A Note on Learning with Integral Operators

Lorenzo Rosasco[1,2], **Mikhail Belkin**[3], **Ernesto De Vito**[4] *
1- Università degli studi di Genova - DISI, via Dodecaneso 35, Genova - Italy
2 - CBCL - MIT, Cambridge, MA 02139 - USA
3- Department of Computer Science and Engineering, Ohio State University, USA
4- DSA, Università di Genova, Italy & INFN, Sezione di Genova, Italy
lrosasco@mit.edu, mbelkin@cse.ohio-state.edu, devito@dima.unige.it

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 [27].

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. 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 for example [3] and references therein. Here we build 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 this problem. In [15] the authors study the relation between the spectrum of an integral operator defined by a symmetric, measurable kernel and its (modified) empirical counterpart in the framework of U-statistics. In particular concentration results for the ℓ_2 distance between the two (ordered) spectra are proven. The results are based on inequalities due to Lidskii and to Wielandt for finite dimensional matrices and the Marcinkiewicz law of large numbers. In [14] the analysis is extended to concentration of eigenfunctions and, using the triangle inequality, for spectral projections. The study of the concentration of empirical eigenvalues is continued in [19, 20], where it is shown that, for kernels of positive type, the problem of eigenvalue convergence reduces to the study the concentration of the empirical covariance matrix defined by n i.i.d ℓ_2 random vectors. Again for positive definite kernel, spectral projections are further studied in [28] where, by deriving a new perturbation result, 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 $k+1$ eigenvalues affects the estimate. The above studies are related to the problem of consistency of kernel PCA. Towards this end, the deviation of the sum of the all but the largest k eigenvalues of the empirical matrix to its mean is studied in [23, 24] using McDiarmid inequality and in [6] using a localized Rademacher complexities approach.
A second related series of works considered convergence of the graph Laplacian in various settings , see [16, 4, 11, 10, 25, 9]. 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 [11, 25, 16]. Results showing uniform convergence over some function class are presented in [10, 9]. Finally, convergence of eigenvalues and eigenfunctions for the case of the uniform distribution was shown in [5]. 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 [18] studies the convergence of the second eigenvalue which is relevant in spectral clustering problems. These results are extended in [27], where operators are defined on the space of continuous functions. The

*This work has been partially supported by the FIRB project RBIN04PARL and by the EU Integrated Project Health-e-Child IST-2004-027749.

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 by 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 with 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 [28] 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 with it. In this case we have to deal with non-self-adjoint operators and the functional analysis becomes more involved. Comparing to [27], 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 [26].

The plan of the paper follows. We start by introducing the necessary mathematical objects in Section 2. In Section 3, we study integral operators defined by reproducing kernel. In Section 4, we apply our methods to study the asymmetric graph Laplacian defined by a positive weight function.

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. [17]). In the following we let $A : \mathcal{H} \to \mathcal{H}$ be a (linear) bounded operator, where $\mathcal{H}$ is a 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 [17]). 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

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

Spectral Theory for Compact Operators. Recall that the spectrum of a matrix K can be defined as the set of eigenvalues $\lambda \in \mathbb{C}$, 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 $\lambda \in \mathbb{C}$ 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., [12]) 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 with 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 above equation 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 with an eigenvalue.

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 $f(x) \leq C_x \|f\| \qquad$ for some constant C_x, is called a *Reproducing Kernel Hilbert space*. It can be shown that there is a unique conjugate symmetric, positive definite kernel function $K : X \times X \to \mathbb{C}$, called *reproducing kernel*, associated with $\mathcal{H}$ 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 [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[1].

Remark 1 *To use the nice results of spectral theory, it is convenient to work with complex numbers, but one can see that in the following we always have to deal with (possibly non self-adjoint) operators where all the eigenvalues are reals and the eigenfunctions real valued.*

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 $\forall \varepsilon > 0$,

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

[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, [2]. The assumption about continuity is not strictly necessary, but it will simplify some technical part.

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 $1 - 2e^{-\tau}$ [21].

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 $\mathbf{K}_{ij} = \frac{1}{n} K(x_i, x_j)$. 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 with K and define the operators $T_K, T_{K,n} : \mathcal{H} \to \mathcal{H}$ given by,

$$T_K = \int_X \langle \cdot, K_x \rangle K_x d\rho(x), \tag{7}$$

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

Note that T_K 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/sampling operator", $R_n : \mathcal{H} \to \mathbb{C}^n$, $R_n(f) = (f(x_1), \ldots, f(x_n))$. It is not hard to check that the adjoint operator $R_n^*: \mathbb{C}^n \to \mathcal{H}$ 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

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

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

In the next subsection, we discuss a parallel with the Singular Value Decomposition for matrices and show that T_K 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 $T_{K,n}$ 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 $T_K - T_{K,n}$, which is done in the following theorem [8].

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

$$\|T_K - T_{K,n}\|_{HS} \le \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}}.$$

Proof: We introduce a sequence $(\xi_i)_{i=1}^n$ of random variables in the space of Hilbert-Schmidt operators $HS(\mathcal{H})$ by $\xi_i = \langle K_{x_i}, \cdot\rangle K_{x_i} - T_K$. From (7) follows that $E(\xi_i) = 0$. By a direct computation we have that $\|\langle \cdot, K_x\rangle K_x\|_{HS}^2 = \|K_x\|^4 \le \kappa^4$. Hence, using (7), $\|T_K\|_{HS} \le \kappa^2$ and $\|\xi_i\|_{HS} \le 2\kappa^2, \quad i = 1,\ldots,n$. From inequality (4) we have with probability $1 - 2e^{-\tau}$

$$\|\frac{1}{n}\sum_i \xi_i\|_{HS} = \|T_K - T_{K,n}\|_{HS} \le \frac{2\sqrt{2}\kappa^2\sqrt{\tau}}{\sqrt{n}},$$

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 an interpretation of the Nyström extension based on SVD and needed to properly compare to relate empirical operators and their mean.

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 $(u_j)_{j=1}^m \in \mathbb{C}^m$ and $(v_j)_{j=1}^p \in \mathbb{C}^p$ (where k is the rank of A) such that they form orthonormal basis of $\mathbb{C}^m$ and $\mathbb{C}^p$ respectively and such that

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

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 T_K as well as the matrix $\mathbf{K}$ and the operator $T_{K,n}$. 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 with $R_{\mathcal{H}}$ (and proposition 3 considering the SVD decomposition associated with R_n). The proof of the following proposition can be deduced from the results in [8] .

Proposition 2 *The following facts hold true.*

1. *The operators L_K and T_K are positive, self-adjoint and trace class. In particular both $\sigma(L_K)$ and $\sigma(T_K)$ are contained in $[0, \kappa^2]$.*
2. *The spectra of L_K and T_K are the same, possibly up to the zero, moreover if σ is a nonzero eigenvalue and u, v associated eigenfunctions of L_K and T_K (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:*

$$\begin{aligned} L_K g &= \sum_{j\ge1} \sigma_j \langle g, u_j\rangle_\rho u_j \\ T_K f &= \sum_{j\ge1} \sigma_j \langle f, v_j\rangle v_j \end{aligned}$$

the eigenfunctions $(u_j)_{j\ge1}$ of L_K form an orthonormal basis of $\ker {L_K}^\perp$ and the eigenfunctions $(v_j)_{j\ge1}$ of T_K for an orthonormal basis on $ker(T_K)^\perp$.

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 T_K 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 $T_{K,n}$.

Proposition 3 *The following facts hold:*

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

2. *The spectra of* $\mathbf{K}$ *and* $T_{K,n}$ *are the same up to the zero, that is,* $\sigma(\mathbf{K}) \setminus \{0\} = \sigma(T_{K,n}) \setminus \{0\}$. *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* $T_{K,n}$ *(normalized to norm* 1 *in* $\mathbb{C}^n$ *and* $\mathcal{H}$*) respectively, then*

$$\hat{u}^i = \frac{1}{\sqrt{\hat{\sigma}_j}} \hat{v}(x_i)$$

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

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

$$\mathbf{K}w = \sum_{j \geq 1} \hat{\sigma}_j \langle w, \hat{u}_j \rangle \hat{u}_j,$$

$$T_{K,n} f = \sum_{j \geq 1} \hat{\sigma}_j \langle f, \hat{v}_j \rangle_{\mathcal{H}} \hat{v}_j;$$

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

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

Note that when K is real-valued, the eigenfunctions are real valued (and the eigenvectors real). Since we are now dealing with self-adjoint operators, we can replace complex Hilbert spaces with real Hilbert spaces.

3.2 Bounds on eigenvalues and spectral projections.

To discuss the variation of the eigenvalues, we need to recall the notion of *extended enumeration* of discrete eigenvalues. We adapt the definition of [13], given for arbitrary self-adjoint operators, 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 [13] 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* $B - A$*, 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* T_K *and* $(\hat{\sigma}_j)_{j \geq 1}$ *the extended decreasing enumeration of discrete eigenvalues for* $T_{K,n}$. *With confidence* $1 - 2e^{-\tau}$,

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

and

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

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

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

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

Proof: Note that

$$\begin{cases} \operatorname{Tr}(T_{K,n}) & = \frac{1}{n} \sum_{i=1}^{n} K(x_i, x_i) \\ \operatorname{Tr}(T_K) & = \int_X K(x, x) d\rho(x). \end{cases}$$

Then we can define a sequence $(\xi_i)_{i=1}{}^n$ of real-valued random variables by $\xi_i = K(x_i, x_i) - \operatorname{Tr}(T_K)$. 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 $1 - 2e^{-\tau}$

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

■

To control the spectral projections associated with one or more eigenvalues we need the following perturbation result whose proof is given in [28] (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 with eigenvalues* $(\alpha_i)_{i \in I}$. *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.

Note that control of projections associated with simple eigenvalues implies the control one the corresponding eigenvectors, 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 T_K 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 with 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. However, for technical reasons we assume that

$$0 < c \leq W(x, s) \tag{9}$$

for all $x, s \in X$. 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 $\mathbf{L}_r : \mathbb{C}^n \to \mathbb{C}^n$ 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 consider $\mathbf{L}_r$ as complex matrix in order to have a nice spectral theory.

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 with 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. Let $m_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} W(\cdot, x_i)$ be the empirical counterpart of $m(\cdot)$, so that $m_n(x_i) = \mathbf{D}_{ii}$

Assumption 1 (A1) *Assume that $\mathcal{H}$ is a RKHS with bounded continuous kernel $K(x, t)$ such that, for all $x \in X$,*

$$W(x, \cdot)/m(\cdot) \in \mathcal{H},\ W(x, \cdot)/m_n(\cdot) \in \mathcal{H},\ W(x, \cdot) \in \mathcal{H}$$

with $\|W(x, \cdot)/m(\cdot)\|_{\mathcal{H}} \leq C$.

The above assumption is satisfied if $\mathcal{H}$ is a Sobolev space (with sufficiently high smoothness degree) and the weight function is also sufficiently smooth (see Assumption A2).

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

$$\begin{aligned} A_{\mathcal{H}} f &= \frac{1}{m(\cdot)} \int_X \langle f, K(x, \cdot) \rangle W(x, \cdot) d\rho(x) \\ T_r f &= f - A_{\mathcal{H}} f \end{aligned} \tag{10}$$

$$\begin{aligned} A_n f &= \frac{1}{m_n(\cdot)} \frac{1}{n} \sum_{i=1}^{n} \langle f, K(x_i, \cdot) \rangle W(x_i, \cdot) \\ T_{r,n} f &= f - A_n f \end{aligned} \tag{11}$$

It is possible to show that L_r, T_r 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. We define a restriction operator $R_n : \mathcal{H} \to \mathbb{C}^n$, $R_n(f)_i = f(x_i) =$

$\langle f, K(\cdot, x_i,)\rangle$ for all $i = 1, \ldots, n$, and an extension operator $E_n : \mathbb{C}^n \to \mathcal{H}$ 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 $T_{r,n}$ and A_n can still be clarified by means of R_n and 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 T_r. 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 T_r are bounded and have positive eigenvalues.*
2. *The eigenfunctions of $A_{\mathcal{H}}$ and T_r are the same and $\sigma(A_{\mathcal{H}}) = 1 - \sigma(T_r)$.*
3. *The spectra of L_r and T_r are the same. All the eigenvalues are real and the eigenfunctions can be chosen real-valued. If $\sigma \neq 1$ is an eigenvalue and u, v associated eigenfunctions of L_r and T_r respectively, then*
$$u(x) = v(x) \quad \text{for almost all } x \in X$$
$$v(x) = \frac{1}{1-\sigma}\int_X \frac{W(x,t)}{m(x)} u(t)\, d\rho(t)$$
4. *Finally the following decompositions hold*
$$L_r = \sum_{j\geq 1} \sigma_j P_j + P_0, \qquad (12)$$
$$T_r = I - \sum_{j\geq 1}(1-\sigma_j)Q_j + D, \qquad (13)$$
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 T_r associated with 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 $\ker D = \ker(I - T_r)$ *and $Q_j D = D Q_j = 0$ for all $j \geq 1$.*

The proof of the above result is long and quite technical, see [22]. 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 T_r 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, though $\sum_{j\geq 1} P_i = I - P_0$, it is not true that $\sum_{j\geq 1} Q_i$ converges to $I - Q_0$, where Q_0 is the spectral projection of T_r associated with the eigenvalue 1. This is the reason why we need to write the decomposition of T_r as in (13) instead of (12). An analogous result allows us to relate $\mathbf{L}_r$ to $T_{r,n}$ 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 $T_{r,n}$ have non-negative eigenvalues.*
2. *The eigenfunctions of A_n and $T_{r,n}$ are the same and $\sigma(A_n) = 1 - \sigma(T_{r,n})$.*
3. *The spectra of $\mathbf{L}_r$ and $T_{r,n}$ 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 $T_{r,n}$, 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*
$$\mathbf{L}_r = \sum_{j=1}^{m}\hat{\sigma}_j\hat{P}_j + \hat{P}_0,$$
$$T_{r,n} = \sum_{j=1}^{m}\hat{\sigma}_j\hat{Q}_j + \hat{Q}_0 + \hat{D},$$
where $\{\hat{\sigma}_1, \ldots \hat{\sigma}_m\} = \sigma(L_r)\setminus\{1\}$, the projections Q_j, P_j are the spectral projections of $\mathbf{L}_r$ and $T_{r,n}$ associated with the eigenvalue σ_j, $\hat{P}_0$ and $\hat{Q}_0$ are the spectral projections of $\mathbf{L}_r$ and $T_{r,n}$ associated with the eigenvalue 1, and $\hat{D}$ is a quasi-nilpotent operator such that $\ker\hat{D} = \ker(I - T_{r,n})$ *and $\hat{Q}_j\hat{D} = \hat{D}\hat{Q}_j = 0$ for all $j = 1, \ldots, m$.*

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

To estimate the deviation of T_r to $T_{r,n}$ we assume the space $\mathcal{H}$ to be a Sobolev space. 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 $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ 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
$$\|f\|_s := \|f\|_X + \sum_{|\alpha|=s}\|(D^\alpha f)(x)\|_X < \infty,$$
where $\|f\|_X^2 = \int_X |f(x)|^2\, dx$ with dx being the Lebesgue measure (the above definition of $\mathcal{H}^s$ 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
$$\|f\|_\infty \leq C_1\|f\|_s. \qquad (14)$$

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

Then $\mathcal{H}^s$ is a RKHS with reproducing kernel $K^s(x, y)$, so that $f(x) = \langle f, K_x^s \rangle_s$ where $K_x^s := K^s(x, \cdot)$. Moreover we also have

$$\sup_{x \in X} \|K_x^s\|_s = C_1 < \infty. \tag{15}$$

In the following we will need the following result from [7].

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*

$$\|M_g\| \leq a \|g\|_{s'} < \infty, \tag{16}$$

for some positive constant a and $s' > s + d/2$.

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

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

The above assumption quantifies the regularity on the weight function needed for Assumption A1 to hold true.
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 $T_W, T_{W,n} : \mathcal{H}_s \to \mathcal{H}_s$ defined by*

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

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

$$\|T_W - T_{W,n}\|_{HS} \leq \frac{2\sqrt{2} C_1 C_2 \sqrt{2\tau}}{\sqrt{n}}.$$

Proof: Note that $\|\langle \cdot, K_{x_i}^s \rangle_s W_{x_i}\|_{HS} = \|K_{x_i}^s\| \|W_{x_i}\|_s \leq C_1 C_2$ so that $T_{W,n}, T_W$ are Hilbert Schmidt. The random variables $(\xi_i)_{i=1}^n$ defined by $\xi_i = \langle \cdot, K_{x_i}^s \rangle_s W_{x_i} - T_W$ are zero mean and bounded by $2C_1 C_2$. Applying (4) we have with confidence $1 - 2e^{-\tau}$

$$\|T_W - T_{W,n}\|_{HS} \leq \frac{2\sqrt{2} C_1 C_2 \sqrt{\tau}}{\sqrt{n}}. \tag{17}$$

■

We then consider multiplication operators defined by the degree functions.

Proposition 14 *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 $1 - 2e^{-\tau}$*

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

where is a positive constant.

Proof: It follows from (16) 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 the 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 $2C_2 a$. Applying (4) we have with high probability

$$\|m - m_n\|_{s'} \leq \frac{2a C_2 \sqrt{2\tau}}{\sqrt{n}}.$$

It follows from (16) that

$$\|M - M_n\| \leq \frac{2a C_2 \sqrt{2\tau}}{\sqrt{n}}. \tag{18}$$

■

Finally, we can combine the above two propositions to get the proof of Theorem 12.
Proof:[Proof of Theorem 12] From (16) and assumption A2, the multiplication 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 $A_\mathcal{H} = M^{-1} T_W$ and $A_n = M_n^{-1} T_{W,n}$ so that we can consider the following decomposition

$$\begin{aligned} A_\mathcal{H} - A_n &= M_n^{-1} T_{W,n} - M^{-1} T_W \qquad (19) \\ &= (M_n^{-1} - M^{-1}) T_W + M_n^{-1} (T_{W,n} - T_W) \\ &= M_n^{-1} (M - M_n) M^{-1} T_W + \\ &\quad + M_n^{-1} (T_{K,n} - T_W). \end{aligned}$$

Recalling (1), we consider the Hilbert-Schmidt norm of the above expression. From (16), (9) and Assumption A2 one has $\|M^{-1}\| \leq a \|m^{-1}\|_{s'} \leq C_3$, and $\|M_n^{-1}\| \leq a \|m_n^{-1}\|_{s'} \leq C_4$, where the constants C_3, C_4 depend only on a, C_2, c. Using the Propositions 13, 14, and (15) we see that there is a constant C, such that

$$\|M_n^{-1} T_{W,n} - M^{-1} T_W\|_{HS} \leq C \frac{\sqrt{\tau}}{\sqrt{n}}.$$

■

In the next section we discuss the implications of the above result 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. The following theorem is an adaptation of results in [1].

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 with 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

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

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 with $\widehat{\Lambda}$, then*
$$\|\widehat{P} - P\| \le \frac{\ell(\Gamma)}{2\pi\delta} \frac{\|B - A\|}{\delta - \|B - A\|};$$
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

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

See [22] for the proof. 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*
$$|\hat{\sigma}_i - \sigma| \le \varepsilon \qquad \text{for all } i = 1, \ldots, m.$$
2. *for any normalized eigenvector $\hat{u} \in \mathbb{C}^n$ 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*
$$\|E_n(\hat{u}) - u\|_s \le R \frac{\sqrt{\tau}}{\sqrt{n}},$$
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

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

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

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

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 $\hat{u} \in \mathbb{C}^n$ 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 with $\hat{\Lambda}$. Let Q be the spectral projection of $A_{\mathcal{H}}$ associated with $1 - \sigma$ 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

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

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 $\|A_n - A_{\mathcal{H}}\| \le \frac{\delta^2}{\delta + \ell(\Gamma)/2\pi}$. ■

We note that by inspecting the above proof, if $A_{\mathcal{H}}$ is self-adjoint, then $n_0 \ge \frac{C^2\tau}{\varepsilon^2}$ 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 17 *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 $n \ge n_0$,*

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

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

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

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

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

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 EU Integrated Project Health-e-Child IST-2004-027749. Mikhail Belkin is partially supported by the NSF Early Career Award 0643916.

References

[1] 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.

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

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

[4] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. In *Proceedings of the 18th Conference on Learning Theory (COLT)*, pages 486–500, 2005.

[5] 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.

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

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

[8] 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.

[9] E. Giné and V. Koltchinskii. Empirical graph laplacian approximation of laplace-beltrami operators: Large sample results. *High Dimensional Probability*, 51:238?259, 2006.

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

[11] 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.

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

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

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

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

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

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

[18] 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.

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

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

[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.

[22] L. Rosasco, M. Belkin, and E. De Vito. A note on perturbation results for learning empirical operators. Technical Report CBCL paper #274/ CSAIL Technical Report #TR-2008-052, MIT, 2008.

[23] 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.

[24] 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.

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

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

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

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