Statistical properties of Kernel Principal Component Analysis
Laurent Zwald, Olivier Bousquet, Gilles Blanchard

To cite this version:
Laurent Zwald, Olivier Bousquet, Gilles Blanchard. Statistical properties of Kernel Principal Component Analysis. COLT, 2004, Banff, AB, Canada. hal-00373799

HAL Id: hal-00373799
https://hal.science/hal-00373799
Submitted on 7 Apr 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Statistical Properties of Kernel Principal Component Analysis

Laurent Zwald¹, Olivier Bousquet², and Gilles Blanchard³

¹ Département de Mathématiques, Université Paris-Sud, Bat. 425, F-91405 Orsay, France
Laurent.Zwald@math.u-psud.fr
² Max Planck Institute for Biological Cybernetics, Spemannstr. 38, D-72076 Tübingen, Germany
olivier.bousquet@tuebingen.mpg.de
³ Fraunhofer First, Kékuléstr. 7, D-12489 Berlin, Germany
blanchard@first.fhg.de

©Springer-Verlag

Abstract. We study the properties of the eigenvalues of Gram matrices in a non-asymptotic setting. Using local Rademacher averages, we provide data-dependent and tight bounds for their convergence towards eigenvalues of the corresponding kernel operator. We perform these computations in a functional analytic framework which allows to deal implicitly with reproducing kernel Hilbert spaces of infinite dimension. This can have applications to various kernel algorithms, such as Support Vector Machines (SVM). We focus on Kernel Principal Component Analysis (KPCA) and, using such techniques, we obtain sharp excess risk bounds for the reconstruction error. In these bounds, the dependence on the decay of the spectrum and on the closeness of successive eigenvalues is made explicit.

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 [19]. 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.

Principal Component Analysis (PCA), and its non-linear variant, kernel-PCA are widely used algorithms in data analysis. They extract from the vector space where the data lie, a basis which is, in some sense, adapted to the data by looking for directions where the variance is maximized. Their applications are very

* Supported by a grant of the Humboldt Foundation
diverse, ranging from dimensionality reduction, to denoising. Applying PCA to a space of functions rather than to a space of vectors was first proposed by Bese [5] (see also [15] for a survey). Kernel-PCA [16] 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 (see e.g. [1]). In the non-linear setting where one uses positive definite kernels, 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, but it also indicates that the properties of the kernel matrix, in particular its spectrum, play a role in the properties of the kernel-PCA algorithm.

Recently, J. Shawe-Taylor, C. Williams, N. Cristianini and J. Kandola [17] have undertaken an investigation of the properties of the eigenvalues of kernel matrices and related it to the statistical performance of kernel-PCA.

In this work, we mainly extend the results of [17]. In particular we treat the infinite dimensional case with more care and we refine the bounds using recent tools from empirical processes theory. We obtain significant improvements and more explicit bounds.

The fact 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), raises a technical difficulty. We propose to tackle this difficulty by using the framework of Hilbert-Schmidt operators and of random vectors in Hilbert spaces. Under some reasonable assumptions (like separability of the RKHS and boundedness of the kernel), things work nicely but some background in functional analysis is needed which is introduced below.

Our approach builds on ideas pioneered by Massart [13], on the fact that Talagrand's concentration inequality can be used to obtain sharp oracle inequalities for empirical risk minimization on a collection of function classes when the variance of the relative error can be related to the expected relative error itself. This idea has been exploited further in [2].

The paper is organized as follows. Section 2 introduces the necessary background on functional analysis and the basic assumptions. We then present, in Section 3 bounds on the difference between sums of eigenvalues of the kernel matrix and of the associated kernel operator. Finally, Section 4 gives our main results on kernel-PCA.

2 Preliminaries

In order to make the paper self-contained, we introduce some background, and give the notations for the rest of the paper.
2.1 Background Material on Functional Analysis

Let $\mathcal{H}$ be a separable Hilbert space. A linear operator $L$ from $\mathcal{H}$ to $\mathcal{H}$ is called Hilbert-Schmidt if $\sum_{i \geq 1} \|L e_i\|_{\mathcal{H}}^2 < \infty$, 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 $\text{HS}(\mathcal{H})$. Endowed with the following inner product $\langle L, N \rangle_{\text{HS}(\mathcal{H})} = \sum_{i,j \geq 1} \langle L e_i, e_j \rangle \langle N e_i, e_j \rangle$, 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 $\lambda(L)$ denotes its spectrum sorted in non-increasing order, repeated according to their multiplicities ($\lambda_1(A) \geq \lambda_2(A) \geq \ldots$). An operator $L$ is called trace-class if $\sum_{i \geq 1} \|e_i\|^2$ is a convergent series. In fact, this series is independent of the chosen orthonormal basis and is called the trace of $L$, denoted by $\text{tr} L$. By Lidskii’s theorem $\text{tr} L = \sum_{i \geq 1} \lambda_i(L)$.

We will keep switching from $\mathcal{H}$ to $\text{HS}(\mathcal{H})$ and treat their elements as vectors or as operators depending on the context, so we will need the following identities. Denoting, for $f, g \in \mathcal{H}$, by $f \otimes g$ the rank one operator defined as $f \otimes g(h) = \langle g, h \rangle f$, it easily follows from the above definitions that $\|f \otimes g\|_{\text{HS}(\mathcal{H})} = \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}$, and for $A \in \text{HS}(\mathcal{H})$,

$$\langle f \otimes g, A \rangle_{\text{HS}(\mathcal{H})} = \langle Ag, f \rangle_{\mathcal{H}}.$$  \hspace{1cm} (1)

We recall that an orthogonal projector in $\mathcal{H}$ is an operator $U$ such that $U^2 = U$ and $U = U^*$ (hence positive). In particular one has $\|U(h)\|_{\mathcal{H}}^2 = \langle h, U h \rangle_{\mathcal{H}}$ and $U$ has rank $d < \infty$ (i.e. it is a projection on a finite dimensional subspace), i.e. if and only if it is Hilbert-Schmidt with $\|U\|_{\text{HS}(\mathcal{H})} = \sqrt{d}$ and $\text{tr} U = d$. 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 subspaces of $\mathcal{H}$, we denote by $\Pi_V$ the unique orthogonal projector such that $\Pi_V = V$ and $\text{ker} \Pi_V = V^\perp$. When $V$ is of finite dimension, $\Pi_{V+}$ is not Hilbert-Schmidt, but we will denote, for a trace-class operator $A$, $\langle \Pi_{V+}, A \rangle = \text{tr} A - \langle \Pi_V, A \rangle_{\text{HS}(\mathcal{H})}$ with some abuse of notation.

2.2 Kernel and Covariance Operators

We recall basic facts about random elements in Hilbert spaces. A random element $Z$ in a separable Hilbert space has an expectation $\mathbb{E} Z = \mathbb{E} Z$ and is the unique vector satisfying $\langle \varepsilon, f \rangle_{\mathcal{H}} = \mathbb{E} \langle Z, f \rangle_{\mathcal{H}}$, for all $f \in \mathcal{H}$. Moreover, when $\mathbb{E} \|Z\|^2 < \infty$, there exists a unique operator $C : \mathcal{H} \rightarrow \mathcal{H}$ such that $\langle f, C g \rangle_{\mathcal{H}} = \mathbb{E} \langle f, Z \rangle_{\mathcal{H}} \langle g, Z \rangle_{\mathcal{H}}$, for all $f, g \in \mathcal{H}$. $C$ is called the covariance operator of $Z$ and is self-adjoint, positive, trace-class operator, with $\text{tr} C = \mathbb{E} \|Z\|^2$ (see e.g. [4]).

The core property of kernel operators that we will use is its intimate relationship with a covariance operator and it is summarized in next theorem. This
property was first used in a similar but more restrictive context (finite-dimensional) by Shawe-Taylor, Williams, Cristianini and Kandola [17].

**Theorem 1.** Let \((X, \mathcal{P})\) be a probability space, \(\mathcal{H}\) be a separable Hilbert space and \(\Phi\) be a map from \(X\) to \(\mathcal{H}\) such that for all \(h \in \mathcal{H}\), \(\langle h, \Phi(x) \rangle_{\mathcal{H}}\) is measurable and \(\mathbb{E}||\Phi(X)||^2 < \infty\). Let \(C\) be the covariance operator associated to \(\Phi(X)\) and \(K : L^2(\mathcal{P}) \rightarrow L^2(\mathcal{P})\) be the integral operator defined as

\[
(Kf)(x) = \int f(y) \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} \, d\mathcal{P}(y).
\]

Then \(\lambda(K) = \lambda(C)\).

In particular, \(K\) is a positive self-adjoint trace-class operator and \(\text{tr}(K) = \mathbb{E}||\Phi(X)||^2 = \sum_{i \geq 1} \lambda_i(K)\).

### 2.3 Eigenvalues Formula

We denote by \(\mathcal{V}_d\) the set of subspaces of dimension \(d\) of \(\mathcal{H}\). The following theorem whose proof can be found in [18] gives a useful formula to compute sums of eigenvalues.

**Theorem 2 (Fan).** Let \(C\) a compact self-adjoint operator on \(\mathcal{H}\), then

\[
\sum_{i=1}^{d} \lambda_i(C) = \max_{V \in \mathcal{V}_d} \langle HV, C \rangle_{HS(\mathcal{H})},
\]

and the maximum is reached when \(V\) is the space spanned by the first \(d\) eigenvectors of \(C\).

We will also need the following formula for single eigenvalues.

**Theorem 3 (Courant-Fischer-Weyl, see e.g. [9]).** Let \(C\) a compact self-adjoint operator on \(\mathcal{H}\), then for all \(d \geq 1\),

\[
\lambda_d(C) = \min_{V \in \mathcal{V}_d} \max_{f \in V} \frac{\langle f, Cf \rangle}{\|f\|^2},
\]

where the minimum is attained when \(V\) is the span of the first \(d - 1\) eigenvectors of \(C\).

### 2.4 Assumptions and Basic Facts

Let \(\mathcal{X}\) denote the input space (an arbitrary measurable space) and \(\mathcal{P}\) 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 \(f : \mathcal{X} \rightarrow \mathbb{R}\), we may use the notation \(P f := \mathbb{E}[f(X)]\) and \(P_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i)\). Also,
\(c_1, \ldots, c_n\) will denote a sequence of Rademacher random variables (i.e., independent with value +1 or -1 with probability 1/2).

Let \(k\) be a positive definite function on \(\mathcal{X}\) and \(\mathcal{H}_k\) the associated reproducing kernel Hilbert space. They are related by the reproducing property: \(\forall f \in \mathcal{H}_k, \forall x \in \mathcal{X}, \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k} = f(x)\). We denote by \(\mathcal{V}\) the set of all vector subspaces of dimension \(d\) of \(\mathcal{H}_k\).

We will always work with the following assumption.

**Assumption 1** We assume that

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

For \(x \in \mathcal{X}\), we denote \(\varphi_x = k(x, \cdot)\) understood as an element of \(\mathcal{H}_k\).

Let \(C_x\) the operator defined on \(\mathcal{H}_k\) by

\[
\langle f, C_x g \rangle_{\mathcal{H}_k} = f(x)g(x).
\]

It is easy to see that \(C_x = \varphi_x \otimes \varphi_x\) and \(C_x\) is trace-class with \(\text{tr} \, C_x = k(x, x)\) and \(\|C_x\|_{\text{HS}(\mathcal{H}_k)} = k(x, x)^2\).

Also, from the definitions and by (1) we have for example \(\langle C_x, C_y \rangle_{\text{HS}(\mathcal{H}_k)} = k^2(x, y)\) and, for any projector \(U\), \(\|U \varphi_x\|_{\mathcal{H}_k}^2 = \langle U, C_x \rangle_{\text{HS}(\mathcal{H}_k)}\).

We will denote by \(C_1 : \mathcal{H}_k \to \mathcal{H}_k\) (resp. \(C_2 : \text{HS}(\mathcal{H}_k) \to \text{HS}(\mathcal{H}_k)\)) the covariance operator associated to the random element \(\varphi_X\) in \(\mathcal{H}_k\) (resp. \(C_X\) in \(\text{HS}(\mathcal{H}_k)\)). Also, let \(K_1\) (resp. \(K_2\)) be the integral operator with kernel \(k(x, y)\) (resp. \(k(x, y)^2\))

**Lemma 1.** Under Assumption 1 the operators \(C_1, C_2, K_1, K_2\) defined above are trace-class with \(\text{tr} \, C_1 = \mathbb{E}[k(X, X)]\), \(\text{tr} \, C_2 = \mathbb{E}[k^2(X, X)]\). They satisfy the following properties

(i) \(\lambda(C_1) = \lambda(K_1)\) and \(\lambda(C_2) = \lambda(K_2)\).
(ii) \(C_1\) is the expectation in \(\text{HS}(\mathcal{H}_k)\) of \(C_X\).
(iii) \(C_2\) is the expectation in \(\text{HS}(\text{HS}(\mathcal{H}_k))\) of \(C_X \otimes C_X\).

**Proof.** (i) To begin with, we prove that \(\text{tr} \, C_1 = \mathbb{E}k(X, X)\) and \(\lambda(C_1) = \lambda(K_1)\) by applying Theorem 1 with \(\Phi(x) = \varphi_x\): since \(k(x, \cdot)\) is measurable, all linear combinations and pointwise limits of such combinations are measurable, so that all the functions in \(\mathcal{H}_k\) are measurable. Hence measurability, for \(h \in \mathcal{H}_k\) of \(x \mapsto \langle \Phi_x, h \rangle_{\mathcal{H}_k}\) follows and we have \(\mathbb{E}|\Phi_X|^2 = \mathbb{E}k(X, X) < \infty\).

Then, we prove that \(\text{tr} \, C_2 = \mathbb{E}k^2(X, X)\) and \(\lambda(C_2) = \lambda(K_2)\) by applying Theorem 1 with \(\Phi(x) = C_x\): for \(h \in \text{HS}(\mathcal{H}_k)\) with finite rank (i.e., \(h = \sum_{i=1}^n \phi_i \otimes \psi_i\) for an orthonormal set \(\phi_i\) and \(\psi_i = h^* \phi_i\), the function \(x \mapsto \langle C_x, h \rangle_{\text{HS}(\mathcal{H}_k)} = \sum_{i=1}^n \phi_i(x)\psi_i(x)\) is measurable (since \(\phi_i\) and \(\psi_i\) are measurable as elements of \(\mathcal{H}_k\)). Moreover, since the finite rank operators are dense in \(\text{HS}(\mathcal{H}_k)\) and \(h \mapsto \langle C_x, h \rangle_{\text{HS}(\mathcal{H}_k)}\) is continuous, we have measurability for all \(h \in \text{HS}(\mathcal{H}_k)\). Finally, we have \(\mathbb{E}|C_X|^2_{\text{HS}(\mathcal{H}_k)} = \mathbb{E}k^2(X, X) < \infty\).
(ii) Since $\mathbb{E}\|C_X\|_{HS(H_k)} = \mathbb{E}k(X, X) < \infty$ the expectation of $C_X$ is well defined in $HS(H_k)$. Moreover for all $f, g \in H_k$, $\langle EC_X f, g \rangle = \mathbb{E}(C_X f, g) := \mathbb{E}(C_X, g \odot f) = \mathbb{E}(C_X f, g) = \mathbb{E}f(X)g(X) = \langle C_1 f, g \rangle$.

(iii) Using $\|C_X \odot C_X\|_{HS(H_k^2)} = \|C_X\|_{HS(H_k)}^2$ and a similar argument gives the last statement.

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. If we do so, the associated operators are denoted by $K_{1,n}$ (which is identified [12] with the normalized kernel matrix of size $n \times n$, $K_{1,n} \equiv (k(X_i, X_j)/n)_{i,j=1,\ldots,n}$) and $C_{1,n}$ which is the empirical covariance operator (i.e. $\langle f, C_{1,n} g \rangle = \frac{1}{n} \sum_{i=1}^n f(X_i)g(X_i)$). We can also define $K_{2,n}$ and $C_{2,n}$ similarly. In particular, Theorem 1 implies that $\lambda(K_{1,n}) = \lambda(C_{1,n})$ and $\lambda(K_{2,n}) = \lambda(C_{2,n})$ and $\text{tr} K_{1,n} = \text{tr} C_{1,n} = \frac{1}{n} \sum_{i=1}^n k(X_i, X_i)$ while $\text{tr} K_{2,n} = \text{tr} C_{2,n} = \frac{1}{n} \sum_{i=1}^n k^2(X_i, X_i)$.

3 General Results on Eigenvalues of Gram Matrices

We first relate sums of eigenvalues to a class of functions of type $x \mapsto \langle H_V, C_X \rangle$. This will allow us to introduce classical tools of Empirical Processes Theory to study the relationship between eigenvalues of the empirical Gram matrix and of the corresponding integral operator.

**Corollary 1.** Under Assumption 1, we have

$$\sum_{k=1}^d \lambda_k(K_1) = \max_{V \in \mathcal{V}_d} \mathbb{E}[\langle H_V, C_X \rangle] \quad \text{and} \quad \sum_{k \geq d+1} \lambda_k(K_1) = \min_{V \in \mathcal{V}_d} \mathbb{E}[\langle H_V, C_X \rangle].$$

**Proof.** The result for the sums of the largest eigenvalues follows from Theorem 2 applied to $C_1$ and Lemma 1. For the smallest ones, we use the fact that $\text{tr} C_1 = \mathbb{E}\text{tr} C_X = \sum_{k \geq 1} \lambda_k(C_1)$, and $\langle H_V, C_X \rangle + \langle H_V, C_X \rangle = \text{tr} C_X$. 

Notice that similar results hold for the empirical versions (replacing $P$ by $P_n$).

3.1 Global Approach

In this section, we obtain concentration result of the sum of the largest eigenvalues and of the sum of the lowest towards eigenvalues of the integral operator. We start with an upper bound on the Rademacher averages of the corresponding classes of functions.

**Lemma 2.**

$$\mathbb{P}_\varepsilon \left[ \frac{1}{n} \sup_{V \in \mathcal{V}_d} \sum_{j=1}^n \varepsilon_j \langle H_{V^j}, C_{X_j} \rangle \right] = \mathbb{P}_\varepsilon \left[ \frac{1}{n} \sup_{V \in \mathcal{V}_d} \sum_{j=1}^n \varepsilon_j \langle H_{V^j}, C_{X_j} \rangle \right] \leq \frac{2}{n} \text{tr} K_{2,n}$$
Proof. We use the symmetry of $\varepsilon$, Theorem 8 with $r \to \infty$ and $h = 0$, and Lemma 1.

We now give the main result of this section, which consists in data-dependent upper and lower bounds for the largest and smallest eigenvalues.

**Theorem 4.** Under Assumption 1, with probability at least $1 - 3e^{-\xi}$,

$$-M \sqrt{\frac{\xi}{2n}} \leq \sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) \leq 2 \sqrt{\frac{d}{n}} \text{tr} K_2 + 3M \sqrt{\frac{\xi}{2n}}. \tag{2}$$

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

$$-M \sqrt{\frac{\xi}{2n}} \leq \sum_{i=d+1}^{d} \lambda_i(K_{1}) - \sum_{i=d+1}^{d} \lambda_i(K_{1,n}) \leq 2 \sqrt{\frac{d}{n}} \text{tr} K_2 + 3M \sqrt{\frac{\xi}{2n}}. \tag{3}$$

Proof. We start with the first statement. Recall that

$$\sum_{i=1}^{d} \lambda_i(K_{1,n}) - \sum_{i=1}^{d} \lambda_i(K_1) = \max_{V \in \mathcal{V}_d} \langle H_V, C_{1,n} \rangle - \max_{V \in \mathcal{V}_d} \langle H_V, C_1 \rangle.$$

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

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

To prove the upper bound, we use McDiarmid’s inequality and symmetrization as in [3] along with the fact that, for a projector $U$, $\langle U, C_x \rangle \leq \|C_x\|^2 \leq M$. We conclude the proof by using Lemma 2. The lower bound is a simple consequence of Hoeffding’s inequality [10]. The second statement can be proved via similar arguments.

It is important to notice that the upper and lower bounds are different. To explain this, following the approach of [17] where McDiarmid’s inequality is applied to $\sum_{i=1}^{d} \lambda_i(K_{1,n})$ directly, we have with probability at least $1 - e^{-\xi}$,

$$-M \sqrt{\frac{\xi}{2n}} \leq \sum_{i=1}^{d} \lambda_i(K_{1,n}) - \mathbb{E} \left[ \sum_{i=1}^{d} \lambda_i(K_{1,n}) \right] \leq M \sqrt{\frac{\xi}{2n}}.$$

Then by Jensen’s inequality, symmetrization and Lemma 2 we get

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

We see that the empirical eigenvalues are biased estimators of the population ones whence the difference between upper and lower bound in (2). Note that applying McDiarmid’s inequality again would have given precisely (2), but we prefer to use the approach of the proof of Theorem 4 as it can be further refined (see next section).

$^4$ Note that one could actually apply the inequality of [7] to this quantity to obtain a sharper bound. This is in the spirit of next section.
3.2 Local Approach

We now use recent work based on Talagrand’s inequality (see e.g. [13, 2]) to obtain better 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 5. Under Assumption 1, for all \( \alpha > 0 \) and \( \xi > 0 \), with probability at least \( 1 - e^{-\xi} \),

\[
\sum_{k=1}^{d} \lambda_k(K_{1,n}) - (1 + \alpha) \sum_{k=1}^{d} \lambda_k(K_1) \leq 704(1 + \alpha^{-1}) r_d^* + \frac{M \xi (1 + \alpha + 2\alpha^{-1})}{n},
\]

where

\[
r_d^* \leq \inf_{h \geq 1} \left\{ \frac{M h}{n} + 2 \sqrt{\frac{d}{n} \sum_{j > h} \lambda_j(K_2)} \right\}.
\]

Moreover, with probability at least \( 1 - e^{-\xi} \), for all \( \alpha \in (0, 1) \),

\[
\sum_{k=1}^{d} \lambda_k(K_{1,n}) - (1 - \alpha) \sum_{k=1}^{d} \lambda_k(K_1) \geq -\frac{M \xi}{n} \left( \frac{1}{3} + \frac{1}{2\alpha} \right). \quad (5)
\]

Notice that the complexity term obtained here is always better than the one of (2) (take \( h = 0 \)). As an example of how this bound differs from (2), assume that \( \lambda_j(K_2) = O(j^{1/\alpha}) \) with \( \alpha > 1 \), then (2) gives a bound of order \( \sqrt{d/n} \), while the above Theorem gives a bound of order \( d^{1/(1+\alpha)}/n^{\alpha/(1+\alpha)} \) which is better. In the case of an exponential decay \( \lambda_j(K_2) = O(e^{-j\gamma}) \) with \( \gamma > 0 \), the rate even drops to \( \log(nd)/n \).

4 Application to Kernel-PCA

We wish to find the linear space of dimension \( d \) that conserves the maximal variance, i.e. which minimizes the error of approximating the data by their projections.

\[ V_n = \arg\min_{V \in \mathbb{R}^d} \frac{1}{n} \sum_{j=1}^{n} \| \varphi X_j - H_V(\varphi X_j) \|^2. \]

\( V_n \) 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:

\[ R_n(V) = \frac{1}{n} \sum_{j=1}^{n} \| \varphi X_j - H_V(\varphi X_j) \|^2 = P_n(\mathbb{H}_{V^\perp}, C_X). \]

\[ R(V) = \mathbb{E} \left[ \| \varphi X - H_{V\varphi} X \| \right] = P(\mathbb{H}_{V^\perp}, C_X). \]

One has \( R_n(V_n) = \sum_{i > d} \lambda_i(K_{1,n}) \) and \( R(V_d) = \sum_{i > d} \lambda_i(K_1) \).
4.1 Bound on the Reconstruction Error

We give a data dependent bound for the reconstruction error.

**Theorem 6.** Under Assumption 1, with probability at least $1 - 2e^{1/ξ}$,

$$R(V_n) \leq \sum_{i=d+1}^{n} \lambda_i(K_{1,n}) + 2 \sqrt{\frac{d}{n}} \text{tr} K_{2,n} + 3M \sqrt{\frac{ξ}{2n}}.$$ 

**Proof.** We have

$$R(V_n) - R_n(V_n) = (P - P_n) \langle H_{V_d}, CX \rangle \leq \sup_{V \in \mathcal{V}_d} (P - P_n) \langle H_{V_d}, CX \rangle.$$ 

We have already treated this quantity in the proof of Theorem 4.

In order to compare the global and the local approach, we give a theoretical bound on the reconstruction error. By definition of $V_n$, we have $R(V_n) - R(V_d) \leq 2\sup_{V \in \mathcal{V}_d} (R - R_n)[V_n]$ so that from the proof of Theorem 4 one gets

$$R(V_n) - R(V_d) \leq 4 \sqrt{\frac{d}{n}} \text{tr}(K_2) + 2M \sqrt{\frac{ξ}{2n}}.$$ 

4.2 Relative Bound

We now show that when the eigenvalues of the kernel operator are well separated, estimation becomes easier in the sense that the excess error of the best empirical $d$-dimensional subspace over the error of the best $d$-dimensional subspace can decay at a much faster rate.

The following lemma captures the key property which allows this rate improvement.

**Lemma 3.** For any subspace $V \subset \mathcal{H}_k$,

$$\text{Var} \left[ \langle H_{V_d}, CX \rangle - \langle H_{V_d}, CX \rangle \right] \leq \mathbb{E} \left[ \langle H_{V_d} - H_{V_d}, CX \rangle \right]^2,$$

and for all $V \in \mathcal{V}_d$, with $λ_d(C_1) > λ_{d+1}(C_1)$,

$$\mathbb{E} \left[ \langle H_{V_d} - H_{V_d}, CX \rangle \right]^2 \leq \frac{2\sqrt{\mathbb{E}k(X, X')}}{λ_d(C_1) - λ_{d+1}(C_1)} \mathbb{E} \left[ \langle H_{V_d} - H_{V_d}, CX \rangle \right], \quad (6)$$

where $X'$ is an independent copy of $X$.

Here is the main result of the section.

**Theorem 7.** Under Assumption 1, for all $d$ such that $λ_d(C_1) > λ_{d+1}(C_1)$, for all $ξ > 0$, with probability at least $1 - e^{1/ξ}$

$$R(V_n) - R(V_d) \leq 705 \inf_{k \geq d} \left\{ \frac{B_d k_h}{n} + 4 \sqrt{\frac{d}{n}} \sum_{j \geq h+1} \lambda_j(K_2) \right\} + \frac{ξ(22M + 27B_d)}{n},$$

where $B_d = 2\sqrt{\mathbb{E}k(X, X')/(λ_d(C_1) - λ_{d+1}(C_1))}$. 

It is easy to see that the term $\sqrt{\mathbb{E}k^2(X, X')}$ is upper bounded by $\mathbb{E}k^2(X, X)$. Similarly to the observation after Theorem 5, the complexity term obtained here will decay faster than the one of Theorem 6, at a rate which will depend on the rate of decay of the eigenvalues.

5 Discussion

Dauxois and Pousse [8] 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 [6]. However, no finite sample bounds were presented. Compared to the work of [12] and [11], 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 1 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 in [12, 11].

Comparing with [17], 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. But they do not provide two-sided inequalities and they do not introduce Rademacher averages which are natural to measure such a difference. Here we use a direct approach and provide two-sided inequalities with empirical complexity terms and even get refinements. Also, when they provide bounds for KPCA, they use a very rough estimate based on the fact that the functional is linear in the feature space associated to $k^2$. Here we provide more explicit and tighter bounds with a global approach. Moreover, when comparing the expected residual of the empirical minimizer and the ideal one, we exploit a subtle property to get tighter results when the gap between eigenvalues is non-zero.

6 Conclusion

We have obtained sharp bounds on the behavior of sums of eigenvalues of Gram matrices and shown how this entails excess risk bounds for kernel-PCA. In particular our bounds exhibit a fast rate behavior in the case where the spectrum of the kernel operator decays fast and contains a gap. These results significantly improve previous results of [17]. The formalism of Hilbert-Schmidt operator spaces over a RKHS turns out to be very well suited to a mathematically rigorous treatment of the problem, also providing compact proofs of the results. We plan to investigate further the application of the techniques introduced here to the study of other properties of kernel matrices, such as the behavior of single eigenvalues.
instead of sums, or eigenfunctions. This would provide a non-asymptotic version of results like those of [1] and of [6].

Acknowledgements

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

References

A Localized Rademacher Averages on Ellipsoids

We give a bound on Rademacher averages of ellipsoids intersected with balls using a method introduced by Dudley.

**Theorem 8.** Let \( \mathcal{H} \) be a separable Hilbert space and \( Z \) be a random variable with values in \( \mathcal{H} \). Assume \( \mathbb{E}[\|Z\|^2] \leq \infty \), and let \( C \) be the covariance operator of \( Z \). For an i.i.d. sample\(^5\) \( Z_1, \ldots, Z_n \), denote by \( C_n \) the associated empirical covariance operator. Let \( B_\alpha = \{ \|v\| \leq \alpha \}, \mathcal{E}_r = \{ \langle v, C v \rangle \leq r \} \) and \( \mathcal{E}_{\alpha, r} = \{ \langle v, C_n v \rangle \leq r \} \). We have

\[
\mathbb{E} \left[ \sup_{v \in B_\alpha \cap \mathcal{E}_{\alpha, r}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle v, Z_i \rangle \right] \leq \frac{1}{\sqrt{n}} \inf_{\|h\| \leq h} \left\{ \sqrt{hr} + \sqrt{\sum_{j=h+1}^{n} \lambda_j(C_n)} \right\}, \tag{7}
\]

and

\[
\mathbb{E} \left[ \sup_{v \in B_\alpha \cap \mathcal{E}_{r}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \langle v, Z_i \rangle \right] \leq \frac{1}{\sqrt{n}} \inf_{\|h\| \leq h} \left\{ \sqrt{hr} + \sqrt{\sum_{j=h+1}^{n} \lambda_j(C)} \right\}. \tag{8}
\]

**Proof.** We will only prove (8), the same argument gives (7). Let \( \{\Phi_i\}_{i=1}^n \) be an orthonormal basis of \( \mathcal{H} \) of eigenvectors of \( C \). Define \( p = \min\{i : \lambda_i(C) = 0\} \). If we prove the result for \( h < p \) we are done, so we assume \( h < p \). For \( v \in B_\alpha \cap \mathcal{E}_r \), we have

\[
\sum_{i=1}^{n} \varepsilon_i \langle v, Z_i \rangle = \left( \sum_{j=1}^{h} \langle v, \Phi_j \rangle^2 \right) \sum_{i=1}^{n} \varepsilon_i Z_i + \left( \sum_{j=h+1}^{n} \sqrt{\lambda_j(C)} \right) \sum_{i=1}^{n} \langle v, \Phi_j \rangle \varepsilon_j \Phi_j \]

\[
\leq \sqrt{\sum_{j=1}^{h} \frac{1}{\lambda_j(C)} \left( \sum_{i=1}^{n} \varepsilon_i Z_i \Phi_j \right)^2 + \alpha \left( \sum_{i=1}^{n} \langle v, \Phi_j \rangle \varepsilon_i Z_j \Phi_i \right)^2}.
\]

\(^5\) The result also holds if the \( Z_i \) are not independent but have the same distribution.
where we used Cauchy-Schwarz inequality and \( \langle v, C v \rangle = \sum_{i \geq 1} \lambda_i(C) \langle v, \Phi_i \rangle^2 \).

Moreover

\[
\frac{1}{n} \mathbb{E} \left( \sum_{j=1}^{n} \xi_j Z_j, \Phi_i \right)^2 = \mathbb{E} \left( \langle Z, \Phi_i \rangle^2 \right) = \langle \Phi_i, C \Phi_i \rangle = \lambda_i(C).
\]

We finally obtain (8) by Jensen’s inequality. \( \square \)

Notice that Mendelson [14] shows that these upper bounds cannot be improved.

We also need the following lemma. Recall that a sub-root function \( \psi \) is a non-decreasing non-negative function on \( [0, \infty) \) such that \( \psi(x)/\sqrt{x} \) is non-increasing.

**Lemma 4.** Under the conditions of Theorem 8, denoting by \( \psi \) the function

\[
\psi(r) := \frac{1}{\sqrt{n}} \inf_{h \geq 0} \left\{ \sqrt{n\psi} + \alpha \sqrt{\sum_{j \geq h+1} \lambda_j(C)} \right\},
\]

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

\[
r^* \leq \inf_{h \geq 0} \left\{ \frac{\epsilon^2 h}{n} + \frac{2ce}{\sqrt{n}} \sqrt{\sum_{j \geq h+1} \lambda_j(C)} \right\}
\]

**Proof.** It is easy to see that the minimum of two sub-root functions is sub-root, hence \( \psi \) as the minimum of a collection of sub-root function is sub-root. Existence and uniqueness of a solution is proved in [2]. To compute it, we use the fact that \( x \leq A\sqrt{x} + B \) implies \( x \leq A^2 + 2B \).

We finish this section with two corollaries of Theorem 8 and Lemma 4.

**Corollary 2.** Define \( W_d = \left\{ V \in \mathcal{V}_d : \mathbb{E} \left( \| H_{V, \perp} - H_{V, \perp} \|^2, C_X \right)^2 \leq r \right\} \), then

\[
\mathbb{E} \left[ \sup_{V \in W_d} \frac{1}{n} \sum_{i=1}^{n} \xi_i \left( \| H_{V, \perp} - H_{V, \perp} \|^2, C_X \right) \right] \leq \sqrt{n} \inf_{h \geq 0} \left\{ \sqrt{nh + 2d} \sqrt{\sum_{j \geq h+1} \lambda_j(K_2)} \right\}
\]

**Proof.** This is a consequence of Theorem 8 since \( \| H_{V} - H_{V} \|_{\mathcal{H}(K_2)} \leq 4d \), so that for \( V \in W_d, P_V \in B_4 \cap \mathcal{E}_r \) with \( \mathcal{E}_r = \{ v \in \mathcal{H}(K_2), \langle v, C_2v \rangle_{\mathcal{H}(K_2)} \leq r \} \). \( \square \)

**Corollary 3.** Define \( W_d = \left\{ V \in \mathcal{V}_d : \mathbb{E} \langle P_V, C_X \rangle^2 \leq r \right\} \) then,

\[
\mathbb{E} \left[ \sup_{V \in W_d} \frac{1}{n} \sum_{i=1}^{n} \xi_i \langle H_{V}, C_X \rangle \right] \leq \sqrt{n} \inf_{h \geq 0} \left( \sqrt{nh + d} \sqrt{\sum_{k \geq h+1} \lambda_k(K_2)} \right).
\]

**Proof.** Use the same proof as in Corollary 2. \( \square \)
B Proofs

Proof (of Theorem 1). Then $\Phi(X)$ is a random element of $\mathcal{H}$. By assumption, each element $h \in \mathcal{H}$ can be identified to a measurable function $x \mapsto \langle f, \Phi(x) \rangle$. Also, if $\mathbb{E}[\|\Phi(X)\|] < \infty$, $\Phi(X)$ 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 $(Th)(x) = \langle h, \Phi(x) \rangle_H$. By Cauchy-Schwarz inequality, $\mathbb{E}(h, \Phi(X))^2 \leq \|h\|^2 \mathbb{E}\|\Phi(X)\|^2$. Thus, $T$ is well-defined and continuous, thus it has a continuous adjoint $T^*$. Let $f \in L_2(P)$, then $(\mathbb{E}\|f\Phi(X)\|)^2 \leq |f|^2 \mathbb{E}\|\Phi(X)\|^2$. So, the expectation of $f(X)\Phi(X) \in \mathcal{H}$ can be defined. But for all $g \in \mathcal{H}$, $\langle T^*f, g \rangle_H = \mathbb{E}\langle g, f(X)\Phi(X) \rangle_H$ which shows that $T^*(f) = \mathbb{E}(f(X)\Phi(X))$. We now prove that $C = T^*T$ and $K = TT^*$. By the definition of the expectation, for all $h, h' \in \mathcal{H}$, $\langle h, T^*T(h') \rangle = \langle h, \mathbb{E}(\Phi(X)(\Phi(X)) \rangle = \mathbb{E}(\langle h, \Phi(X)(h'\Phi(X)) \rangle = \mathbb{E}(\langle h, \Phi(X)\rangle \langle h', \Phi(X) \rangle)$. Thus, by the uniqueness of a 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 $K = TT^*$. By singular value decomposition, it is easy to see that $\lambda(C) = \lambda(K)$ if $T$ is a compact operator. Actually, $T$ is Hilbert-Schmidt. Indeed, $\|T\|^2_{HS(H)} = \sum_{i \geq 1} |T_{ei}|^2 = \sum_{i \geq 1} \mathbb{E}[|\langle e_i, \Phi(X) \rangle|^2] = \mathbb{E}[\|\Phi(X)\|^2]$. Hence, $T$ is compact, $C$ is trace-class (tr $C = \|T\|^2_{HS(H)}$) and since tr $TT^* = \|T^*T\|$, $K$ is trace-class too.

Proof (of Theorem 5). As in the proof of Theorem 4, we have to bound $\sup_{V \in \mathcal{V}_d} \langle P_n - P \rangle \langle H_V, C_X \rangle$. We will use a slight modification of Theorem 3.3 of [2]. It is easy to see that applying Lemma 3.4 of [2] to the class of functions $\{f' = -f; f \in \mathcal{F}\}$, with the assumption $T(f') \leq -BPf'$, one obtains (with the notations of this lemma),

$$Pf' \leq \frac{K}{K + 1} Pf' + \frac{r}{\lambda BK},$$

so that under the assumptions of Theorem 3.3, one can obtain the following version of the result

$$Pf' \leq \frac{K}{K + 1} Pf' + \frac{704K}{B} \frac{r}{n} + \frac{\xi(11(k - a) + 26Bk)}{n},$$

which shows (for the initial class) that

$$P_n f \leq \frac{K + 1}{K} Pf + \frac{704(K + 1)}{B} \frac{r}{n} + \frac{\xi(11(k - a)(K + 1)/K + 26B(K + 1))}{n}.$$

We apply this result to the class of functions $x \mapsto \langle H_V, C_X \rangle$ for $V \in \mathcal{V}_d$, which satisfies $P \langle H_V, C_X \rangle^2 \leq MP \langle H_V, C_X \rangle$, and $\langle H_V, C_X \rangle \in [0, M]$, and use Lemma 4. We obtain that for all $\alpha > 0$ and $\xi > 0$, with probability at least $1 - e^{-\xi}$, every $V \in \mathcal{V}_d$ satisfies

$$P_n \langle H_V, C_X \rangle \leq (1 + \alpha)P \langle H_V, C_X \rangle + 704(1 + \alpha^{1/1})r \frac{\xi(11(1 + \alpha) + 26(1 + \alpha^{1/1}))}{n}.$$
where $r_\ast = \sqrt{\frac{2}{\pi}}$ and $M\Psi_d(r^\ast) = r^\ast$. $\Psi_d(r)$ is the sub-root function that appeared in Corollary 3. This concludes the proof. Inequality (5) is a simple consequence of Bernstein’s inequality. 

**Proof (of Lemma 3).** The first inequality is clear. For the second we start with

$$
\mathbb{E} \left[ \langle H_{V^\perp} - H_{V^\perp}, C_X \rangle^2 \right] = \langle H_{V^d} - H_V, C_2 H_{V^d} - H_V \rangle_{HS(\mathcal{H})}
\leq ||C_2|| ||H_{V^d} - H_V||^2_{HS(\mathcal{H})}
= 2||C_2||(d - \langle H_V, H_{V^d} \rangle_{HS(\mathcal{H})}).
$$

(9)

By Lemma 1, $\mathbb{E} \left[ \langle H_{V^\perp} - H_{V^\perp}, C_X \rangle \right] = \langle H_{V^d} - H_V, C_1 \rangle$. We now introduce an orthonormal basis $(\hat{f}_i)_{i=1}^d$ of $V$ and the orthonormal basis $(\phi_i)_{i=1}^d$ of the first $d$ eigenvectors of $C_1$. Moreover, we have

$$
\langle H_{V^d} - H_V, C_1 \rangle = \sum_{i=1}^d \lambda_i(C_1) - \sum_{i=1}^d \langle f_i, C_1 f_i \rangle.
$$

We decompose $f_i = \sum_{j=1}^d \langle f_i, \phi_j \rangle \phi_j + g_i$, where $g_i \in \text{span}(\phi_1, \ldots, \phi_d)^\perp$ so that

$$
\langle f_i, C_1 f_i \rangle = \sum_{j=1}^d \lambda_j(C_1) \langle f_i, \phi_j \rangle^2 + \langle g_i, C_1 g_i \rangle.
$$

Theorem 3, implies $\langle g_i, C_1 g_i \rangle \leq \lambda_{d+1}(C_1) (1 - \sum_{j=1}^d \langle f_i, \phi_j \rangle^2)$, hence we get

$$
\langle H_{V^d} - H_V, C_1 \rangle \geq \sum_{i=1}^d \lambda_i(C_1) (1 - \sum_{j=1}^d \langle f_i, \phi_j \rangle^2) - \lambda_{d+1}(C_1) (d - \sum_{i,j=1}^d \langle f_i, \phi_j \rangle^2).
$$

Using $1 - \sum_{j=1}^d \langle f_i, \phi_j \rangle^2 = ||H_{V^\perp}(\phi_i)||^2 \geq 0$, and the fact that the eigenvalues of $C_1$ are in a non-decreasing order we finally obtain

$$
\langle H_{V^d} - H_V, C_1 \rangle \geq (\lambda_d(C_1) - \lambda_{d+1}(C_1)) (d - \sum_{i,j=1}^d \langle f_i, \phi_j \rangle^2).
$$

(10)

Also we notice that $||C_2|| \leq ||C_2||_{HS(\mathcal{H})} = ||K_2||_{HS(L^2(P))}$ (by Lemma 1) and since $K_2$ is an integral operator with kernel $k^2(x, y)$, $||K_2||_{HS(\mathcal{H})} = \int k^2(x, y) dP(x) dP(y)$.

Now, Equation (1) gives $\langle H_V, H_{V^d} \rangle_{HS(\mathcal{H})} = \sum_{i,j=1}^d \langle f_i, \phi_j \rangle^2$. Combining this with Inequalities (9) and (10) we get the result. 

**Proof (of Theorem 7).** We will apply Theorem 3.3 of [2] to the class of functions $f_V : x \mapsto \left\langle H_{V^\perp} - H_{V^\perp}, C_x \right\rangle$ for $V \in \mathcal{V}_d$ and taking $V = V_y$ will give the result.

With the notations of [2], we set $T(f_V) = \mathbb{E} \left[ f_V(X)^2 \right]$ and by Lemma 3 we have $T(f_V) \leq \mathbb{E} f_V(X)^2$. Also, $f_V(x) \in [-M, M]$. Moreover, we can upper bound the localized Rademacher averages of the class $f_V$ using Corollary 2, which combined with Lemma 4 gives the result.