Statistical Properties of Kernel Principal Component Analysis

Gilles Blanchard[1] & Olivier Bousquet & Laurent Zwald[2]

[1] *Fraunhofer First (IDA), Kékuléstr. 7, D-12489 Berlin, Germany.*
[2] *Département de Mathématiques, Université Paris-Sud, Bat.425, F-91405, France.*

Abstract

The main goal of this paper is to prove non-asymptotic inequalities on the reconstruction error for Kernel Principal Component Analysis. Our contribution to this topic is two-fold: (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 show "fast rates" of convergence towards the minimum reconstruction error, more precisely we prove that the convergence rate is related to the decay of eigenvalues and is typically faster than $n^{-1/2}$.

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 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.

1 Introduction

Due to their versatility, kernel methods are currently very popular as data-analysis tools. In such algorithms, the key object is the so-called kernel matrix (the Gram matrix built on the data sample) and it turns out that its spectrum can be related to the performance of the algorithm. This has been shown in particular in the case of Support Vector Machines (Williamson, Shawe-Taylor, Schölkopf, and Smola, 1999). Studying 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.

In the present work we focus on Principal Component Analysis (PCA), and its non-linear variant, kernel-PCA, which are widely used algorithms in data analysis. Their goal is to extract a basis adapted to the data, by looking for directions where the variance is maximized. Their applications are very diverse, ranging from dimensionality reduction to denoising. 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, Smola, and Müller, 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.

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 (Anderson, 1963). For the kernelized version, there is a tight connection between the covariance and the kernel matrix of the data. This is actually at the heart of the kernel-PCA algorithm itself, and also indicates that the properties of the kernel matrix, in particular its spectrum, play a crucial role in the properties of the kernel-PCA algorithm.

Recently Shawe-Taylor, Williams, Cristianini, and Kandola (2002, 2005) have undertaken an investigation 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 their results in several 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 variance precisely into account. We refer the reader to the works of Massart (2000), Bartlett, Bousquet, and Mendelson (2003a), Bartlett, Jordan, and McAuliffe (2003b) (between others). Here we show that this principle leads to improved convergence bounds.

Note that we consider these two types of extension *separately*, not simultaneously. While we believe it possible to combine these two extensions, in the framework of this paper we choose to treat them independently to avoid additional technicalities and leave this issue as an open problem.

To state and prove our results we have chosen to use a functional analysis 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 correponding formalism of Hilbert-Schmidt operators and of random variables in Hilbert spaces. 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.

Finally, let us emphasize some open problems that will be discussed in more detail in the concluding part. We want to underline that in our results we consider the number of components d kept in the PCA procedure (or the number of eigenvalues) as a fixed constant. Our focus here is in the dependence of the bounds in the sample size n. As for the dependence in d for fixed n, unfortunately it is clear that our results do not capture the correct behavior: our bound on the reconstruction error eventually increases as a function of d while it basic considerations show that the true reconstruction error is always decreasing in d. In other words, for fixed n there exists a certain dimension $d(n)$ such that the bound obtained for $d' > d(n)$ is actually *less informative* than the bound obtained for $d(n)$. The same issue surfaces in the work of Shawe-Taylor et al. (2005) and as far as we know, this problem has not been solved. An indirectly linked issue is how define a sensible criterion for what would be an optimal dimension choice in KPCA. Obviously the (true) reconstruction error alone is not enough since it is always a decreasing function of the dimension. We believe these two issues to be the most interesting open problems of this paper.

The paper is organized as follows. Section 2 introduces the necessary background on functional analysis, the basic assumptions and some preliminary fundamental results. Section 3 concentrates on bounding the difference between sums of eigenvalues of the kernel matrix and of the associated kernel operator. Finally, Section 4 gives our main results, bounds on the reconstruction error of kernel-PCA. We conclude with an extended discussion on the open issues sketched above.

2 Preliminaries

The core of our results is concerned with estimating eigenvalues of certain operators in a reproducing Hilbert kernel space $\mathcal{H}_k$. The most convenient way to deal with these objects is to use formalism from functional analysis, and in particular to introduce the space of Hilbert-Schmidt operators on $\mathcal{H}_k$ endowed with a suitable Hilbert structure. The present section is devoted to introducing the necessary notation and base properties that will be used repeatedly.

2.1 The Hilbert space of Hilbert-Schmidt operators

Let $\mathcal{H}$ be a separable Hilbert space. A linear operator L from $\mathcal{H}$ to $\mathcal{H}$ is called Hilbert-Schmidt if $\sum_{i\geq 1} \|Le_i\|^2_{\mathcal{H}} = \sum_{i,j\geq 1} \langle Le_i, e_j\rangle^2 < \infty$, 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 non-negative 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, this series is independent of the chosen orthonormal basis and is called the trace of L, denoted by $\operatorname{tr} L$. By Lidskii's theorem $\operatorname{tr} L = \sum_{i\geq 1} \lambda_i(L)$ 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:

$$\|f \otimes g^*\|_{\mathrm{HS}(\mathcal{H})} = \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}} \; ; \tag{1}$$

$$\operatorname{tr} f \otimes g^* = \langle f, g \rangle_{\mathcal{H}} \; ; \tag{2}$$

$$\langle f \otimes g^*, A \rangle_{\mathrm{HS}(\mathcal{H})} = \langle Ag, f \rangle_{\mathcal{H}} \quad \text{for any } A \in \mathrm{HS}(\mathcal{H}) \, . \tag{3}$$

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

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

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

$$\|U\|_{\mathrm{HS}(\mathcal{H})} = \sqrt{d} \, , \tag{4}$$

$$\operatorname{tr} U = d \, . \tag{5}$$

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 such that $\operatorname{range} \Pi_V = V$ and $\ker \Pi_V = V^\perp$. 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,

$$\langle \Pi_{V^\perp}, A \rangle := \operatorname{tr} A - \langle \Pi_V, A \rangle \; . \tag{6}$$

Eigenvalues formulas. We denote by $\mathcal{V}_d$ the set of subspaces of dimension d of $\mathcal{H}$. The following theorem sums up important formulas concerning the individual or summed eigenvalues of self-adjoint compact operators; the first one is due to Fan (see Torki, 1997, for a proof) while the second is the so-called Courant-Fischer-Weyl formula (see e.g. Dunford and Schwartz, 1963).

Theorem 2.1. *Let C a compact self-adjoint operator on $\mathcal{H}$, then for all $d \geq 0$,*

$$\sum_{i=1}^{d} \lambda_i(C) = \max_{V \in \mathcal{V}_d} \langle \Pi_V, C \rangle_{\mathrm{HS}(\mathcal{H})} \,, \tag{7}$$

$$\text{and} \qquad \lambda_{d+1}(C) = \min_{V \in \mathcal{V}_d} \max_{f \perp V} \frac{\langle f, Cf \rangle}{\|f\|^2} \,, \tag{8}$$

where in both cases, the optimum is attained when V is the span of the first d eigenvectors of C.

2.2 Second order integral operators

We recall basic facts about random elements in Hilbert spaces. A random element Z in a separable Hilbert space has an expectation $e \in \mathcal{H}$ when $\mathbb{E}\|Z\| < \infty$ and e is 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 (noncentered) covariance operator through this theorem and definition:

Theorem 2.2. *If $\mathbb{E}\|Z\|^2 < \infty$, there exists a unique operator $C : \mathcal{H} \to \mathcal{H}$ such that*

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

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 *noncentered covariance* operator of Z.

The core property of covariance operators that we will use is its intimate relationship with another integral operator summarized in the next theorem. This property was first used in a similar but more restrictive context (finite dimensional) by Shawe-Taylor et al. (2002, 2005).

Theorem 2.3. *Let $(\mathcal{X}, P)$ be a probability space, $\mathcal{H}$ be a separable Hilbert space 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 $\operatorname{tr}(K_\Phi) = \mathbb{E}\|\Phi(X)\|^2 = \sum_{i \geq 1} \lambda_i(K_\Phi)$.

If we denote $\langle \Phi(x), \Phi(y) \rangle = k(x, y)$, then K_Φ is called the integral operator with kernel k.

2.3 Main framework and assumptions

Let $\mathcal{X}$ denote the input space (an arbitrary measurable space) and P denote a distribution on $\mathcal{X}$ according to which the data is sampled i.i.d. We will denote by P_n the empirical measure associated to a sample $X_1, \ldots, X_n$ from P, i.e. $P_n = \frac{1}{n} \sum \delta_{X_i}$. With some abuse of notation, for a function $f : \mathcal{X} \to \mathbb{R}$, we may use the notation $Pf := \mathbb{E}[f(X)]$ and $P_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.

Let k be a positive definite function on $\mathcal{X}$ and $\mathcal{H}_k$ the associated reproducing kernel Hilbert space (RKHS for short in the sequel). We recall 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:

$$\forall f \in \mathcal{H}_k \qquad \forall x \in \mathcal{X} \qquad \langle f, k(x, .) \rangle = f(x) \,, \tag{9}$$

and in particular

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

Finally, let $\mathcal{V}_d$ denote the set of all vector subspaces of dimension d of $\mathcal{H}_k$.

We will always work with the following assumptions which we will refer collectively to as "assumption **(A)**" in the sequel:

(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.*

Note that assumptions **(A1)**-**(A2)** ensure the measurability of all functions in $\mathcal{H}_k$ since they are obtained by linear combinations and pointwise limits of functions $k(x, \cdot)$.

Notation for the noncentered case. For $x \in \mathcal{X}$, we denote

$$\begin{aligned} \varphi_x &= k(x, .) \in \mathcal{H} \,, \\ C_x &= \varphi_x \otimes \varphi_x^* \in \mathrm{HS}(\mathcal{H}) \,. \end{aligned}$$

The following properties are then straightforward from the preceding sections:

$$\operatorname{tr} C_x = \|C_x\|_{\mathrm{HS}(\mathcal{H}_k)} = k(x, x) \,, \tag{10}$$

$$\langle C_x, C_y \rangle_{\mathrm{HS}(\mathcal{H})} = k^2(x, y) \,, \tag{11}$$

$$\langle f, C_x g \rangle_{\mathcal{H}} = \langle C_x, f \otimes g^* \rangle_{\mathrm{HS}(\mathcal{H})} = f(x) g(x) \,, \tag{12}$$

and for any orthogonal projector U,

$$\langle U, C_x \rangle_{\mathrm{HS}(\mathcal{H}_k)} = \|U\varphi_x\|^2_{\mathcal{H}_k} . \tag{13}$$

Note incidentally that (11) implies that $\mathrm{HS}(\mathcal{H})$ is actually a natural representation of the RKHS with kernel $k^2(x,y)$. Namely to an operator $A \in \mathrm{HS}(\mathcal{H})$ we can associate the function

$$f_A(x) = \langle A, C_x \rangle_{\mathrm{HS}(\mathcal{H})} = \langle A\varphi_x, \varphi_x \rangle_{\mathcal{H}} = (A\varphi_x)(x) ;$$

with this notation, we have $f_{C_x} = k^2(x,\cdot)$, and one can check that (9) is satisfied in $\mathrm{HS}(\mathcal{H})$ with the kernel $k^2(x,y)$ when identifying an operator to its associated function. Also, paralleling the earlier remark about measurability of functions in $\mathcal{H}_k$, assumptions **(A1)-(A2)** ensure that f_A is measurable for any A.

Now, let us denote $C_1 : \mathcal{H}_k \to \mathcal{H}_k$, resp. $C_2 : \mathrm{HS}(\mathcal{H}_k) \to \mathrm{HS}(\mathcal{H}_k)$, the noncentered covariance operator associated to the random element φ_X in $\mathcal{H}_k$, resp. C_X in $\mathrm{HS}(\mathcal{H}_k)$; and $K_1, K_2 : L_2(P) \to L_2(P)$ the integral operators with kernel $k(x,y)$, resp. $k^2(x,y)$ (Note that all these operators are well-defined due to assumption **(A)**). We then have the following property:

Lemma 2.4. *Under assumption* **(A)** *the operators* C_1, C_2, K_1, K_2 *defined above satisfy the following :*

(i) C_1 *is the expectation in* $\mathrm{HS}(\mathcal{H}_k)$ *of* $C_X = \varphi_X \otimes \varphi_X^*$.

(ii) C_2 *is the expectation in* $\mathrm{HS}(\mathrm{HS}(\mathcal{H}_k))$ *of* $C_X \otimes C_X^*$.

(iii) $\lambda(C_1) = \lambda(K_1)$, *and* $\operatorname{tr} C_1 = \operatorname{tr} K_1 = \mathbb{E}\left[k(X,X)\right]$.

(iv) $\lambda(C_2) = \lambda(K_2)$, *and* $\operatorname{tr} C_1 = \operatorname{tr} K_2 = \mathbb{E}\left[k^2(X,X)\right]$.

This Lemma is a direct consequence of Theorems 2.2 and 2.3. (noting that the measurability conditions have been established in the preceding discussions).

Notation for the recentered case. We will be interested in the sequel in recentered versions of the above quantities (which appear for standard covariance operators and PCA techniques), which we now define accordingly. Let us define for all $x \in \mathcal{X}$

$$\begin{aligned} \mu &= \mathbb{E}\left[\varphi_X\right] , \\ \overline{\varphi}_x &= \varphi_x - \mu \in \mathcal{H} , \\ \overline{C}_x &= \overline{\varphi}_x \otimes \overline{\varphi}_x^* \in \mathrm{HS}(\mathcal{H}) ; \end{aligned}$$

we then have $\|\mu\|^2 = \mathbb{E}k(X,X')$ and

$$\operatorname{tr} \overline{C}_x = \left\|\overline{C}_x\right\|_{\mathrm{HS}(\mathcal{H}_k)} = \|\varphi(x) - \mu\|^2 = k(x,x) + \mathbb{E}k(X,X') - 2\mathbb{E}k(X,x) ,$$

where X' denotes an independent copy of X.

Similarly, let us denote $\overline{C}_1$ the covariance operator associated to $\overline{\varphi}_X$, and $\overline{K}_1$ the integral operator with kernel $\overline{k}(x,y) = \langle \varphi_x - \mu, \varphi_y - \mu \rangle = k(x,y) - \mathbb{E}k(X,x) - \mathbb{E}k(X,y) + \mathbb{E}k(X,X')$; then the following holds:

Lemma 2.5. *Under assumption* **(A)** *the operators $\overline{C}_1, \overline{K}_1$ defined above satisfy the following :*

(i) $\overline{C}_1$ *is the expectation in* $\mathrm{HS}(\mathcal{H}_k)$ *of* $\overline{C}_X = \overline{\varphi}_X \otimes \overline{\varphi}_X^*$ *; moreover one has*

$$\overline{C}_1 = C_1 - \mu \otimes \mu^* \tag{14}$$

(ii) $\lambda(\overline{C}_1) = \lambda(\overline{K}_1)$, *and* $\operatorname{tr} \overline{C}_1 = \operatorname{tr} \overline{K}_1 = \mathbb{E}[k(X,X)] - \mathbb{E}[k(X,X')]$.

Again, this lemma is a direct consequence of Theorems 2.2 and 2.3 and of straightforward computations.

Notations for the empirical case. In the following we will study empirical counterparts of the above quantities. The generality of the above results implies that we can replace the distribution P by the empirical measure P_n associated to an i.i.d. sample $X_1, \ldots, X_n$ without any changes and we merely need to introduce adequate notation.

In the noncentered case, the corresponding operators are denoted by $K_{1,n}$ and $C_{1,n}$; we define $K_{2,n}$ and $C_{2,n}$ similarly. In particular, Lemma 2.4 implies that $\lambda(K_{1,n}) = \lambda(C_{1,n}), \lambda(K_{2,n}) = \lambda(C_{2,n})$, $\operatorname{tr} K_{1,n} = \operatorname{tr} C_{1,n} = \frac{1}{n}\sum_{i=1}^n k(X_i, X_i)$, and $\operatorname{tr} K_{2,n} = \operatorname{tr} C_{2,n} = \frac{1}{n}\sum_{i=1}^n k^2(X_i, X_i)$.

Note that $C_{1,n}$ is the empirical covariance operator, i.e. $\langle f, C_{1,n} g \rangle = \frac{1}{n}\sum_{i=1}^n f(X_i) g(X_i)$. An important point is that $K_{1,n}$ can be identified (as in Koltchinskii and Giné, 2000) with the normalized kernel 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 Euclidian 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).

For the centered case, note that the quantities $\overline{\varphi}_x, \overline{C}_x$ already depend on P through the centering, so that we will define the corresponding quantities for P_n with an index n:

$$\begin{aligned}
\overline{\varphi}_{x,n} &= \varphi_x - \frac{1}{n}\sum_{i=1}^n \varphi_{X_i} , \\
\overline{C}_{x,n} &= \overline{\varphi}_{x,n} \otimes \overline{\varphi}_{x,n}^* , \\
\overline{C}_{1,n} &= \frac{1}{n}\sum_{i=1}^n \overline{\varphi}_{X_i,n} \otimes \overline{\varphi}_{X_i,n}^* = \frac{1}{n}\sum_{i=1}^n \overline{C}_{X_i,n} .
\end{aligned}$$

The associated centered kernel operator is denoted $\overline{K}_{1,n}$ and identified with the following centered kernel matrix :

$$\overline{K}_{1,n} \equiv \left(\left\langle \overline{\varphi}_{X_i}, \overline{\varphi}_{X_j} \right\rangle_{\mathcal{H}_k} \right)_{1 \leq i,j \leq n} = \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) K_{1,n} \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right) .$$

where $\mathbf{1}_n = (1, \cdots, 1)' \in \mathbb{R}^n$. As a consequence of Lemma 2.5, we have $\lambda(\overline{C}_{1,n}) = \lambda(\overline{K}_{1,n})$. Finally, note that $\overline{C}_{1,n}$ is a biased estimator of $\overline{C}_1$, so we will additionally introduce

$$\widetilde{C}_{1,n} = \frac{n}{n-1} \overline{C}_{1,n} = C_{1,n} - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi_{X_j}^* , \tag{15}$$

which satisfies $\mathbb{E}\left[\widetilde{C}_{1,n} \right] = \overline{C}_1$.

3 General Results on Eigenvalues of Gram Matrices

We are now able to proceed to our first goal, the estimation of sums of eigenvalues of kernel operators K_1 or $\overline{K}_1$ from eigenvalues of their empirical counterparts $K_{1,n}$ and $\overline{K}_{1,n}$. For this, we will make use of the preliminary results to relate these sums of eigenvalues to an empirical process on classes of functions of type $x \mapsto \langle \Pi_V, C_x \rangle$. In turn, this will allow us to introduce classical tools of empirical process theory to obtain our results.

Let us formulate precisely this stepping stone in the case of K_1 through the following corollary:

Corollary 3.1. *Under Assumption* **(A)**, *we have*

$$\sum_{k=1}^{d} \lambda_k(K_1) = \max_{V \in \mathcal{V}_d} \mathbb{E}\left[\langle \Pi_V, C_X \rangle \right] , \tag{16}$$

$$\sum_{k \geq d+1} \lambda_k(K_1) = \min_{V \in \mathcal{V}_d} \mathbb{E}\left[\langle \Pi_{V^\perp}, C_X \rangle \right] . \tag{17}$$

The first equality in this corollary is an immediate consequence of (7) and assertions (i), (iii) of Lemma 2.4. The second is a consequence of the first, of definition (6) and of the definition of the trace. Of course, corresponding results for the centered and empirical versions hold as well in a parallel fashion.

3.1 Noncentered Case

In this section we consider the easier case of the noncentered kernel operator.

3.1.1 Global approach

The first result consists in data-dependent upper and lower bounds for the sum of the d largest or smallest eigenvalues of the integral operator. It is essentially the same as the result obtained by Shawe-Taylor et al. (2005), but we give a proof for completeness and to show how it fits in our framework.

Theorem 3.2 (Shawe-Taylor et al.). *Under Assumption* **(A)**, *with probability at least* $1 - 3e^{-\xi}$,

$$-M\sqrt{\frac{\xi}{2n}} \leq \sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) \leq 2\sqrt{\frac{d}{n}\operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}}. \tag{18}$$

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

$$-M\sqrt{\frac{\xi}{2n}} \leq \sum_{i\geq d+1} \lambda_i(K_1) - \sum_{i\geq d+1} \lambda_i(K_{1,n}) \leq 2\sqrt{\frac{d}{n}\operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}}. \tag{19}$$

Proof. We start with the first statement. From equation (16) and its counterpart for $K_{1,n}$ we have

$$\sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) = \max_{V\in\mathcal{V}_d} P_n \left\langle \Pi_V, C_X \right\rangle - \max_{V\in\mathcal{V}_d} P \left\langle \Pi_V, C_X \right\rangle .$$

This gives, denoting by V_d the subspace attaining the second maximum,

$$(P_n - P) \left\langle \Pi_{V_d}, C_X \right\rangle \leq \sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) \leq \sup_{V\in\mathcal{V}_d} (P_n - P) \left\langle \Pi_V, C_X \right\rangle .$$

The lower bound above leads to the lower bound of the theorem by an application of Hoeffding's inequality for the empirical mean of an i.i.d. sample of the bounded random variable $\left\langle \Pi_{V_d}, C_X \right\rangle$, namely

$$0 \leq \left\langle \Pi_{V_d}, C_X \right\rangle = \left\langle \Pi_{V_d}, \varphi_X \otimes \varphi_X^* \right\rangle = \left\| \Pi_{V_d}(\varphi_X) \right\|^2 \leq \left\| \varphi_X \right\|^2 \leq M , \tag{20}$$

where we have used the fact that Π_{V_d} is a projector.

For the upper bound, we use standard techniques of concentration and symmetrization. Since $\left\langle \Pi_{V_d}, C_X \right\rangle \in [0, M]$, we can apply the bounded difference concentration inequality (also known as McDiarmid's or Azuma's inequality) to the variable $\sup_{V\in\mathcal{V}_d}(P_n - P) \left\langle \Pi_V, C_X \right\rangle$. Thus with probability $1 - e^{-\xi}$,

$$\sup_{V\in\mathcal{V}_d} (P_n - P) \left\langle \Pi_V, C_X \right\rangle \leq \mathbb{E}\left[\sup_{V\in\mathcal{V}_d} (P_n - P) \left\langle \Pi_V, C_X \right\rangle \right] + M\sqrt{\frac{\xi}{2n}}.$$

By a standard symmetrization argument,

$$\mathbb{E}\left[\sup_{V\in\mathcal{V}_d}(P_n-P)\langle\Pi_V,C_X\rangle\right]\leq 2\mathbb{E}\mathbb{E}_\varepsilon\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_V,C_{X_j}\right\rangle\right],$$

where $(\varepsilon_i)_{i=1,\ldots,n}$ is an i.i.d. family of Rademacher variables. We can apply the bounded difference inequality a second time to this quantity, so that with probability $1-e^{-\xi}$:

$$\mathbb{E}\mathbb{E}_\varepsilon\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_V,C_{X_j}\right\rangle\right]\leq\mathbb{E}_\varepsilon\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_V,C_{X_j}\right\rangle\right]+M\sqrt{\frac{\xi}{2n}}.$$

The expectation on the right-hand-side is then bounded by an application of Lemma 3.3 below, leading to the conclusion.

The second inequality of the Theorem follows from similar arguments. Equation (17) leads to

$$\sum_{i>d}\lambda_i(K_1)-\sum_{i>d}\lambda_i(K_{1,n})=\min_{V\in\mathcal{V}_d}P\left\langle\Pi_{V^\perp},C_X\right\rangle-\min_{V\in\mathcal{V}_d}P_n\left\langle\Pi_{V^\perp},C_X\right\rangle.$$

and thus, denoting by $\widetilde{V}_d$ the subspace attaining the first minimum,

$$(P-P_n)\left\langle\Pi_{\widetilde{V}_d^\perp},C_X\right\rangle\leq\sum_{i>d}\lambda_i(K_1)-\sum_{i>d}\lambda_i(K_{1,n})\leq\sup_{V\in\mathcal{V}_d}(P-P_n)\left\langle\Pi_{V^\perp},C_X\right\rangle.$$

The rest of the proof parallels exactly the proof of the first part. □

We have used the following Lemma in the completion of the proof:

Lemma 3.3.

$$\mathbb{E}_\varepsilon\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_{V^\perp},C_{X_j}\right\rangle\right]=\mathbb{E}_\varepsilon\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_V,C_{X_j}\right\rangle\right]\leq\sqrt{\frac{d}{n}\operatorname{tr}K_{2,n}}\,.$$

and

$$\mathbb{E}\mathbb{E}_\varepsilon\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_{V^\perp},C_{X_j}\right\rangle\right]=\mathbb{E}\mathbb{E}_\varepsilon\left[\frac{1}{n}\sup_{V\in\mathcal{V}_d}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_V,C_{X_j}\right\rangle\right]\leq\sqrt{\frac{d}{n}\operatorname{tr}K_2}$$

Proof. First note that for the two statements, the first equality is straightforward from the definition and the symmetry of Rademacher variables. We then have

$$\begin{aligned}\sum_{j=1}^n\varepsilon_j\left\langle\Pi_V,C_{X_j}\right\rangle&=\left\langle\Pi_V,\sum_{j=1}^n\varepsilon_j\varphi_{X_j}\otimes\varphi_{X_j}^*\right\rangle_{\mathrm{HS}(\mathcal{H}_k)}\\&\leq\sqrt{d}\left\|\sum_{j=1}^n\varepsilon_j\varphi_{X_j}\otimes\varphi_{X_j}^*\right\|_{\mathrm{HS}(\mathcal{H})}=\sqrt{d\sum_{i,j=1}^d\varepsilon_i\varepsilon_jk^2(X_i,X_j)}\,,\end{aligned}$$

where the inequality is Cauchy-Schwarz's. Finally, by Jensen's inequality,

$$\mathbb{E}_\varepsilon \left[\frac{1}{n} \sup_{V \in \mathcal{V}_d} \sum_{j=1}^{n} \varepsilon_j \left\langle \Pi_{V^\perp}, C_{X_j} \right\rangle \right] \le \sqrt{\frac{d}{n}} \sqrt{\frac{\sum_{i=1}^{n} k^2(X_i, X_i)}{n}} .$$

This concludes the proof of the first statement. The second is obtained by a second application of Jensen's inequality. □

Remark. Notice that the upper and lower bounds in Theorem 3.2 are of a different nature. One way to explain this is to consider directly the expectation of the involved quantities: one has

$$0 \le \mathbb{E}\left[\sum_{i=1}^{d} \lambda_i(K_{1,n})\right] - \sum_{i=1}^{d} \lambda_i(K_1) \le \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} (P_n - P) \langle \Pi_V, C_X \rangle\right] \le 2\sqrt{\frac{d}{n} \operatorname{tr} K_2} ,$$

where the lower bound is a consequence of (16) and Jensen's inequality, and the upper bound follows from arguments similar to the above proof.

We see that the empirical eigenvalues are biased estimators of the population ones (although the above inequality only provides an upper bound on the bias); therefore the difference between upper and lower bound in (18) is to be interpreted as bias rather than estimation error. If we additionally apply McDiarmid's inequality twice to the above bound, on the one hand to the quantity $\sum_{i=1}^{d} \lambda_i(K_{1,n})$, and on the other hand to $\operatorname{tr} K_{2,n}$, then we are lead precisely to (18). This approach was followed by Shawe-Taylor et al. (2002, 2005). We have used the same arguments in the proof of Theorem 3.2, but in a different order, as this allows for further refinement (see next section).

3.1.2 Relative bounds

We now use recent work based on Talagrand's inequality (see e.g. Massart, 2000; Bartlett et al., 2003a) to obtain improved concentration for the large eigenvalues of the Gram matrix. We obtain a better rate of convergence, but at the price of comparing the sums of eigenvalues up to a constant factor.

Theorem 3.4. *Under Assumption* **(A)**, *for all $K > 1$ and $\xi > 0$, with probability at least $1 - e^{-\xi}$, the following holds:*

$$\sum_{k=1}^{d} \lambda_k(K_{1,n}) - \frac{K+1}{K} \sum_{k=1}^{d} \lambda_k(K_1) \\ \le 6K \inf_{h \ge 0} \left\{ \frac{Mh}{n} + 2\sqrt{\frac{d}{n} \sum_{j \ge h+1} \lambda_j(K_2)} \right\} + \frac{M\xi(11 + 5K)}{n} . \quad (21)$$

Also, for all $K > 1$ and $\xi > 0$, with probability at least $1 - 3e^{-\xi}$, we have

$$\sum_{k=1}^{d} \lambda_k(K_{1,n}) - \frac{K+1}{K} \sum_{k=1}^{d} \lambda_k(K_1) \leq 282K \inf_{h \geq 0} \left\{ \frac{2hM}{n} + \sqrt{2} \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_{2,n})} \right\} + \frac{2620MK\xi}{n}. \quad (22)$$

Moreover, with probability at least $1 - e^{-\xi}$, for all $K > 1$,

$$\sum_{k=1}^{d} \lambda_k(K_{1,n}) - \frac{K-1}{K} \sum_{k=1}^{d} \lambda_k(K_1) \geq -\frac{M\xi}{n} \left(\frac{1}{3} + \frac{K}{2} \right). \quad (23)$$

The proof of the Theorem is found in Appendix A.2, using a fundamental deviation inequality recalled in Appendix B and additional auxiliary results in Appendix C.

Comments. A superficial look at this result could lead to conclude that it is of the same form as Theorem 2 of Shawe-Taylor et al. (2005) where an infimum operator also appears in the bound. However, the bounds are really of a different nature. In the latter reference the infimum operation comes from the observation that since obviously the partial sum S_d of the first d eigenvalues is increasing in d, we can lower bound S_d by $S_{d'}$ with $d' < d$; hence the empirical lower bound for $S_{d'}$ is *a fortiori* a lower bound for S_d. We could naturally also take advantage of this observation and introduce in the lower bound an additional maximum operation over $d' < d$ but opted against it for readability.

To illustrate the novelty introduced by our result, first notice that if we disregard the multiplicative constants, the complexity term obtained here is always better (or equal) in order than the one of (18) (take $h = 0$). As an example of how this bound differs from (18), assume that $\lambda_j(K_2) = O(j^{-\alpha})$ with $\alpha > 1$, then (18) gives a bound of order $\sqrt{d/n}$, while Theorem 3.4 gives a bound of order $d^{1/(1+\alpha)} n^{-\alpha/(1+\alpha)}$ – hence a better rate. In the case of an exponential decay ($\lambda_j(K_2) = O(e^{-\gamma j})$ with $\gamma > 0$), the rate even drops to $\log(nd)/n$. If K_2 has a finite number k of non-zeros eigenvalues, the bound is of order $\frac{k}{n}$. Of course this improvement comes at the cost of an additional factor in front of the empirical sum, hence this bound is better understood as a *relative* performance bound.

Finally, Theorem 3.4 only covers the case of the sum of the *bigger* eigenvalues. Unfortunately, unlike in the global case, we were not able to use an identical reasoning for the smallest eigenvalues. It is actually possible to derive a result of a similar form, but with worse constants, as a consequence of our results for the generalization of kernel PCA. For this reason we postpone the statement of this result to section 4.

3.2 Recentered Case

In the following result, we extend Theorem 3.2 to a more general case where the data is first recentered. Let us begin with a control of a suprema of random variables:

Theorem 3.5. *Under Assumption* (**A**)*, with probability at least* $1 - 3e^{-\xi}$*,*

$$\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \widetilde{C}_{1,n} \right\rangle - \left\langle \Pi_V, \overline{C}_1 \right\rangle \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + M\left(5\sqrt{\frac{\xi}{n}} + \frac{4}{\sqrt{n}} + \frac{6}{n-1}\right) ;$$

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

$$\sup_{V \in \mathcal{V}_d} \left\langle \Pi_{V^\perp}, \overline{C}_1 \right\rangle - \left\langle \Pi_{V^\perp}, \widetilde{C}_{1,n} \right\rangle \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + M\left(5\sqrt{\frac{\xi}{n}} + \frac{4}{\sqrt{n}} + \frac{6}{n-1}\right) ;$$

The proof of the theorem is relegated to Appendix A.2. It follows the same principles as for Theorem 3.2, but some additional steps are needed to deal with a U-process arising because of the recentering. From this theorem we deduce the following upper bounds:

Theorem 3.6. *Under Assumption* (**A**)*, for all* $\xi > 1$*, with probability greater than* $1 - 3e^{-\xi}$*,*

$$-2M\sqrt{\frac{\xi}{n}} \leq \frac{n}{n-1}\sum_{i=1}^{d} \lambda_i(\overline{K}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{K}_1) \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 18M\sqrt{\frac{\xi}{n}};$$

and with probability greater than $1 - 3e^{-\xi}$*,*

$$-2M\sqrt{\frac{\xi}{n}} \leq \sum_{i \geq d+1} \lambda_i(\overline{K}_1) - \frac{n}{n-1}\sum_{i \geq d+1} \lambda_i(\overline{K}_{1,n}) \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 18M\sqrt{\frac{\xi}{n}}.$$

Proof. (Majoration) Theorem 2.1 entails

$$\frac{n}{n-1}\sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1) \leq \sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \widetilde{C}_{1,n} \right\rangle - \left\langle \Pi_V, \overline{C}_1 \right\rangle ,$$

and

$$\sum_{i \geq d+1} \lambda_i(\overline{C}_1) - \frac{n}{n-1}\sum_{i \geq d+1} \lambda_i(\overline{C}_{1,n}) \leq \sup_{V \in \mathcal{V}_d} \left\langle \Pi_{V^\perp}, \overline{C}_1 \right\rangle - \left\langle \Pi_{V^\perp}, \widetilde{C}_{1,n} \right\rangle .$$

Theorem 3.5 and Lemma 2.4 allow to conclude.

The minoration part follows from Hoeffding's inequality for U-statistics; details can be found in the Appendix. □

4 Application to Kernel-PCA

4.1 Uncentered Case

We first consider in this section the simpler case of "uncentered Kernel-PCA" where the goal is to reconstruct the signal using principal directions of the noncentered covariance operator.

Remember we assume that the number d of KPCA 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 with the RKHS norm) of approximating the data by their projections. The space $\widehat{V}_d$ minimizing the empirical error is given by

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

$\widehat{V}_d$ is the vector space spanned by the first d eigenfunctions of $C_{1,n}$. Analogously, we denote by V_d the space spanned by the first d eigenfunctions of C_1. We will adopt the following notation for the true and empirical *reconstruction error*:

$$R_n(V) = \frac{1}{n} \sum_{j=1}^{n} \|\varphi_{X_j} - \Pi_V(\varphi_{X_j})\|^2 = P_n \left\langle \Pi_{V^\perp}, C_X \right\rangle .$$

$$R(V) = \mathbb{E}\left[\|\varphi_X - \Pi_V \varphi_X\|^2\right] = P \left\langle \Pi_{V^\perp}, C_X \right\rangle .$$

One has $R_n(\widehat{V}_d) = \sum_{i>d} \lambda_i(K_{1,n})$ and $R(V_d) = \sum_{i>d} \lambda_i(K_1)$.

4.1.1 Bound on the Reconstruction Error: global approach

We give a data dependent bound for the reconstruction error which is a simple consequence of 3.2.

Theorem 4.1. *Under Assumption* **(A)**, *with probability at least* $1 - 2e^{-\xi}$,

$$R(\widehat{V}_d) \leq \sum_{i=d+1}^{n} \lambda_i(K_{1,n}) + 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}} .$$

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

$$R(\widehat{V}_d) - R(V_d) \leq 2\sqrt{\frac{d}{n} \operatorname{tr} K_2} + 2M\sqrt{\frac{\xi}{2n}} .$$

Proof. We have

$$R(\widehat{V}_d) - R_n(\widehat{V}_d) = (P - P_n) \left\langle \Pi_{\widehat{V}_d^\perp}, C_X \right\rangle \leq \sup_{V \in \mathcal{V}_d} (P - P_n) \left\langle \Pi_{V^\perp}, C_X \right\rangle ;$$

we have already treated this quantity in the proof of Theorem 3.2, hence the first part is proved.

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

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

The first term has been dealt with above. We obtain a lower bound for the second term using Hoeffding's inequality (again, exactly as in the proof of the lower bound in Theorem 3.2). This concludes the proof. □

4.1.2 Fast rates via localized approach

We now show that that the excess error of the best empirical d-dimensional subspace with respect to the error of the best d-dimensional subspace can decay at a much faster rate than can be expected from Theorem 4.1. 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 the section:

Theorem 4.2. *Let (λ_i) denote the ordered eigenvalues with multiplicity of C_1, resp. (μ_i) the ordered distinct eigenvalues. Let $\widetilde{d}$ be such that $\lambda_d = \mu_{\widetilde{d}}$. Define*

$$\gamma_d = \begin{cases} \mu_{\widetilde{d}} - \mu_{\widetilde{d}+1} & \text{if } \widetilde{d} = 1 \text{ or } \lambda_d > \lambda_{d+1}, \\ \min\left(\mu_{\widetilde{d}-1} - \mu_{\widetilde{d}}, \mu_{\widetilde{d}} - \mu_{\widetilde{d}+1}\right) & \text{otherwise}; \end{cases} \tag{24}$$

and $B_d = 2\sqrt{\mathbb{E}k^4(X, X')}/\gamma_d$.

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

$$R(\widehat{V}_d) - R(V_d) \le 7 \inf_{h \ge 0} \left\{ \frac{B_d h}{n} + 4\sqrt{\frac{d}{n} \sum_{j \ge h+1} \lambda_j(K_2)} \right\} + \frac{\xi(22M + 6B_d)}{n}. \tag{25}$$

Comments. Similarly to the remarks on Theorem 3.4, the complexity term obtained in Theorem 4.2 has a faster (or equal) decay rate, as a function of the sample size n, than the one of Theorem 4.1; this rate depends on the decay behavior of the eigenvalues.

We do state a fully empirical version of the bound (using only empirical eigenvalues) to avoid additional burden. Let us sketch briefly how this could be obtained: in the proof of the Theorem, we can use the *empirically* localized Rademacher complexity at the price of worse constants (see the proof of Theorem 3.4 to see an example of how this plays out). This has the effect of replacing the true eigenvalues by the empirical ones in the sum appearing in (25). However the constant B_d still depends on the true eigenvalues. For this, we can use a simple convergence result of the empirical eigenvalues to the true ones (as proved for example by Koltchinskii and Giné, 2000), so that for n big enough B_d is bounded by $2\widehat{B}_d$ (its empirical counterpart).

The techniques used to obtain the previous fast rates for reconstruction error of KPCA allows us to get improved bounds for the sum of the *smaller* eigenvalues.

Corollary 4.3. *Under Assumption* **(A)***,with probability at least $1 - e^{-\xi}$, for all $K > 1$,*

$$\sum_{k \ge d} \lambda_k(K_1) - \frac{K}{K+1} \sum_{k \ge d} \lambda_k(K_{1,n}) \ge -\frac{M\xi}{n}\left(\frac{1}{3} + \frac{K}{2}\right). \tag{26}$$

Moreover, if $\lambda_d > \lambda_{d+1}$, for all $K > 1$ and $\xi > 0$, with probability at least $1 - 2e^{-\xi}$, the following holds:

$$\sum_{k \geq d+1} \lambda_k(K_1) - \frac{K}{K-1} \sum_{k \geq d+1} \lambda_k(K_{1,n})$$
$$\leq 6K \inf_{h \geq 0} \left\{ \frac{B_d h}{n} + 4\sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} + \frac{K\xi}{n} \left(5B_d + \frac{5M}{6(K-1)} + 22M \right) . \quad (27)$$

where $B_d = 2\sqrt{\mathbb{E}k^4(X, X')}/(\lambda_d - \lambda_{d+1})$.

4.2 Recentered Case

The goal of this section is to show that the rate of convergence obtained in Theorem 4.1 in the uncentered case is of the same order if we consider the empirical re-centering. In this case the Kernel-PCA algorithm solves the following optimization problem:

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

where $\widehat{\overline{V}}_d$ is the vector space spanned by the first d eigenfunctions of $\overline{C}_{1,n}$. We also denote by $\overline{V_d}$ the space spanned by the first d eigenfunctions of $\overline{C}_1$:

$$\overline{V}_d = \operatorname*{Arg\,Min}_{V \in \mathcal{V}_d} \mathbb{E} \| \varphi_X - \mu - \Pi_V(\varphi_X - \mu) \|^2$$

We will adopt the following notation for the reconstruction error:

$$\overline{R}_n(V) = \frac{1}{n-1} \sum_{j=1}^{n} \| \overline{\varphi}_{X_j} - \Pi_V(\overline{\varphi}_{X_j}) \|^2 = \left\langle \Pi_{V^\perp}, \widetilde{C}_{1,n} \right\rangle .$$

$$\overline{R}(V) = \mathbb{E} \| \varphi_X - \mu - \Pi_V(\varphi_X - \mu) \|^2 = P \left\langle \Pi_{V^\perp}, \overline{C}_X \right\rangle .$$

One has $\overline{R}_n(\widehat{\overline{V}}_d) = \frac{n}{n-1} \sum_{i>d} \lambda_i(\overline{K}_{1,n})$ and $\overline{R}(\overline{V}_d) = \sum_{i>d} \lambda_i(\overline{K}_1)$. Following the same line of reasonning as in Theorem 4.1 and using Theorem 3.5 to control the supremum yields the following result.

Theorem 4.4. *Under Assumption* **(A)**, *for any $\xi > 1$, with probability greater than $1 - 3e^{-x}$,*

$$\overline{R}(\widehat{\overline{V}}_d) \leq \frac{n}{n-1} \sum_{i>d} \lambda_i(\overline{K}_{1,n}) + 2\sqrt{\frac{d}{n} \operatorname{tr} K_{2,n}} + 18M\sqrt{\frac{\xi}{n}} .$$

Note that the leading complexity term is the same as in Theorem 4.1: hence recentering in kernel PCA essentially does not introduce additional complexity to the procedure.

5 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) and Koltchinskii (1998), we are interested in non-asymptotic (i.e. finite sample sizes) results. Also, as we are only interested in the case where $k(x, y)$ is a positive definite function, we have the nice property of Theorem 2.3 which allows to consider the empirical operator and its limit as acting on the same space (since we can use covariance operators on the RKHS). This is crucial in our analysis and makes precise non-asymptotic computations possible unlike in the general case studied by Koltchinskii and Giné (2000); Koltchinskii (1998).

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). Moreover their approach for eigenvalues is based on the concentration around the mean of the empirical eigenvalues and on the relationship between the expectation of the empirical eigenvalues and the operator eigenvalues. Here we used a direct approach and extend their results to the recentered case and proved refined bounds 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 of the d-dimensional projection found by the kernel PCA procedure to the ideal one.

Open issues: the nagging problem of the choice of dimension in PCA. All along this paper, the integer d (the number of eigenvalues summed, or the dimension of the space selected by PCA) was always considered fixed a priori.

It is tempting to interpret the bounds appearing in Theorems 4.1 and 4.2 as a classical statistical tradeoff between approximation error (empirical reconstruction error, decreasing with the dimension d) and estimation error (complexity term, increasing with d). This point of view would suggest to select d as the dimension minimizing the bound. However, this view is an illusion since it is clear that 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 two important points: first, that the (true) reconstruction error is by itself not a good criterion to select the dimension (of course, with this criterion the best choice would be not to project the data at all but to keep the whole space). Hence, an alternative and sensible criterion has to be found to define in a well-founded way what the optimal dimension would be.

A second consequence of this observation is that the bounds we found do not exhibit the correct behavior in terms of the dimension d (for a fixed sample size n), since they become *increasing* in d, for big enough d, while the true error is always *decreasing*. Because of the decreasing property of the true error, any quantity bounding the reconstruction error for dimension d is also a valid bound for any $d' > d$. Hence, if we denote $d(n)$ the dimension realizing the mininum of the bound of Theorem 4.1 (for example) for a fixed sample size n, then the bound obtained for $d(n)$ is also valid for any larger dimension and actually *more informative* than the bounds obtained directly for this larger dimension. This property was also noticed by Shawe-Taylor et al. (2005). To sum up, our bound on the estimation error is too pessimistic for larger dimensions and does not provide a correct qualitative explanation for what is really taking place. Obtaining a better understanding of the behavior of the estimation error for fixed n and varying d is a very interesting open problem, which could also eventually lead to a relevant dimension selection criterion (maybe by comparison of the relative importance of approximation error and estimation error for larger dimensions).

We conclude by mentioning additional open problems: it would be of interest to obtain relative convergence rates for the estimation of single eigenvalues, and to obtain nonasymptotic bounds for eigenspace estimation.

Acknowledgements

The authors are extremely grateful to Stéphane Boucheron for invaluable comments and ideas, as well as for motivating this work.

References

T. W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Stat.*, 34:122–148, 1963.

P. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities, 2003a. Submitted, available at `http://www.kyb.mpg.de/publications/pss/ps2000.ps`.

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

P. Baxendale. Gaussian measures on function spaces. *Amer. J. Math.*, 98:891–952, 1976.

P. Besse. *Etude descriptive d'un processus; approximation, interpolation.* PhD thesis, Université de Toulouse, 1979.

P. Besse. Approximation spline de l'analyse en composantes principales d'une variable aléatoire hilbertienne. *Ann. Fac. Sci. Toulouse (Math.)*, 12(5):329–349, 1991.

O. Bousquet. PhD thesis, Ecole Polytechnique, 2002.

J. Dauxois and A. Pousse. *Les analyses factorielles en calcul des probabilités et en statistique: essai d'étude synthétique.* PhD thesis, Université de Toulouse, 1976.

V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence.* Springer, 1999.

N. Dunford and J. T. Schwartz. *Linear Operators Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space*. Number VII in Pure and Applied Mathematics. John Wiley & Sons, New York, 1963.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

V. Koltchinskii. Asymptotics of spectral projections of some random matrices approximating integral operators. *Progress in Probability*, 43:191–227, 1998.

V. Koltchinskii and E. Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303, 2000.

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B*, 53(3):539–572, 1991.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 327–352. MIT Press, Cambridge, MA, 1999. Short version appeared in *Neural Computation* 10:1299–1319, 1998.

J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. Eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In *Algorithmic Learning Theory : 13th International Conference, ALT 2002*, volume 2533 of *Lecture Notes in Computer Science*, pages 23–40. Springer-Verlag, 2002. Extended version available at `http://www.support-vector.net/papers/eigenspectrum.pdf`.

J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and the generalisation error of kernel pca. *IEEE Transactions on Information Theory 51*, 2005. (To appear).

M. Torki. Etude de la sensibilité de toutes les valeurs propres non nulles d'un opérateur compact autoadjoint. Technical Report LAO97-05, Université Paul Sabatier, 1997. Available at http://mip.ups-tlse.fr/publi/rappLAO/97.05.ps.gz.

R. C. Williamson, J. Shawe-Taylor, B. Schölkopf, and A. J. Smola. Sample-based generalization bounds. *IEEE Transactions on Information Theory*, 1999. Submitted. Also: NeuroCOLT Technical Report NC-TR-99-055.

A Additional proofs

A.1 Proofs for section 2

Proof of Theorem 2.2. 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:

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

this concludes the proof.

Proof of Theorem 2.3. 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 $L_2(\mathcal{X} \times \mathcal{X})$. This is the case here since $k(x,y) \le \|\Phi(x)\| \, \|\Phi(y)\|$ 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, $\mathbb{E}\langle h, \Phi(X)\rangle^2 \le \|h\|^2 \mathbb{E}\|\Phi(X)\|^2$. This shows that T is well-defined and continuous; therefore it has a continuous adjoint T^*. Let $f \in L_2(P)$, then $(\mathbb{E}\,\|f(X)\Phi(X)\|)^2 \le \|f\|^2 \mathbb{E}\,\|\Phi(X)\|^2$. Therefore the variable $f(X)\Phi(X) \in \mathcal{H}$ has a well-defined expectation. 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 $C = T^*T$ and $K_\Phi = TT^*$. By the definition of the expectation, for all $h, h' \in \mathcal{H}$, $\langle h, T^*T(h')\rangle = \langle h, \mathbb{E}[\Phi(X)\langle \Phi(X), h'\rangle]\rangle = \mathbb{E}[\langle h, \Phi(X)\rangle \langle h', \Phi(X)\rangle]$. Thus, by the uniqueness of the covariance operator, we get $C = T^*T$. 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 $K = TT^*$ and f an associated eigenvector. Then $(T^*T)T^* f = T^*(TT^*)f = \lambda T^* f$. This shows that $T^*(E_\lambda(TT^*)) \subset E_\lambda(T^*T)$ and conversely $T(E_\lambda(T^*T)) \subset E_\lambda(TT^*)$. Applying T^* to both terms of the last inclusion implies $E_\lambda(T^*T) \subset T^*(E_\lambda(TT^*))$ since $\lambda \neq 0$, and therefore $T^*T(E_\lambda(T^*T)) = E_\lambda(T^*T)$. Conversely, $E_\lambda(TT^*) \subset T(E_\lambda(T^*T))$ for$\lambda \neq 0$. Thus, $E_\lambda(T^*T) = T^*(E_\lambda(TT^*))$ and $E_\lambda(TT^*) = T(E_\lambda(T^*T))$ and finally $\mathrm{dim}(\mathrm{E}_\lambda(\mathrm{T^*T})) = \mathrm{dim}(\mathrm{E}_\lambda(\mathrm{TT^*}))$. This shows that the multiplicity is the same. This concludes the proof. □

A.2 Proofs for section 3

Proof of Theorem 3.4. As in the proof of Theorem 3.2, we have to consider the empirical process $\langle \Pi_V, C_X\rangle$ for $V \in \mathcal{V}_d$. Let us define

$$\mathcal{F}_d = \{x \mapsto \langle \Pi_V, C_x\rangle \;,\; V \in \mathcal{V}_d\}\,.$$

In order to prove inequality (21), we will apply Theorem B.1 (coming from Bartlett et al. (2003a), and recalled in Appendix B along with some additional notation which we will use here) to the class of functions $M^{-1}\mathcal{F}_d$. From equation (20), it holds that $\forall f \in M^{-1}\mathcal{F}_d\,, f(x) \in [0,1]$, and therefore $Pf^2 \leq Pf$, hence the hypotheses of the theorem are satisfied.

What we need is now to obtain upper bounds for localized Rademacher complexities where the localization is in terms of P or P_n. For this we will need some results about local Rademacher complexities on ellipsoids that are regrouped and shown in Appendix C. Let us first denote the "localized" set

$$S_r = \left\{ g \in \mathrm{star}(M^{-1}\mathcal{F}_d),\ Pg^2 \leq r \right\} = M^{-1} \left\{ g \in \mathrm{star}(\mathcal{F}_d),\ Pg^2 \leq M^2 r \right\} . \tag{28}$$

Corollary C.2 entails

$$\mathbb{E} \sup_{g \in S_r} R_n g \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1} \sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right) := \psi_d(r) \,.$$

We now need to upper-bound the fixed point r_d^* of $\psi_d(r)$. For this we use Lemma C.4 with $c = 1\,, \alpha = M^{-1}$, leading to

$$r_d^* \leq \inf_{h \geq 0} \left\{ \frac{h}{n} + 2M^{-1} \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} . \tag{29}$$

Inequality (37) of Theorem B.1 implies that with probability at least $1 - e^{-\xi}$, every $f \in \mathcal{F}_d$ satisfies

$$P_n f \leq \frac{K+1}{K} Pf + 6KMr_d^* + \frac{M\xi(11+5K)}{n} . \tag{30}$$

Putting in the bound (29), taking the supremum over $f \in \mathcal{F}_d$ on the left-hand, then right-hand side, and using (16), we obtain (21).

In order to prove inequality (22), we apply the second part of theorem B.1, which gives us a confidence bound on r_d^* using the Rademacher complexity localized in terms of the empirical measure. For this we define $\widehat{S}_r$ like S_r in (28) but where P_n takes the role of P. Corollary C.2 entails

$$\mathbb{E}_\varepsilon \sup_{g \in \widehat{S}_r} R_n g \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1} \sqrt{d \sum_{k \geq h+1} \lambda_k(K_{2,n})} \right) := \widehat{\psi}_d(r) \,. \tag{31}$$

Then Theorem B.1 tells us that with probability $1 - 2e^{-\xi}$, r_d^* is upper bounded by the fixed point of $20\widehat{\psi}_d(2r) + 31\xi/n$. To upper bound this quantity in turn, we first apply Lemma C.4 with $c = 2, \alpha = M^{-1}$ as above to obtain a bound on the fixed point of $\widehat{\psi}_d(2r)$; then we apply Lemma B.2 with $K = \frac{7}{6}$. Gathering these inequalities and after straightforward calculations, we finally get that with probability at least $1 - 3e^{-\xi}$, $\forall f \in \mathcal{F}_d$,

$$P_n f \leq \frac{K+1}{K} P f + 282K \inf_{h \geq 0} \left\{ \frac{2hM}{n} + \sqrt{2} \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_{2,n})} \right\} + \frac{2620MK\xi}{n},$$

leading to (22).

Finally, inequality (23) is a simple consequence of Bernstein's inequality. $\square$

Proof of Theorem 3.5. The proof of this Theorem follows the same structure as for Theorem 3.2, but some additional ingredients are needed to control U-processes arising from the recentering.

We prove the first statement of Theorem 3.5: the second one follows from the same arguments. First recall the following decomposition from equations (14) and (15):

$$\overline{C}_1 = C_1 - \mu \otimes \mu^* \quad \text{and} \quad \widetilde{C}_{1,n} = C_{1,n} - \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \varphi_{X_i} \otimes \varphi^*_{X_j}, \tag{32}$$

from which we obtain

$$\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \widetilde{C}_{1,n} - \overline{C}_1 \right\rangle \leq \sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, C_{1,n} - C_1 \right\rangle + \sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi^*_{X_j} \right\rangle. \tag{33}$$

It was shown in the proof of Theorem 3.2 that the following holds with probability greater than $1 - 2e^{-\xi}$:

$$\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, C_{1,n} - C_1 \right\rangle \leq 2\sqrt{\frac{d}{n}} \sqrt{\operatorname{tr} K_{2,n}} + 3M\sqrt{\frac{\xi}{2n}},$$

so we now concentrate on the second term of (33). If we denote $G(x_1, \ldots, x_n) = \left\langle \Pi_V, \mu \otimes \mu^* - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi^*_{X_i} \right\rangle$, then we have for any i_0:

$$\begin{aligned} &\left| G(x_1, \ldots, x_n) - G(x_1, \ldots, x_{i_0 - 1}, x'_{i_0}, x_{i_0+1}, \ldots, x_n) \right| \\ &\qquad \leq \frac{1}{n(n-1)} \left\| \sum_{j \neq i_0} (\varphi_{x_{i_0}} - \varphi_{x'_{i_0}}) \otimes \varphi^*_{x_j} + \varphi_{x_j} \otimes (\varphi^*_{x_{i_0}} - \varphi^*_{x'_{i_0}}) \right\| \\ &\qquad \leq \frac{2}{n(n-1)} \sum_{j \neq i_0} \left\| \varphi_{x'_{i_0}} - \varphi_{x_{i_0}} \right\| \left\| \varphi_{x_j} \right\| \leq \frac{4M}{n}. \end{aligned}$$

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

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

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:

$$S_d = \sup_{V \in \mathcal{V}_d} \frac{2}{n} \sum_{j=1}^{n} \langle \Pi_V, \mu \otimes \mu^* \rangle - \left\langle \Pi_V(\varphi_{X_j}), \mu \right\rangle$$

$$R_d = \sup_{V \in \mathcal{V}_d} \frac{1}{n(n-1)} \sum_{i \neq j} \left(\left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j} \right\rangle - \left\langle \Pi_V(\varphi_{X_j}), \mu \right\rangle \right. \\ \left. - \left\langle \Pi_V(\varphi_{X_i}), \mu \right\rangle + \langle \Pi_V, \mu \otimes \mu^* \rangle \right).$$

It can easily be seen that

$$\mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \mu \otimes \mu - \frac{1}{n(n-1)} \sum_{i \neq j} \varphi_{X_i} \otimes \varphi_{X_j} \right\rangle \right] \leq \mathbb{E}\left[S_d\right] + \mathbb{E}\left[R_d\right].$$

Gathering the different inequalities up to now, we have with probability greater than $1 - 3e^{-\xi}$:

$$\sup_{V \in \mathcal{V}_d} \left\langle \Pi_V, \widetilde{C}_{1,n} - \overline{C}_1 \right\rangle \leq 2\sqrt{\frac{d}{n}} \sqrt{\operatorname{tr} K_{2,n}} + \mathbb{E}[S_d] + \mathbb{E}[R_d] + 7M\sqrt{\frac{\xi}{2n}}. \tag{34}$$

We now bound from above the expectation of S_d and R_d using Lemmas A.1 and A.2 below. This leads to the conclusion. □

Lemma A.1. *The following inequality holds:*

$$\mathbb{E}\left[S_d\right] \leq 4 \frac{\mathbb{E}k(X, X)}{\sqrt{n}}$$

Proof. A standard symmetrization argument leads to

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

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 $\|\mu\|^2 = \mathbb{E}k(X, X') \leq (\mathbb{E}k^{\frac{1}{2}}(X, X))^2$ yields the conclusion. $\square$

Lemma A.2. *The following inequality holds:*

$$\mathbb{E}\left[R_d\right] \leq \frac{6}{n-1} \mathbb{E}k(X, X) \,.$$

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 completeness and for the improved constant.

Proof. Since Π_V is a symmetric operator, using Jensen's inequality ,

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

where

$$f_V(X_i, X_i', X_j, X_j') = \\ \left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j} - \varphi_{X_i'} \otimes \varphi^*_{X_j} - \varphi_{X_i} \otimes \varphi^*_{X_j'} + \varphi_{X_i'} \otimes \varphi^*_{X_j'} \right\rangle .$$

Since $f_V(X_i, X_i', X_j, X_j') = -f_V(X_i', X_i, X_j, X_j')$ and $f_V(X_i, X_i', X_j, X_j') = -f_V(X_i, X_i', X_j', X_j)$, following the proof of the standard symmetrization, we get:

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

Therefore,

$$\mathbb{E}[R_d] \leq \frac{2}{n(n-1)} \left(\mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i \neq j} \varepsilon_i \varepsilon_j \left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j} \right\rangle \right] \right.$$
$$\left. + \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} - \sum_{i \neq j} \varepsilon_i \varepsilon_j \left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X'_j} \right\rangle \right] \right) = \frac{2}{n(n-1)} (A + B) \, ;$$

for the first term above we have

$$A \leq \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i,j} \varepsilon_i \varepsilon_j \left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X_j} \right\rangle \right] = C \, ,$$

while for the second we use

$$B \leq \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} - \sum_{i,j} \varepsilon_i \varepsilon_j \left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X'_j} \right\rangle \right] + \mathbb{E}\left[\sup_{V \in \mathcal{V}_d} \sum_{i} \left\langle \Pi_V, \varphi_{X_i} \otimes \varphi^*_{X'_i} \right\rangle \right]$$
$$= D + E \, .$$

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:

$$C \leq \mathbb{E}_X \mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\| \left\| \sum_j \varepsilon_j \Pi_V(\varphi_{X_j}) \right\| \leq \mathbb{E}_X \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\|^2$$
$$= n \mathbb{E} k(X, X) \, ;$$

$$\begin{aligned}
D &\leq \mathbb{E}_{X,X'} \mathbb{E}_\varepsilon \sup_{V \in \mathcal{V}_d} \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\| \left\| \sum_j \varepsilon_j \Pi_V(\varphi_{X'_j}) \right\| \\
&\leq \mathbb{E}_{X,X'} \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\| \left\| \sum_j \varepsilon_j \varphi_{X'_j} \right\| \\
&\leq \mathbb{E}_{X,X'} \sqrt{ \mathbb{E}_\varepsilon \left\| \sum_i \varepsilon_i \varphi_{X_i} \right\|^2 \mathbb{E}_\varepsilon \left\| \sum_j \varepsilon_j \varphi_{X'_j} \right\|^2 } \\
&\leq \mathbb{E}_{X,X'} \sqrt{ \left(\sum_i k(X_i, X_i) \right) \left(\sum_i k(X'_i, X'_i) \right) } \leq n \mathbb{E} k(X, X) \, ;
\end{aligned}$$

$$
\begin{aligned}
E \le \mathbb{E}_X \sup_{V\in\mathcal{V}_d} \sum_i \left\|\Pi_V(\varphi_{X'_i})\right\| \|\varphi_{X_i}\| &\le \mathbb{E}_X \sum_i \left\|\varphi_{X'_i}\right\| \|\varphi_{X_i}\| \\
&\le \mathbb{E}_X \sum_i \sqrt{k(X'_i,X'_i)k(X_i,X_i)} \\
&= n\mathbb{E}k(X,X)\,.
\end{aligned}
$$

Gathering the previous inequalities, we obtain the conclusion. □

Proof of Theorem 3.6. (Minoration) We prove the lower bound for the largest eigenvalues. A similar proof gives the second statement.
Theorem 2.1 leads to

$$
\sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1) \ge \left\langle \overline{C}_{1,n}, \Pi_{\overline{V}_d}\right\rangle - \left\langle \overline{C}_1, \Pi_{\overline{V}_d}\right\rangle \,.
$$

Using the decomposition (32), we get:

$$
\begin{aligned}
&\sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1) \\
&\qquad\qquad \ge \left\langle C_{1,n} - C_1, \Pi_{\overline{V}_d}\right\rangle - \left\langle \Pi_{\overline{V}_d}, \frac{1}{n(n-1)} \sum_{i\neq j} \varphi_{X_i}\otimes\varphi_{X_j} - \mu\otimes\mu \right\rangle \,.
\end{aligned}
$$

The first term is bounded by Hoeffding's inequality exactly as in the proof of Theorem 18. With probability greater than $1-e^{-x}$,

$$
\left\langle C_{1,n} - C_1, \Pi_{\overline{V}_d}\right\rangle = (P - P_n)\left\langle \Pi_{\overline{V}_d}, C_X\right\rangle \ge -M\sqrt{\frac{\xi}{2n}}\,.
$$

For the second term, we apply Hoeffding's inequality for U-statistics (see e.g. Hoeffding, 1963; de la Peña and Giné, 1999); with probability greater than $1-e^{-\xi}$,

$$
-\left\langle \Pi_{\overline{V}_d}, \frac{1}{n(n-1)} \sum_{i\neq j} \varphi_{X_i}\otimes\varphi_{X_j} - \mu\otimes\mu \right\rangle \ge -M\sqrt{\frac{\xi}{2\lceil\frac{n}{2}\rceil}} \ge -M\sqrt{\frac{\xi}{n}}\,.
$$

We finally obtain

$$
\sum_{i=1}^{d} \lambda_i(\overline{C}_{1,n}) - \sum_{i=1}^{d} \lambda_i(\overline{C}_1) \ge -M\sqrt{\frac{\xi}{n}}\left(1+\frac{1}{\sqrt{2}}\right)\,.
$$

Finally using Lemma 2.5 with true and empirical distributions yields the conclusion.

A.3 Proofs for section 4

A key property necessary for the proof of Theorem 4.2 is established in the following Lemma:

Lemma A.3. *Let (λ_i) denote the ordered eigenvalues with multiplicity of C_1, resp. (μ_i) the ordered distinct eigenvalues, and γ_d be defined as in equation (24). For any $V \in \mathcal{V}_d$, there exists $H_V \in \mathcal{V}_d$ such that*

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

and

$$\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X \right\rangle^2\right] \le 2\gamma_d^{-1}\sqrt{\mathbb{E}\left[k^4(X,X')\right]}\mathbb{E}\left[\left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_X \right\rangle\right]\,.$$

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

Step 1: construction of H_V.

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

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

We now define $H_V = \operatorname{span}\{\phi_i, 1 \le i \le d\}$. Obviously H_V is a minimizer of the reconstruction error over subspaces of dimension d. We have (using Lemma 2.4 (ii) at the first line)

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

and on the other hand, using Lemma 2.4 (i):

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

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 $g_i \in \overline{W}^{\perp}_{\widetilde{d}-1}$. We have

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

and

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

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 $g'_i \in \overline{W}^{\perp}_{\widetilde{d}}$. We have

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

and

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

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

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

Finally, it holds that $\|C_2\|_{op} \leq \|C_2\|_{\mathrm{HS}(\mathrm{HS}(\mathcal{H}_k))} = \|K_2\|_{\mathrm{HS}(L_2(P))}$ by Lemma 2.4 (iv); since K_2 is an integral operator with kernel $k^2(x,y)$, we have $\|K_2\|^2_{\mathrm{HS}(L_2(P))} = \int k^4(x,y) dP(x) dP(y) = \mathbb{E}[k^4(X,X')]$. This concludes the proof of the Lemma when $\widetilde{d} > 1$. If $\widetilde{d} = 1$, 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. $\square$

Proof of Theorem 4.2. We will use here again Theorem B.1. We define the following class of functions:

$$\widetilde{\mathcal{F}}_d = \left\{ x \mapsto \left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_x \right\rangle, V \in \mathcal{V}_d \right\},$$

where for each $V \in \mathcal{V}_d$, H_V is obtained via Lemma A.3. We will apply Theorem B.1 to the class $M^{-1}\widetilde{\mathcal{F}}_d$. For any $f \in M^{-1}\widetilde{\mathcal{F}}_d$, it holds that $f \in [-1,1]$; furthermore, Lemma A.3 entails that $Pf^2 \leq M^{-1}B_d P_f$. To upper bound the local Rademacher complexities of this class we define

$$\widetilde{S}_r = \left\{ g \in \text{star}(M^{-1}\widetilde{\mathcal{F}}_d),\ Pg^2 \leq r \right\} = M^{-1}\left\{ g \in \text{star}(\widetilde{\mathcal{F}}_d),\ Pg^2 \leq M^2 r \right\}.$$

Corollary C.3 entails

$$M^{-1}B_d \mathbb{E} \sup_{g \in \widetilde{S}_r} R_n g \leq \frac{M^{-1}B_d}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + M^{-1}\sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right) := \widetilde{\psi}_d(r).$$

Let $\widetilde{r}_d^*$ denote the solution of equation $\widetilde{\psi}_d(r) = r$. We apply Lemma C.4 with the choice $c = M^{-1}B_d, \alpha = M^{-1}$ to obtain

$$\widetilde{r}^* \leq M^{-2} \inf_{h \geq 0} \left\{ \frac{B_d^2 h}{n} + 4B_d \sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\}.$$

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

$$\forall f \in \widetilde{\mathcal{F}}_d,\ Pf \leq \frac{K}{K-1} P_n f + 6K \inf_{h \geq 0} \left\{ \frac{B_d h}{n} + 4\sqrt{\frac{d}{n} \sum_{j \geq h+1} \lambda_j(K_2)} \right\} + \frac{\xi(11M + 5B_d K)}{n} \tag{35}$$

Choosing $V = \widehat{V}_d$ leads to the result. □

Proof of Theorem 4.3. Inequality (26) is a simple consequence of Bernstein's inequality. We now prove inequality (27).

Since we suppose $\lambda_d > \lambda_{d+1}$, $H_V = V_d$ for all $V \in \mathcal{V}_d$. Moreover,

$$\sum_{k \geq d+1} \lambda_k(K_1) - \frac{K}{K-1} \sum_{k \geq d+1} \lambda_k(K_{1,n}) \leq R(\widehat{V}_d) - \frac{K}{K-1} R_n(\widehat{V}_d).$$

Finally, inequality (27) is obtained by gathering inequality (35) and Bernstein's inequality to control $(P - P_n)\left\langle \Pi_{V_d^\perp}, C_x \right\rangle$. □

B Local Rademacher complexities

In this section we recall a fundamental Theorem that is the key to controlling deviations of empirical processes using local Rademacher averages defined either from the true or the empirical distribution. It is a simplified version of Theorems 3.3 and 4.1 of Bartlett et al. (2003a). In the terminology of the latter reference, a *sub-root* function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is nonnegative, nondecreasing, and such that $\psi(r)/\sqrt{r}$ is nonincreasing. Then it can be shown that the fixed point equation $\psi(r) = r$ has a unique positive solution (except for the trivial case $\psi \equiv 0$). Moreover, this solution r^* satisfies that $r^* \leq r$ if and only if $\Psi(r) \leq r$. Also we need the following notation for Rademacher complexities:

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

where (ε_i) are i.i.d. Rademacher; we finally define the *star-shaped hull*

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

Theorem B.1 (Bartlett, Bousquet and Mendelson). *Let $\mathcal{F}$ be a class of functions with ranges in $[-1,1]$ and assume that there exists some constant $B > 0$ such that for every $f \in \mathcal{F}\,, Pf^2 \leq BPf$. Let ψ be a sub-root function and r^* be the fixed point of ψ. If ψ satisfies*

$$\psi(r) \geq B\mathbb{E}_{X,\varepsilon} R_n \left\{ f \in \mathrm{star}(\mathcal{F}) : Pf^2 \leq r \right\}\,,$$

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

$$\forall f \in \mathcal{F}\,,\quad Pf \leq \frac{K}{K-1} P_n f + \frac{6K}{B} r^* + \frac{x(11+5BK)}{n}\,; \tag{36}$$

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

$$\forall f \in \mathcal{F}\,,\quad P_n f \leq \frac{K+1}{K} Pf + \frac{6K}{B} r^* + \frac{x(11+5BK)}{n}\,. \tag{37}$$

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

$$\widehat{\psi}_n(r) \geq 2(10 \vee B)\mathbb{E}_\varepsilon R_n \left\{ f \in \mathrm{star}(\mathcal{F}) \,:\, P_n f^2 \leq 2r \right\} + \frac{(2(10 \vee B) + 11)x}{n}\,, \tag{38}$$

then with probability $1 - 2e^{-x}$, it holds that $\widehat{r}^* \geq r^*$; as a consequence, with probability $1 - 3e^{-x}$, inequality (36) holds with r^* replaced by $\widehat{r}^*$; similarly for inequality (37).

We complete this section with the following Lemma which can be used to obtain upper bounds on fixed points of functions of the form (38):

Lemma B.2 (inspired by Bousquet (2002)). *Let ϕ be a sub-root function and let $\phi_1(r) = \alpha\phi(r) + \beta$ with $\alpha > 1$ and $\beta > 0$. Let r^* (resp. r_1^*) denote the fixed point of ϕ (resp. ϕ_1). We have:*

$$r_1^* \leq \inf_{K>1} \left(K\alpha^2 r^* + \frac{\sqrt{K}}{\sqrt{K}-1}\beta \right) .$$

Proof. During this proof, we keep using the definition of a sub-root function and his property recalled previously.
If $a > 1$ and $b > 0$,

$$\alpha\phi(ar^* + b\beta) + \beta = \alpha\phi\left(a\left(r^* + \frac{b}{a}\beta\right)\right) + \beta \leq \alpha\sqrt{a}\phi\left(r^* + \frac{b}{a}\beta\right) + \beta\,,$$

thus

$$\alpha\phi(ar^* + b) + \beta \leq \alpha\sqrt{a}r^* + \beta\left(1 + \alpha\frac{b}{\sqrt{a}}\right) .$$

Let $K > 1$. Choosing $a = K\alpha^2$ and $b = \frac{\sqrt{K}}{\sqrt{K}-1}$ yields $\phi_1(ar^* + b\beta) \leq ar^* + b\beta$. This concludes the proof of Lemma B.2. $\square$

C Localized Rademacher Averages on Ellipsoids

In this section we group together results that deal with estimating localized Rademacher complexities of function classes given as ellipsoids of a reproducing kernel Hilbert space. We deduce as corollaries the results necessary for the proofs of Theorems 3.4 and 4.2.

Theorem C.1. *Let $\mathcal{H}$ be a separable Hilbert space and $(Z_i)_{1\leq i\leq n} \in \mathcal{H}^n$. Let A be a compact self-adjoint positive linear operator of $\mathcal{H}$ and $(\Phi_i)_{i\geq 1}$ an orthonormal basis of $\mathcal{H}$ of eigenvectors of A. Denote $B_\alpha = \{\|v\| \leq \alpha\}$, $\mathcal{E}_r = \{\langle v, Av\rangle \leq r\}$ and let (ε_i) be an i.i.d. family of Rademacher random variables. Then for any integer $h \leq Rank(A)$, the following holds:*

$$\mathbb{E}_\varepsilon \sup_{v\in B_\alpha\cap\mathcal{E}_r} \frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle v, Z_i\rangle \leq \frac{\sqrt{r}}{n}\sqrt{\sum_{i=1}^h \frac{1}{\lambda_i(A)}\sum_{j=1}^n \langle Z_j, \Phi_i\rangle^2} + \frac{\alpha}{n}\sqrt{\sum_{i\geq h+1}\sum_{j=1}^n \langle Z_j, \Phi_i\rangle^2} \tag{39}$$

Proof. For $v \in B_\alpha \cap \mathcal{E}_r$, we have

$$\begin{aligned}
\sum_{i=1}^n \varepsilon_i\langle v, Z_i\rangle &= \sum_{j=1}^h \langle v, \Phi_j\rangle \left\langle \Phi_j, \sum_{i=1}^n \varepsilon_i Z_i \right\rangle + \sum_{j>h} \langle v, \Phi_j\rangle \left\langle \Phi_j, \sum_{i=1}^n \varepsilon_i Z_i \right\rangle \\
&\leq \sqrt{r\sum_{i=1}^h \frac{1}{\lambda_i(A)} \left\langle \sum_{j=1}^n \varepsilon_j Z_j, \Phi_i \right\rangle^2} + \alpha\sqrt{\sum_{i\geq h+1} \left\langle \sum_{j=1}^n \varepsilon_j Z_j, \Phi_i \right\rangle^2} ,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality for both terms and the equality $\langle v, Av \rangle = \sum_{i \geq 1} \lambda_i(A) \langle v, \Phi_i \rangle^2$. We now integrate over (ε_i); using Jensen's inequality the square roots are pulled outside of the expectation; finally, we have

$$\mathbb{E}_\varepsilon \left\langle \sum_{j=1}^n \varepsilon_j Z_j, \Phi_i \right\rangle^2 = \sum_{j=1}^n \langle Z_j, \Phi_i \rangle^2 .$$

since by independence the cross-terms vanish. This concludes the proof. □

We deduce the two following corollaries of Theorem C.1:

Corollary C.2. *Define* $\mathcal{F}_d = \{x \mapsto \langle \Pi_V, C_x \rangle, V \in \mathcal{V}_d\}$. *Then the following holds:*

$$\mathbb{E}_{X,\varepsilon} R_n \left\{ f \in \operatorname{star}(\mathcal{F}_d), Pf^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + \sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right) \tag{40}$$

and

$$\mathbb{E}_\varepsilon R_n \left\{ f \in \operatorname{star}(\mathcal{F}_d), Pf^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left(\sqrt{rh} + \sqrt{d \sum_{k \geq h+1} \lambda_k(K_{2,n})} \right) \tag{41}$$

Proof. The proof is the same for the two inequalities. We will apply Theorem C.1 in the Hilbert space $\mathrm{HS}(\mathcal{H})$. We have for any $V \in \mathcal{V}_d$, $\|\Pi_V\|_{\mathrm{HS}(\mathcal{H})} \leq \sqrt{d}$, and hence $\mathcal{F}_d \subset \{x \mapsto \langle \Gamma, C_x \rangle ; \Gamma \in B_{\sqrt{d}}(\mathrm{HS}(\mathcal{H}))\}$. Since the latter set is convex and contains the origin, it therefore also contains $\operatorname{star}(\mathcal{F}_d)$. Furthermore, by Lemma 2.4, $P \langle \Gamma, C_x \rangle^2 = \langle \Gamma, C_2 \Gamma \rangle$.

We can therefore apply Theorem C.1 with $\alpha = \sqrt{d}$, $A = C_2$, $Z_i = C_{X_i}$, $v = \Pi_V$, leading to

$$\mathbb{E}_\varepsilon R_n \left\{ f \in \operatorname{star}(\mathcal{F}_d), Pf^2 \leq r \right\} \leq \frac{\sqrt{r}}{n} \sqrt{\sum_{i=1}^h \frac{1}{\lambda_i(C_2)} \sum_{j=1}^n \left\langle C_{X_j}, \Phi_i \right\rangle^2} + \frac{\sqrt{d}}{n} \sqrt{\sum_{i \geq h+1} \sum_{j=1}^n \left\langle C_{X_j}, \Phi_i \right\rangle^2} .$$

Integrating with respect to Z leads to

$$\mathbb{E}_{X,\varepsilon} R_n \left\{ f \in \operatorname{star}(\mathcal{F}_d), Pf^2 \leq r \right\} \leq \frac{1}{\sqrt{n}} \left(\sqrt{rh} + \sqrt{d \sum_{k \geq h+1} \lambda_k(K_2)} \right) ,$$

since $\mathbb{E}\left[\langle C_X, \Phi_i \rangle^2 \right] = \langle \Phi_i, C_2 \Phi_i \rangle = \lambda_i(C_2)$. We obtain (40) in the same way by taking $A = C_{2,n}$ instead of C_2. □

Corollary C.3. *Define* $\widetilde{\mathcal{F}}_d = \left\{x \mapsto \left\langle \Pi_{V^\perp} - \Pi_{H_V^\perp}, C_x \right\rangle, V \in \mathcal{V}_d\right\}$*, where* H_V *is defined via Lemma A.3. Then the following holds:*

$$\mathbb{E}_{X,\varepsilon} R_n \left\{ f \in \operatorname{star}(\widetilde{\mathcal{F}}_d), Pf^2 \le r \right\} \le \frac{1}{\sqrt{n}} \inf_{h \ge 0} \left\{ \sqrt{rh} + 2\sqrt{d \sum_{j>h} \lambda_j(K_2)} \right\} \tag{42}$$

and

$$\mathbb{E}_{\varepsilon} R_n \left\{ f \in \operatorname{star}(\widetilde{\mathcal{F}}_d), Pf^2 \le r \right\} \le \frac{1}{\sqrt{n}} \inf_{h \ge 0} \left\{ \sqrt{rh} + 2\sqrt{d \sum_{j>h} \lambda_j(K_{2,n})} \right\}. \tag{43}$$

Proof. Note that $\Pi_{V^\perp} - \Pi_{H_V^\perp} = \Pi_{H_V} - \Pi_V$. The proof is then almost the same as for Corollary C.2, with the minor change $\alpha = 2\sqrt{d}$ since $\|\Pi_V - \Pi_{H_V}\|^2_{\mathrm{HS}(\mathcal{H}_k)} \le 4d$. □

We finally give the following Lemma to estimate the fixed points of sub-root functions of the above form.

Lemma C.4. *If* $(\lambda_i)_{i>0}$ *is a positive convergent series, denoting by* ψ *the function*

$$\psi(r) := \frac{1}{\sqrt{n}} \inf_{h \ge 0} \left\{ \sqrt{hr} + \alpha \sqrt{\sum_{j \ge h+1} \lambda_j} \right\},$$

it holds that ψ *is a sub-root function and the unique positive solution* r^* *of* $\psi(r) = r/c$ *where* $c > 0$ *satisfies*

$$r^* \le \inf_{h \ge 0} \left\{ \frac{c^2 h}{n} + \frac{2c\alpha}{\sqrt{n}} \sqrt{\sum_{j \ge h+1} \lambda_j} \right\}$$

Proof. It is easy to see that the minimum of two sub-root functions is sub-root, hence ψ as the minimum of a collection of sub-root function is sub-root. Existence and uniqueness of a solution is proved by Bartlett et al. (2003a). To obtain the announced bound, we solve $r^* \le \frac{c}{\sqrt{n}} \left\{ \sqrt{hr^*} + \alpha \sqrt{\sum_{j \ge h+1} \lambda_j} \right\}$ 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. □