Journal of Machine Learning Research 7 (2006) 2303-2328 Submitted 7/05; Revised 7/06; Published 11/06

Accurate Error Bounds for the Eigenvalues of the Kernel Matrix

Mikio L. Braun MIKIO.BRAUN@FIRST.FRAUNHOFER.DE
Fraunhofer FIRST.IDA
Kekuléstr. 7
12489 Berlin, Germany

Editor: John Shawe-Taylor

Abstract

The eigenvalues of the kernel matrix play an important role in a number of kernel methods, in particular, in kernel principal component analysis. It is well known that the eigenvalues of the kernel matrix converge as the number of samples tends to infinity. We derive probabilistic finite sample size bounds on the approximation error of individual eigenvalues which have the important property that the bounds scale with the eigenvalue under consideration, reflecting the actual behavior of the approximation errors as predicted by asymptotic results and observed in numerical simulations. Such scaling bounds have so far only been known for tail sums of eigenvalues. Asymptotically, the bounds presented here have a slower than stochastic rate, but the number of sample points necessary to make this disadvantage noticeable is often unrealistically large. Therefore, under practical conditions, and for all but the largest few eigenvalues, the bounds presented here form a significant improvement over existing non-scaling bounds.

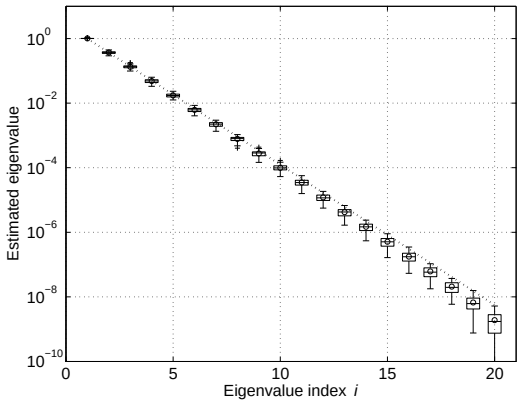
Keywords: kernel matrix, eigenvalues, relative perturbation bounds

1. Introduction

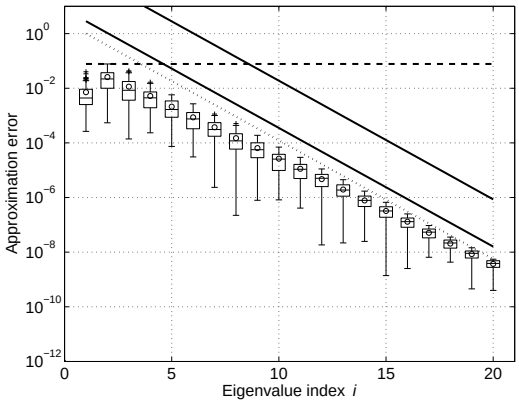
In the theoretical analysis of kernel principal component analysis (Schölkopf et al., 1998), the approximation error between the eigenvalues of the kernel matrix and their asymptotic counterparts plays a crucial role, as the eigenvalues compute the principal component variances in kernel feature space, and these are related to the reconstruction error of projecting to leading kernel principal component directions.

In order to obtain accurate bounds on the approximation error of eigenvalues, it has proven to be of prime importance to derive bounds which scale with the eigenvalue under consideration. The reason is that the approximation error scales with the eigenvalue such that the error is typically much smaller for small eigenvalues. Therefore, non-scaling bounds tend to overestimate the error for small eigenvalues as they are dominated by the largest occurring errors. Now, since smooth kernels usually display rapidly decaying eigenvalues, and such kernels are typically used in machine learning, obtaining accurate bounds in particular for small eigenvalues is highly relevant.

In an asymptotic setting, the effect that the approximation errors scale with the corresponding eigenvalues is well understood. In a paper by Koltchinskii and Giné (2000), a central limit theorem for the distribution of the approximation errors is derived. Considering only a single eigenvalue with multiplicity one, the asymptotic distribution of the properly scaled difference between approximate and true eigenvalue asymptotically approaches a normal distribution with mean zero and variance $\lambda_i^2 \operatorname{Var}_\mu(\psi_i^2)$. Thus, we would expect that the approximation error is of order $O(\lambda_i \operatorname{Std}_\mu(\psi_i^2) n^{-1/2})$,



(a) Approximate eigenvalues (box plots) and the true eigenvalues (dotted line). Note that although the box plots appear to become larger visually, due to the logarithmic scale the approximation error actually becomes small quickly.



(b) Approximation errors (box plots). For orientation, the true eigenvalues (dotted line) have also be included in the plot. The dashed line plots the smallest possible non-scaling bound on the approximation error. The solid lines plot two bounds derived in this paper, the smaller one requiring the knowledge of the true eigenfunctions.

Figure 1: Approximated eigenvalues for kernel matrices with rapidly decaying eigenvalues have an approximation error which scales with the true eigenvalue.

leading to a much smaller approximation error for small eigenvalues than for large eigenvalues (neglecting the variance of ψ_i^2 for the moment).

We are interested in deriving a probabilistic finite sample size bound to show that this effect not only occurs asymptotically, but can already be observed for small sample sizes. The following numerical example illustrates this effect: In Figure 1 we have plotted the approximate eigenvalues and the approximation errors for a kernel function with exponentially decaying eigenvalues constructed from Legendre polynomials (see Section 7.1 for details). The approximation errors scale with the true eigenvalue, and the smallest possible non-scaling bound (dashed line) overestimates the error severely for all but the first four eigenvalues. On the other hand, our bounds (solid lines) scale with the eigenvalues resulting in a bound which matches the true approximation error significantly better.

Such scaling bounds have recently been derived for tail sums of eigenvalues by Blanchard et al. (2006). There, the square root of the considered tail sum occurs in the bound, leading to bounds which correctly predict that the error for tail sums of small eigenvalues is smaller than that for tail sums starting with larger eigenvalues.

However, scaling bounds for the approximation error between individual eigenvalues, as are derived in this work, were not known so far. Note that these two settings are not interchangeable: although bounds on tail sums can be combined (more concretely, subtracted) to obtain bounds for single eigenvalues, the scaling still depends on tail sums, not single eigenvalues.

Note that the error bounds presented in this paper depend on the true eigenvalue. At first, this seems to be an undesirable feature, as this limits the practical applicability of these bounds. However, we have adopted a more theoretical approach in this work with the goal to understand

the underlying principles which permit the derivation of scaling bounds for individual eigenvalues first. In a second step, one could then use these results to construct statistical tests to estimate, for example, the overall decay rate of the eigenvalues based on these bounds. We will briefly discuss the question of constructing confidence bounds again in Section 9.

Overview

This paper is structured as follows: Section 2 contains the statements of the main results and explains the involved quantities. The actual proofs of the results can be found in Sections 3–6. Several numerical examples are discussed in Section 7. The results are compared to existing results in Section 8. Finally, Section 9 summarizes the results and suggests some directions for future work. Supplementary material can be found in the Appendix. References to the Appendix are prefixed by an "A.".

2. The Main Results

The main result consists of three parts: a basic bound, and specialized estimates for two classes of kernel functions. The basic perturbation bound deals with the approximation error based on the norms of certain error matrices. The norms of these error matrices are estimated for kernels with uniformly bounded eigenfunctions, and for kernels with bounded diagonal. Note that the scaling property is already present in the basic perturbation bound, and not a consequence of the estimates of the norms of the error matrices.

2.1 Preliminaries

We consider the following setting: Let k be a Mercer kernel on a probability space $\mathcal{X}$ with probability measure μ. This means that k can be written as

$$k(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y),$$

where $(\lambda_i)_{i\in\mathbb{N}}$ is a sequence of summable non-negative, non-increasing numbers, and $(\psi_i)_{i\in\mathbb{N}}$ is a family of mutually orthogonal unit norm functions with respect to the scalar product $(f,g) \mapsto \int_{\mathcal{X}} fg\,d\mu$. The λ_i are the eigenvalues and the ψ_i the eigenfunctions of the integral operator T_k which maps f to the function $x \mapsto \int_{\mathcal{X}} k(x,y) f(y) \mu(dy)$. Slightly abbreviating the true relationships, we will call λ_i the eigenvalues and ψ_i the eigenfunctions of k.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from μ. The (normalized) kernel matrix is the $n \times n$ matrix $\mathbf{K}_n$ with entries

$$[\mathbf{K}_n]_{ij} := \frac{1}{n} k(X_i, X_j).$$

Denote the (random) eigenvalues of $\mathbf{K}_n$ by $l_1 \geq \ldots \geq l_n \geq 0$. These eigenvalues of $\mathbf{K}_n$ converge to their asymptotic counterparts $(\lambda_i)_{i\in\mathbb{N}}$ (see, for example, the papers by Koltchinskii and Giné, 2000, and Dauxois et al., 1982, or more recently, the Ph.D. thesis of von Luxburg, 2004).

For kernels with an infinite number of non-zero eigenvalues, k can be decomposed into a degenerate kernel $k^{[r]}$ and an error function e^r given a *truncation point* r:

$$\begin{aligned} k^{[r]}(x,y) &:= \sum_{i=1}^{r} \lambda_i \psi_i(x)\psi_i(y), \\ e^r(x,y) &:= k(x,y) - k^{[r]}(x,y). \end{aligned} \tag{1}$$

Note that $k^{[r]}$ and e^r are both Mercer kernels as well. The kernel matrices induced by $k^{[r]}$ and e^r will be denoted by $\mathbf{K}_n^{[r]}$ and $\mathbf{E}_n^r$, respectively, such that $\mathbf{E}_n^r = \mathbf{K}_n - \mathbf{K}^{[r]}$.

Furthermore, let Ψ_n^r be the $n \times r$ matrix with entries

$$[\Psi_n^r]_{i\ell} = \frac{1}{\sqrt{n}} \psi_\ell(X_i).$$

The ℓth column of Ψ_n^r is thus the sample vector of the eigenfunction ψ_ℓ. Therefore, Ψ_n^r is called the *eigenfunction sample matrix*. Using Ψ_n^r, we can write $\mathbf{K}_n^{[r]} = \Psi_n^r \operatorname{diag}(\lambda_1, \ldots, \lambda_r) \Psi_n^{r\top}$ (compare Equation (3)).

The norm of a matrix $\|\mathbf{A}\|$ will always be the operator norm $\max_{\|x\|=1} \|\mathbf{A}x\|$. The ith eigenvalue of a matrix $\mathbf{A}$ in decreasing order will be denoted by $\lambda_i(\mathbf{A})$.

2.2 The Basic Perturbation Bound

The following theorem forms the basis for the finite sample size bounds which we will present. It is a deterministic bound which also holds for non-random choices of points $x_1, \ldots, x_n$.

Theorem 1 (Basic Perturbation Bound) *For $1 \le r \le n$, $1 \le i \le n$,*

$$|l_i - \lambda_i| \le \lambda_i \|\mathbf{C}_n^r\| + \lambda_r + \|\mathbf{E}_n^r\|,$$

with $\mathbf{C}_n^r = \Psi_n^{r\top} \Psi_n^r - \mathbf{I}_r$.

The bound consists of two competing terms. Let us introduce the following symbols and names for the error terms:

$$\begin{aligned} C(r,n) &= \|\mathbf{C}_n^r\|, && \text{(relative error term)} \\ E(r,n) &= \lambda_r + \|\mathbf{E}_n^r\|. && \text{(absolute error term)} \end{aligned}$$

These two terms will be bounded under different assumptions on the kernel matrix.

The relative error term $C(r,n)$ measures the amount of non-orthogonality of the sample vectors of the first r eigenfunctions of k. As $n \to \infty$, $C(r,n) \to 0$ almost surely because the scalar products between the sample vectors converge to the scalar product with respect to μ and the ψ_i form an orthogonal family of unit norm functions with respect to that scalar product. The absolute error term $E(r,n)$ measures the effect of the truncation of the kernel function. Consequently, as $r \to \infty$, $E(r,n) \to 0$. On the other hand, both terms compete against each other, because for $r \to \infty$, $C(r,n) \to \infty$, and $E(r,n)$ does in general not converge to zero as $n \to \infty$. Depending on the choice of r (see below), the bound will have a characteristic shape which first scales with λ_i while the first term dominates, until, for large i (and small eigenvalues), the bound stagnates at a certain level. Also note that if the kernel is degenerate (has only a finite number of non-zero eigenvalues), the bound will be fully relative.

We see that r has to be chosen to balance these two terms. Trivially, the best bound is obtained by minimizing with respect to r, which gives the following corollary.

Corollary 2 *For all* $1 \leq i \leq n$,

$$|l_i - \lambda_i| \leq \min_{1 \leq r \leq n} (\lambda_i C(r,n) + E(r,n)).$$

Note that the optimal choice of r can not be easily computed in general since the choice depends on the true eigenvalues and, as we will see below, the form of the bounds on C and E might not allow to write down the minimizer in closed form.

However, even suboptimal choices of r can lead to meaningful bounds and insights. For the two classes of kernel functions considered below, we will discuss three alternatives with increasing dependency on i, the index of the eigenvalue considered, and the sample size n: (i) *Keep r fixed.* This choice will typically lead to good bounds when $i < r$. However, the bound does not converge to zero as $n \to \infty$. (ii) *Choose r according to i, for example $r = i$.* This choice can be used to show that the bounds decay quickly as i increases. Again, the bound does not converge to zero. (iii) *Choose r according to n.* The goal is to let r grow slowly with n to ensure that the overall bound converges to zero, showing the asymptotic rate of the bound. This case will be discussed in more depth in Section 6.

2.3 Estimates I: Bounded Eigenfunctions

The first class of kernel functions which we consider are Mercer kernels whose eigenfunctions ψ_i are uniformly bounded. An example for this case is given by ψ_i being a sine basis on $\mathcal{X} = [0, 2\pi]$. In the following, let $\Lambda_{>r} = \sum_{i=r+1}^{\infty} \lambda_i$. Convergence of this series follows from the requirement that $(\lambda_i) \in \ell^1$, and $\lambda_i \geq 0$.

Theorem 3 (Bounded Eigenfunctions) *Let k be a Mercer kernel with bounded eigenfunctions, $|\psi_i(x)| \leq M < \infty$ for all $i \in \mathbb{N}$, $x \in \mathcal{X}$. Then, for $1 \leq r \leq n$, with probability larger than $1 - \delta$,*

$$C(r,n) < M^2 r \sqrt{\frac{2}{n} \log \frac{r(r+1)}{\delta}}, \qquad E(r,n) < \lambda_r + M^2 \Lambda_{>r}$$

Consequently, Theorem 1 implies that

$$|l_i - \lambda_i| = O(\lambda_i r \sqrt{\log r} n^{-\frac{1}{2}} + \Lambda_{>r}).$$

Since the eigenfunctions are uniformly bounded, the estimation errors involved in $C(r,n)$ can be bounded conveniently using the Hoeffding inequality uniformly over all $r(r+1)/2$ entries of $\mathbf{C}_n^r$. In particular, in contrast to the bound derived in the next section, $C(r,n)$ does not depend on the eigenvalues. Moreover, $E(r,n)$ can be bounded in a deterministic fashion in this case.

Next we discuss different choices of r as explained at the end of the previous section. For any fixed r, the bound converges to $\lambda_r + M^2 \Lambda_{>r}$ with the usual stochastic convergence rate of $O(n^{-\frac{1}{2}})$. Unless $\Lambda_{>r} = 0$, the bound will not converge to zero.

Setting $r = i$, we see that the bound converges to $\lambda_i + M^2 \Lambda_{>i} = O(\Lambda_{>i})$. This term decays quickly as $i \to \infty$. For example, if $\lambda_i = O(i^{-\alpha})$ for some $\alpha > 1$, then $\Lambda_{>i} = O(i^{1-\alpha})$, and if $\lambda_i = O(e^{-\beta i})$ for some $\beta > 0$, then $\Lambda_{>i} = O(e^{-\beta i})$ (see Theorem A.4 in the Appendix). From these considerations we see that although the bound does not vanish as $n \to \infty$ for this choice of r, the bound scales with the true eigenvalue at a rate which is only slightly slower. This error is still much smaller than that given by non-scaling error bounds, unless the sample size is very large.

Eigenvalues	rate for $r(n)$	error rate
$\lambda_i = O(i^{-\alpha})$, $\alpha > 1$	$r(n) = n^{\frac{1}{2\alpha}}$	$n^{\frac{1-\alpha}{2\alpha}}\sqrt{\log n}$
$\lambda_i = O(e^{-\beta i})$, $\beta > 0$	$r(n) = \log n^{\frac{1}{2\beta}}$	$n^{-\frac{1}{2}}(\log n)^{\frac{3}{2}}$

Table 1: Optimal rates for $r(n)$ and the resulting rates for the upper bound on the approximation error for the case of kernels with bounded eigenfunctions.

Finally, choosing r to grow with n will ensure that the bound vanishes as $n \to \infty$, but this choice will also lower the rate of $C(r,n) \to 0$ such that the resulting overall rate will be sub-stochastic. In Table 2, the optimal rates for $r(n)$ and the resulting rates for the bound are shown for the cases of polynomial and exponential decay of the true eigenvalues (see Section 6 for the proofs). We also see that in the best case (for $\alpha \to \infty$ in the polynomial case, and also for the exponential case), we obtain a rate which is slower than $O(n^{-1/2})$ only by a log-factor, which is almost negligible.

2.4 Estimates II: Bounded Kernel Function

Since the restriction of uniformly bounded eigenfunctions is rather severe, we next consider the case where the kernel function is bounded. More specifically, we will require that the *diagonal* $x \mapsto k(x,x)$ is bounded. A prominent example for such kernel functions are radial-basis kernel functions. Typically, these are kernel functions on normed spaces which are written as

$$k(x,y) = g(\|x-y\|),$$

where g is a bounded function. The choice $g(a) = \exp(-a^2/2\sigma)$ is often simply called the *rbf-kernel with kernel width* σ.

In the following theorem, two independent estimates of the error terms are presented, one based on the Bernstein inequality, and the other based on the Chebychev inequality. The reason for presenting two bounds is that while the bound based on the Bernstein inequality is asymptotically faster, the bound based on the Chebychev inequality usually gives much smaller estimates for small sample sizes since the Bernstein bound contains an $O(n^{-1})$ term which can have a prohibitively large constant if one considers small eigenvalues.

Theorem 4 (Bounded Kernel Function) *Let k be a Mercer kernel with $k(x,x) \le K < \infty$ for all $x \in \mathcal{X}$. Then, for $1 \le r \le n$, with probability larger than $1-\delta$,*

$$C(r,n) < r\sqrt{\frac{2K}{n\lambda_r}\log\frac{2r(r+1)}{\delta}} + \frac{4Kr}{3n\lambda_r}\log\frac{2r(r+1)}{\delta},$$
$$E(r,n) < \lambda_r + \Lambda_{>r} + \sqrt{\frac{2K\Lambda_{>r}}{n}\log\frac{2}{\delta}} + \frac{2K}{3n}\log\frac{2}{\delta}.$$

Consequently, by Theorem 1,

$$|l_i - \lambda_i| = O(\lambda_i\lambda_r^{-\frac{1}{2}} r\sqrt{\log r}\, n^{-\frac{1}{2}} + \Lambda_{>r} + \sqrt{\Lambda_{>r}}\, n^{-\frac{1}{2}} + \lambda_i\lambda_r^{-1} n^{-1} r\log r + n^{-1}).$$

Eigenvalues	rate for $r(n)$	error rate
$\lambda_i = O(i^{-\alpha})$, $\alpha > 1$	$r(n) = n^{\frac{1}{2+3\alpha}}$	$n^{\frac{1-\alpha}{2+3\alpha}}$
$\lambda_i = O(e^{-\beta i})$, $\beta > 0$	$r(n) = \log n^{\frac{1}{3\beta}}$	$n^{-\frac{1}{3}}(\log n)^2$

Table 2: Optimal rates for $r(n)$ and the resulting rates for the upper bound on the approximation error for bounded kernels.

The $\lambda_i\lambda_r^{-1}n^{-1}r\log r$ term in this bound can become prohibitively large for small n and small λ_r. In this case, an alternative bound gives more realistic estimates for moderately small δ:

$$C(r,n) < r\sqrt{\frac{2r(r+1)K}{2\lambda_r n\delta}}, \qquad E(r,n) < \lambda_r + \Lambda_{>r} + \sqrt{\frac{2K\Lambda_{>r}}{n\delta}}. \tag{2}$$

For these bounds,

$$|l_i - \lambda_i| = O(\lambda_i\lambda_r^{-\frac{1}{2}}r^2 n^{-\frac{1}{2}} + \Lambda_{>r} + \sqrt{\Lambda_{>r}}n^{-\frac{1}{2}}).$$

These bounds give a similar picture as those for bounded eigenfunctions. The most significant difference is the occurrence of $\lambda_r^{-1/2}$ and λ_r^{-1} in $C(r,n)$. These terms appear because the eigenfunctions of bounded kernels may have values as large as $\sqrt{K/\lambda_r}$, leading to large second moments of the eigenfunctions and large error terms in $\|\mathbf{C}_n^r\|$.

These observations are also mirrored by the asymptotic rates for the more realistic bound (2) which are summarized in Table 2 (and proved in Section 6). At most, we obtain a rate of $n^{-1/3}$. However, as we will see in Section 7.3, for small sample sizes, the resulting bounds are still much tighter than those for non-scaling bounds.

Overview of Sections 3–6

In the next four sections we will prove the main results. We have tried to make the proofs as self-contained as possible. The derivation of the basic perturbation result relies on several results from the perturbation theory of symmetric matrices which are collected in the Appendix, while the estimates of the norm of the error matrices in Section 4 and 5 rely on standard large deviation bounds. Those two sections could be informative for improving the error estimates in the presence of additional *a priori* information. Readers not interested in the technical details can safely skip to page 2318 where examples are presented and the discussion of the results is continued. That discussion does not refer to details of the proofs.

3. The Basic Perturbation Bound

In this section, we prove the basic perturbation bound (Theorem 1) which derives a bound on the perturbation in terms of the norms of certain error matrices. The proof uses two classic results on the perturbation of symmetric matrices attributed to Weyl and Ostrowski.

Recall that the kernel function k is decomposed into a degenerate kernel $k^{[r]}$ obtained by truncation, and the error term e^r (see Equation (1)). From these functions, we form the $n \times n$ matrices $\mathbf{K}_n^{[r]}$

and $\mathbf{E}_n^r$ with entries

$$[\mathbf{K}_n^{[r]}]_{ij} = \frac{1}{n} k^{[r]}(X_i, X_j), \qquad [\mathbf{E}_n^r]_{ij} = \frac{1}{n} e^r(X_i, X_j).$$

Therefore, $\mathbf{K}_n = \mathbf{K}_n^{[r]} + \mathbf{E}_n^r$, such that $\mathbf{K}_n$ is an additive perturbation of $\mathbf{K}_n^{[r]}$ by $\mathbf{E}_n^r$. The effect on individual eigenvalues of such perturbations is addressed by Weyl's theorem (Theorem A.1).

Lemma 5 *For* $1 \leq i \leq n$, $r \in \mathbb{N}$,

$$|\lambda_i(\mathbf{K}_n^{[r]}) - l_i| \leq \|\mathbf{E}_n^r\|.$$

Proof By Weyl's theorem,

$$|\lambda_i(\mathbf{K}_n^{[r]}) - \lambda_i(\mathbf{K}_n^{[r]} + \mathbf{E}_n^r)| \leq \|\mathbf{E}_n^r\|,$$

and $\mathbf{K}_n^{[r]} + \mathbf{E}_n^r = \mathbf{K}_n$. ■

For the degenerate kernel matrix $\mathbf{K}_n^{[r]}$, we will derive a multiplicative bound on the approximation error of the eigenvalues. The main step is to realize that the kernel matrix of the truncated kernel can be written as the multiplicative perturbation of the diagonal matrix containing the true eigenvalues: Recall that $[\Psi_n^r]_{i\ell} = \psi_\ell(X_i)/\sqrt{n}$ (see Section 2.1) and let $\Lambda^r = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$. Then, we can easily verify that for all $r, n \in \mathbb{N}$, $\mathbf{K}_n^{[r]} = \Psi_n^r \Lambda^r \Psi_n^{r\top}$, since

$$[\Psi_n^r \Lambda^r \Psi_n^{r\top}]_{ij} = \sum_{\ell=1}^{r} [\Psi_n^r]_{i\ell} [\Lambda^r]_{\ell\ell} [\Psi_n^r]_{j\ell} = \frac{1}{n} \sum_{\ell=1}^{r} \psi_\ell(X_i) \lambda_\ell \psi_\ell(X_j) = \frac{1}{n} k^{[r]}(X_i, X_j). \tag{3}$$

Applying Ostrowski's Theorem (Theorem A.2 and its Corollary) leads to a multiplicative bound for the eigenvalues of $\mathbf{K}_n^{[r]}$:

Lemma 6 *For* $1 \leq i \leq r \leq n$,

$$|\lambda_i(\mathbf{K}_n^{[r]}) - \lambda_i| \leq \lambda_i \|\mathbf{C}_n^r\|.$$

Proof By Ostrowski's theorem,

$$|\lambda_i(\Psi_n^r \Lambda^r \Psi_n^{r\top}) - \lambda_i(\Lambda^r)| \leq |\lambda_i(\Lambda^r)| \|\Psi_n^{r\top} \Psi_n^r - \mathbf{I}\| = \lambda_i \|\mathbf{C}_n^r\|,$$

and $\lambda_i(\Psi_n^r \Lambda^r \Psi_n^{r\top}) = \lambda_i(\mathbf{K}_n^{[r]})$, $|\lambda_i(\Lambda^r)| = \lambda_i$, since $\lambda_i \geq 0$. ■

Combining this bound for $\mathbf{K}_n^{[r]}$ with the error induced by the truncation as in Lemma 5 results in the proof of Theorem 1.

Proof (of Theorem 1) For $i \leq r$, by Lemma 6,

$$|\lambda_i(\mathbf{K}_n^{[r]}) - \lambda_i| \leq \lambda_i \|\mathbf{C}_n^r\|.$$

For $i > r$, since $\lambda_i(\mathbf{K}_n^{[r]}) = 0$,

$$|\lambda_i(\mathbf{K}_n^{[r]}) - \lambda_i| = |\lambda_i| = \lambda_i.$$

Thus,

$$|l_i - \lambda_i| \leq |l_i - \lambda_i(\mathbf{K}_n^{[r]})| + |\lambda_i(\mathbf{K}_n^{[r]}) - \lambda_i| \leq \|\mathbf{E}_n^r\| + \begin{cases} \lambda_i \|\Psi_n^{r\top} \Psi_n^r - \mathbf{I}\|, & (1 \leq i \leq r) \\ \lambda_i & (r < i \leq n). \end{cases}$$

where $|l_i - \lambda_i(\mathbf{K}_n^{[r]})|$ has been bounded using Lemma 5. Now, since $\lambda_i \leq \lambda_r$ for $r < i \leq n$,

$$|l_i - \lambda_i| \leq \lambda_i \|\mathbf{C}_n^r\| + \lambda_r + \|\mathbf{E}_n^r\|,$$

and the theorem is proven. ■

4. Estimates I: Bounded Eigenfunctions

In this section, we will prove Theorem 3. We consider the case where the eigenfunctions are uniformly bounded and there exists an $M < \infty$ such that for all $i \in \mathbb{N}$ and $x \in \mathcal{X}$,

$$|\psi_i(x)| \leq M.$$

Lemma 7 *For $1 \leq r \leq n$, with probability larger than $1-\delta$,*

$$\|\mathbf{C}_n^r\| < M^2 r \sqrt{\frac{2}{n} \log \frac{r(r+1)}{\delta}}.$$

Proof Let

$$c_{\ell m} = [\mathbf{C}_n^r]_{\ell m} = \frac{1}{n} \sum_{i=1}^{n} \psi_\ell(X_i)\psi_m(X_i) - \delta_{\ell m}.$$

Note that

$$-M^2 - \delta_{\ell m} \leq \psi_\ell(X_i)\psi_m(X_i) - \delta_{\ell m} \leq M^2 - \delta_{\ell m},$$

such that the range of $\psi_\ell(X_i)\psi_m(X_i) - \delta_{\ell m}$ is given by $2M^2$. Using Hoeffding's inequality, it follows that

$$\mathrm{P}\{|c_{\ell m}| \geq \varepsilon\} \leq 2 \exp\left(-\frac{2n\varepsilon^2}{4M^4}\right). \tag{4}$$

In order to bound $\|\mathbf{C}_n^r\|$, recall that $\|\mathbf{C}_n^r\| \leq r \max_{1\leq \ell,m\leq r} |c_{\ell m}|$ and therefore,

$$\mathrm{P}\{\|\mathbf{C}_n^r\| \geq \varepsilon\} \leq \mathrm{P}\left\{\max_{1\leq \ell,m\leq r} |c_{\ell m}| \geq \frac{\varepsilon}{r}\right\}.$$

Since $c_{\ell m} = c_{m\ell}$, there are $r(r+1)/2$ different elements in the maximum. Thus, by the union bound,

$$\mathrm{P}\left\{\max_{1\leq \ell,m\leq r} |c_{\ell m}| \geq \frac{\varepsilon}{r}\right\} \leq \sum_{\ell \geq m} \mathrm{P}\left\{|c_{\ell m}| \geq \frac{\varepsilon}{r}\right\} \leq r(r+1) \exp\left(-\frac{n\varepsilon^2}{2M^4 r^2}\right)$$

by (4). Equating the right hand side with δ and solving for ε results in the claimed inequality. ■

In order to bound the size of $\|\mathbf{E}_n^r\|$ we use a non-probabilistic upper bound.

Lemma 8 *For $r, n \in \mathbb{N}$,*

$$\|\mathbf{E}_n^r\| \leq M^2 \sum_{i=r+1}^{\infty} \lambda_i.$$

Proof Recall that the entries of $\mathbf{E}_n^r$ are constructed by evaluating the error function $e^r(x,y)$ defined in (1) on all pairs (X_i,X_j) and dividing by n. For $x,y \in \mathcal{X}$,

$$\left|\frac{1}{n}e^r(x,y)\right| = \left|\frac{1}{n}\sum_{i=r+1}^{\infty}\lambda_i\psi_i(x)\psi_i(y)\right| \leq \frac{M^2}{n}\sum_{i=r+1}^{\infty}\lambda_i.$$

Therefore,

$$\|\mathbf{E}_n^r\| \leq n\max_{1\leq i,j\leq n}\left|\frac{1}{n}e^r(X_i,X_j)\right| \leq M^2\sum_{i=r+1}^{\infty}\lambda_i.$$

■

Based on the estimates from these two lemmas, we obtain the final result:

Proof (of Theorem 3) The result is a direct consequence of Theorem 1 and plugging in the estimates from Lemma 7 and 8 for the error terms. ■

5. Estimates II: Bounded Kernel Function

In this section, we treat the case of bounded kernel functions. We have split this section into three subsections, treating the relative error term $\|\mathbf{C}_n^r\|$, the absolute error term $\lambda_r + \|\mathbf{E}_n^r\|$, and the proof of Theorem 4 separately.

Throughout this section, we assume that there exists a $K < \infty$ such that for all $x \in \mathcal{X}$, $k(x,x) \leq K$. From this condition, one can derive upper bounds on individual eigenfunctions ψ_i and the error function e^r. The following easy lemma will prove to be very useful.

Lemma 9 *For $I \subseteq \mathbb{N}$,*

$$0 \leq \sum_{i\in I}\lambda_i\psi_i^2(x) \leq k(x,x) \leq K$$

for all $x \in \mathcal{X}$, and in particular $|\psi_i(x)| \leq \sqrt{K/\lambda_i}$. Consequently, the diagonal of the error function e^r is bounded by $0 \leq e^r(x,x) \leq K$ for all $r \in \mathbb{N}$.

Proof Since all the summands $\lambda_i\psi_i^2(x)$ are positive,

$$K \geq k(x,x) = \sum_{i=1}^{\infty}\lambda_i\psi_i^2(x) \geq \sum_{i\in I}\lambda_i\psi_i^2(x) \geq 0.$$

The bound on ψ_i follows for $I = \{i\}$, and the bound on e^r for $I = \{r+1,\ldots\}$. ■

5.1 The Relative Error Term

We begin by discussing the relative error term. The first step consists in computing an upper bound on the variance of the random variables from which $\mathbf{C}_n^r$ is constructed.

Lemma 10 *For $\ell, m \in \mathbb{N}$,*

$$\begin{aligned} \mathrm{E}_\mu(\psi_\ell^2 \psi_m^2) &\le \min(K/\lambda_\ell, K/\lambda_m), \\ \mathrm{Var}_\mu(\psi_\ell \psi_m - \delta_{\ell m}) &\le \min(K/\lambda_\ell, K/\lambda_m) - \delta_{\ell m}. \end{aligned}$$

Proof By the Hölder inequality,

$$\mathrm{E}_\mu(\psi_\ell^2 \psi_m^2) \le \mathrm{E}_\mu(|\psi_\ell^2|) \sup_{x \in \mathcal{X}} |\psi_\ell^2(x)| \le \frac{K}{\lambda_\ell},$$

because $\mathrm{E}_\mu(|\psi_\ell^2|) = \|\psi_\ell\|^2 = 1$, and by Lemma 9. The same bound holds with ℓ and m interchanged which proves the first inequality.

The second inequality follows from the definition of the variance and the fact that $\mathrm{E}_\mu(\psi_i \psi_j) = \delta_{ij}$:

$$\mathrm{Var}_\mu(\psi_\ell \psi_m - \delta_{\ell m}) = \mathrm{Var}_\mu(\psi_\ell \psi_m) = \mathrm{E}_\mu(\psi_\ell^2 \psi_m^2) - (\mathrm{E}_\mu \psi_\ell \psi_m)^2 \le \min(K/\lambda_\ell, K/\lambda_m) - \delta_{\ell m}.$$

■

Lemma 11 *For $1 \le r \le n$, with probability larger than $1 - \delta$,*

$$\|\mathbf{C}_n^r\| < r \sqrt{\frac{2K}{n\lambda_r} \log \frac{r(r+1)}{\delta}} + \frac{4rK}{3n\lambda_r} \log \frac{r(r+1)}{\delta}$$

Proof Let

$$c_{\ell m} = [\mathbf{C}_n^r]_{\ell m} = \frac{1}{n} \sum_{i=1}^n \psi_\ell(X_i) \psi_m(X_i) - \delta_{\ell m}.$$

Then, for $1 \le \ell \le r$, by Lemma 9, $\sup_{x \in \mathcal{X}} |\psi_\ell(x) \psi_\ell(x)| \le K/\lambda_r$,

$$-\frac{K}{\lambda_r} - \delta_{\ell m} \le c_{\ell m} \le \frac{K}{\lambda_r} - \delta_{\ell m},$$

and the range of $c_{\ell m}$ has size $M := 2K/\lambda_r$.

We can bound the variance of $\psi_\ell(X_i)\psi_m(X_i) - \delta_{\ell m}$ using Lemma 10 as follows:

$$\mathrm{Var}_\mu(\psi_\ell \psi_m - \delta_{\ell m}) \le \frac{K}{\lambda_r} =: \sigma^2.$$

By the Bernstein inequality (see for example van der Vaart and Wellner, 1996),

$$\mathrm{P}\{|c_{\ell m}| \ge \varepsilon\} \le 2 \exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + 2M\varepsilon/3} \right).$$

In the proof of Lemma 7, we showed that

$$\mathrm{P}\{\|\mathbf{C}_n^r\| \ge \varepsilon\} \le \sum_{\ell \ge m} \mathrm{P}\left\{ |c_{\ell m}| \ge \frac{\varepsilon}{r} \right\}.$$

Thus,

$$\mathrm{P}\left\{\|\mathbf{C}_n^r\| \geq \varepsilon\right\} \leq r(r+1)\exp\left(-\frac{n(\varepsilon/r)^2}{2\sigma^2 + 2M\varepsilon/3r}\right).$$

Setting the right hand side equal to δ and solving for ε yields that with probability larger than $1-\delta$,

$$\|\mathbf{C}_n^r\| < \frac{2Mr}{3n}\log\frac{r(r+1)}{\delta} + r\sqrt{\frac{2\sigma^2}{n}\log\frac{r(r+1)}{\delta}}.$$

Substituting the values for σ^2 and M yields the claimed upper bound. ■

Corollary 12 *Alternatively, using the Chebychev inequality instead of the Bernstein inequality, one obtains that for* $1 \leq r \leq n$*, with probability larger than* $1-\delta$*,*

$$\|\mathbf{C}_n^r\| \leq r\sqrt{\frac{r(r+1)K}{2\lambda_r n\delta}}.$$

Proof By the Chebychev inequality,

$$\mathrm{P}\left\{|c_{\ell m}| \geq \varepsilon\right\} < \frac{\mathrm{Var}_\mu(\psi_\ell\psi_m - \delta_{\ell m})}{n\varepsilon^2} \leq \frac{K}{\lambda_r n\varepsilon^2}.$$

Thus,

$$\mathrm{P}\left\{\|\mathbf{C}_n^r\| \geq \varepsilon\right\} \leq \frac{r(r+1)}{2}\frac{Kr^2}{\lambda_r n\varepsilon^2}.$$

Equating the right hand side to δ and solving for ε proves the corollary. ■

5.2 The Absolute Error Term

Next, we study the properties of the random variable $\|\mathbf{E}_n^r\|$. Recall that $\mathbf{E}_n^r$ is obtained by evaluating the error function e^r on all pairs of samples (X_i, X_j). First of all, note that by the definition of e^r, the error function is itself a Mercer kernel such that $\mathbf{E}_n^r$ is positive-semidefinite for all sample realizations. Thus, we can bound $\|\mathbf{E}_n^r\| = \lambda_1(\mathbf{E}_n^r)$ by the trace of $\mathbf{E}_n^r$:

$$\|\mathbf{E}_n^r\| \leq \mathrm{tr}\,\mathbf{E}_n^r = \frac{1}{n}\sum_{i=1}^n e^r(X_i, X_i).$$

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n e^r(X_i, X_i) \to_{\text{a.s.}} \mathrm{E}(e^r(X,X)) =: t_r$$

with $X \sim \mu$, the common distribution of the X_i.

In this section, we will first compute $\mathrm{E}(e^r(X,X))$ in terms of the eigenvalues of k, and then derive a probabilistic bound on $\|\mathbf{E}_n^r\|$.

Lemma 13 *For $r \in \mathbb{N}$,*

$$t_r = \Lambda_{>r} := \sum_{i=r+1}^{\infty} \lambda_i.$$

Proof We compute t_r:

$$t_r = \int_{\mathcal{X}} e^r(x,x)\mu(dx) = \int_{\mathcal{X}} \left(\sum_{\ell=r+1}^{\infty} \lambda_\ell \psi_\ell^2(x) \right) \mu(dx) \overset{(1)}{=} \sum_{\ell=r+1}^{\infty} \lambda_\ell \int_{\mathcal{X}} \psi_\ell^2(x)\mu(dx) \overset{(2)}{=} \sum_{\ell=r+1}^{\infty} \lambda_\ell,$$

where at (1), the integration and summation commute because the function $x \mapsto K$ is an integrable majorant to the sum in parenthesis and Lebesgue's theorem, and (2) holds because $\int \psi_\ell^2(x)\mu(dx) = \|\psi_\ell\|^2 = 1$. ■

Since we are interested in the situation when t_r is much smaller than K, we will use the following bound on the variance.

Lemma 14 *For $r \in \mathbb{N}$,*

$$0 \le e^r(X,X) \le K, \qquad \mathrm{Var}(e^r(X,X)) \le K\mathrm{E}(e^r(X,X)) = Kt_r.$$

Proof The first inequality has been proven in Lemma 9. The variance can be bounded using the Hölder inequality as follows:

$$\begin{aligned}\mathrm{Var}(e^r(X,X)) &= \mathrm{E}(e^r(X,X)^2) - (\mathrm{E}e^r(X,X))^2 \\ &\le \mathrm{E}(|e^r(X,X)|)K - (\mathrm{E}e^r(X,X))^2 \le \mathrm{E}(e^r(X,X))K = t_r K.\end{aligned}$$

■

Lemma 15 *For $r,n \in \mathbb{N}$, with probability larger than $1-\delta$,*

$$\|\mathbf{E}_n^r\| < t_r + \sqrt{\frac{2Kt_r}{n}\log\frac{1}{\delta}} + \frac{2K}{3n}\log\frac{1}{\delta}.$$

Proof In order to apply Bernstein's inequality, we first have to compute the size of the range of $e^r(X_i,X_i)$ and its variance. In Lemma 14, we have proven that the range of $e^r(X_i,X_i)$ has size K, and that $\mathrm{Var}(e^r(X_i,X_i)) \le Kt_r$.

Thus, by the Bernstein inequality, with probability larger than $1-\delta$,

$$\mathrm{P}\left\{\|\mathbf{E}_n^r\| - t_r \ge \varepsilon\right\} \le \exp\left(-\frac{n\varepsilon^2}{2Kt_r + \frac{2K\varepsilon}{3}}\right).$$

Setting the right hand side equal to δ and solving for ε results in the claimed upper bound. ■

Again replacing the Bernstein inequality by the Chebychev inequality, one can show an alternative confidence bound which can be considerably smaller for moderately small δ and small n.

Corollary 16 *For $r,n \in \mathbb{N}$, with probability larger than $1-\delta$,*

$$\|\mathbf{E}_n^r\| < t_r + \sqrt{\frac{Kt_r}{n\delta}}.$$

5.3 The Final Result

We finally combine the estimates from the previous two sections to obtain the bound for bounded kernel functions.

Proof (of Theorem 4) The basic perturbation bound holds by Theorem 1. The upper bounds on $\|\mathbf{C}_n^r\|$ and $\|\mathbf{E}_n^r\|$ were derived in Lemmas 11 and 15. Finally, both estimates can be combined according from the individual bounds at confidence $\delta/2$.[1]

Using the alternative bounds from Corollary 12 and 16, one obtains the bounds from Equation (2). ■

6. Asymptotic Rates

In this section, we derive the optimal growth rates (up to logarithmic factors) for $r(n)$ such that the overall bound converges to zero. The computations have to be carried out for four different settings: kernels with bounded eigenfunctions/bounded kernels, and polynomial decay/exponential decay of eigenvalues.

6.1 Case I: Bounded Eigenfunctions

Polynomial Decay Assume that $\lambda_i = O(i^{-\alpha})$ with $\alpha > 1$. For fixed i, we wish to let r grow with n such that the approximation from Theorem 3 tends to 0. The rate is given as

$$|l_i - \lambda_i| = O(r\sqrt{\log r} n^{-\frac{1}{2}} + \Lambda_{>r}).$$

We omit the $\sqrt{\log r}$ term first. From $\lambda_i = O(i^{-\alpha})$, we obtain the following condition (see Appendix A.2 for rates concerning the tail sums $\Lambda_{>r}$):

$$r n^{-\frac{1}{2}} + r^{1-\alpha} = o(1).$$

We use the following Ansatz: $r = n^{\varepsilon}$ with $\varepsilon > 0$. Thus, we wish to find ε such that

$$n^{\varepsilon - \frac{1}{2}} + n^{\varepsilon(1-\alpha)} = o(1).$$

This condition is obviously met if $\varepsilon < 1/2$. We wish to balance the two terms in order to minimize the overall rate. This rate is attained if

$$\varepsilon - \frac{1}{2} = \varepsilon(1-\alpha) \quad \rightsquigarrow \quad \varepsilon = \frac{1}{2\alpha}.$$

Plugging in this rate shows that

$$|l_i - \lambda_i| = O(n^{\frac{1-\alpha}{2\alpha}} \sqrt{\log n}).$$

1. Let X, X' be positive random variables such that $\mathrm{P}\{X > \varepsilon\} \leq \delta$, $\mathrm{P}\{X' > \varepsilon'\} \leq \delta$. Then, $\mathrm{P}\{X + X' > \varepsilon + \varepsilon'\} \leq 2\delta$, because $\mathrm{P}\{X + X' > \varepsilon + \varepsilon'\} \leq \mathrm{P}\{X > \varepsilon \text{ or } X' > \varepsilon'\} \leq \mathrm{P}\{X > \varepsilon\} + \mathrm{P}\{X' > \varepsilon'\} \leq 2\delta$.

Exponential Decay We assume that $\lambda_i = O(e^{-\beta i})$, $\beta > 0$, such that $\Lambda_{>r} = O(e^{-\beta r})$. We are looking for the slowest rate such that $\Lambda_{>r} = O(n^{-\frac{1}{2}})$. Using the Ansatz $r = \log n^{\varepsilon}$, we obtain the condition

$$e^{-\beta \log n^{\varepsilon}} = n^{-\beta\varepsilon} = O(n^{-\frac{1}{2}}) \qquad \text{if} \qquad -\beta\varepsilon \leq -\frac{1}{2} \quad \rightsquigarrow \quad \varepsilon = \frac{1}{2\beta}.$$

Plugging this choice of ε gives the overall rate of

$$|l_i - \lambda_i| = O(n^{-\frac{1}{2}}(\log n)^{\frac{3}{2}}).$$

6.2 Case II: Bounded Kernel function

In this case, the rate is (using the bound based on the Chebychev inequality)

$$|l_i - \lambda_i| = O(\lambda_r^{-\frac{1}{2}} r^2 n^{-\frac{1}{2}} + \Lambda_{>r} + \sqrt{\Lambda_{>r}} n^{-\frac{1}{2}}).$$

Polynomial Decay Plugging in $\lambda_r = r^{-\alpha}$, $\Lambda_{>r} = r^{1-\alpha}$ (omitting the constants) gives

$$r^{2+\frac{\alpha}{2}} n^{-\frac{1}{2}} + r^{1-\alpha} + r^{\frac{1-\alpha}{2}} n^{-\frac{1}{2}}.$$

We again set $r = n^{\varepsilon}$ and obtain the three terms

$$n^{\varepsilon(2+\frac{\alpha}{2})-\frac{1}{2}} + n^{\varepsilon(1-\alpha)} + n^{\varepsilon(\frac{1-\alpha}{2})-\frac{1}{2}}.$$

First of all, the first term tells us that $\varepsilon \leq \frac{1}{4+\alpha}$, otherwise the bound diverges. Also note that of the three terms, only the first two are relevant, because the third term is always smaller than the first term. They are balanced if

$$\varepsilon\left(\frac{4+\alpha}{2}\right) - \frac{1}{2} = \varepsilon(1-\alpha) \quad \rightsquigarrow \quad \varepsilon = \frac{1}{2+3\alpha},$$

which is also smaller than $\frac{1}{4+\alpha}$ for $\alpha > 1$. Plugging this into either term shows that the resulting rate is

$$|l_i - \lambda_i| = O(n^{\frac{1-\alpha}{2+3\alpha}}).$$

Exponential Decay In this case, $\lambda_r = e^{-\beta r}$, $\Lambda_{>r} = O(e^{-\beta r})$. Therefore, the rate becomes (omitting all constants)

$$e^{\frac{\beta}{2}r} r^2 n^{-\frac{1}{2}} + e^{-\beta r} + e^{-\frac{\beta}{2}r} n^{-\frac{1}{2}}.$$

With the Ansatz $r = \log n^{\varepsilon}$, we get

$$n^{\frac{\beta\varepsilon}{2}-\frac{1}{2}} (\log n^{\varepsilon})^2 + n^{-\beta\varepsilon} + n^{-\frac{\beta\varepsilon}{2}-\frac{1}{2}}.$$

From the first term we get that $\varepsilon \leq 1/\beta$, otherwise it diverges. But for $\varepsilon \leq 1/\beta$, the third term is always smaller than the second term, such that we have to balance the first and the second term. Thus, the optimal rate is given if

$$\frac{\beta\varepsilon}{2} - \frac{1}{2} = -\beta\varepsilon \quad \rightsquigarrow \quad \varepsilon = \frac{1}{3\beta}.$$

This choice results in the overall rate of

$$|l_i - \lambda_i| = O(n^{-\frac{1}{3}}(\log n)^2).$$

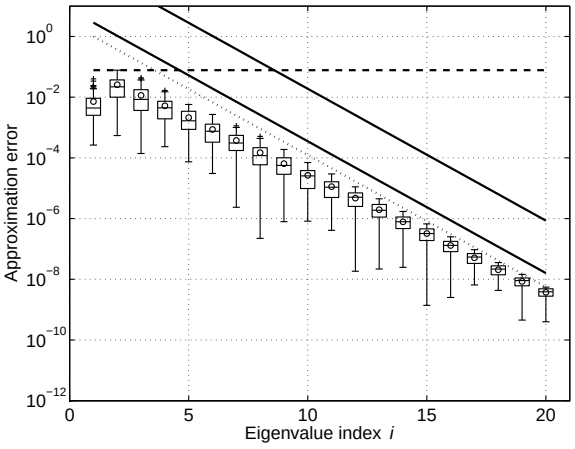


Figure 2: The example from the introduction revisited. The box plots show the distributions of the observed approximation errors for kernel matrices built from $n = 1000$ sample points over 100 re-samples. The two solid lines plot the approximation error bound derived in this work. The upper line uses the bound on $\|\mathbf{C}_n^r\|$ from Theorem 3, while the lower line uses the largest observed value of $\|\mathbf{C}_n^r\|$ on the samples in conjunction with Theorem 1, which requires knowledge of the true eigenfunctions.

7. Examples

We claim that the bounds which we have derived give realistic error estimates already for small sample sizes. In this section, we discuss several examples for both classes of kernels on numerical simulations to support our claim.

7.1 Examples for Kernels with Bounded Eigenfunctions

For the class of Mercer kernels whose eigenfunctions are uniformly bounded, we have been able to derive rather accurate finite sample size bounds. In particular, the truncation error $E(r,n)$ can be bounded in a deterministic fashion. The relative error term $C(r,n)$ scales rather moderately as $r\sqrt{\log r}$ with r, and $E(r,n)$ decays quickly, depending on the rate of decay of the eigenvalues, for both the case of polynomial and exponential decay.

Consider the following example already briefly discussed in the introduction. We construct a Mercer kernel function by specifying an orthogonal set of functions and a sequence of eigenvalues. As orthogonal functions, we use Legendre polynomials $P_n(x)$ (Abramowitz and Stegun, 1972), which are orthogonal polynomials on $[-1,1]$. We take the first 20 polynomials, and set $\lambda_i = \exp(-i)$. Then,

$$k(x,y) = \sum_{i=0}^{19} \nu_i e^{-i} P_i(x) P_i(y)$$

defines a Mercer kernel, where $\nu_i = 1/(2i+1)$ comes from normalization: $\sqrt{\nu_i}P_i$ has unit norm with respect to the probability measure induced by $\mu([a,b]) = |b-a|/2$.

For convenience, the plot from Figure 1(b) is reproduced in Figure 2. Since this kernel has only 20 non-zero eigenvalues, we obtain a purely relative bound (neglecting the round-off errors)

by setting $r = 20$. We see that the bound accurately reflects the true behavior of the approximation error.

We have also marked the smallest possible non-scaling error bound on the maximal observed approximation error. Any non-scaling error bound will necessarily be larger than this observed error with high probability. This plot illustrates the fact that it is essential for obtaining accurate estimates that the error bounds scale with the considered eigenvalue. A non-scaling bound will overestimate the error of smaller eigenvalues significantly.

The plot might suggest that our bound is actually worse for the first few large eigenvalues, but note that the dashed line is only a lower bound to any non-scaling error bound, and actual error bounds will typically be much larger.

Next, we turn to a non-degenerate kernel. In Figure 3, some examples are plotted for the sine-basis kernel, defined as follows. The eigenfunctions are given by

$$\psi_i(x) = \sqrt{2}\sin(ix/2), \qquad i \in \mathbb{N},$$

which form an orthogonal family of functions on the Hilbert space of functions defined on $[0, 2\pi]$ with the scalar product $(f,g) \mapsto \int_0^{2\pi} f(x)g(x)dx/2\pi$. These functions are uniformly bounded by $\sqrt{2}$.

Since we cannot write down the resulting kernel given some choice of eigenvalues in closed form, we truncate the expansion to the first 1000 terms, resulting in a negligible difference to the true kernel function. The resulting kernel function for different choices of eigenvalues are plotted in Figure 3(a). In Figures 3(b)–(d), three such examples are plotted, two for polynomially decaying eigenvalues, and one for exponentially decaying eigenvalues. We plot the bound for different choices of r and see that, with increasing r, the absolute error term becomes smaller such that the bound for small eigenvalues also becomes smaller while, at the same time, the bound for larger eigenvalues becomes larger.

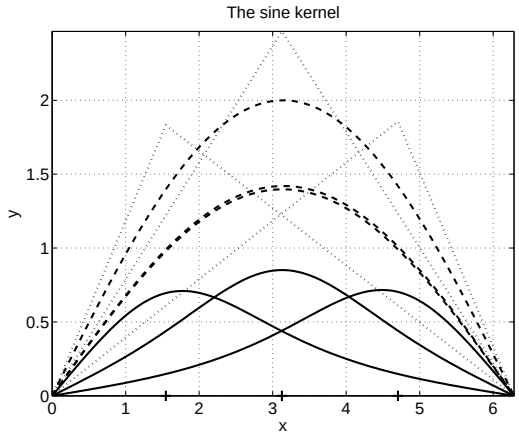
In Figure 3(d), it appears that the bound is actually smaller than the observed eigenvalues. This effect is due to the finite precision arithmetic used in the computations. These rounding errors effectively lead to an additive perturbation of the kernel matrix, which in turn results in an additive perturbation of the eigenvalues of the same magnitude. An interesting observation is that although our bounds fail to be purely relative in the general case, numerically computed eigenvalues will always display a stagnation of small eigenvalues at a certain level due to round-off errors as well. Thus, for numerically computed eigenvalues, fully relative approximation errors are not possible.

7.2 Examples for Kernels with Bounded Kernel Functions

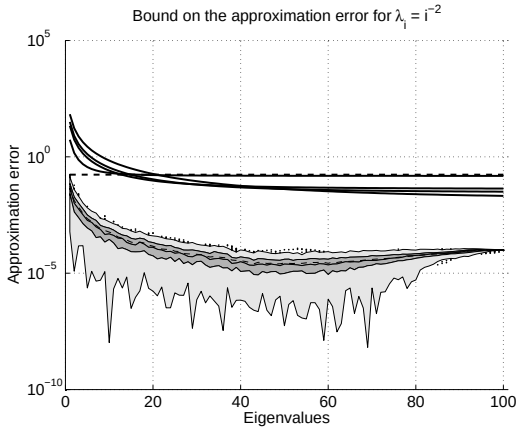
The second class of kernel functions are kernels with bounded diagonal. This class includes the important radial basis function kernels (rbf-kernels). In this case, the eigenfunctions can in principle grow unboundedly as the eigenvalues become smaller, leading to considerably larger error estimates. The most important difference to the previous case is that the relative error term depends on the eigenvalues themselves and scales with the factor $1/\sqrt{\lambda_r}$. Therefore, having smaller eigenvalues can lead to a much larger relative error term (which will nevertheless ultimately decay to zero).

The example we will consider is designed to display this slow rate of convergence. It is well-known that Bernoulli random variables maximize the variance among all bounded random variables taking values in $[0,1]$. We thus consider the following kernel: Let $(A_i)_{i=1}^{\infty}$ be a partition of $\mathcal{X}$ with $\mu(A_1) \geq \mu(A_2) \geq \ldots \geq 0$. Then, set

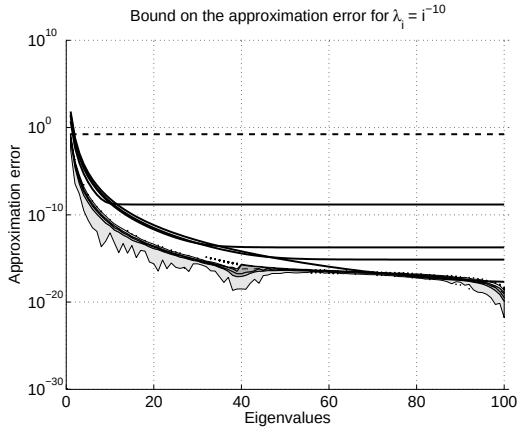
$$\lambda_i = \mu(A_i), \qquad \psi_i(x) = \frac{1}{\sqrt{\lambda_i}} 1_{A_i}(x). \tag{5}$$



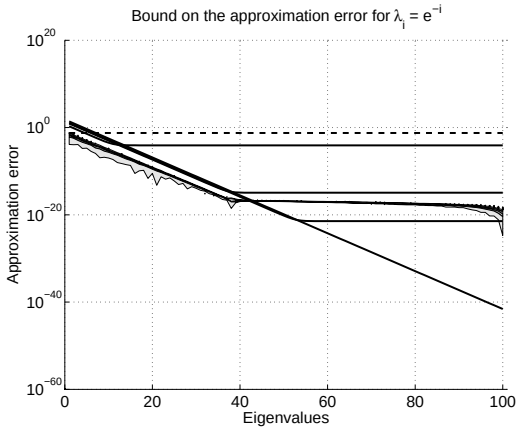
(a) The sine-kernel for different decay rates and at the three points marked by a plus-sign. The decay rates are $\lambda_i = i^{-2}$ (dotted lines), $\lambda_i = i^{-10}$ (dashed lines), $\lambda_i = e^{-i}$ (solid lines). Note that the smoothness depends on rate of decay.



(b) For quadratic decay, the bounds are only slightly better than the best possible non-scaling bound. (Shaded areas correspond to quartile ranges, similar to box plots. See explanation in figure caption.)



(c) For faster polynomial decay, the bounds are much more accurate than the best possible non-scaling bound for small eigenvalues.



(d) For exponential decay, the bounds are much more accurate than the best possible non-scaling bound as observed on the data. In fact, the actual approximation error becomes even larger than the bound due to finite precision arithmetics starting with eigenvalue λ_{40}.

Figure 3: The sine-kernel example. We consider the decay rates $\lambda_i = i^{-2}$, $\lambda_i = i^{-10}$, and $\lambda_i = e^{-i}$. In (a), some example kernel functions are plotted. In (b)–(d), we plot approximation errors as observed over 100 re-samples of $n = 200$ points uniformly sampled from $[0, 2\pi]$, and the bound for $r \in \{10, 35, 50, 100\}$ for confidence $\delta = 0.05$. The dashed line plots the best achievable non-scaling error bound. The distribution of the observed approximation errors is illustrated by differently shaded areas similar to box plots: dark gray area shows lower to upper quartile range, while light gray area shows data points which lie in 1.5 times the interquartile range. Points beyond that are plotted as small dots.

This defines a Mercer kernel

$$k(x,y) = \sum_{i=1}^{\infty} 1_{A_i}(x) 1_{A_i}(y) = \begin{cases} 1 & \text{if there exists an } i \text{ such that } x,y \in A_i, \\ 0 & \text{else.} \end{cases}$$

Note that the matrix $\mathbf{C}_n^r$ is always diagonal for this choice of basis functions because

$$\psi_i(x)\psi_j(x) = \frac{1}{\sqrt{\lambda_i \lambda_j}} \delta_{ij}.$$

Thus, we obtain a slightly improved bound over the one from Corollary 12 because $\|\mathbf{C}_n^r\| = \max_{1 \le i \le r} |[\mathbf{C}_n^r]_{ii}|$ since $\mathbf{C}_n^r$ is diagonal. Then,

$$\mathrm{P}\{\|\mathbf{C}_n^r\| \ge \varepsilon\} \le r \max_{1 \le i \le r} \mathrm{P}\{|[\mathbf{C}_n^r]_{ii}| \ge \varepsilon\} \le \frac{r}{\lambda_r n \varepsilon^2},$$

and consequently, with probability larger than $1-\delta$,

$$\|\mathbf{C}_n^r\| < \sqrt{\frac{r}{\lambda_r n \delta}}. \tag{6}$$

Figure 4 plots the bound for this example. Again, the kernel function cannot be computed in closed form, and we truncate to the first 1000 terms. We plot two different bounds, the bound from (6), and the general result from Theorem 4. Note that the error does not fluctuate after eigenvalue λ_{20}. The reason is that λ_i is so small that not a single point has hit A_i in the sample of $n = 1000$ points; the kernel is effectively degenerate and the approximate eigenvalues beyond l_{20} are equal to zero. In this case, the approximation error is equal to the true eigenvalue which explains the exponential decay. Note though, that for the first 20 eigenvalues, the (slower) rate is actually matched by the bound.

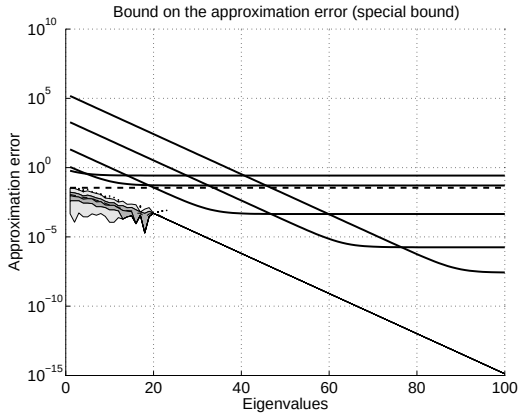
Over all, compared with the examples for bounded kernel functions, the bounds are considerably less tight, but they still correctly predict the scaling of the approximation error with regard to the true eigenvalue.

7.3 Comparisons with a Non-Scaling Hoeffding-Type Bound

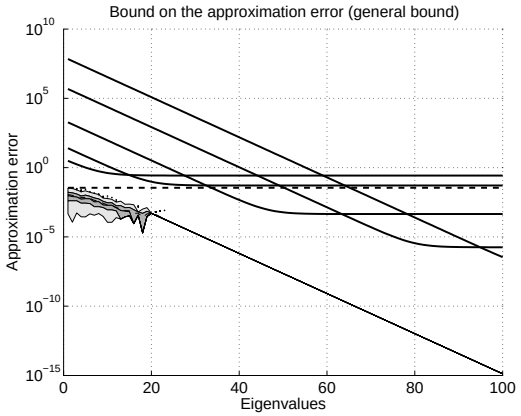
Finally, we would like compare our bound numerically against a non-scaling Hoeffding-type bound. As discussed in Section 6, while the bounds presented in this paper are more accurate for small eigenvalues, the overall rate as $n \to \infty$ is slower than the usual stochastic rate $O(n^{-1/2})$. To illustrate that the bounds can nevertheless be much more accurate even for moderately small sample sizes, we will compare our bounds against a Hoeffding-type bound which does not scale with the eigenvalue under consideration.

We face the problem of choosing an appropriate non-scaling bound. Such bounds exist, but only for tail sums of eigenvalues (see the papers by Shawe-Taylor et al., 2005, and Blanchard et al., 2006, and the discussion in Section 8). However, the full complexity of these bounds is not really necessary for the illustrative purposes we have in mind. In Theorem 6 of the paper by Shawe-Taylor et al. (2005), there is a bound on the concentration of single eigenvalues around their mean: with probability larger than $1-\delta$,

$$|l_i - \mathrm{E}(l_i)| \le K^2 \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}.$$



(a) The bound based on the estimate from Equation (6) for $C(r,n)$ which was specifically derived for this example.



(b) The bound using the general result from Theorem 4. Note that although the bound becomes very large for large r, the minimum over all bounds is the final bound on the approximation error.

Figure 4: The indicator function example (see Equation (5)). This kernel has maximal variance given the constraint that the resulting kernel function is bounded. We consider eigenvalues $\lambda_i = e^{i/3}/Z$, where Z is the normalization constant. The sample size is $n = 1000$, and the bounds are computed for confidence $\delta = 0.05$. The solid lines are the bounds for $r \in \{10, 50, 200, 500\}$, while the dashed line shows the best possible non-scaling error bound.

This bound has the required asymptotic decay rate of $O(n^{-1/2})$. Terms similar to this bound also occur in the more complex bounds on tail sums of eigenvalues, albeit with larger constants. We will therefore pretend that this is an overly optimistic guess of the approximation error and compare our bounds to it.

We are particularly interested in the question if the Hoeffding-type bound, due to its better asymptotic rate, quickly compensates for its non-scaling constant and becomes as small as our bound. We therefore compare the bounds for sample sizes up to $n = 10000$. Figure 5 plots the Hoeffding-type bound with the bound derived in this work. For these plots, since the eigenvalues are known, the optimal r has been computed by numerically minimizing the bound. The optimal r with respect to i have been plotted in Figure 5(d). For polynomial decay of rate $\lambda_i = i^{-2}$, the bounds are clearly inferior to the Hoeffding-type bounds and one can also clearly see that the overall rate is sub-stochastic. However, for faster decay rates, the bounds for smaller eigenvalues are clearly superior, also demonstrating that the number of samples necessary to yield a comparably small bound using the Hoeffding-type bound is fairly large.

So far, we have compared the bounds only for the errors of individual eigenvalues. Let us now compare the bounds for sums of eigenvalues. As we will discuss in Section 8, there exist bounds which directly deal with tail sums of eigenvalues and are more accurate in this case. However, it is instructive to derive a rough estimate for tail sums based on our bounds. We start with bounding the difference between tail sums by summing the individual bounds:

$$\left| \sum_{i=r+1}^{n} l_i - \sum_{i=r+1}^{\infty} \lambda_i \right| \leq \sum_{i=r+1}^{n} |l_i - \lambda_i| + \Lambda_{>n+1} \leq \sum_{i=r+1}^{n} (\lambda_i C(r,n) + E(r,n)) + \Lambda_{>n+1}.$$

Let us roughly estimate the size of the resulting bound. The key to obtain a good estimate lies in choosing a different r for each i. Let us set $r = i$. Then, omitting constants and using the bound for bounded kernel functions which is based on the Chebychev inequality, we get that

$$\begin{aligned} &\sum_{i=r+1} \left(\sqrt{\lambda_i} i^2 n^{-\frac{1}{2}} + \Lambda_{>i} + \sqrt{\Lambda_{>i}} n^{-\frac{1}{2}} \right) + \Lambda_{>n+1} \\ &= n^{-\frac{1}{2}} \left(\sum_{i=r+1}^{n} i^2 \sqrt{\lambda_i} + \sum_{i=r+1}^{n} \sqrt{\Lambda_{>i}} \right) + \sum_{i=r+1}^{n} \Lambda_{>i} + \Lambda_{>n+1}. \end{aligned}$$

Let us consider these tail sums for $i \to \infty$. All of these sums converge if $\alpha > 6$, because

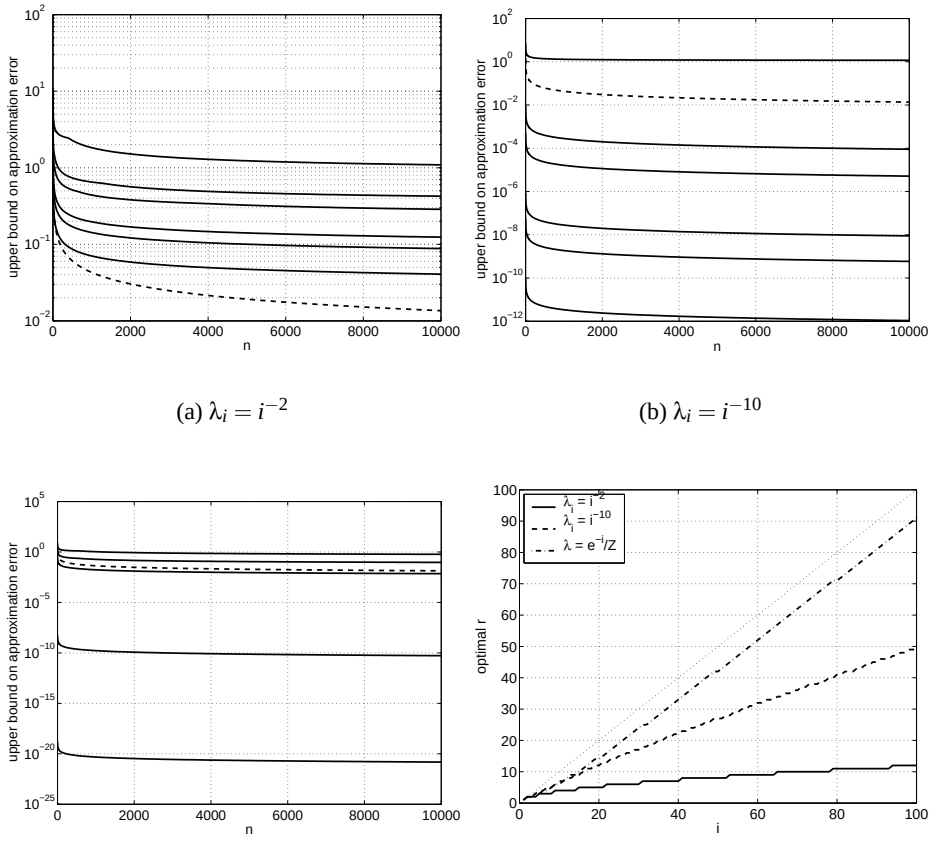
$$i^2 \sqrt{\lambda_i} = O(i^{2-\frac{\alpha}{2}}) = O(i^{-1}), \quad \sqrt{\Lambda_{>i}} = O(i^{\frac{1-\alpha}{2}}) = O(i^{-2\frac{1}{2}}), \quad \Lambda_{>i} = O(i^{1-\alpha}) = O(i^{-5}).$$

Thus, for large i, the bound on the tail sums actually becomes small, giving more accurate bounds than those obtainable by a non-scaling bound.

In Figure 6, we again compare the bound derived in this work against a Hoeffding-type bound for tail sums of eigenvalues. We do not use the rough estimate derived above, but sum the individual approximation error bounds selecting the optimal r in each case. Again we see that these bounds give much smaller estimates than the non-scaling Hoeffding-type bound.

8. Related Work

The bounds presented in this work are the first finite sample size bounds for single eigenvalues which scale with the eigenvalue under consideration. These results contribute to the already existing body of work which we briefly review in this chapter.



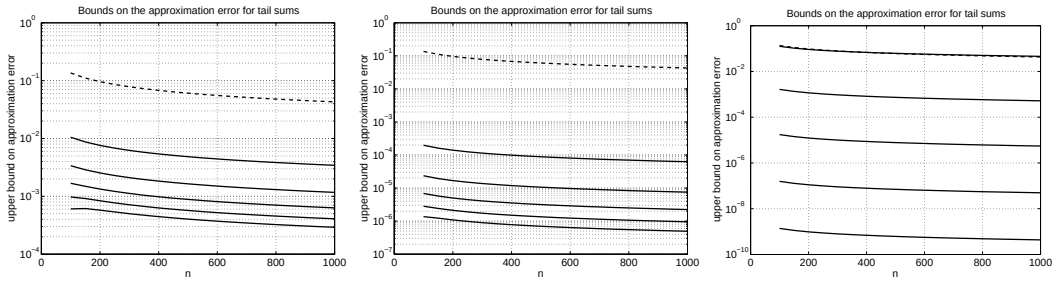
(a) $\lambda_i = i^{-2}$

(b) $\lambda_i = i^{-10}$

(c) $\lambda_i = e^{-i}/Z$ (normalization)

(d) optimal r

Figure 5: Upper bounds with the best choice of r (solid lines) compared with a Hoeffding-type non-scaling $O(n^{-\frac{1}{2}})$ bound (dashed line) for single eigenvalues. The bounds are plotted for eigenvalues $\lambda_1, \lambda_5, \lambda_{10}, \lambda_{50}, \lambda_{100}$, and for the cases of polynomial decay, also for λ_{500}. The truncation point r has been chosen optimally by explicitly minimizing the bound. The confidence was $\delta = 0.05$.



(a) $\lambda = i^{-2}$ (b) $\lambda = i^{-10}$ (c) $\lambda = e^{-i}/Z$ (normalization)

Figure 6: Upper bounds for tail sums (solid lines) compared with a Hoeffding-type non-scaling $O(n^{-\frac{1}{2}})$ bound (dashed line). The tail sums are plotted for $r \in \{10,20,30,40,50\}$. The confidence was set to $\delta = 0.5$.

The asymptotic setting is addressed for example in Dauxois et al. (1982) and Koltchinskii and Giné (2000) where central limit type results for the limit distributions of the eigenvalues are derive. The finite sample setting has been addressed more recently, in particular in Shawe-Taylor et al. (2005) and Blanchard et al. (2006).

The paper by Shawe-Taylor et al. (2005) discusses several aspects of the relation between the approximate eigenvalues and their asymptotic counterparts. These also include concentration inequalities relating the approximate eigenvalues to their expectations (similar inequalities can also be found in the Ph.D. thesis of Mika, 2002). Finally, Theorem 1 and 2 of that paper provide finite sample size bounds on the approximation error for tail sums of the eigenvalues. However, these bounds do not scale with the size of the eigenvalues, leading to the already discussed overestimation of the true approximation error in particular for small eigenvalues.

These results are further refined and extended in the paper by Blanchard et al. (2006). In particular, non-scaling bounds are derived exhibiting fast convergence rates, as well as scaling bounds for tail sums of eigenvalues. As already explained, the latter bounds are particularly important for obtaining accurate estimates for small eigenvalues.

Compared to the results presented in this work, all of these results are either dealing with the asymptotic setting or, in the case of finite sample size bounds, are non-scaling or only deal with tail sums of eigenvalues. Obtaining accurate bounds, in particular bounds which scale with the eigenvalue under consideration, was an open problem so far (see the comments below Theorem 4.2 in the paper by Blanchard et al., 2006). Note that these two problems are not interchangeable: While it is possible to construct bounds for single eigenvalues from bounds of tail sums by subtracting bounds for neighboring indices, and also vice versa by summing up bounds, the resulting scaling factors will not match the quantity under consideration.

From a technical point of view, the approach taken in this work and the one by Blanchard et al. (2006) also differ considerably. While the analysis in the latter is carried out in abstract Hilbert spaces, in this work, the analysis is based in the finite dimensional domain, having the potential advantage that the arguments are somewhat more elementary. However, one could suspect that the absolute terms occurring in our bounds are an artifact of the more elementary approach (in particular

since these terms are a side-effect of the truncation of the kernel matrix). Then, a more abstract approach might be able to obtain fully relative bounds. Note however, that Ostrowski's inequality does not easily extend to the high-dimensional case, as the convergence of the error matrix $\mathbf{C}_n^r$ scales with the dimension of the finite-dimensional case. At any rate, these questions are interesting possible direction for future research.

9. Conclusion and Outlook

We have derived finite sample size bounds on the error between single eigenvalues of the kernel matrix and their asymptotic limits. These bounds scale with the eigenvalue under consideration leading to significantly more accurate bounds for single eigenvalues than previously known bounds in particular for small eigenvalues and small sample sizes. Also for fairly large sample sizes, the bounds can still be superior to existing bounds since the number of samples necessary to make existing non-scaling bounds competitive can be unrealistically large.

For future work, we would like to suggest three possibilities. (i) If additional information on the kernel, or the probability distribution is available, the bounds on the norms of the error matrices could be improved leading to more accurate bounds.

(ii) Note that the resulting bounds require the knowledge of the true eigenvalues. From a theoretical point of view, this approach is acceptable, because we were specifically interested in approximation errors of small eigenvalues, and this assumption is codified into the knowledge of the true eigenvalues. In practical situations, however, one might be interested in obtaining a confidence bound without knowledge of the eigenvalues. This means that one has to derive some property of the true eigenvalues, for example, their rate of decay. Statistical tests could be constructed to this means based on the bounds presented here. Then, the bounds presented in this work predict that the estimated eigenvalues decay at the same rate giving confidence bounds which scale at the correct rate.

(iii) Finally, since the basic perturbation bound also holds for non-random choices of points, the result could be applied in the analysis of the numerical approximation of integral equations. The norms of the error matrices would then be bounded using approximation theory.

Acknowledgments

A substantial part of this research was performed while the author was with the Computer Vision and Pattern Recognition group headed by Professor Joachim Buhmann (now ETH Zürich) at the University of Bonn (see also Braun, 2005). The author wishes to thank Stefan Harmeling, Gilles Blanchard, Joachim Buhmann, Michael Clausen, Motoaki Kawanabe, Tilman Lange, Klaus-Robert Müller, Peter Orbanz, and Axel Munk for fruitful discussion and helpful comments. The author would also like to thank the anonymous referees whose comments helped to improved the paper further, and who made the author aware of an error contained in an earlier version of the paper. In the course of fixing this error, the estimate of the absolute error term could be significantly improved. The author also acknowledges funding received on behalf of DFG grant #Buh 914/5, and BMBF grant 16SV2231.

Appendix A. Supplementary Results

In this section, we collect some supplementary results for reference, which are used in the main text.

A.1 Perturbation of Hermitian Matrices

We use two classical results on the perturbation of eigenvalues for Hermitian matrices.

Theorem A.1 (Weyl) *(Horn and Johnson, 1985, Theorem 4.3.1) Let* $\mathbf{A}, \mathbf{E}$ *be Hermitian* $n \times n$ *matrices. Then, for each* $1 \leq i \leq n$*,*

$$\lambda_i(\mathbf{A}) + \lambda_n(\mathbf{E}) \leq \lambda_i(\mathbf{A}+\mathbf{E}) \leq \lambda_i(\mathbf{A}) + \lambda_1(\mathbf{E}).$$

This implies that

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{A}+\mathbf{E})| \leq \|\mathbf{E}\|.$$

Theorem A.2 (Ostrowski) *(Horn and Johnson, 1985, Theorem 4.5.9., Corollary 4.5.11) Let* $\mathbf{A}$ *be a Hermitian* $n \times n$ *matrix, and* $\mathbf{S}$ *a non-singular* $n \times n$ *matrix. Then, for* $1 \leq i \leq n$*, there exist non-negative real* θ_i *with* $\lambda_n(\mathbf{S}\mathbf{S}^*) \leq \theta_i \leq \lambda_1(\mathbf{S}\mathbf{S}^*)$ *such that*

$$\lambda_i(\mathbf{S}\mathbf{A}\mathbf{S}^*) = \theta_i \lambda_i(\mathbf{A}).$$

Consequently,

$$|\lambda_i(\mathbf{S}\mathbf{A}\mathbf{S}^*) - \lambda_i(\mathbf{A})| \leq |\lambda_i(\mathbf{A})| \|\mathbf{S}^*\mathbf{S} - \mathbf{I}\|.$$

For the case of non-square $\mathbf{S}$, the same result holds as can be shown by extending either $\mathbf{S}$ or $\mathbf{A}$ with zeros until both matrices are square and have the same size and by a continuity argument to extend Ostrowski's theorem to singular $\mathbf{S}$ (Horn and Johnson, 1985, p. 224).

Corollary A.3 *Ostrowski's theorem also holds if* $\mathbf{S}$ *is a (non-square)* $n \times m$ *matrix.*

A.2 Asymptotics of Infinite Sums

For convenience, we collect two elementary computations to estimate the asymptotic rates of tail sums of sequences with polynomial and exponential decay.

Theorem A.4 *For* $\alpha > 1$ *and* $\beta > 0$*,*

$$\sum_{i=r+1}^{\infty} i^{-\alpha} \leq \frac{r^{1-\alpha}}{\alpha - 1} = O(r^{1-\alpha}), \qquad \sum_{i=r+1}^{\infty} e^{-\beta i} = \frac{e^{-\beta(r+1)}}{1 - e^{-\beta}} = O(e^{-\beta r}).$$

To prove these two rates, note that

$$\sum_{i=r+1}^{\infty} i^{-\alpha} \leq \int_{r+1}^{\infty} (x-1)^{-\alpha} dx = \int_{r}^{\infty} x^{-\alpha} dx = \frac{x^{1-\alpha}}{1-\alpha}\bigg|_{r}^{\infty} = 0 - \frac{r^{1-\alpha}}{1-\alpha} = \frac{r^{1-\alpha}}{\alpha - 1}.$$

Furthermore, since $\sum_{i=r+1}^{\infty} (e^{-\beta})^i$ is the tail of a geometric series,

$$\sum_{i=r+1}^{\infty} e^{-\beta i} \leq \frac{1}{1-e^{-\beta}} - \frac{1 - \left(e^{-\beta}\right)^{r+1}}{1-e^{-\beta}} = \frac{\left(e^{-\beta}\right)^{r+1}}{1-e^{-\beta}}.$$

References

Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*, chapter 22, "Legendre Functions", and chapter 8, "Orthogonal Polynomials", pages 331–339, 771–802. Dover, New York, 1972.

Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 2006. (to appear, published online March 30, 2006).

Mikio L. Braun. *Spectral Properties of the Kernel Matrix and their Relation to Kernel Methods in Machine Learning*. PhD thesis, University of Bonn, Germany, 2005. Available electronically at http://hss.ulb.uni-bonn.de/diss_online/math_nat_fak/2005/braun_mikio.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12:136–154, 1982.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

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

Sebastian Mika. *Kernel Fisher Discriminants*. PhD thesis, Technische Universität Berlin, December 2002.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear compoment analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

John Shawe-Taylor, Christopher K. I. Williams, Nello Christianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522, July 2005.

Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, 1996.

Ulrike von Luxburg. *Statistical Learning with Similarity and Dissimilarity Functions*. PhD thesis, Technische Universität Berlin, November 2004.