Available online at www.sciencedirect.com
ScienceDirect
Appl. Comput. Harmon. Anal. 24 (2008) 329–353
Applied and Computational Harmonic Analysis
www.elsevier.com/locate/acha

Diffusion polynomial frames on metric measure spaces

M. Maggioni [a,*,1], H.N. Mhaskar [b,2]

[a] *Department of Mathematics and Computer Science, Duke University, Durham, NC 27708, USA*
[b] *Department of Mathematics, California State University, Los Angeles, CA 90032, USA*

Received 13 January 2007; revised 29 June 2007; accepted 8 July 2007

Available online 26 July 2007

Communicated by Charles K. Chui

Abstract

We construct a multiscale tight frame based on an arbitrary orthonormal basis for the L^2 space of an arbitrary sigma finite measure space. The approximation properties of the resulting multiscale are studied in the context of Besov approximation spaces, which are characterized both in terms of suitable K-functionals and the frame transforms. The only major condition required is the uniform boundedness of a summability operator. We give sufficient conditions for this to hold in the context of a very general class of metric measure spaces. The theory is illustrated using the approximation of characteristic functions of caps on a dumbell manifold, and applied to the problem of recognition of hand-written digits. Our methods outperforms comparable methods for semi-supervised learning.

Keywords: Diffusion multiscales; Polynomial frames; Metric measure spaces; Besov spaces; Semi-supervised learning; Recognition of hand-written digits

1. Introduction

A typical problem in machine learning is to predict the values of a target function given a finite amount of information about the function, such as its values at certain number of points. For example, given a few digitized images of digits, one wishes to develop a model that will predict for any other image whether the corresponding digit is 0. Each image may be viewed as a point in a high dimensional space, and the target function is the characteristic function of the set of points corresponding to the digit 0. An important problem in this theory is to ensure performance guarantees on the model on unseen data. The data is typically noisy, and a great deal of literature is devoted to strategies for approximating the function in spite of this noise. In approximation theory, we consider noise reduction to be a separate statistical issue, and focus on analyzing the intrinsic merits of different models and algorithms with respect to performance guarantees. To use a different language, we focus on analyzing the bias term rather than the variance

* Corresponding author.
E-mail addresses: mauro.maggioni@duke.edu (M. Maggioni), hmhaska@calstatela.edu (H.N. Mhaskar).
[1] The research of this author was supported, in part, by grant DMS-0650413 from the National Science Foundation.
[2] The research of this author was supported, in part, by grant DMS-0605209 from the National Science Foundation and grant W911NF-04-1-0339 from the U.S. Army Research Office.

1063-5203/$ – see front matter
doi:10.1016/j.acha.2007.07.001

term. The issues involved are perhaps best illustrated in the context of approximation of periodic functions by trigonometric polynomials. Our discussion below is based on [16,28,30,33]. The notation we use below is limited only to this discussion, and may be used in a somewhat different sense in the rest of the paper.

The starting point of this problem is $2N+1$ pieces of information about a 2π-periodic continuous function $f:\mathbb{R}\to\mathbb{R}$, where N is a positive integer. The information may consist of either values of f at $2N+1$ given points on $[-\pi,\pi)$ or Fourier coefficients $\hat{f}(k)$ of the function of order up to N. The class $\mathbb{H}_N$ of models which we are interested in consists of all trigonometric polynomials of order N. The performance guarantee by a trigonometric polynomial T of order N is measured in terms of the uniform norm, $\|f-T\|_\infty = \max_{x\in[-\pi,\pi]}|f(x)-T(x)|$. The "dream" performance guarantee is given by the degree of best uniform approximation, which we will denote by

$$E_N(f) := \min_{T\in\mathbb{H}_N} \|f-T\|_\infty.$$

Unfortunately, the construction of the trigonometric polynomial of best approximation is nonlinear as well as difficult. We are not aware of any construction based on a given data. An instinctive approach to approximate f is to find a trigonometric polynomial of order N that interpolates the data: Lagrange interpolation if the values are given, or Fourier projection if the Fourier coefficients are given. However, the trigonometric polynomials I_N obtained in this way satisfy only $\|f-I_N\|_\infty \leqslant c(\log N)E_N(f)$ at best. Here, and in the sequel, c will denote a generic constant, independent of f and N. The performance guarantee might, in fact, be very poor even if the optimal data is perturbed very slightly. For example, if the points at which interpolatory data is available has the form $\{\theta_\ell\}$, where $\cos\theta_\ell$ is a zero of an ultraspherical polynomial of degree N, then $\|f-I_N\|_\infty \leqslant cN^Q E_N(f)$ where Q depends upon the parameters involved in the definition of the ultraspherical polynomial. In view of the uniform boundedness principle of functional analysis, this means that the resulting procedures do not converge for every continuous target function f. There are two alternatives. Either one should allow trigonometric polynomials of higher order to interpolate, or forget about interpolation and focus only on obtaining good approximations. Since the first alternative implies an increase in the complexity of the model, the order of trigonometric polynomials in this context, the second alternative has been investigated far more thoroughly. Thus, for example, if the Fourier information is known, then one considers the operator

$$\sigma_N(H,f,x) := \sum_{|k|\leqslant N} H\big(|k|/N\big)\hat{f}(k)e^{ikx} =: \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_N(H,x-t)f(t)\,\mathrm{d}t,$$

where H is an even function which can be expressed as an indefinite integral of a function of bounded variation, $H(t)=1$ if $0\leqslant|t|\leqslant 1/2$, and $H(t)=0$ if $|t|\geqslant 1$. Then it is known that $\|f-\sigma_N(f)\|_\infty \leqslant cE_{N/2}(f)$ for every continuous, 2π-periodic function f. If the values of the function are known, one replaces $\hat{f}(k)$ above by an approximation using numerical integration. If the quadrature formulas are chosen carefully (but *not otherwise*), the resulting operators also yield the same performance guarantee.

We point out two additional advantages of the operator of the form σ_N. The first advantage is localization. If $S>1$ is an integer, and H is S times continuously differentiable, then

$$\big|\Phi_N(H,t)\big| \leqslant \frac{cn}{(1+n|t|)^S}, \quad t\in[-\pi,\pi).$$

Thus, in particular, if H is infinitely many times differentiable, then the kernel $\Phi_N(H,t)$ decays to 0 faster than any polynomial if $t\neq 0$. This property has the following interesting consequence. Suppose that $r<R$ are integers, and the target function is r times continuously differentiable on one interval $I\subset[-\pi,\pi)$ and R times continuously differentiable on another interval $J\subset[-\pi,\pi)$. In spite of the fact that $V_N(H,f)$ is constructed using the global information in the form of Fourier coefficients, by choosing H to be smooth enough, the performance guarantee of $V_N(H,f)$ is $\mathcal{O}(N^{-r})$ on I and $\mathcal{O}(N^{-R})$ on J. Moreover, one does not need an a priori knowledge of I, J, r, or R in the construction of the operator. In contrast, the Chebyshev alternation theorem implies that the trigonometric polynomial of best approximation cannot be localized in this sense.

In learning theory, one is often interested in finding the solution of a regularization problem, for example, among all the models M under consideration, choose the one that minimizes $\|f-M\|_2+\delta^r\|M^{(r)}\|_2$, where $\|\circ\|_2$ denotes the L^2 norm, and δ is the regularization parameter. In this context, the largest possible set of models is the class W_r^2 of all functions M for which $M^{(r)}$ exists almost everywhere, and $\|M^{(r)}\|_2<\infty$. The classes W_r^p can be defined similarly

using the L^p norm $\|\circ\|_p$. The second advantage of the operator σ_N is that it satisfies for *every* $r \geqslant 1$ and *every* p, $1 \leqslant p \leqslant \infty$,

$$\|f - \sigma_N(H, f)\|_p + N^{-r}\|\sigma_N(H, f)^{(r)}\|_p \leqslant c \inf_{M \in W_r^p} \{\|f - M\|_p + N^{-r}\|M^{(r)}\|_p\}.$$

Accordingly, in all the numerical experiments which we are familiar with, the operator σ_N is highly stable under noise. We note that the construction of the operators σ_N is universal, and does not require any optimization. The last infimum expression on the right-hand side of the above inequality is known as the K-functional for the spaces L^p and W_r^p.

There is a very close connection between the K-functionals and the quantities $E_N(f)$, given by the direct and converse theorems of approximation theory. In particular, the rate at which $E_N(f) \to 0$ as $N \to \infty$ determines the smoothness class to which f belongs. Necessarily, the smoothness class can also be characterized in the same way by the rate of convergence of $\|f - \sigma_N(H, f)\|_p$ or equivalently, by the behavior of the quantities $\sigma_{2^{n+1}}(H, f) - \sigma_{2^n}(f)$. These last quantities also have the frame properties.

There is a completely different perspective to look at this problem, which we mention for the sake of completeness, but will not elaborate upon in this paper. Rather than making trigonometric polynomials as the models of choice, one may wish to ask about the best performance guarantees for universal approximation of a given class of functions, such as the unit ball of W_r^p, given $2N+1$ pieces of information about an unknown target function in this class. This question leads to the idea of nonlinear widths. It is known that if the smoothness is measured in the same L^p norm as the one used to measure the performance guarantee, then the class $\mathbb{H}_N$ (and hence, the operator σ_N) yields asymptotically the best performance with this criterion as well. This fact holds also in multivariate settings. Clearly, the guarantee will be affected both by the class of target functions in question, and the norm in which the approximation is measured. The various "dimension-independent" bounds available in the literature are obtained by choosing a very different class of target functions.

In the context of classical trigonometric polynomial expansions, as well as other classical orthogonal expansions, the problem of constructing approximations localized in both the space and frequency domains, given the coefficients in such expansions, has been investigated a great deal in recent years [30–32]. A survey can be found in [29], where the ideas have also been generalized to construct such approximations and multiscales based on arbitrary orthogonal systems on general sigma finite measure spaces. The purpose of this paper is to extend the entire paradigm described above to the context of approximation by such systems in metric measure spaces, in particular, low dimensional, unknown manifolds embedded in a high dimensional Euclidean space, called the ambient space.

The problem of approximation in this very general context arises in several such practical applications as document analysis [15], face recognition [20], semi-supervised learning [5,36,40], nonlinear image denoising and segmentation [40], processing of articulated images [17], cataloguing of galaxies [18], pattern analysis of brain potentials [23], and the study of brain tumors [8]. Much recent research focuses on developing techniques to take advantage of the lower intrinsic dimensionality in order to learn functions on the data more efficiently than classical techniques developed for learning functions directly on the ambient space. Often, this lower dimensional structure is modeled as a smooth Riemannian manifold or a graph.

Some approaches are based on heat kernels methods [40], on the eigenfunctions of a Laplacian operator naturally defined on the graph/manifold [2,4,5,13,22,26,27,34], on associated multiscale analyses [14,15,24,25,41], and finally on harmonic functions on graphs [21,45]. In particular, ideas from spectral graph theory [11] have recently been applied to function approximation and learning on manifolds and graphs. Analysis by decomposing a function in a series of the eigenfunctions of the graph Laplacian is a classical generalization of Fourier analysis, which has been successful in applications to problems in semi-supervised learning, data mining, and reinforcement learning [5,13, 25,27,34]. The papers [15,40] also discuss some of these and other applications. To the best of our knowledge, the recent construction of diffusion wavelets [15,24] and other multiscale bases [41] on graphs and manifolds are the first attempts to develop tools for the multiscale analysis on these spaces.

The goal of the present paper is to develop novel approximation methods with strong guarantees regarding their asymptotic performance. Initially, we will develop this theory in the generality of arbitrary measure spaces as in [29], based on an orthonormal basis $\{\phi_j\}$ for the corresponding L^2 space and an increasing sequence of numbers $\{\lambda_j\}$ such that $\lim_{j\to\infty} \lambda_j = \infty$. As is customary in learning theory, and approximation theory in general, the guarantees on the approximation properties require an a priori assumption on the smoothness of the target function, measured

in a manner appropriate for the application. Following [29], we measure the smoothness by membership in certain Besov spaces defined using the degree of best possible approximation from spans of $\Pi_\Lambda = \{\phi_j\colon\ \lambda_j \leqslant \Lambda\}$ as $\Lambda \to \infty$, and characterize these spaces in terms of a suitable K-functional, as in [42]. The frames are developed using a small variation of the ideas in [29], and a characterization of the Besov spaces is obtained using the frame transform. Unlike classical multiscale analysis, our theory is applicable to the approximation of functions in the L^p-closures of $\bigcup_{\Lambda\geqslant 0}\Pi_\Lambda$ for *every* p with $1 \leqslant p \leqslant \infty$. Of particular interest is the case $p = \infty$, where the Fourier projections of every function do not necessarily converge. We note that our approximation operators are universal; i.e., their construction does not depend upon any a priori assumptions on the target function, even though the guarantees are necessarily dependent on these assumptions. Further, there is no saturation for the approximation power of our operators: it is possible to obtain an arbitrarily fast rate of decrease for the degrees of approximation, if the target function is sufficiently smooth.

As in [29,42], our constructions assume the uniform boundedness of certain summability operators. In the context of a smooth manifold, when the functions ϕ_j and the numbers λ_j are eigenfunctions and eigenvalues of the Laplace–Beltrami operator, the uniform boundedness of these operators can be proved using the uniform boundedness of the Bochner–Riesz means, proved by Sogge [39]. We prove the necessary bounds in the more general context of quasi-metric spaces and general orthogonal systems, assuming only the finite speed of wave propagation, and the assumption that the measures of balls as well as the Christoffel functions (on-diagonal bounds) based on the system $\{\phi_j\}$ have a polynomial growth. Our proof is similar in spirit to that of Sogge, but the technical details are quite different. In particular, we use detailed bounds on simultaneous approximation by entire functions of finite exponential type rather than an asymptotics for Bessel functions.

Finally, we illustrate our theory using two examples: a dumbell manifold, and the problem of hand-written digit recognition. In the case of the dumbell manifold, we show that our techniques give a sharply localized approximation to the characteristic functions of certain caps, in contrast to ordinary Fourier projections. In the case of the digit recognition problem, we use our techniques for semi-supervised learning based on a standard data set called MNIST, consisting of 60,000 examples. We use a subset of 10,000 of these examples, out of which a very small percentage is used for training. Our approximation methods are used to obtain results which are substantially better than those of comparable methods in the literature available to us.

In Section 2, we define the frames in the context of general measure spaces. In Section 3, we will use the ideas in [42] and [29] to obtain a characterization of certain Besov spaces. These spaces are defined in this context in terms of the degrees of approximation of the target function. The characterizations are given in terms of the smoothness properties of the function, measured by a K-functional, as well as in terms of the frame transforms developed in Section 2. In Section 4, we restrict ourselves to the case of metric measure spaces, and obtain certain mild sufficient conditions on these to ensure that the operators required in the theory in Sections 2 and 3 are uniformly bounded. The numerical experiments are described in Section 5, and the proofs of the new results are given in Section 6.

2. Frames

Let $(\mathbb{X}, \mu)$ be a sigma-finite measure space, $\{\phi_j\}_{j=0}^\infty$ be a complete orthonormal set in $L^2 := L^2(\mathbb{X}, \mu)$, $0 = \lambda_0 \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \cdots$ be a sequence with $\lim_{j\to\infty}\lambda_j = \infty$. In the sequel, we will write L^p in place of $L^p(\mathbb{X}, \mu)$, and assume that each $\phi_j \in L^1 \cap L^\infty$. For any $\lambda \geqslant 0$, let

$$\Pi_\lambda := \operatorname{span}\{\phi_j\colon\ \lambda_j \leqslant \lambda\},$$

and $\Pi_\infty := \bigcup_{\lambda\geqslant 0}\Pi_\lambda$. An element of Π_∞ will be referred to as a *diffusion polynomial*. In most applications, the functions ϕ_j will be the eigenfunctions of a Laplacian operator. We will define

$$E_{\lambda,p}(f) := \min_{P\in\Pi_\lambda}\|f - P\|_p, \quad f \in L^p,\ 1 \leqslant p \leqslant \infty,\ \lambda \geqslant 0. \tag{2.1}$$

We denote by X^p the L^p closure of Π_∞; i.e., the subspace of L^p comprising of functions f such that $E_{\lambda,p}(f) \downarrow 0$ as $\lambda \to \infty$.

In this section, we describe a general construction of certain frame operators. This construction is similar to that in [29], but is slightly different. In the sequel, $c, c_1, \ldots,$ will denote generic constants, whose value may depend upon the fixed parameters in the discussion, such as $\mathbb{X}$, μ, p, the sequences $\{\lambda_j\}$, $\{\phi_j\}$ (but not the individual terms of these),

and the smoothness parameters to be introduced later, but may be different at different occurrences, even within a single formula.

As usual, we define $L^1 + L^\infty$ to be the class of all f such that there exist $f_1 \in L^1$, $f_2 \in L^\infty$, such that $f = f_1 + f_2$. Of course, this decomposition may not be unique. We note that $L^1 \cap L^\infty \subseteq L^p \subseteq L^1 + L^\infty$, $1 \leqslant p \leqslant \infty$. For any $f \in L^1 + L^\infty$, let

$$\hat{f}(\ell) := \int f \phi_\ell \, \mathrm{d}\mu, \quad \ell = 0, 1, \ldots. \tag{2.2}$$

Let $h : [0, \infty) \to \mathbb{R}$, $f \in L^1 + L^\infty$, $x, y \in \mathbb{X}$, and $\lambda > 0$. We define formally the *summability kernel* by

$$\Phi_\lambda(h, x, y) := \sum_{j=0}^{\infty} h(\lambda_j/\lambda)\phi_j(y)\phi_j(x), \tag{2.3}$$

and the *summability operator* by

$$\sigma_\lambda(h, f, x) := \int \Phi_\lambda(h, x, y) f(y) \, \mathrm{d}\mu(y) = \sum_{j=0}^{\infty} h(\lambda_j/\lambda)\hat{f}(j)\phi_j(x). \tag{2.4}$$

This operator is similar to the operators defined in [29], except that the multiplying factors are $h(\lambda_j/\lambda)$ rather than $h(j/\lambda)$. It is convenient to define $\Phi_0(h, x, y) = \sum_{\lambda_j = 0} \phi_j(x)\phi_j(y)$, and

$$\sigma_0(h, f) = \mathbb{P}_0(f) := \sum_{\lambda_j = 0} \hat{f}(j)\phi_j. \tag{2.5}$$

A function $h : [0, \infty) \to \mathbb{R}$ will be called a *multiplier mask* (for the system $(\{\phi_j\}, \{\lambda_j\})$) if

$$\sup_{x \in \mathbb{X},\, \lambda > 0} \int \left|\Phi_\lambda(h, x, y)\right| \mathrm{d}\mu(y) < \infty. \tag{2.6}$$

The class of all multiplier masks for the system $(\{\phi_j\}, \{\lambda_j\})$ will be denoted by $\mathcal{M} := \mathcal{M}(\{\phi_j\}, \{\lambda_j\})$.

Proposition 2.1. *Let $h : [0, \infty) \to \mathbb{R}$. The following are equivalent*:

(a) $h \in \mathcal{M}$.

(b) *For every $f \in L^1$, and $\lambda \geqslant 0$,*

$$\left\|\sigma_\lambda(h, f)\right\|_1 \leqslant c(h)\|f\|_1. \tag{2.7}$$

(c) *For every $1 \leqslant p \leqslant \infty$, $f \in L^p$, and $\lambda \geqslant 0$,*

$$\left\|\sigma_\lambda(h, f)\right\|_p \leqslant c(h)\|f\|_p. \tag{2.8}$$

In order to define our frame operators, we now assume that $h : [0, \infty) \to \mathbb{R}$ is a nonincreasing function such that $h(t) = 1$ if $0 \leqslant t \leqslant 1/2$, and $h(t) = 0$ if $t \geqslant 1$. We define the tight frame kernel by

$$\Psi_n^*(h; x, y) := \begin{cases} \sum_{\lambda_j < 1} \phi_j(x)\phi_j(y), & \text{if } n = 0, \\ \sum_{\lambda_j \geqslant 1} \sqrt{h(\lambda_j/2^n) - h(\lambda_j/2^{n-1})}\phi_j(x)\phi_j(y), & n = 1, 2, \ldots, \end{cases} \tag{2.9}$$

and the tight frame operator by

$$\tau_n^*(h, f, x) := \int \Psi_n^*(h; x, y) f(y) \, \mathrm{d}\mu(y), \quad f \in L^1 + L^\infty, \ x \in \mathbb{X}. \tag{2.10}$$

We will need another frame operator, defined for $f \in L^1 + L^\infty$ by

$$\tau_n(h, f) := \begin{cases} \sigma_1(h, f) & \text{if } n = 0, \\ \sigma_{2^n}(h, f) - \sigma_{2^{n-1}}(h, f), & n = 1, 2, \ldots. \end{cases} \tag{2.11}$$

The following theorem shows, in particular, that the operators τ_n^* provide a tight frame for L^2. One may view (2.12) below as a Littlewood–Paley decomposition of f in any X^p.

Theorem 2.1. *Let $h:[0,\infty)\to\mathbb{R}$ be a nonincreasing function such that $h(t)=1$ if $0\leqslant t\leqslant 1/2$, and $h(t)=0$ if $t\geqslant 1$, and $f\in L^2$. We have*

$$f=\sum_{n=0}^{\infty}\tau_n(h,f)=\sum_{n=0}^{\infty}\int \tau_n^*(h,f,y)\Psi_n^*(h;\circ,y)\,\mathrm{d}\mu(y), \tag{2.12}$$

with convergence in the sense of L^2. Moreover,

$$\|f\|_2^2=\sum_{n=0}^{\infty}\big\|\tau_n^*(h,f)\big\|_2^2, \tag{2.13}$$

and

$$\sum_{n=0}^{\infty}\big\|\tau_n(h,f)\big\|_2^2\leqslant\|f\|_2^2\leqslant 5\sum_{n=0}^{\infty}\big\|\tau_n(h,f)\big\|_2^2. \tag{2.14}$$

If, in addition, $h\in\mathcal{M}$, then for every $1\leqslant p\leqslant\infty$ and $f\in L^p$, we have

$$E_{\lambda,p}(f)\leqslant\big\|f-\sigma_\lambda(h,f)\big\|_p\leqslant cE_{\lambda/2,p}(f),\quad \lambda>0. \tag{2.15}$$

In particular, (2.12) *holds for $f\in X^p$, with convergence in the sense of L^p.*

Even though the proof of Theorem 2.1 does not explicitly require any smoothness condition on h, the proof of the condition (2.6) typically requires that h have a certain number of derivatives with bounded variation. In Theorem 4.1 below, we will demonstrate how the smoothness of h determines the localization properties of the kernels $\sigma_\lambda(h)$, a fact verified already in a number of other cases ([29] and references therein).

For any integer $S\geqslant 1$, the space BV_0^S consists of compactly supported functions $h:[0,\infty)\to\mathbb{R}$ which can be expressed in the form

$$h(x)=\frac{(-1)^S}{S!}\int_x^\infty (t-x)^S\,\mathrm{d}h^{(S)}(t)=\frac{(-1)^S}{S!}\int_0^\infty (t-x)_+^S\,\mathrm{d}h^{(S)}(t),\quad x\geqslant 0,$$

where $h^{(S)}$ is a compactly supported function having bounded variation on $[0,\infty)$, and $y_+:=\max(y,0)$. For functions $h\in BV_0^S$, the summability condition (2.6) can be reduced to the summability of the Bochner–Riesz means of order S. For $\lambda>0$, $S>0$, the Bochner–Riesz means are defined by

$$\mathcal{R}_{\lambda,S}(f):=\sum_{j=0}^{\infty}\left(1-\frac{\lambda_j}{\lambda}\right)_+^S\hat{f}(j)\phi_j. \tag{2.16}$$

We observe that $\mathcal{R}_{\lambda,S}(f)\in\Pi_\lambda$. If $S\geqslant 1$ is an integer, we will say that the *Riesz condition* is satisfied with *exponent* S if

$$\big\|\mathcal{R}_{\lambda,S}(f)\big\|_1\leqslant c\|f\|_1,\quad \lambda>0,\ f\in L^1.$$

We will prove in Proposition 6.1 that the Riesz condition is satisfied with an exponent S if and only if (2.6) is satisfied for every $h\in BV_0^S$.

3. Besov spaces

In this section, we will give different characterizations for Besov spaces. There are many ways of defining Besov spaces for functions defined on subsets of a Euclidean space. Our definitions are motivated by the treatment in [16], where these spaces are defined for real intervals and the periodic setting in terms of moduli of smoothness. Several characterizations of these spaces are given in [16]. Since it is not natural to define moduli of smoothness for functions defined on arbitrary metric spaces, we adopt the characterization of Besov spaces as "approximation spaces" as the definition of Besov spaces.

For a sequence $\mathbf{s} = \{s_n\}_{n=0}^{\infty}$, and $a, \rho > 0$, we write

$$\|\mathbf{s}\|_{\rho,a} := \begin{cases} \{\sum_{n=0}^{\infty}(2^{na}|s_n|)^{\rho}\}^{1/\rho}, & \text{if } 0 < \rho < \infty, \\ \sup_{n\geqslant 0, n\in\mathbb{Z}} 2^{na}|s_n|, & \text{if } \rho = \infty. \end{cases} \tag{3.1}$$

The sequence $\mathbf{s}$ is said to be in the space $\mathsf{b}_{\rho,a}$ if $\|\mathbf{s}\|_{\rho,a} < \infty$. We define the Besov space $B^a_{p,\rho}$ to be the class of all $f \in X^p$ such that $\{E_{2^n,p}(f)\}_{n=0}^{\infty} \in \mathsf{b}_{\rho,a}$.

In order to relate the Besov spaces more explicitly with the smoothness of the function involved, we use the notion of a K-functional. First, we define the operator Δ^* on Π_∞ by $\Delta^*\phi_j = \lambda_j\phi_j$. If $\psi : [0,\infty) \to \mathbb{R}$, we will define $\psi(\Delta^*)\phi_j := \psi(\lambda_j)\phi_j$. If ψ is not defined at 0, but is defined only on $(0,\infty)$ instead, we will extend it to 0 by setting $\psi(0) = 0$. Thus, in particular, if $r < 0$,

$$(\Delta^*)^{-r}\phi_j = \begin{cases} \lambda_j^{-r}\phi_j, & \text{if } \lambda_j \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

For $1 \leqslant p \leqslant \infty$, and $\psi : [0,\infty) \to \mathbb{R}$, the space W^p_ψ consists of all $f \in X^p$ for which there exists a function $\psi(\Delta^*)f \in X^p$ with $\widehat{\psi(\Delta^*)f}(j) = \psi(\lambda_j)\hat{f}(j)$, $j = 0, 1, \ldots$. If $\psi(t) = t^r$ for some $r \in \mathbb{R}$ (with the above convention), we will write W^p_r in place of W^p_ψ. The K-functional (between X^p and W^p_r) is defined by

$$K_r(f,\delta) := \inf_{g\in W^p_r}\left\{\|f-g\|_p + \delta^r\left\|(\Delta^*)^r g\right\|_p\right\}, \quad r > 0,\ \delta > 0,\ f \in X^p. \tag{3.2}$$

The following theorem describes the characterizations of Besov spaces in terms of the K-functionals and also in terms of the frame operators.

Theorem 3.1. *Let $h : [0,\infty) \to \mathbb{R}$ be a nonincreasing function such that $h(t) = 1$ if $0 \leqslant t \leqslant 1/2$, and $h(t) = 0$ if $t \geqslant 1$. Let $a > 0$, $0 < \rho \leqslant \infty$, and $r > a$ be an integer.*

(a) *If $h \in \mathcal{M}$, then $f \in B^a_{p,\rho}$ if and only if $\{\|\tau_n(h,f)\|_p\} \in \mathsf{b}_{\rho,a}$.*
(b) *If the function $t \mapsto \sqrt{h(t) - h(2t)}$ as well as h are in $\mathcal{M}$, then $f \in B^a_{p,\rho}$ if and only if $\{\|\tau^*_n(h,f)\|_p\} \in \mathsf{b}_{\rho,a}$.*
(c) *If $S \geqslant 1$ is an integer, and the Riesz condition is satisfied with exponent S, then $f \in B^a_{p,\rho}$ if and only if $\{K_r(f, 2^{-n})\} \in \mathsf{b}_{\rho,a}$.*

The proof of this theorem has an interesting consequence that brings out the regularization properties of the kernels σ_λ.

Proposition 3.1. *Let $h : [0,\infty) \to \mathbb{R}$ be a nonincreasing function such that $h(t) = 1$ if $0 \leqslant t \leqslant 1/2$, and $h(t) = 0$ if $t \geqslant 1$. Let $1 \leqslant p \leqslant \infty$ and $f \in X^p$. If $h \in \mathcal{M}$ then for every $\delta > 0$,*

$$\left\|f - \sigma_{\delta^{-1}}(h,f)\right\|_p + \delta^r\left\|(\Delta^*)^r\sigma_{\delta^{-1}}(h,f)\right\|_p \leqslant cK_r(f,\delta). \tag{3.3}$$

4. Bounded summability operators

In the case of a smooth manifold with a boundary, Sogge [39] has proved that the Riesz condition holds for all sufficiently large exponents. Our objective in this section is to extend Sogge's argument to the context of certain metric measure spaces.

In the remainder of this section, we find it convenient to write $\ell_j = \sqrt{\lambda_j}$. We would also like to define an adaptation of the operator σ_λ and the kernels Φ_λ. If $H : \mathbb{R} \to \mathbb{R}$, we define formally

$$\tilde{\sigma}_\Lambda(H,f) = \sum_j H(\ell_j/\Lambda)\hat{f}(j)\phi_j, \tag{4.1}$$

and the corresponding kernel by

$$\tilde{\Phi}_\Lambda(H,x,y) = \sum_j H(\ell_j/\Lambda)\phi_j(x)\phi_j(y), \quad x, y \in \mathbb{X}. \tag{4.2}$$

Let $H(x) = h(x^2)$, $x \in \mathbb{R}$. If $h \in BV_0^S$, then H is also an S times iterated integral of a function $H^{(S)}$ having bounded variation on $\mathbb{R}$. We observe that $\sigma_\lambda(h, f) = \tilde{\sigma}_{\sqrt{\lambda}}(H, f)$, and $\Phi_\lambda(h, x, y) = \tilde{\Phi}_{\sqrt{\lambda}}(H, x, y)$.

In this section, we assume that there is a quasi-metric d defined on $\mathbb{X}$, such that the space $\mathbb{X}$ with the corresponding topology is sigma-compact, and the functions $\{\phi_j\}$ are continuous. (A quasi-metric d is a function on $\mathbb{X} \times \mathbb{X}$ satisfying all the conditions of a metric, except that in place of the triangle inequality, one has an estimate of the form $d(x, y) \leqslant c(d(x, z) + d(z, y))$ for all $x, y, z \in \mathbb{X}$.) Also, the measure μ will be assumed to be a complete, regular, Borel measure with respect to this topology. In the sequel, for $x \in \mathbb{X}$ and $r > 0$, let $B(x, r) := \{y \in \mathbb{X}\colon\ d(x, y) \leqslant r\}$, and $\Delta(x, r) := \mathbb{X} \setminus B(x, r)$. An important example, of course, is the case when $\mathbb{X}$ is a smooth manifold, μ is its volume element, d is the geodesic distance, $-\lambda_j$s and ϕ_js are the eigenvalues and the corresponding eigenfunctions of the Laplace–Beltrami operator. Rather than repeating all these assumptions, we will refer to them collectively as the smooth manifold case.

We need two assumptions on the metric, the measure, the sequence $\{\lambda_j\}$, and the system $\{\phi_j\}$. First, the numbers λ_j and the system $\{\phi_k\}$ are related by the following Christoffel function (or "on-diagonal") estimates: There exists a positive constant K, such that for every $x \in \mathbb{X}$ and $r \geqslant 1$,

$$\sum_{\ell_j \leqslant r} \phi_j^2(x) \leqslant c\big(\max(r, 1)\big)^K. \tag{4.3}$$

Next, the metric and the measure are related by the estimate: There exists a positive constant α such that for every $x \in \mathbb{X}$ and $r > 0$,

$$0 < \mu\big(\{y\colon\ d(x, y) \leqslant r\}\big) = \mu\big(\{y\colon\ d(x, y) < r\}\big) \leqslant cr^\alpha.$$

For example, if $\mathbb{X} = [-1, 1]$, $d(x, y) = |x - y|$, μ is the arcsine measure, i.e., $d\mu^*(x) = (1 - x^2)^{-1/2}dx$, ϕ_j is the orthonormalized Chebyshev polynomial of degree j, then $\alpha = 1/2$, $K = 1$. By redefining $d(x, y) = |\arccos x - \arccos y|$, we obtain $\alpha = K = 1$. In the smooth manifold case, $K = \alpha$. In the general case, we may define $d_1(x, y) = d(x, y)^{\alpha/K}$ to obtain another quasi-metric on $\mathbb{X}$, which yields the same topology as d, but for which the parameter $\alpha = K$. Thus, there is no loss of generality in assuming that

$$0 < \mu\big(\{y\colon\ d(x, y) \leqslant r\}\big) = \mu\big(\{y\colon\ d(x, y) < r\}\big) \leqslant cr^K. \tag{4.4}$$

Next, we discuss the wave kernel. For $f_1, f_2 \in L^2$, we will write

$$W(t, f_1, f_2) := \sum_{j=0}^{\infty} \cos(t\ell_j)\hat{f}_1(j)\hat{f}_2(j). \tag{4.5}$$

In view of Schwarz inequality followed by Bessel inequality, we see that for any $f_1, f_2 \in L^2$,

$$\sum_{j=0}^{\infty} \big|\hat{f}_1(j)\big|\big|\hat{f}_2(j)\big| \leqslant \left\{\sum_{j=0}^{\infty} \big|\hat{f}_1(j)\big|^2\right\}^{1/2} \left\{\sum_{j=0}^{\infty} \big|\hat{f}_2(j)\big|^2\right\}^{1/2} \leqslant \|f_1\|_2\|f_2\|_2 < \infty. \tag{4.6}$$

Thus, the expression $W(t, f_1, f_2)$ is well defined for all $f_1, f_2 \in L^2$ and $t \geqslant 0$. If $f : \mathbb{X} \to \mathbb{R}$, we define $\operatorname{supp}(f)$ to be the closure of the set $\{x \in \mathbb{X}\colon\ f(x) \neq 0\}$. The finite speed of wave propagation means that $W(t, f_1, f_2) = 0$ if $t \leqslant \gamma \operatorname{dist}(\operatorname{supp}(f_1), \operatorname{supp}(f_2))$ for some $\gamma > 0$. Again, by redefining the quasi-metric suitably, we may assume that $\gamma = 1$. In the sequel, we will make this assumption in this context. It has been proved that the property of finite speed of wave propagation holds in the smooth manifold case. In general, Sikora [37] has proved an equivalence between this property and the behavior of the corresponding heat kernel. Thus, finite speed of wave propagation does not hold for certain fractal domains considered by Grigor'yan [19]. However, we are not aware of the behavior of the wave kernel is this context. Accordingly, we find it convenient to assume this property in the sequel.

The following theorem states the desired localization estimate.

Theorem 4.1. *Let* (4.4), (4.3)*, and the finite speed of wave propagation hold. Let $S > K$ be an integer, $h \in BV_0^S$, $H(u) := h(u^2)$, $u \in \mathbb{R}$. Then for $\Lambda > 0$, $x, y \in \mathbb{X}$, $x \neq y$,*

$$\big|\Phi_{\Lambda^2}(h, x, y)\big| = \big|\tilde{\Phi}_\Lambda(H, x, y)\big| \leqslant c\big\||H|\big\|_S \Lambda^K \big(\Lambda d(x, y)\big)^{-S}, \tag{4.7}$$

where

$$\big\|\,|H|\,\big\|_S := \sum_{k=0}^{S} \max_{u\in\mathbb{R}} \big|H^{(S)}(u)\big|. \tag{4.8}$$

We note that estimates analogous to (4.7) have been proved in a number of contexts; the survey [29] contains references to some recent works in this direction. The following theorem about the Riesz condition is a standard consequence of the localization theorem above.

Theorem 4.2. *We assume the notations and conditions in Theorem* 4.1. *Then for every* p, $1 \leqslant p \leqslant \infty$, *and* $f \in L^p$,

$$\big\|\tilde{\sigma}_\Lambda(H, f)\big\|_p \leqslant c\big\|\,|H|\,\big\|_S \|f\|_p. \tag{4.9}$$

In particular, every function in BV_0^S *is also a multiplier mask for* $(\{\phi_j\}, \{\lambda_j\})$.

Since space–frequency localized wavelets are also constructed in [15] and [24], we make some remarks comparing our constructions with those in [15,24]. Our scaling spaces are the same as those in [15,24]. While the starting point in [15,24] is a compactly supported "heat kernel" with a exponentially decaying spectrum, the starting point of our paper is the set of eigenvalues and eigenfunctions of this kernel. These eigenfunctions are not localized in the space domain. The functions $\Phi_\lambda(h, \circ, t)$ defined in (2.3) are bandlimited, and become more and more space localized as $\lambda \to \infty$. Moreover, this localization may be controlled by choosing a sufficiently smooth function h. Such a control in localization is only partially achieved in [15], due to a lack of a control parameter similar to λ, and in [24], where localization is achieved only at the expense of the condition number of the bases obtained. We do not define orthogonal wavelet spaces, and do not need the expensive Gram–Schmidt procedure to orthogonalize a natural basis for these, which is needed explicitly in [15] and implicitly in [24]. Instead, here we have defined tight frames, and another set of frames so that their dual frames have similar space–frequency localization. These frames are easy to compute once the eigenvalues and eigenfunctions are given: the computation of these is in general through diagonalization of an integral (e.g. the heat kernel) or differential (e.g. the Laplacian) operator, and therefore is very expensive. The heat kernel needed in [15] is typically fast to construct thanks to its locality: this is the main computational tradeoff between the two constructions. A major difference between this paper and [15] is that we studied the approximation theory for our operators to the full extent as explained in the introduction in the trigonometric case.

5. Numerical examples

In this section, we illustrate and supplement the theory developed in the previous sections. In Section 5.1, we will review some results concerning eigenfunctions of a graph Laplacian. In Section 5.2, we describe our construction of the summability operators used in our numerical experiments. The results of the experiments in the case of the dumbell manifold are described in Section 5.3 and those for the case of recognition of hand-written digits are described in Subsection 5.4. The time complexity of our constructions is described in Section 5.5.

5.1. Eigenfunctions on graphs

Most of the typical applications of our theory involve the computation of the eigenvalues and eigenfunctions of a Laplacian operator defined on a finite graph. For example, if $\mathbb{X}$ is a smooth manifold, and the target function is known at finitely many points X on this manifold, then we may approximate the Laplace–Beltrami operator on $\mathbb{X}$ by a corresponding operator on a graph with X as vertices ([3,22,38] and references therein). In this subsection, we review some of the facts regarding graph Laplacians and their eigenfunctions.

Let (X, W) be a weighted undirected graph, where X is a finite set of vertices, and W is a symmetric $|X| \times |X|$ matrix with nonnegative entries, so that $W(x, y)$ represents the weight of an edge between x and y if $x \neq y$, and $W(x, x) > 0$. The sum of the weights of all edges adjacent to any x is then positive, and defines a function $D(x)$ on X by the formula $D(x) = \sum_{y\in X} W(x, y)$. The graph Laplacian is then defined by

$$\Delta_{(X,W)} f(x) = \sum_{y\in X} \frac{W(x, y)}{D(x)} f(y) - f(x). \tag{5.1}$$

We recall a result in [38]. Let X be a uniform sample from a smooth, compact manifold $\mathbb{X}$ embedded in the Euclidean space $\mathbb{R}^n$ for some integer $n \geqslant 1$, and Δ_B be the Laplace–Beltrami operator of $\mathbb{X}$. For $\varepsilon > 0$, we define the matrix $W = W_\varepsilon$ by

$$W(x, y) = \exp\left(-\frac{\|x - y\|_{\mathbb{R}^n}^2}{2\varepsilon}\right),$$

defining thereby a weighted graph structure on X. If f is any compactly supported, infinitely differentiable function on $\mathbb{X}$, Singer [38] has shown that for every $x \in X$

$$\frac{1}{\varepsilon}\Delta_{(X,W)}(f)(x) = \frac{1}{2}\Delta_B(f)(x) + O\left(\frac{1}{|X|^{\frac{1}{2}}\varepsilon^{\frac{1}{2}+\frac{n}{4}}} + \varepsilon\right). \tag{5.2}$$

5.2. Construction of the summability operators

A standard way to construct the function h which satisfies the conditions of Theorem 2.1 and has a given number of derivatives is the following. We recall that for $S \geqslant 1$, the cardinal B-spline of order S is the function M_S defined by (cf. [7, p. 131], [10, Theorem 4.3])

$$\begin{aligned} M_1(x) &:= \begin{cases} 1, & \text{if } 0 < x \leqslant 1, \\ 0, & \text{otherwise,} \end{cases} \\ M_{S+1}(x) &:= \frac{1}{S}\big\{xM_S(x) + (S + 1 - x)M_S(x - 1)\big\}, \quad S \geqslant 1. \end{aligned} \tag{5.3}$$

It is well known [10, Theorem 4.3] that each M_S is $S-2$ times continuously differentiable on $\mathbb{R}$, $M_S^{(S-1)}$ is a piecewise constant function, $M_S(x) = 0$ if $x \notin [0, S]$, $M_S(x) > 0$ if $s \in (0, S)$, and $\sum_{k\in\mathbb{Z}} M_S(x - k) = 1$ for all $x \in \mathbb{R}$. We consider the function h_S defined by

$$h_S(x) = \sum_{j=-S-1}^{S+1} M_{S+1}\big(2(S + 1)x - j\big), \quad x \geqslant 0. \tag{5.4}$$

It is not difficult to verify that the function h_S is nonincreasing on $[0, \infty)$, is $S - 1$ times continuously differentiable on $\mathbb{R}$, $h_S^{(S)}$ is a piecewise constant function, $h_S(x) = 1$ if $|x| < 1/2$, and $h_S(x) = 0$ if $|x| > 1$.

The operators $\sigma_\lambda(h_S)$ are closely related to the Bochner–Riesz means. Using the relationship [10, Formula 4.1.12]:

$$M_{S+1}(t) = \sum_{r=0}^{S+1} \frac{(-1)^r}{S!}\binom{S+1}{r}(t - r)_+^S, \quad t \in \mathbb{R},$$

we can deduce that for any $f \in L^1 + L^\infty$,

$$\sigma_\lambda(h_S, f) = \sum_{m=-S}^{S} \sum_{r=\max\{0,-m+1\}}^{S+1} \frac{(-1)^{S+1+r}}{S!}\binom{S+1}{r}(r + m)^S \mathcal{R}_{(r+m)\lambda/(2S+2),S}(f). \tag{5.5}$$

In the remainder of this section, we will write $\sigma_{J,S}^*(f) = \sigma_{\lambda_J}(h_S, f)$.

5.3. Dumbell manifold

We consider a dumbell-shaped manifold. We sample 8000 points on this manifold, and approximate the Laplace–Beltrami operator on this manifold as in (5.2), with $\varepsilon = 0.025$. We compute the lowest 800 eigenvalues and the corresponding eigenfunctions, some of which are shown in Fig. 1.

In Fig. 2 we represent two rows of the kernel associated with $\sigma_{800,12}^*$. We now consider the characteristic function f of a "slanted cap" on the dumbell, as in Fig. 3. We approximate f by $\sigma_{800,12}^*(f)$. Since $\sigma_{800,12}^*(P) = P$ if $P \in \Pi_{\lambda_{800}/2}$, we compare this approximation with the approximation by the band-limited projection $P_{\Pi_{\lambda_{800}/2}}$ on the space $\Pi_{\lambda_{800}/2}$.

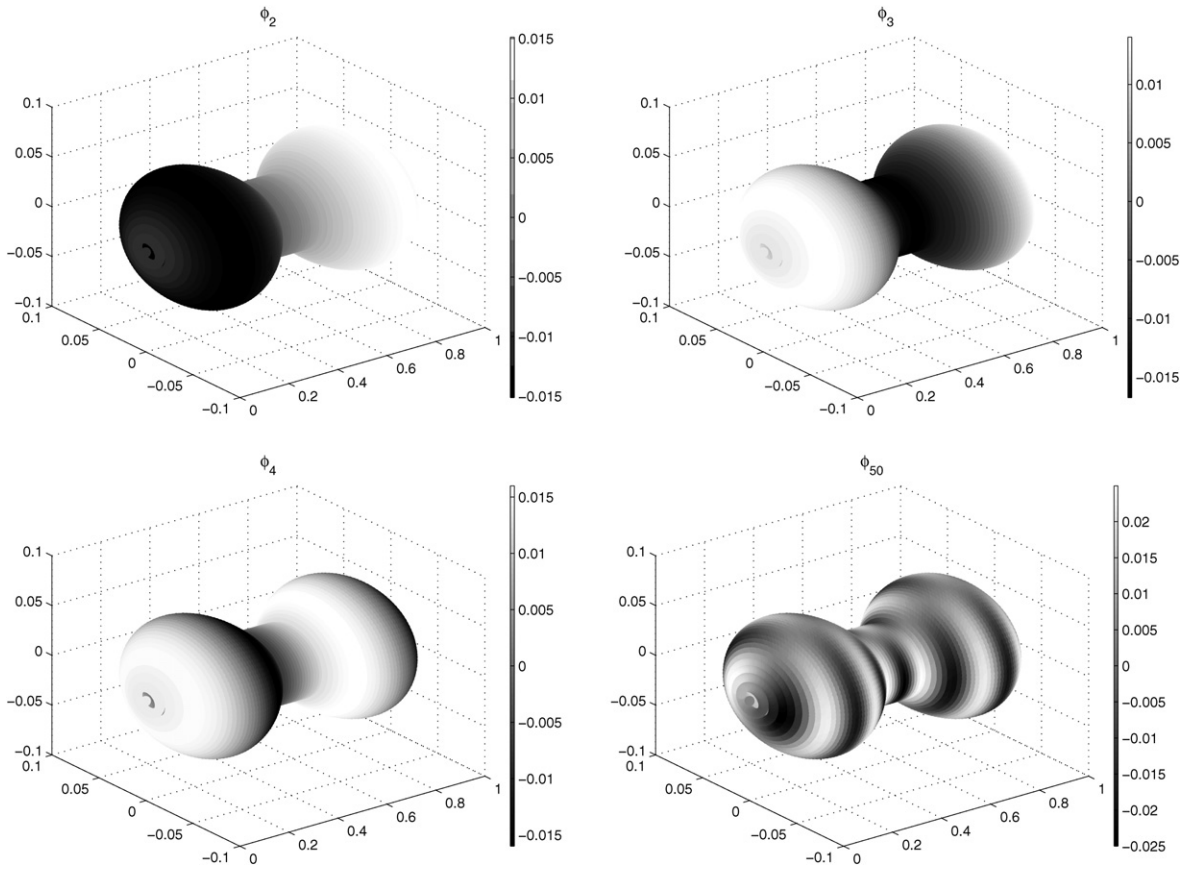

Fig. 1. Eigenfunctions $\varphi_2, \varphi_3, \varphi_4, \varphi_{50}$ of the approximated Laplace–Beltrami on a dumbell manifold sampled at 8000 points.

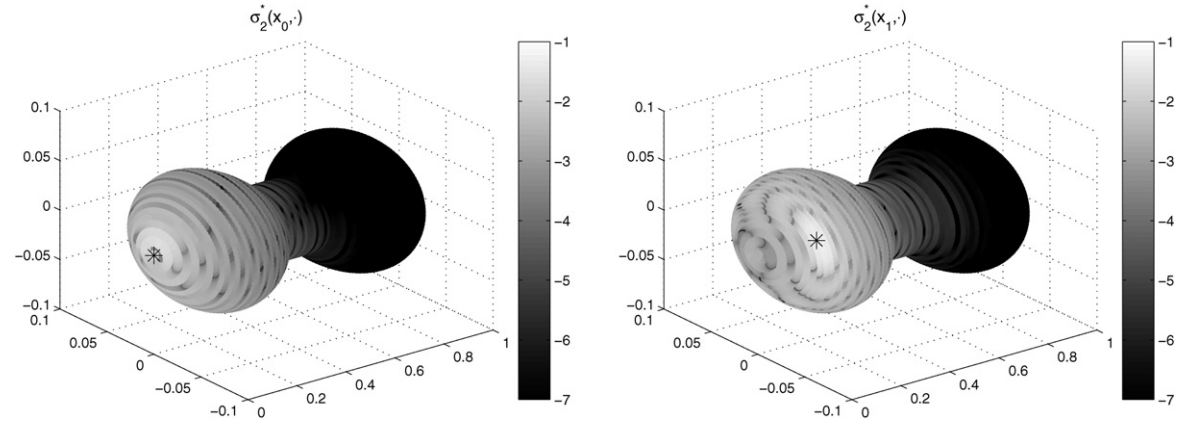

Fig. 2. The kernel $\sigma_\lambda(h_S)$ centered at two different points (logarithmic scale) shows its localization.

Since the spectral projection is the best approximator in L^2, the global L^2 error cannot be smaller than that obtained by the summability kernel. Indeed, we have

$$0.131 = \left\| f - P_{\Pi_{\lambda_{800/2}}}(f) \right\|_{L^2(X)} / \|f\|_{L^2(X)} < \left\| f - \sigma^*_{800,12}(f) \right\|_{L^2(X)} / \|f\|_{L^2(X)} = 0.151.$$

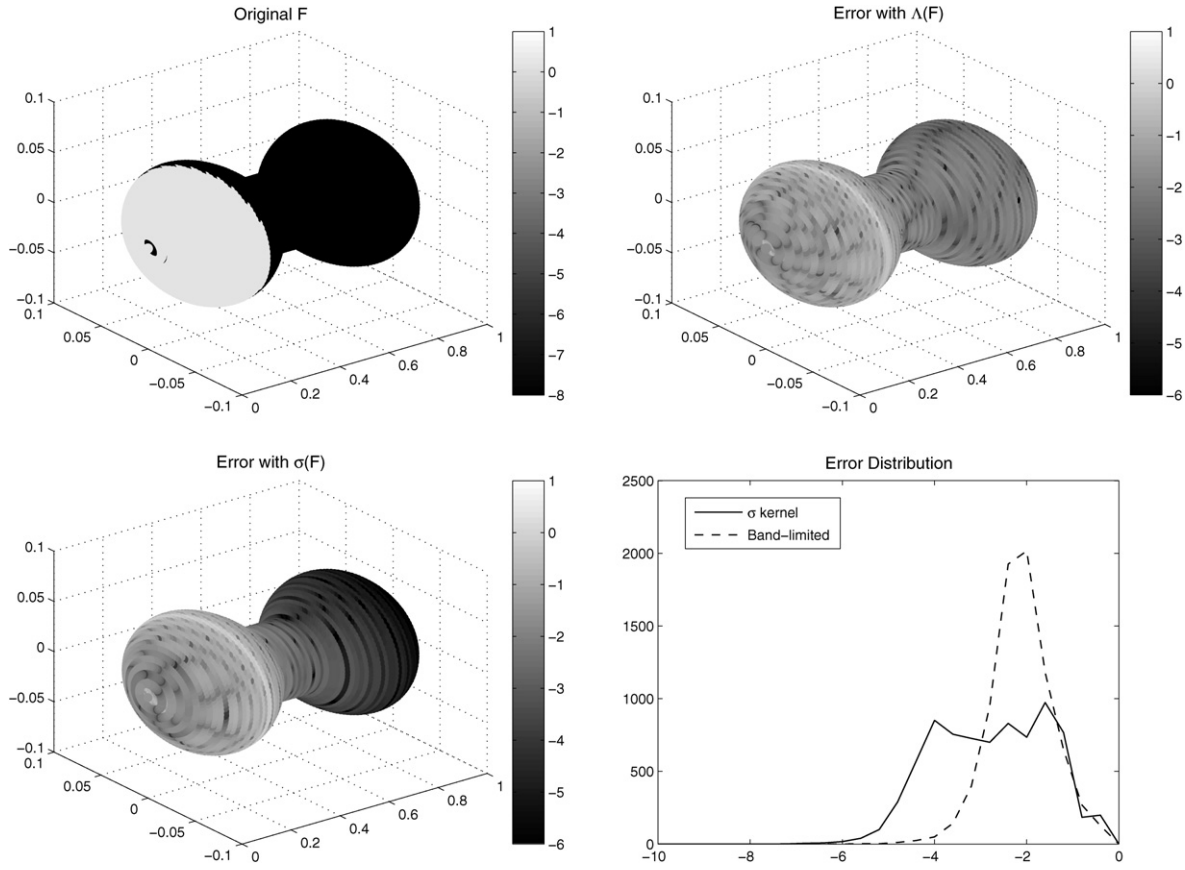

Fig. 3. From left to right: Function F to be approximated, approximation by projecting onto the 800 lowest-frequency eigenfunctions, approximation with the σ-kernel, histogram of errors (x-axis is in logarithmic scale).

However, since f is piecewise smooth and the operator $\sigma^*_{800,12}$ is localized (cf. Theorem 4.1), we expect a larger number of points with smaller errors compared with the corresponding distribution of errors obtained by using the band-limited projection. This is verified by the histogram in Fig. 3.

5.4. Hand-written digit recognition

In this subsection, we describe our experiments concerning the recognition of hand-written digits, based on a data set called MNIST. This data set is standard in the machine learning community, and is publicly available, together with detailed information about the performance of several algorithms, at http://yann.lecun.com/exdb/mnist/. It consists of 60,000 grayscale images of size 28×28 pixels, digitized from hand-written digits $0, \ldots, 9$. An additional set of 10,000 digits is often used for testing any algorithm trained on the previous set (see Fig. 4). Each data point is appropriately labeled. In this supervised setting, an algorithm is trained on the training set and the corresponding labels, and predicts the labels of the test set. Its performance is measured in terms of number of correctly guessed labels. State-of-the-art algorithms achieve errors around 0.4%. These algorithms are often very much tuned to the specific data set in the sense that they take advantage of the knowledge that each data point is an image. Naturally, they use all the training set available, and often even enlarge it by adding more samples obtained by perturbing the samples in the given training set.

In the semi-supervised learning setting, an algorithm is allowed in the training phase to look at all the data points, both training and testing, but it can access only the labels of a *small* percentage of points (e.g. 1%—compared to about 85% in the supervised setting mentioned above). The goal is to predict the labels of all the other data points, as correctly as possible. This framework models situations when collecting data is relatively cheap, while labeling is an

Fig. 4. Examples of test images from the MNIST database.

expensive process. It has been shown that, for this and many other data sets, one can use the large amount of unlabeled data to design learning and approximation procedures that perform better than procedures that do not use the unlabeled data. In this sense, one may expect that smaller amount of unlabeled data makes the learning problem harder. That this is indeed the case for the data set at hand is shown in [1] (in particular, see Section 3.8.1 and Appendix A). For the ease of comparison with [1] (Appendix A, Fig. A.9), we will work with only 10,000 data points, of which 9000 are from the training set and 1000 are from the test set (simply as a precaution, since the sets are supposed to be i.i.d.).

At this point we remark, in passing, that one could work in a supervised learning framework, in which unseen points need to be classified. In this case one can start by observing that it is enough to extend the eigenfunctions used in approximating f to the new points. This can be done in at least two ways: by adding the new points to the graph, and updating the eigenfunctions (for example by feeding the existing eigenfunctions as "initial conditions" to the eigensolver), or by using the geometric harmonic extensions described in [12].

We start by building a graph associated with the set of images, viewed as a subset of $\mathbb{R}^{28\times 28}$. We begin by projecting this set onto its top 50 principal components. This reduces the dimensionality and has a smoothing effect on the images and on the image ensemble. In the remainder of this subsection, we will denote this projection by X. The vertices of the graph are the points in X, and edges will connect very similar images. More precisely, we will use the construction of the self-tuning Laplacian [44] to define the weight matrix W as described in Section 5.1. For each point x_i we consider the 9 nearest neighbors $\mathcal{N}(x_i)$, we let d_i be the distance between x_i and the 8th nearest neighbor, and define

$$W_{ij} := w(x_i, x_j) := \exp\left(-\frac{\|x_i - x_j\|^2_{\mathbb{R}^{50}}}{d_i d_j}\right), \quad x_j \in \mathcal{N}(x_i),$$

$W_{ij} = 0$ otherwise. The result of this computation is a symmetric matrix W of size $|X| \times |X|$, with at most $9|X|$ nonzero entries. Efficient computation of such entries is possible in $O(|X| \log |X|)$ operations by using all-nearest-neighbor searchers, see [9,43].

Having defined the graph, we now define the target function. In this section, let for $i = 0, \ldots, 9$, $x \in X$,

$$\chi_i(x) = \begin{cases} 1, & x \text{ is digit } i, \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

The target function is defined by

$$f(x) = \arg\max\{\chi_i(x):\ i = 0, 1, \ldots, 9\}, \quad x \in X. \tag{5.7}$$

It is clear that the value of $f(x)$ is the digit which x represents.

Our goal is to approximate f using its values on a training set $\tilde{X}$, typically with $|\tilde{X}| \ll |X|$. We will use two different algorithms, one based on bandlimited projections and one based on σ-kernels. To test the performance of each algorithm, we proceed as follows. We select a random training set $\tilde{X}$ of a certain fixed size. We split randomly $\tilde{X}$ further into $\tilde{X}_t$ and $\tilde{X}_v$, with $|\tilde{X}_v| = 0.05|\tilde{X}_t|$. We call $\tilde{X}_v$ the validation set, and will be used to optimize the parameters required by the algorithm. In order to measure the error in our approximation to f, we first construct approximations $\tilde{\chi}_i$ to each χ_i, $i = 0, 1, \ldots, 9$, by using the values of the χ_is on $\tilde{X}_t$ only. We define the approximation to f by

$$\tilde{f}(x) = \arg\max\{\tilde{\chi}_i(x):\ i = 0, 1, \ldots, 9\}, \quad x \in X. \tag{5.8}$$

The error in approximation (or training error) on the validation set is naturally defined by $E_v = \#\{x \in \tilde{X}_v:\ \tilde{f}(x) \neq f(x)\}$. We repeat this procedure several times, for different random choices of $\tilde{X}$, and record the average value of E_v over the different runs. As our approximants, we will use linear combinations of the first J eigenfunctions of the graph Laplacian defined by (5.1), for increasing values of J. The optimal value of J for any algorithm will be the one that gives the minimal average validation error E_v. The error in approximation (or the testing error) is then defined by

$$E_T = \#\{x \in X \setminus \tilde{X}:\ \tilde{f}(x) \neq f(x)\}. \tag{5.9}$$

The objective of our experiments to be described in this section is to compare the least-squared approximation to f in the subspace spanned by the first J eigenfunctions with the approximation given by the summability operators $\sigma^*_{J,S}$. For any test data set $\tilde{X}$, the restrictions of ϕ_j to $\tilde{X}$ are not orthonormalized any more in the sense of the inner product

$$\langle g_1 g_2 \rangle_{\tilde{X}} = \sum_{x\in\tilde{X}} g_1(x)g_2(x), \quad g_1, g_2 : \tilde{X} \to \mathbb{R}.$$

However, we may consider the Gram matrix G defined by $G_{\ell,k} = \langle \phi_\ell \phi_k \rangle_{\tilde{X}}$. The highest number of eigenfunctions we may consider is limited by the size of G for which G is positive definite. Assuming that G is positive definite, and L is a lower triangular matrix such that $L^T L = G^{-1}$, then the functions $\tilde{\phi}_\ell = \sum_j L_{\ell,j}\phi_j$ are orthonormalized with respect to the inner product defined above. It is easy to verify that

$$\sum_{\ell=0}^{J} \tilde{\phi}_\ell(x)\tilde{\phi}_\ell(y) = \sum_{m,j=0}^{J} \left(G^{-1}\right)_{m,j}\phi_m(x)\phi_j(y), \quad x, y \in \tilde{X}.$$

Thus, the least-squared approximation to any χ_i is given by

$$P_{\Pi_{\lambda_J}}(\chi_i, x) = \sum_{m,j} (G^{-1})_{m,j}\langle \chi_i \phi_m \rangle \phi_j(x), \quad x \in X. \tag{5.10}$$

The approximation using the summability operator $\sigma^*_{J,S}(\chi_i)$ is defined by

$$\sigma^*_{J,S}(\chi_i, x) = \sum_{\ell} h_S(\lambda_\ell/\lambda_J)\langle \chi_i \tilde{\phi}_\ell \rangle \tilde{\phi}_\ell(x) = \sum_{m,j} \left\langle \chi_i \left(L^T H L\right)_{m,j}\phi_m \right\rangle \phi_j(x), \quad x \in X, \tag{5.11}$$

where H is the diagonal matrix with $H_{\ell,\ell} = h_S(\lambda_\ell/\lambda_J)$.

We test over several choices of J (ranging from 20 to 500), over different sizes of the training set (ranging from 1% to 30%), different values of S, and each of these tests is performed over 20 random choices of the training set, each such test being called a cross-validation (CV) test. We summarize part of the results in Figs. 5 and 6, and refer the reader to [1] for further details and experiments with the projection onto eigenfunctions. Code for reproducing such results, as well as the pre-computed results, is available at http://www.math.duke.edu/~mauro. In general we observe that the σ-kernel approximation out-performs the bandlimited approximation in terms of classification error, especially for small training sets, and the observed differences are significant (since they hold under cross-validation). Based on our trials for $S = 2, \ldots, 12$, the use of the σ-kernel with different parameters S does not affect significantly the minimal error E_T over all choices of J.

The test error initially decreases as a function of the number of eigenfunctions used, and then starts to increase. As more eigenfunctions are used, the problem of fitting a function to the given values becomes more and more ill-posed, in the sense that more and more functions with the given band can be fit, with a given error, to the training values. This can also be interpreted as a sampling issue, related to the quadrature formulas: higher eigenfunctions require higher sampling rate in order to determine their coefficients. The density of the training samples upper bounds the maximum frequency of the eigenfunctions whose coefficients can be determined. However, this transition does not happen abruptly, but continuously, and can be measured by looking at the condition number of the quadrature formulas (for example, as measured by the condition number of the Cholesky factors above), which degrades as the number of eigenfunctions increases. Finally, from a statistical perspective, this phenomenon can be interpreted as overfitting: it is natural to define the number of degrees of freedom of the model space to be J for the bandlimited projection and $\sum_{m=0}^{J} h_S(\lambda_m/\lambda_J)$ for the $\sigma^*_{J,S}$-method. The optimal J can then be understood in terms of bias-variance tradeoff: higher J corresponds to lower bias (more degrees of freedom, more flexible model) and higher variance.

5.5. Computational considerations

Let $N = |X|$. It is convenient to break up the computational cost as follows:

(i) Construction of the graph and corresponding Laplacian. This operation requires finding the k-nearest neighbors (or the neighbors within a certain distance ε) of any point in the set. A naive approach would require $O(N^2)$

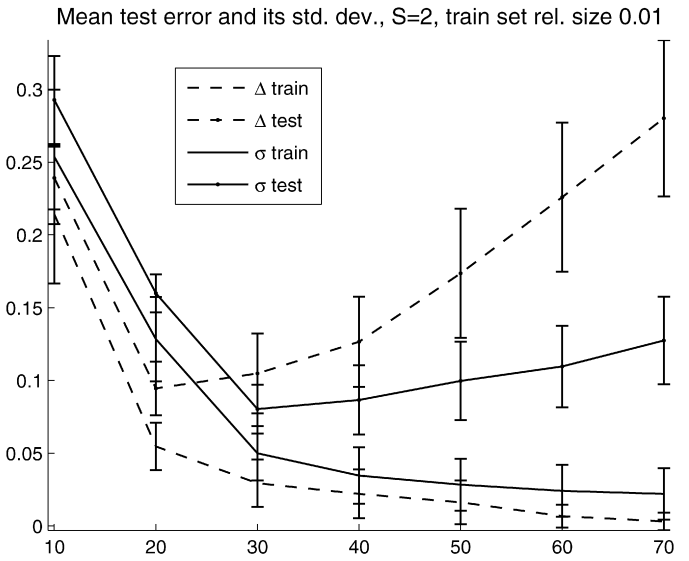

Fig. 5. Training and test errors on the MNIST data set with bandlimited projection and $\sigma^*_{J,S}$-kernel, as a function of the number of eigenvectors (J, on the horizontal axis): the error on the training set is very similar, but the $\sigma^*_{J,2}$-kernel provides better generalization error on the test set.

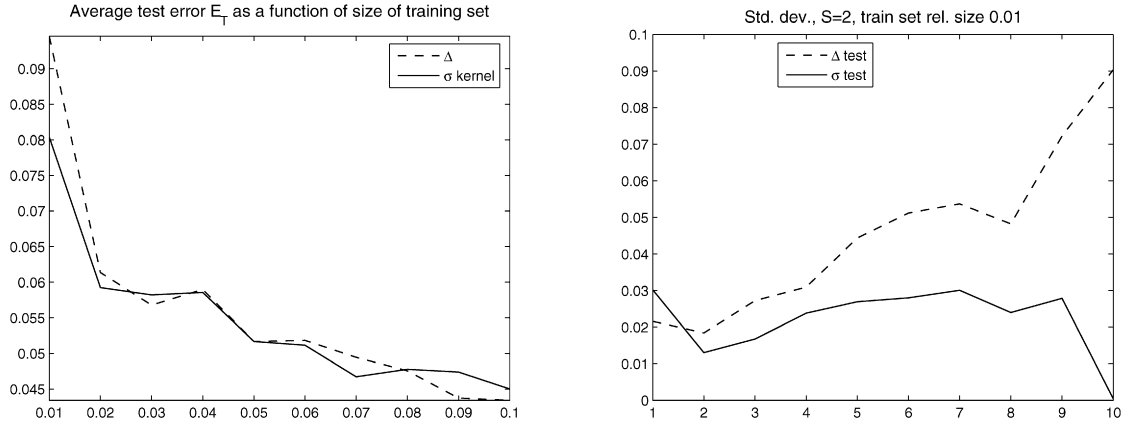

Fig. 6. Left: Minimum error (averaged over CV runs) over all choices of number of eigenvectors to use, for bandlimited projection vs $\sigma^*_{J,2}$-kernel, as a function of the size of the training set. The latter consistently outperforms, especially for small training sets, when it leads to more then 10% improvement in the test error rate. Right: Standard deviations (estimated over CV runs) of the prediction error on the test set, as a function of number of eigenfunctions used.

operations; algorithms based on deterministic or randomized multiscale partitions of the data are available [9, 43], and their cost is $O(N \operatorname{polylog}(N))$. Unfortunately, the constant is often large, being polynomial or even exponential in the dimension of the ambient space, depending on the algorithm chosen and the tradeoff between operation count and storage. The output of this can be regarded as a $N \times N$ sparse matrix W, whose (i, j)-entry is 1 if and only if x_i is a nearest-neighbor of x_j. Each entry is accessible in time $O(1)$. W may be weighted, for example by distance, and this operation is of order $O(N)$. Typically W is also symmetrized (another $O(N \log(N))$ operation). The Laplacian matrix $L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$, where D is the diagonal matrix of row-sums of W, can be constructed from W in time $O(N)$.

(ii) Computation of the M top eigenvalues and eigenvectors of $I - L$. This computation takes $O(M^2N)$, or even $O(MN^2)$ in practice, with symmetric sparse eigensolvers. It is the most expensive part of the algorithm, both theoretically and practically.

(iii) The computation of the σ-kernel applied to a function requires $O(MN)$ operations.

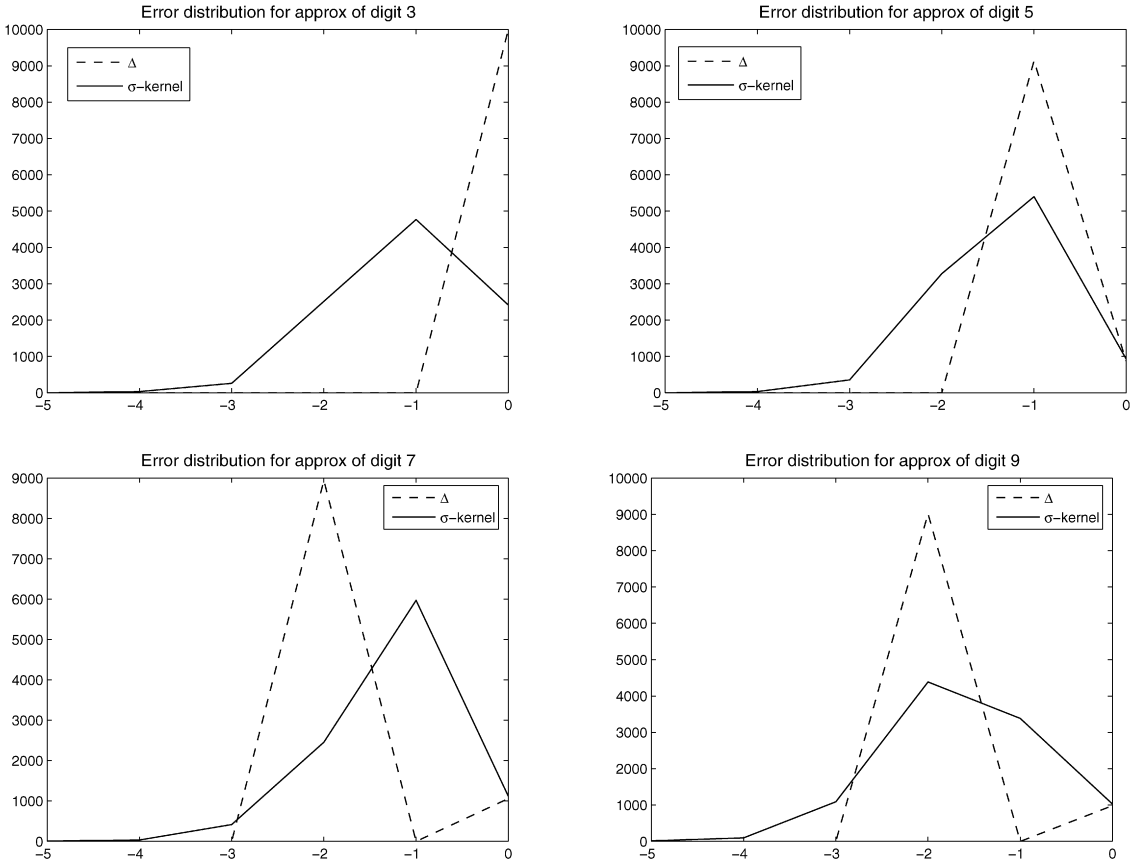

Fig. 7. Distribution of the error (in $\log_{10}$ scale on the horizontal axis) of approximation for the digits $3, 5, 7, 9$, with bandlimited projection and $\sigma^*_{J,2}$-kernel. The training set size was 100. We conjecture the different behavior in approximation observed for different classes corresponds to the complexity of the boundary of the class.

6. Proofs

Proof of Proposition 2.1. If (2.6) holds, then a simple application of Fubini's theorem shows that the part (b) holds. Let (b) hold, $f \in L^\infty$. In view of the Hahn–Banach theorem and the definition of $\sigma_\lambda(h)$, we see that

$$\begin{aligned}\big\|\sigma_\lambda(h, f)\big\|_\infty &= \sup_{\|g\|_1 \leqslant 1} \int \sigma_\lambda(h, f, x) g(x)\, \mathrm{d}\mu(x) = \sup_{\|g\|_1 \leqslant 1} \int f(x) \sigma_\lambda(h, g, x)\, \mathrm{d}\mu(x) \\ &\leqslant \|f\|_\infty \sup_{\|g\|_1 \leqslant 1} \big\|\sigma_\lambda(h, g)\big\|_1 \leqslant c\|f\|_\infty.\end{aligned}$$

Thus, (b) implies (c) for $p = 1, \infty$. An application of the Riesz–Thorin interpolation theorem [6, Theorem 1.1.1] yields (c) in the intermediate cases $1 < p < \infty$.

If $x \in \mathbb{X}$, and $f_{x,\lambda}(y) = \operatorname{sgn} \Phi_\lambda(h, x, y)$, $y \in \mathbb{X}$, then part (c) with $p = \infty$ implies that

$$\int \big|\Phi_\lambda(h, x, y)\big|\, \mathrm{d}\mu(y) = \big\|\sigma_\lambda(h, f_{x,\lambda})\big\|_\infty \leqslant c\|f_{x,\lambda}\|_\infty = c,$$

where the constant c is independent of λ and x. This implies (2.6). □

Proof of Theorem 2.1. Let $f \in L^2$, and $\lambda \geqslant 1$. We note that

$$\big\|f - \sigma_\lambda(h, f)\big\|_2^2 = \sum_{j=0}^{\infty} \big(1 - h(\lambda_j/\lambda)\big)^2 \big(\hat{f}(j)\big)^2.$$

Since $\sum(\hat{f}(j))^2 < \infty$, and $h(\lambda_j/\lambda) \to 1$ as $\lambda \to \infty$, we conclude that $\sigma_\lambda(h, f) \to f$ in L^2 as $\lambda \to \infty$. The first equation in (2.12) is now clear from the definitions. Since $h(t) = 1$ for $t \in [0, 1/2]$ and $h(t) = 0$ for $t \geqslant 1$, for any integer $N \geqslant 1$,

$$\begin{aligned}\sum_{n=0}^{N} \int \tau_n^*(h, f, y)\Psi_n^*(h; \circ, y)\,\mathrm{d}\mu(y) &= \sum_{\lambda_j<1} \hat{f}(j)\phi_j + \sum_{\lambda_j\geqslant 1}\sum_{n=1}^{N}\big(h\big(\lambda_j/2^n\big) - h\big(\lambda_j/2^{n-1}\big)\big)\hat{f}(j)\phi_j \\ &= \sum_{\lambda_j<1} \hat{f}(j)\phi_j + \sum_{\lambda_j\geqslant 1} h\big(\lambda_j/2^N\big)\hat{f}(j)\phi_j \\ &= \sigma_{2^N}(h, f). \end{aligned} \tag{6.1}$$

This implies the second equation in (2.12).

A straightforward computation using the Parseval identity leads to (2.13). Since $h : [0, \infty) \to [0, 1]$, we have for $n \geqslant 1$,

$$\begin{aligned}\big\|\tau_n(h, f)\big\|_2^2 &= \sum_{\lambda_j\geqslant 1} \big(h\big(\lambda_j/2^n\big) - h\big(\lambda_j/2^{n-1}\big)\big)^2\big(\hat{f}(j)\big)^2 \\ &\leqslant \sum_{\lambda_j\geqslant 1} \big(h\big(\lambda_j/2^n\big) - h\big(\lambda_j/2^{n-1}\big)\big)\big(\hat{f}(j)\big)^2 = \big\|\tau_n^*(h, f)\big\|_2^2. \end{aligned} \tag{6.2}$$

Also,

$$\big\|\tau_0(h, f)\big\|_2^2 = \sum_{j=0}^{\infty} h(\lambda_j)^2\big(\hat{f}(j)\big)^2 = \sum_{\lambda_j<1} h(\lambda_j)^2\big(\hat{f}(j)\big)^2 \leqslant \sum_{\lambda_j<1} \big(\hat{f}(j)\big)^2 = \big\|\tau_0^*(h, f)\big\|_2^2. \tag{6.3}$$

The first estimate in (2.14) follows from (6.2), (6.3), and (2.13).

In this proof only, let $g_{j,n} = h(\lambda_j/2^n) - h(\lambda_j/2^{n-1})$, $n \geqslant 1$. Then $g_{j,n} \neq 0$ only when $2^{n-2} < \lambda_j < 2^n$. Therefore, $g_{j,n}g_{j,m} \neq 0$ only when $|n - m| \leqslant 2$. Writing $g_{j,0} = h(\lambda_j)$, we see that $g_{j,0}g_{j,m} \neq 0$ only when $m \leqslant 1$. In the following estimate, we write $\tau_m(h, f) := 0$ if $m < 0$. Using Parseval identity, it follows that

$$\int \tau_n(h, f)\tau_m(h, f)\,\mathrm{d}\mu = 0, \quad |n - m| > 2.$$

Hence, using the first equation in (2.12) and Schwarz inequality, we conclude that

$$\begin{aligned}\|f\|_2^2 &= \sum_{n,m=0}^{\infty} \int \tau_n(h, f)\tau_m(h, f)\,\mathrm{d}\mu = \sum_{\ell=-2}^{2}\sum_{n=0}^{\infty}\int \tau_n(h, f)\tau_{n-\ell}(h, f)\,\mathrm{d}\mu \\ &\leqslant \sum_{\ell=-2}^{2}\sum_{n=0}^{\infty}\big\|\tau_n(h, f)\big\|_2\big\|\tau_{n-\ell}(h, f)\big\|_2 \\ &\leqslant \left(\sum_{n=0}^{\infty}\big\|\tau_n(h, f)\big\|_2^2\right)^{1/2}\sum_{\ell=-2}^{2}\left(\sum_{n=0}^{\infty}\big\|\tau_{n-\ell}(h, f)\big\|_2^2\right)^{1/2} \\ &\leqslant 5\sum_{n=0}^{\infty}\big\|\tau_n(h, f)\big\|_2^2. \end{aligned}$$

This proves the second estimate in (2.14).

Next, let $h \in \mathcal{M}$, $1 \leqslant p \leqslant \infty$, and $f \in L^p$, and $\lambda > 0$. It is easy to verify that (6.1) continues to hold. Since $h(t) = 1$ for $t \in [0, 1/2]$, we see that $\sigma_\lambda(h, P) = P$ for all $P \in \Pi_{\lambda/2}$. Consequently, Proposition 2.1(c) implies that for every $P \in \Pi_{\lambda/2}$,

$$\big\|f - \sigma_\lambda(h, f)\big\|_p = \big\|f - P - \sigma_\lambda(h, f - P)\big\|_p \leqslant \|f - P\|_p + \big\|\sigma_\lambda(h, f - P)\big\|_p \leqslant c\|f - P\|_p.$$

Therefore, taking into account the fact that $\sigma_\lambda(h, f) \in \Pi_\lambda$, we obtain (2.15) by taking the infimum over $P \in \Pi_{\lambda/2}$. If $f \in X^p$, $E_{2^{N-1},p}(f) \to 0$ as $N \to \infty$, and (6.1) implies that (2.12) holds with convergence in the sense of L^p. □

We are now in a position to prove the parts (a) and (b) in Theorem 3.1.

Proof of the parts (a), (b) in Theorem 3.1. Let $h \in \mathcal{M}$. Since $\tau_n(h, f) = \sigma_{2^n}(h, f) - \sigma_{2^{n-1}}(h, f)$ for $n \geqslant 1$, (2.15) implies that

$$\left\|\tau_n(h, f)\right\|_p \leqslant cE_{2^{n-2},p}(f), \quad n \geqslant 3.$$

Hence, if $f \in B^a_{p,\rho}$, then $\{\|\tau_n(h, f)\|_p\} \in \mathsf{b}_{\rho,a}$. It follows from the first equation in (2.12) that for integer $N \geqslant 3$,

$$E_{2^N,p}(f) \leqslant \sum_{n=N}^{\infty} \left\|\tau_n(h, f)\right\|_p.$$

If $\{\|\tau_n(h, f)\|_p\} \in \mathsf{b}_{\rho,a}$, then the discrete Hardy inequality [16, Lemma 3.4, p. 27] now implies that $f \in B^a_{p,\rho}$. This proves part (a).

In this proof only, let $g(t) := \sqrt{h(t) - h(2t)}$, $t \geqslant 0$. Then $\tau^*_n(h, f) = \sigma_{2^n}(g, f)$ for $n \geqslant 1$. Hence, if $g \in \mathcal{M}$ then Proposition 2.1(c) implies that $\|\tau^*_n(h, f)\|_p \leqslant c\|f\|_p$. Moreover, the fact that $g(t) = 0$ for $0 \leqslant t \leqslant 1/4$ implies that $\tau^*_n(h, P) = 0$ for $P \in \Pi_{2^{n-2}}$, $n \geqslant 3$. So, for $n \geqslant 3$ and any $P \in \Pi_{2^{n-2}}$,

$$\left\|\tau^*_n(h, f)\right\|_p = \left\|\tau^*_n(h, f - P)\right\|_p \leqslant c\|f - P\|_p.$$

Thus, taking infimum over $P \in \Pi_{2^{n-2}}$,

$$\left\|\tau^*_n(h, f)\right\|_p \leqslant cE_{2^{n-2},p}(f), \quad n \geqslant 3.$$

Hence, if $f \in B^a_{p,\rho}$, then $\{\|\tau^*_n(h, f)\|_p\} \in \mathsf{b}_{\rho,a}$.

To prove the converse, we observe that

$$\left\|\tau_n(h, f)\right\|_p = \left\|\tau^*_n\big(h, \tau^*_n(h, f)\big)\right\|_p \leqslant c\left\|\tau^*_n(h, f)\right\|_p.$$

Therefore, if $\{\|\tau^*_n(h, f)\|_p\} \in \mathsf{b}_{\rho,a}$ then $\{\|\tau_n(h, f)\|_p\} \in \mathsf{b}_{\rho,a}$. Since $h \in \mathcal{M}$, the part (a) of this theorem implies that $f \in B^a_{p,\rho}$. □

In order to prove the part (c) Theorem 3.1, we recall the following theorem [16, Theorem 9.1, also Chapter 6.7].

Theorem 6.1. *Let $1 \leqslant p \leqslant \infty$, $a, \rho > 0$. Suppose that for every integer $r \geqslant 1$,*

$$E_{n,p}(g) \leqslant cn^{-r}\left\|(\Delta^*)^r g\right\|_p, \quad n = 1, 2, \ldots, \ g \in W^p_r, \tag{6.4}$$

and

$$\left\|(\Delta^*)^r P\right\|_p \leqslant cn^r\|P\|_p, \quad P \in \Pi_n, \ n = 1, 2, \ldots. \tag{6.5}$$

Then $f \in B^a_{p,\rho}$ if and only if for every $r > a$, $K_r(f, 2^{-n}) \in \mathsf{b}_{\rho,a}$.

Thus, we need to prove (6.4) and (6.5). Since we find the estimates of independent interest, we prove them as separate theorems, Theorems 6.2 and 6.3 below. First, we prove a proposition describing the equivalence between the Riesz condition with exponent S and the inclusion $BV^S_0 \subset \mathcal{M}$ [42].

Proposition 6.1. *Let $S \geqslant 1$ be an integer. The Riesz condition with S is satisfied if and only if $BV^S_0 \subset \mathcal{M}$; i.e., for every $h \in BV^S_0$, $1 \leqslant p \leqslant \infty$, $f \in L^p$,*

$$\left\|\sigma_\lambda(h, f)\right\|_p \leqslant c\|f\|_p.$$

Proof. Let the Riesz condition be satisfied with exponent S, $h \in BV^S_0$, and $h(t) = 0$ for $t \geqslant c(h)$, and $f \in L^1$. Then

$$\sigma_\lambda(h,f)=\sum_{j=0}^{\infty}h(\lambda_j/\lambda)\hat{f}(j)\phi_j=\frac{(-1)^S}{S!}\int_0^\infty\left\{\sum_{\lambda_j\leqslant c(h)\lambda}(t-\lambda_j/\lambda)_+^S\hat{f}(j)\phi_j\right\}\mathrm{d}h^{(S)}(t)$$

$$=\frac{(-1)^S}{S!}\int_0^\infty t^S\mathcal{R}_{\lambda t,S}(f)\,\mathrm{d}h^{(S)}(t). \tag{6.6}$$

Recalling that $h^{(S)}$ has a compact support, we conclude using the Minkowski inequality that $\|\sigma_\lambda(h,f)\|_1\leqslant c\|f\|_1$. In view of Proposition 2.1, we have proved that $h\in\mathcal{M}$. The converse statement is obvious, since $\mathcal{R}_{\lambda,S}(f)=\sigma_\lambda(h,f)$ for a special $h\in BV_0^S$. □

Theorem 6.2. *Let $S\geqslant 1$ be an integer, the Riesz condition be satisfied with exponent S, $1\leqslant p\leqslant\infty$, $f\in X^p$, $r\geqslant 1$ be an integer, and $(\Delta^*)^r f\in X^p$. Then for $n\geqslant 1$,*

$$E_{n,p}(f)\leqslant cn^{-r}\left\|(\Delta^*)^r f\right\|_p. \tag{6.7}$$

Proof. In this proof only, let $h=h_S$, where h_S is defined in (5.4), $g(t):=\frac{h(t)-h(2t)}{t^r}$, $t>0$. Then for integer $m\geqslant 1$,

$$\tau_m(h,f)=\sum_{j=0}^{\infty}\big(h\big(\lambda_j/2^m\big)-h\big(\lambda_j/2^{m-1}\big)\big)\hat{f}(j)\phi_j$$

$$=2^{-mr}\sum_{j=0}^{\infty}\frac{h(\lambda_j/2^m)-h(\lambda_j/2^{m-1})}{(\lambda_j/2^m)^r}\lambda_j^r\hat{f}(j)\phi_j=2^{-mr}\sigma_{2^m}\big(g,(\Delta^*)^r f\big).$$

Since $h(t)-h(2t)=0$ for all $t\in[0,1/2]$, $g\in BV_0^S$, and Proposition 6.1 implies that for integer $m\geqslant 3$,

$$\left\|\tau_m(h,f)\right\|_p=2^{-mr}\left\|\sigma_{2^m}\big(g,(\Delta^*)^r f\big)\right\|_p\leqslant c2^{-mr}\left\|(\Delta^*)^r f\right\|_p.$$

Now, let $\nu\geqslant 0$ be the largest integer such that $2^\nu\leqslant n$. Then (2.12) implies that

$$E_{n,p}(f)\leqslant E_{2^\nu,p}(f)\leqslant\sum_{m=\nu+1}^{\infty}\left\|\tau_m(h,f)\right\|_p\leqslant c\left\|(\Delta^*)^r f\right\|_p\sum_{m=\nu+1}^{\infty}2^{-mr}$$

$$\leqslant c2^{-\nu r}\left\|(\Delta^*)^r f\right\|_p\leqslant cn^{-r}\left\|(\Delta^*)^r f\right\|_p. \qquad \square$$

Theorem 6.3. *Let $S\geqslant 1$ be an integer, the Riesz condition be satisfied with exponent S, $1\leqslant p\leqslant\infty$, $n\geqslant 0$, $r\geqslant 1$ be an integer, and $P\in\Pi_n$. Then*

$$\left\|(\Delta^*)^r P\right\|_p\leqslant cn^r\|P\|_p. \tag{6.8}$$

Proof. In this proof only, let $h=h_S$, where h_S is defined in (5.4). In this proof only, let $g(t):=t^r h(t)$. Then

$$(\Delta^*)^r P=\sum_j h\big(\lambda_j/(2n)\big)\lambda_j^r\hat{P}(j)\phi_j=(2n)^r\sum_j g\big(\lambda_j/(2n)\big)\hat{P}(j)\phi_j=(2n)^r\sigma_{2n}(g,P).$$

We note that $g\in BV_0^S$ as well. Since the Riesz condition is satisfied with exponent S, we may apply Proposition 6.1 to conclude that

$$\left\|(\Delta^*)^r P\right\|_p=(2n)^r\left\|\sigma_{2n}(g,P)\right\|_p\leqslant cn^r\|P\|_p. \qquad \square$$

Proof of part (c) in Theorem 3.1. Theorem 3.1(c) follows from Theorems 6.2, 6.3, and 6.1. □

Proof of Proposition 3.1. Let $g\in W_r^p$ be chosen so that

$$\|f-g\|_p+\delta^r\left\|(\Delta^*)^r g\right\|_p\leqslant 2K_r(f,\delta). \tag{6.9}$$

In view of (2.15) and (6.7), we have with $\lambda = \delta^{-1}$,

$$\begin{aligned}\big\|f - \sigma_\lambda(h, f)\big\|_p &\leqslant \big\|f - g + \sigma_\lambda(f - g)\big\|_p + \big\|g - \sigma_\lambda(g)\big\|_p \\ &\leqslant c\big\{\|f - g\|_p + \delta^r \big\|(\Delta^*)^r g\big\|_p\big\} \leqslant cK_r(f, \delta). \end{aligned} \tag{6.10}$$

Taking into account the fact that $(\Delta^*)^r \sigma_\lambda(h, g) = \sigma_\lambda(h, (\Delta^*)^r g)$, we obtain from (6.8) that

$$\begin{aligned}\delta^r \big\|(\Delta^*)^r \sigma_\lambda(h, f)\big\|_p &\leqslant \delta^r \big\{\big\|(\Delta^*)^r \sigma_\lambda(h, f - g)\big\|_p + \big\|\sigma_\lambda\big(h, (\Delta^*)^r g\big)\big\|_p \\ &\leqslant \big\|\sigma_\lambda(h, f - g)\big\|_p + \delta^r \big\|\sigma_\lambda\big(h, (\Delta^*)^r g\big)\big\|_p \leqslant c\big\{\|f - g\|_p + \delta^r \big\|(\Delta^*)^r g\big\|_p\big\} \leqslant cK_r(f, \delta).\end{aligned}$$

Together with (6.10), this yields (3.3). □

We now turn to the proofs of the theorems in Section 4.

In the sequel, we find it convenient to abuse the notation and denote the Fourier transform of a function $\phi : \mathbb{R} \to \mathbb{R}$ by $\hat{\phi}$. It will be clear from the context whether the Fourier transform or the coefficients in the orthogonal expansion are intended. Let $V : \mathbb{R} \to \mathbb{R}$ be chosen such that V is an even function, $\hat{V}$ is an infinitely differentiable function, $\hat{V}(\omega) = 1$ if $|\omega| \leqslant 1/2$, and $\hat{V}(\omega) = 0$ if $|\omega| \geqslant 1$. Then for every $Y > 0$, there exists $H_Y : \mathbb{R} \to \mathbb{R}$ such that

$$\hat{H}_Y(\omega) = \hat{H}(\omega)\hat{V}(\omega/Y), \quad \omega \in \mathbb{R}. \tag{6.11}$$

In the "time domain," one has

$$H_Y(t) = Y \int_{\mathbb{R}} H(u) V\big(Y(t - u)\big)\,\mathrm{d}u. \tag{6.12}$$

We recall from [35, Section 5.2.2, Eq. (4), Section 4.2, Eq. (15)] that

$$\max_{t\in\mathbb{R}} \big|H^{(m)}(t) - H_Y{}^{(m)}(t)\big| \leqslant \frac{c}{Y^{S-m}} \max_{t\in\mathbb{R}} \big|H^{(S)}(t)\big|, \quad m = 0, 1, \ldots, S. \tag{6.13}$$

In the remainder of this section, we will assume that H is a fixed function with

$$\big\||H|\big\|_S = \sum_{k=0}^{S} \max_{u\in\mathbb{R}} \big|H^{(k)}(u)\big| = 1.$$

The following lemma summarizes some of the simple technical details needed in the proof.

Lemma 6.1. (a) *If $G : \mathbb{R} \to \mathbb{R}$ is a bounded function, $\{a_j\}$ is a bounded sequence, then we have for any $C > 0$, $\Lambda \geqslant 1$,*

$$\left|\sum_{\ell_j \leqslant C\Lambda} G(\ell_j/\Lambda) a_j \phi_j(x)\phi_j(y)\right| \leqslant c(C\Lambda)^K \sup_{t\in[0,C]} \big|G(t)\big| \max_{\ell_j \leqslant C\Lambda} |a_j|, \quad x, y \in \mathbb{X}. \tag{6.14}$$

(b) *Let G be a continuous, integrable, even, real valued function on $\mathbb{R}$, vanishing at infinity, such that the Fourier transform $\hat{G}$ is also integrable. Then for every $\Lambda > 0$, $f_1, f_2 \in L^2$,*

$$\sum_j G(\ell_j/\Lambda)\hat{f}_1(j)\hat{f}_2(j) = \frac{\Lambda}{\pi} \int_0^\infty \hat{G}(\Lambda t) W(t, f_1, f_2)\,\mathrm{d}t. \tag{6.15}$$

(c) *For any $x \in \mathbb{X}$, $r > 0$, and a nonincreasing function $g_1 : [0, \infty) \to [0, \infty)$,*

$$\Lambda^K \int_{\Delta(x,r)} g_1\big(\Lambda d(x, y)\big)\,\mathrm{d}\mu(y) \leqslant c \int_{r\Lambda/2}^\infty g(v) v^{K-1}\,\mathrm{d}v. \tag{6.16}$$

Proof. An application of Schwarz inequality and the estimate (4.3) gives for any $x, y \in \mathbb{X}$,

$$\left|\sum_{\ell_j\leqslant C\Lambda} G(\ell_j/\Lambda)a_j\phi_j(y)\phi_j(x)\right| \leqslant \sup_{t\in[0,C]}\left|G(t)\right| \max_{\ell_j\leqslant C\Lambda}|a_j|\left\{\sum_{\ell_j\leqslant C\Lambda}\phi_j^2(x)\right\}^{1/2}\left\{\sum_{\ell_j\leqslant C\Lambda}\phi_j^2(y)\right\}^{1/2}$$
$$\leqslant c(C\Lambda)^K \sup_{t\in[0,C]}\left|G(t)\right| \max_{\ell_j\leqslant C\Lambda}|a_j|. \tag{6.17}$$

This proves part (a).

To prove part (b), we observe first that the Fourier inversion formula holds for G at each point in $\mathbb{R}$. Since G is even and real valued, so is $\hat{G}$. Therefore, the Fourier inversion formula implies that

$$G(u/\Lambda) = \frac{1}{\pi}\int_0^\infty \hat{G}(z)\cos(zu/\Lambda)\,\mathrm{d}z = \frac{\Lambda}{\pi}\int_0^\infty \hat{G}(\Lambda t)\cos(tu)\,\mathrm{d}t. \tag{6.18}$$

In view of (4.6), we may apply Fubini's theorem to conclude from (6.18) that (6.15) holds. This proves part (b).

In this proof only, we will write $\mathcal{A}(x,t) = \{y\in\mathbb{X}:\ t\leqslant d(x,y)\leqslant 2t\}$, and observe that in view of (4.4), $\mu(\mathcal{A}(x,t))\leqslant ct^K$ for $t>0$. Since g nonincreasing, we have

$$\int_{\Delta(x,r)} g_1\big(\Lambda d(x,y)\big)\,\mathrm{d}\mu(y) = \sum_{R=0}^\infty \int_{\mathcal{A}(x,2^Rr)} g_1\big(\Lambda d(x,y)\big)\,\mathrm{d}\mu(y)$$
$$\leqslant \sum_{R=0}^\infty g_1\big(2^Rr\Lambda\big)\mu\big(\mathcal{A}\big(x,2^Rr\big)\big) \leqslant c\sum_{R=0}^\infty g_1\big(2^Rr\Lambda\big)\big(2^Rr\big)^K$$
$$\leqslant cr^K\sum_{R=0}^\infty\int_{2^{R-1}}^{2^R} g_1(ur\Lambda)u^{K-1}\,\mathrm{d}u = cr^K\int_{1/2}^\infty g_1(ur\Lambda)u^{K-1}\,\mathrm{d}u$$
$$= c\Lambda^{-K}\int_{r\Lambda/2}^\infty g_1(v)v^{K-1}\,\mathrm{d}v. \qquad \square$$

The following lemma enables us to establish the fact that the operators $\tilde{\sigma}_L(H_Y)$ are well defined, as well as to obtain a useful estimate on the "tail part."

Lemma 6.2. *Let $\{a_j\}$ be a bounded sequence of complex numbers, and* (4.3) *hold. For $x, y\in\mathbb{X}$, Y, $\Lambda\geqslant 1/2$, $J\geqslant\Lambda$, and integer $N>K-1$, we have*

$$\left|\sum_{\ell_j\geqslant J} H_Y(\ell_j/\Lambda)a_j\phi_j(x)\phi_j(y)\right| \leqslant c(N)J^KY^{-N}(\Lambda/J)^{N+1}\max_j|a_j|. \tag{6.19}$$

In particular, the series $\sum_{j=0}^\infty H_Y(\ell_j/\Lambda)a_j\phi_j(x)\phi_j(y)$ converges uniformly on $\mathbb{X}\times\mathbb{X}$ to a continuous and bounded function, and

$$\left|\sum_{\ell_j\geqslant 2\Lambda} H_Y(\ell_j/\Lambda)a_j\phi_j(x)\phi_j(y)\right| \leqslant c(N)\Lambda^KY^{-N}\max_j|a_j|. \tag{6.20}$$

Proof. Without loss of generality, we may assume that $\max_j|a_j|\leqslant 1$. In this proof only, all constants may depend upon N, and let

$$s(u) := \sum_{\ell_j\leqslant u} a_j\phi_j(x)\phi_j(y), \quad u\geqslant 1.$$

In view of (4.3), Schwarz inequality shows that

$$\left|s(u)\right| \leqslant \sum_{\ell_j\leqslant u}\left|\phi_j(x)\phi_j(y)\right| \leqslant cu^K, \quad u\geqslant 1. \tag{6.21}$$

Since $\hat{V}$ is infinitely differentiable and supported on $[-1,1]$, we see that $|V(u)| \leqslant c(1+|u|)^{-N-3}$ and $|V'(u)| \leqslant c(1+|u|)^{-N-2}$ for all $u \in \mathbb{R}$. Further, recalling that $H(u) = 0$ if $|u| \geqslant 1$, (6.12) shows that for $|t| \geqslant 2$,

$$\big|H_Y(t)\big| \leqslant cY\big(1+Y|t|\big)^{-N-3}, \qquad \big|H_Y'(t)\big| \leqslant cY^2\big(1+Y|t|\big)^{-N-2}.$$

Therefore, if $Y \geqslant 1/2$, $K - N - 2 < -1$, we obtain

$$\begin{aligned}
\left|\sum_{\ell_j \geqslant J} H_Y(\ell_j/\Lambda)a_j\phi_j(x)\phi_j(y)\right| &= \left|\int_J^\infty H_Y(u/\Lambda)\,\mathrm{d}s(u)\right| \\
&= \left|-s(J)H_Y(J/\Lambda) + \frac{1}{\Lambda}\int_J^\infty H_Y'(u/\Lambda)s(u)\,\mathrm{d}u\right| \\
&\leqslant cJ^K Y(YJ/\Lambda)^{-N-3} + c\frac{Y^2}{\Lambda}\int_J^\infty \big|(Yu/\Lambda)^{-N-2}u^K\big|\,\mathrm{d}u \\
&\leqslant cJ^K Y(YJ/\Lambda)^{-N-3} + cY(Y/\Lambda)^{-N-1}J^{K-N-1}.
\end{aligned}$$

This leads to (6.19). The estimate (6.20) is obtained from (6.19) by letting $J = 2\Lambda$. Since the right-hand side of (6.19) tends to 0 as $J \to \infty$, and is independent of $x, y \in \mathbb{X}$, the uniform convergence of the series $\sum_{j=0}^\infty H_Y(\ell_j/\Lambda)a_j\phi_j(x)\phi_j(y)$ is clear. The fact that the resulting continuous function is bounded follows from (6.20) and (6.14) with H_Y in place of G, and $C = 2$. □

The next lemma describes the error in replacing $\tilde{\Phi}_\Lambda(H,x,y)$ by $\tilde{\Phi}_\Lambda(H_Y,x,y)$.

Lemma 6.3. *We assume* (4.3). *Let* $x, y \in \mathbb{X}$, $Y \geqslant 1/2$. *Let* $\{a_j\}$ *be a bounded sequence of complex numbers with* $\max_j |a_j| \leqslant 1$. *Then*

$$\left|\sum_{j=0}^\infty \big(H(\ell_j/\Lambda) - H_Y(\ell_j/\Lambda)\big)a_j\phi_j(x)\phi_j(y)\right| \leqslant c\Lambda^K Y^{-S}. \tag{6.22}$$

Proof. In this proof only, we will write $\hat{U}(t) = \hat{V}(t) - \hat{V}(2t)$, and define for $k = 1, 2, \ldots$, $G_{Y,k} := H_{Y2^k} - H_{Y2^{k-1}}$. We note that $\hat{G}_{Y,k}(t) = \hat{H}(t)\hat{U}(t/Y2^k)$. In view of (6.13), we have for any $Y > 0$,

$$H = H_Y + \sum_{k=1}^\infty G_{Y,k}, \tag{6.23}$$

where the series converges uniformly on $\mathbb{R}$. Moreover,

$$\big|G_{Y,k}(u)\big| \leqslant c\big(Y2^k\big)^{-S}, \quad u \in \mathbb{R},\ k = 0, 1, \ldots. \tag{6.24}$$

Let $k \geqslant 1$ be an integer. In view of (6.24) and (6.14),

$$\left|\sum_{\ell_j \leqslant 2\Lambda} G_{Y,k}(\ell_j/\Lambda)a_j\phi_j(x)\phi_j(y)\right| \leqslant cY^{-S}2^{-kS}\Lambda^K.$$

Hence,

$$\sum_{k=1}^\infty \left|\sum_{\ell_j \leqslant 2\Lambda} G_{Y,k}(\ell_j/\Lambda)a_j\phi_j(x)\phi_j(y)\right| \leqslant cY^{-S}\Lambda^K. \tag{6.25}$$

In view of (6.23) and (6.25), this yields

$$\left|\sum_{\ell_j \leqslant 2\Lambda} \big(H(\ell_j/\Lambda) - H_Y(\ell_j/\Lambda)\big)a_j\phi_j(x)\phi_j(y)\right| \leqslant c\Lambda^K Y^{-S}.$$

Since $H(\ell_j/\Lambda) = 0$ if $\ell_j > 2\Lambda$, (6.20) used with S in place of N gives

$$\left|\sum_{\ell_j>2\Lambda} \big(H(\ell_j/\Lambda) - H_Y(\ell_j/\Lambda)\big)a_j\phi_j(x)\phi_j(y)\right| \leqslant c\Lambda^K Y^{-S}. \qquad \square$$

Proof of Theorem 4.1. In this proof only, let $\alpha \geqslant 1$ be chosen so that

$$d(x, y) \leqslant \alpha\big(d(x, z) + d(z, y)\big), \quad x, y, z \in \mathbb{X}.$$

In light of (6.14), we may assume that $r := (d(x, y)/2\alpha^2) \geqslant 1/\Lambda$. Let $Y = \Lambda r$. Next, let $f, g \in L^1$ be supported on $B(y, d(x, y)/4\alpha(1+\alpha))$, $B(x, d(x, y)/4\alpha(1+\alpha))$, respectively, $\|f\|_1 = \|g\|_1 = 1$, and $1 > \varepsilon > 0$ be arbitrary. Then there exist continuous functions f_1, g_1 such that $\|f_1 - f\|_1 \leqslant \varepsilon$, $\|g_1 - g\|_1 \leqslant \varepsilon$. Multiplying g_1 by a continuous function having range in $[0, 1]$, equal to 1 on $B(x, d(x, y)/4\alpha(1+\alpha))$ and 0 outside $B(x, d(x, y)/4\alpha(1+\alpha))$, we obtain a function $g_2 \in L^1 \cap L^\infty$, such that $\|g - g_2\|_1 \leqslant 2\varepsilon$. Similarly, we obtain a continuous function $f_2 \in L^1 \cap L^\infty$, such that $\|f - f_2\|_1 \leqslant 2\varepsilon$, and f_2 is supported on $B(y, d(x, y)/2\alpha(1+\alpha))$. Now, (6.14) implies that

$$\begin{aligned}
&\left|\sum_j H(\ell_j/\Lambda)\hat{f}(j)\hat{g}(j) - \sum_j H(\ell_j/\Lambda)\hat{f}_2(j)\hat{g}_2(j)\right| \\
&\quad= \left|\sum_j H(\ell_j/\Lambda)\hat{f}(j)\hat{g}(j) - \sum_j H(\ell_j/\Lambda)\hat{f}(j)\hat{g}_2(j)\right| \\
&\qquad+ \left|\sum_j H(\ell_j/\Lambda)\hat{f}(j)\hat{g}_2(j) - \sum_j H(\ell_j/\Lambda)\hat{f}_2(j)\hat{g}_2(j)\right| \\
&\quad\leqslant c\Lambda^K\|g - g_2\|_1 + c\Lambda^K\|f - f_2\|_1 \leqslant c\varepsilon\Lambda^K. \qquad (6.26)
\end{aligned}$$

Next, using Lemma 6.3, we conclude that

$$\left|\sum_j H(\ell_j/\Lambda)\hat{f}_2(j)\hat{g}_2(j) - \sum_j H_Y(\ell_j/\Lambda)\hat{f}_2(j)\hat{g}_2(j)\right| \leqslant c\Lambda^K Y^{-S}\|f_2\|_1\|g_2\|_1 \leqslant c\Lambda^K Y^{-S}. \qquad (6.27)$$

In view of (6.15),

$$\sum_j H_Y(\ell_j/\Lambda)\hat{f}_2(j)\hat{g}_2(j) = \frac{\Lambda}{\pi}\int_0^\infty \hat{H}_Y(\Lambda t)W(t, f_2, g_2)\,\mathrm{d}t = \frac{\Lambda}{2\pi}\int_0^r \hat{H}_Y(\Lambda t)W(t, f_2, g_2)\,\mathrm{d}t.$$

Since $\operatorname{dist}(\operatorname{supp}(f_2), \operatorname{supp}(g_2)) \geqslant (1/2\alpha^2)d(x, y)$, $r \leqslant \operatorname{dist}(\operatorname{supp}(f_2), \operatorname{supp}(g_2))$. Our assumption on the generalized finite speed of wave propagation implies that

$$W(t, f_2, g_2) = 0, \quad t \in [0, r].$$

Consequently, $\sum_j H_Y(\ell_j/\Lambda)\hat{f}_2(j)\hat{g}_2(j) = 0$. The estimates (6.26) and (6.27) now imply that

$$\left|\sum_j H(\ell_j/\Lambda)\hat{f}(j)\hat{g}(j)\right| \leqslant c\Lambda^K\big(\varepsilon + Y^{-S}\big).$$

Since $\varepsilon > 0$ is arbitrary, $f, g \in L^1$ are arbitrary functions supported on the neighborhoods of x and y respectively, and ϕ_js are all continuous, this implies (4.7). $\square$

Proof of Theorem 4.2. It is enough to prove that if $H(t) := h(t^2)$, then

$$\max_{x\in\mathbb{X}} \int \left|\sum_{j=0}^\infty H(\ell_j/\Lambda)\phi_j(x)\phi_j(y)\right| \mathrm{d}\mu(y) \leqslant c, \quad \Lambda \geqslant 1/2. \qquad (6.28)$$

Let $x \in \mathbb{X}$. In this proof only, let $r = 4/\Lambda$. In view of (6.14) and (4.4),

$$\int_{B(x,r)} \left| \sum_{j=0}^{\infty} H(\ell_j/\Lambda)\phi_j(x)\phi_j(y) \right| \mathrm{d}\mu(y) \leqslant c\Lambda^K \mu\big(B(x,r)\big) \leqslant c. \tag{6.29}$$

In view of (4.7) and (6.16),

$$\int_{\Delta(x,r)} \left| \sum_{j=0}^{\infty} H(\ell_j/\Lambda)\phi_j(x)\phi_j(y) \right| \mathrm{d}\mu(y) \leqslant c \int_2^{\infty} v^{-S} v^{K-1}\,\mathrm{d}v \leqslant c.$$

Together with (6.29), this implies (6.28). □

Acknowledgments

We thank Jürgen Prestin, Amit Singer, and the two referees for pointing out a mistake in an earlier version, as well as making many other comments, leading to a substantial improvement of the paper.

References

[1] M. Belkin, Problems of learning on manifolds, PhD thesis, University of Chicago, 2003.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in NIPS, vol. 14, MIT Press, Cambridge, MA, 2001, pp. 585–591.
[3] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 6 (15) (2003) 1373–1396.
[4] M. Belkin, P. Niyogi, Using manifold structure for partially labelled classification, in: Advances in NIPS, vol. 15, MIT Press, Cambridge, MA, 2003.
[5] M. Belkin, P. Niyogi, Semi-supervised learning on Riemannian manifolds, in: Machine Learning, vol. 56, Univ. Chicago, CS Dept., 2004, pp. 209–239 (Invited Special Issue on Clustering), TR-2001-30, 2001.
[6] J. Bergh, J. Löfström, Interpolation Spaces, an Introduction, Springer-Verlag, Berlin, 1976.
[7] C. da Boor, A Practical Guide to Splines, Springer-Verlag, New York, 1978.
[8] M. Chaplain, M. Ganesh, I. Graham, Spatio-temporal pattern formation on spherical surfaces: Numerical simulation and application to solid tumor growth, J. Math. Biol. 42 (2001) 387–423.
[9] E. Chávez, K. Figueroa, and G. Navarro, A fast algorithm for the all k nearest neighbors problem in general metric spaces, preprint.
[10] C.K. Chui, An Introduction to Wavelets, Academic Press, San Diego, 1992.
[11] F.R.K. Chung, Spectral graph theory, in: CBMS Regional Conf. Ser. Math., vol. 92, Conference Board of the Mathematical Sciences, Washington, DC, 1997.
[12] R.R. Coifman, S. Lafon, Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions, Appl. Comput. Harmon. Anal. 21 (1) (2006) 31–52.
[13] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part I: Diffusion maps, Proc. Natl. Acad. Sci. 2 (102) (2005) 7426–7431.
[14] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part II: Multiscale methods, Proc. Natl. Acad. Sci. 2 (102) (2005) 7432–7438.
[15] R.R. Coifman, M. Maggioni, Diffusion wavelets, Appl. Comput. Harmon. Anal. 21 (1) (2006) 53–94 (Tech. Rep., YALE/DCS/TR-1303, Yale Univ., 2004).
[16] R.A. DeVore, G.G. Lorentz, Constructive Approximation, Springer-Verlag, Berlin, 1993.
[17] D.L. Donoho, C. Grimes, Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. (2003) 5591–5596 (Tech. Rep., Statistics Dept., Stanford University).
[18] D.L. Donoho, O. Levi, J.-L. Starck, V.J. Martinez, Multiscale geometric analysis for 3-d catalogues, Technical report, Stanford Univ., 2002.
[19] A. Grigor'yan, Heat kernels and function theory on metric measure spaces, Contemp. Math. 338 (2003) 143–172.
[20] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Machine Intell. 27 (3) (2005) 328–340.
[21] R.I. Kondor, J. Lafferty, Diffusion kernels on graphs and other discrete structures, in: Proc. ICML, Morgan Kaufman, San Mateo, CA, 2002.
[22] S. Lafon, Diffusion maps and geometric harmonics, PhD thesis, Yale University, Dept. of Mathematics & Applied Mathematics, 2004.
[23] P.-C. Lo, Three dimensional filtering approach to brain potential mapping, IEEE Trans. Biomed. Eng. 46 (5) (1999) 574–583.
[24] M. Maggioni, J.C. Bremer Jr., R.R. Coifman, A.D. Szlam, Biorthogonal Diffusion Wavelets for Multiscale Representations on Manifolds and Graphs, vol. 5914, SPIE, 2005, p. 59141M.
[25] M. Maggioni, S. Mahadevan, Fast direct policy evaluation using multiscale analysis of Markov diffusion processes, University of Massachusetts, Department of Computer Science, Technical Report TR-2005-39, accepted at ICML 2006, 2005.
[26] S. Mahadevan, K. Ferguson, S. Osentoski, M. Maggioni, Simultaneous learning of representation and control in continuous domains, in: Proc. AAAI, AAAI Press, Boston, MA, 2006.

[27] S. Mahadevan, M. Maggioni, Value function approximation with diffusion wavelets and Laplacian eigenfunctions, University of Massachusetts, Department of Computer Science, Technical Report TR-2005-38, Proc. NIPS 2005, 2005.
[28] H.N. Mhaskar, D.V. Pai, Fundamentals of Approximation Theory, Narosa Publishing Co., Delhi, 2000.
[29] H.N. Mhaskar, J. Prestin, Polynomial frames: A fast tour, in: L.L. Schumaker, C.K. Chui, M. Neamtu (Eds.), Approximation Theory, vol. XI, Gatliburg, 2004, Nashboro Press, Brentwood, 2005, pp. 287–318.
[30] H.N. Mhaskar, J. Prestin, On local smoothness classes of periodic functions, J. Fourier Anal. Appl. 11 (3) (2005) 353–373.
[31] H.N. Mhaskar, Polynomial operators and local smoothness classes on the unit interval, J. Approx. Theory 131 (2004) 243–267.
[32] H.N. Mhaskar, On the representation of smooth functions on the sphere using finitely many bits, Appl. Comput. Harmon. Anal. 18 (3) (2005) 215–233.
[33] H.N. Mhaskar, J. Prestin, On the detection of singularities of a periodic function, Adv. Comput. Math. 12 (2000) 95–131.
[34] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Proc. NIPS, MIT Press, Cambridge, MA, 2001.
[35] S.M. Nikolskii, Approximation of Functions of Several Variables and Imbedding Theorems, Springer-Verlag, Berlin, 1975.
[36] P. Niyogi, I. Matveeva, M. Belkin, Regression and regularization on large graphs, Technical report, University of Chicago, 2003.
[37] A. Sikora, Riesz transforms, Gaussian bounds, and the method of wave equation, Math. Z. 247 (2004) 643–662.
[38] A. Singer, From graph to manifold Laplacian: The convergence rate, Appl. Comput. Harmon. Anal. 21 (1) (2006) 128–134.
[39] C.D. Sogge, Eigenfunction and Bochner–Riesz estimates on manifolds with boundary, Math. Res. Lett. 9 (2002) 205–216.
[40] A.D. Szlam, M. Maggioni, R.R. Coifman, A general framework for adaptive regularization based on diffusion processes on graphs, Technical report YALE/DCS/TR1365, Yale Univ., 2006.
[41] A.D. Szlam, M. Maggioni, R.R. Coifman, J.C. Bremer Jr., Diffusion-driven Multiscale Analysis on Manifolds and Graphs: Top-down and Bottom-up Constructions, vol. 5914, SPIE, San Diego, 2005, p. 59141D.
[42] H. Triebel, Fourier Analysis and Function Spaces, Teubner Texte Math., vol. 7, Teubner, Leipzig, 1977.
[43] P.M. Vaidya, An $o(n \log n)$ algorithm for the all-nearest-neighbors problem, Discrete Comput. Geom. 4 (10) (1989) 101–115.
[44] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: Advances in NIPS, vol. 17, MIT Press, Cambridge, MA, 2004.
[45] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: Proc. ICML, AAAI Press, Boston, MA, 2003.