Available online at www.sciencedirect.com
SCIENCE DIRECT®
Appl. Comput. Harmon. Anal. 21 (2006) 53–94
Applied and Computational Harmonic Analysis
www.elsevier.com/locate/acha

Diffusion wavelets

Ronald R. Coifman *, Mauro Maggioni

Program in Applied Mathematics, Department of Mathematics, Yale University, New Haven, CT 06510, USA

Received 28 October 2004; revised 11 April 2006; accepted 12 April 2006

Available online 8 June 2006

Communicated by the Editors

Abstract

Our goal in this paper is to show that many of the tools of signal processing, adapted Fourier and wavelet analysis can be naturally lifted to the setting of digital data clouds, graphs, and manifolds. We use diffusion as a smoothing and scaling tool to enable coarse graining and multiscale analysis. Given a diffusion operator T on a manifold or a graph, with large powers of low rank, we present a general multiresolution construction for efficiently computing, representing and compressing T^t. This allows a direct multiscale computation, to high precision, of functions of the operator, notably the associated Green's function, in compressed form, and their fast application. Classes of operators for which these computations are fast include certain diffusion-like operators, in any dimension, on manifolds, graphs, and in non-homogeneous media. We use ideas related to the Fast Multipole Methods and to the wavelet analysis of Calderón–Zygmund and pseudo-differential operators, to numerically enforce the emergence of a natural hierarchical coarse graining of a manifold, graph or data set. For example for a body of text documents the construction leads to a directory structure at different levels of generalization. The dyadic powers of an operator can be used to induce a multiresolution analysis, as in classical Littlewood–Paley and wavelet theory: we construct, with efficient and stable algorithms, bases of orthonormal scaling functions and wavelets associated to this multiresolution analysis, together with the corresponding downsampling operators, and use them to compress the corresponding powers of the operator. While most of our discussion deals with symmetric operators and relates to localization to spectral bands, the symmetry of the operators and their spectral theory need not be considered, as the main assumption is reduction of the numerical ranks as we take powers of the operator.

Keywords: Multiresolution; Multiscale analysis; Wavelets; Wavelets on manifolds; Wavelets on graphs; Diffusion semigroups; Laplace–Beltrami operator; Fast Multipole Method; Matrix compression; Spectral graph theory

1. Introduction

Our goal in this paper is to show that many of the tools of signal processing, adapted Fourier and wavelet analysis can be naturally lifted to the setting of digital data clouds, graphs, and manifolds. We use diffusion as a smoothing and scaling tool to enable coarse graining and multiscale analysis. We introduce a multiresolution geometric construction for the efficient computation of high powers of local diffusion operators, which have high powers with low numerical

* Corresponding author.
E-mail addresses: coifman@math.yale.edu (R.R. Coifman), mauro.maggioni@yale.edu (M. Maggioni).
URL: http://www.math.yale.edu/~mmm82 (M. Maggioni).

1063-5203/$ – see front matter
doi:10.1016/j.acha.2006.04.004

rank. This organization of the geometry of a matrix representing T enables fast direct computation of functions of the operator, notably the associated Green's function, in compressed form. Classes of operators satisfying these conditions include discretizations of differential operators, in any dimension, on manifolds, and in non-homogeneous media. The construction yields multiresolution scaling functions and wavelets and domains, manifolds, graphs, and other general classes of metric spaces. The quantitative requirements of numerical precision in matrix compression enforce the emergence of a natural hierarchical coarse graining of the data set. As an example, for a body of text documents our construction leads to a directory structure at different levels of specificity.

Our construction can be related to Fast Multipole Methods [1] and to the wavelet representation for Calderón–Zygmund integral operators and pseudo-differential operators of [2], but from a "dual" perspective. We start from a semigroup $\{T^t\}$, associated to a diffusion process (e.g., $T = e^{-\varepsilon\Delta}$), rather than from the Green's operator, since the latter is not available in the applications we are most interested in, where the space may be a graph and very little geometrical information is available. We use the semigroup to induce a multiresolution analysis, interpreting the powers of T as dilation operators acting on functions, and constructing precise downsampling operators to efficiently represent the multiscale structure. This yields a construction of multiscale scaling functions and wavelets in a very general setting.

For many examples arising in physics, geometry, and other applications, the powers of the operator T decrease in rank, thus suggesting the compression of the function (and geometric) space upon which each power acts. The scheme we propose, in a nutshell, consists in the following: apply T to a space of test functions at the finest scale, compress the range via a local orthonormalization procedure, represent T in the compressed range and compute T^2 on this range, compress its range and orthonormalize, and so on: at scale j we obtain a compressed representation of T^{2^j}, acting on a family of scaling functions spanning the range of $T^{1+2+2^2+\cdots+2^{j-1}}$, for which we have a (compressed) orthonormal basis, and then we apply T^{2^j}, locally orthonormalize and compress the result, thus getting the next coarser subspace. Among other applications, this allows a generalization of the Fast Multipole algorithm for the efficient computation of the product of the inverse "Laplacian" $(I-T)^{-1}$ (on the complement of the kernel, of course) with an arbitrary vector. This can be carried out (in compressed form) via the Schultz method [2]: we have

$$(I-T)^{-1} f = \sum_{k=1}^{+\infty} T^k f = \prod_{k=0}^{\infty} \bigl(I + T^{2^k}\bigr) f$$

letting $S_K = \sum_{k=1}^{2^K} T^k$,

$$S_{K+1} = S_K + T^{2^K} S_K = \prod_{k=0}^{K} \bigl(I + T^{2^k}\bigr). \tag{1.1}$$

Since the powers T^{2^k} have been compressed and can be applied efficiently to f, we can express $(I-T)^{-1}$ in compressed form and efficiently apply it to any function f.

The interplay between geometry of sets, function spaces on sets, and operators on sets is classical in harmonic analysis. Our construction views the columns of a matrix representing T as data points in Euclidean (or Hilbert) space, for which the first few eigenvectors of T provide coordinates (see [3,4]). The spectral theory of T provides a Fourier analysis on this set relating our approach to multiscale geometric analysis, Fourier analysis, and wavelet analysis. The action of a given diffusion semigroup on the space of functions on the set is analyzed in a multiresolution fashion, where dyadic powers of the diffusion operator correspond to dilations, and projections correspond to downsampling. The localization of the scaling functions we construct allows to reinterpret these operations in function space in a geometric fashion. This mathematical construction has a numerical implementation which is fast and stable. The framework we consider in this work includes at once large classes of manifolds, graphs, spaces of homogeneous type and can be extended even further.

Given a metric measure space and a symmetric diffusion semigroup on it, there is a natural multiresolution analysis associated to it, and an associated Littlewood–Paley theory for the decomposition of the natural function spaces on the metric space. These ideas are classical (see [5] for a specific instance in which the diffusion semigroup is center-stage, but the literature on multiscale decompositions in general settings is vast, see, e.g., [6–10] and references therein). Generalized Heisenberg principles exist [11] that guarantee that eigenspaces are well approximated by scaling functions spaces at the appropriate scale. Our work shows that scaling functions and wavelets can be constructed and

efficient numerical algorithms that allow the implementation of these ideas exist. In effect, this construction allows to "lift" multiscale signal processing techniques to graph and data sets, for compression and denoising of functions and operators on a graph (and of the graph itself), for approximation and learning.

It also allows the introduction of dictionaries of diffusion wavelet packets [12], with the corresponding fast algorithms for best basis selection.

The paper is organized as follows. In Section 2 we present a sketch of the construction of diffusion wavelets in the finite-dimensional case, with minimal assumptions on the operator and the graph on which it acts. We consider the examples of a sampled manifold and of a body of documents. In Section 3 we introduce definitions and notations that will be used throughout the paper. In Section 4 we present the construction of the multiresolution analysis and diffusion wavelets and study some of their properties, in the context of spaces of homogeneous type. In Section 5 we discuss several algorithms for the orthogonalization step, which is crucial to our construction. They also demonstrate how the construction generalizes beyond the setting of spaces of homogeneous type. In Section 6 we discuss how our construction allows to perform certain computations efficiently, in particular the calculation of low-frequency eigenfunctions of the Laplacian in compressed form, and of the Green's function of the Laplacian in compressed form. Section 7 is devoted to two classes of examples to which our construction applies naturally: diffusions on discretized Riemannian manifolds and random walks on weighted graphs. In Section 8 we present several examples of our construction. In Section 9 we discuss how to extend the functions we construct outside the original domain of definition. Section 10 is a discussion of related work. Section 11 concludes the paper, with hints at future work and open questions.

Material related to this paper, such as examples and Matlab scripts for their generation, is made available at: http://www.math.yale.edu/~mmm82/diffusionwavelets.html.

2. The construction in the finite dimensional, discrete setting

In this section we present a particular case of our construction in a finite-dimensional, purely discrete setting, for which only finite dimensional linear algebra is needed.

Here and in the rest of the paper we will use the notation $[L]_{B_1}^{B_2}$ to indicate the matrix representing the linear operator L with respect to the basis B_1 in the domain and B_2 in the range. A set of vectors B_1 represented on a basis B_2 will be written in matrix form $[B_1]_{B_2}$, where the rows of $[B_1]_{B_2}$ are the coordinates of the vectors B_1 in the coordinates B_2.

2.1. Setting and assumptions

We consider a finite weighted graph X and a symmetric positive definite and positive "diffusion" operator T on (functions on) X (for a precise definition and discussion of these hypotheses see Section 4).[1] Without loss of generality, we can assume that $\|T\|_2 \leqslant 1$. Interesting examples include:

(i) The graph could represent a metric space in which points are data (e.g., documents) and edges have positive weights (e.g., a function of the similarity between documents), and $I - T$ could be a Laplacian on X. The Laplacian is associated to a natural diffusion process and random walk P, and T is a self-adjoint operator conjugate to the Markov matrix P. See, for example, [13] and references therein. Also, T could be the heat kernel on a weighted undirected graph.

(ii) X could represent the discretization of a domain or manifold and $T = e^{-\varepsilon(\Delta+V)}$, where Δ is an elliptic partial differential operator such as a Laplacian on the domain, or the Laplace–Beltrami operator on a manifold with smooth boundary, and V is a non-negative potential function. This type of operators are called Schrödinger operators and are well studied objects in mathematical physics. See, for example, [14,15]. A particular example is when X represents a cloud of points generated by a (stochastic) process driven by a Langevin equation (e.g., a protein configuration in a solvent) and the diffusion operator is the Fokker–Planck operator associated to the system.

[1] The construction here described can be immediately generalized to the case of directed graphs and associated non-self-adjoint diffusion operators T, which arise and are important in several applications.

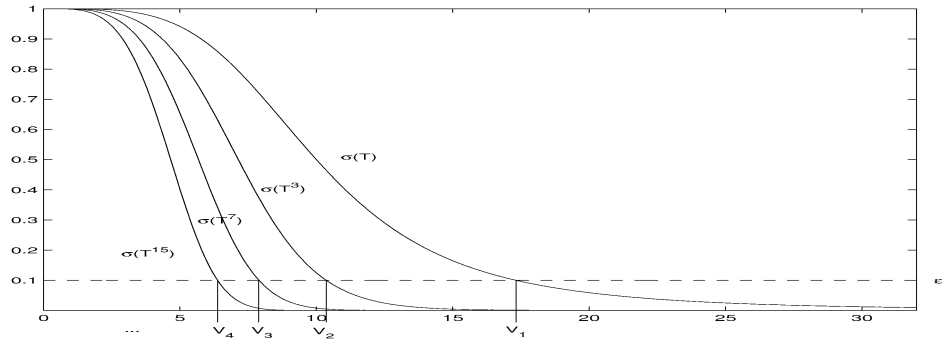

Fig. 1. Spectra of powers of T and corresponding multiscale eigenspace decomposition.

Our main assumptions are that T is local, i.e., $T(\delta_k)$, where δ_k is (a mollification of) the Dirac δ-function at $k \in X$, has small support, and that high powers of T have low numerical rank (see Fig. 1), for example because T is smoothing. Ideally there exists a $\gamma < 1$ such that for every $j \geqslant 0$ we have $\mathrm{Ran}_\varepsilon(T^{2^j}) < \gamma\,\mathrm{Ran}_\varepsilon(T^{2^{j-1}})$, where Ran_ε denotes the ε-numerical rank as in Definition 12.

We want to compute and describe efficiently the powers T^{2^j}, for $j > 0$, which describe the behavior of the diffusion at different time scales. This will allow the computation of functions of the operator in compressed form (notably of the Green's function $(I - T)^{-1}$), as well as the fast computation of the diffusion from any initial conditions. This is of course of great interest in the solution of discretized partial differential equations, of Markov chains (for example arising from physics), but also in learning and classification problems.

The reason why one expects to be able to compress high powers of T is that they are low rank by assumption, so that it should be possible to efficiently represent them on an appropriate basis, at the appropriate resolution. From the analyst's perspective, these high powers are smooth functions with small gradient (or even "band-limited" with small band), hence compressible.

2.2. *Construction of the multi-resolution and compression*

We start by fixing a precision $\varepsilon > 0$; we assume that T is self-adjoint and is represented on the basis $\Phi_0 = \{\delta_k\}_{k\in X}$ and consider the columns of T, which can be interpreted as the set of functions $\tilde{\Phi}_1 = \{T\delta_k\}_{k\in X}$ on X. We refer the reader to the diagram in Fig. 2 for a visual representation of the scheme presented. We use a local multiscale orthogonalization procedure, to be described later, to carefully orthonormalize these columns to get a basis $\Phi_1 = \{\varphi_{1,k}\}_{k\in X_1}$ (X_1 is *defined* as this index set), written with respect to the basis Φ_0, for the range of T up to precision ε (see Definition 12). This information is stored in the sparse matrix $[\Phi_1]_{\Phi_0}$. This yields a subspace that we denote by V_1. Essentially Φ_1 is a basis for the subspace V_1 which is ε-close to the range of T and with basis elements that are well-localized. Moreover, the elements of Φ_1 are coarser than the elements of Φ_0, since they are the result of applying the "dilation" T once. Obviously $|X_1| \leqslant |X|$, but this inequality may already be strict since the numerical range of T may be approximated, within the specified precision ε, by a subspace of smaller dimension. Whether this is the

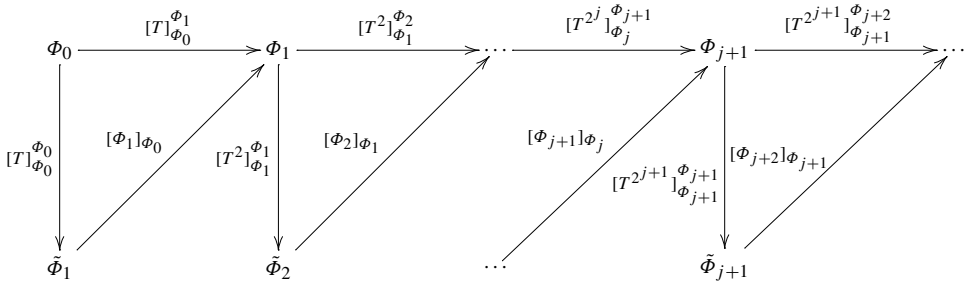

Fig. 2. Diagram for downsampling, orthogonalization, and operator compression. (All triangles are commutative by construction.)

$\{\Phi_j\}_{j=0}^{J}, \{\Psi_j\}_{j=0}^{J-1}, \{[T^{2^j}]_{\Phi_j}^{\Phi_j}\}_{j=1}^{J} \leftarrow$ `DiffusionWaveletTree` $([T]_{\Phi_0}^{\Phi_0}, \Phi_0, J, \mathrm{SpQR}, \varepsilon)$

// Input:
// $[T]_{\Phi_0}^{\Phi_0}$: a diffusion operator, written on the o.n. basis Φ_0
// Φ_0: an orthonormal basis which ε-spans V_0
// J: number of levels to compute
// SpQR: a function compute a sparse QR decomposition, see template below
// ε: precision.

// Output:
// The orthonormal bases of scaling functions, Φ_j, wavelets, Ψ_j, and
// compressed representation of T^{2^j} on Φ_j, for j in the requested range.

for $j = 0$ **to** $J - 1$ **do**

$[\Phi_{j+1}]_{\Phi_j}$, $[T]_{\Phi_0}^{\Phi_1} \leftarrow$`SpQR`$([T^{2^j}]_{\Phi_j}^{\Phi_j}, \varepsilon)$

$T_{j+1} := [T^{2^{j+1}}]_{\Phi_{j+1}}^{\Phi_{j+1}} \leftarrow [\Phi_{j+1}]_{\Phi_j}[T^{2^j}]_{\Phi_j}^{\Phi_j}[\Phi_{j+1}]_{\Phi_j}^{*}$

$[\Psi_j]_{\Phi_j} \leftarrow \mathrm{SpQR}(I_{\langle\Phi_j\rangle} - [\Phi_{j+1}]_{\Phi_j}[\Phi_{j+1}]_{\Phi_j}^{*}, \varepsilon)$

end

Function template for sparse QR factorization:
$Q, R \leftarrow$ `SpQR` (A, ε)

// Input:
// A: sparse $n \times n$ matrix
// ε: precision.

// Output:
// Q, R matrices, possibly sparse, such that $A =_\varepsilon QR$
// Q is $n \times m$ and orthogonal
// R is $m \times n$, and upper triangular up to a permutation
// the columns of Q ε-span the space spanned by the columns of A.

Fig. 3. Pseudo-code for construction of a diffusion wavelet tree.

case or not, we have computed the sparse matrix $[T]_{\Phi_0}^{\Phi_1}$, a representation of an ε-approximation of T with respect to Φ_0 in the domain and Φ_1 in the range. We can also represent T in the basis Φ_1: we denote this matrix by $[T]_{\Phi_1}^{\Phi_1}$ and compute $[T^2]_{\Phi_1}^{\Phi_1} = [\Phi_1]_{\Phi_0}[T^2]_{\Phi_0}^{\Phi_0}[\Phi_1]_{\Phi_0}^{T} = [T]_{\Phi_0}^{\Phi_1}([T]_{\Phi_0}^{\Phi_1})^*$. The last equality holds only when T is self-adjoint and it is the only place where we use self-adjointness.

We proceed now by looking at the columns of $[T^2]_{\Phi_1}^{\Phi_1}$, which are $\tilde{\Phi}_2 = \{[T^2]_{\Phi_1}^{\Phi_1}\delta_k\}_{k\in X_1}$ i.e., by unraveling the bases on which this is happening, $\{T^2\varphi_{1,k}\}_{k\in X_1}$ up to precision ε. Again we can apply a local orthonormalization procedure to this set: this yields an orthonormal basis $\Phi_2 = \{\varphi_{2,k}\}_{k\in X_2}$ for the range of T_1^2 up to precision ε, and also for the range of T_0^3 up to precision 2ε. Observe that Φ_2 is naturally written with respect to the basis Φ_1, and hence encoded in the matrix $[\Phi_2]_{\Phi_1}$. Moreover, depending on the decay of the spectrum of T, $|X_2|$ is in general a fraction of $|X_1|$. The matrix $[T^2]_{\Phi_1}^{\Phi_2}$ is then of size $|X_2| \times |X_1|$, and the matrix $[T^4]_{\Phi_2}^{\Phi_2} = [T^2]_{\Phi_1}^{\Phi_2}([T^2]_{\Phi_1}^{\Phi_2})^*$, a representation of T^4 acting on Φ_2, is of size $|X_2| \times |X_2|$.

After j steps in this fashion, we will have a representation of T^{2^j} onto a basis $\Phi_j = \{\varphi_{j,k}\}_{k\in X_j}$, encoded in a matrix $T_j := [T^{2^j}]_{\Phi_j}^{\Phi_j}$. The orthonormal basis Φ_j is represented with respect to Φ_{j-1} and encoded in the matrix $[\Phi_j]_{\Phi_{j-1}}$. We let $\tilde{\Phi}_j = T_j\Phi_j$ We can represent the next dyadic power of T on Φ_{j+1} on the range of T^{2^j}. Depending on the decay of

the spectrum of T, we expect $|X_j| \ll |X|$, in fact in the ideal situation the spectrum of T decays fast enough so that there exists $\gamma < 1$ such that $|X_j| < \gamma |X_{j-1}| < \cdots < \gamma^j |X|$. This corresponds to downsampling the set of columns of dyadic powers of T, thought of as vectors in $\mathcal{L}^2(X)$. The hypothesis that the rank of powers of T decreases guarantees that we can downsample and obtain coarser and coarser lattices in this spaces of columns.

While Φ_j is naturally identified with the set of Dirac δ-functions on X_j, we can extend these functions living on the "compressed" (or "downsampled") graph X_j to the whole initial graph X by writing

$$[\Phi_j]_{\Phi_0} = [\Phi_j]_{\Phi_{j-1}} [\Phi_{j-1}]_{\Phi_0} = \cdots = [\Phi_j]_{\Phi_{j-1}} [\Phi_{j-1}]_{\Phi_{j-2}} \cdots [\Phi_1]_{\Phi_0} [\Phi_0]_{\Phi_0}. \tag{2.1}$$

Since every function in Φ_0 is defined on X, so is every function in Φ_j. Hence any function on the compressed space X_j can be extended naturally to the whole X. In particular, one can compute low-frequency eigenfunctions on X_j in compressed form and then extend them to the whole X. The elements in Φ_j are at scale $T^{2^{j+1}-1}$ and are much coarser and "smoother," than the initial elements in Φ_0, which is why they can be represented in compressed form. The projection of a function onto the subspace spanned by Φ_j will be by definition an approximation to that function at that particular scale. We refer the reader to Fig. 3 for pseudo-code illustrating the algorithm.

2.3. Scaling function and wavelet transforms

There is an associated fast scaling function transform: suppose we are given f on X and want to compute $\langle f, \varphi_{j,k} \rangle$ for all scales j and corresponding "translations" k. Being given f means we are given $(\langle f, \varphi_{0,k} \rangle)_{k \in X}$. Then we can compute $(\langle f, \varphi_{1,k} \rangle)_{k \in X_1} = [\Phi_1]_{\Phi_0} (\langle f, \varphi_{0,k} \rangle)_{k \in X}$ and so on for all scales. The sparser the matrices $[\Phi_j]_{\Phi_{j-1}}$ (and $[T]_{\Phi_j}^{\Phi_j}$), the faster this computation. This generalizes the classical scaling function transform. We will show later that wavelets can be constructed as well, and that a wavelet transform is also possible.

2.4. Computation of functions of powers of T

In the same way, any power of T can be applied efficiently to a function f. Also, the Green's function $(I - T)^{-1}$ can be applied efficiently to any function, since it can be represented as a (short) product of dyadic powers of T as in (1.1), each of which can be applied efficiently.

We are at the same time compressing the powers of the operator T and the space X itself, at essentially the optimal "rate" at each scale, as dictated by the portion of the spectrum of the powers of T which is above the precision ε.

Observe that each point in X_j can be considered as a "local aggregation" of points in X_{j-1}, which is completely dictated by the action of the operator T on functions on X: the operator itself is dictating the geometry with respect to which it should be analyzed, compressed or applied to any vector.

2.5. Structure of the scaling function bases

The choice of the maps $[\Phi_{j+1}]_{\Phi_j}$, which at the same time orthogonalize and downsample the redundant families $\tilde{\Phi}_{j+1}$, is quite arbitrary, but should be such that the representation of Φ_{j+1} onto Φ_j is sparse.

For the construction of Φ_j we use a multiresolution strategy which could be of independent interest in wavelet theory and in numerical analysis. The details of the construction are given in Section 5. This construction yields an orthonormal basis $\Phi_j = \{\varphi_{j,k}\}_{k \in X_j}$. The elements of this basis can be rearranged in the form $\Phi_j = \{\{\varphi_{j,k,l}\}_{k \in \mathcal{K}_{j,l}}\}_{l=1,\ldots,L_j}$. The index l controls what we call the layer and the index k controls the translation within each layer. The functions $\{\varphi_{j,k,0}\}_{k \in \mathcal{K}_{j,0}}$ have, qualitatively speaking, support of size $\asymp 2^j$, and they are centered at points roughly equispaced at a distance $\asymp 2 \cdot 2^j$, in such a way that their supports are disjoint. The functions $\{\varphi_{j,k,1}\}_{k \in \mathcal{K}_{j,1}}$ have, always qualitatively speaking, support of size $\asymp 2 \cdot 2^j$, and are centered at points roughly equispaced at a distance $\asymp 4 \cdot 2^j$, so that their supports disjoint. These functions of course have been orthogonalized to $\{\varphi_{j,k,0}\}_{k \in \mathcal{K}_{j,0}}$. We proceed by adding one layer of scaling functions at a time, orthogonalizing each time to the previous layers and picking functions that, after orthogonalization, have disjoint supports. In this way the functions $\{\varphi_{j,k,l}\}_{k \in \mathcal{K}_{j,l}}$ have support of size $\asymp 2^l \cdot 2^j$ and are centered at points roughly equispaced at a distance $\asymp 2 \cdot 2^l \cdot 2^j$, so that their supports are disjoint.

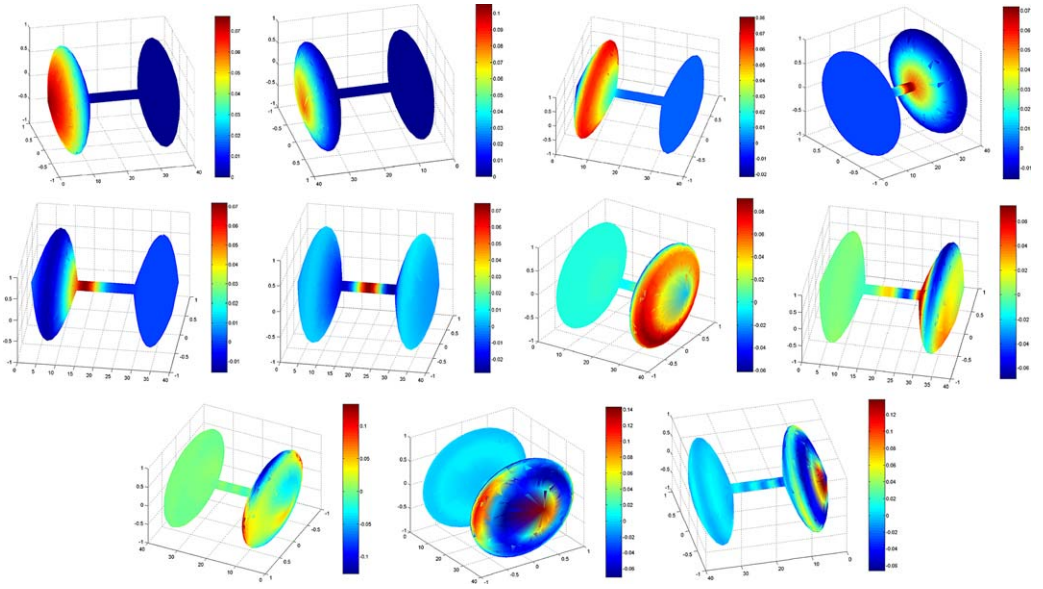

Fig. 4. Some diffusion scaling functions (first 7 pictures, counting left to right, top to bottom) and wavelets (last 4 pictures) at different scales on a dumbell-shaped manifold sampled at 1400 points. Colors are proportional to the function values. Please notice that the color scale may change from picture to picture.

The wavelets Ψ_j are constructed in a similar fashion, simply by continuing the orthogonalization process till the domain (instead of the range) of $[T^{2^j}]_{\Phi_j}^{\Phi_j}$ is exhausted. In practice, to preserve numerical accuracy, this is achieved by starting with the columns of $I_{V_j} - \Phi_{j+1}\Phi_{j+1}^*$.

2.6. Dumbell manifold

As an example, we consider a dumbell-shaped manifold. We sampled the manifold at 1400 points, and we consider the diffusion associated to the (discretized) Laplace–Beltrami operator discussed in Section 7. In Fig. 4 we represent some scaling functions and wavelets, at different scales. Observe how the shape of the scaling functions and wavelets changes with the scale and with the location on the dumbell. This is a consequence of the curvature of the manifold, which varies from point to point and affects the diffusion operator. The scaling functions look like "bump" functions at different locations and have no or few oscillations. Wavelets (last four pictures in Fig. 4) have an oscillatory character. Scaling functions and wavelets on the middle part of the dumbell resemble "standard" scaling functions and wavelets on the interval, rotated around the axis of symmetry of the manifold.

2.7. A corpus of documents

We consider the following cloud of digital data. We are given 1047 articles from Science News and a dictionary of 10,906 words chosen as being relevant for this body of documents. Each document is categorized as belonging to one of the following fields: Anthropology, Astronomy, Social Sciences, Earth Sciences, Biology, Mathematics, Medicine, or Physics. Let $\mathcal{C}$ denotes the set of these categories. This information can be assimilated to the function cat: $X \to \mathcal{C}$, defined by

$$\text{cat}(x) = \{\text{category of the point } x\}.$$

This data has been prepared and is analyzed in the manuscript [16], to which we refer the reader for further information.

A similarity $W_{xy} = W_{yx}$ is given between documents pairs, at least for documents which are very similar. This similarity measure does not depend on the categorization, but only on the word frequencies of the documents. We will report on this and other examples in an upcoming report [17]. Each document has in average 33 neighbors. The (graph) Laplacian is defined as $L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$, where D is the diagonal matrix with entries $D_{xx} = \sum_{y\sim x} W_{xy}$; L is

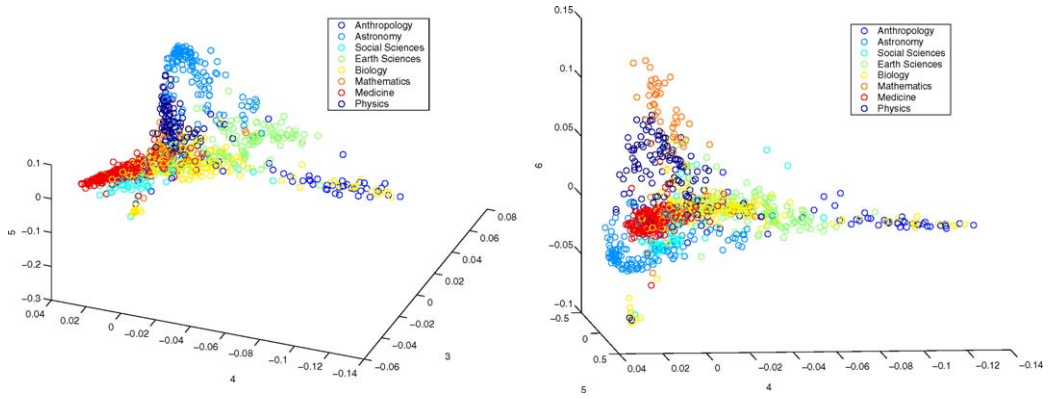

Fig. 5. Embedding $\Xi_6^{(0)}(x) = (\xi_1(x), \ldots, \xi_6(x))$: on the left coordinates 3, 4, 5, and on the right coordinates 4, 5, 6.

self-adjoint and the spectrum of L is contained in $[0, 1]$ [13]. The diffusion kernel we consider is $T = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = I - L$.

We can compute the eigenvalues $0 = \lambda_0 \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_n \leqslant \cdots$ of L and the corresponding eigenvectors $\xi_0, \xi_1, \ldots, \xi_n, \ldots$. As described in [3,18–20] and in Section 4.2, these eigenfunctions can be used to define, for each $n \geqslant 0$ and $t \geqslant 0$, the non-linear embedding (Fig. 5)

$$\Xi_n^{(t)} : X \to \mathbb{R}^n$$

defined by

$$x \mapsto \left(\lambda_i^{\frac{t}{2}} \xi_i(x)\right)_{i=1,\ldots,n}.$$

The properties of these embeddings are discussed in the references above. This embedding seems a posteriori particularly meaningful since it separates quite well among the different categories. For example a simple K-means or hierarchical clustering algorithm ran on $\Xi_n^{(t)}(X)$, yields clusters which match closely the given labels, with some (usually interesting!) outliers. This just corresponds to a particular choice of kernel K-means (or hierarchical clustering), motivated by diffusion distance. We do not give the details of the results here, since work is still in progress [17], and the point of this discussion is that interesting information can be extracted from the multiscale construction presented in this paper. The diffusion kernel is iterated over the set, induces a natural multiscale structure, that gets organized coherently, in space and scale.

We construct the diffusion scaling functions and wavelets on this cloud of points. Some scaling functions are represented in Fig. 6. Scaling functions at different scales represent (from coarse to fine) categories, sub-categories ("topics"), sub-sub-categories ("specialty topics"), and so on. By this we mean that the (essential) support of most scaling functions consists of a set of documents, which are related by diffusion at a certain time (= scale), and have common, well-distinguished topics.

For example, we represent in Fig. 6 some scaling functions at scale 3 and retrieve the documents corresponding to their (essential) supports. We find that:

$\phi_{3,4}$ is about Mathematics, but in particular applications to networks, encryption and number theory;
$\phi_{3,10}$ is about Astronomy, but in particular papers in X-ray cosmology, black holes, galaxies;
$\phi_{3,15}$ is about Earth Sciences, but in particular earthquakes;
$\phi_{3,5}$ is about Biology and Anthropology, but in particular about dinosaurs;
$\phi_{3,2}$ is about Science and talent awards, inventions and science competitions.

These scaling functions (as well as most of the others at scale 3) not only have (essential) support in each of the categories, but on some meaningful, specialized topics. We move then to the next scale: the supports of the scaling functions grow, and so do the corresponding topics. For example $\phi_{4,3}$ now corresponds to a larger portion of Biology and Anthropology, compared to $\phi_{3,5}$, and includes articles on fossils in general, not just dinosaurs. $\phi_{4,2}$ corresponds to a larger portion of Astronomy, compared to $\phi_{3,10}$, and includes articles on comets, asteroids, and space travel. The list of examples could continue. In Fig. 6 we represent scaling functions at even coarser scale and it is obvious how they

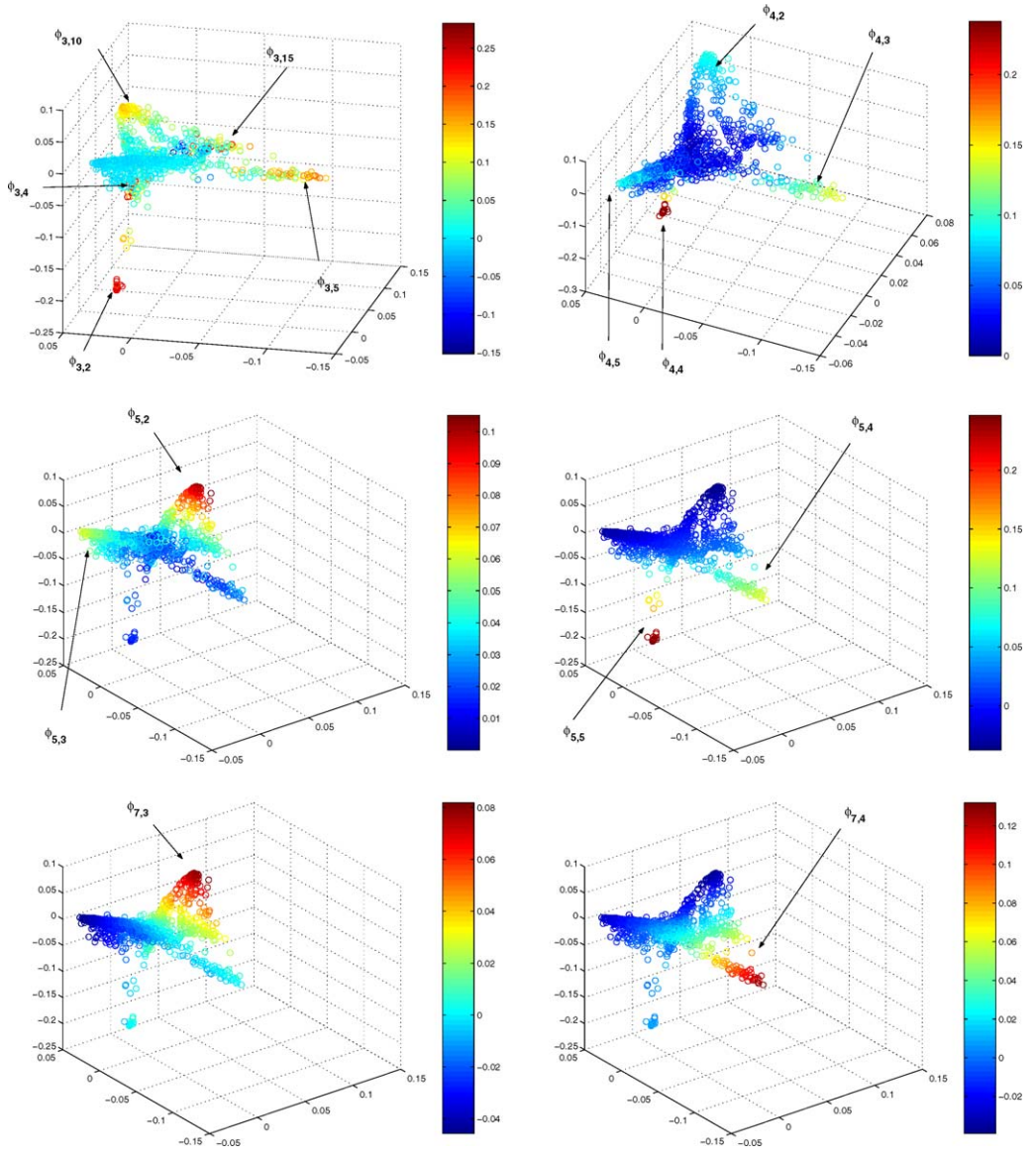

Fig. 6. Scaling functions at different scales represented on the set embedded in $\mathbb{R}^3$ via $(\xi_3(x), \xi_4(x), \xi_5(x))$.

grow, by diffusion, on the set. To show that the coarse scaling functions correspond rather well with the categories we are given, for each category $c \in \mathcal{C}$ we can consider

$$\chi_c(x) = \begin{cases} 1, & x \text{ if } \operatorname{cat}(x) = c, \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

and expand this function onto the scaling functions at a coarse scale, to get a sequence of coefficients $\{\langle \chi_c, \phi_{j,k} \rangle\}_{k \in \mathcal{K}_j}$. For most classes we get a small number of large coefficients, which means that few scaling functions are enough to represent well a given category. The only exception is the category Biology. This could be explained by the fact that Biology is a category having rich and complex relationship with many other categories, as can be seen from the documents themselves, but also from the location of documents related to Biology in the diffusion embedding in Fig. 5: many documents in Biology have strong connections with Medicine, Mathematics, Earth Sciences, and Anthropology. Details are left to [17].

Further examples can be found in Section 8. The example in Section 8.1 is particularly simple and detailed.

3. Notation and definitions

To present our construction in some generality, we will need the following definitions and notations. We start by introducing a model for the spaces X we will consider.

Definition 1. Let X be a set. A function $d : X \times X \to [0, +\infty)$ is called a *quasi-metric* if:

(i) $d(x, y) \geqslant 0$ for every $x, y \in X$, with equality if and only if $x = y$,
(ii) $d(x, y) = d(y, x)$, for every $x, y \in X$,
(iii) there exists $A_X > 0$ such that $d(x, y) \leqslant A_X(d(x, z) + d(z, y))$ for every $x, y, z \in X$ (quasi-triangle inequality).

The pair (X, d) is called a quasi-metric space. (X, d) is a *metric space* if one can choose $A_X = 1$ in (iii).

Example 2. A weighted undirected connected graph (G, E, W), where W is the set of positive weights on the edges in E, is a metric space when the distance is defined by

$$d(x, y) = \inf_{\gamma_{x,y}} \sum_{e \in \gamma_{x,y}} w_e,$$

where $\gamma_{x,y}$ is a path connecting x and y. Often in applications a measure on (the vertices of) G is either uniform or specified by some connectivity properties of each vertex (e.g., sum of weights of the edges concurrent in each vertex).

Let (X, d, μ) be a quasi-metric space with Borel measure μ. For $x \in X$, $\delta > 0$, let

$$B_\delta(x) = \{y \in X\colon\ d(x, y) < \delta\}$$

be the open ball of radius δ around x. For a subset S of X, we will use the notation

$$\mathcal{N}_\delta(S) = \{x \in X\colon\ \exists y \in S\colon\ d(x, y) < \delta\}$$

for the δ-neighborhood of S.

Since we would like to handle finite, discrete and continuous situations in a unique framework, and since many finite-dimensional discrete problems arise from the discretization of continuous and infinite-dimensional problems, we introduce the following definitions.

Definition 3. A quasi-metric measure space (X, d, μ) is said to be of *homogeneous type* [6,21] if μ is a non-negative Borel measure and there exists a constant $C_X > 0$ such that for every $x \in X$, $\delta > 0$,

$$\mu\big(B_{2\delta}(x)\big) \leqslant C_X \mu\big(B_\delta(x)\big). \tag{3.1}$$

We assume $\mu(B_\delta(x)) < \infty$ for all $x \in X, \delta > 0$, and we will work on *connected* spaces X. In the continuous situation, we will assume that X is purely non-atomic, i.e., $\mu(\{x\}) = 0$ for every $x \in X$.

One can replace the quasi-metric d with a quasi-metric ρ, inducing a topology equivalent to the one induced by d, so that the δ-balls in the ρ metric have measure approximately δ: it is enough to define $\rho(x, y)$ as the measure of the smallest ball containing x and y [22]. One can also assume some Hölder-smoothness for ρ, in the sense that

$$\big|\rho(x, y) - \rho(x', y)\big| \leqslant L_X \rho(x, x')^\beta \big[\rho(x, y) + \rho(x', y)\big]^{1-\beta}$$

for some $L_X > 0$, $\beta \in (0, 1)$, $c > 0$, see [7,8,22,23].

Definition 4. Let (X, ρ, μ) be a space of homogeneous type and $\gamma > 0$. A subset $\{x_i\}_i$ of X is a γ-lattice if $\rho(x_i, x_j) > \frac{1}{2} c_{\mathrm{lat}} \gamma$, for all $i \neq j$, and if for every $y \in X$, there exists $i(y)$ such that $\rho(y, x_{i(y)}) < 2 c_{\mathrm{lat}} \gamma$. Here c_{lat} depends only on the constants A_X, C_X of the space of homogeneous type.

Example 5. Examples of spaces of homogeneous type include:

(i) Euclidean spaces of any dimension, with isotropic or anisotropic metrics induced by positive-definite bilinear forms and their powers (see, e.g., [24]).
(ii) Compact Riemannian manifolds of bounded curvature, with the geodesic metric, or also with respect to metrics induced by certain classes of vector fields [25].
(iii) Finite graphs of bounded degree with shortest path distance, in particular k-regular graphs, where the degree of a vertex is defined as

$$d_x = \sum_{y \sim x} w_{yx},$$

where $y \sim x$ means there is an edge between y and x.

In quite general settings, one can construct the following analogue of the dyadic cubes in the Euclidean setting [26,27].

Theorem 6. *Let (X, ρ, μ) be a space of homogeneous type as above. There exists a collection of open subsets*

$$\mathcal{Q} = \left\{ \{Q_{j,k}\}_{k \in \mathcal{K}_j} \right\}_{j \in \mathbb{Z}}$$

and constants $\delta_X > 1$, $\eta > 0$ and $c_1, c_2 \in (0, \infty)$, depending only on A_X, C_X, such that

(i) *for every $j \in \mathbb{Z}$, $\mu(X \setminus \bigcup_{k \in \mathcal{K}_j} Q_{j,k}) = 0$;*
(ii) *for $j \geqslant j_0$ either $Q_{j_0,k} \subseteq Q_{j,k'}$ or $\mu(Q_{j,k} \cap Q_{j_0,k'}) = 0$;*
(iii) *for each $j \in \mathbb{Z}$, $k \in \mathcal{K}_j$ and $j' < j$, there exists a unique k' such that $Q_{j',k} \subseteq Q_{j,k'}$;*
(iv) *each $Q_{j,k}$ contains a point $x_{j,k}$, called center of $Q_{j,k}$, such that*

$$B_{\min\{c_1 \delta_X^j, \operatorname{diam} X\}}(x_{j,k}) \subseteq Q_{j,k} \subseteq B_{c_2 \delta_X^j}(x_{j,k});$$

(v) *for each $j \in \mathbb{Z}$ and each $k \in \mathcal{K}_j$, if we let*

$$\partial_t Q_{j,k} := \left\{ x \in Q_{j,k} \colon \rho(x, X \setminus Q_{j,k}) \leqslant t \delta_X^j \right\},$$

then $\mu(\partial_t Q_{j,k}) \leqslant c_2 t^\eta \mu(Q_{j,k})$.

A set $\mathcal{Q} = \{Q_{j,k}\}_{j,k}$ satisfying the properties (i)–(iv) above is called a family of *dyadic cubes* for X; $\{Q_{j,k}\}_{k \in \mathcal{K}_j}$ is called the set of *dyadic cubes* at scale j, and the set of points $\Gamma_j = \{x_{j,k}\}_{k \in \mathcal{K}_j}$ is called the set of *dyadic centers* at scale j. For each $j \in \mathbb{Z}$ the unique dyadic cube at scale j containing x will be denoted by $Q_j(x)$.

Example 7. In the case of Euclidean space $\mathbb{R}^n$, the classical dyadic cubes correspond to choices $\delta_X = 2$, $\eta = 1$, $c_1 = c_2 = 1$, $\mathcal{K}_j = 2^j \mathbb{Z}$ in the theorem above.

Definition 8. A *center set* $\operatorname{center}(\Phi)$ for a family of functions $\Phi = \{\varphi_k\}_{k \in \mathcal{K}}$ is a set $\{x_k\}_{k \in \mathcal{K}} \subseteq X$ such that there exists $\eta > 0$ such that for every $k \in \mathcal{K}$ we have $\operatorname{supp} \varphi_k \subseteq B_\eta(x_k)$. A set of functions $\{\varphi_k\}_{k \in \mathcal{K}}$ with such a supporting set is called *η-local*.

Notation 9. If Ψ is a family of functions on X and $S \subseteq X$, we let

$$\Psi|_S = \{\psi \in \Psi \colon \operatorname{supp} \psi \subseteq S\}.$$

Notation 10. If V is a closed linear subspace of $\mathcal{L}^2(X, \mu)$, we denote the orthogonal projection onto V by P_V. The closure $\bar{V}$ of any subspace V is taken in $\mathcal{L}^2(X, \mu)$, unless otherwise specified.

Notation 11. If L is a self-adjoint bounded operator on $\mathcal{L}^2(X, \mu)$, with spectrum $\sigma(L)$, and spectral decomposition

$$L = \int_{\sigma(L)} \lambda \, \mathrm{d}E_\lambda,$$

we define

$$L_\varepsilon = \int\limits_{\{\lambda\in\sigma(L):\ |\lambda|>\varepsilon\}} \lambda \,\mathrm{d}E_\lambda.$$

Definition 12. Let $\mathcal{H}$ be a Hilbert space and $\{v_k\}_{k\in\mathcal{K}} \subseteq \mathcal{H}$. Fix $\varepsilon > 0$. A set of vectors $\{\xi_i\}_{i\in\mathcal{I}}$ *ε-spans* $\{v_k\}_{k\in\mathcal{K}}$ if for every $k \in \mathcal{K}$

$$\|P_{\langle\{\xi_i\}_{i\in\mathcal{I}}\rangle} v_k - v_k\|_{\mathcal{H}} \leqslant \varepsilon.$$

We also say, with abuse of notation, that $\langle\{\xi_i\}_{i\in\mathcal{I}}\rangle$ is an *ε-span* of $\langle\{v_k\}_{k\in\mathcal{K}}\rangle$ and write $\langle\{v_k\}_{k\in\mathcal{K}}\rangle \subseteq \langle\{\xi_i\}_{i\in\mathcal{I}}\rangle_\varepsilon$. We let

$$\dim_\varepsilon\big(\{v_k\}_{k\in\mathcal{K}}\big) = \inf\big\{\dim(V'):\ V' \text{ is an } \varepsilon\text{-span of } \big\langle\{v_k\}_{k\in\mathcal{K}}\big\rangle\big\}.$$

The abuse of notation is in the fact that ε-span does not apply to a subspace, but to a set of vectors (spanning some subspace). For instance the ε-spans corresponding to two different set of vectors, spanning the same subspace, are in general different. As an example, if $\{v_1, v_2\}$ are orthonormal and ε is small, $\langle v_1\rangle$ is an ε-span for $\langle\{v_1, \varepsilon v_2\}\rangle$, but not for $\langle\{v_1, v_2\}\rangle$.

4. Multiresolution analysis induced by symmetric diffusion semigroups

4.1. Symmetric diffusion semigroups

We start from the following definition [5,14]:

Definition 13. Let $\{T^t\}_{t\in[0,+\infty)}$ be a family of operators on (X, μ), each mapping $\mathcal{L}^2(X, \mu)$ into itself. Suppose this family is a semigroup, i.e., $T^0 = I$ and $T^{t_1+t_2} = T^{t_1}T^{t_2}$ for any $t_1, t_2 \in [0, +\infty)$, and $\lim_{t\to 0^+} T^t f = f$ in $\mathcal{L}^2(X, \mu)$ for any $f \in \mathcal{L}^2(X, \mu)$.

Such a semigroup is called a *symmetric diffusion semigroup* (or a *Markovian semigroup*) if it satisfies the following:

(i) $\|T^t\|_p \leqslant 1$ for every $1 \leqslant p \leqslant +\infty$ (contraction property).
(ii) Each T^t is self-adjoint (symmetry property).
(iii) T^t is positivity preserving: $Tf \geqslant 0$ for every $f \geqslant 0$ in $\mathcal{L}^2(X)$ (positivity property).
(iv) The semigroup has a positive self-adjoint generator $-\Delta$, so that

$$T^t = e^{-t\Delta}. \tag{4.1}$$

While some of these assumptions are not strictly necessary for our construction, their adoption simplifies this presentation without reducing the types of applications we are interested in at this time.

Some of the hypotheses in the definition above are redundant. For example by interpolation it is easy to see that if T^t is a contraction on $\mathcal{L}^\infty(X, \mu)$ for all times t, then it is a contraction on $\mathcal{L}^p(X, \mu)$ for all $1 \leqslant p \leqslant +\infty$ and all $t \geqslant 0$. Moreover, one can prove that if T^t is compact on $\mathcal{L}^2(X, \mu)$ for all $t > 0$, then it is compact on all $\mathcal{L}^p(X, \mu)$, for all $1 < p < +\infty$ and $t > 0$, and the spectrum, for each $t > 0$, is independent of p and every $\mathcal{L}^2(X, \mu)$ eigenfunction is in $\mathcal{L}^p(X, \mu)$ for $p \in (1, +\infty)$. See [5,14] for these and related properties of semigroups.

We will denote by $\sigma(T)$ the spectrum of T (so that $\{\lambda^t\}_{\lambda\in\sigma(T)}$ is the spectrum of T^t), and by $\{\xi_\lambda\}_{\lambda\in\sigma(T)}$ the corresponding basis (orthogonal since T^t is normal) of eigenvectors, normalized in $\mathcal{L}^2$ (without loss of generality). Here and in all that follows we will use the obvious abuse of notation that implicitly accounts for possible multiplicities of the eigenvalues.

Remark 14. Observe that $\sigma(T) \subseteq [0, 1]$: by the semigroup property and (ii), we have

$$T = T^{\frac{1}{2}}T^{\frac{1}{2}} = \big(T^{\frac{1}{2}}\big)^* T^{\frac{1}{2}}$$

so the spectrum is non-negative. The upper bound obviously follows from condition (i).

4.1.1. Examples

(a) The Poisson semigroup, for example, on the circle or half-space, or anisotropic and/or higher dimensional anisotropic versions (e.g., [24]).

(b) The random walk diffusion induced by a symmetric Markov chain (on a graph, a manifold etc.), or the dual/reversed random walk. In particular Markov chains from statistical physics and dynamics are of great interest.

(c) The semigroup $T^t = e^{tL}$ generated by a second-order differential operator on some interval (a, b) (a and/or b possibly infinite), in the form

$$Lf = a_2(x)\frac{\mathrm{d}^2}{\mathrm{d}x^2}f + a_1(x)\frac{\mathrm{d}}{\mathrm{d}x}f + a_0(x)f(x)$$

with $a_2(x) > 0, a_0(x) \leqslant 0$, acting on a subspace of $\mathcal{L}^2((a, b), q(x)\,\mathrm{d}x)$, where q is an appropriate weight function, given by imposing appropriate boundary conditions, so that L is (unbounded) self-adjoint. Conditions (i) to (iii) are satisfied and (iv) is satisfied if $c = 0$.

This extends to $\mathbb{R}^n$ by considering elliptic partial differential operators in the form

$$Lf = \frac{1}{w(x)}\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}f\right) + c(x)f,$$

where we assume $c(x) \leqslant 0$ and $w(x) > 0$, and we consider this operator as defined on a smooth manifold. L, applied to functions with appropriate boundary conditions, is formally self-adjoint and generates a semigroup satisfying (i) to (iii) and (iv) is satisfied if $c = 0$.

An important case is the Laplace–Beltrami operator on a compact smooth manifold (e.g., a Lie group), with Ricci curvature bounded below [15] and subordinated operators [5].

A potential term V could be added to the Laplacian, to yield the semigroup $e^{t(L+V)}$. The operator $L + V$ is called a Schrödinger operator, a very well-studied object in mathematical physics: see, e.g., [14,15] and references therein, in particular for conditions guaranteeing compactness, self-adjointness and positivity.

(f) If (X, d, μ) is derived from a finite graph (G, W) as described above in Example 2, one can define $D_{ii} = \sum_j W_{ij}$ and then the matrix $D^{-1}W$ is a Markov matrix, which corresponds to the natural random walk on G. The operator $L = D^{-\frac{1}{2}}(I - W)D^{-\frac{1}{2}}$ is the normalized Laplacian on graphs, it is a contraction on $\mathcal{L}^p(G, \mu_G)$, where $\mu_G(\{i\}) = D_{ii}$, and it is self-adjoint. This discrete setting is extremely useful in applications and widely used in a number of fields such as data analysis (e.g., clustering, learning on manifolds, parametrization of data sets), computer vision (e.g., segmentation) and computer science (e.g., network design, distributed processing). We believe our construction will have applications in all these settings. As a starting point, see, for example, [13] for an overview, and [28,29] for particular applications.

(g) One parameter groups (continuous or discrete) of dilations in $\mathbb{R}^n$, or other motion groups (e.g., the Galilei group or Heisenberg-type groups), act on square-integrable functions on these spaces (with the appropriate measures) as isometries or contractions. Continuous wavelets in some of these settings have been widely studied (see, e.g., [30,31] and references therein), but an efficient discretization of such transforms seems in many respects still an open problem.

(h) The random walk diffusion on a hypergroup (see, e.g., [32] and references therein).

Definition 15. A diffusion semigroup $\{T^t\}_{t\geqslant 0}$ is called *compact* if T^t is compact for every $t \geqslant 0$.

Definition 16. A compact diffusion semigroup $\{T^t\}$ with spectrum $\lambda_0 \geqslant \lambda_1 \geqslant \cdots \geqslant \lambda_k \geqslant \cdots \geqslant 0$ is said to have *γ-strong decay*, for some $\gamma > 0$, if there exists a constant $C > 0$ such that for every $\lambda \in (0, 1)$

$$\#\{k\colon \lambda_k \geqslant \lambda\} \leqslant C\log_2^{\gamma}\frac{1}{\lambda}.$$

Remark 17. There are classes of diffusion operators that are not self-adjoint but very important in applications. The construction we propose actually does not depend on this hypothesis, however the interpretation of many results does depend on the spectral decomposition. In a future publication we will address these and related issues in a broader framework.

Definition 18. A positive diffusion semigroup $\{T^t\}$ *acts η-locally*, for some $\eta > 0$, if for every x and every function φ δ-local around x, the function $T\varphi$ is $(\eta + \delta)$-local around x.

Definition 19. A positive diffusion semigroup $\{T^t\}$ is *expanding* if $\operatorname{supp} f \subseteq \operatorname{supp} T^t f$, for every $t > 0$ and every smooth f.

In practice the definition above can be interpreted up to a given precision ε, in the sense that the numerical support of a function φ is the set on which $\varphi \geqslant \varepsilon$.

4.2. Diffusion metrics and embeddings induced by a diffusion semigroup

In all that follows, we will assume, mainly for simplicity, that $\{T^t\}$ is a compact semigroup with γ-strong decay, that acts η-locally on X.

We refer the reader to [3,18,33] and references therein for some motivations and applications of the ideas presented in this section.

Being positive definite, T^t induces the diffusion metric

$$d^{(t)}(x, y) = \sqrt{\sum_{\lambda\in\sigma(T)} \lambda^t \big(\xi_\lambda(x) - \xi_\lambda(y)\big)^2} = \sqrt{\langle \delta_x - \delta_y, T^t(\delta_x - \delta_y)\rangle} = \big\| T^{\frac{t}{2}}\delta_x - T^{\frac{t}{2}}\delta_y \big\|_2. \tag{4.2}$$

Definition 20. The metric defined above is called the *diffusion metric* associated to T, at time t.

If the action of T^t on $\mathcal{L}^2(X, \mu)$ can be represented by a (symmetric) kernel $K^t(x, y)$, then $d^{(t)}$ can be written as

$$d^{(t)}(x, y) = \sqrt{K^t(x, x) + K^t(y, y) - 2K^t(x, y)}, \tag{4.3}$$

since the spectral decomposition of K is

$$K^t(x, y) = \sum_{\lambda\in\sigma(T)} \lambda^t \xi_\lambda(x)\xi_\lambda(y). \tag{4.4}$$

For any subset $\sigma(T)' \subseteq \sigma(T)$ we can consider the map of metric spaces

$$\begin{aligned} \Xi^{(t)}_{\sigma(T)'} : \big(X, d^{(t)}\big) &\to \big(\mathbb{R}^{|\sigma(T)'|}, d_{\text{Euc.}}\big), \\ x &\mapsto \big(\lambda^{\frac{t}{2}}\xi_\lambda(x)\big)_{\lambda\in\sigma(T)'}, \end{aligned} \tag{4.5}$$

which is in particular cases called eigenmap [33,34] and is a form of local multidimensional scaling (see also [3,4, 35]). By the definition of $d^{(t)}$, this map is an isometry when $\sigma(T)' = \sigma(T)$, and an approximation to an isometry when $\sigma(T)' \subsetneq \sigma(T)$.

If $\sigma(T)'$ is the set of the first n top eigenvalues, and if $\int_X K(x, y)\,\mathrm{d}\mu = 1$, then $\Phi_{\sigma(T)'}$ is a minimum local distortion map $(X, d^{(t)}) \to \mathbb{R}^n$, in the sense it minimizes

$$n - \operatorname{tr}\big(P_{\langle\{\xi_\lambda\}_{\lambda\in\sigma(T)'}\rangle} K^t P^*_{\langle\{\xi_\lambda\}_{\lambda\in\sigma(T)'}\rangle}\big) = \frac{1}{2}\sum_{\lambda\in\sigma(T)'} \int_X \int_X \big(\xi_\lambda(x) - \xi_\lambda(y)\big)^2 K^t(x, y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \tag{4.6}$$

among all possible maps $x \mapsto (\phi_1(x), \ldots, \phi_n(x))$ such that $\langle\phi_i, \phi_j\rangle = \delta_{ij}$. This is just a rephrasing, in our situation, of the well-known fact that the top k singular vectors span the best approximating k-dimensional subspace to the domain of a linear operator, in the $\mathcal{L}^2$ sense. See, e.g., [36] and references therein for a comparison between different dimensionality reduction techniques that can be cast in this framework.

These ideas are related to various techniques used for non-linear dimension reduction: see, for example, [3,4,18, 35,37–40] and references therein and http://www.cse.msu.edu/~lawhiu/manifold/ for a list of relevant links.

Example 21. Suppose we have a finite symmetric Markov chain $\mathcal{M}$ on (X, μ), which we can assume irreducible, with transition matrix P. We assume P is positive definite (otherwise we could consider $\frac{1}{2}(I + P)$). Let $\{\lambda_l\}_l$ be the set of

eigenvalues of P, ordered by increasing value, and $\{\xi_l\}_l$ be the corresponding set of right eigenvectors of P, which form an orthonormal basis. Then by functional calculus we have

$$P^t_{i,j} = \sum_l \lambda_l^t \xi_l(i)\xi_l(j)$$

and for an initial distribution f on X, we define

$$P^t f = \sum_l \lambda_l^t \langle f, \xi_l\rangle \xi_l. \tag{4.7}$$

Observe that if f is a probability distribution, so is $P^t f$ for every t since P is Markov. The diffusion map $\Phi_{\sigma(T)'}$ embeds the Markov chain in the Euclidean space $\mathbb{R}^{|\sigma(T)'|}$ in such a way that the diffusion metric induced by $\mathcal{M}$ on X becomes Euclidean distance in $\mathbb{R}^{|\sigma(T)'|}$.

4.3. Multiresolution scaling function spaces

We can interpret T as a *dilation operator* acting on functions in $\mathcal{L}^2(X,\mu)$ and use it to define a multiresolution structure. As in [5] and in classical wavelet theory, we may start by discretizing the diffusion semigroup $\{T^t\}$ at the increasing sequence of times

$$t_j = \sum_{l=0}^{j} 2^l = 2^{j+1} - 1 \tag{4.8}$$

for $j \geqslant 0$. Observe the choice of the factor 2 is arbitrary, and could be replaced by any factor $\lambda > 1$ here and in all that follows. In particular the factor $\delta_X > 1$, associated with geometric properties of the space, as in Theorem 6, will play an important role. Let $\{\lambda_i\}_{i\geqslant 0}$ be the spectrum of T, ordered in decreasing order, and $\{\xi_i\}_i$ the corresponding eigenvectors. We can define "low-pass" portions of the spectrum by letting

$$\sigma_j(T) = \left\{\lambda \in \sigma(T),\ \lambda^{t_j} \geqslant \varepsilon\right\}. \tag{4.9}$$

For a fixed $\varepsilon \in (0,1)$ (which we may think of as our precision), we define the (finite dimensional!) approximation spaces of band-limited functions by

$$V_j = \left\langle \left\{\xi_\lambda\colon \lambda \in \sigma(T),\ \lambda^{t_j} \geqslant \varepsilon\right\}\right\rangle = \left\langle \left\{\xi_\lambda\colon \lambda \in \sigma_j(T)\right\}\right\rangle \tag{4.10}$$

for $j \geqslant 0$. We let $V_{-1} = \mathcal{L}^2(X)$. The set of subspaces $\{V_j\}_{j\geqslant -1}$ is a multiresolution analysis in the sense that it satisfies the following properties:

(i) $V_{-1} = \mathcal{L}^2(X)$, $\lim_{j\to+\infty} V_j = \langle\{\xi_i\colon \lambda_i = 1\}\rangle$.
(ii) $V_{j+1} \subseteq V_j$ for every $j \in \mathbb{Z}$.
(iii) $\{\xi_\lambda\colon \lambda^{t_j} \geqslant \varepsilon\}$ is an orthonormal basis for V_j.

We can also define, for $j \geqslant -1$, the subspaces W_j as the orthogonal complement of V_{j+1} in V_j, so that we have the familiar relation between approximation and detail subspaces as in the classical wavelet multiresolution constructions:

$$V_j = V_{j+1} \oplus^{\perp} W_j. \tag{4.11}$$

The direct orthogonal sum

$$\mathcal{L}^2(X) = \bigoplus_{j\geqslant -1}^{\perp} W_j$$

is a wavelet decomposition of the space, induced by the diffusion semigroup, and related to the Littlewood–Paley decomposition studied in this setting by Stein [5].

Observe that when $\{T^t\}$ has γ-strong decay, then we have

$$\dim V_j \leqslant C\left(2^{-j}\log_2\frac{1}{\varepsilon}\right)^{\gamma} = C2^{-j\gamma}\log_2^{\gamma}\frac{1}{\varepsilon}. \tag{4.12}$$

```
MultiscaleDyadicOrthogonalization (Ψ, Q, J, ε):

// Ψ: a family of functions to be orthonormalized, as in Proposition 22
// Q: a family of dyadic cube on X
// J: finest dyadic scale
// ε: precision

Φ_0 ← Gram-Schmidt_ε(⋃_{k∈K_J} Ψ|_{Q_{J,k}})
l ← 1
do

    for all k ∈ K_{J+l}
        Ψ̃_{l,k} ← Ψ|_{Q_{J+l,k}} \ ⋃_{Q_{J+l-1,k'}⊆Q_{J+l,k}} Ψ|_{Q_{J+l-1,k'}}
        Φ̃_{l,k} ← Gram-Schmidt_ε(Ψ̃_{l,k})
        Φ_{l,k} ← Gram-Schmidt_ε(Φ̃_{l,k})
    end
    l ← l + 1

until Φ_l is empty.
```

Fig. 7. Pseudo-code for the MultiscaleDyadicOrthogonalization of Proposition 22.

While by definition (4.10) we already have an orthonormal basis of eigenfunctions of T for each subspace V_j (and for the subspaces W_j as well), these basis functions are in general highly non-localized, being global Fourier modes of the operator. Our aim is to build localized bases for (ε-approximations of) each of these subspaces, starting from a basis of the fine subspace V_0 and explicitly constructing a downsampling scheme that yields an orthonormal basis for V_j, $j > 0$. This is motivated by general Heisenberg principles (see, e.g., [41] for a setting similar to ours) that guarantee that eigenfunctions have a smoothness or "frequency content" or "scale" determined by the corresponding eigenvalues, and can be reconstructed by maximally localized "bump functions," or atoms, at that scale. Such ideas have many applications in numerical analysis, especially to problems motivated by physics (matrix compression [42–45], multigrid techniques [46,47], etc.).

We avoid the computation of the eigenfunctions, nevertheless the approximation spaces $\tilde{V}_j$ that we build will be ε-approximations of the true V_j's.

4.4. Orthogonalization and downsampling

A crucial ingredient for the construction of the scaling functions will be a scheme that starts with a set of "bump" functions, and constructs a set of well-localized orthonormal functions spanning the same subspace, up to a given precision. In the next section, this will be applied to families of the form $\{T^t\delta_x\}$, where T is a local semigroup, $t+1$ a dyadic power, and δ_x either a set of Dirac δ-functions (in the discrete setting), or a mollification of those (in the continuous case).

In this section we present two ways for orthogonalizing a set of "bump" functions. The first one is based on the following proposition.

Proposition 22 *(Multiscale dyadic orthogonalization). Let (X, ρ, μ) be a space of homogeneous type,* $\operatorname{diam}(X) < +\infty$ *and $\mathcal{Q} = \{\{Q_{j,k}\}_{k\in\mathcal{K}_j}\}_{j\in\mathbb{Z}}$ a family of dyadic cubes, and $\delta_X > 1$, $\eta > 0$, c_1, $c_2 > 0$ as in Theorem 6. Assume $X \in \mathcal{Q}$, more precisely $X = Q_{J+j_X,k}$ for some $j_X \leqslant \log_{\delta_X}(c_1^{-1}\delta_X^{-J}\operatorname{diam}(X))$. Fix $\varepsilon > 0$. Let $\Psi = \{\psi_x\}_{x\in\Gamma}$ be a $\alpha\delta_X^J$-local family, $\alpha \leqslant c_1$, with center set Γ. Suppose Ψ is "uniformly locally finite-dimensional" in the sense that there exist c'_ε, $c''_\varepsilon > 0$ such that for all $k \in \mathcal{K}_J$,*

$$\dim_{c_1\delta_X^J\mu(X)^{-1}\varepsilon}(\Psi|_{Q_{J,k}}) \leqslant c'_\varepsilon\mu(Q_{J,k}),$$

and for all $l \geqslant 0, k \in \mathcal{K}_{J+l}$,

$$\dim_{c_1\delta_X^{J+l}(2\mu(X))^{-1}\varepsilon}\left(\{\psi_x\}_{x\in\Gamma\cap\partial_{\alpha\delta_X^{-l}}Q_{J+l,k}}\right) \leqslant c''_\varepsilon\left(\alpha\delta_X^{-l}\right)^\eta \mu(Q_{J+l,k}).$$

Then there exists an orthonormal basis

$$\Phi = \{\Phi_l\}_{l=0,\dots,L} = \left\{\left\{\{\varphi_{l,k,i}\}_{i\in\mathcal{I}_{l,k}}\right\}_{k\in\mathcal{K}_{J+l}}\right\}_{l=0,\dots,L},$$

with $L \leqslant j_X$ *such that*

(i) $\langle\Psi\rangle \subseteq \langle\Phi\rangle_{(j_X+1)\varepsilon}$.

(ii) *For* $l = 0,\dots,L$, *all* $k \in \mathcal{K}_{J+l}$ *and all* $i \in \mathcal{I}_{l,k}$, $\operatorname{supp}\varphi_{l,k,i} \subseteq Q_{J+l,k}$. *In particular,* $\varphi_{l,k,i}$ *is* $c_2\delta_X^{J+l}$*-local.*

(iii) *For* $l > 0$,

$$\left\langle \bigcup_{k\in\mathcal{K}_{J+l}} \Psi|_{Q_{J+l,k}} \right\rangle \subseteq \left\langle \bigcup_{l'=0}^{l} \Phi_l \right\rangle_{(l+1)\varepsilon},$$

and $\#\mathcal{I}_{l,k} \leqslant c''_\varepsilon(\alpha\min\{\delta_X^{-(l-1)}, 1\})^\eta \mu(Q_{J+l,k})$.

Before presenting the proof, we discuss qualitatively the contents of the proposition. Qualitatively, Ψ is a local family of functions "well-spread" on X, and we orthonormalize it in the following way. At the finest scale J, we consider $\Psi|_{Q_{J,k}}$ for every $k \in \mathcal{K}_J$, and orthonormalize this family, to get Φ_0. At the next scale, we consider, for each $k \in \mathcal{K}_{J+1}$, $\Psi|_{Q_{J+1,k}}$ minus the subset of Ψ already considered, which is $\Psi|_{Q_{J,k'}}$ for $Q_{J,k'} \subseteq Q_{J+1,k}$. Hence the new subset of Ψ to be considered consists of functions in Ψ close to the boundary of $Q_{J+1,k}$: we orthonormalize them to $\langle\Phi_0\rangle$, and orthonormalize what is left to itself, obtaining Φ_1. We proceed in this fashion, at layer l adding a subset of Ψ not considered yet, consisting of function close to the boundary of $Q_{J+l,k}$ for each $k \in \mathcal{K}_{J+l}$. We refer the reader Fig. 7 for pseudo-code implementing the construction in the proposition.

Observe that the condition of "uniform local finite-dimensionality" imposed on Ψ is quite natural. This condition says, roughly speaking, that the set of functions in Ψ which has support in a given set, ε-spans a subspace of ε-dimension at most proportional to the measure of that set. However, we do not need this condition for all sets, but for very special families of sets. First of all it should hold for all the dyadic cubes at the finest scale considered, $Q_{J,k}$, $k \in \mathcal{K}_J$. This is the finest scale considered in the proposition because it is comparable with the size of the supports of the functions in Ψ. Secondly, it should hold for the boundaries of the cubes at all scales. The condition imposed is exactly consistent with the estimate on the volume of this boundaries as given by Theorem 6.

Proof. The construction of the orthonormal basis Φ is done at different scales and locations, corresponding to different dyadic cubes. We will need to orthonormalize various sets of functions $\{v_i\}$, to obtain an orthonormal basis for their ε-span, that contains $\dim_\varepsilon(\{v_i\})$ elements: we will call such a process ε-orthogonalization. For example one can use a modified Gram–Schmidt algorithm described in Section 5 or the orthonormalization method suggested in Proposition 25.

The first "layer" $l = 0$ is constructed as follows: for each dyadic cube $Q_{J,k}$, $k \in \mathcal{K}_J$, consider $\Psi|_{Q_{J,k}}$ and $c_1\delta_X^J\mu(X)^{-1}\varepsilon$-orthonormalize ($c_1$ is as in Theorem 6) this set of functions to obtain $\Phi_0 := \{\varphi_{0,k,i}\}_{i\in\mathcal{I}_{0,k}}$. Property (ii) is satisfied by construction and by the assumption on uniform local finiteness. Observe also that

$$\left\langle \bigcup_{k\in\mathcal{K}_J} \Psi|_{Q_{J,k}} \right\rangle \subseteq \langle\Phi_0\rangle_\varepsilon,$$

since the restriction of a function in the subspace on the right-hand side to every dyadic cube $Q_{J,k}$ is $c_1\delta_X^J\mu(X)^{-1}\varepsilon$-approximated in $\langle\Phi_0\rangle$, and there are at most $(c_1\delta_X^J)^{-1}\mu(X)$ such cubes, by volume considerations.

For the second "layer," $l = 1$, consider, for $k \in \mathcal{K}_{J+1}$,

$$\begin{aligned}\tilde\Psi_{1,k} &:= \Psi|_{Q_{J+1,k}} \setminus \bigcup_{Q_{J,k'}\subseteq Q_{J+1,k}} \Psi|_{Q_{J,k'}} \\ &= \{\psi\in\Psi\colon \operatorname{supp}\psi \subseteq Q_{J+1,k} \text{ but } \operatorname{supp}\psi \nsubseteq Q_{J,k'} \text{ for all } Q_{J,k'} \subseteq Q_{J+1,k}\}\end{aligned}$$

$$\subseteq \bigcup_{Q_{J,k'}\subseteq Q_{J+1,k}} \{\psi_x \in \Psi\colon x \in \partial_{\alpha\delta_X^J\delta_X^{-J}} Q_{J,k'}\}.$$

The last inclusion holds because $\operatorname{supp}\psi_x \subseteq Q_{J+1,k}$ implies $x \in Q_{J,k'}$ for some $Q_{J,k'} \subseteq Q_{J+1,k}$, and $\operatorname{supp}\psi \nsubseteq Q_{J,k'}$ forces $x \in \partial_{\alpha\delta_X^J} Q_{J,k'}$, for otherwise $\operatorname{supp}\psi_x \subseteq B_{\alpha\delta_X^J}(x) \subseteq Q_{J,k'}$ (because ψ_x is $\alpha\delta_X^J$-local). From the assumptions we deduce that

$$\dim_{c_1\delta_X^J(2\mu(X))^{-1}\varepsilon}(\tilde{\Psi}_{1,k}) \leqslant c_\varepsilon''\big(\alpha\delta_X^0\big)^\eta \mu(Q_{J+1,k}).$$

We $c_1\delta_X^{J+1}(2\mu(X))^{-1}\varepsilon$-orthonormalize $\tilde{\Psi}_{1,k}$ to the functions in Φ_0, obtaining $\tilde{\Phi}_{1,k}$, and then let $\Phi_{1,k} := \{\varphi_{1,k,i}\}_{i\in\mathcal{I}_{1,k}}$ be the result of $c_1\delta_X^{J+1}(2\mu(X))^{-1}\varepsilon$-orthonormalizing $\tilde{\Phi}_{1,k}$ for every $k \in \mathcal{K}_{J+1}$. Observe that each function in $\Phi_{1,k}$ has support in the dyadic cube $Q_{J+1,k}$, since so do the functions in $\tilde{\Psi}_{1,k}$ and $\tilde{\Phi}_{1,k}$. This proves property (ii). To see that (iii) also holds, observe that $\langle\bigcup_{k\in\mathcal{K}_{J+1}} \tilde{\Psi}_{1,k}\rangle \subseteq \langle\Phi_1\rangle_\varepsilon$; in fact

$$\left\langle \bigcup_{k\in\mathcal{K}_J} \tilde{\Psi}_{0,k} \cup \bigcup_{k\in\mathcal{K}_{J+1}} \tilde{\Psi}_{1,k} \right\rangle \subseteq \langle\Phi_0\rangle_\varepsilon + \langle\Phi_1\rangle_\varepsilon \subseteq \langle\Phi_0\cup\Phi_1\rangle_{2\varepsilon}.$$

Finally, $\#\mathcal{I}_{1,k} \leqslant \dim_\varepsilon \tilde{\Psi}_{1,k} \leqslant c_\varepsilon''(\alpha\delta_X^0)^\eta\mu(Q_{J+1,k})$.

Proceeding in this fashion, at "layer" $j \geqslant 1$ we consider, for $k \in \mathcal{K}_{J+l}$,

$$\begin{aligned}\tilde{\Psi}_{l,k} &:= \Psi|_{Q_{J+l,k}} \setminus \bigcup_{Q_{J+l-1,k'}\subseteq Q_{J+l,k}} \Psi|_{Q_{J+l-1,k}} \\ &= \{\psi\in\Psi\colon \operatorname{supp}\psi \subseteq Q_{J+l,k} \text{ but } \operatorname{supp}\psi \nsubseteq Q_{J+l-1,k'} \text{ for all } Q_{J+l-1,k'} \subseteq Q_{J+l,k}\} \\ &\subseteq \bigcup_{Q_{J+l-1,k'}\subseteq Q_{J+l,k}} \{\psi_x\in\Psi\colon x\in\partial_{\alpha\delta_X^J\delta_X^{-(J+l-1)}} Q_{J+l-1,k}\}.\end{aligned}$$

As above, the last step follows because $\operatorname{supp}\psi_x \subseteq Q_{J+l,k}$ implies $x \in Q_{J+l-1,k'}$ for some $Q_{J+l-1,k'} \subseteq Q_{J+l,k}$, and $\operatorname{supp}\psi_x \nsubseteq Q_{J+l-1,k'}$ forces $x \in \partial_{\alpha\delta_X^J\delta_X^{-(J+l-1)}} Q_{J+l-1,k'}$, for otherwise $\operatorname{supp}\psi_x \subseteq B_{\alpha\delta_X^J}(x) \subseteq Q_{J+l-1,k'}$ (because ψ_x is $\alpha\delta_X^J$-local). By assumption it follows that

$$\dim_{c_1\delta_X^{J+l}(2\mu(X))^{-1}\varepsilon}(\tilde{\Psi}_{l,k}) \leqslant c_\varepsilon''\big(\alpha\delta_X^{-l}\big)^\eta\mu(Q_{J+l-1,k}).$$

We $c_1\delta_X^{J+l}(2\mu(X))^{-1}\varepsilon$-orthonormalize $\tilde{\Psi}_{l,k}$ to the functions in $\Phi_0,\dots,\Phi_{l-1}$, obtaining $\tilde{\Phi}_{l,k}$, and then let $\Phi_{l,k} := \{\varphi_{l,k,i}\}_{i\in\mathcal{I}_{l,k}}$ be the result of $c_1\delta_X^{J+l}(2\mu(X))^{-1}\varepsilon$-orthonormalizing $\tilde{\Phi}_{l,k}$ for every $k \in \mathcal{K}_{J+l}$. As above, each function in $\Phi_{l,k}$ has support in the dyadic cube $Q_{J+l,k}$, since so do the functions in $\tilde{\Psi}_{l,k}$ and $\tilde{\Phi}_{l,k}$. This proves property (ii). To see that (iii) also holds, first observe that

$$\left\langle \bigcup_{l'=0}^{l} \bigcup_{k\in\mathcal{K}_{J+l'}} \tilde{\Psi}_{l',k} \right\rangle \subseteq \bigoplus_{l'=0}^{l} \langle\Phi_{l'}\rangle_\varepsilon \subseteq \left\langle \bigcup_{l'=0}^{l} \Phi_{l'} \right\rangle_{(l+1)\varepsilon},$$

and secondly, $\#\mathcal{I}_{l,k} \leqslant \dim_\varepsilon \tilde{\Psi}_{l,k} \leqslant c_\varepsilon''(\alpha\delta_X^{-(l-1)})^\eta\mu(Q_{J+l,k})$.

We stop if $\#\mathcal{I}_{l,k} = 0$. Since $X \in \mathcal{Q}$, eventually, for l large enough, X is the only dyadic cube $Q_{J+l,k}$, and we simply finish by orthonormalizing the subset of functions in Ψ which have not been already considered, which finishes the construction. This happens at most at scale $L \leqslant \log_{\delta_X}(c_1^{-1}\delta_X^{-J}\operatorname{diam} X)$. □

Corollary 23. *Let everything be as in Proposition 22. Furthermore assume that Ψ is also "well-distributed" in the sense that there exist $c', c'' > 0$ such that for all $k \in \mathcal{K}_J$*

$$\#(\Psi|_{Q_{J,k}}) \leqslant c'\mu(Q_{J,k}),$$

and for all $l \geqslant 0$ and $k \in \mathcal{K}_{J+l}$,

$$\#\{\Gamma\cap\partial_{\alpha\delta_X^{-l}} Q_{J+l,k}\} \leqslant c''\big(\alpha\delta_X^{-l}\big)^\eta\mu(Q_{J+l,k}).$$

Then the cost of computing $\{\{\varphi_{l,k,i}\}_{i\in\mathcal{I}_{l,k}}\}_{k\in\mathcal{K}_{J+l}}$ *is upper-bounded by*

$$\sim (c_\varepsilon'')^2\alpha^{2\eta}\delta_X^{2l(1-\eta)+2(J+\eta)}\mu(X) + c'c_\varepsilon''\big(\alpha\delta_X^{-(l-1)}\big)^\eta\delta_X^{(2J+l)}\cdot\big(1+(\alpha\delta_X)^\eta c_\eta\delta_X^{(1-\eta)l}\big)\mu(X),$$

where c_η *is a universal constant depending only on* η.

Proof. First observe that from the proof of Proposition 22 it follows that, for $l \geqslant 1$,

$$\#\tilde\Psi_{l,k} \leqslant c'\big(\alpha\delta_X^{-(l-1)}\big)^\eta\mu(Q_{J+l-1,k}).$$

To estimate the computational cost, first recall that the cost of orthonormalizing k vectors to m vectors in n dimensions is in general equal to kmn. When $l=0$ it is easy to see the cost of computing Φ_0 is proportional to $\mu(X)$. For each layer $l>0$, the cost of computing Φ_l can be calculated as follows. First we need to orthonormalize the functions to $\tilde\Psi_{l,k}$, for all $k\in\mathcal{K}_{J+l}$ to the functions already orthonormalized at the previous layers. The result of this operation is $\tilde\Phi_{l,k}$, and the cost is

$$\begin{aligned}
&\sum_{k\in\mathcal{K}_{J+l}} \underbrace{\#\tilde\Psi_{l,k}}_{\#\text{fcns. to orthonormalize}}\cdot\underbrace{\sum_{l'=0}^{l-1}\sum_{k':Q_{J+l',k'}\subseteq Q_{J+l,k}}\#\mathcal{I}_{l',k'}\mu(Q+l',k')}_{\text{cost of projecting off previous layers}}\\
&\quad\leqslant \sum_{k\in\mathcal{K}_{J+l}} c'\big(\alpha\delta_X^{-(l-1)}\big)^\eta\mu(Q_{J+l,k})\\
&\qquad\times\left(\sum_{k':Q_{J,k'}\subseteq Q_{J+l,k}} c_\varepsilon''\mu(Q_{J,k'})^2 + \sum_{l'=1}^{l-1}\sum_{k':Q_{J+l',k'}\subseteq Q_{J+l,k}} c_\varepsilon''\big(\alpha\delta_X^{-(l'-1)}\big)^\eta\mu(Q_{J+l',k'})^2\right)\\
&\quad\sim c'\big(\alpha\delta_X^{-(l-1)}\big)^\eta\mu(X)\cdot\left(c_\varepsilon''\delta_X^{2J+l} + c_\varepsilon''\alpha^\eta\delta_X^{\eta+2J+l}\sum_{l'=1}^{l-1}\delta_X^{l'(1-\eta)}\right)\\
&\quad\sim c'c_\varepsilon''\big(\alpha\delta_X^{-(l-1)}\big)^\eta\mu(X)\delta_X^{2J+l}\cdot\big(1+(\alpha\delta_X)^\eta c_\eta\delta_X^{(1-\eta)l}\big).
\end{aligned}$$

Then we compute the cost of the orthonormalization of $\tilde\Phi_{l,k}$ for all $k\in\mathcal{K}_{J+l}$ to obtain $\Phi_{l,k}$:

$$\sum_{k\in\mathcal{K}_{J+l}}\dim_\varepsilon\tilde\Psi_{l,k}^2\mu(Q_{J+l,k}) \leqslant \big(c_\varepsilon''\big(\alpha\delta_X^{-(l-1)}\big)^\eta\big)^2\big(\delta_X^{J+l}\big)^2\mu(X) \leqslant (c_\varepsilon'')^2\alpha^{2\eta}\delta_X^{2l(1-\eta)+2(J+\eta)}\mu(X). \qquad \square$$

A result in [6] suggests another way of orthonormalizing families of localized functions. This second orthonormalization technique guarantees asymptotic exponential decay on the orthonormal functions we build. We will need the following definitions.

Definition 24. A matrix $(B)_{(j,k)\in J\times J}$ is called η-accretive [6], if there exists an $\eta>0$ such that for every $\xi\in l^2(J)$ we have

$$\Re\mathrm{e}\sum_j\sum_k B_{jk}\xi_j\bar\xi_k \geqslant \eta\sum_j\|\xi\|_{l^2(J)}^2.$$

Proposition 25. *Let* $\Psi=\{\psi_j\}_{j\in\Gamma}$ *be a Riesz basis of some Hilbert space* $\mathcal{H}$, Γ *at most countable. Let* ρ *be a metric on* Γ, *for which there exist* $\nu_\Gamma, E_\Gamma>0$ *such that*

$$\sum_{j\in\Gamma} e^{-\nu_\Gamma\rho(i,j)} < E_\Gamma \tag{4.13}$$

for every $i\in\Gamma$. *Suppose the Gramian matrix* $G_{ij}=\langle\psi_i,\psi_j\rangle$ *is* η*-accretive and there exist* $C>0, \alpha>\nu_X$ *such that for all* $i,j\in\Gamma$

$$|G_{i,j}| \leqslant Ce^{-\alpha\rho(i,j)}.$$

Then there exist $C', \alpha' > 0$ and an orthonormal basis $\{\varphi\}_{j\in\Gamma}$ such that

$$\varphi_j = \sum_{k\in\Gamma} \beta(j,k)\psi_k$$

with

$$\big|\beta(j,k)\big| \leqslant C' e^{-\alpha'\rho(j,k)}.$$

Proof. The proof proceeds exactly as in [6, Proposition 3 in Section 11.4], with the triangle inequality on $\mathbb{Z}^n$ replaced by the quasi-triangle inequality for the metric ρ. □

Condition (4.13) is automatically satisfied on spaces of homogeneous type, up to a change to a topologically equivalent metric:

Lemma 26. *Let (X,ρ,μ) be a space of homogeneous type. There exists a metric ρ', topologically equivalent to ρ, such that for any $\nu > 0$, there exists a constant $E = E(\nu) > 0$ such that*

$$\int_X e^{-\nu\rho'(x,y)}\,\mathrm{d}\mu(y) < E. \tag{4.14}$$

Proof. Let ρ' be a metric, topologically equivalent to ρ, such that $\mu(\{y\in X\colon \rho'(x,y) < r\}) \leqslant C_X r^{\mathrm{d}}$ for some $\mathrm{d} > 0$. The existence of such a ρ' is proved in [48]. We have:

$$\begin{aligned}\int_X e^{-\nu\rho'(x,y)}\,\mathrm{d}\mu(y) &= \int_{\{y:\rho'(x,y)\leqslant 1\}} e^{-\nu\rho'(x,y)}\,\mathrm{d}\mu(y) + \sum_{j\geqslant 0}\int_{\{y:\rho'(x,y)\in(2^j,2^{j+1}]\}} e^{-\nu\rho'(x,y)}\,\mathrm{d}\mu(y)\\ &\leqslant \mu\big(B_1(x)\big) + \sum_{j\geqslant 0} e^{-\nu 2^j} 2^{(j+1)\mathrm{d}} \leqslant E(\nu). \qquad \square\end{aligned}$$

One can apply Proposition 25 for orthogonalizing the set of functions $\tilde\Psi_l$ and $\tilde\Phi_l$ in Proposition 22. This may lead to a lower computational complexity for the multiscale orthogonalization scheme in Proposition 22, possibly down to order $\mu(X)\log_{\delta_X}\mu(X)$, at least when $\eta = 1$. This will be reported in a separate work.

4.5. Construction of the multi-resolution analysis

We can use the orthogonalization procedures in Proposition 22 or Proposition 25 to construct orthonormal scaling functions for the subspaces V_j and generate multiscale bases of orthonormal scaling functions. Other orthogonalization schemes are viable as well. We start with an application of Proposition 22.

Theorem 27. *Suppose we are given*:

(1) *A space of homogeneous type (X,ρ,μ), with* $\operatorname{diam} X < +\infty$*, a family of dyadic cubes $\mathcal{Q}$, and $X = Q_{J_0+j_X,k}$ for some $j_X \geqslant 0$ and $k \in \mathcal{K}_{J_0+j_X}$. Let $\delta_X > 1$, $\eta > 0$ be as in Theorem 6.*
(2) *A Markovian semigroup $\{T^t\}_{t\geqslant 0}$, that acts $\delta_X^{J_0}$-locally on X.*
(3) *A precision $\varepsilon > 0$.*
(4) *An orthonormal basis $\Phi_0 = \{\varphi_x\}_{x\in\Gamma}$, $\alpha_0\delta_X^{J_0}$-local with center set Γ, and such that $V_0 \subseteq \langle\Phi_0\rangle_\varepsilon$.*

Assume there exist constants $c'_\varepsilon, c''_\varepsilon$ such that if $\varepsilon' = c_1\delta_X^J(2\mu(X))^{-1}\varepsilon$, then for all $l \geqslant 0$ and $k \in \mathcal{K}_{J+l}$,

$$\dim_{\varepsilon'}(\Phi_0|_{Q_{J+l,k}}) \leqslant c'_\varepsilon \mu(Q_{J+l,k})$$

and

$$\dim_{\varepsilon'}\big(\{\varphi_x\colon x\in\Gamma\cap\partial_{\alpha_0\delta_X^{-l}}Q_{J+l,k}\}\big) \leqslant c''_\varepsilon\big(\alpha\delta_X^{-l}\big)^\eta\mu(Q_{J+l,k}).$$

Then there exists a sequence of orthonormal scaling function bases $\{\Phi_j\}_{j=1,\dots,j_X}$,

$$\Phi_j := \left\{\left\{\{\varphi_{j,l,k,i}\}_{i\in\mathcal{I}(j,l,k)}\right\}_{k\in\mathcal{K}_{J_0+j+l}}\right\}_{l=0,\dots,j_X-j}$$

with the following properties:

(i) $V_j = \langle T^{t_j}\Phi_0\rangle \subseteq \langle\Phi_j\rangle_{(j_X-j)\varepsilon}$, *where* t_j *is as in* (4.8).
(ii) $\operatorname{supp}\varphi_{j,l,k,i} \subseteq Q_{J_0+j+l,k}$ *for all* $l = 0,\dots,j_x - j$, $k \in \mathcal{K}_{J_0+j+l}$, $i \in \mathcal{I}(j,l,k)$.
(iii) $\#\mathcal{I}(j,l,k) \leqslant c''_\varepsilon(1+\delta_X^{-1}(\alpha_0-1)\min\{\delta_X^{-(l-1)},1\})^\eta \mu(Q_{J_0+j+l,k})$.

Proof. We let, for every $j \geqslant 0$, $\tilde\Phi_j = T^{\delta_X^j-1}\Phi_0$. Since T is a compact contraction and V_0 is an ε-span of Φ_0, V_j is an ε-span of $\tilde\Phi_j$. We would like to apply Proposition 22 to $\tilde\Phi_j$ to obtain an orthonormal basis for V_j, up to ε. We need to check that the hypothesis of the proposition are satisfied by $\tilde\Phi_j$, for $J_j = J_0 + j$, and $\alpha_j = 1 + \delta_X^{-j}(\alpha_0 - 1) < 1$. The ingredients needed to see this are:

(P1) T and its powers are contractions, so $\dim_\varepsilon T(S) \leqslant \dim_\varepsilon S$ for any subspace S.
(P2) T is $\delta_X^{J_0}$-local, so $\tilde\Phi_j$ is $\alpha_j\delta_X^{J_j}$-local.
(P3) T and its powers preserve any chosen center(Φ_0).

For example we want to check that

$$\dim_\varepsilon\left(T^{\delta_X^j-1}\Phi_0|_{Q_{J,k}}\right) < c'_\varepsilon\mu(Q_{J,k}).$$

We have

$$\dim_\varepsilon(\Psi|_{Q_{J,k}}) = \dim_\varepsilon\left(T^{\delta_X^j-1}\Phi_0|_{Q_{J,k}}\right) \leqslant \dim_\varepsilon\left(T^{\delta_X^j-1}(\Phi_0|_{Q_{J,k}})\right) \leqslant \dim_\varepsilon(\Phi_0|_{Q_{J,k}}) \leqslant c'_\varepsilon\mu(Q_{J,k}).$$

The first inequality follows from (P2) and (P3), the second from (P1), and the last from the hypotheses on Φ_0. In an analogous way one checks that

$$\dim_\varepsilon\left(\{\Psi_x\}_{x\in\Gamma\cap\partial_{\alpha_j\delta_X^{-l}}Q_{J+l,k}}\right) \leqslant c''_\varepsilon\left(\alpha_j\delta_X^{-l}\right)^\eta\mu(Q_{J+l,k}).$$

The properties of Φ_j all follow from Proposition 22. □

A. Nahmod shows [11,41] that for large classes of diffusion-like operators on spaces of homogeneous type one can estimate precisely, in an asymptotic sense, the ε-dimensions of the eigenspaces and of the restriction of the eigenspaces to dyadic cubes and lattices. Her results suggest that in those settings one may assume, in the theorem above, that there exist constants $c'_\varepsilon, c''_\varepsilon$ such that if $\varepsilon' = c_1\delta_X^J(2\mu(X))^{-1}\varepsilon$, then for all $l \geqslant 0$ and $k \in \mathcal{K}_{J+l}$,

$$\dim_{\varepsilon'}(\Phi_0|_{Q_{J+l,k}}) \leqslant c'_\varepsilon\delta_X^{-(J+l)}\mu(Q_{J+l,k})$$

and

$$\dim_{\varepsilon'}\left(\{\varphi_x\colon x\in\Gamma\cap\partial_{\alpha_0\delta_X^{-l}}Q_{J+l,k}\}\right) \leqslant c''_\varepsilon\delta_X^{-(J+l)}\left(\alpha\delta_X^{-l}\right)^\eta\mu(Q_{J+l,k}).$$

These connections are currently being investigated and we will report on them in a future work.

The following theorem is an application of Proposition 25. To simplify the notation, we introduce the following definition.

Definition 28. A smooth (Lipschitz) function f on a metric measure space (X,ρ,μ) has (C,α) exponential decay from the center $x_0 \in X$, where C and α are two positive constants, if $|f(y)| \leqslant Ce^{-\alpha\rho(x_0,y)}$ for every $y \in X$.

Theorem 29. *Suppose we are given*:

(1) *A space of homogeneous type* (X,ρ,μ).

(2) *A Markovian semigroup* $\{T^t\}_{t\geqslant 0}$*, that maps functions with exponential decay into functions with exponential decay. More precisely, there exists* $C_T(t) \in (0, C_T^*]$ *and* $\alpha_T(t) \in (0, \alpha_T^*]$ *such that for any smooth (Lipschitz)* f *with* (C, α) *exponential decay from* x_0*,* $\alpha > \alpha_T^*$*,* $T^t f$ *has* $(C_T(t)C, \alpha - \alpha_T(t))$ *exponential decay from* x_0*.*

(3) *A precision* $\varepsilon > 0$*.*

(4) *A set* $\Gamma \subseteq X$ *and constants* $\nu, E > 0$ *such that*

$$\sum_{j\in\Gamma} e^{-\nu\rho(i,j)} < E \tag{4.15}$$

for all $i \in \Gamma$*.*

(5) *An orthonormal basis* $\Phi_0 = \{\varphi_x\}_{x\in\Gamma}$*, such that* φ_x *has* (C_0, α_0) *exponential decay from* x*, for some* $\alpha_0 > \nu + \alpha_T^*$*, and for every* $x \in \Gamma$ *(*C_0, α_0 *independent of* $x \in \Gamma$*). Furthermore, assume* $V_0 \subseteq \langle\Phi_0\rangle_\varepsilon$*.*

Then there exists a family of orthonormal scaling function bases $\{\Phi_j\}_{j\geqslant 1}$*, such that* $V_j = \langle T^{t_j}\Phi_0\rangle \subseteq \langle\Phi_j\rangle_\varepsilon$*. The scaling functions in* $\Phi_j := \{\varphi_{j,i}\}_{i\in\Gamma_j}$*,* $\Gamma_j \subseteq \Gamma$*, have exponential decay. More precisely,* $\varphi_{j,i}$ *has exponential decay from* i*.*

Proof. Let $\tilde{\Phi}_j = (T^{\delta_X^j})_\varepsilon\Phi_0$ (see Notation 11). We would like to apply Proposition 25 to $\tilde{\Phi}_j$ in order to obtain Φ_j. So let us check that the hypotheses of the proposition are satisfied. Clearly $\tilde{\Phi}_j$ is accretive, since it is the image of an orthonormal basis under the strictly positive operator $(T^{\delta_X^j})_\varepsilon$. Secondly, it is easy to see that if f_1 is (C_1, α) exponentially decaying from x_1 and f_2 is (C_2, α) exponentially decaying from x_2, then

$$\langle f_1, f_2\rangle \leqslant CC_1C_2e^{-\alpha\rho(x_1,x_2)},$$

where C depends only on (X, ρ, μ). This implies that the Gramian of $\tilde{\Phi}_j$ satisfies the hypothesis in Proposition 25, with $\alpha = \alpha_0 - \alpha_T(t) \geqslant \alpha_0 - \alpha_T^* > \nu$. Hence we can apply Proposition 25, obtaining an orthonormal basis Φ_j, such that $V_j \subseteq \langle\Phi_j\rangle_\varepsilon$. Each function in Φ_j is exponentially decaying, because of Proposition 25 and the well-separateness hypothesis on Γ. □

Corollary 30. *Let everything be as in Theorem* 29*, except that instead of* (4) *we assume that* Γ *is a lattice in* X*. Then* ρ *can be replaced by a topologically equivalent metric with respect to which the conclusion of the theorem holds.*

Proof. If Γ is a lattice, then $(\Gamma, \rho|_\Gamma)$ is a space of homogeneous type when endowed with the counting measure, and hence condition (4.15) is satisfied for any $\nu > 0$, after replacing the metric $\rho|_\Gamma$ with a topologically equivalent one, for some E dependent on ν. □

4.6. Multiscale construction with compression

The construction in the theorem is not computationally efficient because it requires the computation of T^{2^j} for large j, which is not a sparse matrix. Moreover it does not emphasize the multiresolution structure, in particular the fact that $V_{j+1} \subseteq_\varepsilon V_j$, and consequently that the scaling functions in V_{j+1} can be encoded as linear combinations of scaling functions in V_j. We want to take advantage of the fact that T^{2^j} is low rank by assumption, in order to speed up this computation, and explicitly reveal the multiresolution structure and use it for efficient encoding of the scaling functions and wavelets. We reproduce in Fig. 8 the diagram for the multiscale construction.

To simplify the notation, let us assume, without loss of generality, that the finest scale is $J_0 = 0$. We start from a given orthonormal basis Φ_0 which is 1-local, and ε-dense in V_0. Assume that T is 1-local as well.

We apply T to the basis functions in Φ_0, to obtain $\tilde{\Phi}_1$. Observe that $\tilde{\Phi}_1$ is 2-local. By definition $\langle\tilde{\Phi}_1\rangle = \mathrm{Ran}_\varepsilon(T) = V_1$. Observe that by spectral theory, we know that

$$\dim_\varepsilon(\tilde{\Phi}_1) = \dim_\varepsilon(T\Phi_0) = \dim \mathrm{Ran}_\varepsilon(T) = \#\big\{\lambda \in \sigma(T)\colon\ |\lambda| \geqslant \varepsilon\big\}.$$

We orthogonalize $\tilde{\Phi}_1$ by the method in Proposition 22 or Proposition 25 (if applicable): this yields a factorization $[T]_{\Phi_0}^{\Phi_0} = M_0\Phi_1$, where the columns of Φ_1 are a basis, up to precision, for the range of $[T]_{\Phi_0}^{\Phi_0}$, written on the basis Φ_0,

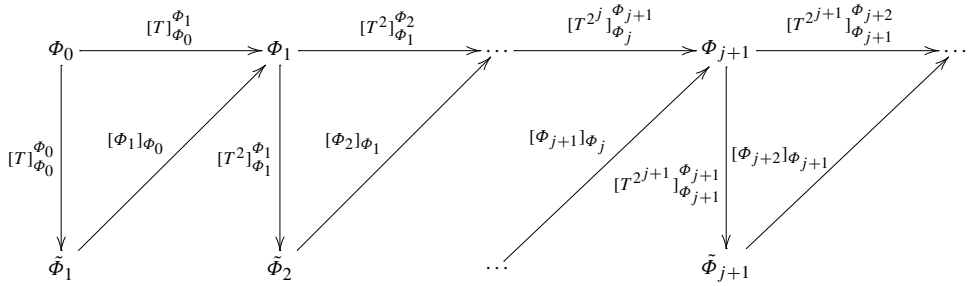

Fig. 8. Diagram for downsampling, orthogonalization and operator compression. (All triangles are commutative by construction.)

and M_0 is the matrix representing T on Φ_0 in the domain and Φ_1 in the range, i.e., $M_0 = [T]^{\Phi_1}_{\Phi_0}$. We now represent T^2 on Φ_1. This matrix is given by

$$\big[T^2\big]^{\Phi_1}_{\Phi_1} = [\Phi_1]_{\Phi_0}\big[T^2\big]^{\Phi_0}_{\Phi_0}\big([\Phi_1]_{\Phi_0}\big)^T = [T]^{\Phi_1}_{\Phi_0}\big([T]^{\Phi_1}_{\Phi_0}\big)^* = M_0 M_0^*,$$

where the second equality follows from the self-adjointness of T. This is the only place where we could use the self-adjointness, which has the advantage that $[T^2]^{\Phi_1}_{\Phi_1}$ computed in this way is automatically symmetric to precision.

If the matrix M_0 in this product is sparse, so is $[T^2]^{\Phi_1}_{\Phi_1}$. Observe that $[T^2]^{\Phi_1}_{\Phi_1}$ is of size $\dim V_1 \times \dim V_1$. We are representing T^2 as an operator acting on an ε-numerical range of T, for which we have constructed the orthogonal and well-localized basis Φ_1. We proceed now by applying $[T^2]^{\Phi_1}_{\Phi_1}$ to Φ_1, and so on.

By induction, at scale j we have a basis Φ_j, and T^{2^j} represented on this basis by a matrix $[T^{2^j}]^{\Phi_j}_{\Phi_j}$. We apply this matrix to Φ_j, obtaining a set of bump functions $\tilde{\Phi}_{j+1}$, which we orthonormalize with the algorithm of Proposition 22, and obtain a basis Φ_{j+1} of scaling functions at the next scale, written on the basis Φ_j. We then represent $(T^{2^j})^2$ on Φ_{j+1}:

$$\big[\big(T^{2^j}\big)^2\big]^{\Phi_{j+1}}_{\Phi_{j+1}} = [\Phi_{j+1}]_{\Phi_j}\big[T^2\big]^{\Phi_j}_{\Phi_j}\big([\Phi_{j+1}]_{\Phi_j}\big)^T = M_j M_j^T, \tag{4.16}$$

where the last equality holds if $T = T^*$ and $M_j = [T]^{\Phi_{j+1}}_{\Phi_j}$ is the analogue of a classical low-pass filter.

Remark 31. We could have used the orthogonalization scheme of Proposition 25 instead of the one of Proposition 22.

We can interpret this construction as having "downsampled" the subspace V_j at the critical "rate" for representing up to the specified precision the action of T^{2^j} on it. This is related to the Heisenberg principle and sampling theory for these operators [11], which implies that low-frequency eigenfunctions are smooth and non-localized, and can be synthesized by coarse (depending on the scale) "bump" functions, while higher-order eigenfunctions required finer and finer "bump" functions. The scaling functions we construct are equivalent to interpolation formulas for functions in V_j. These can be thought as a generalization of interpolation formulas for trigonometric polynomials (eigenfunctions of the Laplacian on the circle). Most of the functions $\{\varphi_{j,l,k}\}$ are essentially as well localized as it is compatible with their being in V_j. Because of this localization property, we can interpret this downsampling in function space geometrically. We have the identifications

$$X_j := \{x_{l,k}\colon\ x_{l,k} \text{ is center of } \varphi_{j,l,k}\} \leftrightarrow \mathcal{K}_{j,l} \leftrightarrow \{\varphi_{j,l,k}\}_{k\in\mathcal{K}_{j,l}}. \tag{4.17}$$

The natural metric on X_j is $d^{(t_j)}$, which is, by (4.2), the distance in $\mathcal{L}^2(X,\mu)$ between the $\varphi_{j,l,k}$'s, and can be computed recursively in each X_j by combining (4.2) with (4.16). This allows us to interpret X_j as a space representing the metric $d^{(t_j)}$ in compressed form. With this in mind, the construction can be interpreted as a scheme for compression of graphs and manifolds.

In our construction we only compute $\varphi_{j,l,k}$ expanded on the basis $\{\varphi_{j-1,l,k}\}_{k\in\mathcal{K}_{j-1,l}}$, i.e., up to the identifications above, we know $\varphi_{j,l,k}$ on the downsampled space X_j only. However we can extend these functions to X_{j-1} and recursively all the way down to $X_0 = X$, just by using the multiscale relations:

$$\begin{aligned}\varphi_{j,l,k}(x) &= [\Phi_j]_{\Phi_{j-1}}\varphi_{j-1,l,k}(x), \quad x \in X_{j-1}\\ &= [\Phi_j]_{\Phi_{j-1}}[\Phi_{j-1}]_{\Phi_{j-2}}\cdots[\Phi_1]_{\Phi_0}\varphi_{0,l,k}(x)\\ &= \left(\prod_{l=0}^{j-1}[\Phi_l]_{\Phi_{l-1}}\right)\varphi_{0,l,k}(x), \quad x \in X_0. \end{aligned} \tag{4.18}$$

This is of course completely analogous to the standard construction of scaling functions in the Euclidean setting [2, 49–51]. This formula also immediately generalizes to arbitrary functions in V_j, extending them from X_j to the whole original space X (see, for example, Fig. 13).

A detailed analysis of computational complexity is left to a forthcoming publication. We mention here that when the spectrum of T has γ-strong decay, with γ large enough, the matrix $M_j = [\Phi_{j+1}]_{\Phi_j}$ is sparse, since T^{2^j} is local (in the compressed space $|X_j|$), so that M_j has $\mathcal{O}(|X_j|\log|X_j|)$ elements above precision ε, at least for ε not too small and j not too large. When j is large the operator is in general not local anymore, but the space X_j on which it is compressed is very small because of the very small rank of T^{2^j}. In this case the algorithms presented can have order $\mathcal{O}(n^2\log^2 n)$, or even $\mathcal{O}(n\log^2 n)$ for ε not too small and γ large enough.

Remark 32. We could have started from $\tilde{V}_0$ and the defined V_j as the result of j steps of our algorithm: in this way we could do without the spectral decomposition for the semigroup. This permits the application of our whole construction and algorithms to the non-self-adjoint case.

Remark 33. Instead of squaring the operator at each level, one could let $T_j = m_j(T_{j-1})$, and $V'_{j+1} = T_j(V'_j)$, $j \geqslant 1$, $T_0 = T$ and $V'_0 = V_0$, for some function m_j for which functional calculus is applicable. In practice one can choose m_j to be a low-order polynomial. This is analogous to certain non-stationary constructions in classical wavelet theory. It allows to sharpen or smooth the spectral projections onto the spaces V_j spanned by the eigenvectors; it requires only minor modifications to the algorithm and to its analysis.

Remark 34 *(Biorthogonal bases).* While at this point we are mainly concerned with the construction of orthonormal bases for the approximation spaces V_j, well-conditioned bases would be just as good for most purposes and would lead to the construction of stable biorthogonal scaling function bases. This could follow the ideas of "dual" operators in the reproducing formula on space of homogeneous type [6,52], and also, exactly as in the classical biorthogonal wavelet construction [53], we would have two ladders of approximation subspaces, with wavelet subspaces giving the oblique projection onto their corresponding duals. Work on this construction is presented in [54] and more is in progress and will be reported in [55]. Different types of multiscale bases on graphs and manifolds are constructed in [56].

4.7. Wavelets

We would like to construct bases $\{\psi_{j,l,k}\}_{k,l}$ for the spaces W_j, $j \geqslant 1$, such that $V_{j+1} \oplus^{\perp} W_j = V_j$. To achieve this, after having built $\{\{\varphi_{j,l,k}\}_{k\in\mathcal{K}_{j,l}}\}_l$ and $\{\{\varphi_{j,l,k}\}_{k\in\mathcal{K}_{j,l}}\}_l$, we can apply our modified multiscale Gram–Schmidt procedure to the set of functions

$$\left\{(P_j - P_{j+1})\varphi_{j,l,k}\right\}_{k\in\mathcal{K}_{j,l}},$$

where P_j is the projection onto V_j, that yields an orthonormal basis Ψ_j of wavelets for the orthogonal complement W_j of V_j in V_{j+1}. Moreover one can easily prove upper bounds for the diameters of the supports of the wavelets so obtained and for their decay.

4.8. Vanishing moments for the scaling functions and wavelets

In Euclidean settings vanishing moments are usually defined via orthogonality relations to subspaces of polynomials up to a certain degree. In our setting the natural subspaces with respect to which to measure orthogonality is the set of eigenfunctions ξ_λ of T. So the number of vanishing moments of an orthonormal scaling function $\varphi_{j,l,k}$ can

be defined as the number of eigenfunctions corresponding to eigenvalues in $\sigma_j(T) \setminus \sigma_{j+1}(T)$ (as defined in (4.9)) to which $T^{2^j}\varphi_{j,l,k}$ is orthogonal up to precision ε. Observe this is comparable to defining the number of vanishing moments based on $\|T^{2^{j+1}}\varphi_{j,l,k}\|_2$, as in classical estimates in multiscale theory of Calderón–Zygmund operators.

In Section 5 we will see that there is an approximately monotonic relationship between the index l for the "layer" of scaling functions and the number of eigenfunctions to which the scaling function are orthogonal, i.e., the number of moments of the scaling functions.

In particular, the mean of the scaling functions is typically roughly constant for the first few layers and then quickly drops below precision as l grows: this would allow to split each scaling function space in two subspaces, one with scaling functions of non-zero mean and one of scaling functions with zero mean, which in fact could be appropriately called wavelets.

5. The orthogonalization step: computational and numerical considerations

In this section, algorithmic in nature, we discuss details of the implementation, comment on the computational complexity of the algorithms and on their numerical stability.

Suppose we are given a δ-local family $\tilde{\Phi} = \{\tilde{\varphi}_k\}_{k\in\tilde{\mathcal{K}}}$ of positive functions on X and let $X_0 = \{\tilde{x}_k\}_{k\in\tilde{\mathcal{K}}}$ be a supporting set for $\tilde{\Phi}$. With obvious meaning, we will sometimes write $\tilde{\varphi}_{\tilde{x}_k}$ for $\tilde{\varphi}_k$. Let $V = \overline{\langle\{\tilde{\varphi}_k\}_{k\in\tilde{\mathcal{K}}}\rangle}$. We want to build an orthonormal basis $\Phi = \{\varphi_k\}_{k\in\mathcal{K}}$ whose span is ε-close to V. Out of the many possible solutions to this standard problem (see, e.g., [57,58] and references therein as a starting point), we seek one for which the φ_k's have small support (ideally of the same order as the support of $\tilde{\varphi}_k$). Standard orthonormalization in general may completely destroy the size the of the support of (most of) the $\tilde{\varphi}_k$'s.

Proposition 22 suggests a possible solution to this problem, at least when the supporting set for $\tilde{\Phi}$ has some regularity, with respect to a dyadic structure in the space. Observe that there are many choices for the dyadic cubes described in Theorem 6, so the assumption is not overly restrictive.

Here we suggest two algorithms that do not assume any regularity on the family $\tilde{\Phi}$, but for which we cannot guarantee good bounds on the numerical stability. We have been using these in practice and in general obtained very satisfying results, but we will point out their possible weaknesses.

The first algorithm uses geometric information on the set (hence outside the scope of linear algebra) to derive a pivoting rule that guarantees that several orthogonal functions have small support; the second one uses uniquely the matrix representing T in a local basis and a non-linear pivoting rule mixing $\mathcal{L}^2$ and $\mathcal{L}^p$ norms, $p < 2$, which favors sparsity.

The computational cost of constructing the multi-resolution analysis described can be as high as $\mathcal{O}(n^3)$ in general, where n is the cardinality of X, if we do not assume that the semigroup is local and has strong decay. It is because of these two distinct types of properties of T that one can dramatically improve the computational burden necessary for the construction. In general:

I. The decay of $\sigma(T)$: the faster the spectrum decays, the smaller the numerical rank of the powers of T, the more these can be compressed, and the faster the construction of Φ_j for large j's.
II. If each basis Φ_j of scaling functions that we build is such that the elements with large support (large "layer index" l) are in a subspace spanned by eigenvectors of T corresponding to very small eigenvalues (depending on l), then these basis functions of large support will not be needed to compute the next (dyadic) power of T. This implies that the matrices representing all the basis transformations will be uniformly sparse.

A combination of these two hypothesis, which are in fact quite natural and verified by several classes of operators arising in practice, allow to perform the construction of the whole multi-resolution analysis in time asymptotically $\mathcal{O}(n^2 \log^2 n)$ or even $\mathcal{O}(n \log^2 n)$.

5.1. Modified Gram–Schmidt with pivoting

Given a set of functions Ψ on X, and $\varepsilon > 0$, the classical algorithm "Modified Gram–Schmidt with Pivoting" computes an orthonormal basis Φ for a subspace ε-close $\langle\tilde{\Psi}\rangle$, as follows.

1. Let $\tilde{\Phi} = \Psi$, $k = 0$.
2. Let φ_k be an element of $\tilde{\Phi}$ with largest $\mathcal{L}^2$ norm. If $\|\varphi_k\|_2 < \frac{\varepsilon}{\sqrt{|\Psi|}}$, stop, otherwise add $\varphi_k/\|\varphi_k\|_2$ to Φ.
3. Orthogonalize all the remaining elements of Ψ to φ_k, obtaining a new set $\tilde{\Phi}$. Increment k by 1. Go to step 2.

At least when $\tilde{\Phi}$ is not too ill-conditioned, the algorithm stops it yields an orthonormal basis Φ whose span is ε-close to $\langle \Psi \rangle$ [57–59]. This classical procedure can be shown to be stable numerically, at least when $\tilde{\Phi}$ is not too ill-conditioned [57,58]. When the problem is very ill-conditioned, a loss of orthogonality in Φ may occur. Reorthogonalizing the elements already selected in Φ is enough to achieve stability, see [57,59] for details and references.

The main drawback of this procedure is that the supports of the functions in Φ can be arbitrarily large even if the supports of $\tilde{\Phi}$ are small and there exists another orthonormal basis for $\langle \tilde{\Phi} \rangle$ made of well-localized functions. This is undesirable, because the size of the supports of the functions in the new basis are crucial for the speed of the algorithm. We suggest two modifications of the modified Gram–Schmidt that seek to obtain orthonormal bases of functions with smaller support.

5.2. *Multiscale modified Gram–Schmidt with geometric pivoting*

We suggest here to use a local Gram–Schmidt orthogonalization with "geometric pivoting": it does not provide guarantees on the size of the support of the orthonormal functions it constructs, but in practice has complexity $\mathcal{O}(n^2 \log^2 n)$, or even $\mathcal{O}(n \log^2 n)$, where n is the cardinality of X, when applied to a local family. It is inspired by the construction in Proposition 22, but it is greedy and can be applied with no assumption on the geometry of the supporting set of the functions to be orthonormalized. Here we are assuming that the approximate (up to precision) ε-neighbors of any point in X are known. After all, we are given T with respect to a local basis, and this already encodes most of the nearest neighbor information needed. This is a highly non-trivial encoding, especially in high dimensions: the literature on fast approximate nearest neighbors and range searches is vast, see, for example, [60] and references therein as a starting point. In the following section we will present an algorithm that does not use the geometry of the supports and does not need any knowledge of nearest neighbors, except what is already encoded in the matrix representing T.

The Gram–Schmidt procedure we present is completely local, in the sense that it is organized so that each of the constructed functions needs to be orthonormalized only to the nearby functions, across scales, at the appropriate scale.

Description of the algorithm. We are given a precision $\varepsilon > 0$ and an $\alpha\delta$-local, $\alpha \leqslant 1$, set of functions $\Psi = \{\psi_x\}_{x \in \Gamma}$ with a supporting set Γ. We pick a maximal net of points $x_{0,0}, \ldots, x_{0,k}$ in Γ which is $2\alpha\delta$-separated (this is similar to the construction of lattices on spaces of homogeneous type [11,26,61]). In order to do this, we pick $x_{0,0} \in \Gamma$, then pick $x_{0,1}$ to be a closest point in Γ which is at distance at least 2δ from $x_{0,0}$, and so on: after $x_{0,0}, \ldots, x_{0,l}$ have been picked, let $x_{0,l+1} \in \Gamma$ be a closest point to $\mathcal{N}_{2\delta}(\{x_{0,0}, \ldots, x_{0,l}\})$. Let $\{x_{0,k}\}_{k \in \mathcal{K}_0}$ be a maximal such collection and let $B_{0,k} = B_{\alpha\delta}(x_{0,k})$ for each $k \in \mathcal{K}_0$, and $\Psi_0 \subseteq \Psi$ be the set of functions having support completely contained in $B_{0,k}$ for some $k \in \mathcal{K}_0$. We ε-orthonormalize this family of functions (i.e., we obtain a basis for an ε-span of Ψ_0, for example with any of the variations of Gram–Schmidt orthogonalization described) to obtain

$$\Phi_0 = \{\varphi_{0,k}\}_{k \in \mathcal{K}_0}.$$

Observe that this orthonormalization procedure is local, in the sense that each function $\psi_x \in \Psi_0$ only interacts with the other functions in Ψ_0 contained in the same ball $B_{0,k}$ containing the support of ψ_x.

We orthogonalize all the remaining functions in Ψ to the functions in Φ_0, and get a new set of functions, Ψ_1. This orthonormalization is also local: each function $\psi_x \in \Psi$ being $\alpha\delta$-local, it interacts only with the functions in Φ_0 which have support in a ball $B_{0,k}$ containing x. Moreover, Ψ_1 is also a local family. In fact it is at least $(\alpha\delta + 2\alpha\delta)$-local. We then proceed in the same way with Ψ_1.

By induction, once $\Phi_0, \ldots, \Phi_l$ have been constructed, so that their union is an orthonormal basis and Ψ_{l+1} is the set of functions in Ψ not yet picked, orthonormalized with respect to $\Phi_0 \cup \cdots \cup \Phi_l$, we pick a maximal subset of functions in Ψ_{l+1} with disjoint supports and ε-orthonormalize them to get Φ_{l+1}. The algorithm terminates when Φ_{l+2} is empty, which happens when $\Phi_0 \cup \cdots \cup \Phi_{l+1}$ already ε-spans $\langle \Psi \rangle$. When the space X is finite, this happens after a finite number of steps since eventually the support of the functions in Ψ_{l+1} is the whole X, in which case the family Ψ_{l+1} consists of all the functions in Ψ which have not been picked already, $\langle \Psi \rangle \subseteq \langle \Phi_0 \cup \cdots \cup \Phi_{l+1} \rangle_\varepsilon$, and Ψ_{l+2} is empty.

The greedy choice of the points at each "layer" can lead to more localized basis than the one obtained by modified Gram–Schmidt. However this algorithm can lead to numerically instabilities. Qualitatively speaking, it is expected to perform well when the $\mathcal{L}^2$ norms of the functions Ψ are within close constants and Γ is a rather dense and uniformly well-distributed set.

Remark 35. The construction in this proposition applies in particular to the fast construction of local orthonormal (multi-)scaling functions in $\mathbb{R}^n$, which may be of independent interest. As an example, this construction can be applied to spline functions of any order, yielding an orthonormal basis of splines different from the classical one due to Strömberg [62].

Remark 36. The cost of running the above algorithm can be as low as $\mathcal{O}(n \log n)$ times the cost of finding (approximate) r-neighbors, in addition to the cost of any pre-processing that may be necessary to perform these searches fast. This follows from the fact that all the orthogonalization steps are local unless Γ is very irregular. A full analysis of the complexity will be presented in a future publication.

5.3. Modified Gram–Schmidt with mixed $\mathcal{L}^2$–$\mathcal{L}^p$ pivoting

In this section we propose a second algorithm for computing an orthogonal basis spanning the same subspace as the one spanned by a given δ-local family Ψ. The main motivation for the algorithm in the previous section was the observation that modified Gram–Schmidt with pivoting in general can generate basis functions with large supports, while it is crucial in our setting to find orthonormal bases with rather small support.

We would like to introduce a term in modified Gram–Schmidt that prefers functions with smaller support to functions with larger support. On the unit ball of $\mathcal{L}^2$, the functions with concentrated support are those in $\mathcal{L}^0$ or, more generally, in $\mathcal{L}^p$ for some $p < 2$.

We can then modify the "modified Gram–Schmidt with pivoting" algorithm of Section 5.1 as follows:

1. Let $\tilde{\Phi} = \Psi$, $k = 0$, $\lambda > 0$, $p \in [0, 2]$.
2. Let $\varphi_{k'}$ the element of $\tilde{\Phi}$ with largest $\mathcal{L}^2$ norm, say $N_k = \|\varphi_{k'}\|_2$. Among all the elements of $\tilde{\Phi}$ with $\mathcal{L}^2$ norm *larger* than N_k/λ, pick the one with the *smallest* $\mathcal{L}^p$ norm: let it be φ_k. If $\|\varphi_k\|_2 < \frac{\varepsilon}{\sqrt{|\Psi|}}$, stop, otherwise add $\varphi_k/\|\varphi_k\|_2$ to Φ.
3. Orthogonalize all the remaining elements of Ψ to φ_k, obtaining a new set $\tilde{\Phi}$. Increment k by 1. Go to step 2.

Choosing the element by bounding from below its $\mathcal{L}^2$ norm is important for numerical stability, but having relaxed the condition of picking the element with largest $\mathcal{L}^2$ norm allows us to pick an element with smaller $\mathcal{L}^p$ norm, yielding potentially much better localized bases. The parameter λ controls this slack. It is easy to construct examples in which the standard modified Gram–Schmidt with $\mathcal{L}^2$ pivoting (which corresponds to $\lambda = 1$) leads to bases with most elements having large support, while our modification with mixed norms yields much better localized bases.

Remark 37. Observe that this algorithm seems not require knowledge of the nearest neighbors. However, in the implementation, the matrix representing Ψ should be in sparse form and queries of non-zero elements by both rows and columns should ideally be $\mathcal{O}(1)$ operations.

We have not investigated theoretically the stability of this algorithm. We just observed that it worked very well in the examples tried, with stability comparable to the standard modified Gram–Schmidt with pivoting and very often much better localization in the resulting basis functions. On the other hand, we can construct quite artificial examples in which any choice of $\lambda < 1$ leads to loss of precision.

5.4. Rank revealing QR factorizations

We refer the reader to [63] and references therein, for details regarding the numerical and algorithmic aspects related to the rank-revealing factorizations discussed in this section.

Definition 38. A partial QR factorization of a matrix $M \in \mathrm{Mat}(n,n)$

$$M\Pi = QR = Q\begin{pmatrix} A_k & B_k \\ 0 & C_k \end{pmatrix}, \tag{5.1}$$

where Q is orthogonal, $A_k \in \mathrm{Mat}(k,k)$ and is upper-triangular with non-negative diagonal elements, $B_k \in \mathrm{Mat}(k, n-k)$, $C_k \in \mathrm{Mat}(n-k, n-k)$, and Π is a permutation matrix, is a *strong rank-revealing QR factorization* if

$$\sigma_{\min}(A_k) \geqslant \frac{\sigma_k(M)}{p_1(k,n)} \quad \text{and} \quad \sigma_j(C_k) \leqslant \sigma_{k+j}(M)\, p_1(k,n) \tag{5.2}$$

and

$$\left|\left(A_k^{-1} B_k\right)\right| \leqslant p_2(k,n), \tag{5.3}$$

where Q in orthogonal, $p_1(k,n)$ and $p_2(k,n)$ are functions bounded by a low-degree polynomial in k and n, and $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote the largest and smallest singular values of a matrix A.

Modified Gram–Schmidt with pivoting, as described in Section 5.1 actually yields a decomposition satisfying (5.2), but with p_1 depending exponentially in n [63], at least for some matrices (which seem to be quite rare).

Let $\mathrm{SRRQR}(M,\varepsilon)$ be an algorithm that computes a rank-revealing QR factorization of M, with $\sigma_{k+1} \leqslant \varepsilon$. Gu and Eisenstat present such an algorithm, that satisfies (5.2) and (5.3) with

$$p_1(k,n) = \sqrt{1 + nk(n-k)} \quad \text{and} \quad p_2(k,n) = \sqrt{n}.$$

Moreover, this algorithm requires in general $\mathcal{O}(n^3)$ operations, but faster algorithms exploiting sparsity of the matrices involved can be devised.

We can then proceed in the construction of the multi-resolution analysis described in Theorem 27 as follows. We replace each orthogonalization step G_j by the following two steps. First we apply $\mathrm{SRRQR}(T_j^{2^j}, \varepsilon)$ to get a rank-revealing QR factorization of T^{2^j}, on the basis Φ_j, which we write as

$$T_j^{2^j}\Pi_j = Q_j R_j = Q_j \begin{pmatrix} A_{j,k} & B_{j,k} \\ 0 & C_{j,k} \end{pmatrix}. \tag{5.4}$$

The columns of Q_j span $\tilde{V}_{j+1}$ up to error $\varepsilon \cdot p(k,n)$, where

$$k = \arg\max_i \left\{ i\colon \lambda_i^{\eta^{j+1-1}} \geqslant \varepsilon \right\} = \dim(\tilde{V}_{j+1}).$$

In fact, the strong rank-revealing QR decomposition above shows that the first k columns of $T^{2^j}\Pi_j$ are a well-conditioned basis that $\varepsilon \cdot p(k,n)$-spans $\tilde{V}_{j+1}$. Now, this basis is a well-selected subset of "bump functions" $T^{2^{j+1}-1}\tilde{\Phi}$ and can be orthogonalized with Proposition 22, with estimates on the supports exactly as claimed Theorem 27.

5.5. Computation of the wavelets

In the computation of wavelets as described in Section 4.7, numerically one has to make sure that, at every orthogonalization step, the construction of the wavelets is made on vectors that are numerically orthogonal to the scaling functions at the previous scale. This can be attained, in a numerically stable way, by repeatedly orthogonalizing the wavelets to the scaling functions. Observe that this orthogonalization is again a local operation and hence the computation is fast. This is also necessary to guarantee that the construction of wavelets will stop exactly at the exhaustion of the wavelet subspace, without "spilling" into the scaling spaces.

6. Wavelet transforms and algorithms

Assume that each orthonormalization step and change of basis can be computed in time $\mathcal{O}(n^k)$, up to logarithmic factors. Usually $k = 1, 2$. Then the computational cost for the algorithms for constructing the whole multi-resolution analysis are as follows.

Proposition 39. *The construction of all the scaling functions and linear transformations* $\{G_j\}$ *and* $\{M_j\}$, *for* $j = 0, \ldots, J$, *as described in Theorem* 27, *can be done in time* $\mathcal{O}(n^k)$.

Corollary 40 *(The fast scaling function transform). Let the scaling function transform of a function* $f \in \mathcal{L}^2(X, \mu)$, $|X| = n$, *be the set of all coefficients*

$$\left\{\langle f, \Phi_j \rangle\right\}_{j=0,\ldots,J}.$$

All these coefficients can be computed in time $\mathcal{O}(n^k)$.

The following is a simple consequence of Corollary 40.

Corollary 41 *(The fast wavelet transform). Let the wavelet transform of a function* $f \in \mathcal{L}^2(X, \mu)$ *be the set of all coefficients*

$$\{\langle f, \Psi_j \rangle\}_{j=0,\ldots,J}.$$

All these coefficients can be computed in time $\mathcal{O}(n^k)$.

Material relevant to the following two sections is in the papers [44,45,64] (and references therein) which present matrix compression techniques with applications to numerical functional calculus and eigenfunction computations.

6.1. Compressed eigenfunction computation

The computation of approximations to the eigenfunctions ξ_λ corresponding to $\lambda > \lambda_0$ can be computed efficiently in V_j, where j is $\min\{j\colon \lambda_0^{t_j} > \varepsilon\}$, and then the result can be extended to the whole space X by using the extension formula (4.18). This technique is equivalent to having constructed integration formulae for the top eigenfunctions of the operator and the linear space they span.

These eigenfunctions can also be used to embed the coarsened space X_j into $\mathbb{R}^n$, along the lines of Section 4.2, or can be extended to X and used to embed the whole space X.

6.2. Direct inversion of "Laplacian-like" operators

Since we can compute efficiently high powers of T, the "Laplacian" $(I - T)$ can be inverted (on the orthogonal complement of the eigenspace corresponding to the eigenvalue 1) via the Schultz method [2], that we exemplify for $\eta = 2$. Since

$$(I - T)^{-1} f = \sum_{k=1}^{+\infty} T^k f$$

and, if $S_K = \sum_{k=1}^{2^K} T^k$, we have

$$S_{K+1} = S_K + T^{2^K} S_K = \prod_{k=0}^{K} \left(I + T^{2^k}\right).$$

Since we can apply efficiently T^{2^k} to any function f and hence the product S_{K+1}, we can apply $(I - T)^{-1}$ in compressed form to any function f. The value of K depends on the gap between 1 and the first eigenvalue smaller than 1, which essentially regulates the speed of convergence of a distribution to its asymptotic distribution.

Observe that we never construct the full matrix representing $(I - T)^{-1}$ (which is general full!), but we only keep a compressed multiscale representation of it, which we can use to compute the action of the operator on any function.

This matrix is of fundamental importance in Markov chains (and it often goes under the name of "fundamental matrix"), discrete and continuous potential theory, study of complex networks, clustering. These and other applications will be discussed in forthcoming publications.

6.3. Relationship with eigenmap embeddings

In Section 4.2 we discussed metrics induced by a symmetric diffusion semigroup $\{T^t\}$ acting on $\mathcal{L}^2(X,\mu)$ and how eigenfunctions of the generator of the semigroup can be used to embed X in Euclidean space. Similar embeddings can be obtained by using diffusion scaling functions and wavelets, since they span subspaces localized around spectral bands. So, for example, the top diffusion scaling functions in V_j approximate well, by the arguments in this section, the top eigenfunctions. See, for example, Fig. 13.

7. Natural multiresolution on sampled Riemannian manifolds and on graphs

The natural construction of Brownian motion and the associated Laplace–Beltrami operator for compact Riemannian manifolds can be approximated on a finite number of points which are realizations of a random variable taking values on the manifold according to a certain probability distribution as in [3,4,65,66]. The construction starts with assuming that we have a data set Γ which is obtained by drawing according to some unknown distribution p and whose range is a smooth manifold $\bar{\Gamma}$. Given a kernel that satisfies some natural mild conditions, for example in the standard Gaussian form

$$K(x,y) = e^{-(\frac{\|x-y\|}{\delta})^2}$$

for some scale factor $\delta > 0$, is normalized twice as follows before computing its eigenfunctions. The main observation is that any integration on the empirical data set Γ of the kernel against a function is in the form

$$\sum_{\Gamma} K(x,y_i) f(y_i)$$

and thus it a Riemann sum associated to the integral

$$\int_{\bar{\Gamma}} K(x,y) f(y) p(y)\,\mathrm{d}y.$$

Hence to capture only the geometric content of the manifold it is necessary to get rid of the measure p on the dataset. This can be estimated at scale δ for example by convolving with some smooth kernel (for example, K itself), thus obtaining an estimated probability density p_δ. One then considers the kernel

$$\tilde{K}(x,y) = \frac{K(x,y)}{p_\delta(y)}$$

as new kernel. This kernel is further normalized so that it becomes averaging on the data set, yielding

$$\bar{K}(x,y) = \frac{\tilde{K}(x,y)}{\sqrt{\int_\Gamma \tilde{K}(x,y) p(y)\,\mathrm{d}y}}.$$

It is shown in [3] that with this normalization, in the limit $|\Gamma| \to +\infty$ and $\delta \to 0$, the kernel $\bar{K}$ thus obtained is the one associated with the Laplace–Beltrami operator on $\bar{\Gamma}$. Applying our construction to this natural kernel, we obtain the natural multiresolution analysis associated to the Laplace–Beltrami operator on a compact Riemannian manifold. This also leads to compressed representation of the heat flow and of functions of it. One can show that Weyl's theorem implies that the heat semigroup has $\frac{n}{2}$-strong decay.

On weighted graphs, the canonical random walk is associated with the matrix of transition probabilities

$$P = D^{-1}W,$$

where W is the symmetric matrix of weights and D is the diagonal matrix defined by $D_{ii} = \sum_j W_{ij}$. P is in general not symmetric, but it is conjugate, via $D^{\frac{1}{2}}$ to the symmetric matrix $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ which is a contraction on $\mathcal{L}^2$. Our construction allows the construction of a multiresolution on the graph, together with downsampled graphs, adapted to the random walk [13].

8. Examples and applications

8.1. A diffusion multiresolution on the circle

For pedagogical purposes, we illustrate the construction in a very simple example. We consider the Laplacian on the circle $\mathbb{T}$, since even in this case the multiresolution analysis we introduce is new.

We let X be a set of 512 equispaced on $\mathbb{R}$, and let T' the standard 3-point discretization of the Laplacian:

$$T' = \begin{pmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \cdots & \cdots \\ & \cdots & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

We let $T = (T')^2$ be our diffusion operator, and let the initial basis Φ_0 is the set of δ-functions on X. We set ε, the precision, to 10^{-4}.

The first 11 levels of our construction leads to subspaces $V_0, V_1, \ldots, V_{11}$ of decreasing dimension, as represented in Fig. 9. Observe that the spectrum does not decay very fast in this case.

We apply T to Φ_0 and after downsampling we obtain a layered orthonormal basis of scaling functions Φ_1 (which, incidentally, are quadratic splines), some of which are represented in Fig. 10. These scaling functions are clearly linear combinations of the δ-functions spanning V_0. The matrix of coefficients is represented in Fig. 11. We then apply T^2,

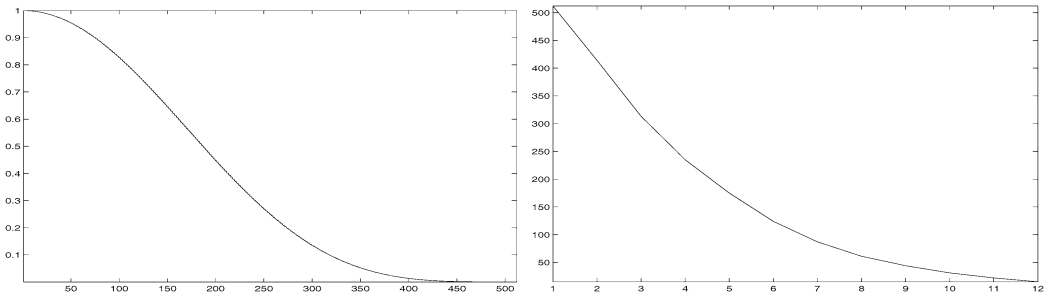

Fig. 9. Left: the spectrum of T. Right: the dimension of V_j as a function j.

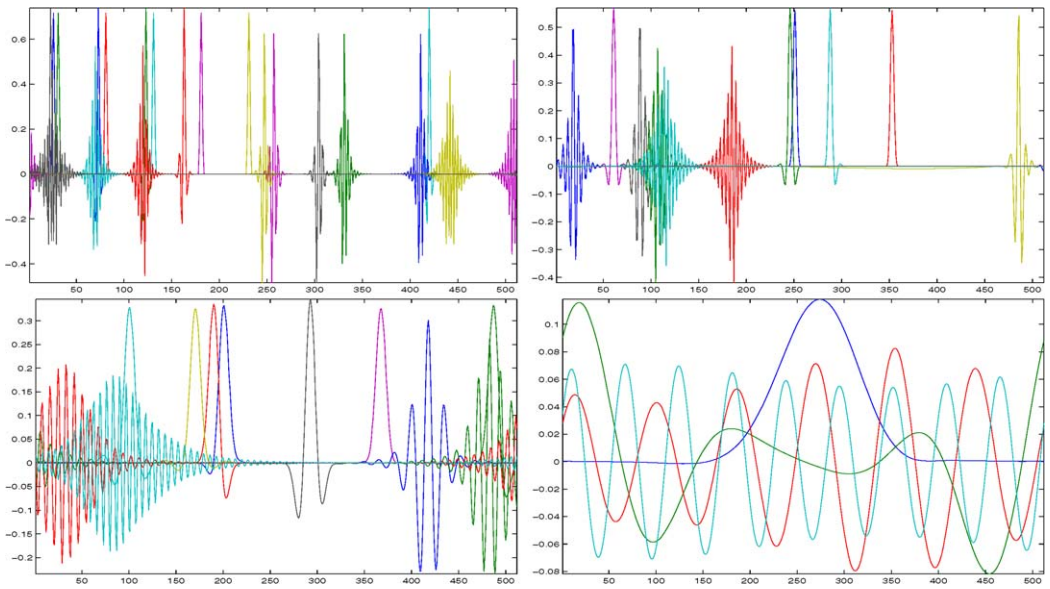

Fig. 10. Some scaling functions in V_1 (top left), in V_3 (top right), V_6 (bottom left), and V_{12} (bottom right).

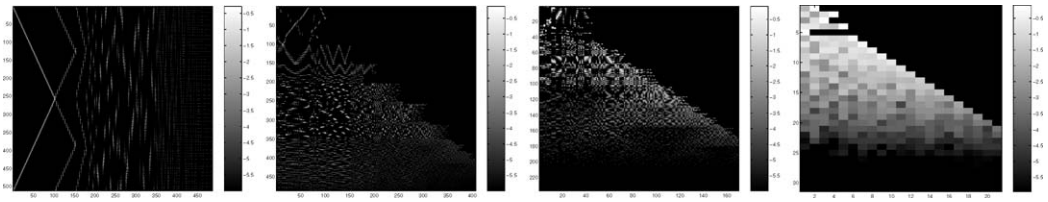

Fig. 11. The scaling function filters M_1 (top left), M_2 (top right), M_5 (bottom left), and M_{11} (bottom right). The images are in logarithmic scale to show entries larger than 10^{-6}.

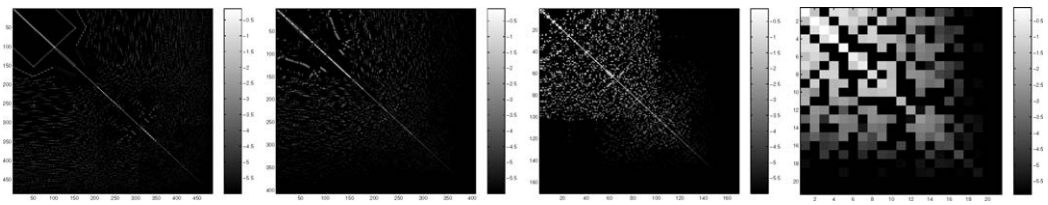

Fig. 12. The compressed powers of T at scale $j = 1, 2, 5, 11$. The images are in logarithmic scale to show entries larger than 10^{-6}.

represented on Φ_1, and so on. We represent in Fig. 10 some scaling functions in V_1, V_2, V_5, and V_{11}, extended to the whole circle. Some observations are in order. In each picture, we represent scaling functions from different layers: some of them maximally concentrated, others more oscillating and with larger support, but still exponential decay. Notice how the scaling functions at very coarse scale, e.g., V_{11} tend to sine and cosine functions. This is because V_{11} is spanned by the few top eigenfunctions of T, which are exactly sine and cosine functions with low frequency, and when only few frequencies are available, these functions cannot be localized very well by taking linear combinations. In V_{15}, an even coarser scale, there are only three scaling functions: the constant function, and sine and cosine.

Observe that the algorithm does not construct them explicitly, but only computes and stores the filter matrices M_j (see Fig. 11), such that

$$\Phi_{j+1} = M_j \Phi_j,$$

as discussed above (see Eq. (2.1) and also the diagram 2). Notice that these matrices have a multi-layer structure and also that the filters they represent are non-stationary both in scale and in location, i.e., they depend on both the scale and the location of the scaling function. The dimension of V_j decreases with j because the part of the spectrum of $T^{2^{j+1}-1}$ above precision gets smaller and smaller with j, and so does the size of the matrices M_j. In Fig. 12 we represent some of the dyadic powers of T represented on the corresponding scaling function bases.

8.2. A simple homogenization problem

We consider the non-homogeneous heat equation on the circle

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(c(x)\frac{\partial}{\partial x}u\right), \tag{8.1}$$

where $c(x)$ is quite non-uniform and we want to represent the large scale/large time behavior of the solution by compressing powers of the operator representing the discretization of the spatial differential operator $\frac{\partial}{\partial x}(c(x)\frac{\partial}{\partial x})$. The spatial operator is of course one of the simplest Sturm–Liouville operators in one dimension.

We choose the function $0 < c < 1$ represented in Fig. 13. We discretize the *spatial* differential operator, thus obtaining a matrix W. In order to have an operator T with $\mathcal{L}^2$-norm 1, we let $D_{ii} = \sum_j (2 - W_{ij})$ and $T = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, which has the desired properties of contraction, self-adjointedness and positiveness.

We discretize the interval at 256 equispaced points, as in the previous example, and the discretization of the right-hand side of (8.1) is a matrix T which, when properly normalized, can be interpreted as a non-translation invariant random walk. Our construction yields a multiresolution associated to this operator that is highly non-uniform, with most scaling functions concentrated around the points where the conductivity is highest, for several scales. The dimension of V_j drops quite fast and V_8 is already reduced to two dimensions. In Fig. 13 we plot, among other things,

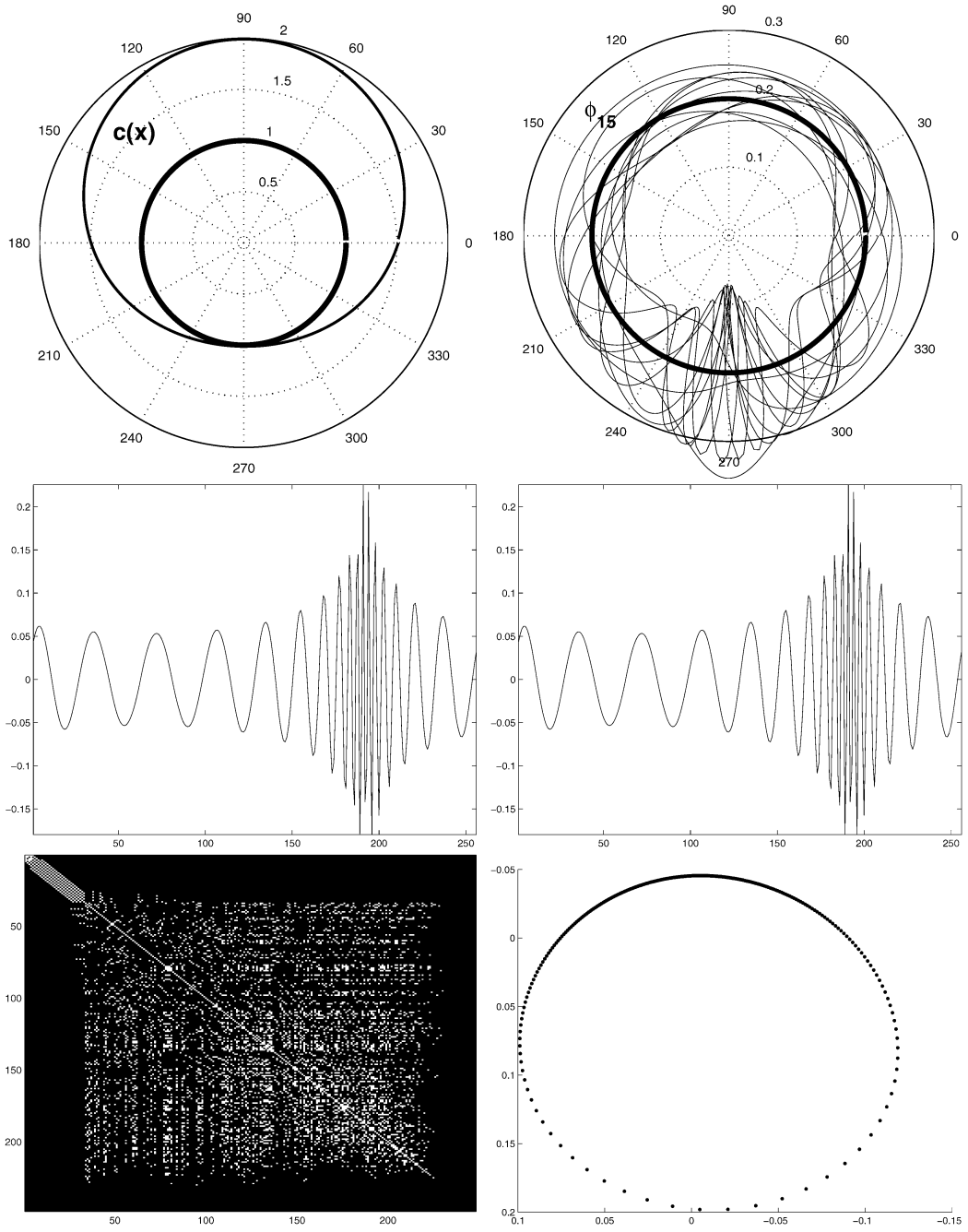

Fig. 13. Non-homogeneous medium with non-constant diffusion coefficient (plotted top left): the scaling functions for V_{15} (top right). In the middle row: an eigenfunction (the 35th) of the diffusion operator (left), and the same eigenfunction reconstructed by extending the corresponding eigenvector of the compressed T_{10} (right): the $\mathcal{L}^2$-error is of order 10^{-5}. The entries above precision of the matrix T^8 compressed on V_4 (bottom left). Bottom: we plot on the right $\{(\varphi_{18,1}(x_i), \varphi_{18,2}(x_i))\}_i$, which are an approximation to the eigenmap corresponding to the first two eigenfunctions, since V_{18} has dimension 3. Points at the bottom of circle are farther in diffusion distance because there the conductivity is lower.

the embedding given by the two scaling functions in V_8: while it differs from an eigenmap because the coordinates are not scaled by the eigenvalues (which we could cheaply compute since it is a 2×2 matrix), it clearly shows the non-uniformity of the heat distance (which is roughly Euclidean distance in the range of this map): the points with very high conductivity are very tightly clustered together, while all the others are further apart and almost equispaced.

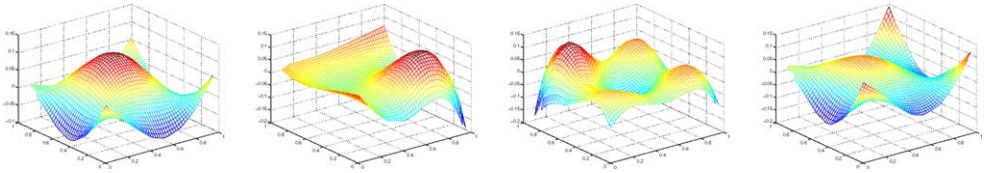

Fig. 14. Example of scaling functions at coarse level associated with a Beltrami diffusion on randomly distributed points on the unit square. For graphical reasons, we are plotting a smooth extension of these scaling functions on a uniform grid by cubic interpolation.

The compressed matrices representing the (dyadic) powers of this operator can be viewed as homogenized versions, at a certain scale which is time and space dependent, of the original operator. The scaling functions at that scale span the domain the homogenized operator at that scale. Further developments in the applications to homogenization problems will be presented in a future work.

8.3. Perturbed lattices

Our construction applies naturally to perturbed lattices in any dimension. Wavelets and multiresolution analysis on irregular lattices have been studied and applied by various researchers, see, for example, [67,68] and references therein. Our construction overcomes the usual difficulties related to local oversampling (too many lattice points accumulating around a single point) and non-uniformities, by automatically computing the local oversampling and locally downsampling and orthogonalizing as requested by the problem.

As an example, we consider a set of 500 randomly drawn from a uniform distribution on the square, and consider the diffusion operator normalized a la Beltrami as described in Section 7. We plot in Fig. 14 some of the scaling functions we obtain.

8.4. A noisy example

We consider a data set X consisting of 1200 points in $\mathbb{R}^3$, which are the union of 400 realizations of thee independent Gaussian normal variables G_1, G_2, G_3, with values in $\mathbb{R}^3$, with means $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, respectively, and standard deviations all equal to 0.4. The data set if depicted in Fig. 15. We consider the graph (G, E, W) built as follows: the vertex set G is X, two points $x, y \in G$ are connected by an edge of weight $W_{x,y} = e^{-(\frac{\|x-y\|}{0.3})^2}$ if such weight is greater than 10^{-3}. We let T be the random walk naturally associated to this graph, which is defined by the row-stochastic matrix of transition probabilities

$$P = D^{-1}W,$$

where $D_{x,x} = \sum_{y\in G} W_{x,y}$ and $D_{x,y} = 0$ for $x \neq y$. We let the diffusion semigroup be generated by $T = P/\|P\|_{2,2}$ and construct the corresponding multiresolution analysis and wavelets.

We can see in Fig. 15 that the eigenfunctions of this operator can tell rather well the three Gaussian clouds apart, since the diffusion inside each cloud is rather fast while the diffusion across clouds is much slower. For the same reason, the diffusion scaling functions, as well as the wavelets, tend to be concentrated on each cloud. The top scaling functions in several scaling subspaces approximate very well the characteristic function of each cloud and the wavelets analyze well oscillating functions which are concentrated on each cloud (local-modes).

In Fig. 16 we show how the diffusion scaling functions compress the data. Let χ_R, χ_B, χ_G be the characteristic functions of each Gaussian cloud and $\chi = \chi_R + \chi_B + \chi_G$. In the figure we represent, for several j's, X_j (notation as in Section 2.2), which is a compressed version of X after diffusing till time $2^{j+1} - 1$. We also compute the multiscale expansion of χ, and use colors to represent the projection onto jth approximation subspace, evaluated on X_j.

8.5. Non-linear analysis of images

We present a toy example of an application to the construction of features and basis functions for the analysis of images. We start with an image I of a white disk on black background, of size 36 by 36, and add a small amount

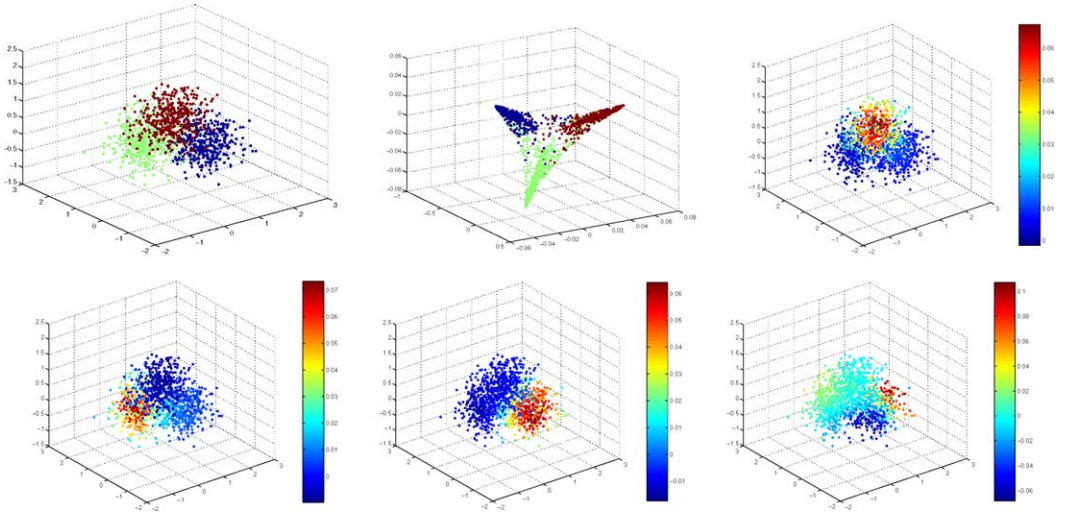

Fig. 15. Top, left to right: the dataset described in Example 8.4 with labels for each Gaussian random variable, the embedding given by the top 3 non-trivial eigenfunction of the graph Laplacian, the values of the scaling function $\varphi_{6,1,1}$. Bottom, left to right: values of the scaling functions $\varphi_{6,1,2}$, $\varphi_{6,1,3}$, and of the wavelet $\psi_{8,5}$.

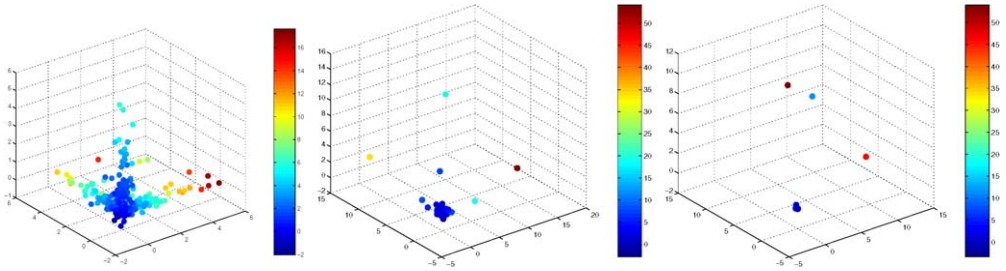

Fig. 16. Left to right, we plot $X_j = M_j \cdots M_1 M_0 X$ for $j = 1, 3, 5$, respectively. The color is proportional to the projection of χ onto V_j evaluated at each point of X_j. The dataset is being compressed, since we have $|X| = 1200$, $|X_1| = 1200$, $|X_3| = 186$, $|X_5| = 39$, and the images of the centers of the classes are preserved as sampling points at each scale.

of Gaussian white noise. We collect all 5 by 5 squares ("patches") from an image, map each of them to a vector in $\mathbb{R}^{25}$ in an arbitrary (but fixed) way, obtaining a point cloud G consisting of 36^2 points in $\mathbb{R}^{25}$. Clearly each $x \in G$ corresponds to a "patch" in the original image. We attach two extra coordinates to each point $x \in G$, which are the x and y coordinates of the center of p, viewed as a "patch" in I. We connect p_1 and p_2 if they both have similar (x, y)-coordinates and are close as vectors in $\mathbb{R}^{25}$. In the specific situation for each p we first consider the set $N(p)$ of the 50 nearest neighbors of p in (x, y) and in $N(p)$ we find the 20 nearest neighbors of p in $\mathbb{R}^{25}$, which we denote by $NN(p)$. We connect p with each element $p' \in NN(p)$ with weight $e^{-(\|p-p'\|_{\mathbb{R}^{25}}/5)^2}$. The weight matrix thus obtained is in general not symmetric: we average with the transpose of the weight matrix in order to obtain symmetric weights. Let (G, E, W) denote the graph thus obtained. We consider the natural diffusion $I - \mathcal{L}$ on G, where $\mathcal{L}$ is the normalized Laplacian and construct the corresponding diffusion scaling functions and wavelets.

Observe that every function on G is actually a 36 by 36 image, because of the natural identifications between vertices of G, "patches," and (x, y)-centers of these "patches." Hence every scaling function and wavelet is actually an image and we see they represent in general interesting multiscale features, such as blobs, edge filters, functions oscillatory along an edge and essentially supported on the edge, and coarser features like Gabor-like filters, and so on. The diffusion process and the multiscale construction allow to discover some of the simplest features of the image, from extremely "local" (in the space of "patches," $\mathbb{R}^{25}$) to more global and complex (see Fig. 17).

Furthermore, the original image I itself can be viewed as a function on G, and hence it can be analyzed as a function on G and denoised by using standard wavelet denoising algorithms [69] adapted to the context of diffusion wavelets, such as best basis algorithms [12]. This denoising *naturally* preserves edges: the "filtering" happens mainly on and

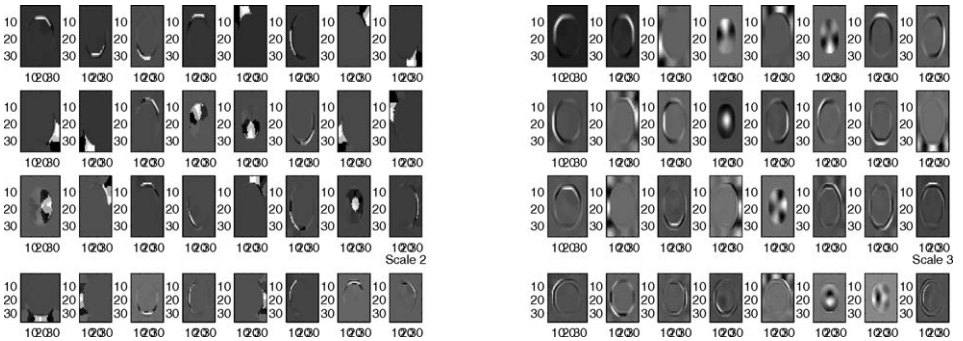

Fig. 17. Scaling functions at scale 2 and 3 on the graph of patches extracted from an image of a white full circle on black background, with noise.

along the edges and not across them, since in "patch" space points corresponding to "patches" with edges are far from points corresponding to "patches" which do not contain edges. Preliminary results are extremely promising [70]; we will report on progress in future publications. This is a highly non-linear analysis of images, since the analyzing elements depend very much on the image itself. Of course the framework can be significantly broadened, in particular in the construction of the graph (other features of images rather than "patches" could be considered in order to build the graph, for example, sets of filter responses) and of the diffusion on the graph (choice of weights, distances, etc.) and extended to movies, hyper-spectral images, and so on. This framework can also be broadened to the joint analysis of families of images, for the construction of features that can be used to analyze new images.

9. Extending the scaling functions outside the set

Suppose there exists a metric measure space $(\bar{X}, \bar{d}, \bar{\mu})$ such that $X \subset \bar{X}$, and suppose that the projection operator

$$R_X^{\bar{X}} : \mathcal{L}^2(\bar{X}, \bar{\mu}) \to \mathcal{L}^2(X, \mu),$$
$$f \mapsto f|_X$$

is bounded. Suppose we have a sequence of subspaces $\{E_n\}_{n \geqslant 0}$, with $E_n \to \mathcal{L}^2(\bar{X}, \bar{\mu})$ as $n \to \infty$, and a singular value decomposition for $R_X^{\bar{X}}|_{E_n}$ that we write as

$$R_X^{\bar{X}}|_{E_n} f = \sum_{i \geqslant 1} \alpha_{n,i} \langle f, \tilde{\theta}_{n,i} \rangle \theta_{n,i},$$

where $\alpha_{n,1} \geqslant \alpha_{n,1} \geqslant \cdots \geqslant 0$.

Given a scaling function $\varphi_{j,k}$, we can find the smallest n such that

$$\varphi_{j,k} \in P_{\overline{\langle \theta_{n,1}, \theta_{n,2}, \ldots \rangle}} E_n \tag{9.1}$$

where in practice this relationship is to be intended in the sense of numerical range of the projection involved. Then we would define the extension of

$$\varphi_{j,k} = \sum_{i \geqslant 1} a_i \theta_{n,i} \tag{9.2}$$

to be

$$\overline{\varphi_{j,k}} = \sum_{i \geqslant 1} \frac{a_i}{\alpha_{n,i}} \tilde{\theta}_{n,i}. \tag{9.3}$$

This makes numerical sense only when the sum is extended over those indices i such that $\alpha_{n,i}$ is well separated below from 0, which is equivalent to the condition that (9.1) holds in the numerical sense. See also [71].

Observe that since the n for which (9.1) holds depends on $\varphi_{j,k}$, each scaling function will be extended in different ways. If we think of the E_n as spaces of functions on $\bar{X}$ with increasing complexity as n grows, and we try to extend each scaling function to a function on $\bar{X}$ with the minimum complexity required, the requirement being (9.1).

Example 42. If X is a subset of $\mathbb{R}^n$, then we can, for example, [3,4] proceed in any of the following ways:

(i) Let

$$BL_c = \left\{ f \in \mathcal{L}^2\left(\mathbb{R}^n\right) \colon \operatorname{supp} \hat{f} \subseteq B_c(0) \right\} = \overline{\left\langle \left\{ e^{i\langle \xi, \cdot \rangle} \colon |\xi| \leqslant c \right\} \right\rangle},$$

the space of band-limited functions with band c and then let E_n be BL_{2^n}. The $\{\tilde{\theta}_{n,i}\}_i$ in this case are generalized prolate spherical functions of [72,73], called *geometric harmonics* in [3,4].

(ii) Let E_n be one the approximation spaces V_n of some classical wavelet multiresolution analysis in $\mathbb{R}^n$, like the one associated to Meyer wavelets.

(iii) Let

$$BG_c = \overline{\left\langle \left\{ e^{-\frac{\|\cdot - y\|}{c}} \right\}_{y \in c\mathbb{Z}^n} \right\rangle},$$

the bump algebra at scale $1/c$, and furthermore let $E_n = BG_{2^{-n}}$. While in this case it is not the case that $E_n \subset E_{n+1}$, it is true that for every $\varepsilon > 0$ there exists an integer $p = p(\varepsilon) > 0$ such that E_{pn} is ε-dense in E_n for every $n \geqslant 0$.

Once all the scaling functions have been extended, any function f in $\mathcal{L}^2(X)$ can be extended to a function $\bar{f}$ in $\mathcal{L}^2(\bar{X})$, by expanding f onto the scaling functions and then extending each term of the expansion. How far the support of $\bar{f}$ extends outside of $X \subset \bar{X}$ is dependent on properties of f, on the embedding of X in $\bar{X}$ and on the function spaces E_n we use to extend. This automatic adaptiveness allows to tune the extension locally to smoothness and frequency properties of f and of the embedding of X in $\bar{X}$. This is a great advantage of this technique compared to many current techniques, which provide extensions of functions at a distance from the set which is either fixed (and usually "small," in the sense that its based on a worst-case estimation of the property of the function on the set), or tend to become numerically unstable when one tries to extend too far away from the set.

When the diffusion operator is induced on the data set by a kernel that is defined outside the data set, it is also possible to use the Nyström extension [74–76], which however has the disadvantages described above.

For a survey of extension techniques in the context of non-linear dimension reduction see [38] and references therein, as well as [3,4].

Example 43. We illustrate the construction above by extending one of the scaling functions on the dumbell-shaped manifold already considered in Section 2.6. We consider the approximate multiresolution of Gaussian kernels as in (iii) above as candidate function space to be used to extend the scaling function in question. Instead of extending to the whole $\mathbb{R}^3$, which would not be useful since we expect the extension to be quite concentrated on the manifold X, we extend to a large set of points Y obtained by adding white noise to the original manifold. See Fig. 18.

10. Comparison with previous work

The last 30 years have seen a flurry of activity around multiscale analysis, in particular in harmonic analysis, scientific computation, image processing, among many others. Much of the development of multiscale analysis and Littlewood–Paley theory can already be found in [5,21]. Here we are merging the two points of views.

Harmonic analysis of clouds of points has been considered in the past by [77–79] where the eigenfunctions of the Laplace operator, or multiresolution constructions, on particular classes of meshes are used to filter, smooth and compress the point set associated with a 3D mesh. This is achieved by filtering the coordinate functions, viewed as functions on the cloud of points. Our multiscale methodology enables adaptive (both in the choice of the diffusion operator and in the choice of scale at each location) filtering, smoothing, and denoising of the point set and in this respect it is most similar to [80]. See also [81] for applications of diffusion to implicit mesh fairing. In image processing and vision, the ideas of scale spaces, through linear and non-linear diffusion, have had a great influence.

In [2] the authors use wavelet bases to compress the matrices representing large classes of operators, including Calderón–Zygmund operators and pseudo-differential operators, by expanding the kernel itself on a wavelet basis, and they show this leads to matrices with $\mathcal{O}(n \log n)$ (or even $\mathcal{O}(n)$ for large classes of operators) entries above precision, allowing for fast matrix operations such as matrix–vector multiplication. The matrices representing the operator in this

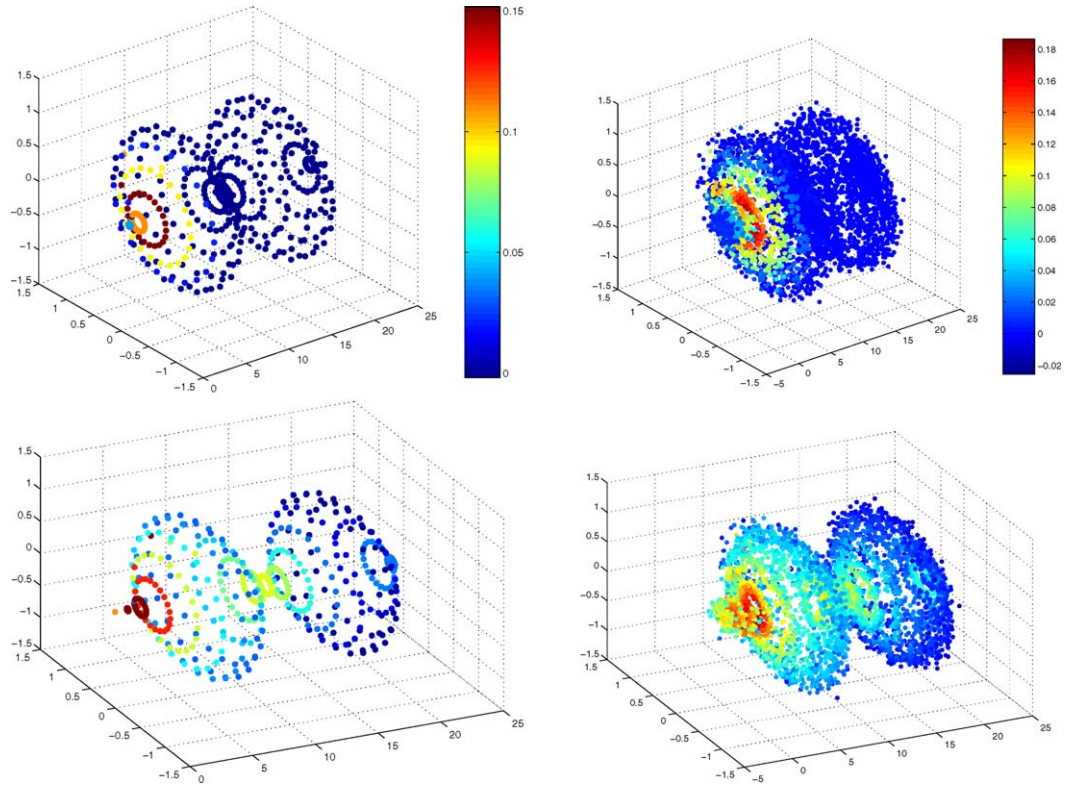

Fig. 18. Top left: the diffusion scaling function $\varphi_{3,3}$ color mapped on the dumbell-shaped manifold sampled at 500 points. Top right: extension of $\varphi_{3,3}$ to 5000 points obtained by adding white Gaussian noise to the points the dumbell. Bottom left and right: as the corresponding pictures above, but for $\varphi_{10,1}$.

"non-standard form," as they call it, are similar to the compressed matrices we obtain with our algorithm. In a way we are generalizing the construction in [2] by building wavelet-like bases adapted to the geometry of the operator.

Our construction is in many ways related to algebraic (or rather mixed algebraic–geometric) multigrid techniques (see the pioneer paper [46], but the literature on the subject is vast). Our construction suggests to look at the powers of the operator and let them dictate the form of the coarsening operators, at different scales, in a natural way and with high precision.

In [44,45] classes of matrices for which fast arithmetic exist are introduced and in subsequent papers various classes of interesting operators arising from partial differential equations are shown to be in these classes. These matrices are not sparse, but admit a sparse multiscale encoding structure; in this work the powers of the operator, and from them a function of the operator, is encoded in multiscale fashion as the action of sparse matrices at different scales.

In [42,43] techniques for compressing low-rank matrices (see also [63]) are applied to the problem of obtaining an explicit sparse inverse, in particular for the solution of integral equations arising from boundary value problems of elliptic PDEs. The intersection with our work lies in the idea of compressing a low-rank matrix in a multiscale fashion; however the emphasis in our work is to obtain efficient representations of powers of an operator, not on the computation of the inverse of a fixed operator.

The construction of wavelets via lifting due to Sweldens [82–86] is in some respects similar to ours. The lifting schemes yields (formally) biorthogonal wavelets by a sequence of "lifting steps," each of which transforms a (formally) biorthogonal pair of bases into another with more desirable properties, e.g., by adding smoothness of the primal wavelets and/or adding vanishing moments to the dual wavelets. The setting is quite general, in particular arbitrary lattices, varying from scale to scale, can be used. What dictates the refinement is not the action of a dilation operator but the subdivision scheme that refines a mesh into a finer one. The geometric refinement implies the functional refinement of the associated scaling functions and hence the multiresolution analysis. It also determines the stability of the obtain system, which in general cannot be guaranteed, unless regularity assumption are made in the

construction and refinement of the meshes. An improved construction of a subdivision scheme with good regularity properties for arbitrary topologies is proposed in [87] and its applications to surface compression studied.

In [88] the authors propose a construction of semi-orthogonal and biorthogonal wavelets based on regular subdivision schemes on samples from surfaces of arbitrary topology. The geometry of the subdivision scheme dictates the refinement equations for the scaling functions and wavelets, and the corresponding multiresolution. The dilation operator is incorporated into the action of the subdivision scheme itself. They suggest applications of their techniques to surface and texture compression. The applications of multiscale techniques have been explored in several papers, please see [88] and references therein.

In [89] the author constructs orthonormal wavelets on stratified Lie groups, which are smooth and well localized in time and frequency. The dilations are given by automorphism of the Lie group, which can be assimilated to a particular case of anisotropic diffusion in our language.

11. Conclusion

We have shown that in the context of manifolds, graphs, data sets, and general metric spaces, diffusion processes and Markov processes can be analyzed in a multiscale fashion very much in the spirit of classical wavelet analysis. We propose fast and stable algorithms for constructing orthonormal bases for the multiscale approximation spaces, and we show that there is an efficient scaling function and wavelet transform.

12. Work in progress

(i) *Biorthogonal multiscale spaces*. Instead of orthonormal bases for the subspaces V_j we could have Riesz–Bases with good constants. This leads to the construction of biorthogonal diffusion wavelets, which is of interest in the multiresolution analysis of diffusion operators which are not self-adjoint and non-symmetric Markov chains [55]. A first construction is presented in [54].

(ii) *Sharper or smoother frequency filters*. Instead of looking at the spectrum of the powers of T and thresholding on it to obtain the approximation subspaces V_j, we consider thresholding a function m_j of the spectrum (which correspond, via the spectral theorem, to a function of the operator T), which "sharpens" (to better localize the spectrum numerically) or smoothness the spectrum of the low-pass frequency portion. We expect sharpening to improve numerical stability and speed, and allow greater control on the subsampling rate and mechanism in general. Smoothing, on the other hand, allows to better control support size and vanishing moments of the scaling functions. Details will be discussed in [55].

(iii) *Wavelet packets*. Wavelet packets can be constructed by considering P_j and $P_j - P_{j+1}$ (and polynomials in them, as suggest in (ii)) to further split the wavelet subspaces. This will allow to have best basis algorithms, for compression, denoising and local discrimination for functions on manifolds, graphs, and so on. The companion paper [12] presents the details of the construction.

(iv) This construction provides an efficient and fast technique for *compressing* and *denoising* functions on the data set, and the data set itself (by viewing the coordinate functions as functions on the data). This has applications to empirical function approximation, learning, denoising, surface compression. The multiscale spaces V_j seem to provide a natural setting for structural complexity minimization and approximation, and the extensions outside the data set, as well as efficient algorithms available for all these computations, seem to hold good promises for problems of regression, classification, and learning. They also provide a natural local multiscale structure that lacks in many learning algorithms or is sought after through more complicate approaches.

(v) Compression of *integral operators* with non-oscillatory kernels, with boundaries of arbitrary geometries, in arbitrary dimension, with no need of providing their parametrization. We do not know how this scheme or some modification of it would perform in compressing oscillatory kernels.

(vi) Randomization. This algorithm can be randomized as soon as a randomized modified Gram–Schmidt with some sparsity-preserving properties (e.g., with geometric pivoting) is available.

(vii) Hierarchical coarse-graining of graphs and document directories. The multiscale organization of graphs that we propose can be used to for matching graphs at different scale or organizational levels. The diffusion distance seems a very natural and promising measure for quantitating relationships between documents and a multiscale analysis of such structures can be expected to provide an interesting organization of such data.

(viii) One can associate a graph with an image (or an hyperspectral image), for example by considering features of the image and similarities between features, and use our methodology for filtering, smoothing, denoising, and compressing, in a flexible and adaptive way, the image and its feature space.

(ix) Multiscale analysis of corpora of documents and web pages. In this context the multiscale scaling functions constructed in this paper are directories of documents and the corresponding directories of words ("concept"), at different scales. The whole multiscale analysis allows to construct a natural hierarchy of directories, topics, and "concepts" for a given corpus of documents. This includes a detailed study of the multiscale construction on directed graphs and associated diffusion processes.

(x) Applications to learning. Learning can often be viewed as an approximation problem under smoothness (or complexity) constraints. Diffusion wavelets can potentially be used as good bases for many interesting smoothness spaces on graphs and manifolds and be successfully employed in learning problems. Their multiscale structure also allows them to be used as multiscale kernels in kernel methods. The multiscale extension presented above is promising in view of extending learned functions to new samples. Applications to reinforcement learning, in particular Markov Decision Processes, are presented in [90–94].

Acknowledgments

We thank James C. Bremer Jr., Stephane Lafon, Martin Mohlenkamp, Raanan Schul, Arthur Szlam, Mark Tygert, and Steven W. Zucker for useful discussions and suggestions during the preparation of the manuscript. We thank the reviewer for several suggestions during the revision process that greatly helped improve the presentation. This work was partially supported from AFOSR/DARPA ISP; M.M. is also grateful for the partial support from NSF (DMS 0501250) and to Institute for Pure and Applied Mathematics, which he was visiting while part of this paper was being written.

References

[1] L. Greengard, V. Rokhlin, A fast algorithm for particle simulations, J. Comput. Phys. 73 (1987) 325–348.
[2] G. Beylkin, R.R. Coifman, V. Rokhlin, Fast wavelet transforms and numerical algorithms, Comm. Pure Appl. Math. 44 (1991) 141–183.
[3] S. Lafon, Diffusion maps and geometric harmonics, Ph.D. thesis, Yale University, 2004.
[4] R. Coifman, S. Lafon, Geometric harmonics, Technical report, Yale University, 2003.
[5] E. Stein, Topics in Harmonic Analysis Related to the Littlewood–Paley Theory, Princeton Univ. Press, 1970.
[6] R.R. Coifman, Y. Meyer, Wavelets: Calderón–Zygmund and Multilinear Operators, Cambridge Univ. Press, 1993.
[7] G. David, J. Journé, A boundedness criterion for generalized Calderón–Zygmund operators, Ann. of Math 120 (2) (1984) 371–391.
[8] G. David, J. Journé, S. Semmes, Opérateurs de Calderón–Zygmund, fonctions para-accrétives et interpolation, Rev. Mat. Iberoamericana 1 (1985) 1–56.
[9] J. Gilbert, Y. Han, J. Hogan, J. Lakey, D. Weiland, G. Weiss, Smooth molecular decompositions of functions and singular integral operators, Mem. Amer. Math. Soc. 156 (2002).
[10] Y. Han, Calderón-type reproducing formula and the Tb theorem, Rev. Mat. Iberoamericana 10 (1) (1994) 51–91.
[11] A. Nahmod, Generalized uncertainty principles on spaces of homogeneous type, J. Funct. Anal. 119 (1994) 171–209.
[12] J.C. Bremer, R.R. Coifman, M. Maggioni, A.D. Szlam, Diffusion wavelet packets, Technical report YALE/DCS/TR-1304, Yale University.
[13] F. Chung, Spectral Graph Theory, vol. 92, CBMS, Amer. Math. Soc., 1997.
[14] E. Davies, Heat Kernels and Spectral Theory, Cambridge Univ. Press, 1989.
[15] P. Li, S.T. Yau, On the parabolic kernel of the Schrödinger operator, Acta Math. 156 (1986) 153–201.
[16] C.E. Priebe, D.J. Marchette, Y. Park, E.J. Wegman, J.L. Solka, D.A. Socolinsky, D. Karakos, K.W. Church, R. Guglielmi, R.R. Coifman, D. Link, D.M. Healy, M.Q. Jacobs, A. Tsao, Iterative denoising for cross-corpus discovery, 2004, in preparation.
[17] R.R. Coifman, M. Maggioni, Multiscale analysis of document corpora, in preparation.
[18] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 6 (15) (2003) 1373–1396.
[19] M. Brand, Continuous nonlinear dimensionality reduction by kernel eigenmaps, in: Proc. IJCAI, 2003.
[20] R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal., in press.
[21] R.R. Coifman, G.L. Weiss, Analyse harmonique noncommutative sur certains espaces homogenes, Springer-Verlag, 1971.
[22] R. Macias, C. Segovia, Lipschitz functions on spaces of homogeneous type, Adv. Math. 33 (1979) 257–270.
[23] R. Macias, C. Segovia, A decomposition into atoms of distribution on spaces of homogeneous type, Adv. Math. 33 (1979) 271–309.
[24] M. Bownik, Anisotropic Hardy spaces and wavelets, Mem. Amer. Math. Soc. 781 (2003).
[25] A. Nagel, E. Stein, S. Wainger, Balls and metrics defined by vector fields I: Basic properties, Acta Math. 155 (1985) 103–147.
[26] M. Christ, A $T(b)$ theorem with remarks on analytic capacity and the Cauchy integral, Colloq. Math. 2 (60/61) (1990) 601–628.
[27] G. David, Wavelets and Singular Integrals on Curves and Surfaces, Springer-Verlag, 1991.
[28] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.

[29] W. Goetzmann, P.W. Jones, M. Maggioni, J. Walden, Beauty is in the eye of the beholder, preprint.
[30] R. Laugesen, E. Wilson, N. Weaver, G. Weiss, A characterization of the higher dimensional groups associated with continuous wavelets, J. Geom. Anal. 12 (1) (2002) 89–102.
[31] M. Maggioni, Wavelet frames on groups and hypergroups via discretization of Calderón formulas, Monatsh. Math. 143 (2004) 299–331.
[32] K. Trimeche, Generalized Wavelets and Hypergroups, Taylor & Francis, 1997.
[33] M. Belkin, P. Niyogi, Using manifold structure for partially labelled classification, in: Adv. NIPS, vol. 15, 2003.
[34] P. Niyogi, X. He, Locality preservity projections, Technical report TR-2002-09, University of Chicago, 2002.
[35] D.L. Donoho, C. Grimes, Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. (2003) 5591–5596; Also technical report, Stanford University.
[36] J. Ham, D. Lee, S. Mika, B. Schölkpf, A kernel view of the dimensionality reduction of manifolds, Technical report, Max-Planck-Institut für Biologische Kybernetik, Tübingen, July 2003.
[37] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[38] Y. Bengio, J. Paiement, P. Vincent, Out-of-sample extensions of lle, isomap, mds, eigenmaps, and spectral clustering, in: NIPS, vol. 16, 2004.
[39] S. Riemenschneider, Z. Shen, Box Splines, Cardinal Series, and Wavelets, Academic Press, 1991.
[40] N. Linial, E. London, Y. Rabinovich, The geometry of graphs and some of its algorithmic applications, Combinatorica 15 (1995) 215–245.
[41] A. Nahmod, Geometry of operators and spectral analysis, Ph.D. thesis, Yale University, 1991.
[42] H. Cheng, Z. Gimbutas, P. Martinsson, V. Rokhlin, On the compression of low rank matrices, Technical report YALEU/DCS/RR-1251, Yale University, July 2003.
[43] P.G. Martinsson, V. Rokhlin, A fast direct solver for boundary integral equations in two dimensions, Technical report YALEU/DCS/RR-1264, Yale University, 2004.
[44] W. Hackbusch, A sparse matrix arithmetic based on H-matrices I. Introduction to H-matrices, Computing 62 (2) (1999) 89–108.
[45] W. Hackbusch, S. Börm, Data-sparse approximation by adaptive H^2-matrices, Computing 69 (1) (2002) 1–35.
[46] A. Brandt, S. McCormick, J. Ruge, Algebraic multigrid (AMG) for automatic multigrid solution with applications to geodetic computations, Technical report, Institute for Computational Studies, Fort Colins, CO, 1982.
[47] O. Livne, Multiscale eigenbasis algorithms, Ph.D. thesis, Department Computer Sciences & Applied Mathematics, Weizmann Institute, Rehovot, 2000.
[48] J. Heinonen, Lectures on Analysis on Metric Spaces, Springer-Verlag, New York, 2001.
[49] S.G. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.
[50] Y. Meyer, Ondelettes et operatéurs, Hermann, Paris, 1990.
[51] I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, 1992.
[52] R.R. Coifman, Multiresolution analysis in nonhomogeneous media, in: J.M. Combes, A. Grossman, Ph. Tchamitchian (Eds.), Wavelets: Time–Frequency Methods and Phase Space, Springer-Verlag, New York, 1989.
[53] A. Cohen, I. Daubechies, J. Faveau, Biorthogonal bases of compactly supported wavelets, Comm. Pure Appl. Math. XLV (1992) 485–560.
[54] M. Maggioni, J.C. Bremer Jr., R.R. Coifman, A.D. Szlam, Biorthogonal diffusion wavelets for multiscale representations on manifolds and graphs, in: Wavelet XI, in: Proc. SPIE, vol. 5914, 2005, 59141M.
[55] J.C. Bremer, R.R. Coifman, M. Maggioni, Biorthogonal diffusion wavelets, in preparation.
[56] M. Maggioni, A.D. Szlam, R.R. Coifman, J.C. Bremer Jr., Diffusion-driven multiscale analysis on manifolds and graphs: Top–down and bottom–up constructions, in: Wavelet XI, in: Proc. SPIE, vol. 5914, 2005, 59141D.
[57] A. Björck, Numerics of Gram–Schmidt orthogonalization, Linear Algebra Appl. 197 (1994) 297–316.
[58] G. Golub, C.V. Loan, Matrix Computations, Johns Hopkins Univ. Press, 1989.
[59] J. Daniel, W. Gragg, L. Kaufman, G. Stewart, Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization, Math. Comput. 30 (1976) 772–795.
[60] P. Indyk, R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, 1998, pp. 604–613; http://citeseer.ist.psu.edu/article/indyk98approximate.html.
[61] R. Krauthgamer, J. Lee, M. Mende, A. Naor, Measured descent: A new embedding method for finite metrics, in preparation.
[62] J. Strömberg, A modified Franklin system and higher order spline systems in $\mathbb{R}^n$ as unconditional bases for Hardy spaces, in: Conference in Honor of A. Zygmund of Wadsworth Math. Ser., 1982, p. 475.
[63] M. Gu, S. Eisenstat, Efficient algorithms for computing a strong rank-revealing QR factorization, SIAM J. Sci. Comput. 17 (4) (1996) 848–869.
[64] L. Grasedyck, W. Hackbusch, S.L. Borne, Adaptive geometrically balanced clustering of h-matrices, Computing (2004), in press.
[65] N. Sidorova, O.G. Smolyanov, H. Weizsäcker, O. Wittich, Brownian motion close to submanifold of Riemannian manifolds, preprint, 2003.
[66] O.G. Smolyanov, H. v Weizsäcker, O. Wittich, Brownian motion on a manifold as limit of stepwise conditioned standard Brownian motions, in: Stochastic Processes, Physics and Geometry: New Interplays, in: Can. Math. Soc. Conf. Proc., vol. 29, Amer. Math. Soc., 2000, pp. 589–602.
[67] I. Daubechies, I. Guskov, P. Schröder, W. Sweldens, Wavelets on irregular point sets, Philos. Trans. R. Soc. London A 357 (1760) (1999) 2397–2413.
[68] J. Simoens, S. Vandewalle, A stabilized lifting construction of wavelets on irregular meshes on the interval, SIAM J. Sci. Comput. 4 (24) (2003) 1356–1378.
[69] D.L. Donoho, I. Johnstone, Ideal denoising in an orthonormal basis chosen from a library of bases, Technical report, Stanford University, 1994.
[70] R.R. Coifman, M. Maggioni, Multiscale data analysis with diffusion wavelets, Technical report YALE/DCS/TR-1335, Yale University, September 2005.

[71] R. Coifman, S. Lafon, Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions, Appl. Comput. Harmon. Anal., in press.
[72] D. Slepian, H. Pollack, Prolate spheroidal wave functions, Fourier analysis and uncertainty I, Bell Syst. Tech. J. 40 (1961) 43–64.
[73] D. Slepian, Prolate spheroidal wave functions, Fourier analysis and uncertainty IV: Extensions to many dimensions; generalized prolate spheroidal wave functions, Bell Syst. Tech. J. 43 (1964) 3009–3058.
[74] C. Fowlkes, S. Belongie, F. Chung, J. Malik, Spectral grouping using the Nyström method, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2) (2004) 214–225.
[75] C. Fowlkes, S. Belongie, J. Malik, Efficient spatiotemporal grouping using the Nyström method, CVPR.
[76] C.K.I. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, in: NIPS, 2000, pp. 682–688.
[77] Z. Karni, C. Gotsman, Spectral compression of mesh geometry, in: Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH'00, ACM Press/Addison–Wesley Publishing Co., 2000, pp. 279–286.
[78] G. Taubin, A signal processing approach to fair surface design, in: Proc. 22nd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH'95, ACM Press, 1995, pp. 351–358.
[79] G. Taubin, T. Zhang, G.H. Golub, Optimal surface smoothing as filter design, in: ECCV, vol. 1, 1996, pp. 283–292.
[80] I. Guskov, W. Sweldens, P. Schröder, Multiresolution signal processing for meshes, in: Proc. 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH'99, ACM Press/Addison–Wesley Publishing Co, 1999, pp. 325–335.
[81] M. Desbrun, M. Meyer, P. Schröder, A.H. Barr, Implicit fairing of irregular meshes using diffusion and curvature flow, in: Proc. 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH'99, ACM Press/Addison–Wesley Publishing Co., 1999, pp. 317–324.
[82] I. Guskov, W. Sweldens, P. Schröder, Multiresolution signal processing for meshes, in: Computer Graphics Proceedings, SIGGRAPH'99, 1999, pp. 325–334.
[83] P. Schröder, W. Sweldens, Spherical wavelets: Efficiently representing functions on the sphere, in: Computer Graphics Proceedings, SIGGRAPH'95, 1995, pp. 161–172.
[84] W. Sweldens, The lifting scheme: A custom-design construction of biorthogonal wavelets, Appl. Comput. Harmon. Anal. 3 (2) (1996) 186–200.
[85] W. Sweldens, The lifting scheme: A construction of second generation wavelets, SIAM J. Math. Anal. 29 (2) (1997) 511–546.
[86] J. Kovačević, W. Sweldens, Wavelet families of increasing order in arbitrary dimensions, IEEE Trans. Image Process. 9 (3) (2000) 480–496.
[87] D. Zorin, P. Schröder, W. Sweldens, Interpolating subdivision for meshes with arbitrary topology, in: Computer Graphics Proceedings, SIGGRAPH'96, 1996, pp. 189–192.
[88] M. Lounsbery, T.D. DeRose, J. Warren, Multiresolution analysis for surfaces of arbitrary topological type, ACM Trans. Graph. 16 (1) (1997) 34–73.
[89] P. Lemarie, Base d'ondelettes sur les groups de Lie stratifies, Bull. Soc. Math. France (117) (1989) 211–232.
[90] S. Mahadevan, M. Maggioni, Value function approximation with diffusion wavelets and Laplacian eigenfunctions, Technical report TR-2005-38, University of Massachusetts, Department of Computer Science, in: Proc. NIPS, 2005.
[91] M. Maggioni, S. Mahadevan, Fast direct policy evaluation using multiscale analysis of Markov diffusion processes, Technical report TR-2005-39, University of Massachusetts, Department of Computer Science, accepted at ICML 2006, 2005.
[92] S. Mahadevan, K. Ferguson, S. Osentoski, M. Maggioni, Simultaneous learning of representation and control in continuous domains, in preparation.
[93] S. Mahadevan, M. Maggioni, Proto-value functions: A spectral framework for solving Markov decision processes, J. Mach. Learn. Res., submitted for publication.
[94] M. Maggioni, S. Mahadevan, Multiscale diffusion bases for policy iteration in Markov decision processes, J. Mach. Learn. Res., submitted for publication.