Risk Bounds for Learning Multiple Components with Permutation-Invariant Losses

Fabien Lauer

Risk Bounds for Learning Multiple Components with Permutation-Invariant Losses

Fabien Lauer
Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

Abstract

This paper proposes a simple approach to derive efficient error bounds for learning multiple components with sparsity-inducing regularization. We show that for such regularization schemes, known decompositions of the Rademacher complexity over the components can be used in a more efficient manner to result in tighter bounds without too much effort. We give examples of application to switching regression and center-based clustering/vector quantization. Then, the complete workflow is illustrated on the problem of subspace clustering, for which decomposition results were not previously available. For all these problems, the proposed approach yields risk bounds with mild dependencies on the number of components and completely removes this dependence for nonconvex regularization schemes that could not be handled by previous methods.

1 INTRODUCTION

This paper focuses on learning problems involving multiple components. A good example is vector quantization (or center-based clustering), in which 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. Other examples include subspace clustering, where the data points are approximated by a collection of subspaces rather than codepoints, and switching regression, that 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 for all these problems which yields bounds with a mild dependence on the number of components for classes of interdependent components. While generalization might not be the primary goal in these problems, such error bounds can lead to model selection strategies and have been the subject of many studies, see, e.g., Bartlett et al. (1998); Biau et al. (2008); Lauer (2019) and references therein.

More precisely, we show how to efficiently take into account the invariance of the loss with respect to permutations of the components to derive risk bounds in multiple component learning problems. The proposed approach is simple and applies to different problems merely by plugging known decomposition results for these problems. For products of independent component classes, a *decomposition result* is one that decomposes the Rademacher complexity of the loss class into a sum of Rademacher complexities over the component classes. Previous works used such decompositions to obtain risk bounds that grow linearly with the number C of components. But for classes constrained in terms of a sparsity-inducing complexity measure, such as an ℓ_p-norm over the complexities of the components, our approach yields risk bounds with a dependence on C that varies for instance between $O(\sqrt{C})$ for $p = 2$ and $O(\log C)$ for $p = 1$. Such sparsity-inducing regularization schemes were already considered by Lei et al. (2015); Maurer (2016), where similar dependencies on the number of categories were obtained for multi-class classification. However, the method of Lei et al. (2015); Maurer (2016) relies on more complex arguments involving structural results on Rademacher and Gaussian complexities, duality, strong convexity and other tools developed by Kakade et al. (2012). Here, we develop the approach in Sect. 2 in a few lines with simple arguments and without invoking other tools. In addition, the proposed method also allows for the use of nonconvex regularization by ℓ_p-quasi-norms with $p \in (0, 1)$, which favors even sparser models. While the analysis of Lei et al. (2015); Maurer (2016) was limited to $p \geq 1$ and a

logarithmic dependence on C, our approach completely removes the dependence on C for nonconvex regularization with $p < 1$.

In Sect. 3, we apply our approach to switching regression, i.e., the problem of learning a collection of regression models from a mixed data set. For sparsity-inducing regularization schemes, this allows us to tighten the bounds of Lauer (2019) from a linear dependence on C to the ones discussed above for the different values of p. Similar results are obtained in Sect. 4 for vector quantization/clustering in Hilbert space, for which the literature only provides error bounds with either a radical dependence on C in the finite-dimensional case (Bartlett et al., 1998) or a linear one for infinite-dimensional Hilbert spaces (Biau et al., 2008). Finally, Section 5 is dedicated to the subspace clustering problem, which has a lot of applications in computer vision, for instance for motion segmentation or face clustering (Vidal, 2011; Vidal et al., 2016), but has not yet received much attention from the viewpoint of learning theory and risk bounds. This offers us the opportunity to illustrate the complete workflow for the application of the proposed approach.

Technically, our bounds are based on the analysis of the Rademacher complexity of the loss class to derive uniform risk bounds. More advanced tools, such as those of Bartlett et al. (2005) or Mendelson (2014), could be used to derive bounds with faster convergence rates or even for unbounded variables. However, these tools are particularly efficient to bound the risk of the empirical risk minimizer, which, for all the multiple component learning problems we consider, cannot be easily computed (and there is no satisfactory convex surrogate loss whose minimizer could be analyzed instead). Therefore, we must focus on *uniform* error bounds in order to apply them to the models returned by practical algorithms.

Notation. We use $[C] = \{1, \dots, C\}$ to refer to the set of integers from 1 to C. Matrices are written in bold and uppercase letters, while vectors are in non-bold lowercase letters. Random variables are written in uppercase letters. Thus, X will refer to a random vector, while $\boldsymbol{X}$ is a matrix. The identity matrix is denoted by $\boldsymbol{I}$. The Frobenius norm $\|\boldsymbol{A}\|_F$ of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of entries A_{ij} is defined as $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$. $\operatorname{Tr}(\boldsymbol{A})$ denotes the trace of the matrix $\boldsymbol{A}$ and we have $\|\boldsymbol{A}\|_F = \sqrt{\operatorname{Tr}(\boldsymbol{A}^\top \boldsymbol{A})}$. For a vector $a \in \mathbb{R}^C$ and any $p \in (0, \infty)$, $\|a\|_p = \left(\sum_{k=1}^C |a_k|^p\right)^{1/p}$ denotes its ℓ_p-norm for $p \geq 1$ or ℓ_p-quasi-norm for $p \in (0, 1)$, while $\|a\|_\infty = \max_{k \in [C]} |a_k|$ is its ℓ_∞-norm. Given two sets, $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{Y}^\mathcal{X}$ stands for the set of functions from $\mathcal{X}$ into $\mathcal{Y}$.

2 GENERAL APPROACH

We focus on learning problems in which the aim is to learn $C \geq 2$ components from a set $\mathcal{V}$ 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 $\ell : \mathcal{V}^C \times \mathcal{Z}$, which measures the pointwise error of a model $f = (f_k)_{1 \leq k \leq C}$ made of C components f_k from $\mathcal{V}$. Then, the aim is to minimize, over a predefined model class $\mathcal{F} \subset \mathcal{V}^C$, the risk

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

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

$$\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, Z_i). \tag{2}$$

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 (Aloise et al., 2009; Lauer, 2016). Instead, we will particularly pay attention to the derivation of upper bounds on the risk that hold *uniformly* over the class $\mathcal{F}$, and thus not only for the empirical risk minimizer which remains elusive in many practical cases.

Before we expose our approach to the derivation of such bounds, we first give a few definitions and start with the one that characterizes the losses considered in this paper.

Definition 1 (Permutation-invariant loss)**.** *A* permutation-invariant loss *over C components from a set $\mathcal{V}$ is a loss functional $\ell : \mathcal{V}^C \times \mathcal{Z}$ such that, for any permutation $(l(k))_{1\leq k\leq C}$ of $[C]$, any $f = (f_k)_{1\leq k\leq C} \in \mathcal{V}^C$ and any $z \in \mathcal{Z}$,*

$$\ell(f, z) = \ell((f_{l(k)})_{1\leq k\leq C}, z).$$

Definition 2 (Loss class)**.** *Given a bounded loss $\ell : \mathcal{V}^C \times \mathcal{Z} \to [0, M]$ and a class $\mathcal{F} \subset \mathcal{V}^C$, the loss class induced by $\mathcal{F}$ is*

$$\mathcal{L}_{\mathcal{F}} = \left\{\ell_f \in [0, M]^{\mathcal{Z}} : \ell_f(z) = \ell(f, z),\ f \in \mathcal{F}\right\}.$$

Definition 3 (Rademacher complexities)**.** *Let T be a random variable with values in $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\boldsymbol{T}_n = (T_i)_{1\leq i\leq n}$ be an n-sample of independent copies of T, let $\boldsymbol{\sigma}_n = (\sigma_i)_{1\leq i\leq n}$ 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* empirical Rademacher complexity *of $\mathcal{F}$ given $\boldsymbol{T}_n = \boldsymbol{t}_n = (t_i)_{1\leq i\leq n}$ is*

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

and its Rademacher complexity*, $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}\hat{\mathcal{R}}_n(\mathcal{F})$, is obtained by taking the expectation wrt. $\boldsymbol{T}_n$.*

The regularization schemes for learning multiple components that we consider are based on two levels of complexity measures. On the first level, let $\omega : \mathcal{V} \to [0, +\infty)$ be a complexity measure for a single component from $\mathcal{V}$ and, for any model $f \in \mathcal{V}^C$, let $\Omega(f) = (\omega(f_k))_{1\leq k\leq C}$ denote the vector of $\mathbb{R}^C$ obtained by a component-wise application of ω to f. Then, at a second level, we measure the complexity of the overall model f by the ℓ_p-(quasi-)norm of $\Omega(f)$. Therefore, in this paper we will focus on the derivation of error bounds for classes

$$\mathcal{F} = \left\{f \in \mathcal{V}^C : \|\Omega(f)\|_p \leq \Lambda\right\}. \tag{3}$$

Definition 4 (Ordered class $\tilde{\mathcal{F}}$)**.** *Given a complexity measure ω as defined above, we denote by $\tilde{f}$ an* ordered version *of $f \in \mathcal{V}^C$ with its components ordered in decreasing order of their complexity:*

$$\forall f = (f_k)_{1\leq k\leq C} \in \mathcal{V}^C, \qquad \tilde{f} = (f_{l(k)})_{1\leq k\leq C},$$

where $l(k)$ is the kth element of a permutation of $[C]$ that ensures

$$\omega(\tilde{f}_1) \geq \cdots \geq \omega(\tilde{f}_C).$$

Then, for any class $\mathcal{F} \subset \mathcal{V}^C$, the ordered class $\tilde{\mathcal{F}}$ *is defined by reordering the elements of $\mathcal{F}$:*

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

Note that for classes built as $\mathcal{F} = \mathcal{F}_0^C = \mathcal{F}_0 \times \cdots \times \mathcal{F}_0$ for some $\mathcal{F}_0 \subset \mathcal{V}$, the ordered class $\tilde{\mathcal{F}}$ is a subset of $\mathcal{F}$: $\forall f \in \mathcal{F}_0^C,\quad \tilde{f} \in \mathcal{F}_0^C$. This is also true for classes $\mathcal{F}$ as in (3), which introduce a dependence between components, encoded in the choice of ℓ_p-(quasi-)norm. For instance, if we let $p = \infty$, then $\mathcal{F}$ in (3) can be written as a product of independent component classes:

$$\begin{aligned}\mathcal{F}_\infty &= \left\{f \in \mathcal{V}^C : \max_{k\in[C]} \omega(f_k) \leq \Lambda\right\} \\ &= \prod_{k=1}^{C} \{f_k \in \mathcal{V} : \omega(f_k) \leq \Lambda\}.\end{aligned} \tag{4}$$

But if we consider $p \in (0, \infty)$, then $\mathcal{F}$ in (3) cannot be written as a mere product, since the complexity $\omega(f_k)$ influences the range of values allowed for $\omega(f_j)$, $j \neq k$. For such classes, the ordered class is a strict subset of $\mathcal{F}$: $\tilde{\mathcal{F}} \subset \mathcal{F}$, and $\tilde{\mathcal{F}} \neq \mathcal{F}$. The inclusion results from the permutation-invariance of the ℓ_p-norm: $\|\Omega(\tilde{f})\|_p = \|\Omega(f)\|_p \leq \Lambda$; and this also implies that there are some $f \in \mathcal{F}$ with $\omega(f_2) > \omega(f_1)$ and thus that do not belong to $\tilde{\mathcal{F}}$.

The interest of the ordered class $\tilde{\mathcal{F}}$ and the fact that it is a subset of $\mathcal{F}$ is highlighted by the following, which shows that for permutation-invariant losses, we can restrict the analysis to this subset of $\mathcal{F}$.

Lemma 1. *Given a bounded permutation-invariant loss $\ell : \mathcal{V}^C \times \mathcal{Z} \to [0, M]$ and a class $\mathcal{F} \subset \mathcal{V}^C$, the risk of any $f \in \mathcal{F}$ can be bounded in terms of the Rademacher complexity of the loss class induced by the ordered class $\tilde{\mathcal{F}}$ instead of $\mathcal{F}$, namely, each of the following holds with probability at least $1 - \delta$:*

$$\forall f \in \mathcal{F}, \quad L(f) \leq \hat{L}_n(f) + 2\mathcal{R}_n(\mathcal{L}_{\tilde{\mathcal{F}}}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2n}},$$

$$\forall f \in \mathcal{F}, \quad L(f) \leq \hat{L}_n(f) + 2\hat{\mathcal{R}}_n(\mathcal{L}_{\tilde{\mathcal{F}}}) + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Proof. By Definitions 1 and 4, $L(f) = L(\tilde{f})$ and $\hat{L}_n(f) = \hat{L}_n(\tilde{f})$. Therefore, the lemma is just a direct consequence of standard error bounds, e.g., Theorem 3.1 in Mohri et al. (2012), holding uniformly over the ordered class $\tilde{\mathcal{F}}$ instead of $\mathcal{F}$. □

For classes $\mathcal{F}$ as in (3) with dependent components, a second interest lies in the fact that $\tilde{\mathcal{F}}$ can be easily embedded in a product of independent component classes with decreasing complexity:

Lemma 2. *Let $\mathcal{F}$ be as in (3) with $p \in (0, \infty]$. Then,*

$$\tilde{\mathcal{F}} \subseteq \Pi_p = \prod_{k=1}^{C} \left\{ f_k \in \mathcal{V} : \omega(f_k) \leq k^{-\frac{1}{p}} \Lambda \right\}.$$

Proof. Assume $p < \infty$ (see (4) for the case $p = \infty$). Then, with $\mathcal{F}$ as in (3), the permutation-invariance of the ℓ_p-norm implies that, for all $f \in \mathcal{F}$,

$$\|\Omega(\tilde{f})\|_p^p = \|\Omega(f)\|_p^p \leq \Lambda^p,$$

while, for any $k \in [C]$,

$$\|\Omega(\tilde{f})\|_p^p = \sum_{l=1}^{C} \omega(\tilde{f}_l)^p \geq \sum_{l=1}^{k} \omega(\tilde{f}_l)^p \geq k\omega(\tilde{f}_k)^p,$$

where the last inequality is due to the ordering of the $\tilde{f}_k$'s in Def. 4. Therefore, for all $\tilde{f} \in \tilde{\mathcal{F}}$ and all $k \in [C]$,

$$\omega(\tilde{f}_k)^p \leq \frac{\Lambda^p}{k},$$

which proves the claimed set inclusion. □

Note that for $p = \infty$ the product class Π_p in Lemma 2 is exactly $\mathcal{F}$ due to (4), whereas for all finite p, Π_p is strictly larger than $\mathcal{F}$ and thus $\tilde{\mathcal{F}}$: there exist f in the product of component classes with $\sum_{k=1}^{C} \omega(f_k)^p > \Lambda^p$ that are thus not in $\mathcal{F}$ and not in $\tilde{\mathcal{F}} \subset \mathcal{F}$. Therefore, the inclusion provided by Lemma 2 is not tight, but its interest lies at another level, namely, the fact that decomposition results available for products of independent classes can help us to bound the Rademacher complexity of $\mathcal{L}_{\tilde{\mathcal{F}}}$.

Instead of deriving a generic framework with cumbersome notations that would encompass many different settings but would also hide the simplicity of the approach, the following illustrates the application of the method on a few examples. In particular, we detail below the settings of switching regression and center-based clustering, for which decomposition results can be found in the literature. Then, we will show in Sect. 5 how to develop the complete workflow for subspace clustering from the definition of the loss function to the derivation of efficient bounds, including the obtention of a decomposition result.

For all these settings we shall derive error bounds with a dependence on C characterized by p via the function

$$\alpha(C, p) = \begin{cases} C, & \text{if } p = \infty \\ \frac{p}{p-1} C^{1-1/p}, & \text{if } 1 < p < \infty \\ 1 + \log C, & \text{if } p = 1 \\ \frac{1}{1-p}, & \text{if } 0 < p < 1. \end{cases} \tag{5}$$

In particular, the dependence on C will be linear for $p = \infty$ (the case of independent component classes), radical for $p = 2$ (the most common case), logarithmic for $p = 1$ (a common choice for sparsity-inducing regularization) and bounded by a constant for $p < 1$ (corresponding to nonconvex regularizers).

3 SWITCHING REGRESSION

In a regression problem, one must learn a model that can accurately predict the real output $Y \in \mathcal{Y} \subset \mathbb{R}$ 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 data points $(x_i, y_i) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to these behaviors is unknown. Thus, the aim is to learn a collection of functions $f_k : \mathcal{X} \to \mathbb{R}$ from a mixed training sample including examples from multiple sources. An important application is that of switched system identification in control theory, see Paoletti et al. (2007); Lauer and Bloch (2019) for an overview.

In such a context, the goal is to find $f \in (\mathbb{R}^{\mathcal{X}})^C$ 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

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

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 increasing the error and compute the loss with respect to the clipped functions as in Lauer (2019):

$$\ell(f, x, y) = \min_{k \in [C]} \left(y - \min \left\{ \frac{1}{2}, \max \left\{ \frac{-1}{2}, f_k(x) \right\} \right\} \right)^2. \tag{6}$$

This ensures that the loss is bounded by 1 for all $y \in \mathcal{Y}$. In addition, it is easy to see that this loss remains permutation-invariant in the sense of Definition 1.

Here, we focus on kernel machines and consider models with components from a reproducing kernel Hilbert space (RKHS) $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ of reproducing kernel K (see Berlinet and Thomas-Agnan (2004) for details). Thus, we set $\mathcal{V} = \mathcal{H}$ and the complexity measure ω to the RKHS norm $\|\cdot\|$ in the approach described above, which yields the risk bound in Theorem 1 below for classes regularized by $\|\Omega(f)\|_p = (\sum_{k=1}^C \|f_k\|^p)^{1/p}$.

Theorem 1. *Let* $\mathcal{F} = \left\{ f \in \mathcal{H}^C : \left\| \begin{bmatrix} \|f_1\| & \dots & \|f_C\| \end{bmatrix} \right\|_p \leq \Lambda \right\}$ *and* $\alpha(C, p)$ *be as in* (5). *Then, with probability at least* $1 - \delta$ *on the random draw of the training sample* $(Z_i)_{1 \leq i \leq n} = ((X_i, Y_i))_{1 \leq i \leq n}$, *the switching regression risk based on the loss* (6) *is uniformly bounded for all* $f \in \mathcal{F}$ *by*

$$L(f) \leq \hat{L}_n(f) + 4\alpha(C, p) \frac{\Lambda \sqrt{\sum_{i=1}^n K(X_i, X_i)}}{n} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Proof. By the permutation-invariance of ℓ in (6), we can apply Lemma 1 and the result follows from the computation of the (empirical) Rademacher complexity of $\mathcal{L}_{\tilde{\mathcal{F}}}$. Then, Lemma 2 gives $\tilde{\mathcal{F}} \subseteq \Pi_p$ and thus $\mathcal{L}_{\tilde{\mathcal{F}}} \subseteq \mathcal{L}_{\Pi_p}$, which further yields

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

Since Π_p is a product of independent component classes, the decomposition result in Theorem 3 of Lauer (2019) then gives

$$\hat{\mathcal{R}}_n(\mathcal{L}_{\tilde{\mathcal{F}}}) \leq 2 \sum_{k=1}^C \hat{\mathcal{R}}_n \left(\left\{ f_k \in \mathcal{H} : \|f_k\| \leq k^{-\frac{1}{p}} \Lambda \right\} \right),$$

while standard computations for RKHS balls (see, e.g., Bartlett and Mendelson (2002)) further ensure that

$$\hat{\mathcal{R}}_n(\mathcal{L}_{\tilde{\mathcal{F}}}) \leq \frac{2\Lambda\sqrt{\sum_{i=1}^n K(X_i,X_i)}}{n} \sum_{k=1}^C k^{-\frac{1}{p}}.$$

Thus, the theorem is proved after a straightforward check that $\sum_{k=1}^C k^{-\frac{1}{p}} \leq \alpha(C,p)$ holds for all $C \geq 2$ and $p \in (0,\infty]$ (see Appendix A for details). □

For independent component classes ($p = \infty$), this result coincides with that in Eq. (18) of Lauer (2019). However, for $p < \infty$, the dependence on C improves according to the definition of $\alpha(C,p)$ in (5). In particular, a radical dependence is obtained for $p = 2$, which could only be obtained in Lauer (2019) through covering numbers and a loss in the order of $\log^{3/2} n$ in terms of convergence rate. In addition, the dependence on C further improves for smaller values of p.

4 VECTOR QUANTIZATION/CLUSTERING

Let $\mathcal{X}$ be a Hilbert space and $\|\cdot\|$ denote its norm. The aim of vector quantization, as described by Bartlett et al. (1998), is to learn a subset $\{f_k\}_{k=1}^C \subset \mathcal{X}$ 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 $f = (f_k)_{1\leq k\leq C}$ is measured via the loss

$$\ell(f,x) = \min_{k\in[C]} \|x - f_k\|^2. \tag{7}$$

Then, the quantity (1) (with $Z = X$) is known as the *distortion* of f 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 f_k's by minimizing the empirical risk (2) based on (7). By considering the Voronoï partition of $\mathcal{X}$ associated to these centers, Biau et al. (2008) interpret the quantity (1) as the *clustering risk* measuring the performance of a particular model $f \in \mathcal{X}^C$.

The setting just described enters our framework in a straightforward manner with $\mathcal{V} = \mathcal{Z} = \mathcal{X}$ and $\omega = \|\cdot\|$. We can thus easily obtain efficient bounds on the clustering risk for regularized classes on the basis of the results of Biau et al. (2008).

Theorem 2. *Let $X \in \mathcal{X}$ be such that $P(\|X\| \leq \Lambda_x) = 1$, $\mathcal{F} = \{f \in \mathcal{X}^C : \left\|\left[\|f_1\| \quad \dots \quad \|f_C\|\right]\right\|_p \leq \Lambda\}$ and $\alpha(C,p)$ be as in (5). Then, with probability at least $1-\delta$ on the random draw of the training sample $(X_i)_{1\leq i\leq n}$, the clustering risk based on the loss (7) is uniformly bounded for all $f \in \mathcal{F}$ by*

$$L(f) \leq \hat{L}_n(f) + 2\alpha(C,p)\left(\frac{2\Lambda\sqrt{\sum_{i=1}^n \|X_i\|^2}}{n} + \frac{\Lambda^2}{\sqrt{n}}\right) + 3(\Lambda_x^2 + \Lambda^2)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Proof. It is easy to see that the clustering loss (7) is permutation-invariant in the sense of Definition 1 and uniformly bounded by $M = \Lambda_x^2 + \Lambda^2$. Thus, as for the switching regression case, we can apply Lemmas 1 and 2. Then, it remains only to show that $\hat{\mathcal{R}}_n(\mathcal{L}_{\Pi_p})$ is smaller than $\sum_{k=1}^C k^{-\frac{1}{p}}$ times a term independent of k, C and p, in order to conclude with the use of $\sum_{k=1}^C k^{-\frac{1}{p}} \leq \alpha(C,p)$ (see Appendix A).

This can be done by following the proof of Theorem 2.1 in Biau et al. (2008), which includes both a decomposition result and the computation of the Rademacher complexity of the loss class for Hilbert space balls. In fact, the statements in Biau et al. (2008) do not concern the empirical version of Rademacher complexity and focus on products of similar classes so that the result is C times the Rademacher complexity wrt. a single component. However, Biau et al. (2008) give all the ingredients to obtain the result in the form stated here. For completeness, we give the details

in Appendix B, which lead to

$$\hat{\mathcal{R}}_n(\mathcal{L}_{\Pi_p}) \leq \sum_{k=1}^{C} \left(\frac{2k^{-\frac{1}{p}}\Lambda\sqrt{\sum_{i=1}^n \|X_i\|^2}}{n} + \frac{k^{-\frac{2}{p}}\Lambda^2}{\sqrt{n}} \right) \tag{8}$$

$$\leq \left(\frac{2\Lambda\sqrt{\sum_{i=1}^n \|X_i\|^2}}{n} + \frac{\Lambda^2}{\sqrt{n}} \right) \sum_{k=1}^{C} k^{-\frac{1}{p}}. \tag{9}$$

□

As for switching regression, this result encompasses for $p = \infty$ the case of independent component classes found in Biau et al. (2008). For $p < \infty$, the improved bound could also have been obtained by following the approach of Lei et al. (2015) or Maurer (2016), which is also very efficient for regularized classes constrained by $\sum_{k=1}^{C} \|f_k\|^p \leq \Lambda^p$. However, as highlighted in the introduction, this would have required $p \geq 1$ and a much heavier machinery, whereas our approach remains simple and provides a proof of Theorem 2 also valid for nonconvex regularizers with $p \in (0, 1)$ as an almost direct consequence of previous decomposition results.

5 SUBSPACE CLUSTERING

Subspace clustering differs from center-based clustering in that the components f_k are subspaces of $\mathcal{X}$ instead of points. In the following, we drop the notation f_k and instead focus on the subspace basis in the form of matrices $\boldsymbol{B}_k \in \mathbb{R}^{d \times d_k}$.

Our starting point in Sect. 5.1 is a uniform bound on the error when learning a single subspace. Then, we extend this to multiple subspaces in Sect. 5.2 and finally tighten the bound for classes defined by ℓ_p-norm regularization in Sect. 5.3.

5.1 Uniform Error Bounds for Subspace Estimation

A d_1-dimensional subspace of $\mathbb{R}^d$ can be represented by a basis $\{b_1, \ldots, b_{d_1}\} \subset \mathbb{R}^d$, i.e., by a matrix $\boldsymbol{B} \in \mathbb{R}^{d \times d_1}$ with $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}$, which yields the projection matrix $\boldsymbol{P} = \boldsymbol{B}\boldsymbol{B}^\top$. Then, the approximation error incurred by the projection of a point x onto the subspace is measured by the loss

$$\ell(\boldsymbol{B}, x) = \|\boldsymbol{P}x - x\|^2 = \|\boldsymbol{B}\boldsymbol{B}^\top x - x\|^2. \tag{10}$$

We are interested here in bounding the expected approximation error (or risk), $L(\boldsymbol{B}) = \mathbb{E}\ell(\boldsymbol{B}, X)$, in terms of its empirical estimation, $\hat{L}_n(\boldsymbol{B}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{B}, X_i)$, for any distribution of X over $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq \Lambda_x\}$ and *uniformly* over the class of d_1-dimensional subspaces of $\mathbb{R}^d$ with basis in

$$\mathcal{B} = \left\{ \boldsymbol{B} \in \mathbb{R}^{d \times d_1} : \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I} \right\}. \tag{11}$$

This can be done as follows (see Appendix C for the proof).

Theorem 3. *Let $X \in \mathbb{R}^d$ be a random vector such that $P(\|X\| \leq \Lambda_x) = 1$. Then, with probability at least $1-\delta$ on the random draw of a data matrix $\boldsymbol{X} = [X_1, \ldots, X_n] \in \mathbb{R}^{d \times n}$ made of n independent copies of X, for any subspace of dimension d_1 and any basis $\boldsymbol{B}$ of that subspace,*

$$L(\boldsymbol{B}) \leq \hat{L}_n(\boldsymbol{B}) + 2\frac{\sqrt{d_1}\|\boldsymbol{X}\|_F}{n} + 3\Lambda_x^2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Note that this bound is uniform and is of the same order as the non-uniform one obtained by Shawe-Taylor et al. (2005).

5.2 Multiple Subspace Learning/Subspace Clustering

We now consider the problem of learning multiple subspaces, represented by basis $\boldsymbol{B}_k \in \mathbb{R}^{d\times d_k}$ and projection matrices $\boldsymbol{P}_k$, $k = 1, \dots, C$, to obtain an approximation of the distribution of X. This setting extends the vector quantization framework to models with subspace components and can be formally encoded by the loss

$$\ell((\boldsymbol{B}_k)_{1\le k\le C}, x) = \min_{k\in[C]} \|\boldsymbol{B}_k\boldsymbol{B}_k^\top x - x\|^2. \tag{12}$$

In this context, the *subspace clustering risk*, $L(\boldsymbol{B}) = \mathbb{E}\ell(\boldsymbol{B}, X)$, of a collection $\boldsymbol{B} = (\boldsymbol{B}_k)_{1\le k\le C}$ of subspace basis $\boldsymbol{B}_k$ can be bounded in terms of the sum of the square roots of the subspace dimensions as follows.

Theorem 4. *Let $X \in \mathbb{R}^d$ be a random vector such that $P(\|X\| \le \Lambda_x) = 1$. Then, with probability at least $1-\delta$ on the random draw of a data matrix $\boldsymbol{X} = [X_1, \dots, X_n] \in \mathbb{R}^{d\times n}$ made of n independent copies of X, for any collection of basis $\boldsymbol{B}$ of subspaces with fixed dimensions d_k,*

$$L(\boldsymbol{B}) \le \hat{L}_n(\boldsymbol{B}) + 2\frac{\sum_{k=1}^C \sqrt{d_k}\|\boldsymbol{X}\|_F}{n} + 3\Lambda_x^2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Proof. Define the loss class $\mathcal{L}_{\mathcal{B}}$ as in Definition 2 from

$$\mathcal{B} = \prod_{k=1}^C \mathcal{B}_k, \quad \text{with } \mathcal{B}_k = \left\{\boldsymbol{B}_k \in \mathbb{R}^{d\times d_k} : \boldsymbol{B}_k^\top \boldsymbol{B}_k = \boldsymbol{I}\right\}.$$

Then, its complexity can be decomposed as a sum of those of classes induced by the $\mathcal{B}_k$'s. To see this, note that, with $\boldsymbol{P}_k = \boldsymbol{B}_k\boldsymbol{B}_k^\top$, the loss can be reformulated as

$$\ell((\boldsymbol{B}_k)_{1\le k\le C}, x) = \|x\|^2 - \max_{k\in[C]} \|\boldsymbol{P}_k x\|^2.$$

Thus, given $(X_i)_{1\le i\le n} = (x_i)_{1\le i\le n}$,

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{B}}) &= \mathbb{E}\sup_{\boldsymbol{B}\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^n \sigma_i \min_{k\in[C]} \|\boldsymbol{P}_k x_i - x_i\|^2 \\
&\le \mathbb{E}\frac{1}{n}\sum_{i=1}^n \sigma_i\|x_i\|^2 + \mathbb{E}\sup_{\boldsymbol{B}\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^n -\sigma_i \max_{k\in[C]} \|\boldsymbol{P}_k x_i\|^2 \\
&= \mathbb{E}\sup_{\boldsymbol{B}\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^n \sigma_i \max_{k\in[C]} \|\boldsymbol{P}_k x_i\|^2 \\
&\le \sum_{k=1}^C \mathbb{E}\sup_{\boldsymbol{B}_k\in\mathcal{B}_k} \frac{1}{n}\sum_{i=1}^n \sigma_i\|\boldsymbol{P}_k x_i\|^2,
\end{aligned}$$

where the third line uses $\mathbb{E}\frac{1}{n}\sum_{i=1}^n \sigma_i\|x_i\|^2 = \frac{1}{n}\sum_{i=1}^n \|x_i\|^2\mathbb{E}\sigma_i = 0$ and the fact that σ_i and $-\sigma_i$ share the same distribution, while the last line is due to Lemma 8.1 in Mohri et al. (2012). Then, similar computations as in the proof of Theorem 3 (see Appendix C) give, for any $k \in [C]$,

$$\mathbb{E}\sup_{\boldsymbol{B}_k\in\mathcal{B}_k} \frac{1}{n}\sum_{i=1}^n \sigma_i\|\boldsymbol{P}_k x_i\|^2 \le \frac{\sqrt{d_k}\|\boldsymbol{X}\|_F}{n}$$

and the result follows from the application of Theorem 3.1 in Mohri et al. (2012) and the fact that the loss defined as the pointwise minimum of losses bounded by Λ_x^2 is also bounded by Λ_x^2. □

Theorem 4 applies to products of independent component classes, which here means that the dimensions of the subspaces do not depend one on the other, and yields a linear dependence on C. The next result below yields tighter bounds by precisely taking dependencies between the dimensions into account.

5.3 Tighter Bounds with ℓ_p-norm Regularization

We have now all the basic building blocks necessary to apply the approach of Sect. 2 and produce tighter bounds for subspace clustering. Specifically, we set $\omega(f_k) = \sqrt{d_k}$ and focus on the set of basis collections with ℓ_p-norm regularization:

$$\mathcal{B}_p = \left\{ \boldsymbol{B} = (\boldsymbol{B}_k)_{1 \leq k \leq C} : \boldsymbol{B}_k \in \mathbb{R}^{d \times d_k}, \boldsymbol{B}_k^\top \boldsymbol{B}_k = \boldsymbol{I}, \left\| \begin{bmatrix} \sqrt{d_1} & \dots & \sqrt{d_C} \end{bmatrix} \right\|_p \leq \Lambda \right\}.$$

Theorem 5. *Let $X \in \mathbb{R}^d$ be a random vector such that $P(\|X\| \leq \Lambda_x) = 1$ and $\alpha(C,p)$ be as in (5). Then, with probability at least $1-\delta$ on the random draw of a data matrix $\boldsymbol{X} = [X_1, \dots, X_n] \subset \mathbb{R}^{d \times n}$ made of n independent copies of X, for any collection of subspace basis $\boldsymbol{B} \in \mathcal{B}_p$,*

$$L(\boldsymbol{B}) \leq \hat{L}_n(\boldsymbol{B}) + 2\alpha(C,p)\frac{\Lambda \|\boldsymbol{X}\|_F}{n} + 3\Lambda_x^2 \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Proof. First, note that the subspace clustering loss (12) is permutation-invariant according to Def. 1. Thus, Lemmas 1 and 2 apply and it remains only to bound $\hat{\mathcal{R}}_n(\mathcal{L}_{\Pi_p})$ with

$$\Pi_p = \prod_{k=1}^{C} \left\{ \boldsymbol{B}_k \in \mathbb{R}^{d \times d_k} : \boldsymbol{B}_k^\top \boldsymbol{B}_k = \boldsymbol{I},\ \sqrt{d_k} \leq k^{-\frac{1}{p}} \Lambda \right\}.$$

Here, the proof of Theorem 4 provides us with

$$\hat{\mathcal{R}}_n(\mathcal{L}_{\Pi_p}) \leq \frac{\sum_{k=1}^{C} \sqrt{d_k} \|\boldsymbol{X}\|_F}{n} \leq \frac{\Lambda \|\boldsymbol{X}\|_F}{n} \sum_{k=1}^{C} k^{-\frac{1}{p}}$$

and plugging $\sum_{k=1}^{C} k^{-\frac{1}{p}} \leq \alpha(C,p)$ (see Appendix A) completes the proof. □

Thus, we recover bounds for subspace clustering with similar dependencies on the main parameters (C and n) as those obtained for switching regression and center-based clustering. Again, we emphasize that once a bound was found for products of independent component classes with a linear dependence on C (Theorem 4), our approach easily yielded mild dependencies for classes with dependent components.

6 CONCLUSIONS

The paper presented a simple approach to derive risk bounds with mild dependence on the number C of components for classes with interdependent components. Only two ingredients are needed to obtain such results with the proposed approached: a permutation-invariant loss and a bound holding for products of independent component classes and providing a decomposition of their Rademacher complexity into a sum of the component complexities.

Future work will consider the application of the proposed approach to other settings and permutation-invariant losses. The new bounds for subspace clustering could also lead to novel model selection strategies in order to tune the number of subspaces and their dimensions from the data.

A USEFUL BOUNDS

We show here that, for any integer $C \geq 2$ and $p \in (0, \infty]$, with $\alpha(C,p)$ as defined in (5),

$$\sum_{k=1}^{C} k^{-\frac{1}{p}} \leq \alpha(C,p).$$

For $p = \infty$, we easily see that $\sum_{k=1}^{C} k^{-\frac{1}{p}} = \sum_{k=1}^{C} 1 = C$. For $p < \infty$, we can write

$$\sum_{k=1}^{C} k^{-\frac{1}{p}} = 1 + \sum_{k=2}^{C} k^{-\frac{1}{p}} \leq 1 + \int_1^C x^{-\frac{1}{p}} dx.$$

Then, for $p = 1$, we have

$$\sum_{k=1}^{C} k^{-\frac{1}{p}} \leq 1 + \int_1^C \frac{1}{x} dx = 1 + \log C - \log 1 = 1 + \log C,$$

while for $p \neq 1$, we have

$$\sum_{k=1}^{C} k^{-\frac{1}{p}} \leq 1 + \frac{p}{p-1}(C^{(p-1)/p} - 1) = \frac{pC^{1-1/p} - 1}{p-1}.$$

So for $p > 1$, we get

$$\sum_{k=1}^{C} k^{-\frac{1}{p}} < \frac{pC^{1-1/p}}{p-1},$$

while for $p < 1$, we obtain

$$\sum_{k=1}^{C} k^{-\frac{1}{p}} \leq \frac{1 - pC^{1-1/p}}{1-p} \leq \frac{1}{1-p}.$$

B COMPLEMENTS FOR THE PROOF OF THEOREM 2

We here restate the results embedded in the proof of Theorem 2.1 in Biau et al. (2008) with empirical Rademacher complexities and a summation over the component classes, as used in the proof of Theorem 2. First, we reformulate the clustering loss as

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

which, for $\Pi_p = \prod_{k=1}^{C} \Pi_{p,k}$ and given $(X_i)_{1 \leq i \leq n} = (x_i)_{1 \leq i \leq n}$, leads to

$$\begin{aligned}\hat{\mathcal{R}}_n\left(\mathcal{L}_{\Pi_p}\right) &= \mathbb{E} \sup_{f \in \Pi_p} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\|x_i\|^2 + \min_{k \in [C]} -2 \langle x_i, f_k \rangle + \|f_k\|^2 \right) \\ &\leq \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \|x_i\|^2 + \mathbb{E} \sup_{f \in \Pi_p} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \min_{k \in [C]} -2 \langle x_i, f_k \rangle + \|f_k\|^2 \\ &= \mathbb{E} \sup_{f \in \Pi_p} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \max_{k \in [C]} 2 \langle x_i, f_k \rangle - \|f_k\|^2 \\ &\leq \sum_{k=1}^{C} \mathbb{E} \sup_{f_k \in \Pi_{p,k}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (2 \langle x_i, f_k \rangle - \|f_k\|^2),\end{aligned}$$

where the last line is due to Lemma 8.1 in Mohri et al. (2012). Then, with $\Lambda_k = k^{-1/p}\Lambda$, for any $k \in [C]$,

$$\begin{aligned}\mathbb{E} \sup_{f_k \in \Pi_{p,k}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (2 \langle x_i, f_k \rangle - \|f_k\|^2) &\leq 2\mathbb{E} \sup_{f_k \in \Pi_{p,k}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \langle x_i, f_k \rangle + \mathbb{E} \sup_{f_k \in \Pi_{p,k}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \|f_k\|^2 \\ &\leq 2\mathbb{E} \sup_{f_k \in \Pi_{p,k}} \frac{1}{n} \left\langle \sum_{i=1}^{n} \sigma_i x_i, f_k \right\rangle + \frac{\Lambda_k^2}{\sqrt{n}} \\ &\leq 2\frac{\Lambda_k}{n} \mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i x_i \right\| + \frac{\Lambda_k^2}{\sqrt{n}} \leq 2\frac{\Lambda_k}{n} \sqrt{\sum_{i=1}^{n} \|x_i\|^2} + \frac{\Lambda_k^2}{\sqrt{n}}.\end{aligned}$$

The second inequality, i.e., (9), is merely due to the fact that $k^{-2/p} \leq k^{-1/p}$ for all $k \geq 1$.

C PROOF OF THEOREM 3

Since $\boldsymbol{P} = \boldsymbol{B}\boldsymbol{B}^\top$ is a projection matrix, it is symmetric and idempotent: $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$. Thus,

$$\begin{aligned}\ell(\boldsymbol{B}, x) &= \|\boldsymbol{P}x - x\|^2 = x^\top \boldsymbol{P}^\top \boldsymbol{P}x - 2x^\top \boldsymbol{P}x + x^\top x \\ &= -x^\top \boldsymbol{P}x + \|x\|^2 = \|x\|^2 - \|\boldsymbol{P}x\|^2.\end{aligned}$$

Hence, the loss is bounded with probability one as $0 \leq \ell(\boldsymbol{B}, X) \leq \|X\|^2 \leq \Lambda_x^2$ and standard error bounds such as Theorem 3.1 in Mohri et al. (2012) apply to the loss class based on (10) and (11),

$$\mathcal{L}_{\mathcal{B}} = \left\{ \ell \in [0, \Lambda_x^2]^{\mathcal{X}} : \ell(x) = \|\boldsymbol{B}\boldsymbol{B}^\top x - x\|^2,\ \boldsymbol{B} \in \mathcal{B} \right\}.$$

Then, the statement is a consequence of the estimation of the empirical Rademacher complexity of $\mathcal{L}_{\mathcal{B}}$ given $(X_i)_{1 \leq i \leq n} = (x_i)_{1 \leq i \leq n}$:

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{B}}) &= \mathbb{E} \sup_{\boldsymbol{B} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\|x_i\|^2 - \|\boldsymbol{P}x_i\|^2) \\ &\leq \mathbb{E} \frac{1}{n} \sum_{i=1}^n \sigma_i \|x_i\|^2 + \mathbb{E} \sup_{\boldsymbol{B} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n -\sigma_i \|\boldsymbol{P}x_i\|^2,\end{aligned}$$

where $\mathbb{E} \frac{1}{n} \sum_{i=1}^n \sigma_i \|x_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|x_i\|^2 \mathbb{E}\sigma_i = 0$ and $-\sigma_i$ has the same distribution has σ_i. Thus, using $\|\boldsymbol{P}x_i\|^2 = x_i^\top \boldsymbol{P}x_i = \mathrm{Tr}(x_i^\top \boldsymbol{P}x_i) = \mathrm{Tr}(\boldsymbol{P}x_i x_i^\top)$, we obtain

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{B}}) &\leq \mathbb{E} \sup_{\boldsymbol{B} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \sigma_i \,\mathrm{Tr}(\boldsymbol{P}x_i x_i^\top) \\ &= \mathbb{E} \sup_{\boldsymbol{B} \in \mathcal{B}} \frac{1}{n} \,\mathrm{Tr}\left(\boldsymbol{P} \left(\sum_{i=1}^n \sigma_i x_i x_i^\top \right) \right) \\ &\leq \mathbb{E} \sup_{\boldsymbol{B} \in \mathcal{B}} \frac{1}{n} \|\boldsymbol{P}\|_F \left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_F,\end{aligned}$$

where

$$\begin{aligned}\left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_F^2 &= \mathrm{Tr}\left(\left(\sum_{i=1}^n \sigma_i x_i x_i^\top \right) \left(\sum_{i=1}^n \sigma_i x_i x_i^\top \right) \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j \,\mathrm{Tr}\left(x_i x_i^\top x_j x_j^\top \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j \,\mathrm{Tr}\left((x_i^\top x_j)^2 \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j (x_i^\top x_j)^2.\end{aligned}$$

In addition, since the trace of an idempotent matrix equals its rank and $\mathrm{rank}(\boldsymbol{B}) = \mathrm{rank}(\boldsymbol{B}\boldsymbol{B}^\top)$, we have

$$\|\boldsymbol{P}\|_F = \sqrt{\mathrm{Tr}(\boldsymbol{P}^\top \boldsymbol{P})} = \sqrt{\mathrm{Tr}(\boldsymbol{P})} = \sqrt{\mathrm{rank}(\boldsymbol{P})} = \sqrt{\mathrm{rank}(\boldsymbol{B})} = \sqrt{d_1}.$$

Thus,

$$\hat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{B}}) \leq \frac{1}{n}\mathbb{E}\sqrt{d_1 \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_i\sigma_j (x_i^\top x_j)^2}$$
$$\leq \frac{1}{n}\sqrt{d_1 \sum_{i=1}^{n} \|x_i\|^2} = \frac{\sqrt{d_1}\|\boldsymbol{X}\|_F}{n}.$$

References

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

Bartlett, P., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

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

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

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

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

Kakade, S., Shalev-Shwartz, S., and Tewari, A. (2012). Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890.

Lauer, F. (2016). On the complexity of switching linear regression. *Automatica*, 74:80–83.

Lauer, F. (2019). Error bounds for piecewise smooth and switching regression. *IEEE Transactions on Neural Networks and Learning Systems*. In press.

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

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

Maurer, A. (2016). A vector-contraction inequality for Rademacher complexities. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 3–17.

Mendelson, S. (2014). Learning without concentration. In *Proc. of the Conference on Learning Theory (COLT)*, pages 25–39.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press, Cambridge, MA.

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

Shawe-Taylor, J., Williams, C., Cristianini, N., and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522.

Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68.

Vidal, R., Ma, Y., and Sastry, S. (2016). *Generalized Principal Component Analysis*. Springer-Verlag New York.