mathcal{F}, n) = 0$.

Appendix B
Proof of Theorem 2

Theorem 2 is a consequence of Theorem 1 applied to the function class

$$\mathcal{L}_{p,\mathcal{F}} = \{\ell \in [0,1]^{\mathcal{X}\times\mathcal{Y}} : \ell(x,y) = |y - \bar{f}(x)|^p, \ \bar{f} \in \bar{\mathcal{F}}\}, \tag{22}$$

whose Rademacher complexity can be related to the one of $\mathcal{F}$ as follows. Let us define the error class as $\mathcal{E} = \{e \in [-2M, 2M]^{\mathcal{X}\times\mathcal{Y}} : e(x,y) = y - \bar{f}(x), \ \bar{f} \in \bar{\mathcal{F}}\}$. Define the function $\phi_p(u) = u^p$ with domain $[0, 2M]$ whose Lipschitz constant is $p(2M)^{p-1}$. Then, $\mathcal{L}_{p,\mathcal{F}} = \phi_p \circ |\cdot| \circ \mathcal{E}$ and, by contraction (see Lemma 4 in Appendix A), we have

$$\mathcal{R}_n(\mathcal{L}_{p,\mathcal{F}}) = \mathcal{R}_n(\phi_p \circ |\cdot| \circ \mathcal{E}) \leq p 2^{p-1} M^{p-1} \mathcal{R}_n(\mathcal{E}).$$

Then, we bound the Rademacher complexity of $\mathcal{E}$ as

$$\begin{aligned}\mathcal{R}_n(\mathcal{E}) &= \mathbb{E}_{\boldsymbol{X}_n \boldsymbol{Y}_n \boldsymbol{\sigma}_n} \sup_{\bar{f}\in\bar{\mathcal{F}}} \frac{1}{n}\sum_{i=1}^n \sigma_i (Y_i - \bar{f}(X_i)) \\ &\leq \mathbb{E}_{\boldsymbol{Y}_n \boldsymbol{\sigma}_n} \frac{1}{n}\sum_{i=1}^n \sigma_i Y_i + \mathbb{E}_{\boldsymbol{X}_n \boldsymbol{\sigma}_n} \sup_{\bar{f}\in\bar{\mathcal{F}}} \frac{1}{n}\sum_{i=1}^n -\sigma_i \bar{f}(X_i)\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{Y}_n \boldsymbol{\sigma}_n} \frac{1}{n}\sum_{i=1}^n \sigma_i Y_i = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y_i \sigma_i} \sigma_i Y_i = 0$, and, since σ_i and $-\sigma_i$ have the same distribution,

$$\mathbb{E}_{\boldsymbol{X}_n \boldsymbol{\sigma}_n} \sup_{\bar{f}\in\bar{\mathcal{F}}} \frac{1}{n}\sum_{i=1}^n -\sigma_i \bar{f}(X_i) = \mathcal{R}_n(\bar{\mathcal{F}})$$

Finally, by contraction, $\mathcal{R}_n(\bar{\mathcal{F}}) \leq \mathcal{R}_n(\mathcal{F})$ and taking $M = \frac{1}{2}$ completes the proof.

Appendix C
Proof of Lemma 1

For each possible classification $\boldsymbol{c} \in \mathcal{G}_{\boldsymbol{x}_n}$ of $\boldsymbol{x}_n$, let $g_{\boldsymbol{c}} \in \mathcal{G}$ be a classifier from $\mathcal{G}$ producing this classification. Then, we build a set $H_{\boldsymbol{c}}$ of functions $h \in \bar{\mathcal{F}}_{\mathcal{G}}$ such that $h(x_i) = h_{g_{\boldsymbol{c}}(x_i)}(x_i) = h_{c_i}(x_i)$ with $(h_k)_{1\leq k\leq C}$ taken from the product of the smallest proper ϵ-nets of the $\bar{\mathcal{F}}_k$'s. Since there are $\mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{q,\boldsymbol{x}_n})$ functions h_k in each one of these ϵ-nets, we have

$$|H_{\boldsymbol{c}}| \leq \prod_{k=1}^C \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{q,\boldsymbol{x}_n})$$

and, since there are at most $\Pi_{\mathcal{G}}(n)$ classifications $\boldsymbol{c} \in \mathcal{G}_{\boldsymbol{x}_n}$, we can build a set $H = \bigcup_{\boldsymbol{c}\in\mathcal{G}_{\boldsymbol{x}_n}} H_{\boldsymbol{c}} \subseteq \bar{\mathcal{F}}_{\mathcal{G}}$ with a cardinality bounded by

$$|H| \leq \sum_{\boldsymbol{c}\in\mathcal{G}_{\boldsymbol{x}_n}} |H_{\boldsymbol{c}}| \leq \Pi_{\mathcal{G}}(n) \prod_{k=1}^C \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{q,\boldsymbol{x}_n}).$$

To conclude, we need to show that H is a $(C^{1/q}\epsilon)$-net of $\bar{\mathcal{F}}_{\mathcal{G}}$ with respect to $d_{q,\boldsymbol{x}_n}$. Given any $f \in \bar{\mathcal{F}}_{\mathcal{G}}$, there is some $\boldsymbol{c} \in \mathcal{G}_{\boldsymbol{x}_n}$ that coincides with the classification of $\boldsymbol{x}_n$ by g in $f(x) = f_{g(x)}(x)$, and thus for which, for $q < \infty$, for all functions $h \in H_{\boldsymbol{c}} \subseteq H$,

$$\begin{aligned}d_{q,\boldsymbol{x}_n}(f,h)^q &= \frac{1}{n}\sum_{i=1}^n |f(x_i) - h(x_i)|^q \\ &= \frac{1}{n}\sum_{i=1}^n |f_{c_i}(x_i) - h_{c_i}(x_i)|^q \\ &= \frac{1}{n}\sum_{k=1}^C \sum_{i:c_i=k} |f_k(x_i) - h_k(x_i)|^q \\ &\leq \sum_{k=1}^C \frac{1}{n}\sum_{i=1}^n |f_k(x_i) - h_k(x_i)|^q.\end{aligned}$$

By construction, among all the functions $h \in H_{\boldsymbol{c}} \subseteq H$, there is at least one such that, for all $k \in [C]$, h_k is the center of an ϵ-ball containing $f_k \in \bar{\mathcal{F}}_k$, i.e., $\frac{1}{n}\sum_{i=1}^n |f_k(x_i) - h_k(x_i)|^q \leq \epsilon^q$. Thus, there is some $h \in H$ such that

$$d_{q,\boldsymbol{x}_n}(f,h)^q \leq \sum_{k=1}^C \epsilon^q = C\epsilon^q.$$

The statement for $q < \infty$ follows by rescaling ϵ by $1/C^{1/q}$. The case $q = \infty$ is proved similarly, but without the need for rescaling:

$$\begin{aligned}d_{\infty,\boldsymbol{x}_n}(f,h) &= \max_{i\in[n]} |f(x_i) - h(x_i)| \\ &= \max_{k\in[C]} \max_{i:c_i=k} |f_k(x_i) - h_k(x_i)| \\ &\leq \max_{k\in[C]} \max_{i\in[n]} |f_k(x_i) - h_k(x_i)| \\ &\leq \max_{k\in[C]} \epsilon = \epsilon.\end{aligned}$$

Appendix D
Proof of Lemma 2

For each possible classification $\boldsymbol{c} \in \mathcal{G}_{\boldsymbol{x}_n}$ of $\boldsymbol{x}_n$, let $n_k = \sum_{i=1}^n \mathbf{1}_{c_i=k}$ denote the number of points classified in category k, define $\mathcal{K}(\boldsymbol{c}) = \{k \in [C] : n_k > 0\}$ as the subset of categories with at least one point, and consider $\mathcal{K}(\boldsymbol{c})$ empirical pseudo-metrics $d_{x_i:c_i=k}$ defined as $d_{2,\boldsymbol{x}_n}$ but on a restricted set of points of cardinality $n_k > 0$: $\forall (f, f') \in \left(\mathbb{R}^{\mathcal{X}}\right)^2$,

$$d_{x_i:c_i=k}(f, f') = \left(\frac{1}{n_k} \sum_{i:c_i=k} (f(x_i) - f'(x_i))^2\right)^{\frac{1}{2}}.$$

For each such distance, we build a proper ϵ-net of $\bar{\mathcal{F}}_k$ of cardinality $\mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{x_i:c_i=k})$. Then, we build a set $H_{\boldsymbol{c}}$ of functions h such that $h(x_i) = h_{c_i}(x_i)$ with $(h_k)_{k\in\mathcal{K}(\boldsymbol{c})}$ taken from the product of these ϵ-nets, so that

$$|H_{\boldsymbol{c}}| \leq \prod_{k\in\mathcal{K}(\boldsymbol{c})} \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{x_i:c_i=k}),$$

where the covering numbers depend on $\boldsymbol{x}_n$ as usual, but also on $\boldsymbol{c}$ via the definition of the pseudo-distances.

Next, we consider a set $H = \bigcup_{\boldsymbol{c}\in\mathcal{G}_{\boldsymbol{x}_n}} H_{\boldsymbol{c}}$ which contains at most

$$|H| \leq \sum_{\boldsymbol{c}\in\mathcal{G}_{\boldsymbol{x}_n}} |H_{\boldsymbol{c}}| \leq \sum_{\boldsymbol{c}\in\mathcal{G}_{\boldsymbol{x}_n}} \prod_{k\in\mathcal{K}(\boldsymbol{c})} \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{x_i:c_i=k})$$

functions. Following the proof of Lemma 1, for any $f \in \bar{\mathcal{F}}_{\mathcal{G}}$, there is some $\boldsymbol{c} \in \mathcal{G}_{\boldsymbol{x}_n}$ such that for all $h \in H_{\boldsymbol{c}} \subseteq H$,

$$\begin{aligned} d_{2,\boldsymbol{x}_n}(f, h)^2 &= \frac{1}{n}\sum_{i=1}^n (f_{c_i}(x_i) - h_{c_i}(x_i))^2 \\ &= \frac{1}{n}\sum_{k\in\mathcal{K}(\boldsymbol{c})} \sum_{i:c_i=k} (f_k(x_i) - h_k(x_i))^2 \\ &= \sum_{k\in\mathcal{K}(\boldsymbol{c})} \frac{n_k}{n}\frac{1}{n_k} \sum_{i:c_i=k} (f_k(x_i) - h_k(x_i))^2 \\ &= \sum_{k\in\mathcal{K}(\boldsymbol{c})} \frac{n_k}{n} d_{x_i:c_i=k}(f_k, h_k)^2. \end{aligned}$$

Thus, by the fact that $\sum_{k\in\mathcal{K}(\boldsymbol{c})} \frac{n_k}{n} = 1$, $d_{2,\boldsymbol{x}_n}(f, h)^2$ is expressed as a convex combination of squared sub-distances $d_{x_i:c_i=k}(f_k, h_k)^2$, which can be bounded by their maximum as

$$d_{2,\boldsymbol{x}_n}(f, h)^2 \leq \max_{k\in\mathcal{K}(\boldsymbol{c})} d_{x_i:c_i=k}(f_k, h_k)^2.$$

By construction, among all the functions h in $H_{\boldsymbol{c}}$, there is at least one such that, for all $k \in \mathcal{K}(\boldsymbol{c})$, h_k is the center of an ϵ-ball containing $f_k \in \bar{\mathcal{F}}_k$, i.e., $d_{x_i:c_i=k}(f_k, h_k)^2 \leq \epsilon^2$. Thus, there is some $h \in H$ such that $d_{2,\boldsymbol{x}_n}(f, h)^2 \leq \epsilon^2$, which proves that H is an ϵ-net of $\bar{\mathcal{F}}_{\mathcal{G}}$.

Now, we can improve the bound on $|H|$ by using uniform covering numbers. In particular, for any $\boldsymbol{c} \in \mathcal{G}_{\boldsymbol{x}_n}$,

$$\mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{x_i:c_i=k}) \leq \sup_{\boldsymbol{x}_{n_k}\subset\mathcal{X}} \mathcal{N}(\epsilon, \bar{\mathcal{F}}_k, d_{2,\boldsymbol{x}_{n_k}}) = \mathcal{N}_2(\epsilon, \bar{\mathcal{F}}_k, n_k).$$

Thus, by using Lemma 12 in Appendix E:

$$\begin{aligned} |H| &\leq \sum_{\boldsymbol{c}\in\mathcal{G}_{\boldsymbol{x}_n}} \prod_{k\in\mathcal{K}(\boldsymbol{c})} \mathcal{N}_2(\epsilon, \bar{\mathcal{F}}_k, n_k) \leq \sum_{\boldsymbol{c}\in\mathcal{G}_{\boldsymbol{x}_n}} \prod_{k\in\mathcal{K}(\boldsymbol{c})} \mathcal{N}_2(\epsilon, \bar{\mathcal{F}}_k, n) \\ &\leq \Pi_{\mathcal{G}}(n) \prod_{k=1}^C \mathcal{N}_2(\epsilon, \bar{\mathcal{F}}_k, n). \end{aligned}$$

Appendix E
Additional results on covering numbers

We need the following result on uniform covering numbers.

Lemma 12. *Let $\mathcal{F}$ be a uniform GC class from $\mathcal{X}$ into $[-M, M]$. Then, for any $\epsilon > 0$ and any $q \in [1, \infty) \cup \{\infty\}$, the uniform covering numbers $\mathcal{N}_q(\epsilon, \mathcal{F}, n)$ form a non-decreasing function of n.*

Proof. Recall from [39] that a class of uniformly bounded real-valued functions is a uniform GC class if and only if its fat-shattering dimension is finite for all $\epsilon > 0$. Note that this also implies the finiteness of its L_q-norm covering numbers by Lemma 7 and, for all $q \in [1, \infty)$, the relation $\mathcal{N}_q(\epsilon, \mathcal{F}, n) \leq \mathcal{N}_\infty(\epsilon, \mathcal{F}, n)$. Thus, for all n, $\mathcal{N}_q(\epsilon, \mathcal{F}, n)$ is finite and, since it is the largest of a finite set of integers, we have $\mathcal{N}_q(\epsilon, \mathcal{F}, n) = \max_{\boldsymbol{x}_n\in\mathcal{X}^n} \mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n})$. Then, there is a sequence $\boldsymbol{x}_n \in \mathcal{X}^n$ on which the maximum (let it be N_n) is attained, i.e., such that $N_n = \mathcal{N}_q(\epsilon, \mathcal{F}, n) = \mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n})$. Note that for any $\boldsymbol{x}_n \in \mathcal{X}^n$,

$$\mathcal{N}_q(\epsilon, \mathcal{F}, n+1) \geq \sup_{x\in\mathcal{X}} \mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n x}),$$

where $d_{q,\boldsymbol{x}_n x}$ is the pseudo-metric defined over the concatenation of the sequence $\boldsymbol{x}_n$ with x. Therefore, it is sufficient to show that $\sup_{x\in\mathcal{X}} \mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n x}) \geq N_n$.

For $q = \infty$, this is a direct consequence of the fact that for all $(f, h) \in \left(\mathbb{R}^{\mathcal{X}}\right)^2$, $d_{\infty,\boldsymbol{x}_n x}(f, h) = \max\{d_{\infty,\boldsymbol{x}_n}(f, h), |f(x) - h(x)|\} \geq d_{\infty,\boldsymbol{x}_n}(f, h)$. For $q \in [1, \infty)$, assume that it is not the case, then $\forall x \in \mathcal{X}$, there is an ϵ-net $\mathcal{H}$ of $\mathcal{F}$ of cardinality $\mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n x}) < N_n$ and, for all $f \in \mathcal{F}$, there is an $h \in \mathcal{H}$ such that $d_{q,\boldsymbol{x}_n x}(f, h) \leq \epsilon$. Since

$$d_{q,\boldsymbol{x}_n x}(f, h)^q = \frac{\sum_{i=1}^n |f(x_i) - h(x_i)|^q + |f(x) - h(x)|^q}{n+1},$$

we have

$$\begin{aligned} d_{q,\boldsymbol{x}_n}(f, h)^q &= \frac{1}{n}\sum_{i=1}^n |f(x_i) - h(x_i)|^q \\ &= \frac{n+1}{n}\left(d_{q,\boldsymbol{x}_n x}(f, h)^q - \frac{1}{n+1}|f(x) - h(x)|^q\right) \\ &\leq \frac{(n+1)\epsilon^q - |f(x) - h(x)|^q}{n}. \end{aligned} \tag{23}$$

In the case where $|f(x_i) - h(x_i)| < \epsilon$, for all $i \in [n]$, then, $d_{q,\boldsymbol{x}_n}(f, h) = \left(\frac{1}{n}\sum_{i=1}^n |f(x_i) - h(x_i)|^q\right)^{\frac{1}{q}} \leq \epsilon$ and $\mathcal{H}$ is also an ϵ-net of $\mathcal{F}$ for the pseudo-metric $d_{q,\boldsymbol{x}_n}$, thus ensuring that $\mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n}) \leq |\mathcal{H}|$, which contradicts the assumptions $\mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n}) = N_n$ and $|\mathcal{H}| < N_n$. If this is not the case,

i.e., if there is some $i \in [n]$, such that $|f(x_i) - h(x_i)| \geq \epsilon$, then choosing $x = x_i$ in (23) yields

$$d_{q,\boldsymbol{x}_n}(f,h)^q \leq \frac{(n+1)\epsilon^q - \epsilon^q}{n} = \epsilon^q$$

and therefore, $\mathcal{H}$ is also an ϵ-net for the pseudo-metric $d_{q,\boldsymbol{x}_n}$ in this case, showing again a contradiction. As a consequence, $\sup_{x\in\mathcal{X}} \mathcal{N}(\epsilon, \mathcal{F}, d_{q,\boldsymbol{x}_n x}) \geq N_n$ and the lemma is proved. □

Using the ideas from the proof of Lemma 1 in [33], we can derive the following structural result on covering numbers.

Lemma 13. *Given a sequence of C real-valued function classes $\mathcal{A}_k$ with domain $\mathcal{Z}$, let $\mathcal{A}$ be either the pointwise maximum class $\{a \in \mathbb{R}^{\mathcal{Z}} : a(z) = \max_{k\in[C]} a_k(z),\ a_k \in \mathcal{A}_k\}$ or the pointwise minimum class $\{a \in \mathbb{R}^{\mathcal{Z}} : a(z) = \min_{k\in[C]} a_k(z),\ a_k \in \mathcal{A}_k\}$. Then, for any $q \geq 1$ and $q = \infty$,*

$$\mathcal{N}(\epsilon, \mathcal{A}, d_{q,\boldsymbol{z}_n}) \leq \prod_{k=1}^{C} \mathcal{N}\left(\frac{\epsilon}{C^{1/q}}, \mathcal{A}_k, d_{q,\boldsymbol{z}_n}\right).$$

Proof. We start with the pointwise maximum case and $q = \infty$. Let $\mathcal{H}_k$ be a minimal proper ϵ-net of $\mathcal{A}_k$ and $\mathcal{H}$ be the pointwise maximum class of $(\mathcal{H}_k)_{k\in[C]}$ (and note that $\mathcal{H}_k \subseteq \mathcal{A}_k$ implies $\mathcal{H} \subseteq \mathcal{A}$). Let $k(f,z) = \operatorname{argmax}_k f_k(z)$. Then, for any $a \in \mathcal{A}$ and $h \in \mathcal{H}$, there are $(a_k)_{k\in[C]}$ and $(h_k)_{k\in[C]}$ such that

$$d_{\infty,\boldsymbol{z}_n}(a,h) = \max_{i\in[n]} |a_{k(a,z_i)}(z_i) - h_{k(h,z_i)}(z_i)|.$$

By using the definition of $k(f,z)$, we can deduce that if $a_{k(a,z_i)}(z_i) \geq h_{k(h,z_i)}(z_i)$, then

$$\begin{aligned}|a_{k(a,z_i)}(z_i) - h_{k(h,z_i)}(z_i)| &= a_{k(a,z_i)}(z_i) - h_{k(h,z_i)}(z_i)\\ &\leq a_{k(a,z_i)}(z_i) - h_{k(a,z_i)}(z_i)\\ &\leq |a_{k(a,z_i)}(z_i) - h_{k(a,z_i)}(z_i)|\end{aligned}$$

and that if $a_{a,k(z_i)}(z_i) < h_{k(h,z_i)}(z_i)$, then

$$\begin{aligned}|a_{k(a,z_i)}(z_i) - h_{k(h,z_i)}(z_i)| &= h_{k(h,z_i)}(z_i) - a_{k(a,z_i)}(z_i)\\ &\leq h_{k(h,z_i)}(z_i) - a_{k(h,z_i)}(z_i)\\ &\leq |h_{k(h,z_i)}(z_i) - a_{k(h,z_i)}(z_i)|.\end{aligned}$$

Thus,

$$|a_{k(a,z_i)}(z_i) - h_{k(h,z_i)}(z_i)| \leq \max_{k\in[C]} |a_k(z_i) - h_k(z_i)|$$

and

$$\begin{aligned}d_{\infty,\boldsymbol{z}_n}(a,h) &\leq \max_{i\in[n]} \max_{k\in[C]} |a_k(z_i) - h_k(z_i)| \qquad (24)\\ &\leq \max_{k\in[C]} d_{\infty,\boldsymbol{z}_n}(a_k,h_k) \leq \max_{k\in[C]} \epsilon \leq \epsilon,\end{aligned}$$

which proves the statement for the pointwise maximum class.

If $\mathcal{A}$ is the pointwise minimum class, let $\mathcal{A}'$ be the pointwise maximum of $(-\mathcal{A}_k)$, and note that $\mathcal{A} = -\mathcal{A}'$. Then the statement follows by using $\mathcal{N}(\epsilon, -\mathcal{A}', d_{p,\boldsymbol{z}_n}) = \mathcal{N}(\epsilon, \mathcal{A}', d_{p,\boldsymbol{z}_n})$, $\mathcal{N}(\epsilon, \mathcal{A}_k, d_{p,\boldsymbol{z}_n}) = \mathcal{N}(\epsilon, -\mathcal{A}_k, d_{p,\boldsymbol{z}_n})$ and the result for the pointwise maximum class $\mathcal{A}'$.

For $q \in [1,\infty)$, the same reasoning applies with (24) replaced by

$$\begin{aligned}d_{q,\boldsymbol{z}_n}(a,h)^q &\leq \frac{1}{n}\sum_{i=1}^{n}\left(\max_{k\in[C]} |a_k(z_i) - h_k(z_i)|\right)^q\\ &\leq \frac{1}{n}\sum_{i=1}^{n}\max_{k\in[C]} |a_k(z_i) - h_k(z_i)|^q\\ &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C} |a_k(z_i) - h_k(z_i)|^q\\ &\leq \sum_{k=1}^{C} d_{q,\boldsymbol{z}_n}(a_k,h_k)^q \leq C\epsilon^q\end{aligned}$$

and a rescaling of ϵ. □

Appendix F
Proof of (21)

Let $\mathcal{F} = \{f = (f_k)_{1\leq k\leq C} : f_k(x) = \langle w_k, x\rangle,\ W_f = \max_{k\in[C]} \|w_k\|_2 < \infty\}$, $\tilde{W}_f = \max\{1, W_f\}$, and $\epsilon_f = \sqrt{\frac{\log\frac{3}{\delta}}{2n}} + \sqrt{\frac{\log\log_2 2\tilde{W}_f}{n}}$. We will prove (21) by showing that

$$P\left\{\exists f \in \mathcal{F},\ L_p^S(\bar{f}) > \hat{L}_{p,n}^S(\bar{f}) + 4pC\frac{R_x\tilde{W}_f}{\sqrt{n}}. + \epsilon_f\right\} \leq \delta \qquad (25)$$

Let $R_j = 2^j$, $\mathcal{F}^j = \{f = (f_k)_{1\leq k\leq C} : f_k(x) = \langle w_k, x\rangle,\ \max_{k\in[C]} \|w_k\|_2 \leq R_j\}$, and, for any $j \geq 1$, set $\epsilon'_j = \epsilon + \sqrt{\frac{\log\log_2 R_j}{n}}$ with $\epsilon = \sqrt{\frac{\log\frac{3}{\delta}}{2n}}$. Thus, we shall prove that (25) holds with $\delta = 3e^{-2n\epsilon^2}$.

For any $f \in \mathcal{F}$, either $f \in \mathcal{F}^0$ with $\tilde{W}_f = 1 \geq 1/2$ and $\epsilon_f = \epsilon$, or there is a $j \geq 1$ such that $f \in \mathcal{F}^j$ with $\tilde{W}_f = W_f$, $R_j \geq W_f > R_{j-1} = R_j/2$ and $\epsilon_f > \epsilon'_j$. Thus, the probability in (25) can be upper bounded by

$$\begin{aligned}&P\left\{\exists j \geq 0, f \in \mathcal{F}^j,\ L_p^S(\bar{f}) > \hat{L}_{p,n}^S(\bar{f}) + 4pC\frac{R_x\tilde{W}_f}{\sqrt{n}} + \epsilon_f\right\}\\ &\leq P\left\{\exists j \geq 1, f \in \mathcal{F}^j,\ L_p^S(\bar{f}) > \hat{L}_{p,n}^S(\bar{f}) + 2pC\frac{R_xR_j}{\sqrt{n}} + \epsilon'_j\right\}\\ &\quad + P\left\{\exists f \in \mathcal{F}^0,\ L_p^S(\bar{f}) > \hat{L}_{p,n}^S(\bar{f}) + 2pC\frac{R_x}{\sqrt{n}} + \epsilon.\right\}\end{aligned}$$

By (18),

$$P\left\{\exists f \in \mathcal{F}^0, L_p^S(\bar{f}) > \hat{L}_{p,n}^S(\bar{f}) + 2pC\frac{R_x}{\sqrt{n}} + \epsilon\right\} \leq e^{-2n\epsilon}$$

and, for any $j \geq 1$,

$$P\left\{\exists f \in \mathcal{F}^j, L_p^S(\bar{f}) > \hat{L}_{p,n}^S(\bar{f}) + 2pC\frac{R_xR_j}{\sqrt{n}} + \epsilon'_j\right\} \leq e^{-2n{\epsilon'_j}^2}.$$

Since $R_j = 2^j$, we have ${\epsilon'_j}^2 = (\epsilon + \sqrt{\log(j)/n})^2 \leq \epsilon^2 + \log(j)/n$ and

$$\begin{aligned}&P\left\{\exists j \geq 1, f \in \mathcal{F}^j,\ L_p^S(\bar{f}) > \hat{L}_{p,n}^S(\bar{f}) + 2pC\frac{R_xR_j}{\sqrt{n}} + \epsilon'_j\right\}\\ &\leq \sum_{j\geq 1} e^{-2n{\epsilon'_j}^2} \leq \sum_{j\geq 1} e^{-2n\epsilon^2}e^{-2\log j} \leq e^{-2n\epsilon^2}\sum_{j\geq 1}\frac{1}{j^2}\\ &= \frac{\pi^2}{6}e^{-2n\epsilon^2} \leq 2e^{-2n\epsilon^2}.\end{aligned}$$

Therefore, the probability in (25) can be upper bounded by $\delta = 3e^{-2n\epsilon^2}$ and the statement is proved.

REFERENCES

[1] R. Quandt, "The estimation of the parameters of a linear regression system obeying two separate regimes," *Journal of the American Statistical Association*, pp. 873–880, 1958.

[2] H. Späth, "Algorithm 39: Clusterwise linear regression," *Computing*, vol. 22, no. 4, pp. 367–373, 1979.

[3] D. Hosmer, "Maximum likelihood estimates of the parameters of a mixture of two regression lines," *Communications in Statistics*, vol. 3, no. 10, pp. 995–1006, 1974.

[4] W. DeSarbo and W. Cron, "A maximum likelihood methodology for clusterwise linear regression," *Journal of Classification*, vol. 5, no. 2, pp. 249–282, 1988.

[5] S. Gaffney and P. Smyth, "Trajectory clustering with mixtures of regression models," in *ACM SIGKDD*, 1999, pp. 63–72.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.

[7] J. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, pp. 1–67, 1991.

[8] A. Rao, D. Miller, K. Rose, and A. Gersho, "A deterministic annealing approach for parsimonious design of piecewise regression models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 2, pp. 159–173, 1999.

[9] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[10] F. Lauer and G. Bloch, *Hybrid System Identification: Theory and Algorithms for Learning Switching Models*. Springer, 2019.

[11] 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 Conf. on Decision and Control (CDC), Maui, Hawaï, USA*, 2003, pp. 167–172.

[12] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari, "A clustering technique for the identification of piecewise affine systems," *Automatica*, vol. 39, no. 2, pp. 205–217, 2003.

[13] J. Roll, A. Bemporad, and L. Ljung, "Identification of piecewise affine systems via mixed-integer programming," *Automatica*, vol. 40, no. 1, pp. 37–50, 2004.

[14] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, "A bounded-error approach to piecewise affine system identification," *IEEE Transactions on Automatic Control*, vol. 50, no. 10, pp. 1567–1580, 2005.

[15] A. L. Juloski, S. Weiland, and W. Heemels, "A Bayesian approach to identification of hybrid systems," *IEEE Transactions on Automatic Control*, vol. 50, no. 10, pp. 1520–1533, 2005.

[16] L. Bako, "Identification of switched linear systems via sparse optimization," *Automatica*, vol. 47, no. 4, pp. 668–677, 2011.

[17] F. Lauer, G. Bloch, and R. Vidal, "A continuous optimization framework for hybrid system identification," *Automatica*, vol. 47, no. 3, pp. 608–613, 2011.

[18] V. L. Le, G. Bloch, and F. Lauer, "Reduced-size kernel models for nonlinear hybrid system identification," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2398–2405, 2011.

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

[20] T. Pham Dinh, H. Le Thi, H. Le, and F. Lauer, "A difference of convex functions algorithm for switched linear regression," *IEEE Transactions on Automatic Control*, vol. 59, no. 8, pp. 2277–2282, 2014.

[21] V. L. Le, F. Lauer, and G. Bloch, "Selective ℓ_1 minimization for sparse recovery," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 3008–3013, 2014.

[22] F. Lauer, "Global optimization for low-dimensional switching linear regression and bounded-error estimation," *Automatica*, vol. 89, pp. 73–82, 2018.

[23] ——, "On the complexity of piecewise affine system identification," *Automatica*, vol. 62, pp. 148–153, 2015.

[24] ——, "On the complexity of switching linear regression," *Automatica*, vol. 74, pp. 80–83, 2016.

[25] M. Jordan and L. Xu, "Convergence results for the EM approach to mixtures of experts architectures," *Neural networks*, vol. 8, no. 9, pp. 1409–1431, 1995.

[26] Y. Chen, X. Yi, and C. Caramanis, "A convex formulation for mixed regression with two components: Minimax optimal rates," in *COLT*, 2014, pp. 560–604.

[27] A. Zeevi, R. Meir, and V. Maiorov, "Error bounds for functional approximation and estimation using mixtures of experts," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1010–1025, 1998.

[28] M. Kearns, R. Schapire, and L. Sellie, "Toward efficient agnostic learning," *Machine Learning*, vol. 17, no. 2-3, pp. 115–141, 1994.

[29] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," *The Annals of Statistics*, vol. 30, no. 1, pp. 1–50, 2002.

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

[31] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, 2012.

[32] V. Kuznetsov, M. Mohri, and U. Syed, "Multi-class deep boosting," in *NIPS 27*, 2014, pp. 2501–2509.

[33] Y. Guermeur, "L_p-norm Sauer-Shelah lemma for margin multi-category classifiers," *Journal of Computer and System Sciences*, vol. 89, pp. 450–473, 2017.

[34] K. Musayeva, F. Lauer, and Y. Guermeur, "Metric entropy and Rademacher complexity of margin multi-category classifiers," in *ICANN*, 2017.

[35] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

[36] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer, 2008.

[37] R. Vidal, S. Soatto, and A. Chiuso, "Applications of hybrid system identification in computer vision," in *European Control Conference*, 2007.

[38] R. Dudley, E. Giné, and J. Zinn, "Uniform and universal glivenko-cantelli classes," *Journal of Theoretical Probability*, vol. 4, no. 3, pp. 485–510, 1991.

[39] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *Journal of the ACM*, vol. 44, no. 4, pp. 615–631, 1997.

[40] M. Kearns and R. Schapire, "Efficient distribution-free learning of probabilistic concepts," *Journal of Computer and System Sciences*, vol. 48, no. 3, pp. 464–497, 1994.

[41] S. Mendelson and R. Vershynin, "Entropy and the combinatorial dimension," *Inventiones mathematicae*, vol. 152, pp. 37–55, 2003.

[42] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[43] S. Mendelson, "Rademacher averages and phase transitions in Glivenko-Cantelli classes," *IEEE Transactions on Information Theory*, vol. 48, no. 1, pp. 251–263, 2002.

[44] P. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525–536, 1998.

[45] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, 2004.

[46] T. Zhang, "Covering number bounds of certain regularized linear function classes," *Journal of Machine Learning Research*, vol. 2, pp. 527–550, 2002.

[47] K. Musayeva, F. Lauer, and Y. Guermeur, "Rademacher complexity and generalization performance of multi-category margin classifiers," *Neurocomputing*, vol. 342, pp. 6–15, 2019.

[48] H. Hang, Y. Feng, I. Steinwart, and J. Suykens, "Learning theory estimates with observations from general stationary stochastic processes," *Neural Computation*, vol. 28, no. 12, pp. 2853–2889, 2016.

[49] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin, 1991.

[50] D. Haussler and P. M. Long, "A generalization of Sauer's lemma," *Journal of Combinatorial Theory, Series A*, vol. 71, no. 2, pp. 219–240, 1995.

[51] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. Long, "Characterizations of learnability for classes of $\{0, \ldots, N\}$-valued functions," *Journal of Computer and System Sciences*, vol. 50, no. 1, pp. 74–86, 1995.

[52] P. Bartlett and J. Shawe-Taylor, "Generalization performance of support vector machines and other pattern classifiers," in *Advances in Kernel Methods – Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. The MIT Press, Cambridge, MA, 1999, ch. 4, pp. 43–54.