Harmonic Analysis of Digital Data Bases

Ronald R. Coifman and Matan Gavish

Abstract Digital databases can be represented by matrices, where rows (say) correspond to numerical sensors readings, or features, and columns correspond to data points. Recent data analysis methods describe the local geometry of the data points using a weighted affinity graph, whose vertices correspond to data points. We consider two geometries, or graphs - one on the rows and one on the columns, such that the data matrix is smooth with respect to the "tensor product" of the two geometries. This is achieved by an iterative procedure that constructs a multiscale partition tree on each graph. We use the recently introduced notion of Haar-like bases induced by the trees to obtain tensor Haar-like bases for the space of matrices, and show that an ℓ_p entropy conditions on the expansion coefficients of the database, viewed as a function on the product of the geometries, imply both smoothness and efficient reconstruction. We apply this methodology to analyze, de-noise and compress a term–document database. We use the same methodology to compress matrices of potential operators of unknown charge distribution geometries and to organize Laplacian eigenvectors, where the data matrix is the "expansion in Laplace eigenvectors" operator.

Key words: Data matrices, Transposable arrays, Haar-like wavelets, Tensor Haar basis, Strömberg's theorem, Coupled diffusion geometry

Ronald R. Coifman
Program in Applied Mathematics, Yale University, 51 Prospect St, New Haven, CT 06511, USA,
e-mail: `coifman-ronald@yale.edu`

Matan Gavish
Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305, USA
and Program in Applied Mathematics, Yale University. e-mail: `gavish@stanford.edu`

1 Introduction

> We should seek out unfamiliar summaries of observational material, and establish their useful properties. - J. W. Tukey [21]

There is a fruitful interplay between Harmonic analysis and the vast data analysis challenges that the scientific community is facing. Ideas that in the past have been applied for the analysis and extraction of information from physical systems are being increasingly applied, in their computational reincarnation, to organize and to extract information from high dimensional digital data sets of arbitrary source. Some examples are Laplacian eigenfunctions [2, 6] and wavelet bases [7].

In [20], J.O. Strömberg showed that the tensor product of Haar bases is extremely efficient in representing functions on product structures. For example, when $f : [0,1]^d \to \mathbb{R}$ has bounded mixed derivatives, he considered the tensor product of Haar bases in all d dimensions of and showed that f can be approximated with L_∞ error $\mathscr{O}\left(\varepsilon \log^{d-1}\left(\frac{1}{\varepsilon}\right)\right)$ by shrinking to zero coefficients of the basis functions with support smaller than ε. As only $\mathscr{O}\left(\frac{1}{\varepsilon}\log^{d-1}\left(\frac{1}{\varepsilon}\right)\right)$ coefficients survive, this yields a remarkable compression scheme.

In fact, *product structures* are among the most common data structures in existence. Consider the usual data matrix, abundant in multivariate statistics. By this we mean a rectangular array with n columns, representing (say) observations or individuals or data points, and p rows, representing attributes or variables measured on each data point. More generally, when each data point responds simultaneously to $d-1$ variable sets, the data is given as a rank-d tensor. While the theory we will develop will include the general case of tensor data structures of arbitrary rank d, the examples discussed are data matrices where $d = 2$. Data matrices arise in text term-document analysis, microarray data analysis, gene association studies, sensor networks, recommendation engines, hyperspectral image processing - to name just a few scenarios. Here, the basic assumption underlying matrix analysis in multivariate statistics, namely that observations are independent and identically distributed, breaks down, since in general correlations exist among both rows and columns of the data matrix. Models for data of this form have been introduced in the statistical literature (e.g. Plaid Models [15] and Transposable Regularized Covariance Models [1]), yet the contrast between the overwhelming wealth of applications and the tools available remains considerable. In this work, we suggest a harmonic analysis approach that leads to a nonparametric model for data tensors, and in particular for data matrices.

Since correlations exist among both rows and columns of the data matrix, there is no longer a preferred dimension among $\{$rows , columns$\}$. Treating rows and columns on equal footing naturally leads to tensor analysis. Strömberg's ideas, namely that tensor product of Haar bases sparsify smooth functions on a product space and that coefficients should be sorted by basis function support size, appears promising for efficiently analyzing matrices and higher order data structures. However, if we are to use this approach to analyze a matrix (a function on an abstract product space $M : \{observations\} \times \{variables\} \to \mathbb{R}$), we must first (1) make sense of the notion of a *geometry* and *Haar basis* on an abstract set such as $\{observations\}$ rather than on $[0,1]$, (2) find an algorithm to construct two such geometries, relative to which a given data matrix is "regular", and (3) find a computationally reasonable way to measure the regularity of a matrix with respect to two given geometries - on $\{observations\}$ and on $\{variables\}$. Only then can we hope to obtain an approximation result similar to Strömberg's.

1.1 Outline

Our goal here is to follow this program. For (1), the local geometry of each set is described by a weighted affinity graph. For (2), in §2.1 we describe a straightforward procedure to integrate an affinity graph into a multiscale partition tree. In §2.2 we then describe an iterative procedure to construct two coupled geometries, namely affinity graphs, given a data matrix.

As an interesting example of organization, in §4 we consider the matrix of Laplacian eigenfunctions on a graph, where the entries represent the value of an eigenfunction at a point. As any two eigenvectors are orthogonal, their mutual distance is constant and a naive attempt to build an eigenvector organization fails. On the other hand, if we organize the graph into a contextual folder hierarchy and relate two eigenvectors also through their correlation on the folders at different scales, we obtain a dual organization of eigenvectors. In the case of a d-dimensional torus $\mathbb{T}^d$, using this procedure one recovers the usual Euclidean geometry of the dual group, namely the lattice $\mathbb{Z}^d$.

In §3 we consider a different operator, namely a potential operators on clouds of unknown geometries. We show that it can be organized to reveal the intrinsic geometry of their domain and range.

In §5 we turn to function bases. We build on [12, 13] where, in collaboration with B. Nadler, we showed that a partition tree on an abstract set X induces a "Haar-like" orthonormal basis for $\{f : X \to \mathbb{R}\}$ with wavelet basis properties. Thus, the construction of a coupled geometry on $\{observations\}$ and $\{variables\}$ also induces Haar-like bases, say $\{\psi_i\}$ and $\{\varphi_j\}$ respectively. We prove an approximation result of Strömberg's type and show that a condition on the ℓ_p sum of expansion coefficients $\sum_{i,j} \left|\left\langle M, \psi_i \otimes \varphi_j \right\rangle\right|^p$ implies both efficiency of reconstruction and "bi-smoothness" of the data matrix with respect to the two trees constructed. Indeed the ℓ_p condition is better suited to this general scenario than a condition on mixed derivatives, which depends on the dimension d. In particular, this means that the ℓ_p condition measures the compatibility of the data matrix with the coupled geometry constructed, thus accomplishing (3) and yielding a stopping condition for the iterative procedure of §2.2. This approximation result leads to a decomposition of the dataset to a "smooth" part and an "irregular" part, the latter having a small support, in the spirit of the classical Calderón-Zygmund decomposition. In §7, we describe how a tensor-Haar basis, induced by a coupled geometry that is compatible with the data matrix, can be used to perform compression and statistical tasks such as denoising and missing value imputation on the data matrix.

1.2 An example

In §7 we study a term-document data set: a matrix M whose entry $M_{i,j}$ is the occurrence frequency of word i in document j. We define a graph structure on the documents, in which we link documents with a preponderance of common vocabulary. This particular graph is crude: highly related documents, which should be nearest neighbors in this graph, share just a small set of highly correlated words - but this correlation may drown in the overall noise. Still this graph allows to organize documents into an inaccurate hierarchy of contextual folders. In order to obtain an improved hierarchy, it is useful to consider conceptual groupings of words, thereby leading us to build an analogous graph structure on the vocabulary, in which we link two words if they occur frequently in the same documents as well as in the contextual folders of documents. This leads to a hierarchy of conceptual word folders, which are then used to refine and rebuild the original document graph and contextual folder structure, and so on (Figure 1).

In each stage of the iteration, the partition trees constructed on the data set induce two Haar-like bases, one on documents and one on words. The original data base can now be viewed as a function on the tensor product of the two graphs, and can therefore be expanded in the tensor product of the two bases. At each stage we compute the ℓ_1 entropy of the tensor Haar coefficients as way of measuring the fit between the geometry and data base, and stop the iterations when no improvement is detected. Observe that by construction this procedure is invariant under permutations of matrix rows and columns, providing a tool for recovering an intrinsic organization of both - even when we are given a matrix with no known order on its rows or columns. Once a coupled geometry that is compatible with the matrix is achieved, we expand the data matrix in the tensor Haar basis and process the data base in the coefficient domain. For instance, the data matrix can now be compressed or de-noised using standard signal processing methodologies. While the Haar bases are not smooth, their construction is random and can be repeated and averaged, in order to eliminate some of the artifacts created, following [5].

In §8 we return to the example of potential operator and show that by transforming to the tensor Haar coefficient matrix, the operator becomes extremely compressed. In other words, by treating a potential operator as a data matrix, recovering its coupled geometry, transforming to a tensor Haar basis and thresholding coefficients, we get a fast numerical scheme.

Note that the ℓ_1 entropy condition of coefficients of an orthogonal expansion is used to quantify the "sparseness" of the expansion - a well known notion in the Compressed Sensing literature [11, 4]. It an easy observation that a function can be recovered to mean square error ε times the ℓ_1 entropy by using only the coefficients larger than ε, whose number does not exceed the ℓ_1 entropy times $\frac{1}{\varepsilon}$. In the special

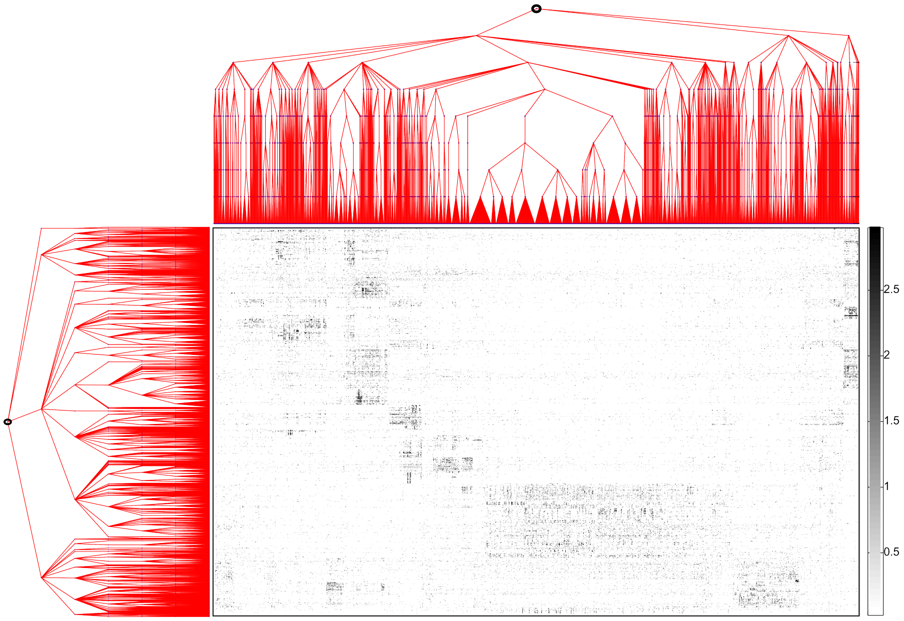


Fig. 1 The Science News data matrix discussed in §7.2. The partition trees on rows and columns are shown. The rows and columns have been re-ordered by the trees depth-first order.

case of the Haar or tensor Haar expansions, *only coefficients of Haar functions with support volume larger than ε are required*. This conveniently eliminates the need to sort the large set of coefficients.

2 The coupled geometry of questionnaires and databases

Suppose that

$$M = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix}$$

is a data matrix, whose columns correspond to observations. In order to organize the observations as the vertices of an affinity graph, a kernel-based approach would be to take the affinity $W_{i,j}$ between observations i and j to be, for instance, $exp\left(-\left\|v_i - v_j\right\|^2\right)$. In modern data analysis scenarios the dimensionality of v_i is often comparable or larger than n. In this scenario, different sets of observations might respond to different sets of variables whereas the correlation between any two complete vectors v_i and v_j is either very large or very small. Consider for example automatic medical tests that measure multiple parameters such as chemical blood composition or brain activity. Different patient groups (corresponding to different medical situations) will have very different response to different variable sets. In these scenarios, quantities such as $exp\left(-\left\|v_i - v_j\right\|^2\right)$ do not provide a useful a affinity among observations. Instead, we should consider partitioning both the observation and variable sets.

2.1 Describing a graph's global structure using a partition tree

The main computational tool we will use to construct a *coupled geometry* and, later, the orthonormal bases they induce, is a *partition tree*. By this we mean a sequence of increasingly refined partitions, readily described by a tree. As an affinity graph only holds "local" geometrical information relating each points to its neighbors, it does not directly allow computations involving the large scale structure of the graph. A convenient way to achieve this is by "integrating" the local distances into a partition tree.

Let $X = \{x_1, \dots x_N\}$ be a finite set. Consider a sequence of L finer and finer partitions of X, denoted $\mathscr{X}^1, \dots \mathscr{X}^L$. For each $1 \leqslant \ell \leqslant L$, the partition at level ℓ is composed of $n(\ell)$ mutually disjoint sets, which we call *folders*,

$$\mathscr{X}^\ell = \left\{X_1^\ell, \dots, X_{n(\ell)}^\ell\right\} \tag{1}$$

such that

$$X = \biguplus_{k=1}^{n(\ell)} X_k^\ell . \tag{2}$$

The finest partition, at level $\ell = L$, is composed of $n(L) = N$ singleton folders: $X_k^L = \{x_k\}$ for $k = 1 \dots N$. The coarsest partition, at level $\ell = 1$, is composed of a single folder, $X_1^1 = X$. The partitions are *nested* in the sense that for $1 < \ell \leqslant L$, each folder $X_k^\ell \in \mathscr{X}^\ell$ is a subset of a folder from $\mathscr{X}^{\ell-1}$. We let $subfolders(\ell,k) \subset \{1 \dots n(\ell+1)\}$ be the indices such that

$$X_k^\ell = \biguplus_{j \in subfolders(\ell,k)} X_j^{\ell+1} . \tag{3}$$

There are many approaches in the literature for constructing a multiscale partition tree given a symmetric adjacency matrix W describing a weighted graph on the vertex set X. We suggest using the following random bottom-up construction. The most refined nontrivial partition, $\mathscr{X}^{L-1}$ is created using a standard "k-means type" approach: For a fixed "radius" $\rho > 0$, a maximal set of centroid points $\{z_i\} \subset X$ such that $i \neq j \Rightarrow W(z_i, z_j) \geqslant \rho$ is selected at random. A partition of X is now obtained by grouping points by affinity from the centroids, namely setting $X_i^{L-1} = \left\{x \in X \,\middle|\, W(x,z_i) > W(x,z_j)\ \forall i \neq j\right\}$ for $i = 1\dots n(L-1)$, where $n(L-1) = \#\{z_i\}$.

In order to group the folder set $\mathscr{X}^{\ell-1} = \left\{X_1^{L-1}\dots X_{n(L-1)}^{L-1}\right\}$ and obtain the next partition, define an affinity between folders by

$$\tilde{W}(i,j) = \left\langle W\mathbf{1}_{X_i^{L-1}}, W\mathbf{1}_{X_j^{L-1}}\right\rangle = \sum_{x\in X_i^{L-1}}\sum_{y\in X_j^{L-1}} W^2(x,y),$$

so that the affinity between folders is measured in the next time-scale. The partition procedure is repeated for the set $\mathscr{X}^{\ell-1} = \left\{X_1^{L-1}\dots X_{n(L-1)}^{L-1}\right\}$ with the affinity $\tilde{W}$ to yield the next partition $\mathscr{X}^{L-2}$, and so on until a trivial partition $X_1^1 = X$ is reached. Below, we refer to the partition sequence $\mathscr{X}^1,\dots\mathscr{X}^L$ as a partition tree $\mathscr{T}$ on X.

We remark that for data sets of size small enough to allow computation of graph Laplacian eigenvectors, it is sometimes preferable to embed the set X in Euclidean space first (e.g. using a Diffusion Embedding [6]) and construct the partition tree there.

2.2 *An iterative procedure to construct a* coupled geometry

Suppose that $f,g : X \to \mathbb{R}$ are two functions on X and $\rho(\cdot,\cdot)$ is an affinity between functions. Typical choices for $\rho(f,g)$ include $exp\left(-\frac{1}{\varepsilon}\sum_{x\in X}(f(x)-g(x))^2\right)$ for some $\varepsilon > 0$. Another example is the absolute value of correlation coefficient, $\left|\frac{cov(f,g)}{\sigma(f)\cdot\sigma(g)}\right|$, whenever it exceeds a given cutoff. If f and g orthogonal, $\rho(f,g)$ would hold no information. But it is possible that f and g are strongly correlated if restricted to part of X. A classical example is sine functions of close frequencies, when compared on a small subset of $[0,1]$.

Given a partition tree $\mathscr{T}$ on X, define the affinity $\rho_{\mathscr{T}}(f,g)$ as follows. Let $\tilde{f}_{\mathscr{T}}\left(X_k^\ell\right) = \frac{1}{\left|X_k^\ell\right|}\sum_{x\in X_k^\ell} f(x)$ denote the average of f on the folder X_k^ℓ of $\mathscr{T}$. In particular $\tilde{f}_{\mathscr{T}}\left(X_k^L\right) = f(x_k)$. Thus $\tilde{f}_{\mathscr{T}} : \mathscr{T} \to \mathbb{R}$ is an extension of f, augmenting its original values by its average values on all folders in all levels of $\mathscr{T}$. By setting $\rho_{\mathscr{T}}(f,g) = \rho\left(\tilde{f}_{\mathscr{T}}, \tilde{g}_{\mathscr{T}}\right)$, we take into account the mutual behavior of f and g on all levels of $\mathscr{T}$.

Now consider a matrix M. Denote the set of columns of M by X, and the set of its rows by Y. When M is a data set, we interpret X as observations and Y as variables, features or sensors. Thus $M : X \times Y \to \mathbb{R}$.

Definition 1. For each $y \in Y$, the row $M_y(x) : x \mapsto M(x,y)$ of M is a function on X . For a given a partition tree $\mathscr{T}_X$ on X, define the *dual affinity* on Y by $W_{\mathscr{T}_X}(y_1,y_2) = \rho_{\mathscr{T}_X}(M_{y_1}, M_{y_2})$. Similarly, for each $x \in X$, the column $M_x(y) : x \mapsto M(x,y)$ of M is a function on Y . For a given partition tree $\mathscr{T}_Y$ on Y , define the *dual affinity* on X by $W_{\mathscr{T}_Y}(x_1,x_2) = \rho_{\mathscr{T}_Y}(M, M_{x_2})$.

We thus arrive at the following procedure to refine a given affinity W on the set X of columns of X:

Algorithm 1. 1. Integrate the affinity W as in §2.1 to obtain a partition tree $\mathscr{T}_X$ on X.
2. Compute the dual affinity $W_{\mathscr{T}_X}$ on Y as in Definition 1.
3. Integrate the affinity $W_{\mathscr{T}_X}$ to obtain a partition tree $\mathscr{T}_Y$ on Y.
4. Compute the *refined version* $\tilde{W}$ of W by setting $\tilde{W}$ to the dual affinity $W_{\mathscr{T}_Y}$ on X .

By iterating this cycle, we obtain stable and refined affinities both on X and on Y. Given a matrix M, the initial affinity W on X is either taken from an external source or taken to be the affinity between the columns of M. This procedure was introduced in [22] and applied for automatic translation using term-document matrices of the same text in different languages.

We now consider two examples from mathematics, rather than data analysis, which illustrate the notion of *coupled geometry* and the possibilities offered by an ability to recover it.

The experiments we describe below suggest that this iterative procedure converges to affinities (W_X, W_Y), which capture what we informally call *the coupled geometry of the matrix M*. However, we still need a quantitative way to determine convergence, and a stopping rule for the iterations. Equally important, we must be able to evaluate the performance of this algorithm and to compare it with other approaches to the same task. This leads to the following key questions:

1. How to quantify and measure the compatibility of the database to a proposed coupled geometry?
2. How can a coupled geometry, which is compatible with the database, be used for analyzing it?

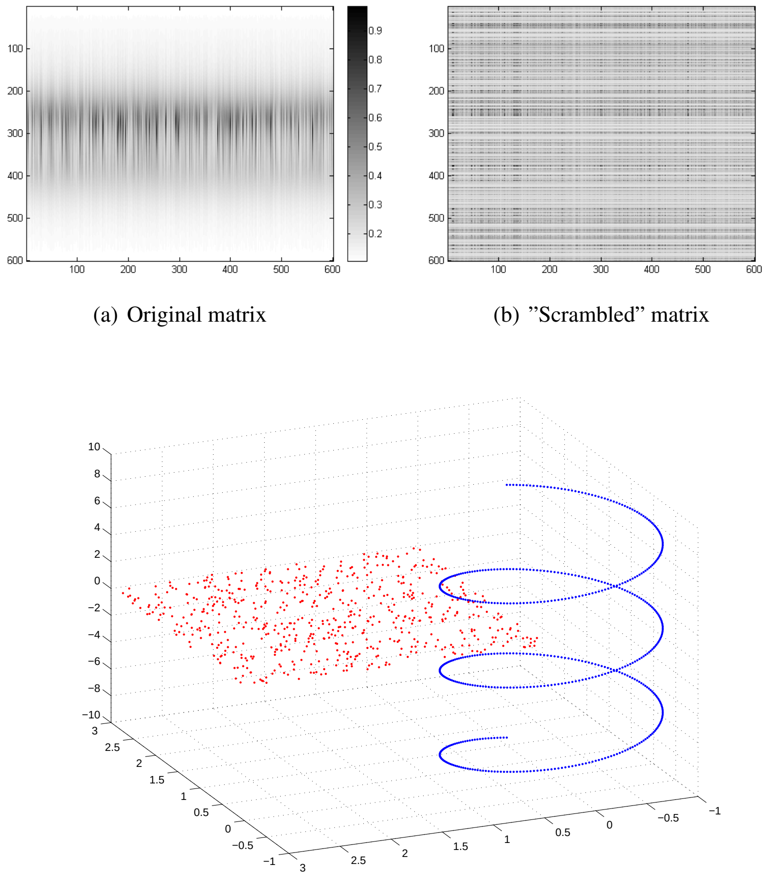
These questions are answered by the theory developed in §6.

3 Example: Numerical compression of a potential operator

As a first example of finding the *coupled geometry* of matrix rows and columns, consider the potential interaction between two point clouds in three dimension as in Fig. 2 (**c**). Let $\{x_i\}_{i=1}^N \subset \mathbb{R}^3$ and $\left\{y_j\right\}_{j=1}^N \subset \mathbb{R}^3$ denote the blue points (one dimensional helix) and the red points (two-dimensional sheet) respectively. The Coulomb potential operator is defined by

$$M_{i,j} = \frac{1}{\left\|x_i - y_j\right\|} .$$

Instead of the potential matrix M (Fig. 2 (**a**)), we are denied the spatial layout of the point clouds and given only the potential matrix with rows and columns in random order.



(a) Original matrix

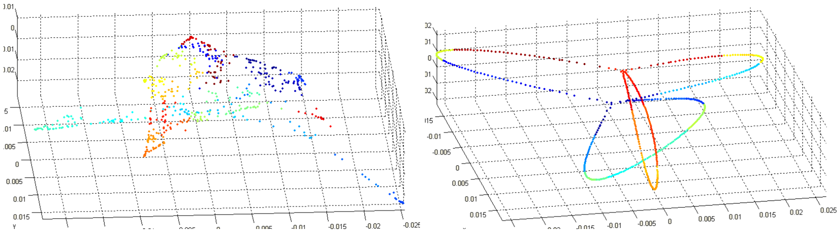
(b) "Scrambled" matrix

(c) Charge layout

Fig. 2 **(a)** Original matrix of potential operator. **(b)** After permuting rows and columns. **(c)** The spatial layout: the helix consists of the points $\{x_i\}_{i=1}^{600}$ and the flat sheet consists of the points $\{y_i\}_{i=1}^{600}$

In this case, after one iteration of the above procedure we recognize that there is a one-dimensional and a two-dimensional structure involved. Fig. 3 shows a diffusion embedding into three dimensional space (as in [6]) of graphs obtained for the matrix columns, corresponding to the two-dimensional charge plate, and of the graph obtained for the matrix rows, corresponding to the one-dimensional charge helix.

In contrast with the fast multipole method [14], which would treat the points in three dimensions, here a point is placed in the overall geometry according to its interactions with other points alone. This suggests a general method for organizing (and, as we will see in §8, for compression and fast calculations) of quite general potential operators on unknown geometries.



(a) column set (b) row set

Fig. 3 The 3-dimensional Diffusion embedding of the "scrambled" potential matrix (§3) row set and column set recovers their correct intrinsic dimensionality and geometry. Symmetries of the spatial layout are present in the potential operator and are reflected in this intrinsic geometry.

4 Example: Recovering the intrinsic geometry of Laplacian eigenfunctions

A second, interesting example for recovering a *coupled geometry* using the iterative procedure of §2 involves organizing a matrix whose rows are samples of some manifold Laplacian eigenfunctions.

On a general manifold, the only a-priori relationship between Laplacian eigenfunctions is through the corresponding eigenvalues, which provide a degenerate one dimensional geometry on the eigenfunctions. However, it is often clear that the set of eigenfunctions carries a much richer structure. As a canonical example, consider the Laplacian eigenvectors on the d-dimensional torus $\mathbb{T}^d$. This set is indexed by the dual group, $\mathbb{Z}^d$. The structure of eigenfunctions can thus be organized using the Euclidean geometry $\mathbb{Z}^d$. This organization, whereby (in $d = 2$, say) the eigenfunctions $\varphi_{k,m}(x,y) = sin(2\pi kx)\,sin(2\pi my)$ and $\varphi_{k,m}(x,y) = sin(2\pi(k+1)x)\,sin(2\pi(m+1)y)$ seem *identical* on scales much larger and much smaller than $\left(\frac{1}{k},\frac{1}{m}\right)$ is fundamental in Fourier analysis. Extending this fact, we claim that by considering the correlation between eigenvectors on different levels of the a multiscale decomposition of the manifold, we can define a "dual geometry" on eigenfunctions. This introduces a variety of questions in analysis, concerning the relation between this dual geometry and the properties of the manifold.

4.1 Recovering the geometry of circle eigenfunctions using ”partition-based correlation”: a calculation

Model the d-dimensional torus $\mathbb{T}^d$ by $[0,1]^d$. The fixed dimension d is suppressed below. For $\ell \geqslant 0$, denote by $K_\ell = 2^{-\ell}\cdot\left\{0,\ldots,2^\ell-1\right\}^d$ the square mesh of resolution $2^{-\ell}$ in $\mathbb{T}^d$. Consider the partition tree $\mathscr{T}$ with partition at level ℓ given by $\mathscr{X}^\ell = \biguplus_{\mathbf{k}\in K_\ell}\left(\mathbf{k}+\left[0,2^{-\ell}\right)^d\right)$. Let ψ be a function such that $\sum_{\mathbf{k}\in K_\ell}\psi\left(2^\ell\mathbf{x}-\mathbf{k}\right)$ approximates the constant function $\mathbf{1}$ on $\mathbb{T}^d$ for each ℓ, $\int\psi(\mathbf{x})\,d\mathbf{x}=1$, $\psi\geqslant 0$, $\frac{\partial}{\partial\xi_i}\hat{\psi}(0)=0$ and $\frac{\partial^2}{\partial\xi_i\partial\xi_j}\hat{\psi}(0)=C\delta_{i,j}$ where $\hat{\psi}$ is the Fourier transform. For example, ψ can be the indicator of $[0,1)^d$.

Definition 2. Define the partition-based correlation (with respect to the partition $\mathscr{X}^\ell$ and weight ψ) of functions $f,g\in L_2\left(\mathbb{T}^d\right)$ by

$$\rho_\ell(f,g)=\sum_{\mathbf{k}\in K_\ell}\left|\int_{\mathbb{T}^d} f(\mathbf{x})\,\overline{g(\mathbf{x})}\,\psi\left(2^\ell\mathbf{x}-\mathbf{k}\right)d\mathbf{x}\right|.$$

For eigenfunctions of $\mathbb{T}^d$ corresponding to eigenvalues $\mathbf{m},\mathbf{m}'\in\mathbb{Z}^d$ we have

$$\rho_\ell\left(e^{i\mathbf{m}\cdot\mathbf{x}},e^{i\mathbf{m}'\cdot\mathbf{x}}\right)=\sum_{\mathbf{k}\in K_\ell}\left|\int_{\mathbb{T}^d} e^{i\left(\mathbf{m}-\mathbf{m}'\right)\cdot\mathbf{x}}\psi\left(2^\ell\mathbf{x}-\mathbf{k}\right)d\mathbf{x}\right|=$$

$$=\sum_{\mathbf{k}\in K_\ell}2^{-\ell d}\left|\hat{\psi}\left(\frac{\mathbf{m}-\mathbf{m}'}{2^\ell}\right)\right|=\left|\hat{\psi}\left(\frac{\mathbf{m}-\mathbf{m}'}{2^\ell}\right)\right|.$$

To see that this recovers the Euclidean affinity of $\mathbb{Z}^d$, recall the an affinity needs hold only for neighbors. Indeed, to second order in $2^{-\ell}\left\|\mathbf{m}-\mathbf{m}'\right\|$, this gives

$$\rho_\ell\left(e^{i\mathbf{m}\cdot\mathbf{x}},e^{i\mathbf{m}'\cdot\mathbf{x}}\right)\approx\left|1-\frac{C}{2}\cdot 2^{-2\ell}\cdot\left\|\mathbf{m}-\mathbf{m}'\right\|^2\right|$$

so that

$$2^{2\ell}\left(1-\rho_\ell\left(e^{i\mathbf{m}\cdot\mathbf{x}},e^{i\mathbf{m}'\cdot\mathbf{x}}\right)\right)\propto\left\|\mathbf{m}-\mathbf{m}'\right\|^2.$$

To improve the approximation, we can define $\rho(f,g)=\sum_{\ell\geqslant 1}2^{-\ell}\rho_\ell(f,g)$, so that

$$\rho\left(e^{i\mathbf{m}\cdot\mathbf{x}},e^{i\mathbf{m}'\cdot\mathbf{x}}\right)=\sum_{\ell\geqslant 1}2^{-\ell}\rho_\ell\left(e^{i\mathbf{m}\cdot\mathbf{x}},e^{i\mathbf{m}'\cdot\mathbf{x}}\right)\approx\left|1-const\cdot\left\|\mathbf{m}-\mathbf{m}'\right\|^2\right|.$$

For example, taking $d=1$ and $\psi=\mathbf{1}_{[0,1)}$ we get

$$\rho_\ell\left(e^{imx},e^{im'x}\right)=\left|sinc\left((m-m')2^{-(\ell+1)}\right)\right|\approx\left|1-\frac{1}{6}\left((m-m')2^{-(\ell+1)}\right)^2\right|.$$

4.2 Recovering the geometry of circle eigenfunctions using ”partition-based correlation”: a computation

Given the above, it is tempting to ask whether one can use partition-based correlation, or similar function affinity notions that are based on a partition tree, to recover useful geometries on sets of manifolds and graphs eigenfunctions. However, it seems that there are very few cases where this question can be tackled analytically. The following simple experiment suggests that the iterative procedure of §2 enables one to study this question empirically.

In an experiment, summarized in Figure 4 below, we recover the dual geometry of the eigenfunctions of the circle $\mathbb{T}^1$. Let $x_1,\dots,x_N$ be equally spaced points in $[0,1]$ and consider the $2m$ by N matrix

$$M_{i,j} = \begin{cases} sin\left(2\pi k\right) & k = 1\dots m \\ cos\left(2\pi k\right) & k = m+1\dots 2m \end{cases}$$

The rows of M are orthogonal. However, eigenfunctions of similar frequencies behave similarly on folders of scale comparable to their frequency. The iterative procedure is designed to detect precisely this kind of similarity. By augmenting the values of each eigenfunctions by its values on folders of all scales and taking correlations along these augmented functions, we calculate a version of the correlation-based affinity and indeed recover the Euclidean affinity between the frequencies. To emphasize the organizational power of this procedure, we apply random permutations to the rows and columns of M before invoking the iterative procedure.

In a work in progress, we are applying the same procedure to investigate dual geometries of eigenfunctions in situations where an analytical approach is unknown, such the connected sum of two tori in $\mathbb{R}^3$.

5 Tensor Haar-like bases

We now proceed to answer the two questions that were stated at the end of §2. The main tool to evaluate the compatibility of a coupled geometry to a given data matrix, and later to process a data matrix, tensor Haar-like bases. A Haar-like bases is a multiscale, localized orthonormal bases induced by a partition tree. These bases were introduced in [12, 13]. An application to semi-supervised learning is included in [12]. A similar construction with applications to pattern detection in networks appears in [18].

As with any hierarchical partition construction on an abstract space, this section is inevitably heavy with notation. In §6.5 we discuss the familiar Euclidean case using more or less the same notation.

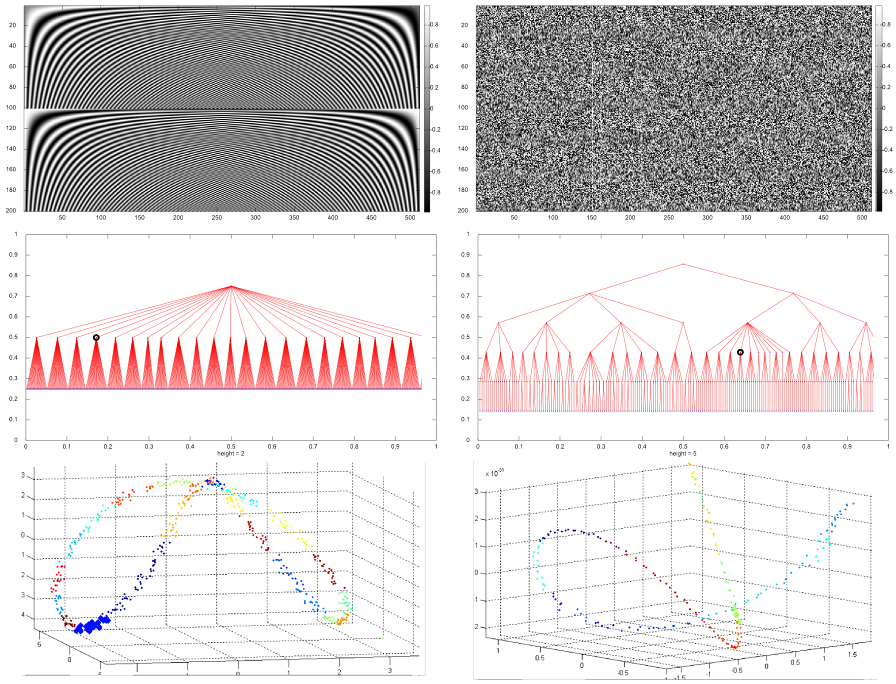


Fig. 4 Recovering the eigenfunction organization of $\mathbb{T}^1$: an experiment for a sines-cosines matrix on $N = 512$ points and frequencies $1 \dots m = 100$ (see §4.2). **Top left:** the original matrix $M_{i,j}$. **Top right:** the matrix after a random permutation of the rows and a random permutation of the columns. **Middle left:** the partition tree on the points $\{x_i\}$ generated by the iterative procedure. Points in the same folder in the middle level have close x values. For example, points in the folder marked by a circle contain all the points in $[0.42, 0.49]$. **Middle right:** the partition tree on the functions generated by the iterative procedure. Functions in the same folder have similar frequencies. For example, the functions in the folder marked by a circle are $sin(2\pi \cdot 81)$, $sin(2\pi \cdot 82)$, $cos(2\pi \cdot 80)$, $cos(2\pi \cdot 81)$. **Bottom left:** A Diffusion Embedding visualization in 3-d space of the of the affinity graph on the points $\{x_i\}$, generated by the iterative procedure. The same points folder as in the middle left panel is highlighted. The curve recovers the geometry of the points $\{x_i\}$ along the circle $\mathbb{R} \backslash \mathbb{Z}$ exactly as points along the curve have increasing x value modulo 1. (The color code represents a partition in one of the tree levels.) **Bottom right:** A Diffusion Embedding [6] visualization in 3-d space of the of the affinity graph on the functions, generated by the iterative procedure. The same points folder as in the middle right panel is highlighted. The curve recovers the geometry of the functions exactly as functions along the curve have increasing frequency. (The color code represents a partition in one of the tree levels.)

5.1 Haar-like bases

Definition 3. Let $\mathscr{T}$ be a partition tree (recall §2.1) with levels $\mathscr{X}^1, \dots \mathscr{X}^L$ on a set X . We say that $\mathscr{T}$ is $(\underline{B}, \overline{B})$-balanced, or $(\underline{B}, \overline{B})$-regular, if

$$\underline{B} \leqslant \frac{\left|X_j^{\ell+1}\right|}{\left|X_k^{\ell}\right|} \leqslant \overline{B} \tag{4}$$

for any $j \in subfolders\,(\ell,k)$ (recall Eq. 3).

There is a discrete analog of multi-resolution analysis associated with a partition tree. Let $V = \left\{f \,\middle|\, f : X \to \mathbb{R}\right\}$. Each partition $\mathscr{X}^\ell$ induces a subspace $V^\ell \subset V$ by $V^\ell = Span_{\mathbb{R}}\left\{\mathbf{1}_{X_1^\ell}, \ldots, \mathbf{1}_{X_{n(\ell)}^\ell}\right\}$. As $V^\ell \subset V^{\ell+1}$, we write W^ℓ $(1 \leqslant \ell < L\ )$ for the orthogonal complement of V^ℓ in $V^{\ell+1}$. Clearly $V^L = \left[\bigoplus_{\ell=1}^{L-1} W^\ell\right] \bigoplus V^1$.

Definition 4. 1. A Haar-like basis Ψ^ℓ for W^ℓ is an orthonormal basis of the form

$$\Psi^\ell = \bigcup_{k=1}^{n(\ell)} \left[\left\{\psi_{\ell,k,j}\right\}_{j=1}^{\#subfolders(\ell,k)-1}\right]$$

where for each $1 \leqslant k \leqslant n(\ell)$, the function $\psi_{\ell,k,j}$ is supported on the folder X_k^ℓ.

2. A Haar-like basis Ψ for V is a union of Haar-like bases for each W^ℓ, $1 \leqslant \ell \leqslant L-1$, together with the normalized constant function on X, $\psi_0 \equiv \frac{1}{\sqrt{N}}$. Namely,

$$\Psi = \{\psi_0\} \cup \bigcup_{\ell=1}^{L-1} \mathfrak{B}^\ell = \{\psi_0\} \cup \bigcup_{\ell=1}^{L-1} \bigcup_{k=1}^{n(\ell)} \left[\left\{\psi_{\ell,k,j}\right\}_{j=1}^{\#subfolders(\ell,k)-1}\right].$$

Fig. 5 illustrates a Haar-like basis induced by a partition tree on a small set.

Remarks:

1. Each basis function $\psi_{\ell,k,j}$ is associated with its support folder X_k^ℓ in the partition $\mathscr{X}^\ell$. The number of basis functions associated to the same folder X_k^ℓ is $\#subfolders(\ell,k) - 1$.
2. These functions resemble the classical Haar functions in the following sense:
   - Since $W^\ell \subset V^{\ell+1}$, each $\psi_{\ell,k,j}$ is piecewise constant on the folders of $\mathscr{X}^{\ell+1}$.
   - Since $\psi_{\ell,k,j}$ is supported on the folder X_k^ℓ, it is nonzero only on these folders of $\mathscr{X}^{\ell+1}$ that are subfolders of X_k^ℓ. In other words, $\psi_{\ell,k,j}$ is a linear combination of $\left\{\mathbf{1}_{X_i^{\ell+1}}\right\}_{i \in subfolders(\ell,k)}$.
   - Since $W^\ell \perp V^\ell$, we have $\left\langle \psi_{\ell,k,j}\,,\,\mathbf{1}_{X_k^\ell}\right\rangle = 0$, so that $\psi_{\ell,k,j}$ is orthogonal to the constant function on X_k^ℓ.

5.2 Tensor product of Haar-like bases

Now suppose that for each $\alpha = 1 \ldots d$, $X[\alpha]$ is a set and that $\mathscr{T}[\alpha]$ is a partition tree on it with levels $\mathscr{X}[\alpha]^\ell = \left\{X[\alpha]_k^\ell\right\}_{k=1}^{n[\alpha](\ell)}$ for $\ell = 1 \ldots L[\alpha]$. Consider the product

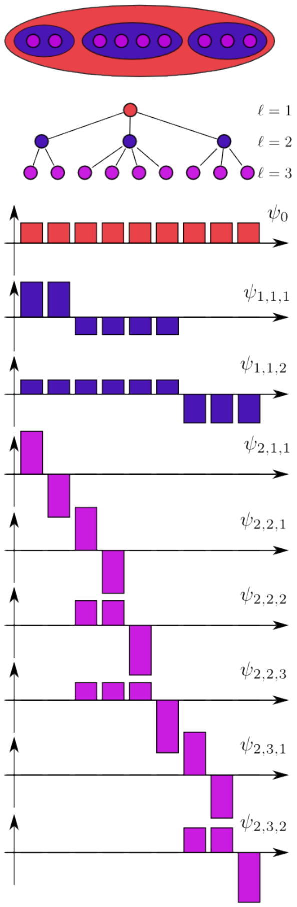


Fig. 5 An illustration of a Haar-like basis on a set of 9 points.

$X = X[1] \times \ldots \times X[d]$ with the normalized counting measure $|\cdot|$. If $\#X[\alpha]^{\ell}_{k}$ is the cardinality of $X[\alpha]^{\ell}_{k}$, then

$$\left| \prod_{\alpha=1}^{d} X[\alpha]^{\ell(\alpha)}_{k(\alpha)} \right| = \left(\prod_{\alpha=1}^{d} \#X[\alpha] \right)^{-1} \cdot \prod_{\alpha=1}^{d} \#X[\alpha]^{\ell(\alpha)}_{k(\alpha)} .$$

Definition 5. 1. Denote by

$$\mathscr{R} = \left\{ \prod_{\alpha=1}^{d} X[\alpha]^{\ell(\alpha)}_{k(\alpha)} \,\middle|\, 1 \leqslant \ell(\alpha) \leqslant L[\alpha] \text{ and } 1 \leqslant k(\alpha) \leqslant n[\alpha](\ell) \right\}$$

the set of all "rectangles", namely all products of folders of the partition trees $\mathscr{T}[1]\ldots\mathscr{T}[d]$.

2. Write $x=(x_1,\ldots,x_d)$ for an element of the product space $X=\prod_{\alpha=1}^{d}X[\alpha]$.
3. If $\Psi[\alpha]=\left\{\psi[\alpha]_{\ell,k,j}\right\}$ is a Haar-like basis induced by the partition tree $\mathscr{T}[\alpha]$ for $\alpha=1\ldots d$, then
$$\Psi=\left\{x\mapsto\prod_{\alpha=1}^{d}\psi[\alpha]_{\ell,k,j}\left(x_\alpha\right)\,\middle|\,\psi[\alpha]_{\ell,k,j}\in\Psi[\alpha]\right\}$$
is the corresponding tensor Haar-like basis. Clearly it is an orthonormal basis for $\{f:X\to\mathbb{R}\}$.
4. Recall that each basis function $\psi[\alpha]_{\ell,k,j}\in\Psi[\alpha]$ is associated with is support folder $X[\alpha]_k^\ell$. We can thus write $R\left(\psi\right)\in\mathscr{R}$ for the support rectangle of a tensor basis function $\psi\in\Psi$.
5. It will be convenient to enumerate the elements of Ψ by $\Psi=\left\{x\mapsto\psi_q(x)\right\}_{q=1}^{N}$, where $N=\prod_{\alpha=1}^{d}\#X[\alpha]$.
6. Assume that for each α the tree $\mathscr{T}[\alpha]$ is $\left(\underline{B}[\alpha],\overline{B}[\alpha]\right)$ - balanced (in the sense of Definition 3). Define $\tau\left(\Psi\right)=\left(\prod_{\alpha=1}^{d}\underline{B}[\alpha]\right)^{-1}$.

To stress the fact that for a data matrix, computing a tensor Haar-like basis function coefficient involves a non-trivial average of parts of the matrix, Fig. 6 shows an example from the Science News dataset discussed in §7. Top, a tensor Haar-like function in color map drawn using the original organization of the data matrix. Bottom, the same function after reordering the rows and columns according to a in-order traversal of the partition trees, which is the ordering used to construct the basis functions.

We conclude this section with an estimate of $\int_X\psi$ for use in the approximation theorem below. Recall that the absolute value of a classical Haar function is constant. This gives $\|\psi\|_\infty^2\cdot|R(\psi)|=1$ and hence
$$\int_X|\psi|^p=\int_X\left|R\left(\psi\right)\right|^{-\frac{p}{2}}=\left|R\left(\psi\right)\right|^{1-\frac{p}{2}}$$
for a classical Haar function ψ. This result extends to our setting using the "balance" constant $\tau\left(\Psi\right)$ above.

Lemma 1. *Let Ψ be a tensor Haar-like basis. Using the notation of Definition 5, for any $\psi\in\Psi$ and any $0<p<2$ we have*
$$\int_X|\psi|^p\leq\tau\left(\Psi\right)^{\frac{p}{2}}\left|R\left(\psi\right)\right|^{1-\frac{p}{2}}$$

Proof. We first show that
$$\max_{\psi\in\Psi}\{\|\psi\|_\infty^2\cdot|R(\psi)|\}\leq\tau(\Psi)\,. \tag{5}$$

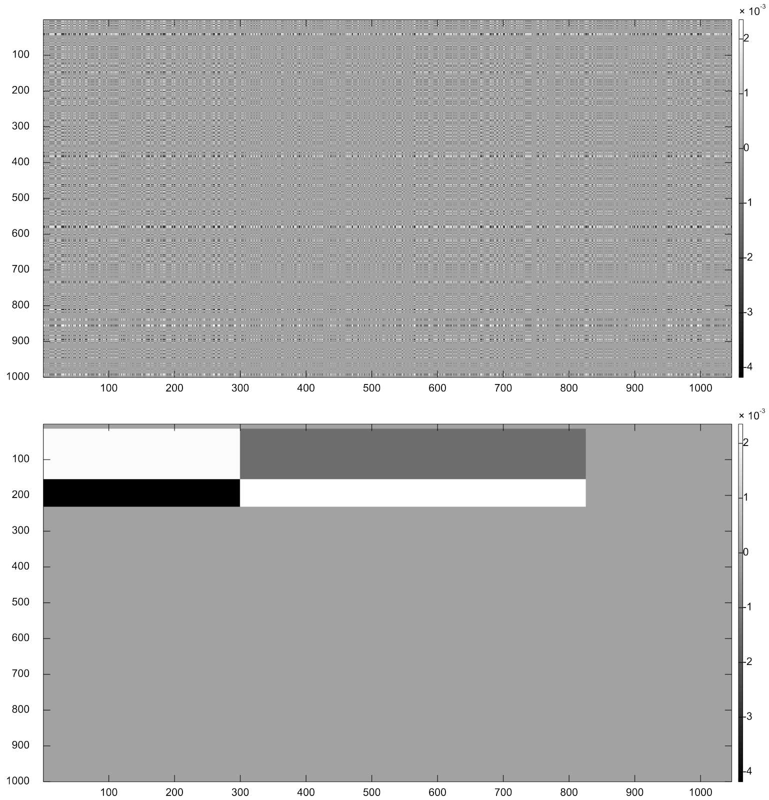


Fig. 6 An example of a tensor Haar-like function on the Science News data matrix. **Top:** In the original rows and columns order of the matrix. **Bottom:** After sorting rows and columns according to the trees which yielded this tensor Haar-like function. In the original order of the rows and columns, namely the order in which the data matrix is observed, an inner product with this basis function is a highly nontrivial averaging operation.

It is enough to show that for each $1 \leq \alpha \leq d$,

$$\max_{\psi[\alpha]\in\Psi[\alpha]} \{\|\psi[\alpha]\|_\infty^2 \cdot |R(\psi[\alpha])|\} \leq \underline{B}[\alpha].$$

Indeed, let $\psi[\alpha] = \psi[\alpha]_{\ell,k,j}$ be associated with the folder $X[\alpha]_k^\ell$. As in (3), we have $X[\alpha]_k^\ell = \biguplus_{j\in subfolders[\alpha](\ell,k)} X[\alpha]_j^{\ell+1}$. By definition, $\psi[\alpha]$ is constant on each set $X[\alpha]_j^{\ell+1}$, with value we denote by $\psi[\alpha]\left(X[\alpha]_j^{\ell+1}\right)$. Let

$$j^* = argmax_{j\in subfolders[\alpha](\ell,k)} \left|\psi[\alpha]_{\ell,k,j}\left(X[\alpha]_j^{\ell+1}\right)\right| ,$$

so that $\|\psi[\alpha]\|_\infty = \psi[\alpha]\left(X[\alpha]_{j^*}^{\ell+1}\right)$. Since

$$1 = \int\limits_X |\psi[\alpha]_{\ell,k,j}|^2 = \sum_{j \in subfolders[\alpha](\ell,k)} \left(\psi[\alpha]\left(X[\alpha]_j^{\ell+1}\right)\right)^2 \left|X[\alpha]_j^{\ell+1}\right|$$

we have by (4)

$$\|\psi[\alpha]\|_\infty^2 = \psi[\alpha]\left(X[\alpha]_{j^*}^{\ell+1}\right)^2 \leqslant \frac{1}{\left|X[\alpha]_{j^*}^{\ell+1}\right|} \leqslant \frac{1}{\underline{B}[\alpha] \cdot \left|X[\alpha]_k^\ell\right|} .$$

Multiplying over $\alpha = 1 \ldots d$ yields 5.
The lemma now follows as

$$\int\limits_X |\psi|^p \leqslant \left(\|\psi\|_\infty^2\right)^{\frac{p}{2}} \cdot |R(\psi)| \leqslant \left(\frac{\tau(\Psi)}{|R(\psi)|}\right)^{\frac{p}{2}} |R(\psi)| = \tau(\Psi)^{\frac{p}{2}} |R(\psi)|^{1-\frac{p}{2}}$$

□

6 Bounded ℓ_p-entropy implies function approximation and decomposition

> ...this is the first indication that a powerful theory for high dimensions exists. - J. O. Strömberg [20]

6.1 Problem statement: Approximating a function on a product space by a few terms in its tensor Haar expansion

Let X be a product space and Ψ a tensor Haar-like basis as in §5.2. We are interested in conditions on a function $f : X \longrightarrow \mathbb{R}$ that will allow approximating in using a small number terms in its tensor Haar expansion, $\sum_{i=1}^N \langle f, \psi_i \rangle \psi_i$.

In [20], J. O. Strömberg, addressing the continuous version of this question, observed it is very useful to sort the tensor Haar functions according to $|R|$, the volume of their support. He proved that if $f \in L_1\left([0,1]^d\right)$ has bounded mixed derivatives, then

$$\sup_{x\in[0,1]^d} \left| f(x) - \sum_{R \in \mathscr{R} \text{ s.t} |R| > \varepsilon} \langle f, \psi_R \rangle \psi_R(x) \right| < const \cdot \varepsilon \cdot \log^{d-1}\left(\frac{1}{\varepsilon}\right) ,$$

where ψ_R is the unique classical tensor Haar basis function supported on the dyadic rectangle R (see §6.5 below for the details of the continuous Euclidean version). However, the mixed derivative condition has certain disadvantages as a measure of function regularity: (1) It is not invariant to rotations and other simple coordinate changes, and (2) it does not generalize to general datasets. We are interested in weaker conditions, which would lead to data analysis algorithms.

We now show that one such condition is that the quantity $\sum_{i=1}^{N} |\langle f,\psi_i\rangle|^p$ (for some $0<p<2$), which we call the "ℓ_p entropy of the tensor Haar coefficients", is small. This condition is easy to check and appears to be much better adapted to our setting than the mixed derivative condition.

6.2 An approximation theorem

All the integrals on X are with respect to the normalized counting measure $|\cdot|$.

Theorem 1. *Fix $0<p<2$ and $f:X\to\mathbb{R}$.*

$$e_p(f)=\sum_{i=1}^{N} |\langle f,\psi_i\rangle|^p \, .$$

Let $\varepsilon>0$ and denote by Af an approximation of f obtained by retaining only coefficients of tensor Haar functions, which are (i) large, and (ii) correspond to basis functions supported on large folders:

$$A_\varepsilon f=\sum_{\substack{1\leqslant i\leqslant N \; s.t \; |\langle f,\psi_i\rangle|>\varepsilon^{\frac{1}{p}} \\ and \; |R(\psi_i)|>\varepsilon}} \langle f,\psi_i\rangle \, \psi_i(x) \, .$$

Then -

1. *The number of coefficients retained in $A_\varepsilon f$ does not exceed $\varepsilon^{-1}e_p(f)$. In particular, it depends on the dimension d only through $e_p(f)$.*
2. *Approximation in the mean when $0<p\leqslant 1$: if $e_p(f)\leqslant 1$ then*

$$\left(\int\limits_X |A_\varepsilon f-f|^p\right)^{\frac{1}{p}} \leqslant \left(\tau(\Psi)^{\frac{p}{2}}+1\right)^{\frac{1}{p}}\cdot\varepsilon^{\left(\frac{1}{p}-\frac{1}{2}\right)} \, . \qquad (6)$$

3. *Approximation in L_p when $1\leqslant p<2$:*

$$\left(\int\limits_X |A_\varepsilon f-f|^p\right)^{\frac{1}{p}} \leqslant \left(\tau(\Psi)^{\frac{1}{p}-\frac{1}{2}}+1\right)\cdot\varepsilon^{\left(\frac{1}{p}-\frac{1}{2}\right)}\cdot(e_p(f))^{\frac{1}{p}} \, , \qquad (7)$$

where $(e_p(f))^{\frac{1}{p}}$ is the ℓ_p norm of the coefficient vector $\{\langle f,\psi_i\rangle\}_{i=1}^{N}$.

4. *Uniform pointwise approximation on a set of large measure: For any $\lambda>0$ and $1\leqslant p<2$ we have*

$$|A_\varepsilon f(x)-f(x)|<\lambda\cdot\varepsilon^{\frac{1}{p}-\frac{1}{2}}$$

for any x outside an exceptional set E_λ with

$$|E_\lambda| \leqslant \frac{1}{\lambda^p}\left(\tau(\Psi)^{\frac{1}{p}-\frac{1}{2}}+1\right)^p \cdot e_p(f)\,.$$

Proof. For part (1), the number of coefficients, we have

$$\#\left\{\, 1 \leqslant i \leqslant N \,\middle|\, |\langle f,\psi_i\rangle| > \varepsilon^{\frac{1}{p}} \,\right\} \cdot \varepsilon \leqslant \sum_{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|>\varepsilon^{\frac{1}{p}}} |\langle f,\psi_i\rangle|^p \leqslant e_p(f)$$

and hence

$$\#\left\{\, 1 \leqslant i \leqslant N \,\middle|\, |\langle f,\psi_i\rangle| > \varepsilon \,\right\} \leqslant \varepsilon^{-1} \cdot e_p(f)\,.$$

The rest of our proof relies on the following key inequalities.
First key inequality: For $0 < p < 2$ we have

$$\int_X \Bigg| \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} \langle f,\psi_i\rangle\, \psi_i \Bigg|^p \leqslant \varepsilon^{1-\frac{p}{2}} \cdot (e_p(f))^{\frac{p}{2}} \tag{8}$$

Second key inequality: for $0 < p \leqslant 1$ we have

$$\int_X \Bigg| \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} \langle f,\psi_i\rangle\, \psi_i \Bigg|^p \leqslant \tau(\Psi)^{\frac{p}{2}} \cdot \varepsilon^{1-\frac{p}{2}} \cdot e_p(f) \tag{9}$$

Third key inequality (3): for $1 \leqslant p < 2$ we have

$$\int_X \Bigg| \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} \langle f,\psi_i\rangle\, \psi_i \Bigg|^p \leqslant (\tau(\Psi)\cdot\varepsilon)^{1-\frac{p}{2}} \cdot e_p(f) \tag{10}$$

Let us first deduce the theorem from these inequalities.

For part (2), approximation in the mean, assume $0 < p \leqslant 1$. Recall that $|x+y|^p \leqslant |x|^p + |y|^p$. We have

$$\begin{aligned}\int\limits_X |A_\varepsilon f - f|^p &= \int\limits_X \Bigg| f - \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|>\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} \langle f,\psi_i\rangle\,\psi_i \Bigg|^p = \\ &= \int\limits_X \Bigg| \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{or } |R(\psi_i)|\leqslant\varepsilon}} \langle f,\psi_i\rangle\,\psi_i \Bigg|^p = \\ &= \int\limits_X \Bigg| \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} \langle f,\psi_i\rangle\,\psi_i + \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} \langle f,\psi_i\rangle\,\psi_i \Bigg|^p \leqslant \\ &\leqslant \int\limits_X \Bigg| \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} \langle f,\psi_i\rangle\,\psi_i \Bigg|^p + \int\limits_X \Bigg| \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} \langle f,\psi_i\rangle\,\psi_i \Bigg|^p .\end{aligned}$$

As $e_p(f) \leqslant 1$, it now follows from (9) and (8) that

$$\begin{aligned}\int\limits_X |A_\varepsilon f - f|^p &\leqslant \tau(\Psi)^{\frac{p}{2}} \cdot \varepsilon^{1-\frac{p}{2}} \sum_{\substack{1\leqslant i\leqslant N \text{ s.t} \\ |R(\psi_i)|\leqslant\varepsilon}} |\langle f,\psi_i\rangle|^p + \varepsilon^{1-\frac{p}{2}} \left(\sum_{\substack{1\leqslant i\leqslant N \text{ s.t} \\ |R(\psi_i)|>\varepsilon}} |\langle f,\psi_i\rangle|^p \right)^{\frac{p}{2}} \leqslant \\ &\leqslant \left(\tau(\Psi)^{\frac{p}{2}} + 1\right) \cdot \varepsilon^{\left(1-\frac{p}{2}\right)}\end{aligned}$$

which is equivalent to (6).

Turning to part (3), approximation in L_p, assume $1 \leqslant p < 2$ and let $f : X \to \mathbb{R}$ such that $e_p(f) \leqslant 1$. Consider the function space $L_p = L_p(X)$ (w.r.t the normalized product counting measure $|\cdot|$) and let ℓ_p denote the sequence space $\mathbb{R}^N$ with the norm $\|(a_1 \dots a_N)\|_{\ell_p} = \left(\sum_{i=1}^N |a_i|^p\right)^{\frac{1}{p}}$. Denote by $\|f\|_{L_p} = \left(\int_X |f|^p\right)^{\frac{1}{p}}$ the norm of $L_p(X)$. Then by the inequalities (8) and (10),

$$\begin{aligned}\|A_\varepsilon f - f\|_{L_p} &\leqslant \left\| \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} \langle f,\psi_i\rangle\,\psi_i \right\|_{L_p} + \left\| \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} \langle f,\psi_i\rangle\,\psi_i \right\|_{L_p} \\ &\leqslant (\tau(\Psi)\cdot\varepsilon)^{\frac{1}{p}-\frac{1}{2}} + \varepsilon^{\frac{1}{p}-\frac{1}{2}} .\end{aligned}$$

It follows that the norm of linear operator $\ell_p \to L_p$ defined by $\Big(\langle f,\psi_1\rangle,\dots\langle f,\psi_N\rangle\Big) \mapsto (A_\varepsilon f - f)$ is bounded by $(\tau(\Psi)+1)^{\frac{1}{p}-\frac{1}{2}}\cdot\varepsilon^{\frac{1}{p}-\frac{1}{2}}$, as required. Finally, for part (4), uniform pointwise approximation on a set of large measure, let $\lambda > 0$ and $0 < p < 2$. Define

$$E_{\lambda,p} = \left\{ x\in X \,\middle|\, |A_\varepsilon f(x) - f(x)| \geqslant \lambda\varepsilon^{\frac{1}{p}-\frac{1}{2}} \right\}.$$

By Markov's inequality and part (3) above we have

$$\begin{aligned} \left|E_{\lambda,p}\right| &\leqslant \frac{\int_X |A_\varepsilon f - f|^p}{\lambda^p\cdot\varepsilon^{1-\frac{p}{2}}} \leqslant \\ &\leqslant \frac{1}{\lambda^p}\left(\tau(\Psi)^{\frac{1}{p}-\frac{1}{2}}+1\right)^p\cdot e_p(f). \end{aligned}$$

To complete the proof, we turn to the three key inequalities (8), (9) and (10).

To see the first key inequality (8), by Parseval's identity we have

$$\begin{aligned} \int_X \Bigg| \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} \langle f,\psi_i\rangle\,\psi_i \Bigg|^2 &= \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} |\langle f,\psi_i\rangle|^2 \leqslant \\ &= \left(\varepsilon^{\frac{1}{p}-\frac{1}{2}}\right)^2\cdot \sum_{\substack{1\leqslant i\leqslant N \text{ s.t } |\langle f,\psi_i\rangle|\leqslant\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} |\langle f,\psi_i\rangle|^p \leqslant \\ &\leqslant \left(\varepsilon^{\frac{1}{p}-\frac{1}{2}}\right)^2\cdot \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|>\varepsilon} |\langle f,\psi_i\rangle|^p . \end{aligned}$$

Now, Hölder's inequality implies that $(\int_X |g|^p)^{\frac{1}{p}} \leqslant \left(\int_X |g|^2\right)^{\frac{1}{2}}$, and (8) follows.

To see the second key inequality (9), assume that $0 < p \leqslant 1$. Since $|x+y|^p \leqslant |x|^p + |y|^p$, by Lemma 1 we have

$$\begin{aligned} \int_X \Bigg| \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} \langle f,\psi_i\rangle\,\psi_i \Bigg|^p &\leqslant \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} |\langle f,\psi_i\rangle|^p \int_X |\psi_i|^p \leqslant \\ &\leqslant \tau(\Psi)^{\frac{p}{2}}\cdot \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} |\langle f,\psi_i\rangle|^p\,|R(\psi_i)|^{1-\frac{p}{2}} \leqslant \\ &\leqslant \tau(\Psi)^{\frac{p}{2}}\cdot\varepsilon^{1-\frac{p}{2}}\cdot \sum_{1\leqslant i\leqslant N \text{ s.t } |R(\psi_i)|\leqslant\varepsilon} |\langle f,\psi_i\rangle|^p \end{aligned} \tag{11}$$

which is (9). Finally, to see the third key inequality (10), let $1 \leqslant p < 2$. Consider the function space $L_p = L_p(X)$ and the sequence space ℓ_p as in the proof of part (3) above. For a sequence $\mathbf{a} = (a_1 \dots a_N)$, define

$$T(\mathbf{a}) = \sum_{1 \leqslant i \leqslant N \text{ s.t } |R(\psi_i)| \leqslant \varepsilon} a_i \psi_i \,.$$

For $0 < p \leqslant 2$, let $\|T\|_{p,p}$ denote the operator norm of the linear operator $T : \ell_p \to L_p$. Clearly $\|T\|_{2,2} \leqslant 1$. By the inequality (9), we have $\|T\|_{1,1} \leqslant (\tau(\Psi) \cdot \varepsilon)^{\frac{1}{2}}$. We now appeal to the Riesz-Thorin Interpolation Theorem (see e.g. [19] pp. 179, theorem 1.3), whereby for any $1 \leqslant p < 2$,

$$\|T\|_{p,p} \leqslant (\tau(\Psi) \cdot \varepsilon)^{\frac{1}{2}(1-t)} \,,$$

whenever $0 \leqslant t \leqslant 1$ satisfies $\frac{1}{p} = 1 - t + \frac{t}{2}$, or equivalently $1 - t = \frac{2}{p} - 1$. It follows that $\|T\|_{p,p} \leqslant (\tau(\Psi) \cdot \varepsilon)^{\frac{1}{p} - \frac{1}{2}}$. Let $1 \leqslant p < 2$. We have

$$\begin{aligned} \left(\int_X \Big| \sum_{1 \leqslant i \leqslant N \text{ s.t } |R(\psi_i)| \leqslant \varepsilon} \langle f, \psi_i \rangle \, \psi_i \Big|^p \right)^{\frac{1}{p}} &= \Big\| T\Big(\langle f, \psi_1 \rangle, \ldots \langle f, \psi_N \rangle \Big) \Big\|_{L_p} \leqslant \\ &\leqslant \|T\|_{p,p} \big\| \big(\langle f, \psi_1 \rangle, \ldots \langle f, \psi_N \rangle \big) \big\|_{\ell_p} = \\ &= (\tau(\Psi) \cdot \varepsilon)^{\frac{1}{p} - \frac{1}{2}} \cdot (e_p(f))^{\frac{1}{p}} \,. \end{aligned}$$

which is equivalent to (10).

□

In order to find the coefficients retained in the approximation above computationally, namely to locate $i \in \{1 \ldots N\}$ such that $|\langle f, \psi_i \rangle| > \varepsilon^{\frac{1}{p}}$ and $R(\psi_i) > \varepsilon$, we must check the size of coefficients $\langle f, \psi_i \rangle$ such that $R(\psi_i) > \varepsilon$. In order to bound the number of coefficients whose size must be examined, we now show that

$$\#\left\{\, 1 \leqslant i \leqslant N \,\middle|\, R(\psi_i) \,| > \varepsilon \,\right\} \leqslant (\beta - 1)^d \beta^2 \cdot \left(\frac{1}{\varepsilon}\right) \cdot \left(\log_\beta \left(\frac{1}{\varepsilon}\right) + 1\right)^{d-1}$$

where $\mathscr{T}[i]$ is $\left(\underline{B}[i], \overline{B}[i]\right)$ -balanced as in (4), and $\beta = \max_{1 \leqslant i \leqslant d} \left\{ \frac{1}{\underline{B}[i]} \right\}$. Indeed, by (4), for any $\left(\underline{B}, \overline{B}\right)$ partition tree, $n(\ell)$ (the number of folders in level ℓ) satisfies $n(\ell) \leqslant \left(\frac{1}{\underline{B}}\right)^{\ell - 1}$. Denote by $\left(\underline{B}[i], \overline{B}[i]\right)$ the parameters for the partition tree $\mathscr{T}[i]$ and let $\beta = \max_{1 \leqslant i \leqslant d} \left\{ \frac{1}{\underline{B}[i]} \right\}$. If $\beta^{-L} \leqslant \varepsilon < \beta^{-L+1}$, then

$$\#\left\{\, R \in \mathscr{R} \, \text{s.t} \, |R| \geqslant \beta^{-L} \,\right\} = \sum_{\substack{(r_1, \ldots, r_d) \in \mathbb{N}^d \\ \beta^{-\Sigma r_i} \geqslant \beta^{-L}}} \beta^{\Sigma r_i} = \sum_{\substack{(r_1, \ldots, r_d) \in \mathbb{N}^d \\ r_1 + \ldots + r_d \leqslant L}} \beta^{\Sigma r_i} \,.$$

Let us show by induction that $\sum_{\substack{(r_1,\ldots,r_d)\in\mathbb{N}^d \\ r_1+\ldots+r_d\leqslant L}} \beta^{\Sigma r_i} \leqslant \beta^{L+1}L^{d-1}$. Indeed, for $d=1$ we have $\sum_{r=0}^{L}\beta^r = \beta^{L+1}-1$. Assuming this for d, we have

$$\sum_{\substack{(r_1,\ldots,r_{d+1})\in\mathbb{N}^{d+1} \\ r_1+\ldots+r_{d+1}\leqslant L}} \beta^{(r_1+\ldots+r_{d+1})} = \sum_{r_{d+1}=0}^{L} \beta^{r_{d+1}} \sum_{\substack{(r_1,\ldots,r_d)\in\mathbb{N}^d \\ r_1+\ldots+r_d\leqslant L-r_{d+1}}} \beta^{(r_1+\ldots+r_d)} \leqslant$$

$$\leqslant \sum_{r_{d+1}=0}^{L} \beta^{r_{d+1}} \left(\beta^{L-r_{d+1}+1}\cdot(L-r_{d+1})^{d-1}\right) =$$

$$= \beta^{L+1}\sum_{k=0}^{L-1} k^{d-1} \leqslant \beta^{L+1}L\cdot L^{d-1}\,.$$

Therefore, $\#\left\{\, R\in\mathscr{R}\,\text{s.t}\,|R|>\varepsilon \,\right\} \leqslant \beta\cdot\beta^L\cdot L^{d-1} \leqslant \beta^2\cdot\left(\frac{1}{\varepsilon}\right)\cdot\left(\log_\beta\left(\frac{1}{\varepsilon}\right)+1\right)^{d-1}$. Since up to $(\beta-1)^d$ can be associated with any rectangle $R\in\mathscr{R}$, we finally obtain that

$$\#\left\{\, 1\leqslant i\leqslant N \,\middle|\, R(\psi_i)\,| > \varepsilon \,\right\} \leqslant (\beta-1)^d\beta^2\cdot\left(\frac{1}{\varepsilon}\right)\cdot\left(\log_\beta\left(\frac{1}{\varepsilon}\right)+1\right)^{d-1}.$$

A tighter bound can be obtained by considering each $\underline{B}[i]$ separately instead of using the uniform bound β.

6.3 ℓ_1-entropy interpreted as smoothness

Consider a data matrix and suppose we have constructed a two coupled graphs, one on the rows and one of the columns. Did we do a good job? Theorem 1 implies that the ℓ_p entropy of the matrix in the tensor Haar-like basis, induced by two partition trees describing the two geometries, provides a way to quantify the usefulness of the coupled geometry at hand. In general, we would like to be able to say that a function is adapted to a given geometry if the function is smooth and simple in this geometry. Instead of smoothness in the Euclidean sense of differentiability, in our general setting it becomes natural to quantify *smoothness* using a Haar-like basis in terms of pointwise exponential decay of coefficients (see [12] for a theorem of this kind), while *simplicity* means a small number of non-negligible tensor Haar-like coefficients, which is, in some cases, intimately related to small ℓ_1-entropy. The above inequalities for Haar-like bases relate function smoothness in the geometry, as it is captured by the basis, to sparsity of the Haar expansion. The unique feature of the tensor Haar-like basis is that *only coefficients of Haar functions ψ_i with support volume $|R(\psi_i)|>\varepsilon$ are required*, eliminating the need to consider coefficients that correspond to basis functions with small support.

In particular, we find an answer to Question 1 from §2.2.

6.4 A Decomposition theorem

The relationship between coefficient decay and smoothness is well-known in harmonic analysis. In [12, 13] it is shown that, for Haar-like bases, coefficient decay is equivalent to Hölder property, just as in the classical case. We now extend this fact to tensor Haar-like bases and deduce a decomposition theorem of Calderón-Zygmund type.

For a set $X[i]$ with a partition tree $\mathscr{T}[i]$, we define a metric on $X[i]$ by

$$\rho(x,y) = \begin{cases} |folder(x,y)| & x \neq y \\ 0 & x = y \end{cases} \tag{12}$$

where $folder(x,y)$ is the smallest folder in any level of $\mathscr{T}[i]$ containing both x and y.

Claim. Let $f : X[i] \to \mathbb{R}$ and $p > 0$. Let Ψ be a Haar-like basis corresponding to the partition tree $\mathscr{T}[i]$. Consider the conditions:

1. There is a constant C such that for any $\psi \in \Psi$ associated with the folder X_k^ℓ,
$$|\langle f\,,\,\psi\rangle| \leqslant C \cdot \left|X_k^\ell\right|^{p+1/2}.$$
2. f is p-Hölder with constant C in the metric ρ. That is, $|f(x) - f(y)| \leqslant C \cdot \rho(x,y)^p$ for all $x, y \in X$.

Then (1) with constant $C \Longrightarrow$(2) with constant γC, and (2) with constant $C \Longrightarrow$(1) with constant δC. The numbers γ and δ depend on p and the partition tree $\mathscr{T}$ alone.

See [13] for the proof.

In our setting of a product space, a similar phenomenon occurs. We discuss the case of matrices, where $d = 2$.

Theorem 2. $X = X[1] \times X[2]$. *Suppose that* $\mathscr{T}[i]$ *is a partition tree on* $X[i]$ *and* Ψ_i *is an induced Haar-like basis,* $i = 1, 2$. *Let* Ψ *be the tensor Haar-like basis. Let* $f : X \to \mathbb{R}$ and $p > 0$. Consider the following conditions:

1. *There is a constant* C *such that for any* $\psi \in \Psi$ *associated with the rectangle* $R(\psi)$,
$$|\langle f\,,\,\psi\rangle| \leqslant C \cdot |R(\psi)|^{p+1/2}.$$
2. f *is* p*– bi-Hölder with constant* C *in the metric* ρ*, in the sense that*

$$\left|f(x,y) - f(x',y) - f(x,y') + f(x',y')\right| \leqslant C \cdot \rho[1](x,x')^p \rho[2](y,y')^p$$

for some constant C, all $x,x' \in X[1]$ and all $y,y' \in X[2]$. Here, $\rho[i]$ is the metric induced by $\mathscr{T}[i]$ on $X[i]$, as in (12), for $i=1,2$.

Then (1) with constant $C \Longrightarrow$(2) with constant γC, and (2) with constant $C \Longrightarrow$(1) with constant δC, where the numbers γ, δ depends on p and on the partition trees $\mathscr{T}[1]$ and *$\mathscr{T}[2]$ alone.*

Proof. $\Longrightarrow$ Assume (1) holds with constant C. Choose $x,x' \in X[1]$ and $y,y' \in X[2]$. Write the tensor Haar-like function explicitly as products: $\Psi = \left\{\psi_i(x)\varphi_j(y)\right\}_{i,j}$ for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. Then

$$f(x,y) = \sum_{i,j} \left\langle f, \psi_i\varphi_j\right\rangle \psi_i(x)\varphi_j(y) = \sum_i a_i(y)\psi_i(x)\,,$$

where $a_i(y) = \sum_j \left\langle f, \psi_i\varphi_j\right\rangle \varphi_j(y)$. Since $\left|\left\langle f, \psi_i\varphi_j\right\rangle\right| \leqslant C\cdot|R(\psi_i)|^{p+1/2}\left|R(\varphi_j)\right|^{p+1/2}$, where $R(\varphi_j)$ is the folder in $\mathscr{T}[2]$ associated with φ_j, by Theorem 2 above for $X[2]$, we have

$$\left|a_i(y)-a_i(y')\right| \leqslant \gamma_2\cdot C\cdot|R(\psi_i)|^{p+1/2}\cdot\rho[2]\left(y,y'\right)^p\,,$$

where γ_2 depends on p and $\mathscr{T}[2]$ alone. Now, define $g(x) = f(x,y)-f(x,y') = \sum_i\left[a_i(y)-a_i(y')\right]\psi_i(x)$. Again by Theorem 2, this time for $X[1]$,

$$\left|g(x)-g(x')\right| \leqslant \gamma_1\cdot\gamma_2\cdot C\cdot\rho[1]\left(x,x'\right)^p\cdot\rho[2]\left(y,y'\right)^p\,.$$

$\Longleftarrow$Assume (2) holds with constant C. As $\int_{X_1}\psi_i(x)=0$ and similarly for X_2, we have for any $\psi=\psi_i(x)\varphi_j(y)$ and any $(x',y')\in R(\psi)$ that

$$\begin{aligned}\left\langle f, \psi_i\varphi_j\right\rangle &= \int_X f(x,y)\psi_i(x)\varphi_j(y)dxdy\\ &= \int_X \left[f(x,y)-f(x',y)-f(x,y')+f(x',y')\right]\psi_i(x)\varphi_j(y)dxdy\,,\end{aligned}$$

hence

$$\left|\left\langle f, \psi_i\varphi_j\right\rangle\right| \leqslant C\cdot\rho[1](x,x')^p\rho[2](y,y')^p\cdot\int_X\left|\psi_i(x)\varphi_j(y)\right|dxdy\,.$$

But by the definition of $\rho[1]$, we have $\rho[1](x,x')\leqslant R(\psi_i)$ and similarly, $\rho[2](y,y')\leqslant R(\varphi_j)$. Finally, by Lemma (1), we have $\int_X\left|\psi_i(x)\varphi_j(y)\right|dxdy\leqslant\sqrt{\tau(\Psi)}\,|R(\psi_i)|^{\frac{1}{2}}\left|R(\varphi_j)\right|^{\frac{1}{2}}$. In summary,

$$\left|\left\langle f, \psi_i\varphi_j\right\rangle\right| \leqslant \sqrt{\tau(\Psi)}\cdot C\cdot|R(\psi)|^{p+\frac{1}{2}}\,.$$

□

We can decompose a given function on X to a regular part and a part with small support, à la Calderón-Zygmund. In the context of a data matrix or tensor, this would mean that if we can describe f efficiently using a tensor Haar-like basis, then it decomposes into a “typical” matrix, which is regular with respect to the coupled geometry we constructed, and an “outlier” matrix or potentially irregular behavior but with small support.

Theorem 3. *Let $f : X \to \mathbb{R}$ and $0 < p < 2$. There is a decreasing sequence of sets $E_\ell \subset X$ where $|E_\ell| < \frac{e_p(f)}{2^\ell}$ for any ℓ, and a decomposition $f = g_\ell + b_\ell$, such that $x \mapsto g_\ell(x)$ (the “good” function) is $\frac{1}{p} - \frac{1}{2}$ - Hölder with constant $2^{\frac{\ell}{p}}$ and $x \mapsto b_\ell(x)$ (the “bad” function) is supported on E_ℓ. The functions g_ℓ and b_ℓ, as well as the set E_ℓ, all have explicit form.*

Proof. Define

$$S_p(x) = \sum_{i=1}^{N} |\langle f, \psi_i \rangle|^p \, \frac{\chi_{R(\psi_i)}(x)}{|R(\psi_i)|}$$

where $\chi_{R(\psi_i)}(x)$ is the indicator function of $R(\psi_i)$. (This is an analog of the Littlewood-Paley function). A Fubini argument gives

$$\int_X S_p(x)\, dx = \sum_{i=1}^{N} |\langle f, \psi_i \rangle|^p = e_p(f)\,.$$

As in the proof of Theorem 1, define an “exceptional set”

$$E_{\ell,p} = \left\{ x \in X \,\middle|\, S_p(x) > 2^\ell \right\}.$$

Clearly $E_{\ell+1,p} \subset E_{\ell,p}$. Markov’s inequality gives

$$\left| E_{\ell,p} \right| \leqslant \frac{e_p(f)}{2^\ell}\,.$$

Now decompose the set of all rectangles $\mathscr{R}$ as follows. Define

$$\mathscr{R}_\ell = \left\{ R \in \mathscr{R} \,\middle|\, R \subseteq E_{\ell,p} \text{ and } R \nsubseteq E_{\ell+1,p} \right\}$$

and observe that $\mathscr{R} = \biguplus_{\ell \in \mathbb{Z}} \mathscr{R}_\ell$. We can thus decompose $f = \sum_{\ell \in \mathbb{Z}} f_\ell$, where

$$f_\ell = \sum_{1 \leqslant i \leqslant N \text{ with } R(\psi_i) \in \mathscr{R}_\ell} \langle f, \psi_i \rangle \, \psi_i\,.$$

Fix $\ell \in \mathbb{Z}$ and $R \in \mathscr{R}_\ell$. By the definition of $\mathscr{R}_\ell$, there exists $x \in R$ with $x \in E_{\ell,p} \setminus E_{\ell+1,p}$. For this x we have

$$\sum_{i=1}^{N} |\langle f, \psi_i \rangle|^p \, \frac{\chi_{R(\psi_i)}(x)}{|R(\psi_i)|} = S_p(x) \leqslant 2^{\ell+1}\,.$$

Choose $1 \leqslant i \leqslant N$ with $R(\psi_i) = R$. As $\chi_{R(\psi_i)}(x) = 1$, we get

$$\frac{|\langle f,\psi_i\rangle|^p}{|R(\psi_i)|} \leqslant S_p(x) \leqslant 2^{\ell+1}\,,$$

namely $|\langle f,\psi_i\rangle| \leqslant 2^{\frac{\ell+1}{p}}\,|R(\psi_i)|^{\frac{1}{p}}$. Define $g_\ell = \sum_{k=1}^{\ell-1} f_k$ and $b_\ell = \sum_{k\geqslant\ell} f_k$. By Theorem 2, g_ℓ is as required. Clearly b_ℓ is supported on $E_{\ell,p}$. □

We note that the above is not an existence result, but rather an *explicit* formula for decomposing a given function.

6.5 The Euclidean analog

We briefly translate the above general, discrete results to the Euclidean case of $f \in L_1\left([0,1]^d\right)$ and the tensor product of classical Haar bases. This yields an approximation theorem in high dimensional Euclidean spaces.

Consider the unit interval $[0,1] \subset \mathbb{R}$ and the classical Haar basis in $[0,1]$. Each Haar functions corresponds to its support, a dyadic interval. We index the basis functions by $\{\psi_I\}_{I\in\mathscr{I}}$ where $\mathscr{I} = \left\{[n2^{-k},(n+1)2^{-k}]\right\}_{n,k}$. Consider the unit cube $[0,1]^d$ in $\mathbb{R}^d$. The *tensor Haar basis* on $[0,1]^d$, namely $\left\{\psi_{I_1}\otimes\ldots\otimes\psi_{I_d}\right\}_{I_i\in\mathscr{I}}$, is an orthonormal basis for $L_2\left([0,1]^d\right)$. Let $\mathscr{R}$ denote the set of dyadic rectangles on $[0,1]^d$, $\mathscr{R} = \left\{I_1\times\ldots\times I_d \,\middle|\, I_i \in \mathscr{I}\right\}$. As each function in the tensor Haar basis corresponds to its support, a dyadic rectangle, we index the tensor Haar basis by$\{\psi_R\}_{R\in\mathscr{R}}$, where for a dyadic rectangle $R = I_1\times\ldots\times I_d$, we have $\psi_R = \psi_{I_1}\otimes\ldots\otimes\psi_{I_d}$. Write $|R|$ for the d-dimensional Lebesgue measure of R.

The proof of Theorem 1 can be used verbatim to prove its Euclidean version:

Theorem 4. *For $f \in L_1\left([0,1]^d\right)$ and $0 < p < 2$, write*

$$e_p(f) = \sum_{R\in\mathscr{R}} |\langle f,\psi_R\rangle|^p\,.$$

Let $\varepsilon > 0$ and denote by Af an approximation of f obtained by retaining only large coefficients of tensor Haar functions supported on large folders:

$$A_\varepsilon f = \sum_{\substack{R\in\mathscr{R}\ s.t\ |\langle f,\psi_R\rangle|>\varepsilon^{\frac{1}{p}} \\ and\ |R|>\varepsilon}} \langle f,\psi_R\rangle\,\psi_R(x)\,.$$

Then -

1. *The number of coefficients retained in $A_\varepsilon f$ does not exceed $\varepsilon^{-1}e_p(f)$.*
2. *Approximation in the mean when $0 < p \leqslant 1$: if $e_p(f) \leqslant 1$ then*

$$\left(\int_{[0,1]^d} |A_\varepsilon f - f|^p\right)^{\frac{1}{p}} \leqslant 2^{\frac{1}{p}} \cdot \varepsilon^{\left(\frac{1}{p}-\frac{1}{2}\right)} . \tag{13}$$

3. Approximation in L_p when $1 \leqslant p < 2$:

$$\left(\int_{[0,1]^d} |A_\varepsilon f - f|^p\right)^{\frac{1}{p}} \leqslant 2 \cdot \varepsilon^{\left(\frac{1}{p}-\frac{1}{2}\right)} \cdot (e_p(f))^{\frac{1}{p}} . \tag{14}$$

4. Uniform pointwise approximation on a set of large measure: For any $\lambda > 0$ and $1 \leqslant p < 2$ we have

$$|A_\varepsilon f(x) - f(x)| < \lambda \cdot \varepsilon^{\frac{1}{p}-\frac{1}{2}}$$

for any x outside an exceptional set E_λ with

$$|E_\lambda| \leqslant \frac{2^p}{\lambda^p} e_p(f) .$$

5. We only need to evaluate the size of

$$\#\left\{R \in \mathscr{R} \,\middle|\, |R| > \varepsilon\right\} \leqslant 4 \cdot \left(\frac{1}{\varepsilon}\right) \cdot \left(\log\left(\frac{1}{\varepsilon}\right) + 1\right)^{d-1}$$

coefficients.

We remark that the class of functions, for which the Haar expansion has finite entropy has been characterized by [8] [1] .

The relation between smoothness and tensor Haar coefficient size of Theorem 2 is well known in the Euclidean case. Theorem 3 extends to the Euclidean case without change in its statement and proof.

[1] It is shown there that this class is independent of the wavelet used, and is equivalent to having a Harmonic extension whose derivative is integrable in the disk (or bi-disk). They also characterize the dual spaces as Bloch spaces, which in our case are just functions with bounded Haar coefficients. Observe further that

$$f = \sum_R |R|^{\frac{1}{2}} \langle f, \psi_R \rangle |R|^{-\frac{1}{2}} \psi_R(x)$$

is a special atomic decomposition of $\left|\frac{d}{dx}\right|^{\frac{1}{2}} \left|\frac{d}{dy}\right|^{\frac{1}{2}} f$, which is therefore in the Hardy space H^1 of the bi-disk. A similar result holds for other entropies, implying a fractional derivative in the Hardy space.

7 Database analysis

The iterative procedure for recovering the coupled geometry of a data matrix, together with the approximation theorem, suggest a family of data analysis schemes. The general recipe is as follows.

Algorithm 2. Given a data matrix M,

1. Compute an initial affinity W_0 on the columns of M.
2. Apply an iteration of Algorithm 1 and obtain partition trees $\mathscr{T}_{rows}$ and $\mathscr{T}_{cols}$ on the rows and columns of M, respectively.
3. Construct induced Haar-like bases Ψ_{rows} and Ψ_{cols} and the tensor Haar-like basis $\Psi = \{\psi_i\}_{i=1}^{N}$.
4. Compute the ℓ_1 entropy $e_1(M,\Psi) = \sum_{i=1}^{N} |\langle M, \psi_i\rangle|$
5. Repeat steps 2-4 until $e_1(M,\Psi)$ converges.
6. Transform[2] M into the tensor Haar-like basis Ψ to obtain the coefficient matrix $\widetilde{M}$.
7. Process M in the coefficient domain $\widetilde{M}$ (see below) and transform back.

This data driven geometric "self-organization" allows to analyze any data matrix according to its intrinsic row or column structure. While this presentation focuses on data matrices, there is nothing special about order-2 tensors: this approach generalizes to data tensors of order d.

The specialization of this general procedure depends on the data analysis task at hand. Each application calls for a detailed treatment. This is beyond the scope of this introduction.

1. **Compression**: to compress M, store both partition trees $\mathscr{T}_{rows}$ and $\mathscr{T}_{cols}$, and only coefficients of tensor Haar-like functions with $R(\psi_i) > const \cdot \varepsilon^2$, where ε is the desired error.
2. **Denoising**: sort coefficients $\{\langle M, \psi_i\rangle\}_{i=1}^{N}$ according to $R(\psi_i)$ and apply thresholding.
3. **Matrix completion and missing value imputation**: If we are given the trees $\mathscr{T}_{rows}$ and $\mathscr{T}_{cols}$ from an external source or prior knowledge, or if the number of missing values in M allows reliable construction of $\mathscr{T}_{rows}$ and $\mathscr{T}_{cols}$, we can estimate the coefficients $\langle M, \psi_i\rangle$ using available entries. Using a tensor Haar-like basis has the distinct advantage that we need only care about averages of the available points on large sets, leading to small estimator variance. See [12] for a detailed analysis in the case $d = 1$.
4. **Anomaly detection**: the residual matrix

$$\sum_{1\leqslant i\leqslant N \text{ with } |R(\psi_i)|\leqslant\varepsilon} \langle M, \psi_i\rangle\, \psi_i$$

[2] Note that a fast algorithm for computing the coefficients $\langle M, \psi_i\rangle$, or equivalently, transforming M into $\widetilde{M}$ exists. We do not pursue this here.

contains the deviations from the average matrix. This yields a method for detecting anomalous observations, as they are different from the "average matrix" in their respective folders.

7.1 Eliminating Haar artifacts

While Haar-like bases are simple to construct and to use in our general setting, their lack of smoothness introduce artifacts into any computation that is taking place in the coefficient domain. This phenomenon was studied in [5] in signal processing setting of classical Haar basis on $[0,L]$. Indeed, the location of the Haar discontinuities, namely the alignment of the dyadic grid on $[0,L]$, is arbitrary. When denoising a function $f:[0,L]\to\mathbb{R}$ using coefficient shrinkage in a given Haar basis, artifacts appear along this arbitrarily shifted dyadic grid. Therefore, they suggest to denoise f using several differently shifted Haar bases, and average the resulting denoised versions of f in order to eliminate the artifacts.

This observation is ideal for eliminating artifacts in our setting. In the general data-analysis recipe 2, after completing step 5, namely after stabilizing the coupled geometry of the data matrix M, we have the affinities W_{rows} and W_{cols} on the rows and columns of M, respectively. Recall that the procedure for integrating an affinity into a partition tree, described in 2.1, is random. In order to reduce the artifacts caused by processing M in a tensor Haar-like basis, we choose r and construct partition trees $\left\{\mathscr{T}^{(i)}_{rows}\right\}_{i=1}^{r}$ using W_{rows} and $\left\{\mathscr{T}^{(i)}_{cols}\right\}_{i=1}^{r}$ using W_{cols}. Computing induced Haar-like bases $\left\{\Psi^{(i)}_{rows}\right\}_{i=1}^{r}$ and $\left\{\Psi^{(j)}_{cols}\right\}_{j=1}^{r}$, this yields r^2 different tensor Haar-like bases $\left\{\Psi^{(i,j)}\right\}_{i,j=1}^{r}$ by setting $\Psi^{(i,j)}=\Psi^{(i)}_{rows}\bigotimes\Psi^{(j)}_{cols}$. Each basis is used to produced a processed version of M (as in the last step of the general recipe 2), and the r^2 versions are averaged to produce the final result.

We remark that one can modify this basic construction of a hierarchical scale decomposition, in order to build wavelets that provide filters restricting the frequency content of a function to bands of eigenfunctions of the diffusion or Laplace operator on the graph. See for example the constructions in [9, 10].

7.2 Example: the Science News database

As a concrete example of data analysis using the general recipe 2, we consider a term-document data matrix. The documents are abstracts of 1047 articles obtained from the Science News journal website, each under one of the following scientific fields: Anthropology, Astronomy, Social Sciences, Earth Sciences, Biology, Mathematics, Medicine, or Physics. This data set was prepared and preprocessed for [17], where additional information can be found, and kindly provided by J. Solka. In the preprocessing step, a dictionary of 10,906 words were chosen as relevant for this body of documents. Of these, the 1,000 words with the highest correlation to article subject classification where selected. The data matrix is thus $(M_{i,j})_{i=1\ldots1{,}000\,;\,j=1\ldots1024}$ where the entry $M_{i,j}$ is the relative frequency of the i-th word in the j-th document.

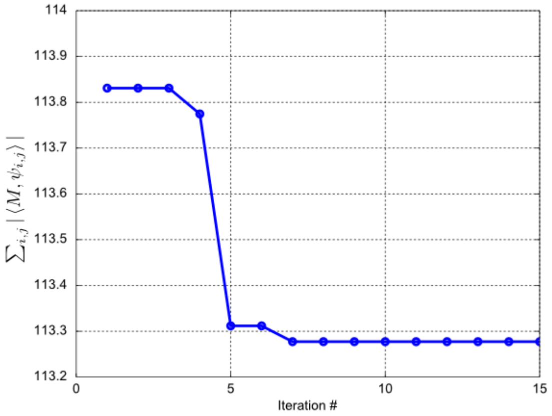


Fig. 7 Iterative procedure: ℓ_1 entropy over iteration number

The original data matrix makes little sense to the naked eye and is not shown. When running steps 1-5 in Algorithm 2 we observe decrease and convergence of the ℓ_1 entropy $e_1(M,\Psi) = \sum_{i=1}^{N} |\langle M, \psi_i\rangle|$ (Fig. 7). In Fig. 8 (left) the data matrix is shown with rows and columns re-organized using depth-first order of the obtained trees. Fig. 8 compares the original matrix with an approximation obtained by retaining those 15% of the tensor Haar-like coefficients corresponding to $\varepsilon = 9.5 \cdot 10^{-5}$. We remark that for this data matrix, as well as for other data matrices with ”intrinsic organization of the row and column sets”, this approximation is superior to counterparts from classical multivariate analysis, such as retaining the 15% largest singular values in the SVD decomposition of the data matrix). Fig. 9 shows the approximation quality $\int_X |A_\varepsilon - A|$ as function of ε, together with the fraction of coefficients retained for the approximation and the theoretical bound of Theorem 1. Evidently, our bound is pessimistic - the approximation obtained is much better.

8 Example: Numerical compression of a potential operator - continued

Operator compression is the ability to store, apply and compute functions of very large matrices. As these numerical tasks determine the limit of many scientific com-

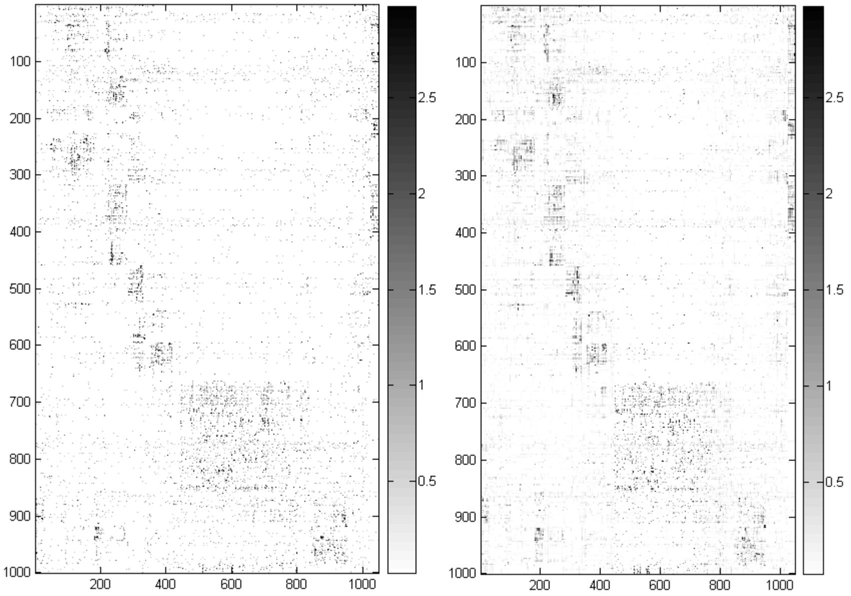


Fig. 8 Science News matrix, rows and columns reordered by trees. **Left:** Original data matrix. **Right:** Approximated matrix using 0.15 of the tensor Haar-like coefficients.

putations and simulations, it is a fundamental problem of numerical analysis. When the matrix at hand describes particle interaction, such as Columbic interaction in three-dimensional space, multiscale [14] and wavelet [3] methods have proved extremely useful. (See [16] for a survey). In §3 we saw that we can organize a potential matrix even when the geometries involved are unknown and the rows and columns are given in a random order. We now return to this example and show that, having organized the matrix, it can be compressed efficiently in the tensor Haar basis. Indeed, a procedure consisting of recovering the coupled geometry of an operator and compressing it in the induced tensor Haar basis is a natural extension of the ideas of [14, 3] to the setting where one is given only the operator describing unknown interaction on unknown geometries.

We observed that the approximation theorem yields an operator approximation scheme. Suppose that $M_{i,j}$ is the matrix of an operator and suppose we can construct a tree $\mathscr{T}_{rows}$ on the rows of M and a tree $\mathscr{T}_{cols}$ on the columns of M such that the ℓ_1 entropy of $M = \sum_{i=1}^{N} \langle M, \psi_i \rangle \, \psi_i$ in the induces tensor Haar-like basis $\Psi = \{\psi_i\}_{i=1}^{N}$, namely $e_1 (M) = \sum_{i=1}^{N} |\langle M, \psi_i \rangle|$, is small. (As before, $N = \#rows \cdot \#cols$ is the number of entries in M). Denote by M_ε the approximation

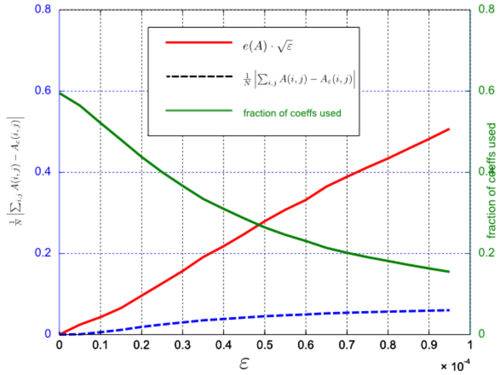


Fig. 9 Science News ℓ_1 approximation error, theoretical bound and fraction of coefficients retained over ε

$$M_\varepsilon = \sum_{\substack{1\leqslant i\leqslant N \text{ with } |\langle f,\psi_i\rangle|>\varepsilon^{\frac{1}{p}} \\ \text{and } |R(\psi_i)|>\varepsilon}} \langle M,\psi_R\rangle\,\psi_R$$

for M. From the approximation theorem, Theorem 1, we obtain that

$$\frac{1}{N}\sum_{x,y}|M(x,y)-M_\varepsilon(x,y)| \leqslant \sqrt{\varepsilon}\cdot\left(\sqrt{\tau(\Psi)}+1\right)e_1(M)\,,$$

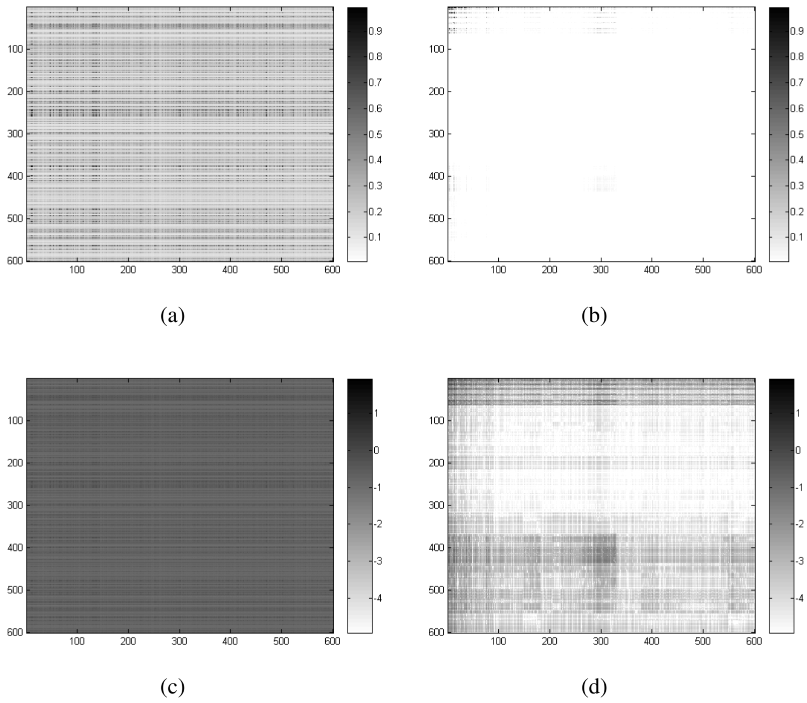
where

$$e_1(M) = \sum_{i=1}^{N}|\langle M,\psi_i\rangle|\,.$$

Returning to the potential operator example from §3, we find that this yields an efficient compression scheme. The tensor Haar-like coefficients matrix is shown in Fig. 10. Fig. 11 shows the $\ell_2 \to \ell_2$ operator norm of the residual $M_\varepsilon - M$ for different values of ε. Fig. 12 shows the ℓ_1 norm of the residual and the bound from Theorem 1. It seems that this bound is too pessimistic.

9 Conclusion

Data matrices, for which both the variable/feature/covariate set and the observation set have an intrinsic organization, are very common. The framework described here proposes a data-driven geometric approach for data "self-organization" and nonparametric statistical analysis, which applies to any data set given as a product structure (of order 2 or higher). The basic assumption underlying this method is the



(a)

(b)

(c)

(d)

Fig. 10 **(a)** "Scrambled" potential operator matrix from Fig. 2. **(b)** Absolute value of the corresponding coefficient matrix in a tensor Haar-like basis (color saturation set to 1 for better visibility). **(c)** The "Scrambled" matrix on $\log_{10}$ scale. **(d)** The Absolute value of coefficient matrix on $\log_{10}$ scale.

existence of intrinsic row and column structure. This framework relies on several of observations:

1. A partition tree on a single graph, or data set, leads to Haar-like multiscale analysis of function on it [12, 13].
2. When the variables set is not arbitrary but rather carries intrinsic structure, it is useful to consider the *coupled* structure of the data matrix. Formally, we view the matrix itself as a function on the product of its row and columns spaces. This extends naturally to higher order data tensors (§2).
3. The first role of tensor product of Haar-like bases is that the absolute sum of expansion coefficients quantifies the regularity of the data set in a proposed coupled geometry (§6).
4. Their second role is for analyzing the data set. By expanding the data matrix in a "well-adapted" tensor Haar-like basis, we can analyze the data in the coefficient domain. Tensor product of Haar-like basis offers a compact representation of data tensors of rank d that is, in some sense, not affected by d. Expansion co-

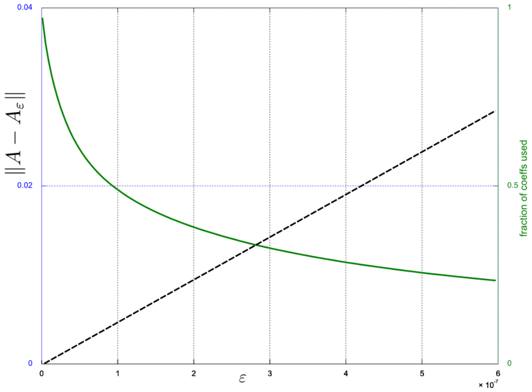


Fig. 11 $\ell_2 \to \ell_2$ operator norm of the residual and number of coefficients retained over ε

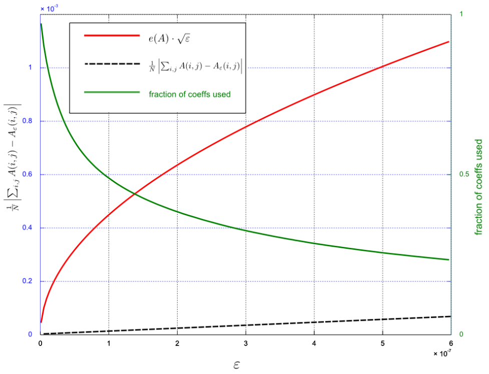


Fig. 12 ℓ_1 norm of the error, theoretical bound and number of coefficients retained over ε

efficients in a tensor Haar-like have a *one-dimensional* organization that relies on the support size $|R(\psi_i)|$, thus allowing direct use of classical signal processing ideas for high dimensional data sets (§7).

We remark that the tensor Haar-like basis used here may be replaced by a tensor product of other constructions, such as scaling functions. In particular, two distinct bases can be used for the column set and for the row set of a matrix, including any Laplacian eigenfunctions of the row or columns affinity graph. From this perspective, the Singular Value Decomposition is a special case of this construction.

Acknowledgements We are indebted to our collaborator Boaz Nadler for many insights regarding Haar-like bases. We also thank Robert Tibshirani and Fred Warner for their helpful comments. M.G. is supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship and is grateful for the hospitality of the Yale Applied Math program during the preparation of this work.

References

1. Allen, G.I., Tibshirani, R.: Transposable regularized covariance models with an application to missing data imputation. To appear in Annals of Applied Statistics (2010)
2. Belkin, M., Niyogi, P., Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **13**, 1373-1397 (2003)
3. Beylkin, G., Coifman, R., Rokhlin, V., Wavelets in numerical analysis. In: Ruskai, M.B., Beylkin, G., Coifman, R. (eds.) Wavelets and their applications, pp. 181-210. Jones and Bartlett, Boston (1992)
4. Candès, E.J., Tao, T., Near-optimal signal recovery from random projections: Universal encoding strategies?. IEEE Trans. Inform. Theory, **52**(12), 5406-5425 (2006)
5. Coifman, R.R., Donoho D.L.: Translation invariant de-noising. In: Antoniadis, A., Oppenheim, G. (eds.) Wavelets and Statistics, pp. 125-150. Springer, New York (1995)
6. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5-30 (2006)
7. Coifman, R.R, Maggioni, M.: Diffusion wavelets. Appl. Comput. Harmon. Anal. **21**(1), 54-95 (2006)
8. Coifman, R.R., Rochberg, R.: Another characterization of B.M.O. Proc. Amer. Math. Soc. **79**, 249–254 (1980)
9. Coifman, R.R., Weiss, G.: Analyse Harmonique Noncommutative sur Certains Espaces Homogenes. Springer-Verlag (1971)
10. Coifman, R.R., Weiss, G.: Extensions of Hardy spaces and their use in analysis. Bul. Of the A.M.S., **83**(4), 569–645 (1977)
11. Donoho, D.L.: Compressed Sensing. IEEE Trans. Inform. Theory **52**(4), 1289–1306 (2006)
12. Gavish, M., Nadler, B., Coifman, R.R.: Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. Proceedings of the 27th International Conference on Machine Learning, ICML (2010)
13. Gavish, M., Nadler, B., Coifman, R.R.: Inference by Haar-like wavelet analysis. preprint (2010)
14. Greengard, L., Rokhlin, V.: A fast algorithm for particle simulations, J. Comput. Phys. **73**, 325-348 (1987)
15. Lazzeroni, L., and Owen, A., Plaid models for gene expression data. Statistica Sinica **12**(1), 61–86 (2002)
16. Martinsson, P., Tygert, M., Multilevel Compression of Linear Operators: Descendants of Fast Multipole Methods and Calderón-Zygmund Theory. Lecture notes, Yale University and Courant Institute (2009) Available at `http://cims.nyu.edu/~tygert/gradcourse/survey.pdf`. Cited 30 May 2010
17. Priebe, C.E., Marchette, D.J., Park, Y., Wegman, E.J., Solka, J.L., Socolinsky, D.A., Karakos, D., Church, K.W., Guglielmi, R., Coifman, R.R., Link, D., Healy, D.M., Jacobs, M.Q., Tsao, A.: Iterative denoising for cross-corpus discovery. Proceedings of COMPSTAT 2004, Physica-Verlag/Springer (2004)
18. Singh, A., Nowak, R., Calderbank, R.: Detecting weak but hierarchically-structured patterns in networks. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, AISTATS (2010)
19. Stein, E.M., Weiss, G.: Fourier analysis on Euclidean spaces. Princeton University Press, Princeton (1971)
20. Strömberg, J.O.: Wavelets in higher dimensions. Documenta Mathematica Extra Volume ICM-1998(3), 523–532 (1998)
21. Tukey, J.W.: The future of data analysis, Ann. Math. Statist. **33**(1), 1–67 (1962)
22. Wallmann, D.M.: Multiscale diffusion coordinate refinement. Ph.D thesis, Yale University (2009)