Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2008-052 August 19, 2008
CBCL-274

A Note on Perturbation Results for Learning Empirical Operators

Rosasco, Lorenzo, Belkin, Mikhail, and De Vito, Ernesto

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

A NOTE ON PERTURBATION RESULTS FOR LEARNING EMPIRICAL OPERATORS

L. ROSASCO, M. BELKIN, E. DE VITO

ABSTRACT. A large number of learning algorithms, for example, spectral clustering, kernel Principal Components Analysis and many manifold methods are based on estimating eigenvalues and eigenfunctions of operators defined by a similarity function or a kernel, given empirical data. Thus for the analysis of algorithms, it is an important problem to be able to assess the quality of such approximations. The contribution of our paper is two-fold:
1. We use a technique based on a concentration inequality for Hilbert spaces to provide new much simplified proofs for a number of results in spectral approximation.
2. Using these methods we provide several new results for estimating spectral properties of the graph Laplacian operator extending and strengthening results from [26].

1. INTRODUCTION

A broad variety of methods for machine learning and data analysis from Principal Components Analysis (PCA) to Kernel PCA to Laplacian-based spectral clustering and manifold methods, rely on estimating eigenvalues and eigenvectors of certain data-dependent matrices. In many cases these matrices can be interpreted as empirical versions of underlying integral operators or closely related objects, such as continuous Laplace operators. Thus establishing connections between empirical operators and their continuous counterparts is essential to understanding these algorithms. In recent years there has been a considerable amount of theoretical work on building these connections. One of the first studies of this problem was conducted in [16, 15], where the authors consider integral operators defined by a kernel. This investigation was continued in [19, 20] (see also references therein). Convergence of Kernel PCA was addressed in [22] and refined in [6, 28]. Convergence of the graph Laplacian in various settings was addressed in [4, 11, 12, 23, 17, 10, 5, 26]. In particular, the last two papers considered spectral convergence.

This paper proposes a method based on analyzing the empirical operators in the Hilber-Schmidt norm and using concentration inequalities in Hilbert spaces of operators. This technique together with some standard perturbation theory allows us to derive a number of results on spectral convergence in an exceptionally simple way. We note that the approach using concentration inequalities in a Hilbert space has already been proved useful for analyzing supervised kernel algorithms, see [9, 27, 3, 24].

We start with introducing the necessary mathematical objects in Section 2. We introduce basic operator and spectral theory and discuss concentration inequalities in Hilbert spaces. This technical summary section aims at making this paper self-contained and provide the reader with a (hopefully useful) overview of the needed tools and results.

In Section 3, we study the spectral properties of kernel matrices generated from random data. We study concentration of operators obtained by an out-of-sample extension using the kernel function and obtain considerably simplified derivations of several existing results on eigenvalues and eigenfunctions. We expect that these techniques will be useful in analyzing algorithms requiring spectral convergence. In fact, in Section 4, we apply these methods to prove convergence of eigenvalues and eigenvectors of the asymmetric graph Laplacian defined by a fixed weight function. We provide stronger convergence results than results than [26], which, to the best of our knowledge, is the only other paper to consider the problem so far. Similar results to the one we prove here have been independently derived by Smale and Zhou and are contatined in the following preprint [25].

2. Notation and preliminaries.

In this section we will discuss various preliminary results necessary for the further development.

Operator theory. We first recall some basic notions from the operator theory (see, e.g. [18]). In the following we let $A : \mathcal{H} \to \mathcal{H}$ be a linear operator, where $\mathcal{H}$ is a (in general complex) Hilbert space with scalar product (norm) $\langle \cdot, \cdot \rangle$ ($\|\cdot\|$) and $(e_j)_{j\geq 1}$ a Hilbert basis in $\mathcal{H}$. We often use the notation $j \geq 1$ to denote a sequence or a sum from 1 to p where p can be infinite. The set of bounded operators on $\mathcal{H}$ is a Banach space with respect to the operator norm $\|A\| = \sup_{\|f\|=1} \|Af\|$. If A is a bounded operator, we let A^* is adjoint, which is also a bounded operator with $\|A^*\| = \|A\|$.

A bounded operator A is Hilbert-Schmidt if $\sum_{j\geq 1} \|Ae_j\|^2 < \infty$ for some (any) Hilbert basis $(e_j)_{j\geq 1}$. The space of Hilbert-Schmidt operators is also Hilbert space (a fact which will be key in our development) endowed with the scalar product $\langle A, B\rangle_{HS(\mathcal{H})} = \sum_j \langle Ae_j, Be_j\rangle$ and we denote by $\|\cdot\|_{HS(\mathcal{H})}$ the corresponding norm. In particular, Hilbert-Schmidt operators are compact.

A closely related notion is that of a *trace class* operator. We say that a bounded operator A is trace class, if $\sum_{j\geq 1} \left\langle (A^*A)^{1/2} e_j, e_j \right\rangle < \infty$ for some (any) Hilbert basis $(e_j)_{j\geq 1}$. In particular, $\mathrm{Tr}(A) = \sum_{j\geq 1} \langle Ae_j, e_j\rangle < \infty$ and $\mathrm{Tr}(A)$ is called the trace of A. The space of trace class operators is a Banach space endowed with the norm $\|A\|_{TC} = \mathrm{Tr}(\sqrt{A^*A})$. Trace class operators are also Hilbert Schmidt (hence compact). The following inequalities relate the different operator norms:

$$\|A\| \leq \|A\|_{HS} \leq \|A\|_{TC}.$$

It can also be shown that for any $A \in HS(\mathcal{H})$ and bounded operator B

$$\begin{aligned} \|AB\|_{HS(\mathcal{H})} &\leq \|A\|_{HS(\mathcal{H})}\|B\| \\ \|BA\|_{HS(\mathcal{H})} &\leq \|B\|\|A\|_{HS(\mathcal{H})}. \end{aligned} \tag{1}$$

Spectral Theory for Compact Operators. Recall that the spectrum of a matrix K can be defined as the set of (in general, complex) eigenvalues λ, s.t. $\det(K - \lambda I) = 0$, or, equivalently, such that $\lambda I - K$ does not have a (bounded) inverse. This definition can be generalized directly to operators. Let $A : \mathcal{H} \to \mathcal{H}$ be a bounded operator, we say that λ belongs to the spectrum $\sigma(A)$, if $(A - \lambda I)$ does not have a bounded inverse. For any $\lambda \notin \sigma(A)$, $R(\lambda) = (A - \lambda I)^{-1}$ is the *resolvent operator*, which is by definition a bounded operator.

It can be shown (e.g., [13]) that if A is a compact operator, then its spectrum is discrete with the only accumulation point at zero. That means that the spectrum consists of isolated points with finite multiplicity $|\lambda_1| \geq |\lambda_2| \geq \cdots$, such that $\lim_{n\to\infty} \lambda_n = 0$.

If the operator A is self-adjoint ($A = A^*$, analogous to a symmetric matrix in the finite-dimensional case), the eigenvalues are real. To each eigenvalue λ, we can associate its *eigenspace*, the set of eigenvectors with this eigenvalue. The corresponding *projection operator* P_λ is defined as the projection onto the span of eigenvectors associated to λ. It can be shown that a self-adjoint compact operator A can be decomposed as follows:

$$A = \sum_{i=1}^{\infty} \lambda_i P_{\lambda_i},$$

the key result known as the *Spectral Theorem*. Moreover, it can be shown that the projection P_λ can be written explicitly in terms of the resolvent operator. Specifically, we have the following remarkable equality:

$$P_\lambda = \frac{1}{2\pi i} \int_{\Gamma\subset\mathbb{C}} (\gamma I - A)^{-1} d\gamma$$

where the integral can be taken over any closed curve in $\mathbb{C}$ containing λ and no other eigenvalue. We note that while an integral of an operator-valued function may seem unfamiliar, it is defined along the same lines as an integral of an ordinary real-valued function. Despite the initial technicality, the equation above allows for far simpler analysis of eigenprojections than other seemingly more direct methods.

This analysis can be extended for operators, which are not self-adjoint, to obtain a decomposition parallel to the Jordan canonical form for matrices. In the case of non-self-adjoint operators the projections are to *generalized eigenspaces* associated to an eigenvalue. To avoid overloading this section, we relegate the precise technical statements for that case to the Appendix A.

Reproducing Kernel Hilbert Space. Let X be a subset[1] of $\mathbb{R}^d$. An Hilbert space $\mathcal{H}$ of functions $f : X \to \mathbb{C}$ such that all the evaluation functionals are bounded, that is,

$$f(x) \leq C_x \|f\| \qquad \text{for some constant } C_x,$$

is called a *Reproducing Kernel Hilbert space.* It can be shown that there is unique symmetric, positive definite kernel function $K : X \times X \to \mathbb{C}$, *a reproducing kernel*, associated to $\mathcal{H}$ and the following reproducing property holds

repro
$$f(x) = \langle f, K_x \rangle \,, \tag{2}$$

where $K_x := K(\cdot, x)$. It is also well known [2] that each given reproducing kernel K uniquely defines a reproducing kernel Hilbert space $\mathcal{H} = \mathcal{H}_K$. We denote the scalar product and norm in $\mathcal{H}$ with $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. We will assume that the kernel is continuous and bounded[2].

Concentration Inequalities in Hilbert spaces. We recall that if $\xi_1, \ldots, \xi_n$ are independent (real-valued) random variables with zero mean and such that $|\xi_i| \leq C,\ i = 1, \ldots, n$, then Höeffding inequality ensures that $\forall \varepsilon > 0$

$$\mathrm{P}\left[\left|\frac{1}{n}\sum_i \xi_i\right| \geq \varepsilon\right] \leq 2e^{-\frac{n\varepsilon^2}{2C^2}}$$

If we set $\tau = \frac{n\varepsilon^2}{2C^2}$ then we can express the above inequality saying that with probability at least (with confidence) $1 - 2e^{-\tau}$

hoff
$$\left|\frac{1}{n}\sum_i \xi_i\right| \leq \frac{C\sqrt{2\tau}}{\sqrt{n}}. \tag{3}$$

Similarly if $\xi_1, \ldots, \xi_n$ are zero mean independent random variables with values in a Hilbert space and such that $\|\xi_i\| \leq C,\ i = 1, \ldots, n$, then the same inequality holds with the absolute value replaced by the norm in the Hilbert space. The following inequality

vec_hoff
$$\left\|\frac{1}{n}\sum_i \xi_i\right\| \leq \frac{C\sqrt{2\tau}}{\sqrt{n}}. \tag{4}$$

is given in [21].

[1]For technical reasons X needs can be taken to be an intersection of an open and a closed subset.

[2]This implies that the elements of $\mathcal{H}$ are bounded continuous functions, the space $\mathcal{H}$ is separable and is compactly embedded in $\mathcal{C}(X)$, with the compact-open topology, [2]. The assumption about continuity is not strictly necessary, but it will simplify some technical part.

3. Integral Operators defined by a Reproducing Kernel

Let the set $X \subset \mathbb{R}^d$ and the reproducing kernel K as above. We endow X with a probability measure ρ, we let $L^2(X, \rho)$ be the space of square integrable functions with norm $\|f\|_\rho^2 = \langle f, f \rangle_\rho = \int_X |f(x)|^2 d\rho(x)$, and

c1
$$\sup_{x \in X} K(x, x) \leq \kappa^2, \tag{5}$$

we define $L_K : L^2(X, \rho) \to L^2(X, \rho)$ to be the corresponding integral operator defined by

int_op
$$L_K f(x) = \int_X K(x, s) f(s) d\rho(s). \tag{6}$$

Suppose we are now given a set of points $\mathbf{x} = (x_1, \ldots, x_n)$ sampled i.i.d. according to ρ. Many problems in statistical data analysis and machine learning deal with the empirical kernel $n \times n$-matrix $\mathbf{K}$ given by $\mathbf{K}_{ij} = \frac{1}{n} K(x_i, x_j)$.

The question we want to discuss is to which extent we can use the kernel matrix $\mathbf{K}$ (and the corresponding eigenvalues, eigenvectors) to estimate L_K (and the corresponding eigenvalues, eigenfunctions). Answering this question is important as it guarantees that the computable empirical proxy is sufficiently close to the ideal infinite sample limit.

The first difficulty in relating L_K and $\mathbf{K}$ is that they operate on different spaces. By default, L_K is an operator on $L^2(X, \rho)$, while $\mathbf{K}$ is a finite dimensional matrix.

To overcome this difficulty we let $\mathcal{H}$ be the RKH space associated to K and define the extension operators $L_{K,\mathcal{H}}, L_{K,n} : \mathcal{H} \to \mathcal{H}$

$$L_{K,\mathcal{H}} = \int_X \langle \cdot, K_x \rangle_{\mathcal{H}} K_x d\rho(x), \tag{7}$$

T
$$L_{K,n} = \frac{1}{n} \sum_{i=1}^{n} \langle \cdot, K_{x_i} \rangle_{\mathcal{H}} K_{x_i}. \tag{8}$$

Note that $L_{K,\mathcal{H}}$ is the integral operator with kernel K with range and domain $\mathcal{H}$ rather than in $L^2(X, \rho)$. In next subsection, we show that $L_{K,\mathcal{H}}$ and L_K have the same eigenvalues (except zeros) and the corresponding eigenfunctions are closely related, and a similar relation holds for $L_{K,n}$ and $\mathbf{K}$. Thus to establish a connections between the spectral properties of $\mathbf{K}/n$ and L_K, it is sufficient to bound the difference $L_{K,\mathcal{H}} - L_{K,n}$, which is done in the following theorem.

op_bound **Theorem 1.** *The operators $L_{K,\mathcal{H}}$ and $L_{K,n}$ are Hilbert-Schmidt. Under the above assumption with confidence $1 - 2e^{-\tau}$*

$$\|L_{K,\mathcal{H}} - L_{K,n}\|_{HS} \leq \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}}.$$

Proof. We introduce a sequence $(\xi_i)_{i=1}^n$ of random variables in the space of Hilbert-Schmidt operators $HS(\mathcal{H})$ by

$$\xi_i = \langle K_{x_i}, \cdot \rangle K_{x_i} - L_{K,\mathcal{H}}.$$

From (8) follows that $E(\xi_i) = 0$. By a direct computation we have that $\|\langle \cdot, K_x \rangle K_x\|_{HS}^2 = \|K_x\|^4 \leq \kappa^4$ so that using again (8) we have

$$\|\xi\|_{HS} \leq 2\kappa^2, \ i = 1, \dots, n.$$

From inequality (4) we have with probability $1 - 2e^{-\tau}$

$$\|\frac{1}{n}\sum_i \xi_i\|_{HS} = \|L_{K,\mathcal{H}} - L_{K,n}\|_{HS} \leq \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}},$$

which establishes the result. □

As an immediate corollary of Theorem 1 we obtain several concentration results for eigenvalues and eigenfunctions discussed in subsectiion 3.2. However before that we provide a discussion of the Nyström extension needed to properly compare the above operators.

3.1. **Extension operators.** To compare the spectral properties of L_K and $L_{K,\mathcal{H}}$ we recall the following result, whose proof can be found in [9, 8].

Proposition 1. *The following facts hold:*

(1) *The operators L_K and $L_{K,\mathcal{H}}$ are positive, self-adjoint and trace class. In particular both $\sigma(L_{K,n})$ and $\sigma(\mathbf{K})$ are contained in $[0, \kappa]$.*
(2) *The spectra of L_K and $L_{K,\mathcal{H}}$ are the same, possibly except zeros, moreover if σ_j is nonzero eigenvalue and u_j, v_j associated eigenfunctions of L_K and $L_{K,\mathcal{H}}$ (normalized to norm 1 in $L^2(X, \rho)$ and $\mathcal{H}$) respectively, then*

$$\begin{aligned} u_j(x) &= \frac{1}{\sqrt{\sigma_j}} v_j(x) \qquad \text{for } \rho\text{-almost all } x \in X \\ v_j(\cdot) &= \frac{1}{\sqrt{\sigma_j}} \int_X K(\cdot, x) u_j(x) d\rho(x) \end{aligned}$$

(3) *Also for all $g \in L^2(X, \rho)$ and $f \in \mathcal{H}$ the following decompositions hold:*

$$\begin{aligned} L_K g &= \sum_{j\geq 1} \sigma_j \langle g, u_j \rangle_\rho u_j \\ L_{K,\mathcal{H}} f &= \sum_{j\geq 1} \sigma_j \langle f, v_j \rangle_{\mathcal{H}} v_j \end{aligned}$$

the eigenfunctions $(u_j)_{j\geq 1}$ of L_K form an orthonormal basis of $\ker {L_K}^{\perp}$ *and the eigenfunctions $(v_j)_{j\geq 1}$ of $L_{K,\mathcal{H}}$ for an orthonormal basis on $ker(L_{K,\mathcal{H}})^{\perp}$.*

Note that the RKHS $\mathcal{H}$ does not depend on the measure ρ. If the support of the measure ρ is only a subset of X (e.g., a finite set of points or a submanifold), then functions in $L^2(X,\rho)$ are only defined on the support of ρ whereas function in $\mathcal{H}$ are defined on the whole space X. The eigenfunction u of L_K and $L_{K,\mathcal{H}}$ coincide (up-to a scaling factor) on the support of the measure, and v is an *extension* of u outside of the support of ρ.
An analogous result relates the matrix $\mathbf{K}$ and the operator $L_{K,n}$.

Proposition 2. *The following facts hold:*

(1) *The finite rank operator $L_{K,n}$ is and the matrix $\mathbf{K}$ are positive, selfadjoint. In particular the spectrum $\sigma(L_{K,n})$ has only finitely many nonzero elements and is contained in $[0,\kappa]$.*
(2) *The spectra of $\mathbf{K}$ and $L_{K,n}$ are the same up to the zero, that is, $\sigma(\mathbf{K})\setminus\{0\} = \sigma(L_{K,n})\setminus\{0\}$. Moreover, if $\hat{\sigma}_j$ is a non zero eigenvalue and $\hat{u}_j, \hat{v}_j$ are the corresponding eigenvector and eigenfunction of $\mathbf{K}/n$ and $L_{K,n}$ then (normalized to norm 1 in $\mathbb{R}^n$ and $\mathcal{H}$)*

$$\begin{aligned} \hat{u}_j^i &= \frac{1}{\sqrt{\hat{\sigma}_j}} \hat{v}_j(x_i) \\ \hat{v}_j(\cdot) &= \frac{1}{\sqrt{\hat{\sigma}_j}} \left(\frac{1}{n} \sum_{i=1}^n K(\cdot, x_i) \hat{u}_j^i \right) \end{aligned}$$

(3) *Also for all $w \in \mathbb{R}^n$ and $f \in \mathcal{H}$ the following decompositions hold:*

$$\begin{aligned} \mathbf{K}w &= \sum_{j\geq 1} \hat{\sigma}_j \left\langle w, \hat{u}_j \right\rangle \hat{u}_j, \\ L_{K,n}f &= \sum_{j\geq 1} \hat{\sigma}_j \left\langle f, \hat{v}_j \right\rangle_{\mathcal{H}} \hat{v}_j; \end{aligned}$$

where the sum runs on the nonzero eigenvalues, the family $(\hat{u}_j)_{j\geq 1}$ is an othonormal basis in $\ker \mathbf{K}^{\perp} \subset \mathbb{R}^n$ and the family $(\hat{v}_j)_{j\geq 1}$ of $L_{K,n}$ form an orthonormal basis for the space $\ker(L_{K,n})^{\perp} \subset \mathcal{H}$, where

$$\ker(L_{K,n}) = \{ f \in \mathcal{H} \mid f(x_i) = 0 \ \forall i = 1, \dots, n \}$$

sec_bounds

3.2. **Bounds on eigenvalues and spectral projections.** To estimate the variation of the eigenvalues, we need to recall the notion of *extended enumeration* of discrete eigenvalues. We adapt the definition of [14], which is given for an arbitrary selfadjoint operator, to positive compact operators If A is a compact operator, an extended enumeration is a sequence of real numbers where every nonzero eigenvalue of A appears exactly as its multiplicity and the other values (if any) are zero. A nenumeration is an extended numeration where any element of the sequence is an isolated eigenvalue with finite multiplicity. If the sequence is infinite, this last condition is equivalent to the fact that any element is non zero.

The following result, due to Kato [14], is an extension to infinite dimensional operators of an inequality due to Lidskii for finite rank operator.

Theorem 2 (Kato 1987). *Let $\mathcal{H}$ be a separable Hilbert space with A, B self-adjoint compact operators. Let $(\gamma_j)_{j\geq 1}$, be an enumeration of discrete eigenvalues of C, then there exist extended enumerations $(\beta_j)_{j\geq 1}$ and $(\alpha_j)_{j\geq 1}$ of discrete eigenvalues of B and A respectively such that,*

$$\sum_{j\geq 1} \phi(|\beta_j - \alpha_j|) \leq \phi(\sum_{j\geq 1} \gamma_j).$$

where ϕ is any nonnegative convex function with $\phi(0) = 0$.

If A and B are positive operators and ϕ is an increasing function, it is possible to choose either $(\beta_j)_{j\geq 1}$ or $(\alpha_j)_{j\geq 1}$ as the decreasing enumeration, and the other sequence as the decreasing extended enumeration. In particular we have

$$(\sum_{j\geq 1} |\beta_i - \alpha_j|^p)^{1/p} \leq (\sum_{j\geq 1} |\gamma_j|^p)^{1/p},\ p \geq 1,$$

so that

$$(\sum_{j\geq 1} |\beta_j - \alpha_j|^2)^{1/2} \leq \|B - A\|_{HS}$$

and

$$\sup_{j\geq 1} |\beta_i - \alpha_j| \leq \|B - A\|.$$

The above results together with Theorem 1 immediately yields the following result.

eigs **Proposition 3.** *There exist extended enumerations $(\sigma_j)_{j\geq 1}$ and $(\hat{\sigma}_j)_{j\geq 1}$ of discrete eigenvalues for $L_{K,\mathcal{H}}$ and $L_{K,n}$, respectively. With confidence $1 - 2e^{-\tau}$*

$$\sup_{j\geq 1} |\sigma_j - \hat{\sigma}_j| \leq \|L_{K,\mathcal{H}} - L_{K,n}\| \leq \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}}$$

and

$$\left(\sum_{j\geq 1} (\sigma_j - \hat{\sigma}_j)^2\right)^{1/2} \leq \|L_{K,\mathcal{H}} - L_{K,n}\|_{HS} \leq \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}}.$$

If we are interested into concentration of the sum of the eigenvalues we can give a straightforward proof.

conc_trace **Proposition 4.** *Under the assumption of Proposition 3 with confidence $1 - 2e^{-\tau}$*

$$|\sum_j \sigma_j - \sum_j \hat{\sigma}_j| = |\operatorname{Tr}(L_{K,\mathcal{H}}) - \operatorname{Tr}(L_{K,n})| \leq \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}}.$$

Proof. Note that

$$\operatorname{Tr}(L_{K,n}) = \frac{1}{n}\sum_{i=1}^{n} K(x_i, x_i), \qquad \text{and} \qquad \operatorname{Tr}(L_{K,\mathcal{H}}) = \int_X K(x,x)d\rho(x).$$

Then we can define a sequence $(\xi_i)_{i=1}{}^n$ of real-valued random variables by $\xi_i = K(x_i, x_i) - \operatorname{Tr}(L_{K,\mathcal{H}})$. Clearly $\mathbb{E}[\xi_i] = 0$ and $|\xi_i| \leq 2\kappa^2$, $i = 1, \ldots, n$ so that Höeffding inequality (3) yields with confidence $1 - 2e^{-\tau}$

$$\left|\frac{1}{n}\sum_i \xi_i\right| = |\operatorname{Tr}(L_{K,\mathcal{H}}) - \operatorname{Tr}(L_{K,n})| \leq \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}}.$$

□

Eigenfunctions and Spectral Projections To control the spectral projections associated to one or more eigenvalues we need the following perturbation result, whose proof is given in [28] (see also Theorem 6 below). If A is a positive compact operator such that $\#\sigma(A) = \infty$, for an $N \in \mathbb{N}$, let P_N^A be the orthogonal projection on the eigenvectors corresponding to the top N eigenvalues.

Proposition 5. *Let A be a compact positive operator. Given an integer N, let $\delta = \frac{\alpha_N - \alpha_{N+1}}{2}$. If B is another compact positive operator such that $\|A - B\| \leq \frac{\delta}{2}$, then*

$$\|P_D^B - P_N^A\| \leq \frac{\|A - B\|}{\delta}$$

where the integer D is such that the dimension of the range of P_D^B is equal to the dimension of the range of P_N^A. If A and B are Hilbert-Schmidt, in the above bound the operator norm can be replaced by the Hilbert-Schmidt norm.

The proof of the above result can be found in the appendix.

We note that control of projections associated to simple eigenvalues implies that the corresponding eigenvectors are close since if u and v are taken to be normalized and such that $\langle u, v\rangle > 0$ then the following inequality holds

$$\|P_u - P_v\|_{HS}^2 \geq 2(1 - \langle u, v\rangle) = \|u - v\|_{\mathcal{H}}^2.$$

We are ready to state a probabilistic bound on eigenprojections. Assume that $\#\sigma(L_K) = \infty$

Theorem 3. *Let N be an integer and $g_N = \sigma_N - \sigma_{N+1}$. Given $\tau > 0$, if the number n of examples satisfies*

$$\frac{g_N}{2} > \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}},$$

then with probability greater than $1 - 2e^{-\tau}$

$$\|P_N - \hat{P}_D\|_{HS} \leq \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{g_N\sqrt{n}},$$

where $P_N = P_N^{L_K}$, $\hat{P}_D = P_D^{\mathbf{K}}$ *and the integer* D *is such that the dimension of the range of* P_D *is equal to the dimension of the range of* $\hat{P}_N$.

4. Asymmetric Graph Laplacian

As before X is a subset of $\mathbb{R}^d$ endowed with a probability measure ρ and $L^2(X,\rho)$ the space of square integrable functions with respect to ρ. Moreover, $W : X \times X \to \mathbb{R}^+$ is a symmetric, positive *weight* contiunuous function so that $W(x,s) \geq 0$ for all $x, s \in X$. Note that here we *do not* require W to be a positive definite kernel. Let $L_r : L^2(X,\rho) \to L^2(X,\rho)$ be defined by

$$L_r f(x) = f(x) - \int_X \frac{W(x,s)f(s)}{m(x)} d\rho(s)$$

where $m(x) = \int_X W(x,s)d\rho(s)$, is called the *degree function*.

Note that a set $\mathbf{x} = (x_1, \ldots, x_n) \in X$ sampled i.i.d. according to ρ defines a weighted undirected graph with the weight matrix $\mathbf{W}$ given by $\mathbf{W}_{ij} = \frac{1}{n} W(x_i, x_j)$. The (asymmetric) normalized graph Laplacian $\mathbf{L}_r : \mathbb{C}^n \to \mathbb{C}^n$ is an $n \times n$ matrix given by

$$\mathbf{L}_r = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W},$$

where the *degree* matrix $\mathbf{D}$ is diagonal with $\mathbf{D}_{ii} = \frac{1}{n}\sum_{j=1}^n W(x_i, x_j)$.

We will view $\mathbf{L}_r$ as a perturbation of L_r due to finite sampling and will extend the approach developed in this paper to relate their spectral properties. Note the methods in from the previous section are not directly applicable in this setting since W does not have to be a positive definite kernel so there is no RKHS associated to it. Moreover, even if W is positive definite, L_r involves division by a function, and may not be a map from the RKHS to itself.

To overcome this difficulty in our theoretical analysis, we will rely on an auxiliary RKHS (which eventually will be taken to be an appropriate Sobolev space). Interestingly enough, this space will play no role from the algorithmic point of view, but only enters the theoretical analysis. More precisely, let $m_n(\cdot) = \frac{1}{n}\sum_{i=1}^n W(\cdot, x_i)$, be the *empirical degree function*, we will need the following

cond2 **Assumption 1** (A1)**.** *Assume that* $\mathcal{H}$ *is a RKHS with bounded continuous kernel and, for all* $x \in X$, *that* $W(x,\cdot)/m(\cdot) \in \mathcal{H}$, $W(x,\cdot)/m_n(\cdot) \in \mathcal{H}$ *and also that for all* $x, s \in X$, $0 < c \leq W(x,s) < \infty$.

We can now consider the following extension operators: $L_{r,\mathcal{H}}, L_{r,\mathcal{H},n}, A_{\mathcal{H}}, A_n : \mathcal{H} \to \mathcal{H}$

op1
$$L_{r,\mathcal{H}} f = f - A_{\mathcal{H}} f = f - \frac{1}{m(\cdot)} \int_X \langle f, K(x,\cdot)\rangle_{\mathcal{H}} W(x,\cdot) d\rho(x), \tag{9}$$

op2
$$L_{r,\mathcal{H},n} f = f - A_n f = f - \frac{1}{m_n(\cdot)} \frac{1}{n} \sum_{i=1}^n \langle f, K(x_i,\cdot)\rangle_{\mathcal{H}} W(x_i,\cdot), \tag{10}$$

It can be seen (see the next subsection, where a detailed analysis is given) that L_r, $L_{r,\mathcal{H}}$ and $A_{\mathcal{H}}$ have related eigenvalues and eigenfunctions and that eigenvalues and eigenfunctions (eigenvectors) of A_n and $\mathbf{L}$ are also closely related. In particular we will see in the following that to relate the spectral properties of L_r and $\mathbf{L}$ it suffices to control the deviation $A_{\mathcal{H}} - A_n$. However before doing this we make the above statements precise in the following subsection.

4.1. **Extension Operators.** The following proposition relates the operators L_r, $A_{\mathcal{H}}$ and $L_{r,\mathcal{H}}$.

ext_lap

Proposition 6. *The following facts hold:*

(1) *The operator $A_{\mathcal{H}}$ is Hilbert-Schmidt, the operators L_r and $L_{r,\mathcal{H}}$ are bounded and have positive eigenvalues.*
(2) *The eigenfunctions of $A_{\mathcal{H}}$ and $L_{r,\mathcal{H}}$ are the same and $\sigma(A_{\mathcal{H}}) = \sigma(L_{r,\mathcal{H}}) - 1$.*
(3) *The spectra of L_r and $L_{r,\mathcal{H}}$ are the same, moreover if $\sigma \neq 1$ is an eigenvalue and the u, v eigenfunctions of L_r and $L_{r,\mathcal{H}}$ (normalized to norm 1 in $L^2(X,\rho)$ and $\mathcal{H}$) respectively, then*

$$
\begin{aligned}
u(x) &= \frac{1}{\sqrt{1-\sigma}} v(x) \\
v(\cdot) &= \frac{1}{\sqrt{1-\sigma}} \int_X d\rho(x) \frac{W(x,\cdot)}{m(\cdot)} u(x)
\end{aligned}
$$

(4) *Finally the following decompositions hold*

$$
\begin{aligned}
L_r &= I - \sum_{j\geq 1} \sigma_j P_j, \\
L_{r,\mathcal{H}} &= I - \left(\sum_{j\geq 1} \sigma_j Q_j + D \right),
\end{aligned}
$$

where Q_j, P_j are the spectral projection in $L^2(X,\rho)$ and $\mathcal{H}$ associated to the eigenvalues σ_j and e eigenvalues σ_j and D is a quasi-nilpotent operator such that

$$Q_j D = D Q_j = 0.$$

As in Section 3,we have an analogous result allowing us to relate $\mathbf{L}_r$ to $L_{r,\mathcal{H},n}$ and A_n.

ext_lap_emp

Proposition 7. *The following facts hold:*

(1) *The operator A_n is Hilbert-Schmidt, the matrix $\mathbf{L}_r$ and the operator $L_{r,\mathcal{H},n}$ have non-negative eigenvalues.*
(2) *The eigenfunctions of A_n and $L_{r,\mathcal{H},n}$ are the same and $\sigma(A_n) = \sigma(L_{r,\mathcal{H},n}) - 1$.*

(3) *The spectra of* $\mathbf{L}_r$ *and* $L_{r,\mathcal{H},n}$ *are the same up to the eigenvalue* 1, *moreover if* $\hat{\sigma} \neq 1$ *is an eigenvalue and the* $\hat{u}, \hat{v}$ *eigenvector and eigenfunction of* $\mathbf{L}_r$ *and* $L_{r,\mathcal{H},n}$, *(normalized to norm* 1 *in* $\mathbb{C}^n$ *and* $\mathcal{H}$*) respectively, then*

$$\begin{aligned}
\hat{u}^i &= \frac{1}{\sqrt{1-\hat{\sigma}}}\hat{v}(x_i) \\
\hat{v}(\cdot) &= \frac{1}{\sqrt{1-\hat{\sigma}}}\sum_{i=1}^{n}\frac{W(x_i,\cdot)}{m_n(\cdot)}\hat{u}^i
\end{aligned}$$

where $\hat{u}^i$ *is the* $i-th$ *component of the eigenvector* $\hat{u}$.

(4) *Finally the following decompositions hold*

$$\begin{aligned}
\mathbf{L}_r &= I - \sum_{j=1}^{n}\hat{\sigma}_j\hat{P}_j, \\
L_{r,\mathcal{H},n} &= I - \left(\sum_{j=1}^{n}\hat{\sigma}_j\hat{Q}_j + \hat{D}\right),
\end{aligned}$$

where $\hat{Q}_j, \hat{P}_j$ *are spectral projections, in* $\mathbb{C}^n$ *and* $\mathcal{H}$, *associated to the eigenvalues* $\hat{\sigma}_j$ *and* $\hat{D}$ *is a quasi nilpotent operator such that* $\hat{Q}_j\hat{D} = \hat{D}\hat{Q}_j = 0$.

The last decomposition is parallel to the Jordan canonical form for (non-symmetric) matrices.

The assumption $W(x,\cdot)/m(\cdot) \in \mathcal{H}$ is crucial and is not satisfied in general. For example, it is not necessarily the case even if $W(x,y)$ is a reproducing kernel. however it is the case when the RKH space $\mathcal{H}$ is a Sobolev space with sufficiently high smoothness degree and the weight function is also sufficiently smooth. To estimate the deviation of $L_{r,\mathcal{H}}$ to $L_{r,\mathcal{H},n}$ we consider this latter situation.

4.2. **Convergence for Operators in Sobolev Spaces and Smooth Weight Function.** We briefly recall some basic definitions as well some connection between Sobolev spaces and RKHS. For the sake of simplicity, X can be assumed to be a bounded open subset of $\mathbb{R}^d$ or a compact smooth manifold and ρ a probability measure with density (wrt to the uniform measure) bounded away from zero.

Recall that for $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ and $|\alpha| = \alpha_1 + \cdots + \alpha_d$, we denote with $D^\alpha f$ the (weak) derivative of f on X. For any $s \in \mathbb{N}$, the Sobolev space $\mathcal{H}^s$ is defined as the space of square integrable functions having weak derivatives on X for all $|\alpha| = s$ and such that

$$\|f\|_s = \|f\|_\rho + \sum_{|\alpha|=s}\|(D^\alpha f)(x)\|_\rho < \infty,$$

the above definition of H^s can be generalized allowing $s \in]0, +\infty[$.

The Sobolev Embedding theorem ensures[3] that, for $s > d/2$ the inclusion $\mathcal{H}^s \hookrightarrow \mathcal{C}(X)$ is well defined and bounded or in other words we have

$$\|f\|_\infty \leq C_1 \|f\|_s.$$

Then $\mathcal{H}^s$ is a RKHS with reproducing kernel $K^s(x, y)$ and $f(x) = \langle f, K^s_x \rangle_s$ where $K^s_x := K^s(x, \cdot)$. Moreover we also have $\sup_{x \in X} \|K^s_x\|_s = C_1 < \infty$.
In the following we will need the following result from [7].

Lemma 1. *Let $g \in \mathcal{C}^s(X)$ where all derivatives are bounded up to order s. The operator $M_g : \mathcal{H}^s \to \mathcal{H}^s$ defined by $M_g f(x) = g(x) f(x)$ is a well defined bounded operator with norm*

molt_norm
$$\|M_g\| \leq a \|g\|_{s'} < \infty. \tag{11}$$

In view of the relation between L_r, $L_{r,\mathcal{H}}$ and $A_\mathcal{H}$ (and their empirical counterparts) to relate the spectral properties of L_r and $\mathbf{L}$ it suffices to control the deviation $A_\mathcal{H} - A_n$. To this aim we make the following assumption.

cond3 **Assumption 2** (A2)**.** *Let $\mathcal{H}_{s'}, \mathcal{H}_s$ to be a Sobolev spaces such that $s' > s + d/2$. We assume that* $\sup_{x \in X} \|K^s_x\|_s = C_1$, $\sup_{x \in X} \|W_x\|_{s'} \leq C_2$, $\|m^{-1}\|_{s'} \leq C_3$, $\|m_n^{-1}\|_{s'} \leq C_4$.

The following theorem establishes the desired result.

lapla_main **Theorem 4.** *If assumption A2 holds, then for some positive constant C with confidence $1 - 2e^{-\tau}$ we have*

$$\|A_\mathcal{H} - A_n\|_{HS} \leq C \frac{\sqrt{\tau}}{\sqrt{n}}$$

The proof of the above result is postponed to section 4.4 and in the next section we consider concentration for the spectra.

4.3. **Bounds on eigenvalues and spectral projections.** Since the operators are no longer self-adjoint the perturbation results in section 3.2 cannot be used. The following result is a reformulation of a result [26], see also [1].

Theorem 5. *Let B, B_n be bounded compact operators such that $B_n \to B$ in operator norm as $n \to \infty$. Let α be a non zero eigenvalue of B with finite multiplicity m. Let Γ be a simple closed curve enclosing α, then there exist N such that for $n > N$ there is a finite set of eigenvalues of B_n with multiplicity summing up to m enclosed in Γ. Moreover every eigenvalue α_n of B_n in such finite set converges to α as $n \to \infty$.*

This result together with Theorem 4 allows to derive convergence of eigenvalues in the sense of the above proposition. To obtain bounds on spectral projections we can use the following result.

---

[3] Under mild conditions on the boundary of X for the case of domain in $\mathbb{R}^d$.

caldo **Theorem 6.** *Let A be a compact operator. Given $1 \leq N \leq p_A$, there exist two costants δ and $R > 0$, depending on N and $\sigma(A)$, such that for any compact operator B satisfying $\|A - B\| \leq \delta/2$, then*

2pert
$$\|P_N^B - P_D^A\| \leq R\|A - B\|, \tag{12}$$

for a suitable integer D, depending on B. If $R\|A - B\| \leq 1$, D is such that the sum of the multiplicity of the first D eigenvectors of B is equal to the sum of the multiplicity of the first N eigenvectors of A.
If A and B are Hilbert-Schmidt operator, in (12) it is possible to replace the operator norm with the Hilbert-Schmidt norm.
Finally, if A is a positive operator, $\delta = \frac{\alpha_N - \alpha_{N+1}}{2}$, $R = \frac{1}{\delta}$ and D is always such that the sum of the multiplicity of the first D eigenvectors of B is equal to the sum of the multiplicity of the first N eigenvectors of A.

Then we can immediately derive the following result.

Theorem 7. *Consider the first N eigenvalues of $A_{\mathcal{H}}$. There exist two costants δ and $R > 0$, depending on N and $\sigma(A_{\mathcal{H}})$, such that if $\frac{\delta}{2} > C\frac{\sqrt{\tau}}{\sqrt{n}}$ with C as in Theorem 4, then with confidence $1 - 2e^{-\tau}$,*

$$\|P^N - \hat{P}^D\|_{HS} \leq R\frac{C\sqrt{\tau}}{\sqrt{n}}.$$

where $P^N, \hat{P}^D$ are the eigenprojections corresponding to the N eigenvalues of $A_{\mathcal{H}}$ and D eigenvalues of A_n. If $RC\frac{\sqrt{\tau}}{\sqrt{n}} \leq 1$, then D is such that the sum of the multiplicity of the first D eigenvectors of A_n is equal to the sum of the multiplicity of the first N eigenvectors of $A_{\mathcal{H}}$.

sec_proof

4.4. **Proofs.** We start giving the proof of Proposition 6.

Proof of Proposition 6. We first need some preliminary observations. Note that $\rho_W = m\rho$ defines a finite measure on X having density w.r.t. ρ. The measures ρ_W, ρ are equivalent[4] and the spaces $L^2(X, \rho)$ and $L^2(X, \rho_W)$ (square integrable functions with respect to ρ_W) are the same vector space but they are endowed with different norm/scalar product. Functions that are orthogonal in one space might not be orthogonal in the other. In particular if L_r is regarded as an operator from and to $L^2(X, \rho_W)$, the eigenvalues and eigenfunctions are the same. The operator $U_W : L^2(X, \rho) \to L^2(X, \rho_W)$ defined by $U_W f(x) = m(x)^{-1/2} f(x)$ is unitary.

[4]Two measures are equivalent if they have the same null sets. In terms of absolute continuity of measures, two measures are equivalent if and only if each is absolutely continuous with respect to the other.

Note that the operator $I_K : \mathcal{H} \to L^2(X, \rho_W)$ defined by $I_K f(x) = \langle f, K_x \rangle_{\mathcal{H}}$ is linear and Hilbert-Schmidt since

$$\begin{aligned} \|I_K\|_{HS}^2 = \sum_{j\geq 1} \|I_K e_j\|_{\rho_W}^2 &= \int_X d\rho_W(x) \sum_{j\geq 1} \langle K_x, e_j \rangle_{\mathcal{H}}^2 \\ &= \int_X K(x,x) m(x) d\rho(x) \leq \|K_x\|^2 \|m\|_\infty. \end{aligned}$$

The operator $I_W^* : L^2(X, \rho_W) \to \mathcal{H}$ defined by

$$I_W^* f = \int_X d\rho(x) \frac{W(x, \cdot)}{m(\cdot)} f(x)$$

is linear and bounded by $\|W_x/m\|_{\mathcal{H}}^2$. A direct computation shows that

$$I_W^* I_K = A_{\mathcal{H}} = I - L_{r,\mathcal{H}}$$

and

$$I_K I_W^* = I - L_r,$$

where $L_r : L^2(X, \rho_W) \to L^2(X, \rho_W)$. Both the above operators are linear and Hilbert-Schmidt since they are composition of a bounded operator and Hilbert-Schmidt operator. Again by a direct computation we have that

$$\sigma(I_K I_W^*) = \sigma(I_W^* I_K) = \sigma(L_r) - 1 = \sigma(L_{r,\mathcal{H}}) - 1.$$

Moreover if $\sigma \neq 1$ and $v \in \mathcal{H}, v \neq 0$ such that $L_{r,\mathcal{H}} v = \sigma v$, then if we let $\sqrt{1-\sigma} u = I_K v$ then $u \neq 0$, u is an eigenfunction of L_r and $\sqrt{1-\sigma} v = I_W^* u$. Similarly we can prove that if $\sigma \neq 1$ and $u \in L^2(X, \rho), u \neq 0$ such that $L_r u = \sigma u$, then if we let $\sqrt{1-\sigma} v = I_W^* u$ then $v \neq 0$, v is an eigenfunction of $L_{r,\mathcal{H}}$ and $\sqrt{1-\sigma} u = I_K v$.

We now show that L_r and $L_{r,\mathcal{H}}$ have positive eigenvalues. To this aim note that

$$L_r = U_W L_s U_W^{-1}.$$

where $L_s : L^2(X, \rho) \to L^2(X, \rho)$ is defined by

$$L_s f(s) = f(s) - \int_X \frac{W(x,s)}{\sqrt{m(x)}\sqrt{m(s)}} f(x) d\rho(x).$$

The operator L_r is positive since $\forall f \in L^2(X, \rho)$,

$$\begin{aligned} \langle L_s f, f \rangle_\rho &= \int_X |f(x)|^2 d\rho(x) - \int_X \int_X \frac{W(x,s)}{\sqrt{m(x)}\sqrt{m(s)}} f(x) f(s) d\rho(x) d\rho(s) \\ &= \frac{1}{2} \int_X \int_X \left[\frac{|f(x)|^2}{m(x)} - 2 \frac{|f(x)||f(s)|}{\sqrt{m(x)}\sqrt{m(s)}} - \frac{|f(s)|^2}{m(s)} \right] W(x,s) d\rho(x) d\rho(s) \\ &= \frac{1}{2} \int_X \int_X W(x,s) \left[\frac{|f(x)|}{\sqrt{m(x)}} - \frac{|f(s)|}{\sqrt{m(s)}} \right]^2 > 0, \end{aligned}$$

where we used

$$\int_X |f(x)|^2 d\rho(x) = \int_X |f(x)|^2 d\rho(x) \frac{\int_X W(x,s)d\rho(s)}{\int_X W(x,s)d\rho(s)} = \int_X \int_X \frac{|f(x)|^2}{m(x)^2} W(x,s)d\rho(x)d\rho(s).$$

Finally we prove that both L_r and $L_{r,\mathcal{H}}$ admits a decomposition in terms of spectral projections.
To this aim we note that since $I_K I_W^*$ is a self adjoint operator, it can be decomposed as

$$I_K I_W^* = \sum_{j\geq 1} \sigma_j P_j$$

where for all j, $P_j : L^2(X, \rho_W) \to L^2(X, \rho_W)$ is the spectral projection associated to an eigenvalue $\sigma_j \neq 0$. Note that by definition P_j satisfies:

$$\begin{aligned} P_j^2 &= P_j, \\ P_j^* &= P_j, \\ P_j P_i &= 0, \; i \neq j, \\ P_j P_{ker(I_K I_W^*)} &= 0 \\ \sum_{j\geq 1} P_j &= I - P_{ker(I_K I_W^*)} \end{aligned}$$

where $P_{ker(I_K I_W^*)}$ is the projection on the kernel of $I_K I_W^*$ and the sum in the last equation converges in the strong operator topology. In particular we have

$$I_K I_W^* P_j = P_j I_K I_W^* = \sigma_j P_j.$$

Let $Q_j : \mathcal{H} \to \mathcal{H}$ be defined by

$$Q_j = \frac{1}{\sigma_j} I_W^* P_j I_K.$$

Then from the properties of the projections P_j we have,

$$\begin{aligned} Q_j^2 &= \frac{1}{\sigma_j^2} I_W^* P_j I_K I_W^* P_j I_K = \frac{1}{\sigma_j} I_W^* P_j P_j I_K = Q_j, \\ Q_j Q_i &= \frac{1}{\sigma_j \sigma_i} I_W^* P_j I_K I_W^* P_i I_K = \frac{1}{\sigma_i} I_W^* P_j P_i I_K = 0. \end{aligned}$$

Moreover, since

$$\sum_{j\geq 1} \sigma_j Q_j = \sum_{j\geq 1} \sigma_j \frac{1}{\sigma_j} I_W^* P_j I_K = I_W^* (\sum_{j\geq 1} P_j) I_K = I_W^* I_K - I_W^* P_{ker(I_K I_W^*)} I_K$$

so that

$$I_K I_W^* = \sum_{j\geq 1} \sigma_j Q_j + I_W^* P_{ker(I_K I_W^*)} I_K,$$

where again all the sums are to be intended as converging in the strong operator topology. If we let $D = I_W^* P_{ker(I_K I_W^*)} I_K$ then

$$Q_j D = \frac{1}{\sigma_j} I_W^* P_j I_K I_W^* P_{ker(I_K I_W^*)} = I_W^* P_j P_{ker(I_K I_W^*)} = 0,$$

and, similarly,$DQ_j = 0$. By construction $\sigma(D) = 0$, that is, D is a quasi-nilpotent operator. □

The proof of Proposition 7 is the essentially the same.

Proof of Proposition 7. The proof is the same as the above proposition by replacing ρ with the empirical measure $\frac{1}{n}\sum_{i=1}^n \delta x_i$. □

Next we prove Theorem 6.

Proof of Theorem 6. Recall that for any bounded operator T with norm strictly less than 1, then Neumann series converges so that

$$\sum_{i=0}^{\infty} T^i = (I - T)^{-1}$$

and this allows to show that if T is invertible and L is another operator such that $q = \|T - L\|\|T^{-1}\| < 1$, then L is also invertible with $\|L^{-1}\| \leq 1/(1-q)\|T^{-1}\|$. This can be seen writing $L = T(I - (I - T^{-1}L)) = T(I - T^{-1}(T - L))$ so that

$$L^{-1} = (I - T^{-1}(T - L))^{-1} T^{-1} = \sum_{i=0}^{\infty} [T^{-1}(T - L)]^i T^{-1}.$$

Let Γ be a (counterclockwise) closed simple curve enclosing $\alpha_1, \ldots, \alpha_N$, but no other points of $\sigma(A)$. The existence is ensured by the fact that all the α_i are isolated points. Since the spectrum of A is a subset of the real line, we can always choose Γ in such a way that Γ intersects the positive real axis only in two points: first point is between α_N and α_{N+1}, whereas the second point is strictly bigger that $2\|A\|$. Let

$$\delta^{-1} = \sup_{\lambda \in \Gamma} \|(\lambda I - A)^{-1}\|,$$

which is finite since Γ is compact curve in the resolvent set $\rho(A)$. Possibly redefining δ we can also assume that $\delta \leq \|A\|$.

For $\lambda \in \Gamma$, we can take $T = \lambda I - A$ and $L = \lambda I - B$ so that, if

radius (13)
$$\|(\lambda I - A)^{-1}(A - B)\| \leq \delta^{-1}\|(A - B)\| = \frac{1}{2} < 1 \tag{13}$$

then the above results give

neu (14)
$$\begin{aligned}(\lambda I - B)^{-1} &= \sum_{i=0}^{\infty} (-1)^k [(\lambda I - A)^{-1}(A - B)]^i (\lambda I - A)^{-1} \\ &= (\lambda I - A)^{-1} + \sum_{i=1}^{\infty} (-1)^k [(\lambda I - A)^{-1}(A - B)]^i (\lambda I - A)^{-1}. \qquad (14)\end{aligned}$$

In particular, Γ does not intersect the spectrum of B. Moreover, since $\sigma(B) \subset [0, \|B\|]$ and, by assumption,

$$\|B\| \leq \|B - A\| + \|A\| < 2\|A\|,$$

Γ is a (counterclockwise) closed simple curve enclosing the first D eigenvalues $\beta_1, \ldots, \beta_D$ of B, but no other points of $\sigma(B)$. Recall that the spectral projections can be written as

$$P_\Gamma^A = \frac{1}{2\pi i}\int_\Gamma (\lambda I - A)^{-1} d\lambda,\; P_\Gamma^B = \frac{1}{2\pi i}\int_\Gamma .(\lambda I - B)^{-1} d\lambda$$

The above expression and (14) give

$$\begin{aligned} P_\Gamma^B - P_\Gamma^A &= \frac{1}{2\pi i}\int_\Gamma ((\lambda I - B)^{-1} - (\lambda I - A)^{-1}) d\lambda \\ &= \frac{1}{2\pi i}\int_\Gamma \sum_{i=1}^{\infty} (-1)^k [(\lambda I - A)^{-1}(A - B)]^i (\lambda I - A)^{-1} d\lambda \end{aligned}$$

so that taking $q \leq 1/2$ the following inequalities hold

$$\|P_\Gamma^B - P_\Gamma^A\| \leq 2\delta^{-1}\frac{\ell(\Gamma)}{2\pi} \leq \frac{\|A - B\|}{\delta^2}\frac{\ell(\Gamma)}{\pi} = C\|A - B\|$$

where $\ell(\Gamma)$ is the length of Γ. In the above computations, one can replace the operator norm with the Hilbert-Schmidt norm. Notice that, if $C\|A - B\| \leq 1$, $\|P_N^A - P_D^B\| \leq 1$. Since both are projections, it follows that the corresponding ranges have the same dimensions.

If A and B are positive, one can choose the curve Γ as in [28] and, following their proof, this gives the explicit form for δ and C. $\square$

To prove Theorem 4 we need the following preliminary estimates.

op_bound2 **Proposition 8.** *The operators $L_{W,\mathcal{H}}, L_{W,n} : \mathcal{H}_s \to \mathcal{H}_s$ defined by*

$$\begin{aligned} L_{W,\mathcal{H}} &= \int_X \langle \cdot, K^s(x, \cdot)\rangle_s W(x, \cdot) d\rho(x), \; \boxed{\texttt{op1}} \\ L_{W,n} &= \frac{1}{n}\sum_{i=1}^{n} \langle \cdot, K^s(x_i, \cdot)\rangle_s W(x_i, \cdot) \boxed{\texttt{,op2}} \end{aligned}$$

are Hilbert Schmidt and with confidence $1 - 2e^{-\tau}$

$$\|L_{W,\mathcal{H}} - L_{W,n}\|_{HS} \leq \frac{2\sqrt{2}C_1C_2\sqrt{2\tau}}{\sqrt{n}}.$$

Proof. Note that $\|\langle \cdot, K_{x_i}^s\rangle_s W_{x_i}\|_{HS} = \|K_{x_i}^s\|\|W_{x_i}\|_s \leq C_1C_2$ so that $L_{W,n}, L_{W,\mathcal{H}}$ are Hilbert Schmidt. The random variables $(\xi_i)_{i=1}^n$ defined by $\xi_i = \langle \cdot, K_{x_i}^s\rangle_s W_{x_i} - L_{W,\mathcal{H}}$ are zero mean and bounded by $2C_1C_2$. Applying (4) we have with confidence $1 - 2e^{-\tau}$

op_bound2 (15)
$$\|L_{W,\mathcal{H}} - L_{W,n}\|_{HS} \leq \frac{2\sqrt{2}C_1C_2\sqrt{\tau}}{\sqrt{n}}.$$

$\square$

Next the multiplication operators defined by the degree functions are considered.

mul_bound **Proposition 9.** *Let $M, M_n : \mathcal{H}^s \to \mathcal{H}^s$ be defined by $Mf(x) = m(x)f(x)$ and $M_n f(x) = m_n(x)f(x)$. Then M, M_n are linear operators bounded by C_2 and with confidence $1 - 2e^{-\tau}$*

$$\|M - M_n\| \leq \frac{2C_2 a\sqrt{2\tau}}{\sqrt{n}}.$$

where is a positive constant.

Proof. It follows from (11) that under assumption A2 M, M_n are well defined operators whose norm is bounded by $2aC_2$ (we assume a is the same for sake of simplicty).

The random variables $(\xi_i)_{i=1}^n$, defined by $\xi_i = W_{x_i} - m$ are zero mean and bounded by $2C_2 a$. Applying (4) we have with high probability

$$\|m - m_n\|_{s'} \leq \frac{2aC_2\sqrt{2\tau}}{\sqrt{n}}.$$

It follows from (11) that

molt_bound

$$\|M - M_n\| \leq \frac{2aC_2\sqrt{2\tau}}{\sqrt{n}}. \tag{16}$$

□

We can combine the above two propositions to get the proof of Theorem (4).

Proof of Theorem 4. It follows from (11) and by assumption A3 that the operators $M^{-1}, M_n^{-1} : \mathcal{H}^s \to \mathcal{H}^s$ defined by $M^{-1}f(x) = m(x)^{-1}f(x)$ and $M_n^{-1}f(x) = m_n^{-1}(x)f(x)$ are linear operators bounded by C_3, C_4 respectively. Then $A_{\mathcal{H}} = M_n^{-1}L_{W,\mathcal{H}}$ and $A_n = M^{-1}L_{W,n}$ so that we can consider the following decomposition

deco

$$\begin{aligned} L_{r,\mathcal{H}} - L_{r,\mathcal{H},n} &= M_n^{-1}L_{W,n} - M^{-1}L_{W,\mathcal{H}} \\ &= (M_n^{-1} - M^{-1})L_{W,\mathcal{H}} + M_n^{-1}(L_{W,n} - L_{W,\mathcal{H}}) \\ &= M_n^{-1}(M - M_n)M^{-1}L_{W,\mathcal{H}} + M_n^{-1}(L_{K,n} - L_{W,\mathcal{H}}). \end{aligned} \tag{17}$$

Recalling (1), we consider the Hilbert-Schmidt norm of the above expression. Using the inequalities (9), (15), (11) and the assumption A3 we see that there is a constant C, such that

$$\|M_n^{-1}L_{K,n} - M^{-1}L_{K,\mathcal{H}}\|_{HS} \leq C\frac{\sqrt{\tau}}{\sqrt{n}}.$$

□

Appendix A. Spectral theorem for non-self-adjoint compact operators

Let $A : \mathcal{H} \to \mathcal{H}$ be a compact operator. The spectrum $\sigma(A)$ of A is defined as the set of complex number such that the operator$(A - \lambda I)$ does not admit a bounded inverse, whereas the complement of $\sigma(A)$ is called the resolvent and denoted by $\rho(A)$. For any $\lambda \in \rho(A)$, $R(\lambda) = (A - \lambda I)^{-1}$ is the resolvent operator, which is by definition a bounded operator. We recall the main results about the spectrum of a compact operator, [13]

Proposition 10. *The spectrum of a compact operator A is a countable compact subset of $\mathbb{C}$ with no accumulation point different from zero, that is,*

$$\sigma(A)\backslash\{0\} = \{\lambda_i \mid i \geq 1,\ \lambda_i \neq \lambda_j \text{ if } i \neq j\} \qquad \text{with} \qquad \lim_{i\to\infty} \lambda_i = 0 \text{ if } |\sigma(A)| = \infty$$

For any $i \geq 1$, λ_i is an eigenvalue of A, that is, there exists a nonzero vector $u \in \mathcal{H}$ such that $Au = \lambda_i u$. Let Γ_i be a (counterclockwise) closed simple curve enclosing λ_i, but no other points of $\sigma(A)$, then the operator defined by

$$P_{\lambda_i} = \frac{1}{2\pi i} \int_{\Gamma_i} (\lambda I - A)^{-1} d\lambda$$

satisfies

$$P_{\lambda_i} P_{\lambda_j} = \delta_{ij} P_{\lambda_i} \qquad \text{and} \qquad (A - \lambda_i) P_{\lambda_i} = D_{\lambda_i} \qquad \text{for all } i, j \geq 1,$$

where D_{λ_i} is a nilpotent operator such that $P_{\lambda_i} D_{\lambda_i} = D_{\lambda_i} P_{\lambda_i} = D_{\lambda_i}$. In particular the dimension of the range of P_{λ_i} is always finite.

We notice that P_{λ_i} is a projection onto a finite dimensional space $\mathcal{H}_{\lambda_i}$, which is left invariant by T. A nonzero vector u belongs to $\mathcal{H}_{\lambda_i}$ if and only if there exists an integer $m \leq \dim \mathcal{H}\lambda_i$ such that $(A - \lambda)^m u = 0$, that is, u is a generalized eigenvector of A. However, if A is symmetric, for all $i \geq 1$, $\lambda_i \in \mathbb{R}$, P_{λ_i} is an orthogonal projection and $D_{\lambda_i} = 0$ and it holds that

$$A = \sum_{i \geq 1} \lambda_i P_{\lambda_i}$$

where the series converges in operator norm. Moreover, if $\mathcal{H}$ is infinite dimensional, $\lambda = 0$ is always in $\sigma(A)$, but it can be or not an eigenvalue of A.

If A be a compact operator with $\sigma(A) \subset [0, \|A\|]$, we introduce the following notation. Denoted by p_A the cardinality of $\sigma(A) \setminus \{0\}$ and given an integer $1 \leq N \leq p_A$, let $\lambda_1 > \lambda_2 > \ldots, \lambda_N > 0$ be the first N nonzero eigenvalues of A, sorted in a decreasing way. We denote by P_N^A the spectral projection onto all the generalized eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_N$. The range of P_N^A is a finite-dimensional vector space, whose dimension is the sum of the algebraic multiplicity of the first N eigenvalues.

Moreover

$$P_N^A = \sum_{j=1}^{N} P_{\lambda_j} = \frac{1}{2\pi i}\int_\Gamma (\lambda I - A)^{-1} d\lambda$$

where Γ is a (counterclockwise) closed simple curve enclosing $\lambda_1, \ldots, \lambda_N$, but no other points of $\sigma(A)$.

References

anse [1] P. M. Anselone. *Collectively compact operator approximation theory and applications to integral equations*. Prentice-Hall Inc., Englewood Cliffs, N. J., 1971. With an appendix by Joel Davis, Prentice-Hall Series in Automatic Computation.

aron50 [2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

bapero07 [3] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

belkin03 [4] M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, 2003.

belniy06 [5] M. Belkin and P. Niyogi. Convergence of laplacian eigenmaps. New York, 2006. Springer.

blbozw06 [6] G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 0885-6125 (Print) 1573-0565 (Online), 2006.

bur98 [7] V. Burenkov. *Sobolev spaces on domains*. B. G. Teubuer, Stuttgart-Leipzig, 1998.

deroca06 [8] E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for Tikhonov regularization. *Anal. Appl. (Singap.)*, 4(1):81–99, 2006.

devroscap05 [9] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, May 2005.

ginkol06 [10] E. Gine' and V. Koltchinskii. Empirical graph laplacian approximation of laplace-beltrami operators: Large sample results. *High Dimensional Probability*, 51:238259, 2006.

hein2006 [11] M. Hein. Uniform convergence of adaptive graph-based regularization. pages 50–64, New York, 2006. Springer.

headlu [12] M. Hein, J. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. pages 470–485, 2005. Student Paper Award.

kato66 [13] T. Kato. *Perturbation theory for linear operators*. Springer, Berlin, 1966.

kato87 [14] T. Kato. Variation of discrete spectra. *Commun. Math. Phys.*, III:501–504, 1987.

koltch98 [15] V. Koltchinskii. Asymptotics of spectral projections of some random matrices approximating integral operators. *Progress in Probabilty*, 43, 1998.

kolgin00 [16] V. Koltchinskii and E. Gine'. Random matrix approximation of spectra of integral op- erators. *Bernoulli*, 6:113–167, 2000.

lafon04 [17] S. Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, 2004.

lang93 [18] S. Lang. *Real and Functional Analysis*. Springer, New York, 1993.

menpaj05 [19] S. Mendelson and A. Pajor. Ellipsoid approximation with random vectors. pages 429–433, New York, 2005. Springer.

menpaj06 [20] S. Mendelson and A. Pajor. On singular values of matrices with independent rows. *Bernoulli*, 12(5):761–773, 2006.

pinelis92 [21] I. Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. *Probability in Banach Spaces, 8, Proceedings of the 8th International Conference*, pages 128–134, 1992.

shcrka02 [22] J. Shawe-Taylor, N. Cristianini, and J. Kandola. On the concentration of spectral properties. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in*

Neural Information Processing Systems 14, pages 511–517, Cambridge, MA, 2002. MIT Press.

singer06 [23] A. Singer. From graph to manifold laplacian: The convergence rate. *Appl. Comput. Harmon. Anal.,*, 21:128–134, 2006.

smazho05 [24] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *submitted*, 2005. retrievable at http://www.tti-c.org/smale.html.

smazho08 [25] S. Smale and D. Zhou. Geometry of probability spaces. *preprint*, 2008. retrievable at http://www.tti-c.org/smale.html.

vobebo04 [26] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.

yaroca07 [27] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

zwabla06 [28] L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *NIPS*, 2006.

Mikhail Belkin, Department of Computer Science and Engineering, Ohio State University

E-mail address: `mbelkin@cse.ohio-state.edu`

Ernesto De Vito, D.S.A., Università di Genova, Stradone Sant'Agostino, 37, 16123, Genova, Italy and INFN, Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy

E-mail address: `devito@dima.unige.it`

Lorenzo Rosasco, Center for Biological and Computational Learning, MIT, DISI, Università di Genova

E-mail address: `rosasco@disi.unige.it`