Error bounds with almost radical dependence on the number of components for multi-category classification, vector quantization and switching regression

Fabien Lauer

Error bounds with almost radical dependence on the number of components for multi-category classification, vector quantization and switching regression

Fabien Lauer*[1]

[1]Université de Lorraine, CNRS, LORIA, F-54000 Nancy

June 4, 2018

Abstract

This paper presents a simple approach for obtaining efficient error bounds in learning problems involving multiple components. In particular, we obtain error bounds with a close to radical dependency on the number of components for multi-category classification, vector quantification and switching regression when the regularization scheme relies on a sum of norms over the components. These results are obtained thanks to the combination of a number of structural results on Rademacher complexities and a suitable handling of the structure of the regularizer.

Keywords: Learning theory, Rademacher complexity, Kernel methods, Regularization.

1 Introduction

This paper focuses on learning problems involving multiple components. A good example is multi-category classification with margin classifiers based on one score function (here called component) per category. Other examples include vector quantization/clustering and switching regression. In vector quantization, one is interested in estimating a model (or codebook) made of a finite number of components (or codepoints) that can well approximate the observations of a random variable. Switching regression works similarly, but with random input–output pairs and components that are functions approximating the output given the input. In this paper, we propose a unified approach to derive generalization error bounds with sublinear dependency on the number of components for all these problems.

*email: fabien.lauer@loria.fr

For such problems, the literature provides error bounds based on the analysis of Rademacher complexities [BM02]. These bounds are linear in the number C of components as they are derived from a decomposition of the Rademacher complexity of interest into a sum of Rademacher complexities over the component function classes, see [KMS14] for multi-category classification, [BDL08] for clustering and [Lau17] for switching regression. This dependency on the number of components can be made sublinear by using chaining arguments and covering numbers, see, e.g., [Gue17, MLG18, Lau17]. However, this comes at the cost of slower convergence rates in the sample size n, except for linear models in a finite-dimensional space where a dependency on the dimension is introduced, see, e.g., [BLL98].

These results typically consider that the model belongs to a product of independent component classes, usually containing elements with bounded norm, while many practical methods implement a regularization scheme corresponding to a bound on a sum of norms over the components. By exploiting this form of regularization more precisely, we show how to derive error bounds with a $O(\sqrt{C \log C})$ growth rate with respect to C without hindering the convergence rate. More precisely, our approach uses previous results by [KMS14, BDL08, Lau17] on the decompositions of the Rademacher complexities and leads to convergence rates similar to those obtained in these references, i.e., in $O(1/\sqrt{n})$.

For multi-category classification, our assumption on the regularization scheme was previously used in [LDBK15] in a more efficient manner, resulting in a growth rate that varies between $O(\log C)$ and $O(\sqrt{C})$ depending on the precise form of the regularization.

However, the derivation of this result requires additional tools and a more involved analysis, while our proofs remain shorter and more simple. In addition, the simplicity of the proposed approach allows its easy extension to other settings, whereas it is not known yet whether the technique of [LDBK15] also applies to vector quantization or switching regression.

For clustering/vector quantization and switching regression, it appears that such assumptions on the model class have not been considered in previous work to tighten the error bounds. However, note that for data living in a finite-dimensional Euclidean space, known bounds, see e.g., [BLL98], are in $O(\sqrt{C})$, so that we here focus on Hilbert spaces of large or infinite dimension. Specifically, when considering the finite-dimensional case we assume that the input data live in $\mathbb{R}^d$ with $d > \log C$. Indeed, our analysis shows that we can remove the radical dependency on d at the cost of an additional $\sqrt{\log C}$ factor.

We claim to provide a "simple" approach to derive our bounds. However, since this approach easily applies to many different settings, we present it in a rather abstract framework before instantiating it for classification, clustering or switching regression. As a result, the paper might not read as smoothly as it could have, but this is mostly due to notational complexities implied by the abstraction rather than true technical difficulties.

Paper organization. We start in Section 2 by formalizing our general framework and how it can be specified for the three problems of interest. Then, Section 3 presents the derivation of the error bounds, again, first in the general framework and next in a dedicated form for classification, vector quantization and switching regression. Concluding remarks are given in Section 4.

Notation. We denote by $[C]$ the set of integers from 1 to C. For two sets, $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{B}^{\mathcal{A}}$ denotes the the set of functions from $\mathcal{A}$ into $\mathcal{B}$. The notation $\|\cdot\|$ is used for the norm in the Hilbert space $\mathcal{H}$ induced by the inner product $\langle\cdot,\cdot\rangle$ in $\mathcal{H}$.

2 General setting

We focus on learning problems in which the aim is to learn $C \geq 2$ components from a Hilbert space $\mathcal{H}$ on the basis of data points $z_i \in \mathcal{Z}$, $i = 1, \dots n$. In the following, $\mathcal{Z}$ will be instantiated either as $\mathcal{X} \times \mathcal{Y}$ for problems with input space $\mathcal{X}$ and output space $\mathcal{Y}$ or just as $\mathcal{X}$ in contexts without outputs.

Specifically, let Z be a random variable taking values in $\mathcal{Z}$. A particular problem is characterized by a loss functional $\ell : \mathcal{Z} \times \mathcal{H}^C$, which measures the pointwise error of a model $g = (g_k)_{1 \leq k \leq C}$ made of C components g_k from $\mathcal{H}$. Then the aim is to minimize, over a predefined model class $\mathcal{G} \subset \mathcal{H}^C$, the risk

$$L(g) = \mathbb{E}\ell(Z, g) \tag{1}$$

on the basis of a sample of n independent copies Z_i of Z. In particular, we concentrate on the standard strategy minimizing the empirical risk

$$\frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, g). \tag{2}$$

However, we here focus on statistical aspects of learning and will not discuss algorithmic issues related to the actual minimization of this quantity, which can be highly nontrivial [AK98, ADHP09, Lau16].

For our approach to apply, we require that the loss can be computed (or upper bounded) through a convenient mapping $f : \mathcal{Z} \times \mathcal{H}^C \to \mathbb{R}$ of g, i.e., such that

$$\forall (z, g) \in \mathcal{Z} \times \mathcal{H}^C, \quad \ell(z, g) \leq \phi \circ f(z, g) \tag{3}$$

for some Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$. The meaning of "convenient" here shall become clear when we state our main technical result. Basically, the aim of f is twofold: to formulate the analysis in terms of Lipschitz functions and to allow efficient decompositions of the capacity measures.

Given (3), the empirical risk (2) is equal to or upper bounded by

$$\hat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \phi \circ f(Z_i, g),$$

which will be used as a proxy for (2) in the subsequent analysis. Similarly, the risk of any $g \in \mathcal{G}$ is upper bounded by $\mathbb{E}\phi \circ f(Z, g)$, which, in turn, can be bounded uniformly over $\mathcal{G}$ via an analysis of the complexity of the real-valued function class

$$\mathcal{F}_{\mathcal{G}} = \{f_g \in \mathbb{R}^{\mathcal{Z}} : f_g(z) = f(z, g),\ g \in \mathcal{G}\}. \tag{4}$$

In particular, our results will be targeted at model classes of the form

$$\mathcal{G} = \left\{ g \in \mathcal{H}^C : \sum_{k=1}^{C} \|g_k\|^p \leq \Lambda^p \right\}, \tag{5}$$

in which the vector of the norms of the components has a bounded ℓ_p-norm. Note that such constraints implement a coupling between the components and imply

that $\mathcal{G}$ cannot be written as a mere product of C independent classes.

We now describe a few example applications of this setting.

2.1 Margin multi-category classification

For classification problems with C categories, we focus our attention on the risk of margin classifiers $h : \mathcal{X} \to \mathcal{Y} = [C]$. Such classifiers are classifiers whose output in $[C]$ is based on a score per category computed by a component function $g_k : \mathcal{X} \to \mathbb{R}$ (thus, $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ in this setting). Specifically, we have

$$h(x) = \operatorname*{argmax}_{k\in[C]} g_k(x)$$

and, with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, the standard classification loss,

$$\begin{aligned}\ell_h(h(x), y) &= \mathbb{1}_{h(x)\neq y} \\ &= \mathbb{1}_{\operatorname{argmax}_{k\in[C]} g_k(x)\neq y} = \ell(z, g),\end{aligned} \tag{6}$$

can be upper bounded using the Lipschitz function

$$\phi(u) = \begin{cases} 1, & \text{if } u \leq 0 \\ 1 - \frac{u}{\gamma}, & \text{if } u \in (0, \gamma) \\ 0, & \text{if } u \geq \gamma \end{cases} \tag{7}$$

as

$$\begin{aligned}\ell(z, g) &\leq \mathbb{1}_{g_y(x) - \max_{k\neq y} g_k(x) \leq 0} \\ &\leq \phi\left(g_y(x) - \max_{k\neq y} g_k(x)\right).\end{aligned}$$

Note that the output of ϕ in (7) is not influenced by values of its argument outside of $[0, \gamma]$. Therefore, we can clip these values and define

$$f(z, g) = \min\left\{\gamma, \max\left\{0, g_y(x) - \max_{k\neq y} g_k(x)\right\}\right\} \tag{8}$$

to satisfy the requirement in (3).

Finally, note that the regularization scheme hinted at by (5) is commonly implemented in this context, for instance by multi-class support vector machines [WW98, CS01, LLW04, GM11, LDBK15], and most often with $p = 2$.

2.2 Vector quantization / clustering

Let $\mathcal{X}$ be a Hilbert space. The aim of vector quantization, as described in [BLL98], is to learn a subset $\{g_k\}_{k=1}^C \subset \mathcal{X}^C$ of C elements from $\mathcal{X}$, called codepoints, that can well represent the observations of the random variable $X \in \mathcal{X}$. Specifically, we can limit the analysis to nearest neighbors quantizers, for which the error of a model $g = (g_k)_{1\leq k\leq C}$ is measured via the loss

$$\ell(x, g) = \min_{k\in[C]} \|x - g_k\|^2. \tag{9}$$

Then, the quantity (1) (with $Z = X$) is known as the *distortion* of g for which upper bounds are of primary importance.

This problem can also be seen as a center-based clustering one, in which the goal is to divide the observations of X into C groups centered at the g_k's by minimizing the empirical risk (2) based on (9). By considering the Voronoï partition of $\mathcal{X}$ associated to these centers, [BDL08] interprets the quantity (1) as the *clustering risk* measuring the performance of a particular model $g \in \mathcal{X}^C$.

The setting just described enters our framework in a straightforward manner with $\mathcal{Z} = \mathcal{X}$, $\mathcal{H} = \mathcal{X}$, $f = \ell$ in (9) and ϕ set as the identity.

2.3 Switching regression

In a regression problem, one must learn a model that can accurately predict the real output $Y \in \mathcal{Y} \subseteq \mathbb{R}$ given the input $X \in \mathcal{X}$. Switching regression refers to the specific case where the process generating Y can arbitrarily switch between different behaviors. The difficulty then comes from the fact that the switchings are not observed and the association of the observations to these behaviors is unknown. Thus, the aim is to learn a collection of functions $g_k : \mathcal{X} \to \mathbb{R}$ from a mixed training sample including examples from multiple sources. An important application is that of switched system identification in control theory, see [PJFTV07, GPV12, LG18] for an overview.

In such a context, the goal is to find $g \in (\mathbb{R}^{\mathcal{X}})^C$ so that at least one of its components can accurately estimate the output Y given X. The loss can thus be defined on the basis of[1]

$$\min_{k\in[C]} (y - g_k(x))^2.$$

More precisely, we assume that $\mathcal{Y}$ is bounded and, without loss of generality, that $\mathcal{Y} = [-1/2, 1/2]$. Thus, we can clip the outputs of the components at $1/2$ without

[1] Other choices involving for instance absolute deviations instead squared errors are also possible here; see [Lau17] for a description of switching regression in a more general setting.

increasing the error and compute the loss with respect to the clipped functions as in [Lau17], i.e., by

$$\ell(z,g) = \min_{k\in[C]} \left(y - \min\left\{ \frac{1}{2}, \max\left\{ \frac{-1}{2}, g_k(x) \right\} \right\} \right)^2. \tag{10}$$

This ensures that the loss is bounded by 1 for all $y \in \mathcal{Y}$, which simplifies the analysis below.

The switching regression problem just described can be characterized in our framework with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, ϕ set as the identity and $f = \ell$ in (10).

In this context, the learning algorithms in, e.g., [LBL11, PLLL14], implement a regularization scheme in accordance with (5).

3 Error bounds

Here, we first derive our main technical result in a rather abstract framework before instantiating it for the three settings presented above in dedicated subsections. This result will allow us to bound the complexity of a function class under sum-of-norms constraints as in (5) via the maximum complexity of function classes related to the components (to be defined precisely below).

First, we need the following result showing that $\mathcal{G}$ can be embedded in a class $\overline{\mathcal{G}}$ built as a union of product classes with decaying components. More precisely, we mean that the Rademacher complexities of the component classes (see Definition 1 below) are decaying with the component index k. This result will allow us to work on a class with an easier to handle structure than that of $\mathcal{G}$.

The idea, made precise in the lemma below, is that if a function g belongs to $\mathcal{G}$, a large norm for one component must be compensated by a small norm for another. Thus, the component functions can be ordered so as to belong to a sequence of balls of decreasing radius. The union in the definition of $\overline{\mathcal{G}}$ implements the fact that different functions of $\mathcal{G}$ result in different orderings of their components that must all be taken into account when building an embedding of $\mathcal{G}$.

Lemma 1. *Let $\mathcal{G}$ be as in (5) and $(k_l(j))_{1\le l\le C} \in [C]^C$ denote the jth permutation of $[C]$ among $J = C!$, the first of which is just $[C]$. Then, for all $p > 0$, there are classes*

$$\overline{\mathcal{G}}_{1,k} = \left\{ g_k \in \mathcal{H} : \|g_k\| \le k^{-\frac{1}{p}}\Lambda \right\},\ k = 1,\dots,C,$$
$$\overline{\mathcal{G}}_j = \prod_{l=1}^{C} \overline{\mathcal{G}}_{1,k_l(j)},\ j = 1,\dots,J,$$

such that

$$\mathcal{G} \subset \overline{\mathcal{G}} = \bigcup_{j=1}^{J} \overline{\mathcal{G}}_j.$$

Proof. For any g, let $\tilde{g}$ be a version of g with its components reordered such that

$$\|\tilde{g}_1\| \ge \|\tilde{g}_2\| \ge \dots \ge \|\tilde{g}_C\|.$$

Then, the construction of $\overline{\mathcal{G}}$ is such that $\tilde{g} \in \overline{\mathcal{G}}$ implies that $\tilde{g}$ belongs to a product of some permutation of the $\overline{\mathcal{G}}_{1,k}$'s, which makes g also belong to a product of a (possibly different) permutation of the $\overline{\mathcal{G}}_{1,k}$'s. Thus, $\tilde{g} \in \overline{\mathcal{G}}$ implies $g \in \overline{\mathcal{G}}$. As a result, showing that for all $g \in \mathcal{G}$, $\tilde{g} \in \overline{\mathcal{G}}$ is sufficient to prove that $\mathcal{G} \subset \overline{\mathcal{G}}$.

With $\mathcal{G}$ as in (5), we have, for all $g \in \mathcal{G}$,

$$\sum_{k=1}^{C} \|g_k\|^p = \sum_{k=1}^{C} \|\tilde{g}_k\|^p \le \Lambda^p.$$

Thus, and by the definition of $\tilde{g}$, for any $l \in [C]$,

$$\Lambda^p \ge \sum_{k=1}^{C} \|\tilde{g}_k\|^p \ge \sum_{k=1}^{l} \|\tilde{g}_k\|^p \ge l\|\tilde{g}_l\|^p$$

and

$$\|\tilde{g}_l\| \le l^{-1/p}\Lambda,$$

which implies $\tilde{g}_l \in \overline{\mathcal{G}}_{1,l}$. Therefore, $\tilde{g} \in \overline{\mathcal{G}}_1 \subset \overline{\mathcal{G}}$ and the statement is proved.

□

Given this embedding of $\mathcal{G}$, we now aim at error bounds that hold uniformly over $\overline{\mathcal{G}}$ instead of $\mathcal{G}$. Despite the apparent loss due to the increase of the size of the class, the union structure of $\overline{\mathcal{G}}$ will help us to optimize the bound thanks to the following lemma, which basically states the simple fact that if a risk bound holds uniformly over $\overline{\mathcal{G}}_j$ for each j, then a bound of the same flavor holds uniformly over $\overline{\mathcal{G}}$. Specifically, we consider bounds based on Rademacher complexities.

Definition 1 (Rademacher complexity)**.** *Let T be a random variable with values in $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\boldsymbol{T}_n = (T_i)_{1\le i\le n}$ be an n-sample of independent copies of 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 with domain $\mathcal{T}$. The* Rademacher complexity *of $\mathcal{F}$ is*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(T_i),$$

where the expectation is taken over both $\boldsymbol{\sigma}_n$ and $\boldsymbol{T}_n$.

This measure of capacity allows for the derivation of the following generic bound.

Theorem 1 (After, e.g., Theorem 3.1 in [MRT12])**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{Z}$ into $[0,1]$ and $(Z_i)_{1\leq i\leq n}$ 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 $f \in \mathcal{F}$,*

$$\mathbb{E}f(Z) \leq \frac{1}{n}\sum_{i=1}^{n} f(Z_i) + 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Bounds of this flavor are used in the following to bound the risk uniformly over $\overline{\mathcal{G}}$.

Lemma 2. *Let $\overline{\mathcal{G}} = \bigcup_{j=1}^{J} \overline{\mathcal{G}}_j$ be a union of J sets $\overline{\mathcal{G}}_j$. Given a mapping $f : \mathcal{Z} \times \mathcal{H}^C \to \mathbb{R}$, define, for each one of these sets, $\mathcal{F}_{\overline{\mathcal{G}}_j}$ as in (4). If, for some constant $\alpha > 0$, for all $j \in [J]$ and all $\delta_j \in (0,1)$, the bound*

$$\forall g \in \overline{\mathcal{G}}_j, \quad L(g) \leq \hat{L}_n(g) + \alpha\mathcal{R}_n(\mathcal{F}_{\overline{\mathcal{G}}_j}) + \sqrt{\frac{\log\frac{1}{\delta_j}}{2n}} \quad (11)$$

holds with probability at least $1-\delta_j$, then, for any $\delta \in (0,1)$, the bound

$$\forall g \in \overline{\mathcal{G}}, \quad L(g) \leq \hat{L}_n(g) + \alpha \max_{j\in[J]} \mathcal{R}_n(\mathcal{F}_{\overline{\mathcal{G}}_j}) + \sqrt{\frac{\log\frac{J}{\delta}}{2n}}$$

holds with probability at least $1-\delta$.

Proof. Setting $\delta_j = \delta/J$ and using the union bound, we obtain that all the J bounds (11) hold simultaneously with probability at least $1 - \sum_{j=1}^{J}\delta_j = 1-\delta$. Then, for any $g \in \overline{\mathcal{G}}$, there is some $j \in [J]$ such that $g \in \overline{\mathcal{G}}_j$ and for which we can apply the bound (11). Upper bounding the corresponding Rademacher complexity by the maximum complexity over all $j \in [J]$ completes the proof. □

We can now state the main theorem, in which the "convenient" form of f that we require appears as a condition on the decomposition of its Rademacher complexity.

Theorem 2. *Let f be as in (3) and such that for any $\mathcal{G} = \prod_{k=1}^{C}\mathcal{G}_k$, for $\mathcal{F}_{\mathcal{G}}$ as in (4),*

$$\mathcal{R}_n(\mathcal{F}_{\mathcal{G}}) \leq c\sum_{k=1}^{C}\mathcal{R}_n(\Psi_{\mathcal{G}_k}),$$

where $c > 0$ is a constant and $\Psi_{\mathcal{G}_k}$ is a real-valued function class deduced from $\mathcal{G}_k$ via a mapping $\psi : \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ as

$$\Psi_{\mathcal{G}_k} = \{\psi_{g_k} \in \mathbb{R}^{\mathcal{Z}} : \psi_{g_k}(z) = \psi(z, g_k),\ g_k \in \mathcal{G}_k\}. \quad (12)$$

If ψ is such that, for any $\mathcal{G}_k = \{g_k \in \mathcal{H} : \|g_k\| \leq \Lambda\}$, the bound

$$\mathcal{R}_n\left(\Psi_{\mathcal{G}_k}\right) \leq \Lambda^q R(n) \quad (13)$$

holds with $q \geq 1$ and a function $R(n)$ that does not depend on Λ, then, for $\mathcal{G}$ as in (5), all $p \in (0,2]$ and any $\delta \in (0,1)$, the bound

$$\forall g \in \mathcal{G}, \quad L(g) < \hat{L}_n(g) + 4cL_\phi\Lambda^q\sqrt{C}R(n) + \sqrt{\frac{C\log C + \log\frac{1}{\delta}}{2n}}, \quad (14)$$

in which L_ϕ is the Lipschitz constant of ϕ, holds with probability at least $1-\delta$.

Proof. By Lemma 1, we have $\mathcal{G} \subset \overline{\mathcal{G}} = \bigcup_{j=1}^{J}\overline{\mathcal{G}}_j$ with $J = C! < C^C$. For any $\overline{\mathcal{G}}_j$ of Lemma 1, we have by (3) and Theorem 1 that, with probability at least $1-\delta$, for any $g \in \overline{\mathcal{G}}_j$,

$$L(g) \leq \frac{1}{n}\sum_{i=1}^{n}\phi\circ f(Z_i, g) + 2\mathcal{R}_n(\phi\circ\mathcal{F}_{\overline{\mathcal{G}}_j}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

In addition, by the contraction principle, see e.g., Lemma 4.2 in [MRT12],

$$\mathcal{R}_n(\phi\circ\mathcal{F}_{\overline{\mathcal{G}}_j}) \leq L_\phi\mathcal{R}_n(\mathcal{F}_{\overline{\mathcal{G}}_j}).$$

Thus, we can apply Lemma 2 with $\alpha = 2L_\phi$ and it remains to show that for all $p \in (0,2]$, $\max_{j\in[J]}\mathcal{R}_n(\mathcal{F}_{\overline{\mathcal{G}}_j}) \leq 2c\Lambda^q\sqrt{C}R(n)$. Using the assumptions and the fact that, for any $j \in [J]$, $\overline{\mathcal{G}}_j$ is the product $\prod_{k=1}^{C}\overline{\mathcal{G}}_{j,k}$, we have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_{\overline{\mathcal{G}}_j}) &\leq c\sum_{k=1}^{C}\mathcal{R}_n(\Psi_{\overline{\mathcal{G}}_{j,k}}) \\
&= c\sum_{l=1}^{C}\mathcal{R}_n(\Psi_{\overline{\mathcal{G}}_{1,k_l(j)}}) \\
&\leq c\Lambda^q R(n)\sum_{l=1}^{C}k_l(j)^{-q/p} \\
&= c\Lambda^q R(n)\sum_{k=1}^{C}k^{-q/p}.
\end{aligned}$$

Thus, for all $q \geq 1$ and $p \in (0,2]$,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_{\overline{\mathcal{G}}_j}) &\leq c\Lambda^q R(n)\sum_{k=1}^{C}\frac{1}{\sqrt{k}} \\
&\leq c\Lambda^q R(n)\int_1^{C+1}\frac{1}{\sqrt{t-1}}dt \\
&= 2c\Lambda^q R(n)\sqrt{C},
\end{aligned}$$

which completes the proof. □

Note that for specific values of $p < 2$, tighter bounds on the second term in the right-hand side of (14) are typically available. For instance, for $p = 1$, the $\sqrt{C}$ can be replaced by $(1 + \log C)$. However, the growth rate with respect to C remains determined by the last term in (14).

In the next subsections, we detail the consequences of Theorem 2 for the three considered settings.

3.1 Multi-category classification

In the multi-category classification setting of Sect. 2.1, the conditions of Theorem 2 are verified for kernel machines thanks to two results from the literature. The first one is a classical bound on the Rademacher complexity of balls in a reproducing kernel Hilbert space (RKHS) (see [BTA04] for definitions and properties of RKHSs):

Lemma 3 (After Lemma 22 in [BM02])**.** *Given an RKHS $\mathcal{H}$ of reproducing kernel K,*

$$\mathcal{R}_n(\{g_k \in \mathcal{H} : \|g_k\| \leq \Lambda\}) \leq \frac{\Lambda_x \Lambda}{\sqrt{n}},$$

where $\Lambda_x = \sup_{x \in \mathcal{X}} \|K(x, \cdot)\| = \sup_{x \in \mathcal{X}} \sqrt{K(x, x)}$.

The second one is a decomposition result due to [KMS14]:

Lemma 4 (After Theorem 2 in [KMS14])**.** *Let the mapping f be as in (8) and $\mathcal{F}_\mathcal{G}$ be defined as in (4). If $\mathcal{G} = \prod_{k=1}^C \mathcal{G}_k$, then* [2]

$$\mathcal{R}_n\left(\mathcal{F}_\mathcal{G}\right) \leq \sum_{k=1}^C \mathcal{R}_n\left(\mathcal{G}_k\right).$$

Therefore, with $\psi(z, g_k) = g_k(x)$, $q = 1$ and $L_\phi = \frac{1}{\gamma}$, Theorem 2 yields the following bound.

Theorem 3. *Let $\mathcal{H}$ be an RKHS of reproducing kernel K. Then, for the class $\mathcal{G}$ as in (5) and any $\delta \in (0, 1)$, the multi-class risk based on the loss (6) is bounded with probability at least $1 - \delta$, uniformly for all $g \in \mathcal{G}$, by*

$$L(g) \leq \hat{L}_n(g) + \frac{4\Lambda\Lambda_x}{\gamma}\sqrt{\frac{C}{n}} + \sqrt{\frac{C \log C + \log \frac{1}{\delta}}{2n}},$$

with Λ_x as in Lemma 3.

[2] Theorem 2 in [KMS14] actually states that $\mathcal{R}_n\left(\mathcal{F}_\mathcal{G}\right) \leq C\mathcal{R}_n\left(\bigcup_{k=1}^C \mathcal{G}_k\right)$, but the proof essentially allows for the derivation of the bound stated here.

Note that, under the assumptions of Theorem 3, a growth rate in C between logarithmic and radical (depending on p) can be obtained as detailed in [LDBK15] so that Theorem 3 does not constitute an improvement over the literature. However, for the most common case of $p = 2$, our result only adds a $\sqrt{\log C}$ factor compared to the one of [LDBK15] with a much shorter and simpler proof.

3.2 Clustering

We now consider the clustering case and the setting of Sect. 2.2. In this context, Theorem 2 yields a bound with almost radical dependency on C. The remaining ingredients can be found in the derivation of Theorem 2.1 in [BDL08].

Theorem 4. *Let $\mathcal{G}$ be as in (5) and assume that $P(\|X\| \leq \Lambda_x) = 1$. Then, for any $\delta \in (0, 1)$, the clustering risk based on the loss (9) is, with probability at least $1 - \delta$, uniformly bounded for all $g \in \mathcal{G}$ by*

$$L(g) \leq \hat{L}_n(g) + 4(2\Lambda_x + 1)\Lambda^q\sqrt{\frac{C}{n}} + \sqrt{\frac{C \log C + \log \frac{1}{\delta}}{2n}}$$

with $q = 1$ if $\Lambda \leq 1$ and $q = 2$ otherwise.

Proof. Given that $f = \ell$ (9), we can rewrite the functions $f_g \in \mathcal{F}_\mathcal{G}$ (4) as

$$\begin{aligned} f_g(x) &= \min_{k \in [C]} \|x - g_k(x)\|^2 \\ &= \|x\|^2 + \min_{k \in [C]} -2\langle x, g_k\rangle + \|g_k\|^2, \end{aligned}$$

which leads to

$$\begin{aligned} &\mathcal{R}_n\left(\mathcal{F}_\mathcal{G}\right) \\ &= \mathbb{E}\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i \left(\|X_i\|^2 + \min_{k \in [C]} -2\langle X_i, g_k\rangle + \|g_k\|^2\right) \\ &\leq \mathbb{E}\frac{1}{n}\sum_{i=1}^n \sigma_i \|X_i\|^2 \\ &\quad + \mathbb{E}\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i \min_{k \in [C]} -2\langle X_i, g_k\rangle + \|g_k\|^2. \end{aligned}$$

The fact that the σ_i's are centered and independent of the X_i's yields

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^n \sigma_i \|X_i\|^2 = 0.$$

Then, by following the proof of Lemma 4.3 in [BDL08] (or by using Lemma 8.1 in [MRT12]), we obtain that

$$\mathcal{R}_n\left(\mathcal{F}_{\mathcal{G}}\right) \leq \sum_{k=1}^{C} \mathcal{R}_n\left(\Psi_{\mathcal{G}_k}\right),$$

where $\Psi_{\mathcal{G}_k}$ is built as in (12) with $\psi(x, g_k) = -2\langle x, g_k\rangle + \|g_k\|^2$. In addition, [BDL08] also shows that for $\mathcal{G}_k = \{g_k \in \mathcal{H} : \|g_k\| \leq \Lambda\}$,

$$\begin{aligned}\mathcal{R}_n\left(\Psi_{\mathcal{G}_k}\right) &\leq 2\Lambda\sqrt{\frac{\mathbb{E}\|X\|^2}{n}} + \frac{\Lambda^2}{\sqrt{n}} \\ &\leq \Lambda^q \frac{2\sqrt{\mathbb{E}\|X\|^2}+1}{\sqrt{n}}\end{aligned}$$

with $q = 1$ if $\Lambda \leq 1$ and $q = 2$ otherwise.

Therefore, since $P(\|X\| \leq \Lambda_x) = 1$, $\mathbb{E}\|X\|^2 \leq \Lambda_x^2$ and Theorem 2 applies to this setting to yield the claimed risk bound. □

3.3 Switching regression

For the switching regression setting of Sect. 2.3, the conditions of Theorem 2 are fulfilled by making use of the analysis in [Lau17]. In particular, we rely on the following decomposition result.

Lemma 5 (After Theorem 3 in [Lau17])**.** *Let* $\mathcal{G} = \prod_{k=1}^{C} \mathcal{G}_k \subset \mathcal{H}^C$ *and* $\mathcal{F}_{\mathcal{G}}$ *be as in* (4) *with* $f = \ell$ (10)*. Then,*

$$\mathcal{R}_n\left(\mathcal{F}_{\mathcal{G}}\right) \leq 2\sum_{k=1}^{C} \mathcal{R}_n(\Psi_{\mathcal{G}_k}),$$

where $\Psi_{\mathcal{G}_k}$ *is as in* (12) *for the clipping function* ψ*:*

$$\psi(z, g_k) = \min\left\{\frac{1}{2}, \max\left\{\frac{-1}{2}, g_k(x)\right\}\right\}.$$

The condition (13) is in turn satisfied with a contraction argument and the application of Lemma 3, which yields, for any RKHS $\mathcal{H}$ and all $\mathcal{G}_k = \{g_k \in \mathcal{H} : \|g_k\| \leq \Lambda\}$,

$$\mathcal{R}_n(\Psi_{\mathcal{G}_k}) \leq \mathcal{R}_n(\mathcal{G}_k) \leq \frac{\Lambda\Lambda_x}{\sqrt{n}}.$$

Therefore, we can apply Theorem 2 with $c = 2$ and $q = 1$ to obtain a risk bound with almost radical dependence on C.

Theorem 5. *Let* $\mathcal{H}$ *be an RKHS of kernel* K *and set* $\Lambda_x = \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}$*. Then, for the class* $\mathcal{G}$ *as in* (5) *and any* $\delta \in (0,1)$*, the switching regression risk based on the minimal squared loss* (10) *is bounded with probability at least* $1-\delta$*, uniformly for all* $g \in \mathcal{G}$*, by*

$$L(g) \leq \hat{L}_n(g) + 8\Lambda_x\Lambda\sqrt{\frac{C}{n}} + \sqrt{\frac{C\log C + \log\frac{1}{\delta}}{2n}}.$$

4 Conclusions

The paper presented a unified approach to the derivation of error bounds for problems in which one is to learn C components from a Hilbert space. With this approach, and under appropriate assumptions on the model class, we could obtain an almost radical dependence on C (up to logarithmic factors) for multi-category classification, vector quantization and switching regression.

Future work will study the applicability of this approach to other problems involving multiple components, such as multi-task learning and structured prediction. In addition, recall that for multi-category classification, logarithmic to radical dependency on C can be obtained as in [LDBK15] via a somewhat more involved analysis of Gaussian complexities (that can be used to bound the Rademacher complexity). As hinted at in the introduction, future work will aim at extending this analysis to the other settings of clustering and switching regression in the hope of further reducing the influence of the number of components on the bounds. On the algorithmic side, a practical evaluation of methods implementing the regularization scheme suggested by our analysis in vector quantization should also be conducted.

References

[ADHP09] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

[AK98] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1–2):237–260, 1998.

[BDL08] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 2008.

[BLL98] P.L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44(5):1802–1813, 1998.

[BM02] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexi-

ties: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[BTA04] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, 2004.

[CS01] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[GM11] Y. Guermeur and E. Monfrini. A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica*, 22(1):73–96, 2011.

[GPV12] A. Garulli, S. Paoletti, and A. Vicino. A survey on switched and piecewise affine system identification. In *Proc. of the 16th IFAC Symp. on System Identification (SYSID)*, pages 344–355, 2012.

[Gue17] Y. Guermeur. L_p-norm Sauer-Shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017.

[KMS14] V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In *Advances in Neural Information Processing Systems 27*, pages 2501–2509, 2014.

[Lau16] F. Lauer. On the complexity of switching linear regression. *Automatica*, 74:80–83, 2016.

[Lau17] F. Lauer. Error bounds for piecewise smooth and switching regression. *arXiv preprint*, arXiv:1707.07938, 2017.

[LBL11] V.L. Le, G. Bloch, and F. Lauer. Reduced-size kernel models for nonlinear hybrid system identification. *IEEE Transactions on Neural Networks*, 22(12):2398–2405, 2011.

[LDBK15] Y. Lei, U. Dogan, A. Binder, and M. Kloft. Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems 28*, pages 2035–2043, 2015.

[LG18] F. Lauer and Bloch G. *Hybrid system identification: Theory and algorithms for learning switching models*. Springer, 2018. (to appear).

[LLW04] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[MLG18] K. Musayeva, F. Lauer, and Y. Guermeur. A sharper bound on the Rademacher complexity of margin multi-category classifiers. In *Proc. of the 28th Eur. Symp. on Artificial Neural Networks (ESANN)*, pages 503–508, 2018.

[MRT12] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, 2012.

[PJFTV07] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: a tutorial. *European Journal of Control*, 13(2-3):242–262, 2007.

[PLLL14] T. Pham Dinh, H.A. Le Thi, H.M. Le, and F. Lauer. A difference of convex functions algorithm for switched linear regression. *IEEE Transactions on Automatic Control*, 59(8):2277–2282, 2014.

[WW98] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, 1998.