Upper and Lower Bounds on the Performance of Kernel PCA

Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, John Shawe-Taylor

Upper and Lower Bounds on the Performance of Kernel PCA

Maxime Haddouche MAXIME.HADDOUCHE@ENS-PARIS-SACLAY.FR
Inria and ENS Paris-Saclay
London, United Kingdom

Benjamin Guedj BENJAMIN.GUEDJ@INRIA.FR
Inria and University College London
London, United Kingdom

Omar Rivasplata O.RIVASPLATA@UCL.AC.UK
DeepMind and University College London
London, United Kingdom

John Shawe-Taylor J.SHAWE-TAYLOR@UCL.AC.UK
University College London
London, United Kingdom

Abstract

Principal Component Analysis (PCA) is a popular method for dimension reduction and has attracted an unfailing interest for decades. Recently, kernel PCA has emerged as an extension of PCA but, despite its use in practice, a sound theoretical understanding of kernel PCA is missing. In this paper, we contribute lower and upper bounds on the efficiency of kernel PCA, involving the empirical eigenvalues of the kernel Gram matrix. Two bounds are for fixed estimators, and two are for randomized estimators through the PAC-Bayes theory. We control how much information is captured by kernel PCA on average, and we dissect the bounds to highlight strengths and limitations of the kernel PCA algorithm. Therefore, we contribute to the better understanding of kernel PCA. Our bounds are briefly illustrated on a toy numerical example.

Keywords: Statistical learning theory, kernel PCA, PAC-Bayes, dimension reduction.

1. Introduction

Principal Component Analysis (PCA) is a celebrated dimension reduction method. It was first described by Pearson (1901); and it was developed further by several authors (see *e.g.* Jolliffe, 1986, and references therein). In a nutshell, PCA summarises high-dimensional data $(x_1, ..., x_m) \in \mathbb{R}^d$, $m \in \mathbb{N}^*$, into a smaller space, which is designed to be 'meaningful' and more easily interpretable. By 'meaningful' we mean that this new subspace still captures efficiently the correlations between data points, while at the same time reducing drastically the dimension of the space. A popular tool to design this meaningful subspace is the *Gram Matrix* of the data, defined as $(\langle x_i, x_j \rangle)_{i,j}$. PCA then considers the eigenvectors of this matrix. Note that this is a linear operation, in the sense that PCA consists of an orthogonal transformation of the coordinate system in which we describe our data, followed by a projection onto the first k directions in the new system, corresponding to the largest k eigenvalues of the Gram matrix.

Over the past two decades, PCA has been studied and enriched (*e.g.*, principal curves as a nonlinear extension of PCA, as done by Guedj and Li, 2018). The particular extension of PCA that we focus on is 'kernel PCA' (which may be traced back to Schölkopf et al., 1998). Using a kernel, we map our data into a reproducing kernel Hilbert space[1] (RKHS). The linear PCA then operates in this Hilbert space to yield a finite-dimensional subspace onto which we project new data points. The final step is to assess how close from the original data is this projection. Kernel PCA is widely used in the machine learning literature (*e.g.*, Kim and Klabjan, 2020; Xu et al., 2019; Vo and Durlofsky, 2016, to name but a few recent works) which makes the need of a better theoretical understanding even more pressing.

A first theoretical study has been made in Shawe-Taylor et al. (2005) who derived PAC (Probably Approximately Correct) guarantees for kernel PCA. The PAC bounds proposed by Shawe-Taylor et al. (2005) were set up to control the averaged projection of new data points onto a finite-dimensional subspace of the RKHS into which data is embedded.

Bounds in Shawe-Taylor et al. (2005) include a Rademacher complexity. Rademacher complexity terms are known to be challenging to compute in many settings as they typically blow up combinatorially. To provide more numerically friendly results, we investigate a different route than Shawe-Taylor et al. (2005) and introduce the first PAC-Bayesian study of kernel PCA which, as a byproduct, allows to replace the Rademacher term by a Kullback-Leibler divergence (which has a closed form when distributions are Gaussian, and can be approximated by Monte Carlo in other cases). PAC-Bayes theory is a powerful framework to study generalisation properties of randomised predictors, and was introduced in the seminal works of Shawe-Taylor and Williamson (1997); McAllester (1998, 1999). PAC-Bayes theory has then been developed further by Seeger (2002); McAllester (2003); Maurer (2004); Catoni (2007) among others. PAC-Bayes has emerged in the past few years as one of the promising leads to study generalisation properties of deep neural networks (Dziugaite and Roy, 2017; Letarte et al., 2019). A surge of recent papers in a variety of different settings illustrates the flexibility and relevance of PAC-Bayes as a principled tool (Amit and Meir, 2018; Dziugaite and Roy, 2018a,b; Rivasplata et al., 2018, 2020; Holland, 2019; Haddouche et al., 2020; Nozawa et al., 2020; Mhammedi et al., 2019, 2020; Cantelobre et al., 2020). We refer to the recent survey Guedj (2019) and tutorial Guedj and Shawe-Taylor (2019), and the paper Rivasplata et al. (2020), for details on PAC-Bayes theory.

Our contributions. We aim at PAC and PAC-Bayesian bounds on the performance of kernel PCA. We provide empirical PAC bounds which improve on those from Shawe-Taylor et al. (2005). We introduce the first PAC-Bayes lower and upper bounds for kernel PCA, which clarify the merits and limitations of the overall method. These results are unprecedented, to the best of our knowledge.

Outline. We introduce our notation and recall existing theoretical results on kernel PCA in Section 2. Section 3 contains two new PAC bounds for kernel PCA, and Section 4 is devoted to two new PAC-Bayes bounds, along with our proofs. The paper closes with a brief illustration of the numerical value of our bounds (Section 5) and concluding remarks (Section 6). We gather proofs of technical results in Section 7.

1. We refer the reader to Hein and Bousquet (2004) or Hofmann et al. (2005) for an introduction to RKHS and their uses in machine learning.

2. Notation and preliminaries

We let $\mathbb{R}^{m\times n}$ denote the space of matrices of shape $m \times n$ with real entries. The data space is $\mathcal{X} \subseteq \mathbb{R}^d$. We assume to have access to $s = (x_1, \ldots, x_m) \in \mathcal{X}^m$, a realisation of the size-m random vector $S = (X_1, \ldots, X_m) \in \mathcal{X}^m$.

We let $\mathcal{M}_1(\mathcal{X})$ denote the space of probability distributions over $\mathcal{X}$ and $\mu \in \mathcal{M}_1(\mathcal{X})$ stands for the distribution that generates one random example $X \in \mathcal{X}$. Its empirical counterpart is given by $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$, *i.e.*, the empirical distribution defined by the random sample. We assume the collected sample to be independent and identically distributed (iid): $S \sim \mu^m$, where $\mu^m = \mu \otimes \cdots \otimes \mu$ (m copies).

$\mathbb{E}_\nu[f] = \mathbb{E}_{X\sim\nu}[f(X)] = \int_{\mathcal{X}} f(x)\nu(dx)$ denotes the expectation under $\nu \in \mathcal{M}_1(\mathcal{X})$, for $f : \mathcal{X} \to \mathbb{R}$. We denote by $\mathcal{H}$ a (separable) Hilbert space, equipped with an inner product $\langle\cdot,\cdot\rangle$. We let $\|u\| = \langle u, u\rangle^{1/2}$ be the norm of $u \in \mathcal{H}$. The operator $P_V : \mathcal{H} \to \mathcal{H}$ is the orthogonal projection onto a subspace V, and $P_v = P_{\mathrm{span}\{v\}}$. In what follows, $\mathcal{F}$ is a set of predictors, and $\pi, \pi^0 \in \mathcal{M}_1(\mathcal{F})$ represent probability distributions over $\mathcal{F}$. Finally, $\mathbb{E}_\pi[L] = \mathbb{E}_{f\sim\pi}[L(f)] = \int_{\mathcal{F}} L(f)\pi(df)$ is the expectation under $\pi \in \mathcal{M}_1(\mathcal{F})$, for $L : \mathcal{F} \to \mathbb{R}$.

On Reproducing Kernel Hilbert Spaces (RKHS). We recall results from Hein and Bousquet (2004) on the links between RKHS and different mathematical structures.

Let us start by a primer on kernels. The key idea is that while data belongs to a data space $\mathcal{X}$, a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ implicitly embeds data into a Hilbert space (of real-valued functions), where there is an abundance of structure to exploit. Such a function κ is required to be *symmetric* in the sense that $\kappa(x_1, x_2) = \kappa(x_2, x_1)$ for all $x_1, x_2 \in \mathcal{X}$.

Definition 1 (PSD kernels) *A symmetric real-valued function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a* positive semi definite *(PSD) kernel if $\forall n \geq 1$, $\forall x_1, ..., x_n \in \mathcal{X}$, $\forall c_1, ..., c_n \in \mathbb{R}$:*

$$\sum_{i,j=1}^{n} c_i c_j \kappa(x_i, x_j) \geq 0.$$

If the inequality is strict (for non-zero coefficients $c_1, \ldots, c_n$), then the kernel is said to be positive definite *(PD).*

For instance, polynomial kernels $\kappa(x, y) = (x^T y + r)^n$, and Gaussian kernels $\kappa(x, y) = \exp(-||x - y||^2/2\sigma^2)$ (for $n \geq 1, (x, y) \in (\mathbb{R}^d)^2, r \geq 0, \sigma > 0$) are PD kernels.

Definition 2 (RKHS) *A* reproducing kernel Hilbert space *(RKHS) on $\mathcal{X}$ is a Hilbert space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ (functions from $\mathcal{X}$ to $\mathbb{R}$) where all evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, defined by $\delta_x(f) = f(x)$, are continuous.*

Note that, since the evaluation functionals are linear, an equivalent condition to continuity (of all the δ_x's) is that for every $x \in \mathcal{X}$, there exists $M_x < +\infty$ such that

$$\forall f \in \mathcal{H}, \quad |f(x)| \leq M_x ||f||.$$

This condition is the so-called *reproducing property*. When $\mathcal{H}$ is an RKHS over $\mathcal{X}$, there is a kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a mapping $\psi : \mathcal{X} \to \mathcal{H}$ such that $\kappa(x_1, x_2) = \langle \psi(x_1), \psi(x_2)\rangle_{\mathcal{H}}$.

Intuitively: the 'feature mapping' ψ maps data points $x \in \mathcal{X}$ to 'feature vectors' $\psi(x) \in \mathcal{H}$, while the kernel computes the inner products between those feature vectors without needing explicit knowledge of ψ.

The following key theorem from Aronszajn (1950) links PD kernels and RKHS.

Theorem 3 (Moore-Aronszajn, 1950) *If κ is a positive definite kernel, then there exists a unique reproducing kernel Hilbert space $\mathcal{H}$ whose kernel is κ.*

In Hein and Bousquet (2004), a sketch of the proof is provided: from a PD kernel κ, we build an RKHS from the pre-Hilbert space

$$V = \text{span}\left\{\kappa(x, .) \mid x \in \mathcal{X}\right\}.$$

We endow V with the following inner product:

$$\left\langle \sum_i a_i \kappa(x_i, .), \sum_j b_j \kappa(x_j, .) \right\rangle_V = \sum_{i,j} a_i b_j \kappa(x_i, x_j). \tag{1}$$

It can be shown that this is indeed a well-defined inner product. Thus, the rest of the proof consists in the completion of V into an Hilbert space (of functions) verifying the reproducing property, and the verification of the uniqueness of such an Hilbert space.

A important special case is when $|\mathcal{X}| < +\infty$, then V is a finite-dimensional vector space. Thus if we endow it with an inner product, V is already an Hilbert space (it already contains every pointwise limits of all Cauchy sequences of elements of V). As a consequence, the associated RKHS is finite-dimensional in this case.

Definition 4 (Aronszajn mapping) *For a fixed PD kernel κ, we define the* Aronszajn mapping $\psi : \mathcal{X} \to (\mathcal{H}, \langle \cdot, \cdot \rangle)$ *such that*

$$\forall x \in \mathcal{X}, \quad \psi(x) = \kappa(x, .),$$

where we denote by $\mathcal{H}$ the RKHS given by the Moore-Aronszajn theorem and $\langle \cdot, \cdot \rangle$ is the inner product given in (1). In the sequel, we refer to the Aronszajn mapping as the pair $(\psi, \mathcal{H})$ when it is important to highlight the space $\mathcal{H}$ into which ψ embeds the data.

When it comes to embedding points of $\mathcal{X}$ into a Hilbert space through a feature mapping ψ, several approaches have been considered (see Hein and Bousquet, 2004, Section 3.1). The Aronszajn mapping is one choice among many.

On kernel PCA. We present here the results from Shawe-Taylor et al. (2005). Fix a PD kernel κ. We denote by $(\psi, \mathcal{H})$ the Aronszajn mapping of $\mathcal{X}$ into $\mathcal{H}$.

Definition 5 *The kernel Gram matrix of a data set $s = (x_1, \ldots, x_m) \in \mathcal{X}^m$ is the element $K(s)$ of $\mathbb{R}^{m \times m}$ defined as*

$$K(s) = (\kappa(x_i, x_j))_{i,j}.$$

When the data set is clear from the context, we will shorten this notation to $K = (\kappa(x_i, x_j))_{i,j}$.

Note that for a random sample $S = (X_1, \ldots, X_m)$, the corresponding $K(S)$ is a random matrix.

The goal of kernel PCA is to analyse K by putting the intricate data sample of size m from the set $\mathcal{X}$ into $\mathcal{H}$, where data are properly separated, and then find a small (in terms of dimension) subspace V of $\mathcal{H}$ which catches the major part of the information contained in the data. We define $\mu \in \mathcal{M}_1(\mathcal{X})$, a probability measure over $\mathcal{X}$, as the distribution representing the way data are spread out over $\mathcal{X}$.

In other words, we want to find a subspace $V \subseteq \mathcal{H}$ such as

$$\forall x \in \mathcal{X}, \quad \left| ||P_V(\psi(x))||^2 - ||\psi(x)||^2 \right| \approx 0$$

where P_V is the orthogonal projection over the subspace V.

The notation $[m] = \{1, \ldots, m\}$ is convenient. Recall that ψ and $\mathcal{H}$ are defined such that we can express the elements of K as a scalar product in $\mathcal{H}$:

$$\forall i, j \in [m], \quad K_{i,j} = \kappa(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{H}}.$$

Definition 6 *For any probability distribution ν over $\mathcal{X}$, we define the self-adjoint operator on $L^2(\mathcal{X}, \nu)$ associated to the kernel function κ as:*

$$\mathcal{K}_\nu(f)(x) = \int_{\mathcal{X}} f(x')\kappa(x, x')\nu(dx').$$

Definition 7 *We use the following conventions:*

- *If μ is the data-generating distribution, then we rename $\mathcal{K} := \mathcal{K}_\mu$.*
- *If $\hat{\mu}$ is the empirical distribution of our m-sample $(x_i)_i$, then we name $\hat{\mathcal{K}} := \mathcal{K}_{\hat{\mu}}$.*
- *$\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of the operator $\mathcal{K}$.*
- *$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_m \geq 0$ are the eigenvalues of the kernel matrix K.*

More generally, let $\lambda_1(A) \geq \lambda_2(A) \geq \cdots$ be the eigenvalues of a matrix A, or a linear operator A. Then in Definition 7, we use the shortcuts $\lambda_i = \lambda_i(\mathcal{K})$ and $\hat{\lambda}_i = \lambda_i(K)$. Notice that $\forall i \in \{1, \ldots, m\}$, we have $\lambda_i(\hat{\mathcal{K}}) = \frac{\hat{\lambda}_i}{m}$.

Definition 8 *For a given sequence of real scalars $(a_i)_{i \geq 1}$ of length m, which may be infinity, we define for any k the initial sum and the tail sum as*

$$a^{\leq k} := \sum_{i=1}^{k} a_i \quad \text{and} \quad a^{>k} := \sum_{i=k+1}^{m} a_i.$$

Definition 9 *The* sample covariance matrix *of a random data set $S = (X_1, \ldots, X_m)$ is the element $C(S)$ of $\mathbb{R}^{m \times m}$ defined by*

$$C(S) = \frac{1}{m} \sum_{i=1}^{m} \psi(X_i)\psi(X_i)',$$

where $\psi(x)'$ denotes the transpose of $\psi(x)$. Notice that this is the sample covariance in feature space. When S is clear from the context, we will shorten $C(S)$ to C.

One could object that C may not be finite-dimensional, because $\mathcal{H}$ is not (in general). However, notice that the subspace of $\mathcal{H}$ spanned by $\psi(x_1), ..., \psi(x_m)$ is always finite-dimensional, hence by choosing a basis of this subspace, C becomes effectively a finite-dimensional square matrix (of size no larger than $m \times m$).

Definition 10 *For any probability distribution ν over $\mathcal{X}$, we define $\mathcal{C}_\nu : \mathcal{H} \to \mathcal{H}$ as the mapping $\alpha \mapsto \mathcal{C}_\nu(\alpha)$ where:*

$$\mathcal{C}_\nu(\alpha) = \int_{\mathcal{X}} \langle \psi(x), \alpha \rangle \psi(x) \nu(dx).$$

If ν has density $v(x)$, then we write $\mathcal{C}_v(\alpha) = \int_{\mathcal{X}} \langle \psi(x), \alpha \rangle \psi(x) v(x) dx$. Notice that the eigenvalues of $\mathcal{K}_\nu$ and $\mathcal{C}_\nu$ are the same for any ν, the proof of this fact may be found in Shawe-Taylor et al. (2005). We similarly define the simplified notations $\mathcal{C} := \mathcal{C}_\mu$ (when μ is the population distribution) and $\hat{\mathcal{C}} = \mathcal{C}_{\hat{\mu}}$ (when $\hat{\mu}$ is the empirical distribution). We then define for any $k \in \{1, \dots, m\}$

- *V_k the subspace spanned by the k-first eigenvectors of $\mathcal{C}$,*
- *$\hat{V}_k$ the subspace spanned by the k-first eigenvectors of $\hat{\mathcal{C}}$.*

Notice that $\hat{\mathcal{C}}$ coincides with the sample covariance matrix C, *i.e.* we have

$$\hat{\mathcal{C}}(\alpha) = C\alpha, \qquad \forall \alpha \in \mathcal{H}.$$

So for any $k > 0$, $\hat{V}_k$ is the subspace spanned by the first k eigenvectors of the matrix C.

Proposition 11 (Courant-Fischer's corollary) *If $(u_i)_i$ are the eigenvectors associated to $(\lambda_i(\mathcal{K}_\nu))_i$ and V_k is the space spanned by the k first eigenvectors:*

$$\lambda_k(\mathcal{K}_\nu) = \mathbb{E}_\nu[||P_{u_k}(\psi(x))||^2] = \max_{\dim(V)=k} \min_{0 \neq v \in V} \mathbb{E}_\nu[||P_v(\psi(x))||^2],$$

$$\lambda^{\leq k}(\mathcal{K}_\nu) = \max_{\dim(V)=k} \mathbb{E}_\nu[||P_V(\psi(x))||^2],$$

$$\lambda^{>k}(\mathcal{K}_\nu) = \min_{\dim(V)=k} \mathbb{E}_\nu[||P_V^\perp(\psi(x))||^2].$$

Now we will denote by $\mathbb{E}_\mu$ the expectation under the true data-generating distribution μ and by $\mathbb{E}_{\hat{\mu}}$ the expectation under the empirical distribution of an m-sample S. Combining the last properties gives us the following equalities.

Proposition 12 *We have*

$$\mathbb{E}_{\hat{\mu}}[||P_{\hat{V}_k}(\psi(x))||^2] = \frac{1}{m} \sum_{i=1}^{m} ||P_{\hat{V}_k}(\psi(x_i))||^2 = \frac{1}{m} \sum_{i=1}^{k} \hat{\lambda}_i,$$

$$\mathbb{E}_\mu[||P_{V_k}(\psi(x))||^2] = \sum_{i=1}^{k} \lambda_i.$$

We now recall the main results from Shawe-Taylor et al. (2005) before introducing our own results.

With the notation introduced in Definition 8, when projecting onto the subspace $\hat{V}_k$ spanned by the first k eigenvectors of $\hat{\mathcal{C}} = \hat{\mathcal{K}}$, the tail sum $\lambda^{>k} = \sum_{i>k} \lambda_i$ lower-bounds the expected squared residual.

Theorem 13 (Shawe-Taylor et al., 2005, Theorem 1) *If we perform PCA in the feature space defined by the Aronszjan mapping $(\psi, \mathcal{H})$ of a kernel κ, then with probability of at least $1-\delta$ over random m-samples S we have for all $k \in [m]$ if we project new data x onto the space $\hat{V}_k$, the expected square residual satisfies:*

$$\lambda^{>k} \leq \mathbb{E}_\mu[||P^\perp_{\hat{V}_k}(\psi(x))||^2] \leq \min_{1\leq \ell \leq k} \left[\frac{1}{m}\hat{\lambda}^{>\ell}(S) + \frac{1+\sqrt{\ell}}{\sqrt{m}} \sqrt{\frac{2}{m}\sum_{i=1}^m \kappa(x_i,x_i)^2} \right] + R^2 \sqrt{\frac{18}{m}\log\left(\frac{2m}{\delta}\right)}.$$

This holds under the assumption that $\|\psi(x)\| \leq R$, for any $x \in \mathcal{X}$.

Theorem 14 (Shawe-Taylor et al., 2005, Theorem 2) *If we perform PCA in the feature space defined by the Aronszjan mapping $(\psi, \mathcal{H})$ of a kernel κ, then with probability of at least $1-\delta$ over random m-samples S we have for all $k \in [m]$ if we project new data x onto the space $\hat{V}_k$, the expected square projection satisfies:*

$$\lambda^{\leq k} \geq \mathbb{E}_\mu[||P_{\hat{V}_k}(\psi(x))||^2] \geq \max_{1\leq \ell \leq k} \left[\frac{1}{m}\hat{\lambda}^{\leq \ell}(S) - \frac{1+\sqrt{\ell}}{\sqrt{m}} \sqrt{\frac{2}{m}\sum_{i=1}^m \kappa(x_i,x_i)^2} \right] - R^2 \sqrt{\frac{19}{m}\log\left(\frac{2(m+1)}{\delta}\right)}.$$

This holds under the assumption that $\|\psi(x)\| \leq R$, for any $x \in \mathcal{X}$.

Notice that the purpose of those two theorems is to control, by upper and lower bounds, the theoretical averaged squared norm projections of a new data point x into the empirical subspaces $\hat{V}_k$ and $\hat{V}_k^\perp$: $\mathbb{E}_\mu[||P_{\hat{V}_k}(\psi(x))||^2]$ and $\mathbb{E}_\mu[||P^\perp_{\hat{V}_k}(\psi(x))||^2]$. Let us note that for each theorem, only one side of the inequality is empirical (while the other one consists in an unknown quantity, $\lambda^{\leq k}$ or $\lambda^{>k}$, respectively).

Our contribution is twofold:

- We first propose two empirical PAC bounds improving (in some cases) the results of Shawe-Taylor et al. (2005). These are collected in Section 3.
- Casting this onto the PAC-Bayes framework, we then move on to two more sophisticated empirical bounds which are replacing the theoretical quantities $\lambda^{>k}$ and $\lambda^{\leq k}$ in Theorems 13 and 14. This is the core of the paper (Section 4).

3. A theoretical analysis of kernel PCA: PAC bounds

We present in this section two PAC bounds, which are directly comparable to those of Shawe-Taylor et al. (2005) which were recalled in Theorem 13 and Theorem 14 (see also Section 5 for a brief numerical comparison). These bounds exploit the classical idea of splitting a data set in half, one being used as a training set and the other as a test set.

Theorem 15 *If we perform PCA in the feature space defined by the Aronszjan mapping $(\psi, \mathcal{H})$ of a kernel κ, then with probability at least $1-\delta$ over random $2m$-samples $S = S_1 \cup S_2$ (where $S_1 = \{x_1, ..., x_m\}$, $S_2 = \{x_{m+1}, ..., x_{2m}\}$ are two disjoints m-samples, we have for all $k \in [m]$, if we project new data x onto the space $\hat{V}_k(S_1)$, the expected square projection is bounded by :*

$$\mathbb{E}_\mu \left[||P_{\hat{V}_k(S_1)}(\psi(x))||^2 \right] \geq \frac{1}{m} \sum_{i=1}^{m} ||P_{\hat{V}_k(S_1)}(\psi(x_{m+i}))||^2 - R^2 \sqrt{\frac{2}{m} \log\left(\frac{1}{\delta}\right)}.$$

Where $\hat{V}_k(S_1)$ is the subspace spanned by the k eigenvectors of the covariance matrix $C(S_1)$ corresponding to the k largest eigenvalues of $C(S_1)$. This holds under the assumption that $\|\psi(x)\| \leq R$, for any $x \in \mathcal{X}$.

This result provides an empirical lower bound for the theoretical expected squared projection. In other words, it quantifies how accurate the projection on a new data point onto our empirical subspace is.

Theorem 16 *If we perform PCA in the feature space defined by the Aronszjan mapping $(\psi, \mathcal{H})$ of a kernel κ, then with probability at least $1-\delta$ over random $2m$-samples $S = S_1 \cup S_2$ (where $S_1 = \{x_1, ..., x_m\}$, $S_2 = \{x_{m+1}, ..., x_{2m}\}$ are two disjoints m-samples, we have for all $k \in [m]$, the expected square residual is bounded by:*

$$\mathbb{E}_\mu \left[||P_{\hat{V}_k^\perp(S_1)}(\psi(x))||^2 \right] \leq \frac{1}{m} \sum_{i=1}^{m} ||P_{\hat{V}_k^\perp(S_1)}(\psi(x_{m+i}))||^2 + R^2 \sqrt{\frac{2}{m} \log\left(\frac{1}{\delta}\right)}.$$

Where $\hat{V}_k^\perp(S_1)$ is the orthogonal complement of $\hat{V}_k(S_1)$, the subspace spanned by the k eigenvectors of the covariance matrix $C(S_1)$ corresponding to the k largest eigenvalues of $C(S_1)$. This holds under the assumption that $\|\psi(x)\| \leq R$, for any $x \in \mathcal{X}$.

Theorem 16 provides an upper bound on the residual squared projection. It therefore measures how much information is lost by the projection of a new data point onto our empirical subspace.

The rest of this section is devoted to the proofs of Theorem 15 and Theorem 16. Numerical implementations of both theorems are gathered in Section 5.

We first recall a classical concentration inequality of McDiarmid (1989).

Theorem 17 (Bounded differences, McDiarmid) *Let $X_1, ..., X_n$ be independent random variables taking values in $\mathcal{X}$ and $f : \mathcal{X}^n \to \mathbb{R}$. Assume that for all $1 \leq i \leq n$ and all $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in \mathcal{X}$ we have:*

$$\sup_{x_i, \hat{x}_i} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i.$$

Then for all $\delta > 0$:

$$\mathbb{P}\left(f(X_1, ..., X_n) - \mathbb{E}[f(X_1, ..., X_n)] > \delta\right) \leq \exp\left(-\frac{2\delta^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof [Proof of Theorem 15]

Let $S = S_1 \cup S_2$ a size-$2m$ sample. We recall that our data are iid. We first apply Proposition 11 and Proposition 12 to S_1.

We define $\hat{V}_k(S_1)$ the subspace spanned by the k eigenvectors of the covariance matrix $C(S_1)$ corresponding to the top k eigenvalues of $C(S_1)$. We need now to study $\mathbb{E}_\mu\left[||P_{\hat{V}_k(S_1)}(\psi(x))||^2\right]$.

Note that

$$\mathbb{E}_\mu\left[||P_{\hat{V}_k(S_1)}(\psi(x))||^2\right] = \left(\mathbb{E}_\mu\left[||P_{\hat{V}_k(S_1)}(\psi(x))||^2\right] - \frac{1}{m}\sum_{i=1}^m ||P_{\hat{V}_k(S_1)}(\psi(x_{m+i}))||^2\right) + \frac{1}{m}\sum_{i=1}^m ||P_{\hat{V}_k(S_1)}(\psi(x_{m+i}))||^2.$$

Because our data are iid, following the distribution μ, we know that S_1 and S_2 are independent, hence

$$\mathbb{E}_{S_2}\left[\frac{1}{m}\sum_{i=1}^m ||P_{\hat{V}_k(S_1)}(\psi(x_{m+i})||^2\right] = \mathbb{E}_\mu\left[||P_{\hat{V}_k(S_1)}(\psi(x))||^2\right].$$

We can now apply McDiarmid's inequality: with probability $1-\delta$,

$$\mathbb{E}_\mu\left[||P_{\hat{V}_k(S_1)}(\psi(x))||^2\right] - \frac{1}{m}\sum_{i=1}^m ||P_{\hat{V}_k(S_1)}(\psi(x_{m+i}))||^2 \geq -R^2\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)}.$$

■

Proof [Proof of Theorem 16] The proof is similar to the previous one, just replace $P_{\hat{V}_k(S_1)}$ by $P_{\hat{V}_k^\perp(S_1)}$ and use McDiarmid's inequality. ■

4. A theoretical analysis of kernel PCA: PAC-Bayes bounds

This section contains our main results which harness the power and flexibility of PAC-Bayes. We bring bounds of a new kind for kernel PCA: while our PAC bounds (in Section 3) were assessing that kernel PCA was efficient with a certain amount of confidence, the two next theorems, on the contrary, explicit the limitations we face when projecting onto an empirical subspace, therefore contributing to a better overall theoretical understanding.

Theorem 18 *For a finite data space $\mathcal{X}$, for $\alpha \in \mathbb{R}$, $\delta \in]0, 1]$, if we perform PCA in the feature space defined by the Aronszjan mapping $(\psi, \mathcal{H})$ of a kernel κ, then with probability*

of at least $1-\delta$ *over random* m*-samples* S*, we have for all* $k \in [m]$*, the expected square projection is bounded by:*

$$\mathbb{E}_\mu[||P_{\hat{V}_k}(\psi(x))||^2] \leq \frac{1}{m}\hat{\lambda}^{\leq k} + \frac{\log(1/\delta)}{m^\alpha} + \frac{R^4}{2m^{1-\alpha}}.$$

and the optimal value for α *is* $\alpha_0 = \frac{1}{2} + \frac{1}{2\log(m)}\log\left(\frac{2\log(1/\delta)}{R^4}\right)$*. This holds under the assumption that* $\|\psi(x)\| \leq R$*, for any* $x \in \mathcal{X}$*.*

This theorem shows that the expected squared projection cannot improve on a quantity close to the partial sum of empirical eigenvalues. This demonstrates that measuring the efficiency of our projection through this empirical sum is therefore relevant.

The next theorem is intended in the same spirit, but holds for empirical squared residuals.

Theorem 19 *For a finite data space* $\mathcal{X}$*, for* $\alpha \in \mathbb{R}$*,* $\delta \in]0,1]$*, if we perform PCA in the feature space defined by the Aronszjan mapping* $(\psi, \mathcal{H})$ *of a kernel* κ*, then with probability at least* $1-\delta$ *over random* m*-samples* S*, we have for all* $k \in [m]$*, the expected square residual is bounded by:*

$$\mathbb{E}_\mu[||P_{\hat{V}_k^\perp}(\psi(x))||^2] \geq \frac{1}{m}\hat{\lambda}^{>k} - \frac{\log(1/\delta)}{m^\alpha} - \frac{R^4}{2m^{1-\alpha}}.$$

and the optimal value for α *is* $\alpha_0 = \frac{1}{2} + \frac{1}{2\log(m)}\log\left(\frac{2\log(1/\delta)}{R^4}\right)$*. This holds under the assumption that* $\|\psi(x)\| \leq R$*, for any* $x \in \mathcal{X}$*.*

Remark 20 *The assumption* $|\mathcal{X}| < +\infty$ *may appear to be restrictive at first glance. As a matter of fact, it covers the case of* $\mathcal{X} \subseteq \mathbb{R}^d$ *bounded if one decides to discretise* $\mathcal{X}$ *into a finite grid* $\mathcal{G}$*. With a large number of points on the grid, one can approximate* $\mathcal{X}$ *efficiently and also apply Theorems 18 and 19 and those theorems provide solid bounds independent of the number of points inside* $\mathcal{G}$*.*

Numerical implementations of those two bounds are gathered in Section 5.

We now move to the proofs of Theorem 18 and Theorem 19. We start with additional technical background.

Self-bounding functions. We use the notion of *self-bounding function* (presented for instance in Boucheron et al., 2009, Definition 2) which allows to deal with a certain type of exponential moment. This tool is at the core of the recent work from Haddouche et al. (2020), to obtain PAC-Bayesian generalisation bounds when the loss is unbounded (as typically assumed in most of the PAC-Bayes literature, see the discussion of Haddouche et al., 2020, and references therein).

Definition 21 (Boucheron et al., 2009, Definition 2) *A function* $f : \mathcal{X}^m \to \mathbb{R}$ *is said to be* (a,b)*-self-bounding with* $(a,b) \in (\mathbb{R}^+)^2 \setminus \{(0,0)\}$*, if there exists* $f_i : \mathcal{X}^{m-1} \to \mathbb{R}$ *for every* $i \in [m]$ *such that for all* $i \in [m]$ *and* $x \in \mathcal{X}$*:*

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1$$

and

$$\sum_{i=1}^{m} f(x) - f_i(x^{(i)}) \leq af(x) + b.$$

Where for all $1 \leq i \leq m$*, the removal of the* i*th entry is* $x^{(i)} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_m)$*. We denote by* $\texttt{SB}(a, b)$ *the class of functions that satisfy this definition.*

In Boucheron et al. (2009), the following bound has been presented to deal with the exponential moment of a self-bounding function. Let $c_+ := \max(c, 0)$ denote the positive part of $c \in \mathbb{R}$. We define $c_+^{-1} := +\infty$ when $c_+ = 0$.

Theorem 22 (Boucheron et al., 2009, Theorem 3.1) *Let* $Z = g(X_1, ..., X_m)$ *where* $X_1, ..., X_m$ *are independent (not necessarily identically distributed)* $\mathcal{X}$*-valued random variables. We assume that* $\mathbb{E}[Z] < +\infty$*. If* $g \in \texttt{SB}(a, b)$*, then defining* $c = (3a - 1)/6$*, for any* $s \in [0; c_+^{-1})$ *we have:*

$$\log\left(\mathbb{E}\left[e^{s(Z-\mathbb{E}[Z])}\right]\right) \leq \frac{(a\mathbb{E}[Z] + b)\, s^2}{2(1 - c_+ s)}.$$

PAC-Bayes. We will consider a fixed learning problem with data space $\mathcal{Z}$, set of predictors $\mathcal{F}$, and loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}^+$. We denote by μ a probability distribution over $\mathcal{Z}$, $s = (z_1, \ldots, z_m)$ is a size-m sample. We denote as $\Sigma_{\mathcal{F}}$ the considered σ-algebra on $\mathcal{F}$. Finally, we define for any $f \in \mathcal{F}$, $L(f) = \mathbb{E}_{z\sim\mu}[\ell(f, z)]$ and $\hat{L}_s(f) = \frac{1}{m}\sum_{i=1}^{m} \ell(f, z_i)$.

Let us highlight that in supervised learning, for instance, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subset \mathbb{R}^d$ is a space of inputs, and $\mathcal{Y}$ a space of labels. In this case predictors are functions $f : \mathcal{X} \to \mathcal{Y}$. One may be interested in applying PCA over the input space $\mathcal{X}$ to reduce the input dimension.

We use the PAC-Bayesian bound given by Rivasplata et al. (2020), in which the authors developed a PAC-Bayesian bound allowing the use of a data-dependant prior (we refer to the discussion and references therein for an introduction to data-dependent prior distributions in PAC-Bayes).

Definition 23 (Stochastic kernels) *A* stochastic kernel *from* $\mathcal{Z}^m$ *to* $\mathcal{F}$ *is defined as a mapping* $Q : \mathcal{Z}^m \times \Sigma_{\mathcal{F}} \to [0; 1]$ *where*

- *For any* $B \in \Sigma_{\mathcal{F}}$*, the function* $s = (z_1, ..., z_m) \mapsto Q(s, B)$ *is measurable,*
- *For any* $s \in \mathcal{Z}^m$*, the function* $B \mapsto Q(s, B)$ *is a probability measure over* $\mathcal{F}$*.*

We will denote by $\texttt{Stoch}(\mathcal{Z}^m, \mathcal{F})$ *the set of all stochastic kernels from* $\mathcal{Z}^m$ *to* $\mathcal{F}$*.*

Definition 24 *For any* $Q \in \texttt{Stoch}(\mathcal{Z}^m, \mathcal{F})$ *and* $s \in \mathcal{Z}^m$*, we define* $Q_s[L] := \mathbb{E}_{f\sim Q_s}[L(f)]$ *and* $Q_s[\hat{L}_s] := \mathbb{E}_{f\sim Q_s}[\hat{L}_s(f)]$*. We generalise this definition to the case where we consider* $S = (Z_1, ..., Z_m)$ *a random sample. Then we consider the random quantities* $Q_S[L]$ *and* $Q_S[\hat{L}_S]$*.*

For the sake of clarity, we now present a slightly less general version of one of the main theorems of Rivasplata et al. (2020). We define $\mu^m := \mu \otimes \cdots \otimes \mu$ (m times).

Theorem 25 (Rivasplata et al., 2020) *For any $F : \mathbb{R}^2 \to \mathbb{R}$ convex, for any $Q^0 \in \texttt{Stoch}(\mathcal{Z}^m, \mathcal{F})$, we define*

$$\xi := \mathbb{E}_{S\sim\mu^m}\mathbb{E}_{f\sim Q_S^0}\left[\exp\left(F(\hat{L}_S(f), L(f))\right)\right].$$

Then for any $Q \in \texttt{Stoch}(\mathcal{Z}^m, \mathcal{F})$, any $\delta \in]0;1]$, with probability at least $1-\delta$ over the random draw of the sample $S \sim \mu^m$

$$F(Q_S[(\hat{L}_S, L]) \leq \mathrm{KL}(Q_S, Q_S^0) + \log(\xi/\delta).$$

For a fixed PD kernel κ, $(\psi, \mathcal{H})$ is the associated Aronszajn mapping. Let us now consider a finite data space $\mathcal{X}$. We therefore have

$$N_{\mathcal{H}} := \dim(\mathcal{H}) < +\infty.$$

For the sake of clarity, we will assume that $\mathcal{H} = \mathbb{R}^{N_{\mathcal{H}}}$ endowed with the Euclidean inner product. The space is provided with the Borelian σ-algebra $\mathcal{B}(\mathbb{R}^{N_{\mathcal{H}}})$.

We assume that our data distribution $\psi(\mu)$ over $\mathcal{H}$ is bounded in $\mathcal{H}$ by a scalar R:

$$\forall x \in \mathcal{X}, \quad ||\psi(x)|| \leq R.$$

Remark 26 *Note that this assumption is not especially restrictive. Indeed, if κ is bounded by C, then for all $x \in \mathcal{X}, ||\psi(x)||^2 = \kappa(x,x) \leq C$.*

First, note that the function $||P_{V_k}(\psi(x))||^2$ is expressed with a quadratic dependency over the coordinates of $\psi(x)$. However, it would be far more convenient better to consider linear functions.

Proposition 27 (Shawe-Taylor et al., 2005, Prop. 11) *Let $\hat{V}$ be the subspace spanned by some subset I of the set of eigenvectors of the correlation matrix $C(S)$ associated to the kernel matrix $K(S)$. There exists an Hilbert space mapping $\hat{\psi} : \mathcal{X} \to \mathbb{R}^{N_{\mathcal{H}}^2}$ such that the projection norm $||P_{\hat{V}}(\psi(x))||^2$ is a linear function $\hat{f}$ in $\mathbb{R}^{N_{\mathcal{H}}^2}$ such that, for all $(x,z) \in \mathcal{X}^2$*

$$\langle\hat{\psi}(x), \hat{\psi}(z)\rangle = \kappa(x,z)^2.$$

Furthermore, if $|I| = k$, then the 2-norm of the function $\hat{f}$ is $\sqrt{k}$.

Proof Let $X = U\Sigma V'$ be the SVD of the sample matrix X (where each column represents a data point). The projection norm is then given by:

$$||P_{\hat{V}}(\psi(x)||^2 = \psi(x)'U(I)U(I)'\psi(x)$$

where $U(I) \in \mathbb{R}^{N_{\mathcal{H}}\times k}$ containing the k columns of U in the set I (or equivalently $U(I) \in \mathbb{R}^{N_{\mathcal{H}}\times N_{\mathcal{H}}}$ filled with zeros). If we define $w := U(I)U(I)' \in \mathbb{R}^{N_{\mathcal{H}}^2}$ then we have:

$$||P_{\hat{V}}(\psi(x)||^2 = \sum_{i,j=1}^{N_{\mathcal{H}}} w_{i,j}\psi(x)_i\psi(x)_j = \sum_{i,j=1}^{N_{\mathcal{H}}} w_{i,j}\hat{\psi}(x)_{i,j} = \langle w, \hat{\psi}(x)\rangle,$$

where for all i, $\psi(x)_i$ represents the i-th coordinate of $\psi(x)$ and for all $x \in \mathcal{X}, \hat{\psi}(x) = (\psi(x)_i \psi(x)_j)_{i,j=1..N_{\mathcal{H}}}$.

Thus, the projection norm is a linear function $\hat{f}$ in $\mathbb{R}^{N_{\mathcal{H}}^2}$ with the mapping $\hat{\psi}$. Note that considering the 2-norm of $\hat{f}$ is equivalent to consider the 2-norm of w. Then, if we denote by $(u_i)_i$ the columns of U, we have:

$$\begin{aligned}
||\hat{f}||^2 &= \sum_{i,j=1}^{N_{\mathcal{H}}} w_{i,j}^2 = ||U(I)U'(I)||^2 \\
&= \left\langle \sum_{i \in I} u_i u_i', \sum_{j \in I} u_j u_j' \right\rangle \\
&= \sum_{i,j \in I} (u_i' u_j)^2 \\
&= k.
\end{aligned}$$

Hence the 2-norm of $\hat{f}$ is $\sqrt{k}$. Finally, for $(x, z) \in \mathcal{X}^2$:

$$\begin{aligned}
\kappa(x,z)^2 = \langle \psi(x), \psi(z) \rangle^2 &= \left(\sum_{i=1}^{N_{\mathcal{H}}} \psi(x)_i \psi(z)_i \right)^2 \\
&= \sum_{i,j=1}^{N_{\mathcal{H}}} \psi(x)_i \psi(z)_i \psi(x)_j \psi(z)_j \\
&= \sum_{i,j=1}^{N_{\mathcal{H}}} (\psi(x)_i \psi(x)_j)(\psi(z)_i \psi(z)_j) \\
&= \langle \hat{\psi}(x), \hat{\psi}(z) \rangle.
\end{aligned}$$

■

Now, recall that for a fixed k, P_{V_k} minimises the shortfall between the squared norm projection of $\psi(x)$ and $||\psi(x)||^2$ for any x over the space of projection functions over a subspace of dimension at most k. We therefore introduce the following learning framework.

The data space is $\mathcal{X}$ with the probability distribution μ. The space $\mathcal{X}$ is endowed with a σ-algebra $\Sigma_{\mathcal{X}}$.

The goal is to minimise the loss over the set of linear predictors in $\mathbb{R}^{N_{\mathcal{H}}^2}$ corresponding to projections into a k-dimensional subspace of $\mathbb{R}^{N_{\mathcal{H}}}$

$$\mathcal{F}_k := \left\{ f_w : x \mapsto \langle w, \hat{\psi}(x) \rangle \mid \exists V \subseteq \mathbb{R}^{N_{\mathcal{H}}}, \dim(V) = k,\ \forall x \in \mathcal{X}, f_w(x) = ||P_V(\psi(x)||^2 \right\}.$$

Because k may variate between 1 and $N_{\mathcal{H}}$, we also define the space $\mathcal{F}$ such that for all k, $\mathcal{F}_k \subseteq \mathcal{F}$:

$$\mathcal{F} := \left\{ f_w : x \mapsto \langle w, \hat{\psi}(x) \rangle \mid \exists V \subseteq \mathbb{R}^{N_{\mathcal{H}}}, \forall x \in \mathcal{X}, f_w(x) = ||P_V(\psi(x)||^2 \right\}.$$

When needed, we will assimilate f_w to w.

The loss function is $\ell : \mathcal{F} \times \mathcal{X} \to [0, R^2]$ such that for $f_w \in \mathcal{F}$ and $x \in \mathcal{X}$:

$$\ell(f, x) = ||\psi(x)||^2 - f_w(x).$$

This loss is indeed non-negative because all the vectors of $\mathcal{F}$ are representing the squared norm of projectors in $\mathbb{R}^{N_{\mathcal{H}}}$ so we have for all $f_w \in \mathcal{F}$, $x \in \mathcal{X}$, $f_w(x) \leq ||\psi(x)||^2$.

Remark 28 *In other contexts and tasks other loss functions are used. The reason that the loss function just defined above makes sense in our task is given by Proposition 11. Indeed, we know that if f is a projector over a space V of dimension k, then we have $\ell(f, x) = ||P_V^{\perp}(\psi(x)||^2$ and moreover the expected squared residual $\lambda^{>k}$ satisfies:*

$$\lambda^{>k} = \min_{\dim(V)=k} \mathbb{E}_{\mu}[||P_V^{\perp}(\psi(x))||^2] = \mathbb{E}_{\mu}[||P_{V_k}^{\perp}(\psi(x))||^2].$$

We assume $|\mathcal{X}| < +\infty$. To apply our PAC-Bayesian theorem (Theorem 25) we need to introduce an adapted stochastic kernel. For a fixed k, we set Q^k, P^k as follows, $\forall s \in \mathcal{X}^m$, $\forall B \in \Sigma_{\mathcal{F}_k}$:

$$Q^k(s, B) = \mathbb{1}\left\{f_{\hat{w}_k(s)} \in B\right\},$$

$$P^k(s, B) = \mathbb{1}\left\{f_{\hat{w}_k^{\perp}(s)} \in B\right\},$$

where $\Sigma_{\mathcal{F}_k}$ is the σ-algebra on $\mathcal{F}_k$, and:

- For all k, the vector $\hat{w}_k(s)$ of $\mathbb{R}^{N_{\mathcal{H}^2}}$ is such that $f_{\hat{w}_k}(x) = ||P_{\hat{V}_k(s)}(.)||^2$ where $\hat{V}_k(s)$ is the $k-$dimensional subspace defined in Definition 10 obtained from the sample s;
- For all k, the vector $\hat{w}_k^{\perp}(s)$ of $\mathbb{R}^{N_{\mathcal{H}^2}}$ is such that $f_{\hat{w}_k^{\perp}}(x) = ||P_{\hat{V}_k^{\perp}(s)}(.)||^2$ where $\hat{V}_k{}^{\perp}(s)$ is the orthogonal of $\hat{V}_k(s)$.

We need the following technical results.

Theorem 29 *For all k, Q^k is a stochastic kernel.*

Proof The proof is postponed to Section 7. ■

Theorem 30 *For all k, P^k is a stochastic kernel.*

Proof The proof is postponed to Section 7. ■

Remark 31 *The proof of Theorem 29 exploits the hypothesis $|\mathcal{X}| < +\infty$. Indeed, the fact that $\mathcal{H}$ is finite-dimensional allows to consider well-defined symmetric matrices and to exploit a result from Wilcox (1972).*

Now we prove Theorem 18, which we recall here for convenience: for a finite data space $\mathcal{X}$, for $\alpha \in \mathbb{R}$, $\delta \in]0,1]$, for any $1 \leq k \leq m$, we have with probability $1-\delta$ over the random m-sample S

$$\mathbb{E}_\mu[||P_{\hat{V}_k}(\psi(x))||^2] \leq \frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i + \frac{\log(1/\delta)}{m^\alpha} + \frac{R^4}{2m^{1-\alpha}}.$$

and the optimal value for α is $\alpha_0 = \frac{1}{2} + \frac{1}{2\log(m)}\log\left(\frac{2\log(1/\delta)}{R^4}\right)$.

Proof [Proof of Theorem 18]

Let $k \in [m]$. We first apply Theorem 25 with probability $1-\delta$, $F : (x,y) \mapsto m^\alpha(y-x)$ and the stochastic kernel P^k (thanks to Theorem 30) as prior and posterior which nullify the KL-divergence term. We then have, with probability $1-\delta$,

$$m^\alpha(P_S^k(L(f)) - P_S^k(\hat{L}_S(f)) \leq \mathrm{KL}(P_S^k, P_S^k) + \log(\xi/\delta)$$

where $\xi := \mathbb{E}_{S\sim\mu^m}\mathbb{E}_{f\sim P_S^k}\left[\exp\left(m^\alpha(\hat{L}_S(f) - (L(f))\right)\right]$.

Notice that for any sample S, P_S^k is the Dirac measure in $f_{\hat{w}_k^\perp(S)}$ *i.e.* the function representing the squared norm projection $||P_{\hat{V}_k^\perp(S)}(.)||^2$. Hence $P_S^k(L(f)) = L(f_{\hat{w}_k^\perp(S)})$, $Q_S^k(\hat{L}_S(f)) = \hat{L}_S(f_{\hat{w}_k^\perp(S)})$.

Finally we have

$$m^\alpha\left(L(f_{\hat{w}_k^\perp(S)}) - \hat{L}_S(f_{\hat{w}_k^\perp(S)})\right) \leq \log(\xi/\delta)$$

hence:

$$m^\alpha\left(\mathbb{E}_\mu\left[||\psi(x)||^2 - ||P_{\hat{V}_k^\perp(S)}(\psi(x))||^2\right] - \mathbb{E}_{\hat{\mu}}\left[||\psi(x)||^2 - ||P_{\hat{V}_k^\perp(S)}(\psi(x))||^2\right]\right) \leq \log(\xi/\delta)$$

where $\hat{\mu}$ is the empirical distribution over the m-sample S.

Furthermore, because we are considering orthogonal projections, we can see that for any $x \in \mathcal{X}$, we have

$$||\psi(x)||^2 - ||P_{\hat{V}_k^\perp(S)}(\psi(x))||^2 = ||P_{\hat{V}_k(S)}(\psi(x))||^2.$$

So

$$\mathbb{E}_\mu\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] \leq \mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] + \frac{\log(\xi/\delta)}{m^\alpha}.$$

Thanks to Proposition 12, we know that $\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k^\perp(S)}(\psi(x))||^2\right] = \frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i$, which yields

$$\mathbb{E}_\mu\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] \leq \frac{1}{m}\sum_{i=1}^{k}\hat{\lambda}_i + \frac{\log(\xi/\delta)}{m^\alpha}.$$

Finally, we need to control $\log(\xi)$. To do so, we use Theorem 22. We first recall that

$$\begin{aligned}
\xi &:= \mathbb{E}_{S\sim\mu^m}\mathbb{E}_{f\sim P_S^k}\left[\exp\left(m^\alpha(L(f) - \hat{L}_S(f))\right)\right] \\
&= \mathbb{E}_{S\sim\mu^m}\left[\exp\left(m^\alpha(L(f_{\hat{w}_k^\perp}) - \hat{L}_S(f_{\hat{w}_k^\perp}))\right)\right] \\
&= \mathbb{E}_{S\sim\mu^m}\left[\exp\left(m^\alpha(\mathbb{E}_\mu\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] - \mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right])\right)\right].
\end{aligned}$$

Yet, thanks to Propositions 11 and 12, we know that:

$$\mathbb{E}_\mu\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] \leq \max_{\dim(V)=k} \mathbb{E}_\mu[||P_V(\psi(x))||^2] = \mathbb{E}_\mu\left[||P_{V_k(S)}(\psi(x))||^2\right]$$

and

$$\begin{aligned}\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right]] &= \max_{\dim(V)=k} \mathbb{E}_{\hat{\mu}}[||P_V(\psi(x))||^2] \\ &\geq \mathbb{E}_{\hat{\mu}}\left[||P_{V_k}(\psi(x))||^2\right].\end{aligned}$$

Thus we have

$$\mathbb{E}_\mu\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] - \mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] \leq \mathbb{E}_\mu\left[||P_{V_k}(\psi(x))||^2\right] - \mathbb{E}_{\hat{\mu}}\left[||P_{V_k}(\psi(x))||^2\right]$$

Remark 32 *The interest of this maneuver is to replace the projector $P_{\hat{V}_k(S)}$, which is data-dependent, by P_{V_k}, which is not. Doing so, if we set $Y = \frac{1}{m}\sum_{i=1}^m Y_i$ and for all i, $Y_i = ||P_{V_k}(\psi(x_i))||^2$ (where the Y_i are iid), we can write properly:*

$$\mathbb{E}_\mu\left[||P_{V_k}(\psi(x))||^2\right] - \mathbb{E}_{\hat{\mu}}\left[||P_{V_k}(\psi(x))||^2\right] = \mathbb{E}_{\mathcal{S}}[Y] - Y$$

while $\mathbb{E}_\mu\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] \neq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right]\right]$.

So, finally we have

$$\xi \leq \mathbb{E}_{S\sim\mu^m}\left[\exp\left(m^\alpha(\mathbb{E}_{\mathcal{S}}[Y] - Y)\right)\right].$$

We define the function $f : \mathcal{X}^m \to \mathbb{R}$ as

$$f : z \mapsto \frac{1}{R^2}\sum_{i=1}^m \left(R^2 - ||P_{V_k}(\psi(z_i))||^2\right) \qquad \text{for } z = (z_1, ..., z_m) \in \mathcal{X}^m.$$

We define $Z = f(X_1, ..., X_m)$ and notice that $\mathbb{E}_{\mathcal{S}}[Y] - Y = \frac{R^2}{m}\left(Z - \mathbb{E}_{\mathcal{S}}[Z]\right)$. We first prove that $f \in \texttt{SB}(\beta, 1-\beta)$ (cf. Def 21) for any $\beta \in [0,1]$. Indeed, for all $1 \leq i \leq m$, we define

$$f_i(z^{(i)}) = \frac{1}{R^2}\sum_{j\neq i}\left(R^2 - ||P_{V_k}(\psi(z_i))||^2\right)$$

where $z^{(i)} = (z_1, ..., z_{i-1}, z_{i+1}, ..., z_m) \in \mathcal{X}^{m-1}$ for any $z \in \mathcal{X}^m$ and for any i. Then, since $0 \leq ||P_{V_k}(\psi(z_i))||^2 \leq R^2$ for all i, we have

$$0 \leq f(z) - f_i(z^{(i)}) = \frac{R^2 - ||P_{V_k}(\psi(z_i))||^2}{R^2} \leq 1.$$

Moreover, because $f(z) \leq m$ for any $z \in \mathcal{X}^m$, we have

$$\begin{aligned}\sum_{i=1}^m f(z) - f_i(z^{(i)}) &= \sum_{i=1}^m \frac{R^2 - ||P_{V_k}(\psi(z_i))||^2}{R^2} \\ &= f(z) = \beta f(z) + (1-\beta)f(z) \leq \beta f(z) + (1-\beta)m.\end{aligned}$$

Since this holds for any $x \in \mathcal{X}^m$, this proves that f is $(\beta, 1-\beta)$-self-bounding.

Now, to complete the proof, we will use Theorem 22. Because Z is $(1/3, (2/3)m)$-self-bounding, we have for all $s \in \mathbb{R}^+$

$$\log\left(\mathbb{E}_{\mathcal{S}}\left[e^{s(Z-\mathbb{E}_{\mathcal{S}}[Z])}\right]\right) \leq \frac{\left(\frac{1}{3}\mathbb{E}_{\mathcal{S}}[Z] + \frac{2m}{3}\right)s^2}{2}.$$

And since $Z \leq m$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[e^{m^\alpha(\mathbb{E}_{\mathcal{S}}[Y]-Y)}\right] &= \mathbb{E}_{\mathcal{S}}\left[e^{\frac{R^2}{m^{1-\alpha}}(Z-\mathbb{E}_{\mathcal{S}}[Z])}\right] \\
&\leq \exp\left(\frac{\left(\frac{1}{3}\mathbb{E}_{\mathcal{S}}[Z] + \frac{2m}{3}\right)R^4}{2m^{2-2\alpha}}\right) && (\textit{Theorem 22}) \\
&\leq \exp\left(\frac{R^4}{2m^{1-2\alpha}}\right). && (\text{since } \mathbb{E}_{\mathcal{S}}[Z] \leq m).
\end{aligned}$$

So, finally, we have

$$\frac{\log(\xi)}{2m^\alpha} \leq \frac{R^4}{m^{1-\alpha}},$$

hence the final result.

To obtain the optimal value of α, we simply study the derivative of the univariate function $f_{R,\delta}(\alpha) := \frac{\log(1/\delta)}{m^\alpha} + \frac{R^4}{m^{1-\alpha}}$. ■

We now prove Theorem 19 which deals with the expected square residuals. We recall the theorem: for a finite data space $\mathcal{X}$, for $\alpha \in \mathbb{R}$, $\delta \in]0,1]$, for any $1 \leq k \leq m$, we have with probability $1-\delta$ over the random m-sample S:

$$\mathbb{E}_\mu[||P_{\hat{V}_k^\perp}(\psi(x))||^2] \geq \frac{1}{m}\sum_{i=k+1}^{m}\hat{\lambda}_i - \frac{\log(1/\delta)}{m^\alpha} - \frac{R^4}{2m^{1-\alpha}}$$

and the optimal value for α is $\alpha_0 = \frac{1}{2} + \frac{1}{2\log(m)}\log\left(\frac{2\log(1/\delta)}{R^4}\right)$.

Proof [Proof of Theorem 19] The proof is similar to the one of Theorem 18 but it rather involves the stochastic kernel Q^k. Let $k \in [m]$. We first apply Theorem 25 with probability $1-\delta$, $F : (x,y) \mapsto m^\alpha(x-y)$ and the stochastic kernel Q^k (cf. Theorem 29) as prior and posterior which nullify the KL-divergence term. We then have with probability $1-\delta$:

$$m^\alpha(Q_S^k(\hat{L}_S(f)) - Q_S^k(L(f))) \leq \mathrm{KL}(Q_S^k, Q_S^k) + \log(\xi/\delta)$$

where $\xi := \mathbb{E}_{S\sim\mu^m}\mathbb{E}_{f\sim Q_S^k}\left[\exp\left(m^\alpha(\hat{L}_S(f) - (L(f))\right)\right]$.

We notice that for any sample S, Q_S^k is the Dirac measure in $f_{\hat{w}_k(S)}$ i.e. the function representing the squared norm projection $||P_{\hat{V}_k(S)}(.)||^2$. Hence $Q_S^k(L(f)) = L(f_{\hat{w}_k(S)})$, $Q_S^k(\hat{L}_S(f)) = \hat{L}_S(f_{\hat{w}_k(S)})$. Finally we have:

$$m^\alpha\left(\hat{L}_S(f_{\hat{w}_k(S)}) - L(f_{\hat{w}_k(S)})\right)) \leq \log(\xi/\delta)$$

hence

$$m^{\alpha}\left(\mathbb{E}_{\hat{\mu}}\left[||\psi(x)||^2 - ||P_{\hat{V}_k(S)}(\psi(x))||^2\right] - \mathbb{E}_{\mu}\left[||\psi(x)||^2 - ||P_{\hat{V}_k(S)}(\psi(x))||^2\right]\right) \leq \log(\xi/\delta),$$

where $\hat{\mu}$ is the empirical distribution over the m-sample S.

Furthermore, because we are considering orthogonal projections, we remark that for any $x \in \mathcal{X}$, we have:

$$||\psi(x)||^2 - ||P_{\hat{V}_k(S)}(\psi(x))||^2 = ||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2.$$

So we have by multiplying by -1:

$$\mathbb{E}_{\mu}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right] \geq \mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right] - \frac{\log(\xi/\delta)}{m^{\alpha}}.$$

Thanks to Proposition 12, we know that $\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k(S)}(\psi(x))||^2\right] = \frac{1}{m}\sum_{i=k+1}^{m}\hat{\lambda}_i$ and then we have:

$$\mathbb{E}_{x\sim\mu}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right] \geq \frac{1}{m}\sum_{i=k+1}^{m}\hat{\lambda}_i - \frac{\log(\xi/\delta)}{m^{\alpha}}.$$

Finally we need to control $\log(\xi)$. To do so, we use Theorem 22. We first recall that:

$$\begin{aligned}
\xi &:= \mathbb{E}_{S\sim\mu^m}\mathbb{E}_{f\sim Q_S^k}\left[\exp\left(m^{\alpha}(\hat{L}_S(f) - L(f))\right)\right] \\
&= \mathbb{E}_{S\sim\mu^m}\left[\exp\left(m^{\alpha}(\hat{L}_S(f_{\hat{w}_k}) - L(f_{\hat{w}_k}))\right)\right] \\
&= \mathbb{E}_{S\sim\mu^m}\left[\exp\left(m^{\alpha}\left(\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right] - \mathbb{E}_{\mu}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right]\right)\right)\right].
\end{aligned}$$

Yet, thanks to Propositions 11 and 12, we know that:

$$\mathbb{E}_{\mu}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right] \geq \min_{\dim(V)=k}\mathbb{E}_{\mu}[||P_{V^{\perp}}(\psi(x))||^2] = \mathbb{E}_{\mu}\left[||P_{V_k^{\perp}(S)}(\psi(x))||^2\right]$$

and

$$\begin{aligned}
\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right]] &= \min_{\dim(V)=k}\mathbb{E}_{\hat{\mu}}[||P_{V^{\perp}}(\psi(x))||^2] \\
&\leq \mathbb{E}_{\hat{\mu}}\left[||P_{V_k^{\perp}}(\psi(x))||^2\right].
\end{aligned}$$

Thus we have

$$\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V^{\perp}}_k(S)}(\psi(x))||^2\right] - \mathbb{E}_{\mu}\left[||P_{\hat{V^{\perp}}_k(S)}(\psi(x))||^2\right] \leq \mathbb{E}_{\hat{\mu}}\left[||P_{V_k^{\perp}}(\psi(x))||^2\right] - \mathbb{E}_{\mu}\left[||P_{V_k^{\perp}}(\psi(x))||^2\right].$$

Remark 33 *The interest of this maneuver is to replace the projector $P_{\hat{V}_k(S)}$, which is data-dependent, by P_{V_k}, which is not. Doing so, if we set $Y = \frac{1}{m}\sum_{i=1}^{m}Y_i$ and for all i, $Y_i = ||P_{V_k^{\perp}}(\psi(x_i))||^2$ (where the Y_i are iid), we can write properly:*

$$\mathbb{E}_{\hat{\mu}}\left[||P_{V_k^{\perp}}(\psi(x))||^2\right] - \mathbb{E}_{\mu}\left[||P_{V_k^{\perp}}(\psi(x))||^2\right] = Y - \mathbb{E}_{\mathcal{S}}[Y]$$

while $\mathbb{E}_{\mu}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right] \neq \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\hat{\mu}}\left[||P_{\hat{V}_k^{\perp}(S)}(\psi(x))||^2\right]\right]$.

So, finally, we have:

$$\xi \leq \mathbb{E}_{S\sim\mu^m}\left[\exp\left(m^\alpha(Y - \mathbb{E}_{\mathcal{S}}[Y])\right)\right].$$

We define the function $f : \mathcal{X}^m \to \mathbb{R}$ as

$$f : z \mapsto \frac{1}{R^2}\sum_{i=1}^{m} ||P_{V_k^\perp}(\psi(z_i))||^2 \qquad \text{for } z = (z_1, ..., z_m) \in \mathcal{X}^m.$$

We define $Z = f(X_1, ..., X_m)$ and notice that $\mathbb{E}_{Y-\mathcal{S}}[Y] = \frac{R^2}{m}\left(Z - \mathbb{E}_{\mathcal{S}}[Z]\right)$. We first prove that $f \in \mathtt{SB}(\beta, 1-\beta)$ (cf. Def 21) for any $\beta \in [0, 1]$.

Indeed, for all $1 \leq i \leq m$, we define:

$$f_i(z^{(i)}) = \frac{1}{R^2}\sum_{j\neq i} ||P_{V_k}(\psi(z_i))||^2$$

where $z^{(i)} = (z_1, ..., z_{i-1}, z_{i+1}, ..., z_m) \in \mathcal{X}^{m-1}$ for any $z \in \mathcal{X}^m$ and for any i. Then, since $0 \leq ||P_{V_k^\perp}(\psi(z_i))||^2 \leq R^2$ for all i, we have

$$0 \leq f(z) - f_i(z^{(i)}) = \frac{||P_{V_k^\perp}(\psi(z_i))||^2}{R^2} \leq 1.$$

Moreover, because $f(z) \leq m$ for any $z \in \mathcal{X}^m$, we have:

$$\begin{aligned}\sum_{i=1}^{m} f(z) - f_i(z^{(i)}) &= \sum_{i=1}^{m} \frac{||P_{V_k^\perp}(\psi(z_i))||^2}{R^2} \\ &= f(z) = \beta f(z) + (1-\beta)f(z) \leq \beta f(z) + (1-\beta)m.\end{aligned}$$

Since this holds for any $x \in \mathcal{X}^m$, this proves that f is $(\beta, 1-\beta)$-self-bounding.

Now, to complete the proof, we use Theorem 22. Because Z is $(1/3, (2/3)m)$-self-bounding, we have for all $s \in \mathbb{R}^+$:

$$\log\left(\mathbb{E}_{\mathcal{S}}\left[e^{s(Z-\mathbb{E}_{\mathcal{S}}[Z])}\right]\right) \leq \frac{\left(\frac{1}{3}\mathbb{E}_{\mathcal{S}}[Z] + \frac{2m}{3}\right)s^2}{2}.$$

And since $Z \leq m$:

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left[e^{m^\alpha(\mathbb{E}_{\mathcal{S}}[Y]-Y)}\right] &= \mathbb{E}_{\mathcal{S}}\left[e^{\frac{R^2}{m^{1-\alpha}}(Z-\mathbb{E}_{\mathcal{S}}[Z])}\right] \\ &\leq \exp\left(\frac{\left(\frac{1}{3}\mathbb{E}_{\mathcal{S}}[Z] + \frac{2m}{3}\right)R^4}{2m^{2-2\alpha}}\right) && (\textit{Theorem 22}) \\ &\leq \exp\left(\frac{R^4}{2m^{1-2\alpha}}\right). && (\text{since } \mathbb{E}_{\mathcal{S}}[Z] \leq m).\end{aligned}$$

So, finally, we have

$$\frac{\log(\xi)}{m^\alpha} \leq \frac{R^4}{2m^{1-\alpha}}.$$

Hence the final result.

To obtain the optimal value of α, we simply study the derivative of the univariate function $f_{R,\delta}(\alpha) := \frac{\log(1/\delta)}{m^\alpha} + \frac{R^4}{m^{1-\alpha}}$. ■

5. A brief numerical illustration

In this section we briefly illustrate the numerical behaviour of our lower and upper bounds, with respect to Shawe-Taylor et al. (2005). We conduct two experiments below.

Experiment 1 We exploit the dataset used in Shawe-Taylor et al. (2005) to compare our bound in Theorem 15, which we recall here:

$$\frac{1}{m}\sum_{i=1}^{m}||P_{\hat{V}_k(S_1)}(\psi(x_{m+i}))||^2 - R^2\sqrt{\frac{2}{m}\log\left(\frac{1}{\delta}\right)},$$

with the one from Shawe-Taylor et al. (2005), given by

$$\max_{1\leq\ell\leq k}\left[\frac{1}{m}\hat{\lambda}^{\leq\ell}(S) - \frac{1+\sqrt{\ell}}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^{m}\kappa(x_i,x_i)^2}\right] - R^2\sqrt{\frac{19}{m}\log\left(\frac{2(m+1)}{\delta}\right)}.$$

We choose $\delta = 0.05$, and the dataset size is $N = 696$. We take $m = \frac{N}{4} = 174$, and $k \in \{1,\dots,100\}$.

We generate a $2m$-sized dataset $S = S_1 \cup S_2$, we apply our kernel PCA method over S_1 (to obtain the projection space $\hat{V}_{\mathcal{S}_\infty}$), then we compute our bound by using $S_2 = \{x_{m+1},\cdots,x_{2m}\}$. This is the blue curve in Figure 1.

We then compute the bound from Shawe-Taylor et al. (2005) by using the eigenvalues of $K(S)$ (the orange curve in Figure 1). Finally, we draw $\lambda^{\leq k}$ (green curve).

Clearly, on this specific instance of the kernel PCA problem, our Theorem 15 leads to a much tighter bound than the one of Shawe-Taylor et al. (2005). Let us stress here that this is merely a safety check on a specific example.

Experiment 2 We are using the same experimental framework as in Experiment 1. We now compute four curves: the theoretical eigenvalues, the bound from Theorem 18 with the 'naive' choice of $\alpha = 1/2$ and also with the optimised α_0. We also compute the bound from Theorem 15. Results are shown in Figure 2. Clearly, the choice of α significantly influences the tightness of the bound.

6. Conclusion

We provided empirical bounds for two quantities: the expected squared norm and the expected residual of a new data point projected onto the empirical (small) subspaces given by the kernel PCA method. This outperforms (as illustrated on an example) the existing bounds given by Shawe-Taylor et al. (2005). Another improvement on the seminal work of Shawe-Taylor et al. (2005) is that we provide both lower and upper empirical bounds for each studied quantity. Doing so, we contribute a better theoretical understanding of kernel PCA, which we hope will translate into practical insights on strengths and limitations of kernel PCA for machine learning theoreticians and practitioners.

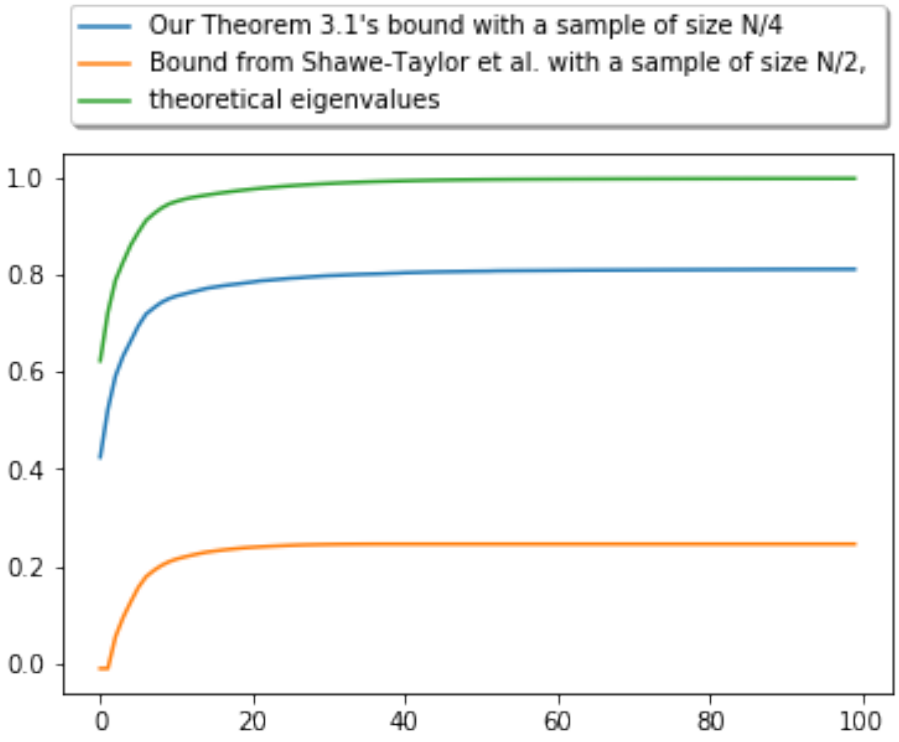

Figure 1: Evaluation of Theorem 15. The x-axis is the number k of considered eigenvalues to compute the bound, the y-axis is the amount of information contained in the projection, from 0 to 1.

7. Proofs – technical results

7.1 Proof of Theorem 29

First, for all k and $s \in \mathcal{X}^m$, the function $B \mapsto Q_s^k(B)$ is the Dirac in $||P_{\hat{V}_k(s)}(.)||^2 \in \mathcal{F}_k$ hence it is a well-defined probability law. Now, we fix $k \in \{1..N_{\mathcal{H}}\}$ and $B \in \Sigma_{\mathcal{F}_k}$. We need to prove that the function $A : s \mapsto Q^k(s, B)$ is measurable. We first decompose A into several functions:

$$s \overset{\psi}{\longmapsto} (\psi(x_1) \cdots \psi(x_m)) \overset{A_1}{\longmapsto} C(s) \overset{A_2}{\longmapsto} \text{eigenvectors of } C(s)$$
$$\overset{A_3}{\longmapsto} f_{\hat{w}_k} \overset{A_4}{\longmapsto} \mathbb{1}\{f_{\hat{w}_k} \in B\}$$

For all the intermediate spaces (which are all finite-dimensional vector spaces, or subsets of them), we will consider them with their Lebesgue $\sigma-$algebra (or the one induced on the corresponding subset), which will allow us to consider the usual notion of continuity onto those spaces. Then for every s, we have $A(s) = A_4 \circ A_3 \circ A_2 \circ A_1 \circ \psi(s)$.
Now our goal is to prove that all those functions are measurable, doing so, A will be measurable as composition of measurable functions.

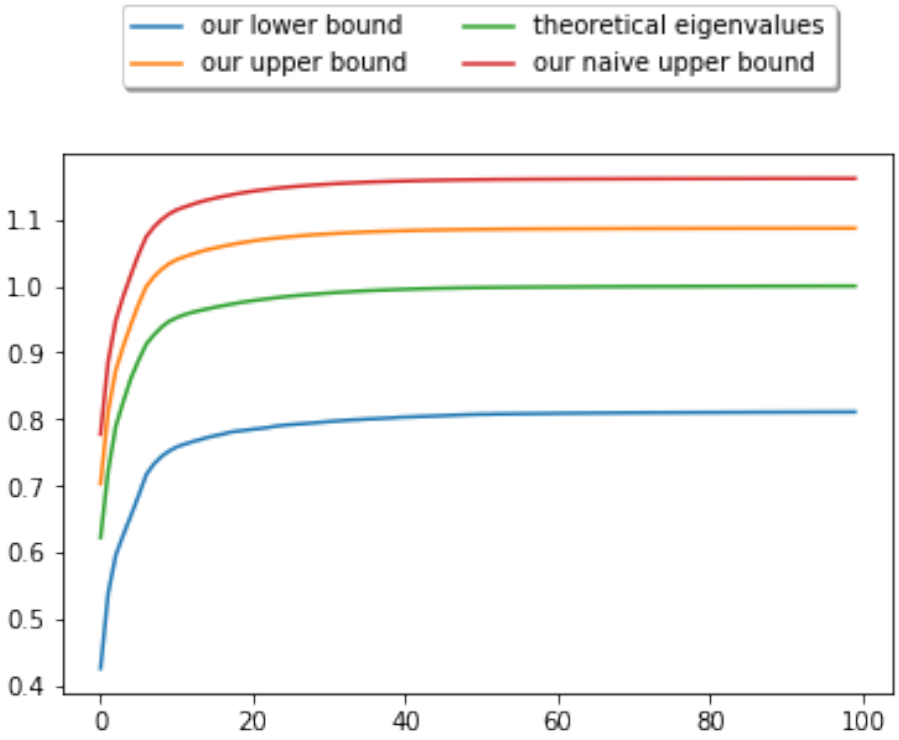

Figure 2: The x-axis is the number k of considered eigenvalues to compute the bound, the y-axis is the amount of information contained in the projection, from 0 to 1.

First, because the σ-algebra on $\mathcal{X}$ is $\mathcal{P}(\mathcal{X})$, we know that ψ is measurable.

Measurablity of A_1. Thanks to the definition of $C(S)$, we know that every coordinate of $C(s)$ consists in a linear combination of the coordinates of $(\psi(x_1), \cdots, \psi(x_m)$. Thus, A_1 is continuous therefore measurable.

Measurablity of A_2. To prove that A_2 is measurable, we need to show that the eigenvectors from a symmetric matrix ($C(s)$ is indeed symmetric) are a measurable function of this matrix. This result is true: to prove it, we will detail the problem treated in Wilcox (1972). Let us consider a polynomial

$$M(p) = \sum_{|\alpha| \leq q} M_\alpha p^\alpha$$

where $q \in \mathbb{N}$ $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n, |\alpha| = \alpha_1 + \cdots + \alpha_n$. Also, $p = (p_1, ..., p_n) \in \mathbb{R}^n, p^\alpha = p_1^{\alpha 1} \cdots p_n^{\alpha_n}$. Finally, for any α, M_α is an Hermitian matrix of size $d \times d$. Let us consider the following eigenvalue problem for E an Hermitian positive definite matrix:

$$M(p)x = \lambda E x \quad x \in \mathbb{R}^d.$$

If we denote by $\lambda_1(p) \geq \cdots \geq \lambda_d(p)$ the d eigenvalues of $M(p)$ in descending order, we can use the following:

Theorem 34 (Wilcox, 1972, Theorem 2) *There exists functions $v_i : \mathbb{R}^n \to \mathbb{R}^d$, $i \in [d]$, such that*

- *For all $i \in [d]$ and $p \in \mathbb{R}^n$, it holds that $M(p)v_i(p) = \lambda_i(p)v_i(p)$;*
- *For all $1 \in [d]$, the function $v_i(p)$ is Lebesgue measurable.*

We now prove the following lemma:

Lemma 35 *Let $M = (m_{i,j})_{i,j} \in \mathbb{R}^{d\times d}$ be a symmetric matrix, then the eigenvectors of M are measurable functions on the coordinates of M.*

Proof We set for $(i,j) \in [d] \times [d]$, $E_{i,j}$ the matrix with value 1 in coordinate (i,j) and 0 everywhere else. Then we take $n = d(d+1)/2$ and we define $p = (m_{i,j})_{i\leq j} \in \mathbb{R}^n$. We also define for $i < j$, $M_{i,j} := E_{i,j} + E_{j,i}$ and if i=j $M_{i,i} = E_{i,i}$. In any case, $M_{i,j}$ is an Hermitian matrix.

We set for $i \leq j$, $\alpha_{i,j}$ the vector of $\mathbb{R}^n$ whom all the coordinates are null except the (i,j)-th which is equal to 1. Doing so, we have for all (i,j), $p^{\alpha_{i,j}} = m_{i,j}$. Finally we prove that we have the equality

$$M = M(p) = \sum_{i\leq j} M_{i,j} p^{\alpha_{i,j}}$$

because M is symmetric. Now, we take $E = I_d$, then the considered eigenvalue problem is for any $x \in \mathbb{R}^d$:

$$M(p)x = \lambda x.$$

Now we can apply Theorem 34, we obtain that the eigenvectors of M are measurable on p, *i.e.* the coordinates of M. ■

Finally we just have to apply that last lemma to $M = C(s) \in \mathbb{R}^{N_\mathcal{H}\times N_\mathcal{H}}$ to conclude that A_2 is measurable in $C(s)$.

Hence A_2 is measurable.

Measurability of A_3. Thanks to the proof of Proposition 27, we know that if U is the matrix whom the i-th column correspond to $v_i(s)$ the i-th eigenvector of $C(s)$, then we have, if $I_k = \{1..k\}$:

$$\forall x \in \mathcal{X}, f_{\hat{w}_k(s)}(x) = ||P_{\hat{V}_k(s)}(\psi(x))||^2 = \psi(x)'U(I_k)U(I_k)'\psi(x)$$

where $U(I_k) \in \mathbb{R}^{N_\mathcal{H}\times N_\mathcal{H}}$ consists in the k first columns of U filled with $\mathcal{N}_\mathcal{H} - k$ columns of 0. Hence, if we reordinate the coordinate one can affirm that $\hat{w}_k = U(I_k)U(I_k)'$.

Then the coordinates of $\hat{w}_k(s)$ consists in a linear combination of the coordinates of the $k-$first eigenvectors of $C(s)$. So the function from $(\mathbb{R}^{N_\mathcal{H}}) \times N_\mathcal{H}$ to $\mathbb{R}^{N_\mathcal{H}^2}$ which transform the eigenvectors into $\hat{w}_k$ is continuous therefore measurable. Furthermore, $w \mapsto f_w$ is an isomorphism between $\mathbb{R}^{N_\mathcal{H}^2}$ and its dual which are two finite-dimensional spaces, it is then continuous, so measurable.

Finally we can conclude that A_3 is measurable.

Measurability of A_4. Because we chose on $\mathcal{F}_k$ the σ-algebra $\mathcal{P}(\mathcal{F}_k)$, we know that B is measurable on $\mathcal{F}_k$ and so is the indicator function $\mathbb{1}\{\cdot \in B\}$. So A_4 is measurable, hence the final result.

7.2 Proof of Theorem 30

The proof is similar to the one of Theorem 29. Indeed; for some fixed k and $B \in \Sigma_{\mathcal{F}_k}$, we have the following decomposition of $A' : s \mapsto P^k(s, B)$:

$$s \overset{\psi}{\longmapsto} (\psi(x_1) \cdots \psi(x_m)) \overset{A_1}{\longmapsto} C(s) \overset{A_2}{\longmapsto} \text{eigenvectors of } C(s)$$
$$\overset{A'_3}{\longmapsto} f_{\hat{w}_k^{\perp}} \overset{A_4}{\longmapsto} \mathbb{1}\left\{f_{\hat{w}_k^{\perp}} \in B\right\}$$

Where the functions ψ, A_1, A_2, A_4 are the same than those in the proof of Theorem 29 and we already proved their measurability. Thus, because $A' = A_4 \circ A'_3 \circ A_2 \circ A_1 \circ \psi$, we just have to prove that $A'3$ is measurable. For that we suppose that we have $(v_1, ..., v_m)$ the eigenvectors of $C(S)$ (ordinate according to the value of their asociated eigenvalue), if we take as in Proposition 27, U the matrix of the eigenvectors then we have, if $J_k = \{k+1, ..., m\}$:

$$\forall x \in \mathcal{X}, f_{\hat{w}_k^{\perp}(s)}(x) = ||P_{\hat{V}_k^{\perp}(s)}(\psi(x))||^2 = \psi(x)' U(J_k) U(J_k)' \psi(x)$$

where $U(J_k) \in \mathbb{R}^{N_{\mathcal{H}} \times N_{\mathcal{H}}}$ consists in the $N_{\mathcal{H}} - k$ last columns of U filled with k columns of 0. Hence, if we reordinate the coordinate one can affirm that $\hat{w}_k = U(J_k)U(J_k)'$.

Then the coordinates of $\hat{w}_k^{\perp}(s)$ consists in a linear combination of the coordinates of the $k-$first eigenvectors of $C(s)$. So the function from $(\mathbb{R}^{N_{\mathcal{H}}}) \times N_{\mathcal{H}}$ to $\mathbb{R}^{N_{\mathcal{H}}^2}$ which transform the eigenvectors into $\hat{w}_k$ is continuous therefore measurable. Furthermore, $w \mapsto f_w$ is an isomorphism between $\mathbb{R}^{N_{\mathcal{H}}^2}$ and its dual, which are two finite-dimensional spaces, it is then continuous, so measurable.

Finally we can claim that A'_3 is measurable. Hence A' is measurable, which concludes the proof and this paper.

References

R. Amit and R. Meir. Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory. In *International Conference on Machine Learning*, pages 205–214. PMLR, 2018.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

S. Boucheron, G. Lugosi, P. Massart, et al. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

T. Cantelobre, B. Guedj, M. Pérez-Ortiz, and J. Shawe-Taylor. A PAC-Bayesian Perspective on Structured Prediction with Implicit Loss Embeddings. Preprint, 2020. URL https://arxiv.org/abs/2012.03780. Under review. Accessed from arXiv:2012.03780.

O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. Institute of Mathematical Statistics, 2007.

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Uncertainty in Artificial Intelligence*, 2017. Accessed from arXiv:1703.11008.

G. K. Dziugaite and D. M. Roy. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors. In *International Conference on Machine Learning*, pages 1377–1386. PMLR, 2018a.

G. K. Dziugaite and D. M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, pages 8430–8441, 2018b.

B. Guedj. A Primer on PAC-Bayesian Learning. In *Proceedings of the second congress of the French Mathematical Society*, 2019. URL https://arxiv.org/abs/1901.05353.

B. Guedj and L. Li. Sequential learning of principal curves: Summarizing data streams on the fly. arXiv:1805.07418, 2018. URL https://arxiv.org/abs/1805.07418. Under review.

B. Guedj and J. Shawe-Taylor. A Primer on PAC-Bayesian Learning. ICML 2019 Tutorial, 2019. URL https://bguedj.github.io/icml2019/material/main.pdf.

M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor. PAC-Bayes unleashed: generalisation bounds with unbounded losses. Preprint, 2020. URL https://arxiv.org/abs/2006.07279. Under review. Accessed from arXiv:2006.07279.

M. Hein and O. Bousquet. Kernels, associated structures and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, 2004.

T. Hofmann, B. Schölkopf, and A. J. Smola. A tutorial review of RKHS methods in machine learning. Unpublished, 2005. Accessed from http://alex.smola.org.

M. Holland. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems*, pages 2715–2724, 2019.

I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

C. Kim and D. Klabjan. A simple and fast algorithm for L1-Norm Kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1842–1855, 2020.

G. Letarte, P. Germain, B. Guedj, and F. Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6872–6882, 2019.

A. Maurer. A note on the PAC Bayesian theorem. arXiv:cs/0411099, 2004. URL https://arxiv.org/abs/cs/0411099.

D. A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational Learning Theory*, pages 230–234. ACM, 1998. Also one year later in *Machine Learning* 37(3), pages 355–363, 1999.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999.

D. A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(1): 148–188, 1989.

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes Un-Expected Bernstein Inequality. In *Advances in Neural Information Processing Systems*, pages 12202–12213, 2019.

Z. Mhammedi, B. Guedj, and R. C. Williamson. PAC-Bayesian Bound for the Conditional Value at Risk. *Advances in Neural Information Processing Systems*, 2020. To appear.

K. Nozawa, P. Germain, and B. Guedj. PAC-Bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 21–30. PMLR, 2020.

K. Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

O. Rivasplata, E. Parrado-Hernández, J. S. Shawe-Taylor, S. Sun, and C. Szepesvári. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems*, volume 31, pages 9214–9224, 2018.

O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor. PAC-Bayes Analysis Beyond the Usual Bounds. *Advances in Neural Information Processing Systems*, 2020. To appear.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

M. Seeger. PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification. *Journal of Machine Learning Research*, 3(Oct):233–269, 2002.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the tenth annual conference on Computational Learning Theory*, pages 2–9, 1997.

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

H. X. Vo and L. J. Durlofsky. Regularized kernel PCA for the efficient parameterization of complex geological models. *Journal of Computational Physics*, 322:859–881, 2016.

C. H. Wilcox. Measurable eigenvectors for Hermitian matrix-valued polynomials. *Journal of Mathematical Analysis and Applications*, 40(1):12–19, 1972.

Z. Xu, J. Liu, X. Luo, Z. Yang, Y. Zhang, P. Yuan, Y. Tang, and T. Zhang. Software defect prediction based on kernel PCA and weighted extreme learning machine. *Information and Software Technology*, 106:182–200, 2019.