Mach Learn (2007) 66:259–294
DOI 10.1007/s10994-006-6895-9

Statistical properties of kernel principal component analysis

Gilles Blanchard · Olivier Bousquet · Laurent Zwald

Received: April 2, 2005 / Revised: December 16, 2005 / Accepted: December 22, 2005 / Published online: March 30, 2006

Abstract The main goal of this paper is to prove inequalities on the reconstruction error for kernel principal component analysis. With respect to previous work on this topic, our contribution is twofold: (1) we give bounds that explicitly take into account the empirical centering step in this algorithm, and (2) we show that a "localized" approach allows to obtain more accurate bounds. In particular, we show faster rates of convergence towards the minimum reconstruction error; more precisely, we prove that the convergence rate can typically be faster than $n^{-1/2}$. We also obtain a new *relative* bound on the error.

A secondary goal, for which we present similar contributions, is to obtain convergence bounds for the partial sums of the biggest or smallest eigenvalues of the kernel Gram matrix towards eigenvalues of the corresponding kernel operator. These quantities are naturally linked to the KPCA procedure; furthermore these results can have applications to the study of various other kernel algorithms.

The results are presented in a functional analytic framework, which is suited to deal rigorously with reproducing kernel Hilbert spaces of infinite dimension.

Keywords Kernel principal components analysis · Fast convergence rates · Kernel spectrum estimation · Covariance operator · Kernel integral operator

Editor: Nicolo Cesa-Bianchi

G. Blanchard (✉)
Fraunhofer FIRST (IDA), Kékuléstr. 7, D-12489 Berlin, Germany
e-mail: blanchar@first.fhg.de

O. Bousquet
Pertinence, France
e-mail: oliver.bousquet@tuebingen.mpg.de

L. Zwald
Département de Mathématiques, Université Paris-Sud, Bat.425, F-91405, France
e-mail: andre.elisseeff@tuebingen.mpg.de

1 Introduction

1.1 Goals of this paper

The main focus of this work is principal component analysis (PCA), and its 'kernelized' variant, kernel PCA (KPCA). PCA is a linear projection method giving as an output a sequence of nested linear subspaces which are adapted to the data at hand. This is a widely used preprocessing method with diverse applications, ranging from dimensionality reduction to denoising. Various extensions of PCA have been explored; applying PCA to a space of functions rather than a space of vectors was first proposed by Besse (1979) (see also the survey of Ramsay and Dalzell, 1991). Kernel PCA (Schölkopf et al., 1999) is an instance of such a method which has boosted the interest in PCA, as it allows to overcome the limitations of linear PCA in a very elegant manner by mapping the data to a high-dimensional feature space.

For any fixed d, PCA finds a linear subspace of dimension d such that the data linearly projected onto it have maximum variance. This is obtained by performing an eigendecomposition of the empirical covariance matrix and considering the span of the eigenvectors corresponding to the leading eigenvalues. This sets the eigendecomposition of the *true* covariance matrix as a natural 'idealized' goal of PCA and begs the question of the relationship between this goal and what is obtained empirically. However, despite being a relatively old and commonly used technique, little has been done on analyzing the statistical performance of PCA. Most of the previous work has focused on the asymptotic behavior of empirical covariance matrices of Gaussian vectors (e.g., Anderson, 1963). Asymptotic results for PCA have been obtained by Dauxois and Pousse (1976), and Besse (1991) in the case of PCA in a Hilbert space.

There is, furthermore, an intimate connection between the covariance operator and the Gram matrix of the data, and in particular between their spectra. In the case of KPCA, this is a crucial point at two different levels. From a practical point of view, this connection allows to reduce the eigendecomposition of the (infinite dimensional) empirical kernel covariance operator to the eigendecomposition of the kernel Gram matrix, which makes the algorithm feasible. From a theoretical point of view, it provides a bridge between the spectral properties of the kernel covariance and those of the so-called *kernel integral operator*.

Therefore, theoretical insight on the properties of kernel PCA reaches beyond this particular algorithm alone: it has direct consequences for understanding the spectral properties of the kernel matrix and the kernel operator. This makes a theoretical study of kernel PCA all the more interesting: the kernel Gram matrix is a central object in all kernel-based methods and its spectrum often plays an important role when studying various kernel algorithms; this has been shown in particular in the case of support vector machines (Williamson et al., 2001). Understanding the behavior of eigenvalues of kernel matrices, their stability and how they relate to the eigenvalues of the corresponding kernel integral operator is thus crucial for understanding the statistical properties of kernel-based algorithms.

Asymptotical convergence and central limit theorems for estimation of integral operator eigenspectrum by the spectrum of its empirical counterpart have been obtained by Koltchinskii and Giné (2000). Recent work of Shawe-Taylor et al. (2002, 2005) (see also the related work of Braun, 2005) has put forward a finite-sample analysis of the properties of the eigenvalues of kernel matrices and related it to the statistical performance of kernel PCA. Our goal in the present work is mainly to extend the latter results in two different directions:

- In practice, for PCA or KPCA, an (empirical) recentering of the data is generally performed. This is because PCA is viewed as a technique to analyze the *variance* of the data; it is often desirable to treat the mean independently as a preliminary step (although, arguably, it is also feasible to perform PCA on uncentered data). This centering was not considered in the cited previous work while we take this step into account explicitly and show that it leads to comparable convergence properties.
- to control the estimation error, Shawe-Taylor et al. (2002, 2005) use what we would call a *global approach* which typically leads to convergence rates of order $n^{-1/2}$. Numerous recent theoretical works on M-estimation have shown that improved rates can be obtained by using a so-called *local approach*, which very coarsely speaking consists in taking the estimation error variance precisely into account. We refer the reader to the works of Massart (2000), Bartlett et al. (2005, 2003), Koltchinskii (2004) (among others). Applying this principle to the analysis of PCA, we show that it leads to improved bounds.

Note that we consider these two types of extension *separately*, not simultaneously. While we believe it possible to combine these two results, in the framework of this paper we choose to treat them independently to avoid additional technicalities. We therefore leave the local approach in the recentered case as an open problem.

To state and prove our results we use an abstract Hilbert space formalism. Its main justification is that some of the most interesting positive definite kernels (e.g., the Gaussian RBF kernel) generate an infinite dimensional reproducing kernel Hilbert space (the "feature space" into which the data is mapped). This infinite dimensionality potentially raises a technical difficulty. In part of the literature on kernel methods, a matrix formalism of finite-dimensional linear algebra is used for the feature space, and it is generally assumed more or less explicitly that the results "carry over" to infinite dimension because (separable) Hilbert spaces have good regularity properties. In the present work, we wanted to state rigorous results directly in an infinite-dimensional space using the corresponding formalism of Hilbert-Schmidt operators and of random variables in Hilbert spaces. This formalism has been used in other recent work related to ours (Mendelson and Pajor, 2005; Maurer, 2004). We hope the necessary notational background which we introduce first will not tax the reader excessively and hope to convince her that it leads to a more rigorous and elegant analysis.

One point we want to stress is that, surprisingly maybe, our results are essentially independent of the "kernel" setting. Namely, they hold for any bounded variable taking values in a Hilbert space, not necessarily a kernel space. This is why we voluntarily delay the introduction of kernel spaces until Section 4, *after* stating our main theorems. We hope that this choice, while possibly having the disadvantage of making the results more abstract at first, will also allow the reader to distinguish more clearly between the mathematical framework needed to prove the results and the additional structure brought forth by considering a kernel space, which allows a richer interpretation of these results, in particular in terms of estimation of eigenvalues of certain integral operators and their relationship to the spectrum of the kernel Gram matrix. In a sense, we take here the exact counterpoint of Shawe-Taylor et al. (2005) who started with studying of the eigenspectrum of the Gram matrix to conclude on the reconstruction error of kernel PCA.

The paper is therefore organized as follows. Section 2 introduces the necessary background on Hilbert spaces, Hilbert-Schmidt operators, and random variables in those spaces. Section 3 presents our main results on the reconstruction error of PCA applied to such variables. In Section 4 show how these results are to be interpreted in the framework of a reproducing kernel Hilbert space and the relation to the eigenspectrum of the kernel Gram matrix. In Section 5,

we compute numerically the different bounds obtained on two 'theoretical' examples in an effort to paint a general picture of their respective merits. Finally, we conclude in Section 6 with a discussion of various open issues.

1.2 Overview of the results

Let us give a quick non-technical overview of the results to come. Let Z be a random variable taking values in a Hilbert space $\mathcal{H}$. If we fix the target dimension d of the projected data, the goal is to recover an optimal d-dimensional space V_d such that the average squared distance between a datapoint Z and its projection on V_d is minimum. This quantity is called the (true) *reconstruction error* and denoted $R(V_d)$. Using available data, this optimal subspace is estimated by $\widehat{V}_d$ using the PCA procedure, which amounts to minimizing the *empirical* reconstruction error. One of the quantities we are interested in is to upper bound the so-called (true) *excess error* of $\widehat{V}_d$ as compared to the optimal V_d, that is, $R(\widehat{V}_d) - R(V_d)$ (which is always nonnegative, by definition). Note that the bounds we obtain are only valid with high probability, since $R(\widehat{V}_d)$ is a random quantity.

Our reference point is an inequality, here dubbed "global bound", obtained by Shawe-Taylor et al. (2005), taking the form

$$R(\widehat{V}_d) - R(V_d) \lesssim \sqrt{\frac{d}{n} \operatorname{tr} C_2'}, \tag{1}$$

where tr denotes the trace, and C_2' is a certain operator related to the fourth moments of the variable. By the symbol $\lesssim$ we mean that we are forgetting (for the purposes of this section) about some terms considered lower-order, and that the inequality is true up to a finite multiplicative constant. This inequality is recalled in Theorem 3.1, with some minor improvements over the original bound of Shawe-Taylor et al. As a first improvement obtained in Theorem 3.5, we prove that this bound also holds if the data is empirically recentered in the PCA procedure (which is often the case, but was not taken into account in the above bound).

Next, we prove two different inequalities improving on the bound (1). Both of them rely on a certain quantity $\rho(d, n)$, which depends on the decay of the eigenvalues of operator C_2', and is *always smaller* than the right-hand side of (1). The first inequality, dubbed "excess bound", reads (Theorem 3.2)

$$R(\widehat{V}_d) - R(V_d) \lesssim B_d \rho(d, n), \tag{2}$$

where $B_d \lesssim (R(V_d) - R(V_{d-1}))^{-1}$. The second inequality, dubbed "relative bound", reads (Theorem 3.4)

$$R(\widehat{V}_d) - R(V_d) \lesssim \sqrt{R(V_d)\rho(d, n)} + \rho(d, n). \tag{3}$$

It is valid under the stronger assumption that the variable Z has a constant norm a.s.: this is the case in particular for kernel PCA when a translation invariant kernel is used. Typically, we expect that (2) exhibits a better behavior than (1) for fixed d when n grows large, while the converse is true for (3) (it will be better than (1) for fixed n and large d). To illustrate what amounts to a possibly confusing picture, we plotted these different bounds on two examples (details are given in Section 5). The result appears in Fig. 1. The conclusion is that, at least when n is large enough, the best bound between (2) and (3) always beats the original bound

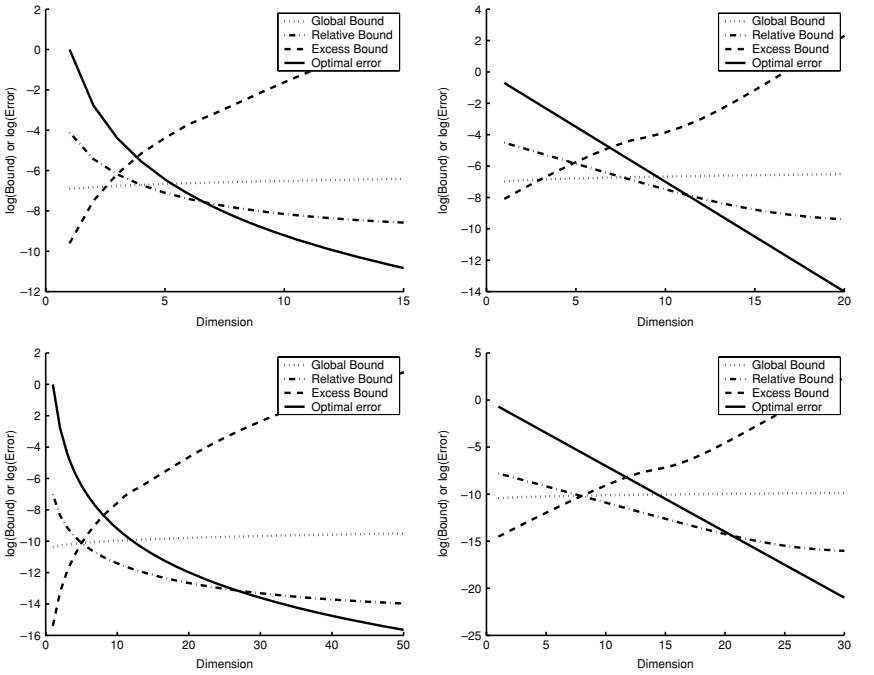

Fig. 1 Comparison of the different (log-) bounds on the excess error in different settings. Left: power decay of eigenvalues. Right: exponential decay of eigenvalues. Top: $n = 10^7$, bottom: $n = 10^{10}$. For details, see Section 5 a final point

(1) (and this, by orders of magnitude). Finally, we show that all bounds except (2) have also an *empirical* counterpart, by which we mean that we obtain bounds of a similar form using purely empirical (hence accessible) quantities.

Finally, when the Hilbert space $\mathcal{H}$ is, additionally, assumed to be a kernel space with reproducing kernel k, our results take a richer interpretation in terms of spectrum estimation of a certain integral operator. Namely, it is known that $R(V_d)$ is exactly equal to the sum of the eigenvalues of rank larger than d of the so-called *kernel integral operator*, while the empirical reconstruction error $R_n(\widehat{V}_d)$ is equal to a similar tail sum of the spectrum of the kernel matrix of the data. This is explained in detail in Section 4.

2 Mathematical preliminaries

Our results revolve around orthogonal projections of a random variable taking values in a Hilbert space $\mathcal{H}$ onto finite dimensional subspaces. Since the space $\mathcal{H}$ is infinite-dimensional, the usual matrix notation used for finite dimensional linear algebra is inappropriate and the most convenient way to deal rigorously with these objects is to use formalism from functional analysis, and in particular to introduce the space of Hilbert-Schmidt operators on $\mathcal{H}$ endowed with a suitable Hilbert structure. The present section is devoted to introducing the necessary notation and basic properties that will be used repeatedly. We first start with generalities on Hilbert-Schmidt operators on Hilbert spaces. We then define more precisely the probabilistic framework used throughout the paper.

2.1 The Hilbert space of Hilbert-Schmidt operators

This section is devoted to recalling some reference material concerning analysis on Hilbert spaces (see, e.g., Dunford & Schwartz, 1963). Let $\mathcal{H}$ be a separable Hilbert space. A linear operator L from $\mathcal{H}$ to $\mathcal{H}$ is called Hilbert-Schmidt if $\sum_{i\geq 1} \|Le_i\|_{\mathcal{H}}^2 = \sum_{i,j\geq 1} \langle Le_i, e_j\rangle^2 < \infty$, where $(e_i)_{i\geq 1}$ is an orthonormal basis of $\mathcal{H}$. This sum is independent of the chosen orthonormal basis and is the squared of the Hilbert-Schmidt norm of L when it is finite. The set of all Hilbert-Schmidt operators on $\mathcal{H}$ is denoted by $\mathrm{HS}(\mathcal{H})$. Endowed with the following inner product $\langle L, N\rangle_{\mathrm{HS}(\mathcal{H})} = \sum_{i\geq 1} \langle Le_i, Ne_i\rangle = \sum_{i,j\geq 1} \langle Le_i, e_j\rangle\langle Ne_i, e_j\rangle$, it is a separable Hilbert space.

A Hilbert-Schmidt operator is compact, it has a countable spectrum and an eigenspace associated to a non-zero eigenvalue is of finite dimension. A compact, self-adjoint operator on a Hilbert space can be diagonalized, i.e., there exists an orthonormal basis of $\mathcal{H}$ made of eigenfunctions of this operator. If L is a compact, positive self-adjoint operator, we will denote $\lambda(L) = (\lambda_1(L) \geq \lambda_2(L) \geq \ldots)$ the sequence of its *positive* eigenvalues sorted in non-increasing order, repeated according to their multiplicities; this sequence is well-defined and contains all nonzero eigenvalues since these are all nonnegative and the only possible limit point of the spectrum is zero. Note that $\lambda(L)$ may be a finite sequence. An operator L is called trace-class if $\sum_{i\geq 1}\langle e_i, Le_i\rangle$ is a convergent series. In fact, the sum of this series is independent of the chosen orthonormal basis and is called the trace of L, denoted by $\mathrm{tr}\, L$. Moreover, $\mathrm{tr}\, L = \sum_{i\geq 1} \lambda_i(L)$ for a self-adjoint operator L.

We will keep switching from $\mathcal{H}$ to $\mathrm{HS}(\mathcal{H})$ and treat their elements as vectors or as operators depending on the context. At times, for more clarity we will index norms and dot products by the space they are to be performed in, although this should always be clear from the objects involved. The following summarizes some notation and identities that will be used in the sequel.

Rank one operators. For $f, g \in \mathcal{H}\backslash\{0\}$ we denote by $f \otimes g^*$ the rank one operator defined as $f \otimes g^*(h) = \langle g, h\rangle f$. The following properties are straightforward from the above definitions:

$$\|f \otimes g^*\|_{\mathrm{HS}(\mathcal{H})} = \|f\|_{\mathcal{H}}\, \|g\|_{\mathcal{H}}\,; \tag{4}$$

$$\mathrm{tr}\, f \otimes g^* = \langle f, g\rangle_{\mathcal{H}}; \tag{5}$$

$$\langle f \otimes g^*, A\rangle_{\mathrm{HS}(\mathcal{H})} = \langle Ag, f\rangle_{\mathcal{H}} \text{ for any } A \in \mathrm{HS}(\mathcal{H})\,. \tag{6}$$

Orthogonal projectors. We recall that an orthogonal projector in $\mathcal{H}$ is an operator U such that $U^2 = U = U^*$ (and hence positive). In particular one has

$$\|U(h)\|_{\mathcal{H}}^2 = \langle h, Uh\rangle_{\mathcal{H}} \leq \|h\|_{\mathcal{H}}^2\,;$$
$$\langle f \otimes g^*, U\rangle_{\mathrm{HS}(\mathcal{H})} = \langle Uf, Ug\rangle_{\mathcal{H}}.$$

U has rank $d < \infty$ (i.e., it is a projection on a finite dimensional subspace), if and only if it is Hilbert-Schmidt with

$$\|U\|_{\mathrm{HS}(\mathcal{H})} = \sqrt{d}, \tag{7}$$

$$\mathrm{tr}\, U = d. \tag{8}$$

In that case it can be decomposed as $U = \sum_{i=1}^{d} \phi_i \otimes \phi_i^*$, where $(\phi_i)_{i=1}^d$ is an orthonormal basis of the image of U.

If V denotes a closed subspace of $\mathcal{H}$, we denote by Π_V the unique orthogonal projector having range V and null space $V^\perp$. When V is of finite dimension, $\Pi_{V^\perp}$ is not Hilbert-Schmidt, but we will denote (with some abuse of notation), for a trace-class operator A,

$$\langle \Pi_{V^\perp}, A \rangle := \operatorname{tr} A - \langle \Pi_V, A \rangle . \tag{9}$$

2.2 Random variables in a Hilbert space

Our main results relate to a bounded variable Z taking values in a (separable) Hilbert space $\mathcal{H}$. In the application we have in mind, kernel PCA, $\mathcal{H}$ is actually a reproducing kernel Hilbert space and Z is the kernel mapping of an input space $\mathcal{X}$ into $\mathcal{H}$. However, we want to point out that these particulars—although of course of primary importance in *practice*, since the reproducing property allows the computation of all relevant quantities—are essentially irrelevant to the nature of our results. This is why we rather consider this abstract framework.

Expectation and covariance operators in a Hilbert space. We recall basic facts about random variables in Hilbert spaces. A random variable Z in a separable Hilbert space is well-defined iff every continuous linear form $\langle e, Z \rangle$, $e \in \mathcal{H}$ is measurable. It has an expectation $e \in \mathcal{H}$ whenever $\mathbb{E}\,\|Z\| < \infty$ and e is then the unique vector satisfying $\langle e, f \rangle_{\mathcal{H}} = \mathbb{E}\langle Z, f \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$. We now introduce the (non-centered) covariance operator through this theorem and definition (a shortened proof can be found in the Appendix):

Theorem 2.1. *If $\mathbb{E}\,\|Z\|^2 < \infty$, there exists a unique operator $C : \mathcal{H} \to \mathcal{H}$ such that*

$$\langle f, Cg \rangle_{\mathcal{H}} = \mathbb{E}\,[\langle f, Z \rangle_{\mathcal{H}} \langle g, Z \rangle_{\mathcal{H}}], \quad \forall f, g \in \mathcal{H} .$$

This operator is self-adjoint, positive, trace-class with $\operatorname{tr} C = \mathbb{E}\,\|Z\|^2$, *and satisfies*

$$C = \mathbb{E}\,[Z \otimes Z^*].$$

We call C the *non-centered covariance* operator of Z.

Let P denote the probability distribution of Z. We assume $Z_1, \ldots, Z_n$ are sampled i.i.d. according to P and we will denote by P_n the empirical measure associated to this sample, i.e., $P_n = \frac{1}{n} \sum \delta_{Z_i}$. With some abuse, for an integrable function $f : \mathcal{H} \to \mathbb{R}$, we will at times use the notation $Pf := \mathbb{E}\,[f(Z)]$ and $P_n f := \frac{1}{n} \sum_{i=1}^n f(Z_i)$.

Let us from now on assume that $\mathbb{E}\left[\|Z\|^4\right] < \infty$. For $z \in \mathcal{H}$, we denote $C_z = z \otimes z^* \in \mathrm{HS}(\mathcal{H})$. Now, let us denote $C_1 : \mathcal{H} \to \mathcal{H}$, respectively $C_2 : \mathrm{HS}(\mathcal{H}) \to \mathrm{HS}(\mathcal{H})$, the non-centered covariance operator associated to the random element Z in $\mathcal{H}$, respectively to $C_Z = Z \otimes Z^*$ in $\mathrm{HS}(\mathcal{H})$. By a direct consequence of Theorem 2.1 we obtain that C_1 is the expectation in $\mathrm{HS}(\mathcal{H})$ of $C_Z = Z \otimes Z^*$ while C_2 is the expectation in $\mathrm{HS}(\mathrm{HS}(\mathcal{H}))$ of $C_Z \otimes C_Z^*$.

In the following we will study empirical counterparts of the above quantities and introduce the corresponding notation: $C_{1,n} = \frac{1}{n} \sum_{i=1}^n Z_i \otimes Z_i^*$ denotes the empirical covariance operator while $C_{2,n} = \frac{1}{n} \sum_{i=1}^n C_{Z_i} \otimes C_{Z_i}^*$. It is straightforward to check that $\operatorname{tr} C_{1,n} = \frac{1}{n} \sum_{i=1}^n \|Z_i\|^2$ and $\operatorname{tr} C_{2,n} = \frac{1}{n} \sum_{i=1}^n \|Z_i\|^4$.

3 Main results

3.1 Framework for PCA in a hilbert space

For all of the results to come, we assume that we are dealing with a *bounded* random variable Z taking values in $\mathcal{H}$, i.e. $\|Z\|^2 \leq M$ a.s. This ensures that $\mathbb{E}\|Z\|^4 < \infty$ and hence the existence of operators C_1 and C_2. This is actually, of course, a much stronger hypothesis than the mere existence of a fourth moment, but we will need it to make use of various concentration theorems.

In this section we first recall the result obtained by Shawe-Taylor et al. (2002, 2005) for which we give a proof for the sake of completeness. This is what we refer to as the "global approach in the uncentered case". We then present our two new contributions: (1) Faster rates of convergence via the local approach in the uncentered case and (2) Study of the empirically recentered case (global approach only).

In the case of "uncentered PCA", the goal is to reconstruct the signal using principal directions of the non-centered covariance operator. Remember we assume that the number d of PCA directions kept for projecting the observations has been fixed *a priori*. We wish to find the linear space of dimension d that conserves the maximal norm, i.e., which minimizes the error (measured through the averaged squared Hilbert norm) of approximating the data by their projections. We will adopt the following notation for the true and empirical *reconstruction error* of a subspace V:

$$R_n(V) = \frac{1}{n}\sum_{j=1}^{n} \|Z_j - \Pi_V(Z_j)\|^2 = P_n\langle \Pi_{V^\perp}, C_Z\rangle = \langle \Pi_{V^\perp}, C_{1,n}\rangle\,,$$

and

$$R(V) = \mathbb{E}[\|Z - \Pi_V(Z)\|^2] = P\langle \Pi_{V^\perp}, C_Z\rangle = \langle \Pi_{V^\perp}, C_1\rangle\,.$$

Let us denote $\mathcal{V}_d$ the set of all vector subspaces of dimension d of $\mathcal{H}_k$. It is well known that the d-dimensional space V_d attaining the best reconstruction error, that is,

$$V_d = \operatorname*{Arg\,Min}_{V\in\mathcal{V}_d} R(V),$$

is obtained as the span of the first d eigenvectors of operator C_1. This definition is actually abusive if the above Arg Min is not reduced to a single element, i.e. the eigenvalue $\lambda_d(C_1)$ is multiple. In this case, unless said otherwise any arbitrary choice of the minimizer is fine. Its empirical counterpart, the space $\widehat{V}_d$ minimizing the empirical error,

$$\widehat{V}_d = \operatorname*{Arg\,Min}_{V\in\mathcal{V}_d} R_n(V)\,, \tag{10}$$

is the vector space spanned by the first d eigenfunctions of the empirical covariance operator $C_{1,n}$. Finally, it holds that $R_n(\widehat{V}_d) = \sum_{i>d}\lambda_i(C_{1,n})$ and $R(V_d) = \sum_{i>d}\lambda_i(C_1)$.

3.2 The global approach in the uncentered case

We first essentially reformulate a theorem proved by Shawe-Taylor et al. (2005), while adding some minor refinements. The proof will allow us to introduce the main important quantities that will be used in the results to come in the next sections.

Theorem 3.1. *Assume* $\|Z\|^2 \leq M$ *a.s. and that* $Z \otimes Z^*$ *belongs a.s. to a set of* $\mathrm{HS}(\mathcal{H})$ *with bounded diameter* L. *Then for any* $n \geq 2$, *with probability at least* $1 - 3e^{-\xi}$,

$$|R(\widehat{V}_d) - R_n(\widehat{V}_d)| \leq \sqrt{\frac{d}{n-1} \operatorname{tr} C'_{2,n}} + (M \wedge L)\sqrt{\frac{\xi}{2n}} + L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}}\,. \tag{11}$$

Also, with probability at least $1 - 2e^{-\xi}$,

$$0 \leq R(\widehat{V}_d) - R(V_d) \leq \sqrt{\frac{d}{n} \operatorname{tr} C'_2} + 2(M \wedge L)\sqrt{\frac{\xi}{2n}}\,, \tag{12}$$

where $C'_2 = C_2 - C_1 \otimes C_1^*$ *and* $C'_{2,n} = C_{2,n} - C_{1,n} \otimes C_{1,n}^*$.

Comments. (1) It should be clear from the proof that the right-hand side members of the two above inequalities are essentially interchangeable between the two bounds (up to changes of the constant in front of the deviation term). We picked this particular formulation choice in the above theorem with the following thought in mind: we interpret inequality (11) as a confidence interval on the true reconstruction error that can be computed from purely empirical data. On the other hand, inequality (12) concerns the *excess error* of $\widehat{V}_d$ with respect to the optimal V_d. The optimal error is *not* available in practice, which means that this inequality is essentially useful to study from a theoretical point of view the convergence properties of $\widehat{V}_d$ to V_d (in the sense of reconstruction error). In this case we would typically be more interested to relate this convergence to intrinsic properties of P, not P_n.

(2) With respect to Shawe-Taylor et al. (2005), we introduce the following minor improvements: (a) the main term involves $C'_2 = C_2 - C_1 \otimes C_1^*$ instead of C_2 (note that $\operatorname{tr}(C_1 \otimes C_1^*) = \|C_1\|^2$, but we chose the former—if perhaps less direct—formulation for an easier comparison to Theorems 3.2 and 3.4, to come in the next section); (b) the factor in front of the main term is 1 instead of $\sqrt{2}$; (c) we can take into account additional information on the diameter L (note that $L \leq 2M$ always holds) of the support of C_Z if it is available. For example, if the Hilbert space is a kernel space with kernel k on a input space $\mathcal{X}$ (see Section 4 for details), then $L^2 = \sup_{x,y\in\mathcal{X}}(k^2(x,x) + k^2(y,y) - 2k^2(x,y))$; in the case of a Gaussian kernel with bandwidth σ over a input space of diameter D, this gives $L^2 = 2(1 - \exp(-D^2/\sigma^2))$ which can be smaller than $M = 1$.

Proof: We have

$$R(\widehat{V}_d) - R_n(\widehat{V}_d) = (P - P_n)\big\langle \Pi_{\widehat{V}_d^\perp}, C_Z \big\rangle \leq \sup_{V \in \mathcal{V}_d} (P - P_n)\langle \Pi_{V^\perp}, C_Z \rangle\,. \tag{13}$$

For any finite dimensional subspace V, we have by definition

$$\langle \Pi_{V^\perp}, C_z \rangle = \operatorname{tr} C_z - \langle \Pi_V, z \otimes z^* \rangle = \|z\|^2 - \|\Pi_V(z)\|^2 = \|\Pi_{V^\perp}(z)\|^2 \,, \tag{14}$$

which implies in turn that $\langle \Pi_{V^\perp}, C_Z \rangle \in [0, M]$ a.s.

However, another inequality is also available from the assumption about the support of C_Z. Namely, let z, z' belong to the support of the variable Z; and let $z_\perp, {z'}_\perp$ denote their orthogonal projections on $V^\perp$. Then $z_\perp \otimes z_\perp^*$ is the orthogonal projection of $z \otimes z^*$ on $V^\perp \otimes V^{\perp *}$. By the contractivity property of an orthogonal projection, we therefore have

$$\begin{aligned}
\|z \otimes z^* - z' \otimes z'^*\| &\geq \|z_\perp \otimes z_\perp^* - {z'}_\perp \otimes {z'}_\perp^*\| \\
&\geq |\|z_\perp \otimes z_\perp^*\| - \|{z'}_\perp \otimes {z'}_\perp^*\|| \\
&= |\|z_\perp\|^2 - \|{z'}_\perp\|^2| \\
&= |\langle \Pi_{V^\perp}, z \otimes z^* - z' \otimes z'^* \rangle| \,,
\end{aligned}$$

so that we get in the end

$$|\langle \Pi_{V^\perp}, C_z - C_{z'} \rangle| \leq \|C_z - C_{z'}\| \leq L,$$

by assumption on the diameter of the support of C_Z. Finally, we have $|\langle \Pi_{V^\perp}, C_z - C_{z'} \rangle| \leq L \wedge M$. We can therefore apply the bounded difference concentration inequality (Theorem B.1 recalled in the Appendix) to the variable $\sup_{V \in \mathcal{V}_d} (P_n - P)\langle \Pi_V, C_Z \rangle$, yielding that with probability $1 - e^{-\xi}$,

$$\sup_{V \in \mathcal{V}_d} (P_n - P)\langle \Pi_{V^\perp}, C_Z \rangle \leq \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} (P_n - P)\langle \Pi_{V^\perp}, C_Z \rangle \right] + (M \wedge L)\sqrt{\frac{\xi}{2n}} \,. \tag{15}$$

Naturally, the same bound holds when replacing $(P_n - P)$ by $(P - P_n)$.

We now bound the above expectation term:

$$\begin{aligned}
\mathbb{E}\left[\sup_{V \in \mathcal{V}_d} (P_n - P)\langle \Pi_{V^\perp}, C_Z \rangle \right] &= \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \frac{1}{n} \sum_i C_{Z_i} - \mathbb{E}\,[C_{Z'}] \right\rangle \right] \\
&\leq \sqrt{d}\, \mathbb{E}\left[\left\| \frac{1}{n} \sum_i C_{Z_i} - \mathbb{E}\,[C_{Z'}] \right\| \right] \\
&\leq \sqrt{d}\, \mathbb{E}\left[\left\| \frac{1}{n} \sum_i C_{Z_i} - \mathbb{E}\,[C_{Z'}] \right\|^2 \right]^{\frac{1}{2}} \\
&= \sqrt{\frac{d}{n}} \sqrt{\mathbb{E}[\|C_Z - \mathbb{E}[C_{Z'}]\|^2]} \,,
\end{aligned}$$

where we have used first Cauchy-Schwarz's inequality and the fact that $\|\Pi_V\| = \sqrt{d}$, then Jensen's inequality. It holds that $\mathbb{E}[\|C_Z - \mathbb{E}[C_{Z'}]\|^2] = \frac{1}{2}\mathbb{E}[\|C_Z - C_{Z'}\|^2]$, where Z' is an independent copy of Z. Therefore, we can apply Hoeffding's inequality (Theorem B.2 of the

Appendix, used with parameter $r = 2$) to obtain that with probability at least $1 - e^{-\xi}$, the following bound holds:

$$\mathbb{E}[\|C_Z - \mathbb{E}[C_{Z'}]\|^2] \leq \frac{1}{2n(n-1)} \sum_{i \neq j} \left\| C_{Z_i} - C_{Z_j} \right\|^2 + L^2 \sqrt{\frac{\xi}{n}};$$

finally it can be checked that

$$\frac{1}{n^2} \sum_{i \neq j} \left\| C_{Z_i} - C_{Z_j} \right\|^2 = 2 \operatorname{tr}(C_{2,n} - C_{1,n} \otimes C_{1,n}^*),$$

which leads to the first part of the theorem after applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.

For the second part, the definition of $\widehat{V}_d$ implies that

$$0 \leq R(\widehat{V}_d) - R(V_d) \leq (R(\widehat{V}_d) - R_n(\widehat{V}_d)) - (R(V_d) - R_n(V_d)).$$

The first term is controlled as above, except that we don't apply Hoeffding's inequality but write directly instead

$$\mathbb{E}[\|C_Z - \mathbb{E}[C_{Z'}]\|^2] = \operatorname{tr}(C_2 - C_1 \otimes C_1^*).$$

We obtain a lower bound for the second term using Hoeffding's inequality (Theorem B.2 this time with $r = 1$). This concludes the proof. □

3.3 Localized approach I: Fast rates

The so-called *localized* approach gives typically more accurate results than the global approach by taking into account the *variance* of the empirical processes involved. When the variance can in turn be upper-bounded by some multiple of the expectation, this generally gives rise to more precise bounds.

Interestingly, it turns out that we can obtain different inequalities depending on the function class to which we apply the "localization" technique. In this first section we will apply it to the *excess loss* class; in the next section we will obtain a different result by applying the technique to the loss class itself.

A similar key quantity appears in these two different applications and we will define it here beforehand:

$$\rho(A, d, n) = \inf_{h \geq 0} \left\{ A \frac{h}{n} + \sqrt{\frac{d}{n} \sum_{j > h} \lambda_j(C_2')} \right\}, \tag{16}$$

where we recall that $C_2' = C_2 - C_1 \otimes C_1^*$. As this quantity will appear in the main terms of the bounds in several results to come, is it relevant to notice already that it is always smaller than the quantity $\sqrt{\frac{d}{n} \operatorname{tr} C_2'}$ appearing in Theorem 3.1. In fact, depending on the behavior of the eigenvalues of c_2', the behavior of ρ as a power of n can vary from $n^{-\frac{1}{2}}$ to n^{-1} (when C_2' is finite dimensional). We give some examples in Section 5.

In the first application, we will obtain a result showing an improved convergence rate (as a function of n, and for fixed d) of the reconstruction error of $\widehat{V}_d$ to the optimal one, that is,

a bound improving on (12). This however comes at the price of an additional factor related to the size of the gap between two successive distinct eigenvalues.

Here is the main result of this section:

Theorem 3.2. *Assume $\|Z\|^2 \leq M$ a.s. Let (λ_i) denote the ordered eigenvalues with multiplicity of C_1, resp. (μ_i) the ordered distinct eigenvalues. Let $\tilde{d}$ be such that $\lambda_d = \mu_{\tilde{d}}$. Define*

$$\gamma_d = \begin{cases} \mu_{\tilde{d}} - \mu_{\tilde{d}+1} & \text{if } \tilde{d} = 1 \text{ or } \lambda_d > \lambda_{d+1}, \\ \min\left(\mu_{\tilde{d}-1} - \mu_{\tilde{d}}, \mu_{\tilde{d}} - \mu_{\tilde{d}+1}\right) & \text{otherwise}; \end{cases} \tag{17}$$

and $B_d = \left(\mathbb{E}\langle Z, Z'\rangle^4\right)^{\frac{1}{2}} \gamma_d^{-1}$ (where Z' is an independent copy of Z).

Then for all d, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$ the following holds:

$$R(\widehat{V}_d) - R(V_d) \leq 24\rho(B_d, d, n) + \frac{\xi(11M + 7B_d)}{n}. \tag{18}$$

Comments. As a consequence of the earlier remarks about ρ, the complexity term obtained in Theorem 3.2 has a faster (or equal) decay rate, as a function of the sample size n, than the one of Theorem 3.1; this rate depends on the decay behavior of the eigenvalues.

Note that in contrast to the other theorems, we do not state an empirical version of the bound (that would use only empirical quantities). It is possible (up to worse multiplicative constants) to replace the operator C_2' appearing in ρ by the empirical $C_{2,n}'$ (see Theorem 3.4 below for an example of how this plays out). However, to have a fully empirical quantity, the constant B_d would also have to be empirically estimated. We leave this point as an open problem here, although we suspect simple convergence result of the empirical eigenvalues to the true ones (as proved for example by Koltchinskii and Giné, 2000) may be sufficient to obtain a fully empirical result.

At the core of the proof of the theorem we use general results due to Bartlett et al. (2005) using localized Rademacher complexities. We recall a succinct version of these results here. We first need the following notation: let $\mathcal{X}$ be a measurable space and $X_1, \ldots, X_n$ a n-uple of points in X; for a class of functions $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$, we denote

$$\mathcal{R}_n\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i),$$

where $(\varepsilon_i)_{i=1\ldots n}$ are i.i.d. Rademacher variables. The *star-shaped hull* of a class of functions $\mathcal{F}$ is defined as

$$\text{star}(\mathcal{F}) = \{\lambda f \mid f \in \mathcal{F},\ \lambda \in [0,1]\}.$$

Finally, a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is called *sub-root* if it is nonnegative, nondecreasing, and such that $\psi(r)/\sqrt{r}$ is nonincreasing. It can be shown that the fixed point equation $\psi(r) = r$ has a unique positive solution (except for the trivial case $\psi \equiv 0$).

Theorem 3.3 (Bartlett, Bousquet and Mendelson). *Let $\mathcal{X}$ be a measurable space, P be a probability distribution on $\mathcal{X}$ and $X_1, \ldots, X_n$ an i.i.d. sample from P. Let $\mathcal{F}$ be a class of functions on X ranging in $[-1, 1]$ and assume that there exists some constant $B > 0$ such*

that for every $f \in \mathcal{F}$, $Pf^2 \le BPf$. *Let* ψ *be a sub-root function and* r^* *be the fixed point of* ψ. *If* ψ *satisfies*

$$\psi(r) \ge B\mathbb{E}_{X,\varepsilon}\mathcal{R}_n\{f \in \mathrm{star}(\mathcal{F}) \mid Pf^2 \le r\},$$

then for any $K > 1$ *and* $x > 0$, *with probability at least* $1 - e^{-x}$,

$$\forall f \in \mathcal{F}, \quad Pf \le \frac{K}{K-1}P_n f + \frac{6K}{B}r^* + \frac{x(11+5BK)}{n}; \tag{19}$$

also, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F}, \quad P_n f \le \frac{K+1}{K}Pf + \frac{6K}{B}r^* + \frac{x(11+5BK)}{n}. \tag{20}$$

Furthermore, if $\widehat{\psi}_n$ *is a data-dependent sub-root function with fixed point* $\widehat{r}^*$ *such that*

$$\widehat{\psi}_n(r) \ge 2(10 \vee B)\mathbb{E}_\varepsilon\mathcal{R}_n\{f \in \mathrm{star}(\mathcal{F}) \mid P_n f^2 \le 2r\} + \frac{(2(10 \vee B)+11)x}{n}, \tag{21}$$

then with probability $1 - 2e^{-x}$, *it holds that* $\widehat{r}^* \ge r^*$; *as a consequence*, *with probability* $1 - 3e^{-x}$, *inequality* (19) *holds with* r^* *replaced by* $\widehat{r}^*$; *similarly for inequality* (20).

Proof of Theorem 3.2. The main idea of the proof is to apply Theorem 3.3 to the class of excess losses $f(z) = \langle \Pi_{V^\perp} - \Pi_{V_d^\perp}, C_z\rangle$, $V \in \mathcal{V}_d$. However, at this point already we find ourselves in a quagmire from the fact that V_d, the optimal d-dimensional space, is actually not always uniquely defined in the case the eigenvalue $\lambda_d(C_1)$ has multiplicity greater than 1. Up until now, in this situation we have let the actual choice of $V_d \in \mathrm{Arg\,Min}_{V \in \mathcal{V}_d} R(V)$ unspecified since it did not alter the results. However, for the present proof this choice does matter, because although the choice of V_d has no influence on the expectation Pf of the above functions, it changes the value of Pf^2, which is of primary importance in the assumptions of Theorem 3.3: more precisely we need to ensure that $Pf^2 \le BPf$ for some constant B.

It turns out that in order to have this property satisfied, we need to pick a minimizer of the true loss, $H_V \in \mathrm{Arg\,Min}_{V' \in \mathcal{V}_d} R(V')$ *depending* on V. More precisely, for each $V \in \mathcal{V}_d$ it is possible to find an element $H_V \in \mathcal{V}_d$ such that:

$$R(H_V) = \min_{H \in \mathcal{V}_d} R(H) = R(V_d), \tag{22}$$

and

$$\mathbb{E}\big[\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z\rangle^2\big] \le 2B_d\,\mathbb{E}\big[\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z\rangle\big], \tag{23}$$

where B_d is defined in the statement of the theorem. This property is proved in Lemma A.1 in the Appendix.

We now consider the class of functions

$$\mathcal{F}_d = \big\{ f_V : x \mapsto \langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_x\rangle \,\big|\, V \in \mathcal{V}_d \big\},$$

where for each $V \in \mathcal{V}_d$, H_V is obtained via the above. We will apply Theorem 3.3 to the class $M^{-1}\mathcal{F}_d$. For any $f \in M^{-1}\mathcal{F}_d$, it holds that $f \in [-1, 1]$; furthermore, inequality (23) entails that $Pf^2 \leq M^{-1}B_d Pf$.

We now need to upper bound the local Rademacher complexities of the star-shaped hull of $\mathcal{F}_d$. We first note that $\Pi_{V^\perp} - \Pi_{H_V^\perp} = \Pi_{H_V} - \Pi_V$ and $\|\Pi_V - \Pi_{H_V}\|^2 \leq 4d$, where we have used the triangle inequality and the fact that $\|\Pi_V\|^2 = \dim(V)$. Therefore,

$$\mathcal{F}_d \subset \{x \mapsto \langle \Gamma, C_x \rangle \mid \Gamma \in \mathrm{HS}(\mathcal{H}), \|\Gamma\|^2 \leq 4d\}.$$

Since the latter set is convex and contains the origin, it therefore also contains $\mathrm{star}(\mathcal{F}_d)$. On the other hand, for a function of the form $f(x) = \langle \Gamma, C_x \rangle$, it holds true that $Pf^2 = \mathbb{E}[\langle \Gamma, C_X \rangle^2] = \langle \Gamma, C_2 \Gamma \rangle$ by definition of operator C_2. Hence, we have

$$\begin{aligned}\{g \in \mathrm{star}(M^{-1}\mathcal{F}_d) \mid Pg^2 \leq r\} &= M^{-1}\{g \in \mathrm{star}(\mathcal{F}_d) \mid Pg^2 \leq M^2 r\} \\ &\subset M^{-1}\{x \mapsto \langle \Gamma, C_x \rangle \mid \|\Gamma\|^2 \leq 4d, \langle \Gamma, C_2 \Gamma \rangle \leq M^2 r\} := \mathcal{S}_r.\end{aligned}$$

The goal is now to upper bound $\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_r$. For this we first decompose each function in this set as $\langle \Gamma, C_x \rangle = \langle \Gamma, C_x - C_1 \rangle + \langle \Gamma, C_1 \rangle$, so that

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_r \leq \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{1,r} + \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r},$$

defining the set of constant functions

$$\mathcal{S}_{1,r} = M^{-1}\{x \mapsto \langle \Gamma, C_1 \rangle \mid \langle \Gamma, C_2 \Gamma \rangle \leq M^2 r\};$$

and the set of centered functions

$$\mathcal{S}_{2,r} = M^{-1}\{x \mapsto \langle \Gamma, C_x - C_1 \rangle \mid \|\Gamma\|^2 \leq 4d, \langle \Gamma, (C_2 - C_1 \otimes C_1^*)\Gamma \rangle \leq M^2 r\},$$

note that in these set definitions we have relaxed some conditions on the functions in the initial set $\mathcal{S}_r$, keeping only what we need to obtain the desired bound: for $\mathcal{S}_{r,1}$ we dropped the condition on $\|\Gamma\|$ and for $\mathcal{S}_{r,2}$ we replaced C_2 by $C_2' = C_2 - C_1 \otimes C_1^*$. Remark that this last operator is still positive, since by definition

$$\langle \Gamma, C_2 \Gamma \rangle = \mathbb{E}[\langle C_Z, \Gamma \rangle^2] \geq \mathbb{E}[\langle C_Z, \Gamma \rangle]^2 = \langle \Gamma, C_1 \rangle^2 = \langle \Gamma, \left(C_1 \otimes C_1^*\right) \Gamma \rangle. \tag{24}$$

Bounding the Rademacher complexity of $\mathcal{S}_{1,r}$ is relatively straightforward since it only contains constant functions, and one can check easily that for a set of scalars $A \subset \mathbb{R}$,

$$\mathbb{E}\left[\sup_{a \in A}\left(a \sum_{i=1}^n \varepsilon_i\right)\right] = \frac{1}{2}(\sup A - \inf A)\,\mathbb{E}\left[\left|\sum_{i=1}^n \varepsilon_i\right|\right] \leq \frac{1}{2}(\sup A - \inf A)\sqrt{n},$$

leading to

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{1,r} \leq M^{-1} n^{-\frac{1}{2}} \sup\{\langle \Gamma, C_1 \rangle \mid \langle \Gamma, C_2 \Gamma \rangle \leq M^2 r\} \leq \sqrt{\frac{r}{n}},$$

where we have used (24). To deal with the Rademacher complexity of $S_{2,r}$, we introduce an orthonormal basis (Φ_i) of eigenvectors of operator C_2'. Let Γ be any element of $\mathrm{HS}(\mathcal{H})$ such that

$$\|\Gamma\|^2 = \sum_i \langle \Gamma, \Phi_i\rangle^2 \le 4d, \quad \text{and} \quad \langle \Gamma, C_2'\Gamma\rangle = \sum_i \lambda_i(C_2')\langle \Gamma, \Phi_i\rangle^2 \le M^2 r.$$

Now, for any integer $h \le Rank(C_2')$,

$$\begin{aligned}
&\sum_{i=1}^n \varepsilon_i \langle \Gamma, C_{Z_i} - C_1\rangle \\
&\quad = \sum_{j=1}^h \langle \Gamma, \Phi_j\rangle \left\langle \Phi_j, \sum_{i=1}^n \varepsilon_i \left(C_{Z_i} - C_1\right)\right\rangle + \sum_{j>h} \langle \Gamma, \Phi_j\rangle \left\langle \Phi_j, \sum_{i=1}^n \varepsilon_i \left(C_{Z_i} - C_1\right)\right\rangle \\
&\quad \le M\left(r \sum_{i=1}^h \frac{1}{\lambda_i(C_2')} \left\langle \sum_{j=1}^n \varepsilon_j \left(C_{Z_j} - C_1\right), \Phi_i\right\rangle^2\right)^{1/2} \\
&\qquad + 2\left(d \sum_{i\ge h+1} \left\langle \sum_{j=1}^n \varepsilon_j \left(C_{Z_j} - C_1\right), \Phi_i\right\rangle^2\right)^{1/2}, \qquad (25)
\end{aligned}$$

where we used the Cauchy-Schwarz inequality for both terms. We now integrate over (ε_i) and (Z_i); using Jensen's inequality the square roots are pulled outside of the expectation; and we have, for each $i \ge 1$,

$$\begin{aligned}
&\mathbb{E}\mathbb{E}_\varepsilon \left\langle \sum_{j=1}^n \varepsilon_j \left(C_{Z_j} - C_1\right), \Phi_i\right\rangle^2 \\
&\quad = \mathbb{E}\sum_{j=1}^n \langle C_{Z_j} - C_1, \Phi_i\rangle^2 = n\mathbb{E}\langle \Phi_i, \left(C_{2,n} - C_1 \otimes C_1^*\right)\Phi_i\rangle = n\langle \Phi_i, C_2'\Phi_i\rangle = n\lambda_i(C_2').
\end{aligned}$$

Because (25) is valid for any $h \le Rank(C_2')$, we finally obtain the following inequality:

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{S}_{2,r} \le \frac{1}{\sqrt{n}} \inf_{h\ge 0} \left(\sqrt{rh} + 2M^{-1}\sqrt{d \sum_{j\ge h+1} \lambda_j(C_2')}\right) := \psi_0(r),$$

(where the extension of the infimum to $h > Rank(C_2')$ is straightforward). It is easy to see any infimum of sub-root functions is sub-root, hence ψ_0 is sub-root. To conclude, we need to upper bound the fixed point of the sub-root function $\psi(r) = M^{-1}B_d(\psi_0(r) + \sqrt{r/n})$.

To obtain a bound, we solve $r^* \le \frac{2M^{-1}B_d}{\sqrt{n}}\{(h^{\frac{1}{2}} + 1)\sqrt{r^*} + 2M^{-1}\sqrt{d\sum_{j\ge h+1}\lambda_j}\}$ for each $h \ge 0$ (by using the fact that $x \le A\sqrt{x} + B$ implies $x \le A^2 + 2B$), and take the infimum over h, which leads to

$$r^* \le 8M^{-2}\left(\inf_{h\ge 0}\left\{\frac{B_d^2 h}{n} + B_d\sqrt{\frac{d}{n}\sum_{j\ge h+1}\lambda_j(C_2')}\right\} + \frac{B_d^2}{n}\right).$$

We can now apply Theorem 3.3 at last, obtaining that for any $K > 1$ and every $\xi > 0$, with probability at least $1 - e^{-\xi}$:

$$\forall V \in \mathcal{V}_d, \quad Pf_V \leq \frac{K}{K-1} P_n f_V + 24K\rho(B_d, d, n) + \frac{\xi(11M + 7B_d K)}{n}. \tag{26}$$

We now choose $V = \widehat{V}_d$ in the above inequality; we have $R(H_{\widehat{V}_d}) = R(V_d)$ and the definition (10) of $\widehat{V}_d$ entails $P_n f_{\widehat{V}_d} \leq 0$. Letting $K \to 1$, we have a family of increasing sets whose probability is bounded by $e^{-\xi}$, so that this also holds for the limiting set $K = 1$: this leads to the announced result. □

3.4 Localized approach II: relative bound

We now apply the localization technique directly to the initial loss class. This gives rise to a *relative bound*, where the bounding quantity also depends on the value of the loss itself: the smaller the loss, the tighter the bound. Unfortunately, we were only able to prove this result under the stronger assumption that the variable Z has a constant norm a.s. (instead of, previously, a bounded norm). Here is the result of this section:

Theorem 3.4. *Assume Z takes values on the sphere of radius $\sqrt{M}$, i.e. $\|Z\|^2 = M$ a.s. Then for all d, $n \geq 2$, $\xi > 0$, with probability at least $1 - 4e^{-\xi}$ the following holds:*

$$|R(\widehat{V}_d) - R_n(\widehat{V}_d)| \leq c\left(\sqrt{R_n(\widehat{V}_d)\left(\rho_n(M, d, n) + M\frac{(\xi + \log n)}{n}\right)} + \rho_n(d, n) + \frac{M(\xi + \log n)}{n}\right), \tag{27}$$

where c is a universal constant ($c \leq 1.2 \times 10^5$). Also, with probability at least $1 - 2e^{-\xi}$,

$$R(\widehat{V}_d) - R(V_d) \leq c\left(\sqrt{R(V_d)\left(\rho(M, d, n) + M\frac{\xi}{n}\right)} + \rho(d, n) + M\frac{\xi}{n}\right), \tag{28}$$

where c is a universal constant ($c \leq 80$), the quantity ρ is defined by (16), and ρ_n is defined similarly by (16) *where the operator C_2' is replaced by its empirical counterpart $C_{2,n}'$.*

Comments. In contrast to Theorem 3.2, the behavior of the above inequalities for fixed d and n tending do infinity is actually *worse* than the original global bound of Theorem 3.1. (The order $\rho(M, d, n)^{\frac{1}{2}}$ as a function of n is typically between $n^{-\frac{1}{2}}$ and $n^{-\frac{1}{4}}$: some more specific examples are given in Section 5.) On the other hand, the behavior for fixed n and varying d is now of greater interest, since $R(\widehat{V}_d)$ goes to zero as d increases. If $R(\widehat{V}_d)$ decreases quickly enough, the bound is actually *decreasing* as a function of d (at least for values of d such that the first term is dominant). This is the only bound which exhibits this behavior.

Proof: In this proof, c will denote a real constant whose exact value can be different from line to line. We start by proving the second part of the theorem. We will apply Theorem 3.3 to the class of functions $M^{-1}\mathcal{G}_d$, where $\mathcal{G}_d$ is the loss class

$$\mathcal{G}_d = \{g_V : x \mapsto \langle \Pi_{V^\perp}, C_x \rangle \mid V \in \mathcal{V}_d\}.$$

From Eq. (14), we know that $\forall g \in M^{-1}\mathcal{G}_d$, $g(x) \in [0,1]$, and therefore $Pg^2 \leq Pg$, hence the first assumptions of Theorem 3.3 are satisfied with $B = 1$.

Hence, we have

$$\begin{aligned}&\{g \in \text{star}(M^{-1}\mathcal{G}_d) \,|\, Pg^2 \leq r\}\\ &\quad = \{g : x \mapsto \lambda M^{-1}(\|x\|^2 - \langle \Pi_V, C_x\rangle) \,|\, V \in \mathcal{V}_d, Pg^2 \leq r, \lambda \in [0,1]\} := \mathcal{L}_r.\end{aligned}$$

The goal is now to upper bound $\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_r$. For this we first decompose each function in this set as

$$M^{-1}\lambda(\|x\|^2 - \langle \Pi_V, C_x\rangle) = M^{-1}\lambda(\|x\|^2 - \langle \Pi_V, C_1\rangle) + M^{-1}\langle \lambda \Pi_V, C_1 - C_x\rangle.$$

Notice that, since we assumed that $\|x\|^2 = M$ a.s., the first above term is a.s. a positive constant equal to $\lambda(1 - M^{-1}\langle \Pi_V, C_1\rangle)$. Furthermore, the L_2 norm of any $g \in \mathcal{L}_r$ can be rewritten as

$$\begin{aligned}Pg^2 &= M^{-2}\lambda^2 P(\|x\|^2 - \langle \Pi_V, C_x\rangle^2)\\ &= \lambda^2(1 - 2M^{-1}\langle \Pi_V, C_1\rangle) + M^{-2}\langle \Pi_V, C_2 \Pi_V\rangle\\ &= (\lambda\left(1 - M^{-1}\langle \Pi_V, C_1\rangle\right))^2 + M^{-2}\langle \lambda \Pi_V, (C_2 - C_1 \otimes C_1^*)\lambda \Pi_V\rangle.\end{aligned}$$

From the two last displays, we can write

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_r \leq \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{1,r} + \mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{2,r},$$

defining the set of constant functions

$$\mathcal{L}_{1,r} = \{x \mapsto c \,|\, 0 \leq c \leq \sqrt{r}\}\,,$$

and the set of centered functions

$$\mathcal{L}_{2,r} = \{x \mapsto M^{-1}\langle \Gamma, C_1 - C_x\rangle \,|\, \|\Gamma\|^2 \leq d, \langle \Gamma, (C_2 - C_1 \otimes C_1^*)\Gamma\rangle \leq M^2 r\}.$$

We can now apply the same device as in the proof of Theorem 3.2 to obtain that

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{1,r} \leq \sqrt{\frac{r}{n}},$$

and

$$\mathbb{E}\mathbb{E}_\varepsilon \mathcal{R}_n \mathcal{L}_{2,r} \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1} \sqrt{d \sum_{j \geq h+1} \lambda_j(C_2')} \right);$$

again following the proof of Theorem 3.2, we obtain by application of Theorem 3.3 that for any $K > 1$ and every $\xi > 0$, with probability at least $1 - e^{-\xi}$:

$$\forall V \in \mathcal{V}_d, \qquad R(V) \leq \frac{K}{K-1} R_n(V) + 12K\rho(M, d, n) + \frac{\xi M(11 + 7K)}{n}, \tag{29}$$

and similarly with probability at least $1 - e^{-\xi}$:

$$\forall V \in \mathcal{V}_d, \qquad R_n(V) \leq \frac{K+1}{K} R(V) + 12K\rho(M, d, n) + \frac{\xi M(11 + 7K)}{n}. \tag{30}$$

We now apply (29) to $\widehat{V}_d$ and (30) to V_d to conclude, using $R_n(\widehat{V}_d) \leq R_n(V_d)$, that with probability at least $1 - 2e^{-\xi}$, for any $K > 2$:

$$R(\widehat{V}_d) - R(V_d) \leq 36\left(\frac{1}{K} R(V_d) + K\left(\rho(M, d, n) + M\frac{\xi}{n}\right)\right),$$

We now choose $K = \max(2, (\rho(M, d, n) + M\frac{\xi}{n})^{-\frac{1}{2}} R(V_d)^{\frac{1}{2}})$; this leads to the conclusion of the last part of the theorem.

For the first part of the theorem, we basically follow the same steps, except that we additionally use (21) of Theorem (3.3) to obtain empirical quantities. It can be checked that if ψ is a sub-root function with fixed point r^* and $\psi_1 = \alpha\psi(r) + \beta$ for nonnegative α, β then the fixed point r_1^* of ψ_1 satisfies $r_1^* \leq 4(\alpha^2 r^* + \beta)$, see for example Lemma 4.10 of Bousquet (2002). So, we can unroll the same reasoning as in the first part of the present proof, except that the covariance operators are replaced by their empirical counterparts and we consider directly the empirical Rademacher complexities without expectation over the sample. Finally we conclude that for any $K > 2$, with probability at least $1 - 4e^{-\xi}$,

$$\forall V \in \mathcal{V}_d, \qquad |R(V) - R_n(V)| \leq c\left(\frac{1}{K} R_n(V) + KM\left(\rho_n(d, n) + \frac{\xi}{n}\right)\right).$$

Using the union bound, we can make this bound uniform over positive integer values of K in the range $[2 \dots n]$ at the price of replacing ξ by $\xi + \log n$. We then apply this inequality to $\widehat{V}_d$ and pick $K = \max(2, \lceil(\rho_n(M, d, n) + M\frac{(\xi + \log n)}{n})^{-\frac{1}{2}} R(V_d)^{\frac{1}{2}}\rceil)$, which, for any $n \geq 3$, is an integer belonging to the integer interval $[2 \dots \sqrt{n}]$ since $R_n(\widehat{V}_d) \leq M$. This leads to the first inequality of the theorem. □

3.5 Recentered case

In this section we extend the results of Theorem 3.1 in a different direction. Namely, we want to prove that a bound of the same order is available if we include empirical re-centering in the procedure, which is commonly done in practice.

For this we first need to introduce additional notation:

$$\bar{Z} = Z - \mathbb{E}[Z] \in \mathcal{H}_k,$$
$$\bar{C}_Z = \bar{Z} \otimes \bar{Z}^* \in \mathrm{HS}(\mathcal{H});$$

Similarly, let us denote $\bar{C}_1$ the covariance operator associated to $\bar{Z}$; therefore, $\bar{C}_1$ is the expectation in HS($\mathcal{H}$) of $\bar{C}_Z$ and satisfies $\bar{C}_1 = C_1 - \mathbb{E}[Z] \otimes \mathbb{E}[Z]^*$.

The quantities $\bar{Z}, \bar{C}_z$ already depend on P through the centering, so that we will define the corresponding quantities for P_n corresponding to an empirical recentering:

$$\widehat{Z} = Z - \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

$$\bar{C}_{Z,n} = \hat{Z} \otimes \hat{Z}^*,$$

$$\overline{C}_{1,n} = \frac{1}{n-1}\sum_{i=1}^{n} \bar{C}_{Z_i,n} = C_{1,n} - \frac{1}{n(n-1)}\sum_{i \neq j} Z_i \otimes Z_j^*.$$

Note that the specific normalization for $\bar{C}_{1,n}$ is chosen so that it is an unbiased estimator of $\bar{C}_1$, that is, $\mathbb{E}\left[\bar{C}_{1,n}\right] = \bar{C}_1$.

In this case the PCA algorithm finds the d-dimensional space minimizing the empirical reconstruction error of the empirically recentered data:

$$\widehat{W}_d = \operatorname*{Arg\,Min}_{V \in \mathcal{V}_d} \frac{1}{n}\sum_{j=1}^{n} \|\widehat{Z}_j - \Pi_V(\widehat{Z}_j)\|^2,$$

and $\hat{W}_d$ is the vector space spanned by the first d eigenfunctions of $\bar{C}_{1,n}$. We also denote by W_d the space spanned by the first d eigenfunctions of $\bar{C}_1$, which minimizes the true average reconstruction error of the truly recentered data:

$$W_d = \operatorname*{Arg\,Min}_{V \in \mathcal{V}_d} \mathbb{E}\|\bar{Z} - \Pi_V(\bar{Z})\|^2.$$

We will adopt the following notation for the reconstruction errors, true and empirical:

$$\bar{R}_n(V) = \frac{1}{n-1}\sum_{j=1}^{n} \|\widehat{Z}_j - \Pi_V(\widehat{Z}_j)\|^2 = \langle \Pi_{V^\perp}, \bar{C}_{1,n} \rangle.$$

$$\bar{R}(V) = \mathbb{E}\|\bar{Z} - \Pi_V(\bar{Z})\|^2 = \langle \Pi_{V^\perp}, \bar{C}_1 \rangle.$$

Again, the reason for the specific normalization of $\bar{R}_n(V)$ is to make it an unbiased estimator of $\bar{R}(V)$.

In this situation we have the following theorem similar to Theorem 3.1:

Theorem 3.5. *Assume that $\|Z\|^2 \leq M$ a.s. Then for any $\xi > 1$ and $n \geq 10$, with probability greater than $1 - 5e^{-\xi}$, the following inequality holds:*

$$|\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)| \leq \sqrt{\frac{d}{n}\operatorname{tr}(C_{2,n} - C_{1,n} \otimes C_{1,n}^*)} + 14M\sqrt{\frac{\xi}{2n}} + 2M\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}};$$

also, with probability at least $1 - 3e^{-\xi}$,

$$0 \leq \bar{R}(\widehat{W}_d) - \bar{R}(W_d) \leq \sqrt{\frac{d}{n} \operatorname{tr}\left(C_2 - C_1 \otimes C_1^*\right)} + 17M\sqrt{\frac{\xi}{n}}$$

The proof of this theorem follows the same structure as for Theorem 3.1, but some additional ingredients are needed to control U-processes arising from the empirical recentering. Note that the leading complexity term is the same as in Theorem 3.1: hence recentering in kernel PCA essentially does not introduce additional complexity to the procedure. A minor downside with respect to Theorem 3.1 is that we lose the refinement introduced by considering the diameter of the support of C_Z.

Proof: We have

$$|\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)| = |\langle \widehat{W}_d, \bar{C}_1 - \bar{C}_{1,n}\rangle| \leq \sup_{V \in \mathcal{V}_d} |\langle \Pi_{V^\perp}, \bar{C}_1 - \bar{C}_{1,n}\rangle|.$$

Denoting $\mu = \mathbb{E}[Z]$, recall the following identities:

$$\bar{C}_1 = C_1 - \mu \otimes \mu^* \quad \text{and} \quad \bar{C}_{1,n} = C_{1,n} - \frac{1}{n(n-1)} \sum_{i \neq j}^{n} Z_i \otimes Z_j^*, \tag{31}$$

from which we obtain

$$\begin{aligned} \sup_{V \in \mathcal{V}_d} |\langle \Pi_{V^\perp}, \bar{C}_{1,n} - \bar{C}_1\rangle| \leq & \sup_{V \in \mathcal{V}_d} |\langle \Pi_{V^\perp}, C_{1,n} - C_1\rangle| \\ & + \sup_{V \in \mathcal{V}_d} \left| \left\langle \Pi_{V^\perp}, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} Z_i \otimes Z_j^* \right\rangle \right|. \end{aligned} \tag{32}$$

It was shown in the proof of Theorem 3.1 that the following holds with probability greater than $1 - 3e^{-\xi}$:

$$\sup_{V \in \mathcal{V}_d} |\langle \Pi_{V^\perp}, C_{1,n} - C_1\rangle| \leq \sqrt{\frac{d}{n}} \sqrt{\operatorname{tr}\left(C_{2,n} - C_{1,n} \otimes C_{1,n}^*\right)} + M\sqrt{\frac{\xi}{2n}} + 2M \frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}},$$

so we now focus on the second term of (32). If we denote
$G(z_1, \ldots, z_n) = \langle \Pi_{V^\perp}, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} z_i \otimes z_j^* \rangle$, then we have for any i_0 :

$$\begin{aligned} & \left| G(z_1, \ldots, z_n) - G\left(z_1, \ldots, z_{i_0 - 1}, z'_{i_0}, z_{i_0 + 1}, \ldots, z_n\right) \right| \\ & \quad \leq \frac{1}{n(n-1)} \left\| \sum_{j \neq i_0} \left(\left(z_{i_0} - z'_{i_0}\right) \otimes z_j^* + z_j \otimes \left(z_{i_0}^* - z'^*_{i_0}\right) \right) \right\| \\ & \quad \leq \frac{2}{n(n-1)} \sum_{j \neq i_0} \left\| z'_{i_0} - z_{i_0} \right\| \left\| z_j \right\| \leq \frac{4M}{n}. \end{aligned}$$

Therefore we can apply the bounded difference inequality (Theorem B.1) to G, so that with probability greater than $1 - e^{-\xi}$,

$$\sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^* - \frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle$$
$$\leq \mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^* - \frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle\right] + 4M\sqrt{\frac{\xi}{2n}}.$$

To deal with the above expectation, we consider Hoeffding's decomposition (see de la Peña and Giné, 1999, p. 137) for U-processes. To this end, we define the following quantities:

$$S_d = \sup_{V\in\mathcal{V}_d}\frac{2}{n}\sum_{j=1}^{n}(\langle\Pi_{V^\perp}, \mu\otimes\mu^*\rangle - \langle\Pi_{V^\perp}(Z_j), \mu\rangle)$$

$$R_d = \sup_{V\in\mathcal{V}_d} -\frac{1}{n(n-1)}\sum_{i\neq j}(\langle\Pi_{V^\perp}, Z_i\otimes Z_j^*\rangle - \langle\Pi_{V^\perp}(Z_j), \mu\rangle$$
$$-\langle\Pi_{V^\perp}(Z_i), \mu\rangle + \langle\Pi_{V^\perp}, \mu\otimes\mu^*\rangle).$$

It can easily be seen that

$$\mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\left\langle \Pi_{V^\perp}, \mu\otimes\mu^* - \frac{1}{n(n-1)}\sum_{i\neq j} Z_i\otimes Z_j^*\right\rangle\right] \leq \mathbb{E}[S_d] + \mathbb{E}[R_d].$$

Gathering the different inequalities up to now, we have with probability greater than $1 - 5e^{-\xi}$:

$$\sup_{V\in\mathcal{V}_d}|\langle\Pi_{V^\perp}, \bar{C}_{1,n} - \bar{C}_1\rangle| \leq \sqrt{\frac{d}{n}}\sqrt{\operatorname{tr}(C_{2,n} - C_{1,n}\otimes C_{1,n}^*)} + 5M\sqrt{\frac{\xi}{2n}} + 2M\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}}$$
$$+\mathbb{E}[S_d] + \mathbb{E}[R_d].$$

We now bound from above the expectation of S_d and R_d using Lemmas 3.6 and 3.7 below, which leads to

$$\mathbb{E}[S_d] \leq 4\frac{\mathbb{E}\|Z\|^2}{\sqrt{n}} \leq 6M\sqrt{\frac{\xi}{2n}},$$

and

$$\mathbb{E}[R_d] \leq \frac{6}{n-1}\mathbb{E}\|Z\|^2 \leq 3M\sqrt{\frac{\xi}{2n}},$$

where we have used the assumptions $\xi > 1$ and $n \geq 10$. This leads to the first inequality of the theorem.

For the second part of the theorem, the definition of $\widehat{W}_d$ implies that

$$0 \leq \bar{R}(\widehat{W}_d) - \bar{R}(W_d) \leq \left(\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d)\right) - \left(\bar{R}(W_d) - \bar{R}_n(W_d)\right).$$

For the first term, we proceed as above except we consider only one-sided bounds and, for the main term, use instead the proof of the second part of Theorem 3.1. We thus obtain that with probability at least $1 - 2e^{-\xi}$,

$$\bar{R}(\widehat{W}_d) - \bar{R}_n(\widehat{W}_d) \leq \sqrt{\frac{d}{n} \operatorname{tr}\left(C_2 - C_1 \otimes C_1^*\right)} + 15M\sqrt{\frac{\xi}{n}}.$$

As for the second term,

$$\bar{R}(W_d) - \bar{R}_n(W_d) = \mathbb{E}\left[\left\langle \Pi_{W_d^\perp}, \bar{C}_{1,n} \right\rangle\right] - \left\langle \Pi_{W_d^\perp}, \bar{C}_{1,n} \right\rangle;$$

and we can write

$$\langle \Pi_{W_d^\perp}, \bar{C}_{1,n} \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} g(Z_i, Z_j),$$

with

$$g(z_1, z_2) = \frac{1}{2}\left\langle z_1 - z_2, \Pi_{W_d^\perp}(z_1 - z_2) \right\rangle.$$

whenever $\|z_1\|^2$ and $\|z_2\|^2$ are bounded by M, we have $g(z_1, z_2) \in [0, M]$, therefore we can apply Hoeffding's inequality (Theorem B.2 with $r = 2$) to conclude that with probability at least $1 - e^{-\xi}$,

$$\bar{R}_n(W_d) - \bar{R}(W_d) \leq M\sqrt{\frac{\xi}{n}}.$$

□

Lemma 3.6. *The random variable S_d defined above satisfies the following inequality:*

$$\mathbb{E}\left[S_d\right] \leq 4\frac{\mathbb{E}\|Z\|^2}{\sqrt{n}}.$$

Proof: A standard symmetrization argument leads to

$$\begin{aligned}
\mathbb{E}[S_d] &\leq \mathbb{E}\mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \frac{4}{n} \sum_{j=1}^n \varepsilon_j \langle \Pi_{V^\perp}(Z_j), \mu \rangle \\
&\leq \frac{4}{n}\mathbb{E}\mathbb{E}_\varepsilon \left\| \Pi_{V^\perp}\left(\sum_{j=1}^n \varepsilon_j Z_j\right)\right\| \|\mu\| \\
&\leq \frac{4}{n}\mathbb{E}\mathbb{E}_\varepsilon \left\| \sum_{j=1}^n \varepsilon_j Z_j \right\| \|\mu\| \\
&\leq \frac{4}{\sqrt{n}}\mathbb{E}\sqrt{\operatorname{tr} C_{1,n}}\, \|\mu\|,
\end{aligned}$$

where we successively applied the Cauchy-Schwarz inequality, the contractivity of an orthogonal projector, and Jensen's inequality. Applying Jensen's inequality again, and the fact that $\|\mu\|^2 = \|\mathbb{E}[Z]\|^2 \leq \mathbb{E}\|Z\|^2$ yields the conclusion. □

Lemma 3.7. *The random variable R_d defined above satisfies the following inequality:*

$$\mathbb{E}R_d \leq \frac{6}{n-1}\mathbb{E}\|Z\|^2 .$$

Remark. The proof uses techniques developed by de la Peña and Giné (1999). Actually, we could directly apply Theorems 3.5.3 and 3.5.1 of this reference, getting a factor 2560 instead of 6. We give here a self-contained proof tailored for our particular case for the sake of completeness and for the improved constant.

Proof: Let us denote (Z'_i) an independent copy of (Z_i). Since $\Pi_{V^\perp}$ is a symmetric operator, using Jensen's inequality ,

$$\mathbb{E}[R_d] \leq \frac{1}{n(n-1)}\mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\sum_{i\neq j} f_V(Z_i, Z_{i'}, Z_j, Z_{j'})\right],$$

where

$$f_V(Z_i, Z_{i'}, Z_j, Z_{j'}) = \langle \Pi_{V^\perp}, Z_i \otimes Z_j^* - Z_{i'} \otimes Z_j^* - Z_i \otimes Z_{j'}^* + Z_{i'} \otimes Z_{j'}^* \rangle.$$

Since $f_V(Z_i, Z_{i'}, Z_j, Z_{j'}) = -f_V(Z_{i'}, Z_i, Z_j, Z_{j'})$ and $f_V(Z_i, Z_{i'}, Z_j, Z_{j'}) = -f_V(Z_i, Z_{i'}, Z_{j'}, Z_j)$, following the proof of the standard symmetrization, we get:

$$\mathbb{E}[R_d] \leq \frac{1}{n(n-1)}\mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\sum_{i\neq j}\varepsilon_i\varepsilon_j f_V(Z_i, Z_{i'}, Z_j, Z_{j'})\right]$$

Therefore,

$$\mathbb{E}[R_d] \leq \frac{2}{n(n-1)}\left(\mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\sum_{i\neq j}\varepsilon_i\varepsilon_j\langle\Pi_{V^\perp}, Z_i \otimes Z_j^*\rangle\right] + \mathbb{E}\left[\sup_{V\in\mathcal{V}_d} -\sum_{i\neq j}\varepsilon_i\varepsilon_j\langle\Pi_{V^\perp}, Z_i \otimes Z_{j'}^*\rangle\right]\right) = \frac{2}{n(n-1)}(A+B);$$

for the first term above we have

$$A \leq \mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\sum_{i,j}\varepsilon_i\varepsilon_j\left\langle\Pi_{V^\perp}, Z_i \otimes Z_j^*\right\rangle\right] = C,$$

while for the second we use

$$B \leq \mathbb{E}\left[\sup_{V\in\mathcal{V}_d} -\sum_{i,j}\varepsilon_i\varepsilon_j\left\langle\Pi_{V^\perp}, Z_i \otimes Z_{j'}^*\right\rangle\right] + \mathbb{E}\left[\sup_{V\in\mathcal{V}_d}\sum_{i}\left\langle\Pi_{V^\perp}, Z_i \otimes Z_{i'}^*\right\rangle\right]$$
$$= D + E.$$

We bound terms C, D, E by the following similar chains of inequalities where we successively use the Cauchy-Schwarz inequality, the contractivity of an orthogonal projector and a standard computation on sums of weighted Rademacher:

$$C \leq \mathbb{E}_Z \mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \left\| \sum_i \varepsilon_i Z_i \right\| \left\| \sum_j \varepsilon_j \Pi_{V^\perp}(Z_j) \right\| \leq \mathbb{E}_Z \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i Z_i \right\|^2 = n\mathbb{E}\,\|Z\|^2 ;$$

$$\begin{aligned} D &\leq \mathbb{E}_{Z,Z'} \mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \left\| \sum_i \varepsilon_i Z_i \right\| \left\| \sum_j \varepsilon_j \Pi_{V^\perp}(Z_{j'}) \right\| \\ &\leq \mathbb{E}_{Z,Z'} \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i Z_i \right\| \left\| \sum_j \varepsilon_j Z_{j'} \right\| \\ &\leq \sqrt{\mathbb{E}_{Z,Z'} \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i Z_i \right\|^2 \mathbb{E}_\varepsilon \left\| \sum_j \varepsilon_j Z_{j'} \right\|^2} = n\mathbb{E}\,\|Z\|^2 ; \end{aligned}$$

$$E \leq \mathbb{E}_{Z,Z'} \sup_{V \in \mathcal{V}_d} \sum_i \|\Pi_{V^\perp}(Z_{i'})\| \, \|Z_i\| \leq \sum_i \mathbb{E}_{Z,Z'} \|Z_{i'}\| \, \|Z_j\| \leq n\mathbb{E}\,\|Z\|^2 .$$

Gathering the previous inequalities, we obtain the conclusion. □

4 Kernel PCA and eigenvalues of integral operators

4.1 Kernel PCA

In this section we review briefly how our results are interpreted in the case where the Hilbert space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with kernel function k. This is the standard framework of kernel PCA. The reason why we mention it only at this point on the paper is to emphasize that our previous results are, actually, largely independent of the RKHS setting and could be expressed for any bounded random variable in an abstract Hilbert space.

In this framework the input space $\mathcal{X}$ is an arbitrary measurable space and X is a random variable on $\mathcal{X}$ with probability distribution P. Let k be a positive definite function on $\mathcal{X}$ and $\mathcal{H}_k$ the associated RKHS. We recall (see, e.g., Aronszajn, 1950) that $\mathcal{H}_k$ is a Hilbert space of real functions on $\mathcal{X}$, containing functions $k(x, \cdot)$ for all $x \in \mathcal{H}_k$ and such that the following *reproducing property* is satisfied:

$$\forall f \in \mathcal{H}_k \qquad \forall x \in \mathcal{X} \qquad \langle f, k(x, .)\rangle = f(x), \tag{33}$$

and in particular

$$\forall x, \quad y \in \mathcal{X} \qquad \langle k(x, \cdot), k(y, \cdot)\rangle = k(x, y).$$

The space $\mathcal{X}$ can be mapped into $\mathcal{H}_k$ via the so-called *feature mapping* $x \in \mathcal{X} \mapsto \Phi(x) = k(x, \cdot) \in \mathcal{H}_k$. The reproducing property entails that $\langle \Phi(x), \Phi(y)\rangle = k(x, y)$ so that we can basically compute all dot products involving images of points of $\mathcal{X}$ in $\mathcal{H}_k$ (and linear combinations thereof) using the kernel k. The kernel PCA procedure then consists in applying PCA to the variable $Z = \Phi(X)$.

We make the following assumptions on the RKHS, which will allow to apply our previous results:

(A1) *$\mathcal{H}_k$ is separable.*
(A2) *For all $x \in \mathcal{X}$, $k(x, .)$ is P-measurable.*
(A3) *There exists $M > 0$ such that $k(X, X) \leq M$ P-almost surely.*

Assumption **(A1)** is necessary in order to apply the theory we developed previously. Typically, a sufficient condition ensuring **(A1)** is that $\mathcal{X}$ is compact and k is a continuous function. Assumption **(A2)** ensures the measureability of all functions in $\mathcal{H}_k$ since they are obtained by linear combinations and pointwise limits of functions $k(x, \cdot)$; it also ensures the measureability of Z. It holds in particular in the case where k is continuous. Finally, assumption **(A3)** ensures that the variable Z is bounded a.s. since $\|Z\|^2 = \|\Phi(X)\|^2 = k(X, X)$. Note that we also required the stronger assumption of $\|Z\|^2 = k(X, X) = M$ a.s. for Theorem 3.2. Although this clearly is a strong assumption, it still covers at least the important class of *translation invariant* kernels of the form $k(x, y) = k(x - y)$ (where $\mathcal{X}$ is in this case assumed to be a Euclidean space), the most prominent of which is the Gaussian kernel $k(x, y) = \exp\left(-\|x - y\|^2/(2\sigma^2)\right)$.

For the computations in $\mathrm{HS}(\mathcal{H}_k)$, the following equalities are available:

$$\operatorname{tr} C_{\Phi(x)} = \left\| C_{\Phi(x)} \right\|_{\mathrm{HS}(\mathcal{H}_k)} = k(x, x), \tag{34}$$

$$\langle C_{\Phi(x)}, C_{\Phi(y)} \rangle_{\mathrm{HS}(\mathcal{H}_k)} = k^2(x, y), \tag{35}$$

$$\langle f, C_{\Phi(x)} g \rangle_{\mathcal{H}_k} = \langle C_{\Phi(x)}, f \otimes g^* \rangle_{\mathrm{HS}(\mathcal{H}_k)} = f(x)g(x). \tag{36}$$

Note incidentally that (35) implies that $\mathrm{HS}(\mathcal{H}_k)$ is actually a natural representation of the RKHS with kernel $k^2(x, y)$. Namely to an operator $A \in \mathrm{HS}(\mathcal{H}_k)$ we can associate the function

$$f_A(x) = \left\langle A, C_{\Phi(x)} \right\rangle_{\mathrm{HS}(\mathcal{H}_k)} = \langle A \cdot \Phi(x), \Phi(x) \rangle_{\mathcal{H}_k} = (A \cdot \Phi(x))(x);$$

with this notation, we have $f_{C_{\Phi(x)}} = k^2(x, \cdot)$, and one can check that (33) is satisfied in $\mathrm{HS}(\mathcal{H}_k)$ with the kernel $k^2(x, y)$ when identifying an operator to its associated function.

Finally, the trace of operators C_2, $C_1 \otimes C_1^*$ and $C_{2,n}$, $C_{1,n} \otimes C_{1,n}^*$ appearing in Theorems 3.1 and 3.5 satisfy the following identities:

$$\operatorname{tr} C_2 = \operatorname{tr} \mathbb{E}\left[C_{\Phi(X)} \otimes C_{\Phi(X)}^*\right] = \mathbb{E}\left[\left\| C_{\Phi(X)} \right\|^2\right] = \mathbb{E}[k^2(X, X)];$$

$$\operatorname{tr} C_{2,n} = \frac{1}{n} \sum_{i=1}^{n} k^2(X_i, X_i);$$

$$\operatorname{tr}\left(C_1 \otimes C_1^*\right) = \|C_1\|^2 = \mathbb{E}\left[k^2(X, Y)\right] \quad \text{(where } Y \text{ is an independent copy of } X\text{)},$$

$$\operatorname{tr}\left(C_{1,n} \otimes C_{1,n}^*\right) = \frac{1}{n^2} \sum_{i=1}^{n} k^2(X_i, X_j).$$

4.2 Eigenvalues of integral operators

We now review the relation of Kernel PCA to eigenvalues and eigenfunctions of the kernel integral operator. Again, this relation is well-known and is actually central to the KPCA procedure; we now expose it here to explicitly show how to formulate it in our abstract

setting and how our results can be interpreted in that interesting light, although their initial formulation was independent of it.

The intimate relationship of the covariance operator with another relevant integral operator is summarized in the next theorem. This property was stated in a similar but more restrictive context (finite dimensional) by Shawe-Taylor et al. (2002, 2005).

Theorem 4.1. *Let $(\mathcal{X}, P)$ be a probability space, $\mathcal{H}$ be a separable Hilbert space, X be a $\mathcal{X}$-valued random variable and Φ be a map from $\mathcal{X}$ to $\mathcal{H}$ such that for all $h \in \mathcal{H}$, $\langle h, \Phi(.)\rangle$ is measurable and $\mathbb{E}\,\|\Phi(X)\|^2 < \infty$. Let C_Φ be the covariance operator associated to $\Phi(X)$ and $K_\Phi : L_2(P) \to L_2(P)$ be the integral operator defined as*

$$(K_\Phi f)(t) = \mathbb{E}\,[f(X)\langle\Phi(X), \Phi(t)\rangle] = \int f(x)\langle\Phi(x), \Phi(t)\rangle dP(x).$$

Then K is a Hilbert-Schmidt, positive self-adjoint operator, and

$$\lambda(K_\Phi) = \lambda(C_\Phi).$$

In particular, K_Φ is a trace-class operator and $\mathrm{tr}(K_\Phi) = \mathbb{E}\,\|\Phi(X)\|^2 = \sum_{i\geq 1}\lambda_i(K_\Phi)$.

This result is proved in the appendix. Note that we actually have $\langle\Phi(x), \Phi(y)\rangle = k(x, y)$, so that K_Φ is really the integral operator with kernel k. We chose the above formulation in the theorem to emphasize that the reproducing property is not essential to the result.

Furthermore, as should appear from the proof, the theorem can be easily extended to find an explicit correspondence between the eigenvectors of C_Φ and the eigenfunctions of K_Φ. This is an essential point for kernel PCA, as it allows to reduce the problem of finding the eigenvectors of the "abstract" operator $C_{1,n}$ to finding the eigenfunctions the kernel integral operator $K_{1,n}$ defined as above, with P taken as the empirical measure; $K_{1,n}$ can then be identified (as in Koltchinskii and Giné, 2000) to the normalized kernel Gram matrix of size $n \times n$, $K_{1,n} \equiv (k(X_i, X_j)/n)_{i,j=1,\ldots,n}$. This comes from the fact that $L_2(P_n)$ is a finite-dimensional space so that any function $f \in L_2(P_n)$ can be identified to the n-uple $(f(X_i))_{i=1,\ldots,n}$; this way the Hilbert structure of $L_2(P_n)$ is isometrically mapped into $\mathbb{R}^n$ embedded with the standard Euclidean norm rescaled by n^{-1}. (Note that this mapping may not be *onto* in the case where two datapoints are identical, but this does not cause a problem.)

A further consequence of Theorem 4.1 and of the above remarks is the following identification of (positive part of) spectra:

$$\begin{aligned}
\lambda(C_1) &= \lambda(K_1)\,;\\
\lambda(C_{1,n}) &= \lambda(K_{1,n})\,;\\
\lambda(C_{2,n}) &= \lambda(K_{2,n})\,;\\
\lambda(C'_2) &= \lambda(K'_2)\,;\\
\lambda(C'_{2,n}) := \lambda(C_{2,n} - C_{1,n}\otimes C^*_{1,n}) &= \lambda\left(\left(I_n - \frac{1}{n}\mathbf{1}\right)K_{2,n}\left(I_n - \frac{1}{n}\mathbf{1}\right)\right) =: \lambda(K'_{2,n}),
\end{aligned}$$

where K_1 denotes the kernel integral operator with kernel k and the true probability distribution P ; $K_{1,n}, K_{2,n}$ are identified to the matrices $(k(X_i, X_j)/n)_{i,j=1,\ldots,n}$,

$(k^2(X_i, X_j)/n)_{i,j=1,\dots,n}$, respectively; I_n denotes the identity matrix of order n; $\mathbf{1}$ denotes the square $n \times n$ matrix whose entries are all ones; and K_2' is the kernel operator with kernel $\tilde{k}_2(x, y) = k^2(x, y) - \mathbb{E}_X[k^2(X, y)] - \mathbb{E}_Y[k^2(x, Y)] + \mathbb{E}_{X,Y}[k^2(X, Y)]$. To understand the two last identities of the above display, first note that $C_2 - C_1 \otimes C_1^* = \mathbb{E}\left[(C_Z - \mathbb{E}[C_Z]) \otimes (C_Z - \mathbb{E}[C_Z])^*\right]$ is the covariance operator for the variable $\bar{C}_Z = C_Z - \mathbb{E}[C_Z]$. The identities follow by (a) applying Theorem 4.1 to this variable (when P is the true distribution or the empirical measure, respectively) and (b) some simple algebra, omitted here, to identify the the corresponding operators (this is similar to Kernel PCA with recentering, see, e.g., Schölkopf et al., 1999).

These identities have two interesting consequences:

- all quantities involving empirical operators appearing in the bounds of Theorems 3.1, 3.2, 3.5 can be computed from the finite-dimensional kernel matrices $K_{1,n}$, $K_{2,n}$. In the last section we had already obtained the expressions for the traces by elementary calculations; further, the above spectra identities allow to identify also partial sums of eigenvalues appearing in the bounds.
- The optimal reconstruction error $R(V_d)$ coincides with the tail sum of eigenvalues $\sum_{i>d} \lambda_i$ of the integral operator K_1, while the empirical construction error $R_n(\widehat{V}_d)$ coincides with the tail sum of eigenvalues of the kernel Gram matrix $K_{1,n}$. Therefore, our results also allow to bound the error made when estimating eigenvalues of K_1 by the eigenvalues of its empirical counterpart $K_{1,n}$. More precisely, minor modifications of the proofs of Theorems 3.1, 3.4, 3.5 result in bounds on the difference between these tail sums: global bound, relative bound and global bound for the recentered operators, respectively. (However, note that Theorem 3.2 has no direct interpretation in this framework: it only focuses on convergence of the reconstruction error.) Similar techniques apply also for dealing with partial sums $\sum_{i \le d} \lambda_i$. Approximating the integral operator K_1 by its empirical counterpart $K_{1,n}$ is known as the *Nyström method* (see, e.g., Williams and Seeger, 2000). We collect the resulting inequalities in the following theorem.

Theorem 4.2. *Assume* **(A1), (A2)** *are satisfied. Let $\mathcal{X}_0$ be the support of distribution P on $\mathcal{X}$; assume $\sup_{x \in \mathcal{X}_0} k(x, x) \le M$ and $\sup_{x,y \in \mathcal{X}_0} \left(k^2(x, x) + k^2(y, y) - 2k^2(x, y)\right) \le L^2$. Denote $R(d) = \sum_{i>d} \lambda_d(K_1)$ and $R_n(d) = \sum_{i>d} \lambda_d(K_{1,n})$. Then for any $n \ge 2$, either of the following inequalitites holds with probability at least $1 - e^{-\xi}$:*

$$R(d) - R_n(d) \le \sqrt{\frac{d}{n} \operatorname{tr} K_2'} + (M \wedge L)\sqrt{\frac{\xi}{2n}}; \tag{37}$$

$$R(d) - R_n(d) \le \sqrt{\frac{d}{n-1} \operatorname{tr} K_{2,n}'} + (M \wedge L)\sqrt{\frac{\xi}{2n}} + L\frac{\sqrt{d}\xi^{\frac{1}{4}}}{n^{\frac{3}{4}}}; \tag{38}$$

$$R(d) - R_n(d) \ge -\sqrt{\frac{2\xi}{n}(M \wedge L)R(d)} - (M \wedge L)\frac{\xi}{3n}; \tag{39}$$

$$R(d) - R_n(d) \ge -\sqrt{\frac{2\xi}{n}(M \wedge L)\left(R_n(d) - (M \wedge L)\frac{\xi}{3n}\right)_+} - (M \wedge L)\frac{\xi}{3n}. \tag{40}$$

Under the stronger condition $k(x, x) = M$ for all $x \in \mathcal{X}_0$, either of the following inequalities holds with probability at least $1 - e^{-\xi}$:

$$R(d) - R_n(d) \le c\left(\sqrt{R(d)\left(\rho(M, d, n) + M\frac{\xi}{n}\right)} + \rho(d, n) + \frac{M\xi}{n}\right); \tag{41}$$

$$R(d) - R_n(d) \le c\left(\sqrt{R_n(d)\left(\rho_n(M,d,n) + M\frac{(\xi+\log n)}{n}\right)} + \rho_n(d,n) + \frac{M(\xi+\log n)}{n}\right). \tag{42}$$

Comments. A consequence of this theorem worth noticing is that by combining (42) and (40) applied to $d, d+1$ respectively (or vice-versa), we obtain a (fully empirical) *relative* bound for estimating *single* eigenvalues. However the relative factor in the main term of the bound is the tail sum of eigenvalues rather than the single eigenvalue itself. Also, similar bounds are available for the partial sums $\sum_{i\le d}\lambda_i$; however in that case the relative bounds lose most of their interest since the "relative" factor appearing in the bound is then typically not close to zero.

Finally, using Theorem 3.5 inequalities similar to (37) and (38) can be proved for bounding the difference between the sum of eigenvalues of the "recentered" integral operator $\bar{K}_1$ with kernel $\bar{k}(x,y) = k(x,y) - \mathbb{E}_X[k(X,y)] - \mathbb{E}_Y[k(x,Y)] + \mathbb{E}_{X,Y}[k(X,Y)]$ and the sum of eigenvalues of the recentered kernel matrix $\bar{K}_{1,n} = (I_n - \frac{1}{n}\mathbf{1})K_{1,n}(I_n - \frac{1}{n}\mathbf{1})$. The principle is exactly similar to the above and we omit the exact statements.

Proof: Bounds (37), (38) are almost direct consequences of Theorem 3.1, and (41), (42) of Theorem 3.2, respectively. More precisely, we know that $R(d) = R(V_d)$ and $R_n(d) = R_n(\widehat{V}_d)$. Theorems 3.1, 3.2 provide upper bounds for $R(\widehat{V}_d) - R_n(\widehat{V}_d)$ (here we need only one-sided bounds, hence the inequalities are valid with slightly higher probability), and we furthemore have $R(V_d) \le R(\widehat{V}_d)$ by definition.

Concerning the "relative" lower bounds (39) and (40), we start with the following fact:

$$R(d) - R_n(d) = R(V_d) - R_n(\widehat{V}_d) \ge R(V_d) - R_n(V_d) = (P_n - P)\langle \Pi_{V_d}, C_Z\rangle;$$

Consider now the function $f : z \to \langle \Pi_{V_d}, C_z\rangle$. Using the same arguments as in the beginning of the proof of Theorem 3.1, we conclude that a.s. $f(Z) \in [a,b]$ for some interval $[a,b]$ with $a \ge 0$ and $|a-b| \le M \wedge L$. We now apply Bernstein's inequality (Theorem B.3) to the function $(f-a) \in [0, M\wedge L]$, obtaining that with probability at least $1 - e^{-\xi}$, we have

$$(P_n - P)\langle \Pi_{V_d}, C_Z\rangle \ge -\sqrt{\frac{2\xi P(f-a)^2}{n}} - (M\wedge L)\frac{\xi}{3n}.$$

Now, note that

$$P(f-a)^2 \le (M\wedge L)(Pf - a) \le (M\wedge L)Pf.$$

This proves (39). Inequality (40) follows by using the fact that $x \ge 0$ and $x^2 + ax + b \ge 0$ with $a \ge 0$ implies $x^2 \ge -b - a\sqrt{-(b\wedge 0)}$ (here applied to $x = \sqrt{R(d)}$ and the corresponding terms coming from (39). □

5 Comparison of the bounds

Of interest is to understand how the different bounds obtained here compare to each other. In this short section we will present two different simplified example benchmark settings where we assume that the true distribution, and in particular the eigenvalues of C_1 and C_2, are known,

and visualize the different bounds. We do not consider here the bound for the recentered case (Theorem 3.5) as it is, up to worse multiplicative constants, essentially equivalent to the non-centered case of Theorem 3.1, as far as the bounding quantity is concerned.

We therefore focus on Theorems 3.1, 3.2, 3.4, more precisely on the excess error inequalities bounding $R(\widehat{V}_d) - R(V_d)$. (Since Theorem 3.2 only deals with this quantity, this is the one we must consider if we want to compare the different theorems.) In general, we expect the following general picture:

(1) The global bound of Theorem 3.1 results in a bound of order $\sqrt{\frac{d}{n}}$.
(2) The excess bound of Theorem 3.2 will result in a bound that decays faster than the global bound as a function of n for fixed d, but has a worse behavior as a function of d for fixed n, because of the factor B_d which will grow rapidly as d increases.
(3) The relative bound of Theorem 3.4 will result in a bound that decays slower than the global bound as a function of n, but we expect a better behavior as a function of d for fixed n, because the risk $R(V_d)$ enters as a factor into the main term of the bound. Actually, we expect that this bound is the only one to be *decreasing* as a function of d, at least for values of d such that the other terms in the bound are not dominant.

Example 1. For this first case we consider a case where the eigenvalues of C_1 and C_2 decay as a power of n. More precisely, suppose that $M = 1$, $R(V_d) = \sum_{i>d} \lambda_i(C_1) = ad^{-\gamma}$ and $\sum_{i>d} \lambda_i(C_2) = a'd^{-\alpha}$ (with $\alpha, \gamma \geq 0$ and $2\gamma \geq \alpha - 1$). In this case, we have $\rho(A, d, n) \lesssim (A^{\frac{-\alpha}{2+\alpha}} d^{\frac{1}{2+\alpha}} n^{-\frac{1+\alpha}{2+\alpha}}) \wedge d^{\frac{1}{2}} n^{-\frac{1}{2}}$, while $B_d = \mathcal{O}(d^{2+\gamma})$.

Example 2. In this case we assume an exponential decay of the eigenvalues: $M = 1$, $R(V_d) = \sum_{i>d} \lambda_i(C_1) = ae^{-\gamma d}$ and $\sum_{i>d} \lambda_i(C_2) = a'e^{-\alpha d}$ (with the same constraints on γ, α as in the first example). In this case, we have $\rho(A, d, n) \lesssim (An^{-1}(1 \vee \log(A^{-1} d^{\frac{1}{2}} n^{\frac{1}{2}}))) \wedge d^{\frac{1}{2}} n^{-\frac{1}{2}}$, while $B_d = \mathcal{O}(e^{\gamma d})$.

We display the (log-)bounds for $R(\widehat{V}_d) - R(V_d)$ for these two examples in Fig. 1, with the choices $\alpha = \gamma = 4$ for example 1, and $\alpha = \gamma = 0.7$ for example 2; we picked $a = 1, a' = 0.5, \xi = 3$, $n \in \{10^7, 10^{10}\}$ for both cases. The bounds are plotted as given in the text including the multiplicative constants; for the relative bound of Theorem 3.4 we strived to pick the best multiplicative constant c that was still compatible with a rigorous mathematical proof. We included in the figure a plot of the (log-)optimal reconstruction error itself $R(V_d)$, which allows to compare the magnitude of the bounds to the magnitude of the target quantity (or, speaking with some abuse, the magnitude of the "bias" and of the bound on the "estimation error").

Note that our goal here is merely to visualize the behavior of the bounds, so that we do not claim that the above choice of parameters correspond to any "realistic" situation (in particular we had to choose a unrealistically high values for n to try to exhibit the trend behavior of the bounds for large n despite the loose multiplicative constants involved). However, the two above general behaviors of the eigenvalues can be exhibited for the Gaussian kernel and some choices of the generating distribution on the real line, as reported for example by Bach and Jordan (2002), so that we trust these examples are somewhat representative.

In both cases we observe, as expected from the above remarks, that the excess bound of Theorem 3.2 gives a much more accurate result when d is small. Quickly however, as d increases, this bound becomes essentially uninformative due to its bad scaling as a function of d, while the relative bound of Theorem 3.4 becomes better. Finally, we can observe a small region on the d-range where the initial global bound is better. This is mainly due

to the worse multiplicative constants arising when applying the localized approach. As n increases, the influence of these constants becomes less important and this region eventually vanishes.

6 Conclusion and discussion

Comparison with previous work. Dauxois and Pousse (1976) studied asymptotic convergence of PCA and proved almost sure convergence in operator norm of the empirical covariance operator to the population one. These results were further extended to PCA in a Hilbert space by Besse (1991). However, no finite sample bounds were presented. Moreover, the centering of the data was not considered.

Compared to the work of Koltchinskii and Giné (2000), we are interested in non-asymptotic (i.e., finite sample sizes) results; furthermore our emphasis is on reconstruction error for PCA while these authors were focusing only on eigenspectra estimation. It is however noteworthy that the recentered fourth moment operator C_2' appearing in our finite sample bounds also surfaces naturally as the covariance operator of the limiting Gaussian process appearing in the central limit theorem proved by the above authors.

Comparing with Shawe-Taylor et al. (2002, 2005), we overcome the difficulties coming from infinite dimensional feature spaces as well as those of dealing with kernel operators (of infinite rank). They also start from results on the operator eigenvalues on a RKHS to conclude on the properties of kernel PCA. Here we used a more direct approach, extended their results to the recentered case and proved refined bounds and possible faster convergence rates for the uncentered case. In particular we show that there is a tight relation between how the (true or empirical) eigenvalues decay and the rate of convergence of the reconstruction error.

Asymptotic vs. non-asymptotic. A point of controversy that might be raised is the following: what is the interest in non-asymptotic bounds if they give informative results only for unreasonably high values of n, as is the case in our examples of Section 5? In this case, why not consider directly the asymptotic results (e.g., central limit theorems) cited above, which surely should be more accurate in the limit? The answer to this is that ideally, our goal would be to understand the behavior of PCA (or of the eigenspectrum of the Gram matrix) for a *fixed* (although, for the time, possibly large) value of n and *across* values of d. This could, for example, help answering the question of how to choose the projection dimension d in a suitable way (we discuss this issue below). As far as we know, central limit theorems, even concerning the eigenspectrum as a whole, are not precise enough to capture this type of behavior. This is illustrated at the very least in the fact that for any value of n, all empirical eigenvalues of rank $d > n$ are zero, which of course is *always* far from the "asymptotic gaussian" behavior given by the CLT. In all honesty, as will appear more clearly below, our bounds are also quite inaccurate for the "very high dimension" regime where d is of same order as n, but might be interesting for intermediate regimes (e.g., d growing as a root power of n). While we are still far from a full understanding of possible regimes across values of (n, d), we hope to have shown that our results present interesting contributions in this general direction.

The nagging problem of the choice of dimension in PCA. Even if we had a full, exact picture of how the estimation error behaves for arbitrary (n, d), the choice of the projection dimension in PCA poses problems of its own. It is tempting to see the reconstruction error $R(V)$

as an objective criterion to minimize and interpret Theorems 3.1 or 3.4 as a classical statistical tradeoff between empirical 'model' error $R_n(\widehat{V}_d)$ (here the 'model' is the set of linear subspaces of dimension d) and estimation error $(R(\widehat{V}_d) - R(V_d))$, for which explicit bounds are provided by the theorems. The sum $S_{n,d}$ of these two contributions is a bound on $R(\widehat{V}_d)$, which would suggest to select the dimension d minimizing $S_{n,d}$ as the best possible guess for the choice of the dimension. However, even if the bound $S_{n,d}$ presents a minimum at a certain $d_0(n)$, this whole view is an illusion: it is clear that, in this case, the *true* reconstruction error $R(\widehat{V}_d)$ of the subspace selected empirically is a decreasing function of d (since $\widehat{V}_d \subset \widehat{V}_{d+1}$). This emphasizes that the (true) reconstruction error is by itself not a good criterion to select the dimension: as far as reconstruction error is concerned, the best choice would be not to project the data at all but to keep the whole space; there is no "overfitting regime" for that matter. This also shows, incidentally, that for $d > d_0(n)$, bounding $R(\widehat{V}_d)$ by $S_{n,d}$ is totally off mark, since $S_{n,d} \geq S_{n,d_0(n)} \geq R(\widehat{V}_{d_0(n)}) \geq R(\widehat{V}_d)$. In other words, for $d > d_0(n)$ the bound fails to capture any information additional to that obtained for $d = d_0(n)$ (this was also noted by Shawe-Taylor et al., 2005).

Hence, an alternative and sensible criterion has to be found to define in a well-founded way what the optimal dimension should be. Up to some point, the nature of the optimal choice depends on what kind of processing is performed next on the data *after* applying PCA. The further processing might suggest its own specific tradeoff between projection dimension (which might result in some complexity penalty) and allowed error. Another, more "agnostic" possibility, is to choose the dimension for which the "approximation error" $R(V_d)$ and the "estimation error" $R(\widehat{V}_d) - R(V_d)$ are approximately of the same order. (We expect in general that the approximation error is dominating for low dimensions, while the converse holds for high dimensions.) If we trust the relative bound of Theorem 3.4, a possible (empirical) criterion would then be to choose d such that $R_n(\widehat{V}_d)$ is of the same order as $\rho_n(M, d, n)$. These different possibilities illustrate at any rate the interest of understanding correctly the behavior of the estimation error across d for a given n.

Finally, additional open problems include obtaining relative convergence rates for the estimation of single eigenvalues, and nonasymptotic bounds for eigenspace estimation.

Appendix A: Additional proofs

A.1 Proof of Theorem 2.1.

For the existence of operator C and its basic properties, see, e.g., Baxendale (1976). We proceed to prove the last part of the theorem. First, we have $\mathbb{E}\|Z \otimes Z^*\| = \mathbb{E}\|Z\|^2 < \infty$, so that $\mathbb{E}[Z \otimes Z^*]$ is well-defined. Now, for any $f, g \in \mathcal{H}$ the following holds by the definition of C and of the expectation operator in a Hilbert space:

$$\langle f, \mathbb{E}\left[Z \otimes Z^*\right] g\rangle = \mathbb{E}\left[\langle Z \otimes Z^*, f \otimes g^*\rangle\right] = \mathbb{E}[\langle Z, f\rangle\langle Z, g\rangle] = \langle f, Cg\rangle\,;$$

this concludes the proof.

A.2 Additional proof for Section 3

A key property necessary for the proof of Theorem 3.2 is established in the following Lemma:

Lemma A.1. *Let and γ_d be defined as in Eq.* (17). *For any $V \in \mathcal{V}_d$, there exists $H_V \in \mathcal{V}_d$ such that*

$$R(H_V) = \min_{H \in \mathcal{V}_d} R(H), \tag{43}$$

and

$$\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \right\rangle^2\right] \leq 2\gamma_d^{-1} \sqrt{\mathbb{E}_{Z,Z'}[\langle Z, Z' \rangle^4]}\, \mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \right\rangle\right],$$

where Z' is an independent copy of Z.

Proof: Recall the following notation: let (λ_i) denote the ordered eigenvalues with multiplicity of C_1, resp. (μ_i) the ordered distinct eigenvalues, and let $\tilde{d}$ be the integer such that $\lambda_d = \mu_{\tilde{d}}$.

Let us denote W_i the eigenspace associated to eigenvalue μ_i and $\bar{W}_j = \bigoplus_{i=1}^{j} W_i$. We first assume $\tilde{d} > 1$ and denote k, ℓ the fixed integers such that $\lambda_{d-\ell} = \mu_{\tilde{d}-1}, \lambda_{d-\ell+1} = \ldots = \lambda_d = \ldots = \lambda_{d+k} = \mu_{\tilde{d}}$ and $\lambda_{d+k+1} = \mu_{\tilde{d}+1}$.

Step 1: construction of H_V. Let $(\phi_1, \ldots, \phi_{d-\ell})$ be an orthonormal basis of $\bar{W}_{\tilde{d}-1}$. Let $V^{(1)}$ denote the orthogonal projection of $\bar{W}_{\tilde{d}-1}$ on V ; in other words, the space spanned by the projections of $(\phi_i)_{i \leq d-\ell}$ on V. The space $V^{(1)}$ is of dimension $d - \ell' \leq d - \ell$; let $(f_1, \ldots, f_{d-\ell'})$ denote an orthonormal basis of $V^{(1)}$. We complete this basis arbitrarily to an orthonormal basis $(f_i)_{i \leq d}$ of V.

Denote now $V^{(2)} = \operatorname{span}\{f_{d-\ell+1}, \ldots, f_d\}$. Note that by construction, $V^{(2)} \perp \bar{W}_{\tilde{d}-1}$. Let $W_{\tilde{d}}^{(2)}$ be the orthogonal projection of $V^{(2)}$ on $W_{\tilde{d}}$. The space $W_{\tilde{d}}^{(2)}$ is of dimension $\ell'' \leq \ell$; let $(\phi_{d-\ell+1}, \ldots, \phi_{d+\ell''-\ell})$ be an orthogonal basis of $W_{\tilde{d}}^{(2)}$. We finally complete this basis arbitrarily to an orthonormal basis $(\phi_i)_{d-\ell+1 \leq i \leq d+k}$ of $W_{\tilde{d}}$. Note that by construction, in particular $V^{(2)} \perp \operatorname{span}\{\phi_{d+1}, \ldots, \phi_{d+k}\}$.

We now define $H_V = \operatorname{span}\{\phi_i, 1 \leq i \leq d\}$. Obviously H_V is a minimizer of the reconstruction error over subspaces of dimension d. We have, using the definition $C_2 = \mathbb{E}[C_Z \otimes C_Z^*]$:

$$\begin{aligned}
\mathbb{E}\left[\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \rangle^2\right] &= \left\langle \Pi_{H_V} - \Pi_V, C_2 \Pi_{H_V} - \Pi_V \right\rangle_{\mathrm{HS}(\mathcal{H})} \\
&\leq \|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}))} \|\Pi_{H_V} - \Pi_V\|^2_{\mathrm{HS}(\mathcal{H})} \\
&= 2\|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}))}(d - \langle \Pi_V, \Pi_{H_V} \rangle_{\mathrm{HS}(\mathcal{H})}) \\
&= 2\,\|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}))} \left(d - \sum_{i,j=1}^{d} \langle f_i, \phi_j \rangle^2 \right);
\end{aligned}$$

and on the other hand, using the definition $C_1 = \mathbb{E}C_Z$:

$$\mathbb{E}\left[\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_Z \rangle\right] = \langle \Pi_{H_V} - \Pi_V, C_1 \rangle = \sum_{i=1}^{d} (\lambda_i - \langle f_i, C_1 f_i \rangle).$$

We will decompose the last sum into two terms, for indices i smaller or greater than $d - \ell$, and bound these separately.

Step 2a: indices $i \leq d-\ell$. In this case we decompose $f_i = \sum_{j\leq d-\ell}\langle f_i, \phi_j\rangle\phi_j + g_i$, with $g_i \in \bar{W}_{\tilde{d}-1}^{\perp}$. We have

$$\langle g_i, C_1 g_i\rangle \leq \mu_{\tilde{d}}\,\|g_i\|^2 = \mu_{\tilde{d}}\left(1 - \sum_{j\leq d-\ell}\langle f_i, \phi_j\rangle^2\right),$$

and

$$\begin{aligned}\sum_{i=1}^{d-\ell}(\lambda_i - \langle f_i, C_1 f_i\rangle) &\geq \sum_{i=1}^{d-\ell}\lambda_i\left(1 - \sum_{j=1}^{d-\ell}\langle f_i, \phi_j\rangle^2\right) - \sum_{i=1}^{d-\ell}\mu_{\tilde{d}}\left(1 - \sum_{j\leq d-\ell}\langle f_i, \phi_j\rangle^2\right)\\ &\geq (\mu_{\tilde{d}-1} - \mu_{\tilde{d}})\left(d - \ell - \sum_{i,j=1}^{d-\ell}\langle f_i, \phi_j\rangle^2\right).\end{aligned}$$

Step 2b: indices $i > d-\ell$. In this case remember that $f_i \perp \phi_j$ for $1 \leq j \leq d-\ell$ and $d+1 \leq j \leq d+k$. We can therefore decompose $f_i = \sum_{j=d-\ell+1}^{d}\langle f_i, \phi_j\rangle\phi_j + g_i'$ with $g_i' \in \bar{W}_{\tilde{d}}^{\perp}$. We have

$$\langle g_i', C_1 g_i'\rangle \leq \mu_{\tilde{d}+1}\|g_i'\|^2 = \mu_{\tilde{d}+1}\left(1 - \sum_{j=d-\ell+1}^{d}\langle f_i, \phi_j\rangle^2\right),$$

and

$$\begin{aligned}\sum_{i=d-\ell+1}^{d}(\lambda_i - \langle f_i, C_1 f_i\rangle) &= \mu_{\tilde{d}}\left(\ell - \sum_{i,j=d-\ell+1}^{d}\langle f_i, \phi_j\rangle^2\right) - \sum_{i=d-\ell+1}^{d}\langle g_i', C_1 g_i'\rangle\\ &\geq (\mu_{\tilde{d}} - \mu_{\tilde{d}+1})\left(\ell - \sum_{i,j=d-\ell+1}^{d}\langle f_i, \phi_j\rangle^2\right).\end{aligned}$$

Finally collecting the results of steps 2a-b we obtain

$$\begin{aligned}\left\langle \Pi_{H_V} - \Pi_V, C_1\right\rangle &\geq \min(\mu_{\tilde{d}-1} - \mu_{\tilde{d}}, \mu_{\tilde{d}} - \mu_{\tilde{d}+1})\left(d - \sum_{i,j=1}^{d-\ell}\langle f_i, \phi_j\rangle^2 - \sum_{i,j=d-\ell+1}^{d}\langle f_i, \phi_j\rangle^2\right)\\ &\geq \min\left(\mu_{\tilde{d}-1} - \mu_{\tilde{d}}, \mu_{\tilde{d}} - \mu_{\tilde{d}+1}\right)(2\,\|C_2\|)^{-1}\,\mathbb{E}\left[\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X\rangle^2\right].\end{aligned}$$

Finally, it holds that

$$\begin{aligned}\|C_2\|^2_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}))} &= \mathbb{E}_{Z,Z'}\left[\langle C_Z \otimes C_Z^*, C_{Z'} \otimes C_{Z'}^*\rangle_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}))}\right]\\ &= \mathbb{E}_{Z,Z'}\left[\langle C_Z, C_{Z'}\rangle^2_{\mathrm{HS}(\mathcal{H})}\right]\\ &= \mathbb{E}_{Z,Z'}\left[\langle Z, Z'\rangle_H^4\right].\end{aligned}$$

This concludes the proof of the Lemma when $\tilde{d} > 1$. If $\tilde{d} = 1$, the proof can be adapted with minor modifications, essentially removing step **(2a)**, so that in the final inequality only the second term of the minimum appears. □

A.3 Proof of Theorem 4.1

It is a well-known fact that an integral kernel operator such as K_ϕ is Hilbert-Schmidt if and only if the kernel $k(x, y)$ (here equal to $\langle \Phi(x), \Phi(y) \rangle$) is an element of $L_2(\mathcal{X} \times \mathcal{X})$ (endowed with the product measure). This is the case here since $k(x, y) \leq \|\Phi(x)\| \, \|\Phi(y)\|$ and $\mathbb{E}\|\Phi(X)\|^2 < \infty$ by assumption. We now characterize this operator more precisely.

Since $\mathbb{E}\|\Phi(X)\| < \infty$, $\Phi(X)$ has an expectation which we denote by $\mathbb{E}[\Phi(X)] \in \mathcal{H}$. Consider the linear operator $T : \mathcal{H} \to L_2(P)$ defined as $(Th)(x) = \langle h, \Phi(x) \rangle_{\mathcal{H}}$. By the Cauchy-Schwarz inequality, $\mathbb{E}\langle h, \Phi(X) \rangle^2 \leq \|h\|^2 \mathbb{E}\|\Phi(X)\|^2$. This shows that T is well-defined and continuous; therefore it has a continuous adjoint T^*. Let $f \in L_2(P)$, then he variable $f(X)\Phi(X) \in \mathcal{H}$ has a well-defined expectation since f and $\|\Phi\|$ are in $L_2(P)$. But for all $g \in \mathcal{H}$, $\langle T^* f, g \rangle_{\mathcal{H}} = \langle f, Tg \rangle_{L_2(P)} = \mathbb{E}[\langle g, f(X)\Phi(X) \rangle_{\mathcal{H}}]$ which shows that $T^*(f) = \mathbb{E}[\Phi(X) f(X)]$.

We now show that $C_\Phi = T^*T$ and $K_\Phi = TT^*$. By definition, for all $h, h' \in \mathcal{H}$, $\langle h, T^*Th' \rangle = \langle Th, Th' \rangle = \mathbb{E}\big[\langle h, \Phi(X) \rangle \langle h', \Phi(X) \rangle\big]$. Thus, by the uniqueness of the covariance operator, we get $C_\Phi = T^*T$. Similarly, $(TT^* f)(x) = \langle T^* f, \Phi(x) \rangle = \mathbb{E}[\langle f(X)\Phi(X), \Phi(x) \rangle] = \int f(y) \langle \Phi(y), \Phi(x) \rangle dP(y)$ so that $K_\Phi = TT^*$. This also implies that K_Φ is self-adjoint and positive.

We finally show that the nonzero eigenvalues of TT^* and T^*T coincide by a standard argument. Let $E_\mu(A) = \{x, Ax = \mu x\}$ be the eigenspace of the operator A associated with μ. Moreover, let $\lambda > 0$ be a positive eigenvalue of $K_\Phi = TT^*$ and f an associated eigenvector. Then $(T^*T)T^* f = T^*(TT^*)f = \lambda T^* f$. This shows that $T^* E_\lambda(TT^*) \subset E_\lambda(T^*T)$; similarly, $T E_\lambda(T^*T) \subset E_\lambda(TT^*)$. Applying T^* to both terms of the last inclusion implies $T^*T E_\lambda(T^*T) = E_\lambda(T^*T) \subset T^* E_\lambda(TT^*)$ (the first equality holds because $\lambda \neq 0$). By the same token, $E_\lambda(TT^*) \subset T E_\lambda(T^*T)$. Thus, $E_\lambda(T^*T) = T^* E_\lambda(TT^*)$ and $E_\lambda(TT^*) = T E_\lambda(T^*T)$; this finally implies $\dim(\mathrm{E}_\lambda(\mathrm{T}^*\mathrm{T})) = \dim(\mathrm{E}_\lambda(\mathrm{TT}^*))$. This shows that λ is also an eigenvalue for C_Φ with the same multiplicity and concludes the proof.

Appendix B: Concentration inequalities

Some concentration inequalities used all along the paper are recalled here for the sake of completeness.

Theorem B.1 (McDiarmid, 1989). *Let $X_1, \ldots, X_n$ be n independent random variables taking values in $\mathcal{X}$ and let $Z = f(X_1, \ldots, X_n)$ where f is such that:*

$$\sup_{x_1, \ldots, x_n, x_i' \in \mathcal{X}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i, \quad \forall\, 1 \leq i \leq n,$$

then

$$P[Z - \mathbb{E}[Z] \geq \xi] \leq e^{-2\xi^2/(c_1^2 + \ldots + c_n^2)},$$

and

$$P[\mathbb{E}[Z] - Z \geq \xi] \leq e^{-2\xi^2/(c_1^2+\ldots+c_n^2)}.$$

Theorem B.2 (Hoeffding, 1963). *Let* $1 \leq r \leq n$ *and* $X_1, \ldots, X_n$ *be n independent random variables. Denote*

$$U = \frac{1}{n(n-1)\ldots(n-r+1)} \sum_{i_1 \neq \ldots \neq i_r} g(X_{i_1}, \ldots, X_{i_r}).$$

If g has range in $[a, b]$ *then*

$$\mathbb{P}[U - \mathbb{E}_U[\geq] t] \leq e^{-2\lceil n/r \rceil t^2/(b-a)^2},$$

and

$$\mathbb{P}[\mathbb{E}_U[-] U \geq t] \leq e^{-2\lceil n/r \rceil t^2/(b-a)^2}.$$

Theorem B.3 (Bernstein's inequality). *Let* f *be a bounded function. With probability at least* $1 - e^{-\xi}$,

$$(P - P_n)(f) \leq \sqrt{\frac{2\xi P f^2}{n}} + \frac{\|f\|_\infty \xi}{3n}, \tag{44}$$

and with probability at least $1 - e^{-\xi}$,

$$(P_n - P)(f) \leq \sqrt{\frac{2\xi P f^2}{n}} + \frac{\|f\|_\infty \xi}{3n}. \tag{45}$$

Acknowledgments This work was supported in part by the PASCAL Network of Excellence (EU # 506778). The authors are extremely grateful to Stéphane Boucheron for invaluable comments and ideas, as well as for motivating this work. The authors wish to thank the anonymous reviewers for many insightful comments leading to many improvements of the paper, in particular the relative bounds.

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics, 34*, 122–148.

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

Bach, F., & Jordan, M. (2002). Kernel independent component analysis. *Journal of Machine Learning Research, 3*, 1–48.

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

Bartlett, P., Jordan, M., & McAuliffe, J. (2003). Convexity, classification, and risk bounds. Technical report, Department of Statistics, U.C. Berkeley, To appear in *J.A.S.A.*

Baxendale, P. (1976). Gaussian measures on function spaces. *American Journal of Mathematics, 98*, 891–952.
Besse, P. (1979). Etude descriptive d'un processus; approximation, interpolation. PhD thesis, Université de Toulouse.
Besse, P. (1991). Approximation spline de l'analyse en composantes principales d'une variable aléatoire hilbertienne. *Annals of Faculty of Science Toulouse (Mathematics), 12(5)*, 329–349.
Bousquet, O. (2002). Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. PhD thesis, Ecole Polytechnique.
Braun, M. (2005). Spectral properties of the kernel matrix and their relation to kernel methods in machine learning. PhD thesis, Friedrich-Wilhelms-Universität Bonn, Available at http://hss.ulb.uni-bonn.de/diss_online/math_nat_fak/2005/braun_mikio.
Dauxois, J., & Pousse, A. (1976). Les analyses factorielles en calcul des probabilités et en statistique: essai d'étude synthétique. PhD thesis, Université de Toulouse.
de la Peña, V. H., & Giné, E. (1999) Decoupling: From dependence to independence. Springer.
Dunford, N., & Schwartz, J. T. (1963). *Linear operators part II: Spectral theory, self adjoint operators in Hilbert space*. Number VII in Pure and Applied Mathematics. New York: John Wiley & Sons.
Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association, 58*, 13–30.
Koltchinskii, V. (2004). Local rademacher complexities and oracle inequalities in risk minimization. Technical report, Department of mathematics and statistics, University of New Mexico.
Koltchinskii, V., & Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli, 6(1)*, 113–167.
Massart, P. (2000). Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse, IX*, 245–303.
Maurer, A. (2004) Concentration of Hilbert-Schmidt operators and applications to feature learning. Manuscript.
McDiarmid, C. (1989). On the method of bounded differences. Surveys in combinatorics (pp. 148–188). Cambridge University Press.
Mendelson, S., & Pajor, A. (2005). Ellipsoid approximation with random vectors. In P. Auer, & R. Meir, (Eds.), *Proceedings of the 18th annual conference on learning theory (COLT 05) of lecture notes in computer science*, vol. 3559 (pp. 429–433). Springer.
Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B, 53(3)*, 539–572.
Schölkopf, B., Smola, A. J., & Müller, K.-R. (1999) Kernel principal component analysis. In B. Schölkopf, C. J. C. Burges, & A. J. Smola, (Eds.), *Advances in kernel methods—Support vector learning* (pp. 327–352). Cambridge, MA: MIT Press. Short version appeared in *Neural Computation, 10*, 1299–1319, 1998.
Shawe-Taylor, J., Williams, C., Cristianini, N., & Kandola, J. (2002). Eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In *Algorithmic Learning Theory: 13th International Conference, ALT 2002 of lecture notes in computer science*, vol. 2533 (pp. 23–40). Springer-Verlag.
Shawe-Taylor, J., Williams, C., Cristianini, N., & Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalisation error of kernel PCA. *IEEE Transactions on Information Theory 51, (7)*, 2510–2522.
Williams, C. K. I., & Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. In P. Langley, editor, Proceedings of the 17th international conference on machine learning (pp. 1159–1166), San Francisco, California: Morgan Kaufmann.
Williamson, R. C., Smola, A. J., & Schölkopf, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory, 47*(6), 2516–2532.