Geometry on Probability Spaces

Steve Smale
Toyota Technological Institute at Chicago
1427 East 60th Street, Chicago, IL 60637, USA
E-mail: smale@math.berkeley.edu
Ding-Xuan Zhou
Department of Mathematics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, CHINA
E-mail: mazhou@cityu.edu.hk

December 14, 2007

Preliminary Report

1 Introduction

Partial differential equations and the Laplacian operator on domains in Euclidean spaces have played a central role in understanding natural phenomena. However this avenue has been limited in many areas where calculus is obstructed as in singular spaces, and function spaces of functions on a space X where X itself is a function space. Examples of the last occur in vision and quantum field theory. In vision it would be useful to do analysis on the space of images and an image is a function on a patch. Moreover in analysis and geometry, the Lebesgue measure and its counterpart on manifolds are central. These measures are unavailable in the vision example and even in learning theory in general.

There is one situation where in the last several decades, the problem has been studied with some success. That is when the underlying space is finite (or even discrete). The introduction of the graph Laplacian has been a major development in algorithm research

and is certainly useful for unsupervised learning theory.

The point of view taken here is to benefit from both the classical research and the newer graph theoretic ideas to develop geometry on probability spaces. This starts with a space X equipped with a kernel (like a Mercer kernel) which gives a topology and geometry; X is to be equipped as well with a probability measure. The main focus is on a construction of a (normalized) Laplacian, an associated heat equation, diffusion distance, etc. In this setting the point estimates of calculus are replaced by integral quantities. One thinks of secants rather than tangents. Our main result, Theorem 1 below, bounds the error of an empirical approximation to this Laplacian on X.

2 Kernels and Metric Spaces

Let X be a set and $K : X \times X \to \mathbb{R}$ be a reproducing kernel, that is, K is symmetric and for any finite set of points $\{x_1, \cdots, x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semidefinite. The Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined to be the completion of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$. See [1, 10, 11].

We assume that the feature map $\mathcal{F} : X \to \mathcal{H}_K$ mapping $x \in X$ to $K_x \in \mathcal{H}_K$ is injective.

Definition 1. *Define a metric $d = d_K$ induced by K on X as $d(x, y) = \|K_x - K_y\|_K$. We assume that (X, d) is complete and separable (if it is not complete, we may complete it).*

The feature map is an isometry. Observe that

$$K(x, y) = \langle K_x, K_y \rangle_K = \frac{1}{2} \left\{ \langle K_x, K_x \rangle_K + \langle K_y, K_y \rangle_K - (d(x, y))^2 \right\}.$$

Then K is continuous on $X \times X$, and hence is a Mercer kernel.

Let ρ_X be a Borel probability measure on the metric space X.

Example 1. *Let $X = \mathbb{R}^n$ and K be the reproducing kernel $K(x, y) = \langle x, y \rangle_{\mathbb{R}^n}$. Then $d_K(x, y) = \|x - y\|_{\mathbb{R}^n}$ and $\mathcal{H}_K$ is the space of linear functions.*

Example 2. *Let X be a closed subset of $\mathbb{R}^n$ and K be the kernel given in the above example restricted to $X \times X$. Then X is complete and separable, and K is a Mercer kernel. Moreover, one may take ρ_X to be any Borel probability measure on X.*

Remark 1. *In Examples 1 and 2, one can replace $\mathbb{R}^n$ by a separable Hilbert space.*

Example 3. *In Example 2, let $\mathbf{x} = \{x_i\}_{i=1}^m$ be a finite subset of X drawn from ρ_X. Then the restriction $K|_{\mathbf{x}}$ is a Mercer kernel on $\mathbf{x}$ and it is natural to take $\rho_{\mathbf{x}}$ to be the uniform measure on $\mathbf{x}$.*

3 Approximation of Integral Operators

The integral operator $L_K : L^2_{\rho_X} \to \mathcal{H}_K$ associated with the pair (K, ρ_X) is defined by

$$L_K(f)(x) := \int_X K(x,y) f(y) d\rho_X(y), \qquad x \in X. \tag{3.1}$$

It may be considered as a self-adjoint operator on $L^2_{\rho_X}$ or $\mathcal{H}_K$. See [10, 11]. We shall use the same notion L_K for these operators defined on different domains.

Assume that $\mathbf{x} := \{x_i\}_{i=1}^m$ is a sample independently drawn according to ρ_X.

Recall the **sampling operator** $S_{\mathbf{x}} : \mathcal{H}_K \to \ell^2(\mathbf{x})$ associated with $\mathbf{x}$ defined [17, 18] by

$$S_{\mathbf{x}}(f) = \big(f(x_i)\big)_{i=1}^m.$$

By the reproducing property of $\mathcal{H}_K$ taking the form $\langle K_x, f\rangle_K = f(x)$ with $x \in X, f \in \mathcal{H}_K$, the adjoint of the sampling operator, $S^T_{\mathbf{x}} : \ell^2(\mathbf{x}) \to \mathcal{H}_K$, is given by

$$S^T_{\mathbf{x}} c = \sum_{i=1}^m c_i K_{x_i}, \qquad c \in \ell^2(\mathbf{x}).$$

As the sample size m tends to infinity, the operator $\frac{1}{m} S^T_{\mathbf{x}} S_{\mathbf{x}}$ converges to the integral operator $L_{\widetilde{K}}$. We prove the following convergence rate.

Proposition 1. *Assume*

$$\kappa := \sqrt{\sup_{x \in X} K(x,x)} < \infty. \tag{3.2}$$

Let $\mathbf{x}$ be a sample independently drawn from (X, ρ_X). With confidence $1-\delta$, we have

$$\left\| \frac{1}{m} S^T_{\mathbf{x}} S_{\mathbf{x}} - L_K \right\|_{\mathcal{H}_K \to \mathcal{H}_K} \le \frac{4\kappa^2 \log\big(2/\delta\big)}{\sqrt{m}}. \tag{3.3}$$

Proof. The proof follows [18]. We need a probability inequality [15, 18] for a random variable ξ on (X, ρ_X) with values in a Hilbert space $(H, \|\cdot\|)$. It says that if $\|\xi\| \le \widetilde{M} < \infty$ almost surely, then with confidence $1-\delta$,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \big[\xi(x_i) - E(\xi)\big] \right\| \le \frac{4\widetilde{M} \log\big(2/\delta\big)}{\sqrt{m}}. \tag{3.4}$$

We apply this probability inequality to the separable Hilbert space H of Hilbert-Schmidt operators on $\mathcal{H}_K$, denoted as $HS(\mathcal{H}_K)$. The inner product for $A_1, A_2 \in HS(\mathcal{H}_K)$ is defined as $\langle A_1, A_2 \rangle_{HS} = \sum_{j\ge1} \langle A_1 e_j, A_2 e_j \rangle_K$ where $\{e_j\}_{j\ge1}$ is an orthonormal basis of $\mathcal{H}_K$. The space $HS(\mathcal{H}_K)$ is the subspace of the space of bounded linear operators on $\mathcal{H}_K$ with the norm $\|A\|_{HS} < \infty$.

The random variable ξ on (X, ρ_X) in (3.4) is given by

$$\xi(x) = K_x \langle \cdot, K_x \rangle_K, \qquad x \in X.$$

The values of ξ are rank-one operators in $HS(\mathcal{H}_K)$. For each $x \in X$, if we choose an orthonormal basis $\{e_j\}_{j\ge1}$ of $\mathcal{H}_K$ with $e_1 = K_x/\sqrt{K(x,x)}$, we see that $\|\xi(x)\|_{HS}^2 = \sum_{j\ge1} \|\xi(x)e_j\|_K^2 = (K(x,x))^2 \le \kappa^4$. Observe that $\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^m K_{x_i} \langle \cdot, K_{x_i} \rangle_K = \frac{1}{m} \sum_{i=1}^m \xi(x_i)$ and $E(\xi) = L_K$. Then we observe that the desired bound (3.3) follows from (3.4) with $\widetilde{M} = \kappa^2$ and the norm relation

$$\|A\|_{\mathcal{H}_K \to \mathcal{H}_K} \le \|A\|_{HS}. \tag{3.5}$$

This proves the proposition. □

Some analysis related to Proposition 1 can be found in [19, 14, 5, 12].

Definition 2. *Denote $p = \int_X K_x d\rho_X \in \mathcal{H}_K$. We assume that p is positive on X. The normalized Mercer kernel on X associated with the pair (K, ρ_X) is defined by*

$$\widetilde{K}(x,y) = \frac{K(x,y)}{\sqrt{p(x)}\sqrt{p(y)}}, \qquad x, y \in X. \tag{3.6}$$

This construction is in [19], except that we have replaced their positive weighting function by a Mercer kernel.

In the following,

$\widetilde{\ }$ means expressions wrt the normalized kernel $\widetilde{K}$, in particular $\widetilde{K}_{\mathbf{x}} := \Big(\widetilde{K}(x_i, x_j)\Big)_{i,j=1}^m$.

Consider the RKHS $\mathcal{H}_{\widetilde{K}}$, and the integral operator $L_{\widetilde{K}} : \mathcal{H}_{\widetilde{K}} \to \mathcal{H}_{\widetilde{K}}$ associated with the normalized Mercer kernel $\widetilde{K}$.

Denote the sampling operator $\mathcal{H}_{\widetilde{K}} \to \ell^2(\mathbf{x})$ associated with the pair $(\widetilde{K}, \mathbf{x})$ as $\widetilde{S}_{\mathbf{x}}$. If we denote $\mathcal{H}_{\widetilde{K},\mathbf{x}} = \text{span}\{\widetilde{K}_{x_i}\}_{i=1}^m$ in $\mathcal{H}_{\widetilde{K}}$ and its orthogonal complement as $\mathcal{H}^{\perp}_{\widetilde{K},\mathbf{x}}$, then we see that $\widetilde{S}_{\mathbf{x}}(f) = 0$ for any $f \in \mathcal{H}^{\perp}_{\widetilde{K},\mathbf{x}}$. Hence $\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}}|_{\mathcal{H}^{\perp}_{\widetilde{K},\mathbf{x}}} = 0$. Moreover, $\widetilde{K}_{\mathbf{x}}$ is the matrix representation of the operator $\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}}|_{\mathcal{H}_{\widetilde{K},\mathbf{x}}}$.

Proposition 1 yields the following bound with $\widetilde{K}$ replacing K in Proposition 1.

Theorem 1. *Denote $p_0 = \min_{x\in X} p(x)$. For any $0 < \delta < 1$, with confidence $1 - \delta$,*

$$\left\| L_{\widetilde{K}} - \frac{1}{m}\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}} \right\|_{\mathcal{H}_{\widetilde{K}}\to\mathcal{H}_{\widetilde{K}}} \leq \frac{4\kappa^2 \log\big(2/\delta\big)}{\sqrt{m}p_0}. \tag{3.7}$$

Note that $\frac{1}{m}\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}}$ is given from data $\mathbf{x}$ by a linear algorithm [16]. So Theorem 1 is an error estimate for that algorithm.

The analysis in [19] deeply involves the constant $\min_{x,y\in X} K(x, y)$. In the case of the Gaussian kernel $K_\sigma(x, y) = \exp\{-\frac{|x-y|^2}{2\sigma^2}\}$, this constant decays exponentially fast as the variance σ of the gaussian becomes small, while in Theorem 1 the constant p_0 decays polynomially fast, at least in the manifold setting [20].

4 Tame Heat Equation

Define a tame version of Laplacian as $\Delta_{\widetilde{K}} = I - L_{\widetilde{K}}$. The tame heat equation takes the form

$$\frac{\partial u}{\partial t} = -\Delta_{\widetilde{K}} u. \tag{4.1}$$

One can see that the solution to (4.1) with the initial condition $u(0) = u_0$ is given by

$$u(t) = \exp\left\{-t\Delta_{\widetilde{K}}\right\} u_0. \tag{4.2}$$

5 Diffusion Distances

Let $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be the eigenvalues of $L_{\widetilde{K}} : L^2_{\rho_X} \to L^2_{\rho_X}$ and $\{\varphi^{(i)}\}_{i\geq 1}$ be corresponding normalized eigenfunctions which form an orthonormal basis of $L^2_{\rho_X}$. The

Mercer Theorem asserts that

$$\widetilde{K}(x,y) = \sum_{i=1}^{\infty} \lambda_i \varphi^{(i)}(x)\varphi^{(i)}(y)$$

where the convergence holds in $L^2_{\rho_X}$ and uniformly.

Definition 3. *For $t > 0$ we define a reproducing kernel on X as*

$$\widetilde{K}^t(x,y) = \sum_{i=1}^{\infty} \lambda_i^t \varphi^{(i)}(x)\varphi^{(i)}(y). \tag{5.1}$$

Then the tame diffusion distance D^t on X is defined as $d_{\widetilde{K}^t}$ by

$$D^t(x,y) = \|\widetilde{K}^t_x - \widetilde{K}^t_y\|_{\widetilde{K}^t} = \left\{ \sum_{i=1}^{\infty} \lambda_i^t \left[\varphi^{(i)}(x) - \varphi^{(i)}(y)\right]^2 \right\}^{1/2}. \tag{5.2}$$

The kernel $\widetilde{K}^t$ may be interpreted as a tame version of the heat kernel on a manifold. The functions $(\widetilde{K}^t)_x$ are in $\mathcal{H}_{\widetilde{K}}$ for $t \geq 1$, but only in $L^2_{\rho_X}$ for $t \geq 0$.

The quantity corresponding to D^t for the classical heat kernel and other varieties has been developed by Coifman et. al. [8, 9] where it is used in data analysis and more.

Note that $\widetilde{K}^t$ is continuous for $t \geq 1$, in contrast to the classical kernel (Greens function) which is of course singular on the diagonal of $X \times X$.

The integral operator $L_{\widetilde{K}^t}$ associated with the reproducing kernel $\widetilde{K}^t$ is the tth power of the positive operator $L_{\widetilde{K}}$.

Note that for $t = 1$, the distance D^1 is the one induced by the kernel $\widetilde{K}$. When t becomes large, the effect of the eigenfunctions $\varphi^{(i)}$ with small eigenvalues λ_i is little. In particular, when λ_1 has multiplicity 1, since $\varphi^{(1)} = \sqrt{p}$ we have

$$\lim_{t\to\infty} D^t(x,y) = \left|\sqrt{p(x)} - \sqrt{p(y)}\right|.$$

In the same way, since the tame Laplacian $\Delta_{\widetilde{K}}$ has eigenpairs $(1-\lambda_i, \varphi^{(i)})$, we know that the solution (4.2) to the tame heat equation (4.1) with the initial condition $u(0) = u_0 \in L^2_{\rho_X}$ satisfies

$$\lim_{t\to\infty} u(t) = \langle u_0, \sqrt{p}\rangle_{L^2_{\rho_X}} \sqrt{p}.$$

6 Graph Laplacian

Consider the reproducing kernel K and the finite subset $\mathbf{x} = \{x_i\}_{i=1}^m$ of X.

Definition 4. *Let $K_{\mathbf{x}}$ be the symmetric matrix $K_{\mathbf{x}} = (K(x_i, x_j))_{i,j=1}^m$ and $D_{\mathbf{x}}$ be the diagonal matrix with $(D_{\mathbf{x}})_{i,i} = \sum_{j=1}^m (K_{\mathbf{x}})_{i,j} = \sum_{j=1}^m K(x_i, x_j)$. Define a discrete Laplacian matrix $L = L_{\mathbf{x}} = D_{\mathbf{x}} - K_{\mathbf{x}}$. The normalized discrete Laplacian is defined by*

$$\widehat{\Delta}_{K,\mathbf{x}} = I - D_{\mathbf{x}}^{-\frac{1}{2}} K_{\mathbf{x}} D_{\mathbf{x}}^{-\frac{1}{2}} = I - \frac{1}{m}\widehat{K}_{\mathbf{x}}, \tag{6.1}$$

where $\widehat{K}_{\mathbf{x}}$ is the special case of the matrix $\widetilde{K}_{\mathbf{x}}$ when ρ_X is the uniform distribution of the finite set $\mathbf{x}$, that is,

$$\widehat{K}_{\mathbf{x}} := \left(\frac{K(x_i, x_j)}{\sqrt{\frac{1}{m}\sum_{\ell=1}^m K(x_i, x_\ell)}\sqrt{\frac{1}{m}\sum_{\ell=1}^m K(x_j, x_\ell)}} \right)_{i,j=1}^m. \tag{6.2}$$

The above construction is the same as that for graph Laplacians [7]. See [2] for a discussion in learning theory. But the setting here is different: the entries $K(x_i, x_j)$ are induced by a reproducing kernel, they might take negative values satisfying a positive definiteness condition, and are different from weights of a graph or entries of an adjacency matrix.

In the following, $\widehat{\ }$ means expressions with respect to an implicit sample $\mathbf{x}$.

7 Approximation of Eigenfunctions

In this section we study the approximation of eigenfunctions from the bound for approximation of linear operators. We need the following result which follows from the general discussion on perturbation of eigenvalues and eigenvectors, see e.g. [4]. For completeness, we give a detailed proof for the approximation of eigenvectors.

Proposition 2. *Let A and $\widehat{A}$ be two compact positive self-adjoint operators on a Hilbert space H, with nondecreasing eigenvalues $\{\lambda_j\}$ and $\{\widehat{\lambda}_j\}$ with multiplicity. Then there holds*

$$\max_{j\geq 1} |\lambda_j - \widehat{\lambda}_j| \leq \|A - \widehat{A}\|. \tag{7.1}$$

Let w_k be a normalized eigenvector of A associated with eigenvalue λ_k. If $r > 0$ satisfies

$$\lambda_{k-1} - \lambda_k \geq r, \quad \lambda_k - \lambda_{k+1} \geq r, \quad \|A - \widehat{A}\| \leq \frac{r}{2}, \tag{7.2}$$

then

$$\|w_k - \widehat{w}_k\| \leq \frac{4}{r}\|A - \widehat{A}\|,$$

where $\widehat{w}_k$ is a normalized eigenvector of $\widehat{A}$ associated with eigenvalue $\widehat{\lambda}_k$.

Proof. The eigenvalue estimate (7.1) follows from Weyl's Perturbation Theorem (see e.g. [4]).

Let $\{w_j\}_{j\geq 1}$ be an orthonormal basis of H consisting of eigenvectors of A associated with eigenvalues $\{\lambda_j\}$. Then $\widehat{w}_k = \sum_{j\geq 1} \alpha_j w_j$ where $\alpha_j = \langle \widehat{w}_k, w_j\rangle$.

Consider $\widehat{A}\widehat{w}_k - A\widehat{w}_k$. It can be expressed as

$$\widehat{A}\widehat{w}_k - A\widehat{w}_k = \widehat{\lambda}_k\widehat{w}_k - \sum_{j\geq 1}\alpha_j A w_j = \sum_{j\geq 1}\alpha_j\left(\widehat{\lambda}_k - \lambda_j\right) w_j.$$

When $j \neq k$, we see from (7.2) and (7.1) that

$$|\widehat{\lambda}_k - \lambda_j| \geq |\lambda_k - \lambda_j| - |\widehat{\lambda}_k - \lambda_k| \geq \min\{\lambda_{k-1} - \lambda_k, \lambda_k - \lambda_{k+1}\} - \|A - \widehat{A}\| \geq \frac{r}{2}.$$

Hence

$$\|\widehat{A}\widehat{w}_k - A\widehat{w}_k\|^2 = \sum_{j\geq 1}\alpha_j^2\left(\widehat{\lambda}_k - \lambda_j\right)^2 \geq \sum_{j\neq k}\alpha_j^2\left(\widehat{\lambda}_k - \lambda_j\right)^2 \geq \frac{r^2}{4}\sum_{j\neq k}\alpha_j^2.$$

But $\|\widehat{A}\widehat{w}_k - A\widehat{w}_k\| \leq \|\widehat{A} - A\|$. It follows that

$$\left\{\sum_{j\neq k}\alpha_j^2\right\}^{1/2} \leq \frac{2}{r}\|\widehat{A} - A\|.$$

From $\|\widehat{w}_k\|^2 = \sum_{j\geq 1}\alpha_j^2 = 1$, we also have $\alpha_k = \left\{1 - \sum_{j\neq k}\alpha_j^2\right\}^{1/2} \geq 1 - \left\{\sum_{j\neq k}\alpha_j^2\right\}^{1/2}$. Therefore,

$$\begin{aligned}\|w_k - \widehat{w}_k\| &= \|(1-\alpha_k)w_k + \alpha_k w_k - \widehat{w}_k\| = \|(1-\alpha_k)w_k - \sum_{j\neq k}\alpha_j w_j\| \\ &\leq |1-\alpha_k| + \|\sum_{j\neq k}\alpha_j w_j\| \leq 2\left\{\sum_{j\neq k}\alpha_j^2\right\}^{1/2} \leq \frac{4}{r}\|\widehat{A} - A\|.\end{aligned}$$

This proves the desired bound. □

An immediate easy corollary of Proposition 2 and Theorem 1 is about the approximation of eigenfunctions of $L_{\widetilde{K}}$ by those of $\frac{1}{m}\widetilde{S}_{\mathbf{x}}^T\widetilde{S}_{\mathbf{x}}$.

Recall the eigenpairs $\{\lambda_i, \varphi^{(i)}\}_{i\ge 1}$ of the compact self-adjoint positive operator $L_{\widetilde{K}}$ on $L^2_{\rho_X}$. Observe that $\|\varphi^{(k)}\|_{L^2_{\rho_X}} = 1$, but $\|\varphi^{(k)}\|^2_{\widetilde{K}} = \langle \frac{1}{\lambda_k} L_{\widetilde{K}}\varphi^{(k)}, \varphi^{(k)}\rangle_{\widetilde{K}} = \frac{1}{\lambda_k}\|\varphi^{(k)}\|^2_{L^2_{\rho_X}} = \frac{1}{\lambda_k}$. So $\frac{1}{\sqrt{\lambda_k}}\varphi^{(k)}$ is a normalized eigenfunction of $L_{\widetilde{K}}$ on $\mathcal{H}_{\widetilde{K}}$ associated with eigenvalue λ_k.

Also, $\frac{1}{m}\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}}|_{\mathcal{H}^{\perp}_{\widetilde{K},\mathbf{x}}} = 0$ and $\frac{1}{m}\widetilde{K}_{\mathbf{x}}$ is the matrix representation of $\frac{1}{m}\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}}|_{\mathcal{H}_{\widetilde{K},\mathbf{x}}}$.

Corollary 1. *Let $k \in \{1,\dots,m\}$ such that $r := \min\{\lambda_{k-1}-\lambda_k, \lambda_k-\lambda_{k+1}\} > 0$. If $0<\delta<1$ and $m \in \mathbb{N}$ satisfy*

$$\frac{4\kappa^2 \log\big(4/\delta\big)}{\sqrt{m}p_0} \le \frac{r}{2},$$

then with confidence $1-\delta$, we have

$$\left\| \frac{\varphi^{(k)}}{\sqrt{\lambda_k}} - \Phi^{(k)} \right\|_{\widetilde{K}} \le \frac{16\kappa^2 \log\big(4/\delta\big)}{r\sqrt{m}p_0},$$

where $\Phi^{(k)}$ is a normalized eigenfunction of $\frac{1}{m}\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}}$ associated with its kth largest eigenvalue $\lambda_{k,\mathbf{x}}$. Moreover, $\lambda_{k,\mathbf{x}}$ is the kth largest eigenvalue of the matrix $\frac{1}{m}\widetilde{K}_{\mathbf{x}}$, and with an associated normalized eigenvector v, we have $\Phi^{(k)} = \frac{1}{\sqrt{\lambda_{k,\mathbf{x}}}}\frac{1}{\sqrt{m}}\widetilde{S}^T_{\mathbf{x}} v$.

8 Algorithmic Issue for Approximating Eigenfunctions

Corollary 1 provides quantitative understanding of the approximation of eigenfunctions of $L_{\widetilde{K}}$ by those of the operator $\frac{1}{m}\widetilde{S}^T_{\mathbf{x}}\widetilde{S}_{\mathbf{x}}$. The matrix representation $\frac{1}{m}\widetilde{K}_{\mathbf{x}}$ is given by the kernel $\widetilde{K}$ which involves K and the function p defined by ρ_X. Since ρ_X is unknown algorithmically, we want to study the approximation of eigenfunctions by methods involving only K and the sample $\mathbf{x}$. This is done here.

We shall apply Proposition 2 to the operator $A = L_{\widetilde{K}}$ and $\widehat{A}$ defined by means of the matrix $\widehat{K}_{\mathbf{x}}$ given by $(K,\mathbf{x})$ in (6.2), and derive estimates for the approximation of eigenfunctions.

Let $\widehat{\lambda}_{1,\mathbf{x}} \ge \widehat{\lambda}_{2,\mathbf{x}} \ge \ldots \ge \widehat{\lambda}_{m,\mathbf{x}} \ge 0$ be the eigenvalues of the matrix $\frac{1}{m}\widehat{K}_{\mathbf{x}}$.

Theorem 2. *Let $k \in \{1,\dots,m\}$ such that $r := \min\{\lambda_{k-1}-\lambda_k, \lambda_k-\lambda_{k+1}\} > 0$. If $0<\delta<1$ and $m \in \mathbb{N}$ satisfy*

$$\frac{4\kappa^2 \log\big(4/\delta\big)}{\sqrt{m}p_0} \le \frac{r}{8}, \tag{8.1}$$

then with confidence $1-\delta$, we have

$$\left\| \varphi^{(k)} - \frac{1}{\sqrt{m}}\widetilde{S}^T_{\mathbf{x}} v \right\|_{\widetilde{K}} \le \frac{72\sqrt{\lambda_k}\kappa^2 \log\big(4/\delta\big)}{r\sqrt{m}p_0}, \tag{8.2}$$

where $v \in \ell^2(\mathbf{x})$ is a normalized eigenvector of the matrix $\frac{1}{m}\widehat{K}_{\mathbf{x}}$ with eigenvalue $\widehat{\lambda}_{k,\mathbf{x}}$.

Remark 2. *Since $\Delta_{\widetilde{K}} = I - L_{\widetilde{K}}$ and $\widehat{\Delta}_{K,\mathbf{x}} = I - D_{\mathbf{x}}^{-\frac{1}{2}} K_{\mathbf{x}} D_{\mathbf{x}}^{-\frac{1}{2}} = I - \frac{1}{m}\widehat{K}_{\mathbf{x}}$, Theorem 2 provides error bounds for the approximation of eigenfunctions of the tame Laplacian $\Delta_{\widetilde{K}}$ by those of the normalized discrete Laplacian $\widehat{\Delta}_{K,\mathbf{x}}$. Note that $\frac{1}{\sqrt{m}}\widetilde{S}_{\mathbf{x}}^T v = \frac{1}{\sqrt{m}}\sum_{i=1}^m v_i \widetilde{K}_{x_i}$.*

We need some preliminary estimates for the approximation of linear operators.

Applying the probability inequality (3.4) for the random variable ξ on (X, ρ_X) with values in the Hilbert space $\mathcal{H}_K$ given by $\xi(x) = K_x$, we have

Lemma 1. *Denote $p_{\mathbf{x}} = \frac{1}{m}\sum_{j=1}^m K_{x_j} \in \mathcal{H}_K$. With confidence $1 - \delta/2$, we have*

$$\|p_{\mathbf{x}} - p\|_K \le \frac{4\kappa \log\big(4/\delta\big)}{\sqrt{m}}. \tag{8.3}$$

The error bound (8.3) implies

$$\max_{i=1,\dots,m} |p_{\mathbf{x}}(x_i) - p(x_i)| \le \frac{4\kappa^2 \log\big(4/\delta\big)}{\sqrt{m}}. \tag{8.4}$$

This yields the following error bounds between the matrices $\widetilde{K}_{\mathbf{x}}$ and $\widehat{K}_{\mathbf{x}}$.

Lemma 2. *Let $\mathbf{x} \in X^m$. When (8.3) holds, we have*

$$\left\| \frac{1}{m}\widetilde{K}_{\mathbf{x}} - \frac{1}{m}\widehat{K}_{\mathbf{x}} \right\| \le \left(2 + \frac{4\kappa^2 \log\big(4/\delta\big)}{\sqrt{m}p_0} \right) \frac{4\kappa^2 \log\big(4/\delta\big)}{\sqrt{m}p_0}. \tag{8.5}$$

Proof. Let $\Sigma = \operatorname{diag}\{\Sigma_i\}_{i=1}^m$ with $\Sigma_i := \frac{\sqrt{p_{\mathbf{x}}(x_i)}}{\sqrt{p(x_i)}}$. Then $\frac{1}{m}\widetilde{K}_{\mathbf{x}} = \Sigma \frac{1}{m}\widehat{K}_{\mathbf{x}}\Sigma$. Decompose

$$\frac{1}{m}\widetilde{K}_{\mathbf{x}} - \frac{1}{m}\widehat{K}_{\mathbf{x}} = \Sigma \frac{1}{m}\widehat{K}_{\mathbf{x}}(\Sigma - I) + (\Sigma - I)\frac{1}{m}\widehat{K}_{\mathbf{x}}.$$

We see that

$$\left\| \frac{1}{m}\widetilde{K}_{\mathbf{x}} - \frac{1}{m}\widehat{K}_{\mathbf{x}} \right\| \le \|\Sigma\| \left\| \frac{1}{m}\widehat{K}_{\mathbf{x}} \right\| \|\Sigma - I\| + \|\Sigma - I\| \left\| \frac{1}{m}\widehat{K}_{\mathbf{x}} \right\|.$$

Since $\frac{1}{m}\widehat{K}_{\mathbf{x}} \le I$, we have $\left\| \frac{1}{m}\widehat{K}_{\mathbf{x}} \right\| \le 1$. Also, for each i,

$$\left| \frac{\sqrt{p_{\mathbf{x}}(x_i)}}{\sqrt{p(x_i)}} - 1 \right| \le \frac{|p_{\mathbf{x}}(x_i) - p(x_i)|}{p(x_i)}.$$

Hence

$$\|\Sigma - I\| \le \max_{i=1,\dots,m} \frac{|p_{\mathbf{x}}(x_i) - p(x_i)|}{p(x_i)}.$$

This in connection with (8.4) proves the statement. □

Now we can prove the estimates for the approximation of eigenfunctions.

Proof of Theorem 2. First, we define a linear operator which plays the role of $\widehat{A}$ in Proposition 2. This is an operator, denoted as $\widehat{L}_{K,\mathbf{x}}$, on $\mathcal{H}_{\widetilde{K}}$ such that $\widehat{L}_{K,\mathbf{x}}(f) = 0$ for any $f \in \mathcal{H}_{\widetilde{K},\mathbf{x}}^{\perp}$, and $\mathcal{H}_{\widetilde{K},\mathbf{x}}$ is an invariant subspace of $\widehat{L}_{K,\mathbf{x}}$ with the matrix representation equal to the matrix $\frac{1}{m}\widehat{K}_{\mathbf{x}}$ given by (6.2). That is, in $\mathcal{H}_{\widetilde{K},\mathbf{x}} \bigoplus \mathcal{H}_{\widetilde{K},\mathbf{x}}^{\perp}$, we have

$$\widehat{L}_{K,\mathbf{x}}\left[\widehat{K}_{x_1},\ldots,\widehat{K}_{x_m}\right] = \left[\widehat{K}_{x_1},\ldots,\widehat{K}_{x_m}\right]\frac{1}{m}\widehat{K}_{\mathbf{x}}, \quad \widehat{L}_{K,\mathbf{x}}|_{\mathcal{H}_{\widetilde{K},\mathbf{x}}{}^{\perp}} = 0.$$

Second, we consider the difference between the operator $L_{\widetilde{K}}$ and $\widehat{L}_{K,\mathbf{x}}$. By Theorem 1, there exists a subset X_1 of X^m of measure at least $1-\delta/2$ such that (3.7) holds true for $\mathbf{x} \in X_1$. By Lemmas 1 and 2, we know that there exists another subset X_2 of X^m of measure at least $1-\delta/2$ such that (8.3) and (8.5) hold for $\mathbf{x} \in X_2$. Recall that $\widetilde{S}_{\mathbf{x}}^T\widetilde{S}_{\mathbf{x}}|_{\mathcal{H}_{\widetilde{K},\mathbf{x}}^{\perp}} = 0$ and $\widetilde{K}_{\mathbf{x}}$ is the matrix representation of the operator $\widetilde{S}_{\mathbf{x}}^T\widetilde{S}_{\mathbf{x}}|_{\mathcal{H}_{\widetilde{K},\mathbf{x}}}$. This in connection with (8.5) and the definition of $\widehat{L}_{K,\mathbf{x}}$ tells us that for $\mathbf{x} \in X_2$,

$$\left\|\frac{1}{m}\widetilde{S}_{\mathbf{x}}^T\widetilde{S}_{\mathbf{x}} - \widehat{L}_{K,\mathbf{x}}\right\| = \left\|\frac{1}{m}\widetilde{K}_{\mathbf{x}} - \frac{1}{m}\widehat{K}_{\mathbf{x}}\right\| \le \left(2 + \frac{4\kappa^2\log\big(4/\delta\big)}{\sqrt{m}p_0}\right)\frac{4\kappa^2\log\big(4/\delta\big)}{\sqrt{m}p_0}.$$

Together with (3.7), we have

$$\left\|L_{\widetilde{K}} - \widehat{L}_{K,\mathbf{x}}\right\| \le \left(3 + \frac{4\kappa^2\log\big(4/\delta\big)}{\sqrt{m}p_0}\right)\frac{4\kappa^2\log\big(4/\delta\big)}{\sqrt{m}p_0}, \qquad \forall \mathbf{x} \in X_1 \cap X_2. \tag{8.6}$$

Third, we apply Proposition 2 to the operators $A = L_{\widetilde{K}}$ and $\widehat{A} = \widehat{L}_{K,\mathbf{x}}$ on the Hilbert space $\mathcal{H}_{\widetilde{K}}$. Let $\mathbf{x}$ be in the subset $X_1 \cap X_2$ which has measure at least $1-\delta$. Since $r \le 1$, the estimate (8.6) together with the assumption (8.1) tells us that (7.2) holds. Then by Proposition 2, we have

$$\left\|\frac{1}{\sqrt{\lambda_k}}\varphi^{(k)} - \widehat{\varphi}^{(k)}\right\|_{\widetilde{K}} \le \frac{4}{r}\left\|L_{\widetilde{K}} - \widehat{L}_{K,\mathbf{x}}\right\| \le \frac{50\kappa^2\log\big(4/\delta\big)}{r\sqrt{m}p_0}, \tag{8.7}$$

where $\widehat{\varphi}^{(k)}$ is a normalized eigenfunction of $\widehat{L}_{K,\mathbf{x}}$ associated with eigenvalue $\widehat{\lambda}_{k,\mathbf{x}}$.

Finally, we specify $\widehat{\varphi}^{(k)}$. From the definition of the operator $\widehat{L}_{K,\mathbf{x}}$, we see that for $u \in \ell^2(\mathbf{x})$, we have

$$\widehat{L}_{K,\mathbf{x}}\left(\sum_{i=1}^m u_i\widetilde{K}_{x_i}\right) = \sum_{i=1}^m\left(\frac{1}{m}\widehat{K}_{\mathbf{x}}u\right)_i\widetilde{K}_{x_i}.$$

Since $\| \sum_{i=1}^m u_i \widetilde{K}_{x_i} \|_{\widetilde{K}}^2 = u^T \widetilde{K}_{\mathbf{x}} u$, we know that the normalized eigenfunction $\widehat{\varphi}^{(k)}$ of $\widehat{L}_{K,\mathbf{x}}$ associated with eigenvalue $\widehat{\lambda}_{k,\mathbf{x}}$ can be taken as

$$\widehat{\varphi}^{(k)} = \left(v^T \widetilde{K}_{\mathbf{x}} v \right)^{-1/2} \sum_{i=1}^m v_i \widetilde{K}_{x_i} = \left(\frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v \right)^{-1/2} \frac{1}{\sqrt{m}} \widetilde{S}_{\mathbf{x}}^T v$$

where $v \in \ell^2(\mathbf{x})$ is a normalized eigenvector of the matrix $\frac{1}{m}\widehat{K}_{\mathbf{x}}$ with eigenvalue $\widehat{\lambda}_{k,\mathbf{x}}$.

Since (8.5) holds, we see that

$$\left| \frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v - \widehat{\lambda}_{k,\mathbf{x}} \right| = \left| \frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v - \frac{1}{m} v^T \widehat{K}_{\mathbf{x}} v \right| = \left| v^T \left(\frac{1}{m} \widetilde{K}_{\mathbf{x}} - \frac{1}{m} \widehat{K}_{\mathbf{x}} \right) v \right| \le \frac{9\kappa^2 \log\bigl(4/\delta\bigr)}{\sqrt{m} p_0}.$$

Hence

$$\left| \frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v - \lambda_k \right| \le \left| \frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v - \widehat{\lambda}_{k,\mathbf{x}} \right| + |\widehat{\lambda}_{k,\mathbf{x}} - \lambda_k| \le \frac{22\kappa^2 \log\bigl(4/\delta\bigr)}{\sqrt{m} p_0}.$$

Since $\left(\frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v \right)^{-1/2} \left\| \frac{1}{\sqrt{m}} \widetilde{S}_{\mathbf{x}}^T v \right\|_{\widetilde{K}} = 1$, we have

$$\begin{aligned} \left\| \frac{1}{\sqrt{\lambda_k}} \frac{1}{\sqrt{m}} \widetilde{S}_{\mathbf{x}}^T v - \widehat{\varphi}^{(k)} \right\|_{\widetilde{K}} &= \left\| \left(\frac{1}{\sqrt{\lambda_k}} - \frac{1}{\sqrt{\frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v}} \right) \frac{1}{\sqrt{m}} \widetilde{S}_{\mathbf{x}}^T v \right\|_{\widetilde{K}} \\ &= \frac{\left| \frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v - \lambda_k \right|}{\sqrt{\lambda_k} + \sqrt{\frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v}} \frac{1}{\sqrt{\lambda_k} \sqrt{\frac{1}{m} v^T \widetilde{K}_{\mathbf{x}} v}} \left\| \frac{1}{\sqrt{m}} \widetilde{S}_{\mathbf{x}}^T v \right\|_{\widetilde{K}} \le \frac{22\kappa^2 \log\bigl(4/\delta\bigr)}{\sqrt{m} p_0 \lambda_k}. \end{aligned}$$

This in connection with (8.7) tells us that for $\mathbf{x} \in X_1 \cap X_2$ we have

$$\left\| \frac{1}{\sqrt{\lambda_k}} \varphi^{(k)} - \frac{1}{\sqrt{\lambda_k}} \frac{1}{\sqrt{m}} \widetilde{S}_{\mathbf{x}}^T v \right\|_{\widetilde{K}} \le \frac{50\kappa^2 \log\bigl(4/\delta\bigr)}{r\sqrt{m} p_0} + \frac{22\kappa^2 \log\bigl(4/\delta\bigr)}{\sqrt{m} p_0 \lambda_k}.$$

Since $r \le \lambda_k - \lambda_{k+1} \le \lambda_k$, this proves the stated error bound in Theorem 2. □

References

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

[2] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. **15** (2003), 1373–1396.

[3] M. Belkin and P. Niyogi, Convergence of Laplacian eigenmaps, preprint, 2007.

[4] R. Bhatia, Matrix Analysis, Graduate Texts in Mathematics **169**, Springer-Verlag, New York, 1997.

[5] G. Blanchard, O. Bousquet, and L. Zwald, Statistical properties of kernel principal component analysis, Mach. Learn. **66** (2007), 259–294.

[6] S. Bougleux, A. Elmoataz, and M. Melkemi, Discrete regularization on weighted graphs for image and mesh filtering, Lecture Notes in Computer Science **4485** (2007), pp. 128–139.

[7] F. R. K. Chung, Spectral Graph Theory, Regional Conference Series in Mathematics **92**, SIAM, Philadelphia, 1997.

[8] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps, PNAS of USA **102** (2005), 7426–7431.

[9] R. Coifman and M. Maggiono, Diffusion wavelets, Appl. Comput. Harmonic Anal. **21** (2006), 53–94.

[10] F. Cucker and S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. **39** (2001), 1–49.

[11] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.

[12] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. **5** (2005), 59–85.

[13] G. Gilboa and S. Osher, Nonlocal operators with applications to image processing, UCLA CAM Report 07-23, July 2007.

[14] V. Koltchinskii and E. Giné, Random matrix approximation of spectra of integral operators, Bernoulli **6** (2000), 113–167.

[15] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, Ann. Probab. **22** (1994), 1679–1706.

[16] S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. **41** (2004), 279–305.

[17] S. Smale and D. X. Zhou, Shannon sampling II. Connections to learning theory, Appl. Comput. Harmonic Anal. **19** (2005), 285–302.

[18] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. **26** (2007), 153–172.

[19] U. von Luxburg, M. Belkin, and O. Bousquet, Consistency of spectral clustering, Ann. Stat., to appear.

[20] G. B. Ye and D. X. Zhou, Learning and approximation by Gaussians on Riemannian manifolds, Adv. Comput. Math., in press. DOI 10.1007/s10444-007-9049-0

[21] D. Zhou and B. Schölkopf, Regularization on discrete spaces, in *Pattern Recognition*, Proc. 27th DAGM Symposium, Berlin, 2005, pp. 361–368.