Journal of Machine Learning Research () Submitted 12/04; Published

On Learning with Integral Operators

Lorenzo Rosasco, LROSASCO@MIT.EDU
Center for Biological and Computational Learning, MIT,
Cambridge, MA, USA
& Dipartimento di Informatica e Scienze dell'Informazione,
Università di Genova, Italy

Mikhail Belkin, MBELKIN@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering,
Ohio State University, U.S.A.

Ernesto De Vito DEVITO@DIMA.UNIGE.IT
Dipartimento di Scienze per l'Architettura,
Università di Genova, Italy
& INFN, Sezione di Genova, Italy

Editor:

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 (von Luxburg et al., 2008).
Keywords: spectral convergence, empirical operators, learning integral operators, perturbation methods

1. Introduction

A broad variety of methods for machine learning and data analysis from Principal Components Analysis (PCA) to Kernel PCA, 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 this paper, we propose a method for analyzing empirical operators based on concentration inequalities in Hilbert spaces. This

technique together with perturbation theory results 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 (De Vito et al., 2005, Yao et al., 2007, Bauer et al., 2007, Smale and Zhou, 2005). Here we develop on this approach to provide a detailed and comprehensive study of perturbation results for empirical estimates of integral operators as well as empirical graph Laplacians.

In recent years several works started considering these connections. The first study of this problem appeared in (Koltchinskii and Gine', 2000, Koltchinskii, 1998), where the authors consider integral operators defined by a kernel. In (Koltchinskii and Gine', 2000) the authors study the relation between the spectrum of an integral operator with respect to a probability distribution and its (modified) empirical counterpart in the framework of U-statistics. In particular they prove that the ℓ_2 distance between the two (ordered) spectra goes to zero under the assumption that the kernel is symmetric and square integrable. Moreover, under some stronger conditions, rates of convergence and distributional limit theorems are obtained. The results are based on inequalities due to Lidskii and to Wielandt for finite dimensional matrices and the Marcinkiewicz law of large numbers. In (Koltchinskii, 1998) similar results were obtained for convergence of eigenfunctions and, using the triangle inequality, for spectral projections. These investigations were continued in (Mendelson and Pajor, 2005, 2006), where it was shown that, under the assumption that the kernel is of positive type, the problem of eigenvalue convergence reduces to the study of how the random operator $\frac{1}{n}\sum_i X_i \otimes X_i$ deviates from its average $\mathbb{E}[X \otimes X]$, with respect to the operator norm, where $X, X_1, \ldots, X_n$ are i.i.d ℓ_2 random vectors. The result is based on a on a symmetrization technique and on the control of a suitable Radamacher complexity.
The above studies are related to the problem of consistency of kernel PCA considered in (Shawe-Taylor et al., 2002, 2004) and refined in (Blanchard et al., 2006, Zwald and Blanchard, 2006). In particular, (Shawe-Taylor et al., 2002, 2004) study the deviation of the sum of the all but the largest k eigenvalues of the empirical matrix to its mean using McDiarmid inequality. The above result is improved in (Blanchard et al., 2006) where fast rates are provided by means of a localized Rademacher complexities approach. The results in (Zwald and Blanchard, 2006) are a development of the results in (Koltchinskii, 1998). Using a new perturbation results the authors study directly the convergence of the whole subspace spanned by the first k eigenvectors and are able to show that only the gap between the k and $k+1$ eigenvalue affects the estimate. All the above results hold for symmetric, positive definite kernels.

A second related series of works considered convergence of the graph Laplacian in various settings , see for example (Belkin, 2003, Lafon, 2004, **?**, Hein et al., 2005, Hein, 2006, Singer, 2006, Gine' and Koltchinskii., 2006). These papers discuss convergence of the graph Laplacian directly to the Laplace-Beltrami operator. Convergence of the normalized graph Laplacian applied to a fixed smooth function on the manifold is discussed in (Hein et al., 2005, Singer, 2006, Lafon, 2004). Results showing uniform convergence over some function class are presented in (Hein, 2006, Gine' and Koltchinskii., 2006). Finally, convergence of eigenvalues and eigenfunctions for the case of the uniform distribution was shown in (Belkin and Niyogi, 2007).

Unlike these works, where the kernel function is chosen adaptively depending on the number of points, we will be primarily interested in convergence of the graph Laplacian to its continuous (population) counterpart for a *fixed* weight function. The work (Luxburg et al., 2004) studies the convergence of the second eigenvalue which is relevant in spectral clustering problems. These results are extended in (von Luxburg et al., 2008), where operators are defined on the space of continuous functions. The analysis is performed in the context of perturbation theory in Banach spaces and bounds on individual eigenfunctions are derived. The problem of out-of-sample extension is considered via a Nyström approximation argument. Working in Banach spaces the authors have only mild requirements for the weight function defining the graph Laplacian, at the price of having to do fairly complicated analysis.

Our contribution is twofold. In the first part of the paper, we assume that the kernel K is symmetric and positive definite. We start considering the problem of out-of-sample extension of the kernel matrix and discuss a singular value decomposition perspective on Nyström-like extensions. More precisely, we show that a finite rank (extension) operator acting on the reproducing kernel Hilbert space $\mathcal{H}$ defined by K can be naturally associated to the empirical kernel matrix: the two operators have same eigenvalues and related eigenvectors/eigenfunctions. The kernel matrix and its extension can be seen as compositions of suitable restriction and extension operators that are explicitly defined by the kernel. A similar result holds true for the asymptotic integral operator, whose restriction to $\mathcal{H}$ is a Hilbert-Schmidt operator. We can use concentration inequalities for operator valued random variables and perturbation results to derive concentration results for eigenvalues (taking into account the multiplicity), as well as for the sums of eigenvalues. Moreover, using a perturbation result for spectral projections, we derive finite sample bounds for the deviation between the spectral projection associated with the k largest eigenvalues. We recover several known results with simplified proofs, and derive new results.

In the second part of the paper, we study the convergence of the asymmetric normalized graph Laplacian to its continuous counterpart. To this aim we consider a fixed positive symmetric weight function satisfying some smoothness conditions. These assumptions allows us to introduce a suitable intermediate reproducing kernel Hilbert space $\mathcal{H}$, which is, in fact, a Sobolev Space. We describe explicitly restriction and extension operators and introduce a finite rank operator with spectral properties related to those of the graph Laplacian. Again we consider the law of large numbers for operator-valued random variables to derive concentration results for empirical operators. We study behavior of eigenvalues as well as the deviation of the corresponding spectral projections with respect to the Hilbert-Schmidt norm. To obtain explicit estimates for spectral projections we generalize the perturbation result in (Zwald and Blanchard, 2006) to deal with non-self-adjoint operators. From a technical point the main difficulty in studying the asymmetric graph Laplacian is that we no longer assume the weight function to be positive definite so that there is no longer a natural RKH space associated to it. In this case we have to deal with non-self-adjoint operators and the functional analysis becomes more involved. Comparing to (von Luxburg et al., 2008), we note that the RKH space $\mathcal{H}$ replaces the Banach space of continuous functions. Assuming some regularity assumption on the weight functions we can exploit the Hilbert space structure to obtain more explicit results. Among other things, we derive explicit convergence rates for a large class of weight functions. Finally we note that for the case of positive

definite weight function results similar to those presented here have been independently derived in the preprint (Smale and Zhou, 2008).

The plan of the paper follows. We start by 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 refine the results in (von Luxburg et al., 2008), which, to the best of our knowledge, is the only other paper to consider the problem so far.

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 in operator theory (see, e.g. (Lang, 1993)). In the following we let $A : \mathcal{H} \to \mathcal{H}$ be a (linear) bounded 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^* be its adjoint, which is 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 a Hilbert space (a fact which will be a key in our development) endowed with the scalar product $\langle A, B \rangle_{HS} = \sum_j \langle Ae_j, Be_j \rangle$ and we denote by $\|\cdot\|_{HS}$ 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 $\sum_{j\geq 1} \left\langle \sqrt{A^*A} e_j, e_j \right\rangle < \infty$ for some (any) Hilbert basis $(e_j)_{j\geq 1}$ (where $\sqrt{A^*A}$ is the square root of the positive operator A^*A defined by spectral theorem (Lang, 1993)). 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 Hilbert-Schmidt operator A and bounded operator B we have

$$\begin{aligned} \|AB\|_{HS} &\leq \|A\|_{HS}\|B\|, \\ \|BA\|_{HS} &\leq \|B\|\|A\|_{HS}. \end{aligned} \tag{1}$$

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 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., (Kato, 1966)) that if A is a compact operator, then $\sigma(A) \setminus \{0\}$ consists of a countable family of isolated points with finite multiplicity $|\lambda_1| \geq |\lambda_2| \geq \cdots$ and either $\sigma(A)$ is finite or $\lim_{n\to\infty} \lambda_n = 0$. If the operator A is self-adjoint ($A = A^*$, analogous to a symmetric matrix in the finite-dimensional case), the eigenvalues are real. Each eigenvalue λ has an associated *eigenspace* which is the span of the associated eigenvectors. 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 simple rectifiable curve $\Gamma \subset \mathbb{C}$ (with positive direction) 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 to 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 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 a unique symmetric, positive definite kernel function $K : X \times X \to \mathbb{C}$, called *reproducing kernel*, associated

to $\mathcal{H}$ and the following reproducing property holds

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

where $K_x := K(\cdot, x)$. It is also well known (Aronszajn, 1950) 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[1].

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 $|\xi_i| \leq C, \quad i = 1, \ldots, n$, then Hoeffding 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) $1 - 2e^{-\tau}$,

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

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

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

holds true with probability at least $1 - 2e^{-\tau}$ (Pinelis, 1992).

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 $\|f\|_\rho^2 = \langle f, f \rangle_\rho = \int_X |f(x)|^2 d\rho(x)$. If

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

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

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

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 $\mathbf{K}_{ij} = \frac{1}{n} K(x_i, x_j)$.

1. 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, (Aronszajn, 1950). The assumption about continuity is not strictly necessary, but it will simplify some technical part.

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 operators $L_{K,\mathcal{H}}, L_{K,n} : \mathcal{H} \to \mathcal{H}$ given by,

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

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

Note that $L_{K,\mathcal{H}}$ is the integral operator with kernel K with range and domain $\mathcal{H}$ rather than in $L^2(X,\rho)$. The reason for writing it in this seemingly complicated form is to make the parallel with (8) clear. To justify the "extension operator" in 8, consider the natural "restriction operator", $R_n : \mathcal{H} \to \mathbb{R}^n$, $R_n(f) = (f(x_1), \ldots, f(x_n))$. It is not hard to check that the adjoint operator $R_n^* : \mathbb{R}^n \to \mathcal{H}$ can be written as $R_n^*(y_1, \ldots, y_n)(\cdot) = \frac{1}{n}\sum y_i K(\cdot, x_i)$. Indeed, we see that

$$\begin{aligned} \langle f, R_n^*(y_1, \ldots, y_n)\rangle_{\mathcal{H}} &= \langle R_n(f), (y_1, \ldots, y_n)\rangle_{\mathbb{R}^n} \\ &= \frac{1}{n}\sum y_i f(x_i) = \frac{1}{n}\sum y_i \langle f, K(\cdot, x_i)\rangle_{\mathcal{H}}, \end{aligned}$$

where $\mathbb{R}^n$ is endowed with $1/n$ times the euclidean scalar product. Thus, we observe that $L_{K,n} = R_n^* \circ R_n$ is the composition of the restriction operator and its adjoint. On the other hand for the operator $\mathbf{K}$ on $\mathbb{R}^n$ we have that $\mathbf{K} = R_n \circ R_n^*$. Similarly, if $R_{\mathcal{H}}$ denotes the inclusion $\mathcal{H} \hookrightarrow L^2(X,\rho)$, $L_{K,\mathcal{H}} = R_{\mathcal{H}}^* \circ R_{\mathcal{H}}$.

In the next subsection, we discuss a parallel with the Singular Value Decomposition for matrices and demonstrate that $L_{K,\mathcal{H}}$ and L_K have the same eigenvalues (possibly, up to some zero eigenvalues) and the corresponding eigenfunctions are closely related. 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 (De Vito et al., 2005).

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 $\|\langle \cdot, K_x \rangle K_x\|_{HS}^2 = \|K_x\|^4 \leq \kappa^4$. Hence, using (7), $\|L_{K,\mathcal{H}}\|_{HS} \leq \kappa^2$ and

$$\|\xi_i\|_{HS} \leq 2\kappa^2, \qquad i = 1, \ldots, n.$$

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 subsection 3.2. However before doing that we provide a discussion of the Nyström extension needed to properly compare the above operators.

3.1 Extension operators

We will now briefly revisit the Nystorm extension and clarify some connections to the Singular Value Decomposition (SVD) for operators. Recall that applying SVD to a $m \times p$ matrix A produces a *singular system* consisting of singular (strictly positive) values $(\sigma_j)_{j=1}^k$, and vectors $(u_j)_{j=1}^m \in \mathbb{R}^m$ and $(v_j)_{j=1}^p \in \mathbb{R}^p$ (where k is the rank of A) such that they form orthonormal basis of $\mathbb{R}^m$ and $\mathbb{R}^p$ respectively and such that

$$\begin{cases} A^*Au_j = \sigma_j u_j & j = 1, \dots k \\ A^*Au_j = 0 & j = k+1, \dots, m \end{cases} \quad \text{and} \quad \begin{cases} AA^*v_j = \sigma_j v_j & j = 1, \dots k \\ A^*Au_j = 0 & j = k+1, \dots, p \end{cases}$$

It is not hard to see that the matrix A can be written as $A = U\Sigma V$, where U and V are matrices obtained by "stacking" u's and v's, and Σ is a $m \times p$ matrix having the singular values σ_i on the first k-entries on the diagonal (and zero outside), so that $Au_i = \sqrt{\sigma_j}v_j$ and $A^*v_j = \sqrt{\sigma_j}u_j$, which is the formulation we will use in this paper. The same formalism applies more generally to operators and allows us to connect the spectral properties of L_K and $L_{K,\mathcal{H}}$ as well as the matrix $\mathbf{K}$ and the operator $L_{K,n}$. The basic idea is that each of these pairs (as shown in the previous subsection) corresponds to a singular system and thus share eigenvalues (up to some zero eigenvalues) and have eigenvectors related by a simple equation. Indeed the following result can obtained considering the SVD decomposition associated to $R_{\mathcal{H}}$ (and proposition 3 considering the SVD decomposition associated to R_n). The proof of the following proposition can be deduced from the results in (De Vito et al., 2005, 2006) .

Proposition 2 *The following facts hold true.*

1. *The operators L_K and $L_{K,\mathcal{H}}$ are positive, self-adjoint and trace class. In particular both $\sigma(L_K)$ and $\sigma(L_{K,\mathcal{H}})$ are contained in $[0, \kappa^2]$.*

2. *The spectra of L_K and $L_{K,\mathcal{H}}$ are the same, possibly up to the zero, moreover if σ is a nonzero eigenvalue and u, v associated eigenfunctions of L_K and $L_{K,\mathcal{H}}$ (normalized to norm 1 in $L^2(X,\rho)$ and $\mathcal{H}$) respectively, then*

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

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

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

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 eigenfunctions 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. Moreover, the extension/restriction operations preserve both the normalization and orthogonality of the eigenfunctions. An analogous result relates the matrix $\mathbf{K}$ and the operator $L_{K,n}$.

Proposition 3 *The following facts hold:*

1. *The finite rank operator $L_{K,n}$ is Hilbert-Schmidt and the matrix $\mathbf{K}$ are positive, self-adjoint. In particular the spectrum $\sigma(L_{K,n})$ has only finitely many nonzero elements and is contained in $[0,\kappa^2]$.*

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 $\hat{\sigma}$ is a non zero eigenvalue and $\hat{u}, \hat{v}$ are the corresponding eigenvector and eigenfunction of $\mathbf{K}/n$ and $L_{K,n}$ (normalized to norm 1 in $\mathbb{R}^n$ and $\mathcal{H}$) respectively, then*

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

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

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

where the sum runs over the nonzero eigenvalues, the family $(\hat{u}_j)_{j\geq 1}$ is an orthonormal 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

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

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 (Kato, 1987), which is given for an arbitrary selfadjoint operator, to the compact operators If A is such an operator, an extended enumeration is a sequence of real numbers where every nonzero eigenvalue of A appears exactly according to its multiplicity and the other values (if any) are zero. An enumeration 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 (Kato, 1987) is an extension to infinite dimensional operators of an inequality due to Lidskii for finite rank operator.

Theorem 4 (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

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

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.

Proposition 5 *Let $(\sigma_j)_{j\geq 1}$ be the decreasing enumeration of discrete eigenvalues for $L_{K,\mathcal{H}}$ and $(\hat{\sigma}_j)_{j\geq 1}$ the extended decreasing enumeration of discrete eigenvalues for $L_{K,n}$. 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

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

The following result can be deduced by Theorem 4 with $p = 1$ and Theorem 1, however a simpler direct proof is given below.

Proposition 6 *Under the assumption of Proposition 5 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 $1 - 2e^{-\tau}$

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

■

To control the spectral projections associated to one or more eigenvalues we need the following perturbation result, proof is given in (Zwald and Blanchard, 2006) (see also Theorem 15 in Section 4.3). If A is a positive compact operator such that $\sigma(A)$ is infinite, for an $N \in \mathbb{N}$, let P_N^A be the orthogonal projection on the eigenvectors corresponding to the top N eigenvalues.

Proposition 7 *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.

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

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

As a consequence of the above proposition and Theorem 1, we can derive a probabilistic bound on eigen-projections. Assume for the sake of simplicity, that the cardinality of $\sigma(L_K)$ is infinite.

Theorem 8 *Let $(\sigma_j)_{j\geq 1}$ be the decreasing enumeration of discrete eigenvalues for $L_{K,\mathcal{H}}$ and 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 P_N.

4. Asymmetric Graph Laplacian

In this section we will consider the case of the so-called asymmetric normalized graph Laplacian, which is the identity matrix minus the transition matrix for the natural random walk on a graph. In such a random walk, the probability of leaving a vertex along a given edge is proportional to the weight of that edge. As before, we will be interested in a specific class of graphs (matrices) associated to data.

Let $W : X \times X \to \mathbb{R}^+$ be a symmetric continuous (weight) function. Note that we will not require W to be a positive definite kernel, but only a positive function. A set of data points $\mathbf{x} = (x_1, \ldots, x_n) \in X$ 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 $\mathbf{L}_r : \mathbb{R}^n \to \mathbb{R}^n$ 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 $\mathbf{D}_{ii} = \frac{1}{n}\sum_{j=1}^n W(x_i, x_j)$.

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.

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*. We see that when a set $\mathbf{x} = (x_1, \ldots, x_n) \in X$ is sampled i.i.d. according to ρ, the matrix $\mathbf{L}_r$ is an empirical version of the operator L_r.

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 that 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

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

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

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

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

It is possible to show (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 $\mathbf{L}_r$ are also closely related. In particular we will see in the following that to relate the spectral properties of L_r and $\mathbf{L}_r$ 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

In analogy to Section 3.1 we consider the relation between the operators we want to study and their extensions. In this case the SVD argument does not apply in a straightforward way but we can still define a restriction operator $R_n : \mathcal{H} \to \mathbb{R}^n$, $R_n(f)_i = f(x_i) = \langle K(x_i,\cdot), f\rangle$ for all $i = 1, \ldots, n$, and an extension operator $E_n : \mathbb{R}^n \to \mathcal{H}$ that is now written as $E_n(y_1, \ldots, y_n)(\cdot) = \frac{1}{n} \sum y_i W(\cdot, x_i)/m_n(\cdot)$. Clearly the extension operator is no longer the adjoint of R_n but the connection among the operators $\mathbf{L}_r$ to $L_{r,\mathcal{H},n}$ and A_n can still be clarified by means of R_n, E_n. Indeed it is easy to check that $A_n = R_n E_n$ and $\mathbf{D}^{-1}\mathbf{W} = E_n R_n$. Similarly the infinite sample restrictions and extension operators can be defined to relate the operators L_r, $A_{\mathcal{H}}$ and $L_{r,\mathcal{H}}$. The next proposition considers such a connection.

Proposition 9 *The following facts hold true.*

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 $\sigma(A_{\mathcal{H}}) = 1 - \sigma(L_{r,\mathcal{H}})$.*
3. *The spectra of L_r and $L_{r,\mathcal{H}}$ are the same, moreover if $\sigma \neq 1$ is an eigenvalue and u, v associated eigenfunctions of L_r and $L_{r,\mathcal{H}}$ respectively, then*

$$\begin{aligned} u(x) &= v(x) \qquad \text{for almost all } x \in X \\ v(x) &= \frac{1}{1-\sigma} \int_X \frac{W(x,t)}{m(x)} u(t) \, d\rho(t) \end{aligned}$$

4. Finally the following decompositions hold

$$L_r = \sum_{j\geq 1} \sigma_j P_j + P_0, \tag{11}$$

$$L_{r,\mathcal{H}} = I - \sum_{j\geq 1} (1-\sigma_j) Q_j + D, \tag{12}$$

where $\{\sigma_i \mid i \geq i\} = \sigma(L_r) \setminus \{1\}$*, the projections* Q_j, P_j *are the spectral projections of* L_r *and* $L_{r,\mathcal{H}}$ *associated to the eigenvalue* σ_j*,* P_0 *is the spectral projection of* L_r *associated with the eigenvalue* 1*, and* D *is a quasi-nilpotent operator such that* $\ker D = \ker (I - L_{r,\mathcal{H}})$ *and* $Q_j D = D Q_j = 0$ *for all* $j \geq 1$*.*

The proof of the above result is long and quite technical and can be found in Appendix A. Note that, with respect to Proposition 3, neither the normalization nor the orthogonality is preserved by the extension/restriction operations. However, one can easily shows that, if $u_1, \ldots, u_m$ is a linearly independent family of eigenfunctions of L_r with eigenvalues $\sigma_1, \ldots, \sigma_m \neq 1$, then the extension $v_1, \ldots, v_m$ is a linearly independent family of eigenfunctions of $L_{r,\mathcal{H}}$ with eigenvalues $\sigma_1, \ldots, \sigma_m \neq 1$. Finally, we stress that in item 4 both series converge in the strong operator topology, however, tough $\sum_{j\geq 1} P_i = I - P_0$, it is not true that $\sum_{j\geq 1} Q_i$ converges to $I - Q_0$, where Q_0 is the spectral projection of $L_{r,\mathcal{H}}$ associated to the eigenvalue 1. This is the reason why we need to write the decomposition of $L_{r,\mathcal{H}}$ as in (12) instead of (11). An analogous result allows us to relate $\mathbf{L}_r$ to $L_{r,\mathcal{H},n}$ and A_n.

Proposition 10 *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* $\sigma(A_n) = 1 - \sigma(L_{r,\mathcal{H},n})$*.*

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* $L_{r,\mathcal{H},n}$*, then*

$$\hat{u}^i = \hat{v}(x_i)$$
$$\hat{v}(x) = \frac{1}{1-\hat{\sigma}} \sum_{i=1}^{n} \frac{W(x, x_i)}{m_n(x)} \hat{u}^i$$

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

4. *Finally the following decompositions hold*

$$\mathbf{L}_r = \sum_{j=1}^{m} \hat{\sigma}_j \hat{P}_j + \hat{P}_0,$$

$$L_{r,\mathcal{H},n} = \sum_{j=1}^{m} \hat{\sigma}_j \hat{Q}_j + \hat{Q}_0 + \hat{D},$$

where $\{\hat{\sigma}_1, \dots \hat{\sigma}_m\} = \sigma(L_r) \setminus \{1\}$*, the projections* Q_j, P_j *are the spectral projections of* $\mathbf{L}_r$ *and* $L_{r,\mathcal{H},n}$ *associated to the eigenvalue* σ_j*,* $\hat{P}_0$ *and* $\hat{Q}_0$ *are the spectral projections of* $\mathbf{L}_r$ *and* $L_{r,\mathcal{H},n}$ *associated with the eigenvalue* 1*, and* D *is a quasi-nilpotent operator such that* $\ker \hat{D} = \ker (I - L_{r,\mathcal{H},n})$ *and* $\hat{Q}_j \hat{D} = \hat{D} \hat{Q}_j = 0$ *for all* $j = 1, \dots, m$*.*

The last decomposition is parallel to the Jordan canonical form for (non-symmetric) matrices. Notice that, since the sum is finite, $\sum_{j=1}^{m} \hat{Q}_j + \hat{Q}_0 = I$.

4.2 Graph Laplacian Convergence for Smooth Weight Functions

The assumption $W(x, \cdot)/m(\cdot) \in \mathcal{H}$ is crucial and is not satisfied in general. For example, it is not necessarily verified even if $W(x, y)$ is a reproducing kernel. however it holds true 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. 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 (with respect to the uniform measure) bounded away from zero. Recall that for $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$ 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 to allow $s \in]0, +\infty[$.

The Sobolev Embedding theorem ensures[2] 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

$$\|f\|_\infty \leq C_1 \|f\|_s. \tag{13}$$

Then $\mathcal{H}^s$ is a RKHS with reproducing kernel $K^s(x, y)$, so that $f(x) = \langle f, K^s_x \rangle_s$ 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 (Burenkov, 1998).

Lemma 11 *Let* $g \in \mathcal{C}^s(X)$*, where all derivatives are bounded up to order* s*. The multiplication 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*

$$\|M_g\| \leq a \|g\|_{s'} < \infty, \tag{14}$$

for some positive constant a*.*

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

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.

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

The following theorem establishes the desired result.

Theorem 12 *If assumption A2 holds, then for some positive constant C with confidence $1 - 2e^{-\tau}$ we have*

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

To prove Theorem 12 we need the following preliminary estimates.

Proposition 13 *The operators $L_{W,\mathcal{H}}, L_{W,n} : \mathcal{H}_s \to \mathcal{H}_s$ defined by*

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

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

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

■

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

Proposition 14 *Let $M, M_n : \mathcal{H}^s \to \mathcal{H}^s$ be defined by $Mf(x) = m(x)f(x)$ and $M_nf(x) = m_n(x)f(x)$. Then M, M_n are linear operators bounded by C_2 and with confidence $1 - 2e^{-\tau}$*

$$\|M - M_n\| \leq \frac{2C_2a\sqrt{2\tau}}{\sqrt{n}}.$$

where is a positive constant.

Proof It follows from (16), (14) 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 simplicity).
The random variables $(\xi_i)_{i=1}^n$, defined by $\xi_i = W_{x_i} - m$ are zero mean and bounded by $2C_2a$. Applying (4) we have with high probability

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

It follows from (14) that

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

■

Finally, we can combine the above two propositions to get the proof of Theorem 12.
Proof [Proof of Theorem 12] It follows from (16), (14) 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

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

Recalling (1), we consider the Hilbert-Schmidt norm of the above expression. Using the inequalities (14), (15), (14) and the assumption A3 we see that there is a constant C, such that

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

■

In the next section we discuss the implications of the above results in terms of concentration of eigenvalues and spectral projections.

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. See the appendix for a short review about spectral theory for compact (not necessarily self-adjoint) operators. The following theorem is an adaptation of results in (Anselone, 1971).

Theorem 15 *Let A be a compact operator. Given a finite set Λ of non-zero eigenvalues of A, let Γ be any simple rectifiable closed curve (having positive direction) with Λ inside and $\sigma(A) \setminus \Lambda$ outside. Let P be the spectral projection associated to Λ, that is,*

$$P = \frac{1}{2\pi i}\int_\Gamma (\lambda - A)^{-1}\, d\lambda,$$

and define

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

Let B be another compact operator such that

$$\|B - A\| \leq \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi} < \delta,$$

then the following facts hold true.

1. *The curve Γ is a subset of the resolvent of B enclosing a finite set $\widehat{\Lambda}$ of non-zero eigenvalues of B;*
2. *Denoting by $\widehat{P}$ the spectral projection of B associated to $\widehat{\Lambda}$, then*

$$\|\widehat{P} - P\| \leq \frac{\ell(\Gamma)}{2\pi\delta} \frac{\|B - A\|}{\delta - \|B - A\|};$$

3. *The dimension of the range of P is equal to the dimension of the range of $\hat{P}$.*

Moreover, if $B - A$ is a Hilbert-Schmidt operator, then

$$\|\widehat{P} - P\|_{HS} \leq \frac{\ell(\Gamma)}{2\pi\delta} \frac{\|B - A\|_{HS}}{\delta - \|B - A\|}.$$

We postpone the proof of the above result to Appendix A.

Here we note that, if A is self-adjoint, then spectral theorem ensures that

$$\delta = \min_{\lambda \in \Gamma, \sigma \in \Lambda} |\lambda - \sigma|.$$

The above theorem together with the results obtained in the previous section allows to derive several results.

Proposition 16 *Let σ be an eigenvalue of L_r, $\sigma \neq 1$, with multiplicity m. For any $\varepsilon > 0$ and $\tau > 0$, there exists an integer n_0 and a positive constant R such that, if the number of examples is greater than n_0, with probability greater than $1 - 2e^{-\tau}$,*

1. *there are $\hat{\sigma}_1, \ldots, \hat{\sigma}_m$ (possibly repeated) eigenvalues of the matrix $\mathbf{L}_r$ satisfying*

$$|\hat{\sigma}_i - \sigma| \leq \varepsilon \qquad \text{for all } i = 1, \ldots, m.$$

2. *for any normalized eigenvector $\hat{u} \in \mathbb{R}^n$ of $\mathbf{L}_r$ with eigenvalue $\hat{\sigma}_i$ for some $i = 1, \ldots, m$, there exists an eigenfunction $u \in \mathcal{H}^s \subset L^2(X, \rho)$ of L_r with eigenvalue σ, satisfying*

$$\|E_n(\hat{u}) - u\|_s \leq R \frac{\sqrt{\tau}}{\sqrt{n}},$$

where $E_n(\hat{u})(x) = \frac{1}{1-\hat{\sigma}_i} \frac{1}{m_n(x)} \frac{1}{n} \sum_{j=1}^n W(x, x_j) \hat{u}^j$.

Proof We apply Theorem 15 with $A = A_{\mathcal{H}}$, $B = A_n$ and $\Gamma = \{\lambda \in \mathbb{C} \mid |\lambda - (1-\sigma)| = \varepsilon\}$. Since $A_{\mathcal{H}}$ is compact and assuming ε small enough, we have that $\Lambda = \{1-\sigma\}$. Let $n_0 \in \mathbb{N}$ such that

$$\frac{C\sqrt{\tau}}{\sqrt{n_0}} \leq \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi} \qquad \text{where} \quad \delta^{-1} = \sup_{\lambda\in\Gamma} \|(\lambda - A_{\mathcal{H}})^{-1}\|.$$

By Theorem 12, with probability greater than $1 - 2e^{-\tau}$, for all $n \geq n_0$

$$\|A_n - A_{\mathcal{H}}\| \leq \|A_n - A_{\mathcal{H}}\|_{HS} \leq \frac{C\sqrt{\tau}}{\sqrt{n}} \leq \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi}.$$

Item 1 of Theorem 15 with Proposition 10 ensures that $\widehat{\Lambda} = \{1 - \hat{\sigma}_1, \ldots, 1 - \hat{\sigma}_m\}$, so that $|\hat{\sigma}_i - \sigma| < \varepsilon$ for all $i = 1, \ldots, m$.
Let now $\hat{u} \in \mathbb{R}^n$ be a normalized vector such that $\mathbf{L}_r \hat{u} = \hat{\sigma}_i \hat{u}$ for some $i = 1, \ldots, m$. Then from Proposition 10, $\hat{v} = E_n(\hat{u})$ is an eigenfunction of A_n with eigenvalue $1 - \hat{\sigma}$, so that $\hat{Q}\hat{v} = \hat{v}$ where $\hat{Q}$ is the spectral projection of A_n associated to $\hat{\Lambda}$. Let Q be the spectral projection of $A_{\mathcal{H}}$ associated to $1 - \sigma$ and define $u = Q\hat{v} \in \mathcal{H}^s$. By definition of Q, $Au = (1-\sigma)u$. Since $\mathcal{H}^s \subset L^2(X, \rho)$, Proposition 9 ensures that $L_r u = \sigma u$. Item 2 of Theorem 15 gives that

$$\begin{aligned}\|\hat{v} - u\|_s = \|\hat{Q}\hat{v} - Q\hat{v}\|_s &\leq \|\hat{Q} - Q\| \|E_n(\hat{u})\| \leq \|E_n\| \frac{\ell(\Gamma)}{2\pi\delta} \frac{\|A_n - A_{\mathcal{H}}\|}{\delta - \|A_n - A_{\mathcal{H}}\|} \\ &\leq C_2 C_4 \frac{\delta + \ell(\Gamma)/2\pi}{\delta^2} \|A_n - A_{\mathcal{H}}\| \leq R\frac{\tau}{\sqrt{n}},\end{aligned}$$

where $R = C_2 C_4 \frac{\delta + \ell(\Gamma)/2\pi}{\delta^2} C$, C is the constant given in Theorem 12, the constants C_2, C_4 are given in Assumption 2, and we use that $\|A_n - A_{\mathcal{H}}\| \leq \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi}$. ■

We add the following remark.

Remark 17 *By inspecting the above proof, if $A_{\mathcal{H}}$ is selfadjoint, then $n_0 \geq \frac{C^2\tau}{\varepsilon^2}$ provided that $\varepsilon < \min_{\sigma' \in \sigma(L_r), \sigma' \neq \sigma} |\sigma' - \sigma|$.*

Next we consider convergence of the spectral projections of $A_{\mathcal{H}}$ and A_n associated with the first$N$-eigenvalues. For sake of simplicity, we assume that the cardinality of $\sigma(A_{\mathcal{H}})$ is infinite.

Proposition 18 *Consider the first N eigenvalues of $A_{\mathcal{H}}$. There exist an integer n_0 and a constant $\hat{R} > 0$, depending on N and $\sigma(A_{\mathcal{H}})$, such that, with confidence $1 - 2e^{-\tau}$ and for any $n \geq n_0$,*

$$\|P_N - \hat{P}_D\|_{HS} \leq \frac{\hat{R}\sqrt{\tau}}{\sqrt{n}},$$

where $P_N, \hat{P}_D$ are the eigenprojections corresponding to the first N eigenvalues of $A_{\mathcal{H}}$ and D eigenvalues of A_n, and D is such that the sum of the multiplicity of the first D eigenvalues of A_n is equal to the sum of the multiplicity of the first N eigenvalues of $A_{\mathcal{H}}$.

Proof The proof is close to the one of previous proposition. We apply Theorem 15 with $A = A_{\mathcal{H}}$, $B = A_n$ and the curve Γ equal to the boundary of the rectangle

$$\{\lambda \in \mathbb{C} \mid \frac{\alpha_N + \alpha_{N+1}}{2} \leq \Re e(\lambda) \leq \|A\| + 2,\ |\Im m(\lambda)| \leq 1\},$$

where α_N is the N-largest eigenvalue of $A_{\mathcal{H}}$ and α_{N+1} the $N+1$-largest eigenvalue of $A_{\mathcal{H}}$. Clearly Γ encloses the first N largest eigenvalues of $A_{\mathcal{H}}$, but no other points of $\sigma(A)$. Define $\delta^{-1} = \sup_{\lambda\in\Gamma}\|(\lambda - A_{\mathcal{H}})^{-1}\|$ and $n_0 \in \mathbb{N}$ such that

$$\frac{C\sqrt{\tau}}{\sqrt{n_0}} \leq \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi} \qquad \text{and} \qquad \frac{C\sqrt{\tau}}{\sqrt{n_0}} < 1.$$

As in the above corollary, with probability greater than $1 - 2e^{-\tau}$, for all $n \geq n_0$

$$\|A_n - A_{\mathcal{H}}\| \leq \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi} \qquad \text{and} \qquad \|A_n - A_{\mathcal{H}}\| < 1.$$

The last inequality ensures that the largest eigenvalues of A_n is smaller than $1 + \|A_{\mathcal{H}}\|$, so that by Theorem 15, the curve Γ encloses the first D-eigenvalues of A_n, where D is equal to the sum of the multiplicity of the first N eigenvalues of $A_{\mathcal{H}}$. The proof is finished letting $\hat{R} = \frac{\delta+\ell(\Gamma)/2\pi}{\delta^2}C$. ■

Acknowledgments

Ernesto De Vito and Lorenzo Rosasco have been partially supported by the FIRB project RBIN04PARL and by the EU Integrated Project Health-e-Child IST-2004-027749. Mikhail Belkin is partially supported by the NSF Early Career Award 0643916.

Appendix A. Some Proofs

We start giving the proof of Proposition 9.

Proof [of Proposition 9]

We first need some preliminary observations. By Assumption 1 the measure $\rho_W = m\rho$, has density m w.r.t. ρ, and is equivalent[3] to ρ and the spaces $L^2(X,\rho)$ and $L^2(X,\rho_W)$ are the same vector space but they are endowed with different norm/scalar products (so that functions that are orthogonal in one space might not be orthogonal in the other). In particular the eigenvalues of L_r are the same if we regarded it as an operator from and to $L^2(X,\rho_W)$ or as an operator from and to $L^2(X,\rho)$. Moreover 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.

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

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

3. Two measures are equivalent if they have the same null sets.

where $\kappa^2 = \sup_{x\in X} K(x,x)$. The operator $I_W^* : L^2(X,\rho_W) \to \mathcal{H}$ defined by

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

is linear and bounded since, by Assumption 1, $\sup_{x,t\in X} \frac{W(x,t)}{m(x)} < +\infty$. 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 $I_W^* I_K$ and $I_K I_W^*$ are Hilbert-Schmidt operators since they are composition of a bounded operator and Hilbert-Schmidt operator. Again by a direct computation we have that

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

Moreover, let $\sigma \neq 1$ and $v \in \mathcal{H}$ with $v \neq 0$ such that $L_{r,\mathcal{H}} v = \sigma v$. Letting $u = I_K v$, then

$$L_r u = (I - I_K I_W^*) I_K v = I_K L_r v = \sigma u \quad \text{and} \quad I_W^* u = I_W^* I_K v = (1-\sigma) v \neq 0,$$

so that $u \neq 0$ and u is an eigenfunction of L_r with eigenvalue σ. Similarly we can prove that if $\sigma \neq 1$ and $u \in L^2(X,\rho), u \neq 0$ is such that $L_r u = \sigma u$, then $v = \frac{1}{1-\sigma} I_W^* u$ is different from zero and is an eigenfunction of $L_{r,\mathcal{H}}$ with eigenvalue σ,

We now show that L_r and $L_{r,\mathcal{H}}$ have positive eigenvalues. Towards this end, we 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.
Note that since $I_K I_W^*$ is a self adjoint operator on $L^2(X,\rho_W)$, it can be decomposed as

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

where for all j, $P_j : L^2(X,\rho_W) \to L^2(X,\rho_W)$ is the spectral projection of $I_K I_W^*$ associated to the eigenvalue $1-\sigma_j \neq 0$. Moreover note that P_j is also the spectral projection of L_r associated to the eigenvalue $\sigma_j \neq 1$. By definition P_j satisfies:

$$\begin{aligned}
P_j^2 &= P_j, \\
P_j^* &= P_j, \\
P_j P_i &= 0, \quad i \neq j, \\
P_j P_{ker(I_K I_W^*)} &= 0 \\
\sum_{j\geq 1} P_j &= I - P_{ker(I_K I_W^*)} = I - P_0
\end{aligned}$$

where $P_{ker(I_K I_W^*)}$ is the projection on the kernel of $I_K I_W^*$, that is, the projection P_0. Moreover the sum in the last equation converges in the strong operator topology. In particular we have

$$I_K I_W^* P_j = P_j I_K I_W^* = (1-\sigma_j)P_j,$$

so that

$$Ł_r = I - I_K I_W^* = \sum_{j\geq 1}\sigma_j P_j + P_0.$$

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,

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

Moreover,

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

so that

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

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

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

and, similarly $DQ_j = 0$. By construction $\sigma(D) = 0$, that is, D is a quasi-nilpotent operator. Equation (12) is now clear as well as the fact that $\ker D = \ker (I - L_{r,\mathcal{H}})$. ■

Proof [Proof of Proposition 10] 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 15.

Proof [Proof of Theorem 15] We recall the following basic result. Let S and T two bounded operators acting on $\mathcal{H}$ and defined $C = I - ST$. If $\|C\| < 1$, then T has a bounded inverse and

$$T^{-1} - S = (I - C)^{-1} CS$$

where we note that $\|I - C\|^{-1} \leq \frac{1}{1-\|C\|}$ since $\|C\| < 1$.

Let A and B two compact operators. Let Γ be a compact subset of the resolvent of A and define

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

which is finite since Γ is compact. Assume that

$$\|B - A\| < \delta,$$

then for any $\lambda \in \Gamma$

$$\|(\lambda - A)^{-1}(B - A)\| \leq \|(\lambda - A)^{-1}\| \|B - A\| \leq \delta^{-1}\|B - A\| < 1.$$

Hence we can apply the above result with $S = (\lambda - A)^{-1}$, $T = (\lambda - B)$, since

$$C = I - (\lambda - A)^{-1}(\lambda - B)^{-1} = (\lambda - A)^{-1}(B - A).$$

It follows that $(\lambda - B)$ has a bounded inverse and

$$(\lambda - B)^{-1} - (\lambda - A)^{-1} = (I - (\lambda - A)^{-1}(B - A))^{-1}(\lambda - A)^{-1}(B - A)(\lambda - A)^{-1}.$$

In particular, Λ is a subset of the resolvent of B and, if $B - A$ is a Hilbert-Schmidt operator, so is $(\lambda - B)^{-1} - (\lambda - A)^{-1}$.

We choose as Λ be a finite set of non-zero eigenvalues. Let Γ be any simple closed curve with Λ inside and $\sigma(A) \setminus \Lambda$ outside. Let P be the spectral projection associated with Λ, then

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

Applying the above result, it follows that Γ is a subset of the resolvent of B and we let $\widehat{\Lambda}$ be the subset of $\sigma(B)$ inside Γ and $\widehat{P}$ the corresponding spectral projection, then

$$\begin{aligned}\widehat{P} - P &= \frac{1}{2\pi i}\int_\Gamma (\lambda - B)^{-1} - (\lambda - A)^{-1}\, d\lambda \\ &= \frac{1}{2\pi i}\int_\Gamma (I - (\lambda - A)^{-1}(B - A))^{-1}(\lambda - A)^{-1}(B - A)(\lambda - A)^{-1}\, d\lambda.\end{aligned}$$

It follows that

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

In particular if $\|B - A\| \leq \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi} < \delta$, $\|\widehat{P} - P\| \leq 1$ so that the dimension of the range of P is equal to the dimension of the range of $\hat{P}$. It follows that $\widehat{\Lambda}$ is not empty.
If $B - A$ is a Hilbert-Schmidt operator, we can replace the operator norm with the Hilbert-Schmidt norm, and the corresponding inequality is a consequence of the fact that the Hilbert-Schmidt operator are an ideal. ■

Appendix B. 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, (Kato, 1966)

Proposition 19 *The spectrum of a compact operator A is a countable compact subset of $\mathbb{C}$ with no accumulation point different from zero, that is,*

$$\sigma(A) \setminus \{0\} = \{\lambda_i \mid i \geq 1,\ \lambda_i \neq \lambda_j \text{ if } i \neq j\}$$

with $\lim_{i\to\infty} \lambda_i = 0$ if the cardinality of $\sigma(A)$ is infinite. 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 rectifiable closed simple curve (with positive direction) 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 rectifiable closed simple curve (with positive direction) enclosing $\lambda_1, \ldots, \lambda_N$, but no other points of $\sigma(A)$.

References

Philip 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.

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

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, 2003.

Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 129–136. MIT Press, Cambridge, MA, 2007.

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

V.I. Burenkov. *Sobolev spaces on domains*. B. G. Teubuer, Stuttgart-Leipzig, 1998.

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.

Ernesto De Vito, Lorenzo Rosasco, and Andrea Caponnetto. Discretization error analysis for Tikhonov regularization. *Anal. Appl. (Singap.)*, 4(1):81–99, 2006.

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

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

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.

T. Kato. *Perturbation theory for linear operators*. Springer, Berlin, 1966.

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

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

V. Koltchinskii and E. Gine'. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6:113–167, 2000.

S. Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, 2004.

S. Lang. *Real and Functional Analysis*. Springer, New York, 1993.

Ulrike Von Luxburg, Olivier Bousquet, and Mikhail Belkin. On the convergence of spectral clustering on random samples: the normalized case. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT 2004*, pages 457–471. Springer, 2004.

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

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

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.

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.

John Shawe-Taylor, Chris Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and the generalisation error of kernel pca. *to appear in IEEE Transactions on Information Theory*, 51, 2004. URL `http://eprints.ecs.soton.ac.uk/9779/`.

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

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

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

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.

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

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