Les Cahiers du GERAD ISSN: 0711–2440

Construction of Sparse Signal Representations with Adaptive Multiscale Orthogonal Bases

A. Saucier
C. Audet

G–2010–42

July 2010

Construction of Sparse Signal Representations with Adaptive Multiscale Orthogonal Bases

Antoine Saucier
Charles Audet*

Département de mathématiques et de génie industriel
École Polytechnique de Montréal
Montréal (Québec) Canada, H3C 3A7

{antoine.saucier, charles.audet}@polymtl.ca

* *and GERAD*

July 2010

Les Cahiers du GERAD
G–2010–42

Abstract

We propose a new approach to construct adaptive multiscale orthonormal (AMO) bases of $\mathbb{R}^N$ that provide highly sparse signal representations. Our new multilayer AMO basis design produces a high proportion of small scale vectors. The basis vectors are built from small scale to large scales, layer by layer. For each layer, the basis vector maximizes a p-norm measure of sparsity. We compare the sparsity ratios SR, (i.e., the percentage of negligibly small coefficients) obtained with AMO and Daubechies wavelet bases for seven families of piecewise smooth signals with randomly located discontinuities. The signals are composed of polynomials, sinusoids and exponentials pieces. In all cases, AMO bases produce a SR increase ranging from 6% to 37%. AMO bases have three main advantages over wavelets. First, they are found automatically by solving a sequence of optimization problems, which eliminates the problem of selecting a wavelet for a given signal. Second, they can provide a significantly sparser representation. Finally, they have the ability to produce zero coefficients for a larger family of piecewise smooth signals. The drawbacks of AMO bases are computational: the basis computation is more expensive, the basis vectors require storage space and no fast transform is known.

Key Words: Adaptive wavelets; data-adaptive basis; multiscale orthonormal basis of $\mathbb{R}^N$; sparse signal representation; optimal sparsity.

Résumé

Nous proposons une nouvelle approche pour construire des bases Multi-échelles Orthonormales Adaptatives (MOA) de $\mathbb{R}^N$ qui permet d'obtenir des représentations très creuses pour les signaux. Notre nouveau design de base multicouche produit une forte proportion de vecteurs de petite échelle. Les vecteurs de base sont construits des petites échelles vers les grandes échelles, couche par couche. Pour chaque couche, les vecteurs minimisent une mesure du nombre de cœfficients non nuls basée sur une norme p. Nous comparons la proportion de cœfficients nuls (PCN) obtenue avec les bases MOA et les ondelettes de Daubechies pour sept familles de signaux lisses par morceaux avec des discontinuités disposées aléatoirement. Les signaux sont composés de morceaux polynomiaux, sinusoïdaux et exponentiels. Dans tous les cas, les bases MOA produisent une augmentation de la PCN allant de 6% à 37%. Les bases MOA ont trois avantages principaux sur les ondelettes. D'abord, elles sont construites automatiquement en résolvant une série de problèmes d'optimisation, ce qui élimine le problème de choisir une ondelette pour un signal donné. Ensuite, elles produisent une PCN significativement plus élevée. Finalement, elles ont la capacité de produire des représentations creuses pour une famille plus large de signaux lisses par morceaux. Les désavantages des bases MOA sont au niveau des calculs : le calcul de la base est plus coûteux, les vecteurs de base requièrent un espace de stockage et il n'y a pas de transformée rapide connue.

Mots clés : Ondelettes; bases multi-échelles orthonormales adaptatives de $\mathbb{R}^N$; représentations creuses des signaux.

1 Introduction

Since their introduction, a wide variety of wavelets has been created. The choice of a wavelet for a given application has consequently become a significant problem to overcome [15, 14]. In practice, several standard wavelet bases are tried and then the basis producing the best results for a given purpose is selected. However, this approach does not always produce satisfactory results and consequently many authors (e.g. [16, 3, 7, 5]) have proposed methods to design wavelets that match a specified signal. Most applications of wavelet bases take advantage of their capacity to approximate particular classes of functions with few nonzero wavelet coefficients [13, part b]. If the signal is smooth and the wavelet function has enough vanishing moments, then the wavelet coefficients are small at small scales. Several methods [6] have been developed to control the magnitude of wavelet coefficients via vanishing moments. Other methods include the matching pursuit algorithm [12], the spectral approach [11] and wavelet packets [10].

The design of wavelet basis functions was studied via the relationship between wavelet transforms and filter banks. Related work includes the following contributions: [1] studied principal component filter banks; [2] proposed multiscale principal component analysis (PCA), in which one computes the PCA of the wavelet coefficients at each scale, followed by combining the results at relevant scales; [4] proposed a dictionary of orthogonal bases generated by a set of locally adapted version of the Karhunen-Loève transform; [9] proposed a multiresolution form of the singular value decomposition (SVD), which may be viewed as a type of fast, approximate SVD; [8] proposed an optimality theory of optimal filter banks.

The quest for a multiscale orthonormal basis that provides a highly sparse representation for a specific signal family can be approached within the multiresolution framework, which gives a specific structure to the basis (e.g. the self-similarity of the wavelet functions) and provides a significant benefit: the fast wavelet transform algorithm. In this paper, we approach this quest from an optimization perspective, i.e. outside the multiresolution framework. We propose a method which allows the automatic construction of an adaptive multiscale orthogonal (AMO) basis providing an optimally sparse representation for a given family of signals.

Our AMO bases differ from discrete wavelet bases in two main ways. First, we use basis vectors of $\mathbb{R}^N$ instead of continuous functions to represent sampled signals of $\mathbb{R}^N$. Second, our basis vectors are not constrained to be self-similar, which gives them a larger degree of adaptability to the signals of interest.

The paper is divided as follows. Section 2 proposes the AMO bases and describes some of their properties. Section 3 quantifies and optimizes the sparsity of AMO bases. Then, Section 4 details our algorithm to generate AMO bases, which are compared to Daubechies wavelets bases in Section 5. Concluding remarks are presented in Section 6.

2 Design and properties of AMO bases

2.1 Multiscale orthonormal bases in $\mathbb{R}^N$

Each vector $\boldsymbol{x} = (x[0], x[1], \ldots, x[N-1])^T$ of an orthonormal basis in $\mathbb{R}^N$ has a *scale* given by $\ell(\boldsymbol{x}) = j-i+1$, where i is the smallest index such that $x[i] \neq 0$ and j is the largest index such that $x[j] \neq 0$. The interval $[i, j]$ is called the vector *support* of $\boldsymbol{x}$. Two vectors are said to be *disjoint* if their supports are non-intersecting. A vector whose scale equals N is said to be a *large scale* vector, all others are *small scale* vectors. The proportion of small scale vectors (over N) is denoted by P_{ssv}.

The *scales* of a basis $B = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ is the set $\{\ell(\boldsymbol{x}_1), \ell(\boldsymbol{x}_2), \ldots, \ell(\boldsymbol{x}_N)\} = \{\ell_1, \ell_2, \ldots, \ell_{n_{\max}}\}$, where $1 \leq \ell_1 < \ell_2 < \ldots < \ell_{n_{\max}} \leq N$. A basis is said to be *multiscale* if distinct scales exist among the basis vectors, i.e. when $n_{\max} \geq 2$. The index n varying from 1 to $n_{\max}$ is referred to as the *level*. The basis vectors will be partitioned into disjoint subsets according to their level. All vectors at a given level have the same scale.

The next subsections describe a specific family of multiscale orthonormal bases.

2.2 Construction of a multilayer and multiscale basis

For a given level $n \in \{1, 2, \ldots, n_{\max}\}$ of a basis, a *layer* is a set of disjoint vectors with scale equal to ℓ_n that differ only by a shift of their support. In this section, we propose a way to construct a basis level by level, and layer by layer. At level n, the number of layers is denoted by J_n, and each layer contains the same number of vectors, M_n.

In previous work on multiscale principal components [15], we considered basis designs for which there was a single layer per level. In the present work, we consider more general constructions in which each level typically contains several layers. An illustration of a multilayer multiscale basis of $\mathbb{R}^{32}$ with four levels is given in Figure 1, where the support of each vector is represented by a double-arrowed line segment. The values on the left of the figure are the number of degrees of freedom of the basis vectors for each layer, and this is discussed later in this section.

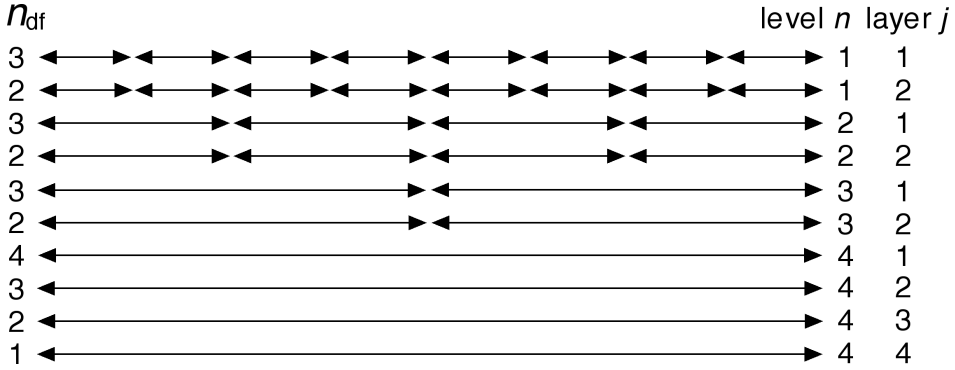

Figure 1: A multiscale basis with $N = 32$, $\ell_1 = 4$, $n_{\max} = 4$ levels and number of layers $J_1 = J_2 = J_3 = 2, J_4 = 4$, with $M_1 = 8, M_2 = 4, M_3 = 2$ and $M_4 = 1$. Supports are represented by a double-arrow line segment. Here the proportion of small scale vectors is $28/32 = 0.875$.

Each basis vector is identified by a level index n, a layer index j and a location index m that determines the vector support location. This structure is made apparent by usage of subscripts and superscripts in the notation. Basis vectors are written $\boldsymbol{x}_{n,m}^{(j)}$, and the complete basis is then given by

$$B = \{\boldsymbol{x}_{n,m}^{(j)} \in \mathbb{R}^N : n = 1, 2, \ldots, n_{\max}, j = 1, 2, \ldots, J_n, m = 1, 2, \ldots, M_n\}.$$

In a given layer, vectors differ only by a shift of their support. More precisely, at level n and layer j, consider the unit-norm vector $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ written as

$$\boldsymbol{x}_n^{(j)} := (x_n^{(j)}[0], x_n^{(j)}[1], \ldots, x_n^{(j)}[\ell_n - 1])^T.$$

The vectors of the layer are recursively constructed in $\mathbb{R}^N$ according to

$$\begin{aligned} \boldsymbol{x}_{n,1}^{(j)} &= (x_n^{(j)}[0], x_n^{(j)}[1], \ldots, x_n^{(j)}[\ell_n - 1], 0, 0, \ldots, 0)^T, \\ \boldsymbol{x}_{n,m}^{(j)} &= D_n \, \boldsymbol{x}_{n,m-1}^{(j)}, \qquad m = 2, 3, \ldots, M_n, \end{aligned}$$

where D_n denotes the operator which performs a shift by ℓ_n units, i.e. for $\boldsymbol{a} \in \mathbb{R}^N$,

$$D_n(a_1, a_2, \ldots, a_N)^T = (\underbrace{0, 0, \ldots, 0}_{\ell_n}, a_1, a_2, \ldots, a_{N-\ell_n})^T.$$

Let $\mathcal{S}_{n,m} := [(m-1)\ell_n, m\,\ell_n - 1]$ denote the support of $\boldsymbol{x}_{n,m}^{(j)}$. This support depends only on n and m, it is independent of the layer j. Observe that this construction ensures that vectors in the same layer are pairwise disjoint.

Our construction of the basis vectors is sequential, from small to large scale, one layer at a time. It produces unit-norm basis vectors orthogonal to all previously constructed ones and ensures that the scale of

the vectors at a specific level is twice that of the previous level, and the scale at the last level is exactly equal to N:

$$\ell_1 \geq 2, \quad \ell_n = 2\,\ell_{n-1}, \quad n = 2, 3, \ldots, n_{\max}, \quad \text{and} \quad \ell_{n_{\max}} = N. \tag{1}$$

The number of degrees of freedom for the choice of the first vector $\boldsymbol{x}_{n,1}^{(j)}$ of level n and layer j is given by $n_{\mathrm{df}}(n, j) := \ell_n - k(n) - j$, where $k(n)$ denotes the number of previously generated vectors of level strictly less than n, whose support intersects $[0, \ell_n - 1]$. Note that $n_{\mathrm{df}}(n, j)$ takes into account both the orthogonality and the normalization constraints. In the example, $k(1) = 0, k(2) = 4, k(3) = 12$ and $k(4) = 28$. The quantity $n_{\mathrm{df}}(n, j)$ is the number of parameters that can be adjusted to obtain a sparse representation. For example, in the spirit of wavelets, we may want each basis vector to be orthogonal to a constant vector and to a linear vector (i.e. $(0, 1, 2, \ldots, \ell_n - 1)^T$ at level n), which would require to have $n_{\mathrm{df}}(n, j) \geq 2$ for each basis vector.

The number of degrees of freedom is an important parameter because it determines the degree of adaptability of the AMO bases. For a given AMO basis, we choose a minimal value $\alpha \in \{1, 2, \ldots, N/2 - 1\}$ for the number of degrees of freedom[1] available for each basis vector up to level $n_{\max} - 1$. The last level uses up all remaining degrees of freedom. The minimal number of degrees of freedom, together with the objective of maximizing the P_{ssv}, determines the number of layers at each level of the AMO basis.

Let us consider again the example in Figure 1 in which the minimal number degree of freedom α is arbitrarily fixed to 2. The construction starts at level $n = 1$, layer $j = 1$. Since $\alpha = 2$, each level-one vector must have a scale greater than or equal to three, i.e. $\ell_1 \geq 3$ (we arbitrarily chose $\ell_1 = 4$). The first line of the figure represents the $M_1 = 8$ disjoint vectors of the layer with a degree of freedom $n_{\mathrm{df}}(1, 1) = \ell_1 - 1 = 3$, exceeding the minimum requirement of $\alpha = 2$. The construction adds a second layer ($j = 2$) of vectors with the same scale. The number of degrees of freedom decreases by one: $n_{\mathrm{df}}(1, 2) = 2 \geq \alpha$ since $\boldsymbol{x}_{1,1}^{(2)}$ has to be orthogonal to $\boldsymbol{x}_{1,1}^{(1)}$. Adding a third layer to the first level is not possible, as the number of degrees of freedom would drop below α.

The construction then proceeds to level $n = 2$ and layer $j = 1$ with the scale $\ell_2 = 2\ell_1 = 8$. The vector $\boldsymbol{x}_{2,1}^{(1)}$ needs to be orthogonal to the four previously constructed vectors $\boldsymbol{x}_{1,1}^{(1)}, \boldsymbol{x}_{1,1}^{(2)}, \boldsymbol{x}_{1,2}^{(1)}$ and $\boldsymbol{x}_{1,2}^{(2)}$ (it is necessarily orthogonal to the twelve other vectors of level 1 since they are disjoint). Hence $n_{\mathrm{df}}(2, 1) = \ell_2 - 1 - 4 = 3$, which satisfies the minimal requirement of $\alpha = 2$. In a second layer ($j = 2$) in level $n = 2$, the vector $\boldsymbol{x}_{2,2}^{(1)}$ must be orthogonal to the four level-1 vectors as well as to the last generated level-2 vector, hence $n_{\mathrm{df}}(2, 2) = \ell_2 - 1 - 5 = 2 = \alpha$. No more layers are possible at this level.

The construction continues up to level $n = n_{\max} = 4$ where all degrees of freedom are used, leading to the complete basis represented in Figure 1. For this multilayer and multiscale basis construction method, it can be shown that all the levels from 2 to $n_{\max} - 1$ contain the same number of layers, equal to the minimal number of degrees of freedom α. The number of layers at the first and last levels may possibly differ from α.

2.3 Respecting the orthogonality constraints via a change of variables

The basis design described above makes it possible for each $\boldsymbol{x}_{n,m}^{(j)}$ to be orthogonal to all previously constructed vectors, including the subset of vectors that have a support which intersects the support $\mathcal{S}_{n,m}$ of $\boldsymbol{x}_{n,m}^{(j)}$. The number of the latter vectors is $k(n, j) = k(n) - j - 1$, which depends on n, j, ℓ_1 and α.

For each vector $\boldsymbol{x}_{n,m}^{(j)}$, we define the set $X_{n,m}^{(j)}$ of all the previously constructed vectors that have a support which intersects $\mathcal{S}_{n,m}$, i.e. a basis vector $\boldsymbol{x}_{\nu,\mu}^{(r)} \in B$ belongs to the set $X_{n,m}^{(j)}$ if and only if $\mathcal{S}_{\nu,\mu} \cap \mathcal{S}_{m,n} \neq \emptyset$ and $\nu \in \{1, 2, \ldots, n-1\}$, $r \in \{1, 2, \ldots, j\}$ (lower levels), or if $\nu = n$, $r \in \{1, 2, \ldots, j-1\}$ (same level, lower layers). The orthogonality of $\boldsymbol{x}_{n,m}^{(j)}$ to all the vectors in $X_{n,m}^{(j)}$ can be written in the form

$$\boldsymbol{x}^T \boldsymbol{x}_{n,m}^{(j)} = 0 \ \forall \boldsymbol{x} \in X_{n,m}^{(j)}. \tag{2}$$

[1] In general the minimal number of degrees of freedom α satisfies $\alpha \leq \ell_1 - 1$. For a multiscale basis with only two levels, i.e. $n_{\max} = 2$, we have $\ell_1 = N/2$ which yields $\alpha \leq N/2 - 1$.

The only components of $\boldsymbol{x}$ and $\boldsymbol{x}_{n,m}^{(j)}$ that contribute to the scalar product $\boldsymbol{x}^T\boldsymbol{x}_{n,m}^{(j)}$ in (2) are associated to the support $\mathcal{S}_{n,m}$ of $\boldsymbol{x}_{n,m}^{(j)}$. To simplify the expression of the orthogonality conditions (2), we introduce for each N-dimensional vector $\boldsymbol{x}_{\mu,\nu}^{(r)} \in X_{n,m}^{(j)}$ a ℓ_n-dimensional vector $\boldsymbol{z}_{\mu,\nu}^{(r)}$ whose components are equal to the components of $\boldsymbol{x}_{\mu,\nu}^{(r)}$ in the support $\mathcal{S}_{n,m}$, i.e. we define the vector set

$$Z_n^{(j)} := \{\boldsymbol{z}_{\mu,\nu}^{(r)} \in \mathbb{R}^{\ell_n} : z_{\mu,\nu}^{(r)}[k] = x_{\mu,\nu}^{(r)}[k], k \in \mathcal{S}_{n,m}, \boldsymbol{x}_{\mu,\nu}^{(r)} \in X_{n,m}^{(j)}\}. \tag{3}$$

By construction, $Z_n^{(j)}$ is independent of m. Since $x_{n,m}^{(j)}[k] = x_n^{(j)}[k]$ for every k in the support $\mathcal{S}_{n,m}$ and equal to zero outside the support, the orthogonality conditions (2) can be rewritten in the simpler form

$$\boldsymbol{z}^T\boldsymbol{x}_n^{(j)} = 0, \quad \forall \boldsymbol{z} \in Z_n^{(j)}. \tag{4}$$

Since $|Z_n^{(j)}| = k(n,j)$, the orthogonality conditions (4) imply that $\boldsymbol{x}_n^{(j)}$ belongs to a vector subspace $\mathcal{V}_n^{(j)}$ of dimension $d(n,j) := \ell_n - k(n,j)$.

Condition (4) may be compactly written as $\mathrm{A}_{n,j}\,\boldsymbol{x}_n^{(j)} = \boldsymbol{0}$, where $\mathrm{A}_{n,j}$ is a matrix whose lines are formed of the coordinates of the vectors $\boldsymbol{z} \in Z_n^{(j)}$. It follows that an orthonormal basis $\beta_n^{(j)} = \{\boldsymbol{b}_k : k = 1, 2, \ldots, d(n,j)\}$ of the subspace $\mathcal{V}_n^{(j)}$ can be obtained as the kernel of $\mathrm{A}_{n,j}$. If the vector $\boldsymbol{y}_n^{(j)} := (y[0], y[1], \ldots, y[d(n,j)-1])^T$ contains the coordinates of $\boldsymbol{x}_n^{(j)}$ in the basis $\beta_n^{(j)}$, then $\boldsymbol{x}_n^{(j)}$ and $\boldsymbol{y}_n^{(j)}$ are related by

$$\boldsymbol{x}_n^{(j)} = P_{n,j}^T\,\boldsymbol{y}_n^{(j)}, \tag{5}$$

where the k^{th} line of the matrix $P_{n,j}$ contains the coordinates of the vector $\boldsymbol{b}_k \in \beta_n^{(j)}$. The relationship (5), from elementary linear algebra, relates the coordinates of a vector expressed in two different bases. The matrix $P_{n,j}$ is rectangular. The orthonormality of the basis $\beta_n^{(j)}$ implies that $P_{n,j}P_{n,j}^T = I_{\ell_n \times \ell_n}$, the $\ell_n \times \ell_n$ identity matrix. It follows that

$$\left(\boldsymbol{x}_n^{(j)}\right)^T \boldsymbol{x}_n^{(j)} \;=\; (P_{n,j}^T\,\boldsymbol{y}_n^{(j)})^T P_{n,j}^T\,\boldsymbol{y}_n^{(j)} \;=\; \left(\boldsymbol{y}_n^{(j)}\right)^T P_{n,j}P_{n,j}^T\,\boldsymbol{y}_n^{(j)} \;=\; \left(\boldsymbol{y}_n^{(j)}\right)^T \boldsymbol{y}_n^{(j)},$$

and therefore any unit-norm vector $\boldsymbol{y}_n^{(j)} \in \mathbb{R}^{d(n,j)}$ allows to generate via $\boldsymbol{x}_n^{(j)} = P_{n,j}^T\,\boldsymbol{y}_n^{(j)}$ a unit-norm vector $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ which is orthonormal to all vectors $\boldsymbol{z} \in Z_n^{(j)}$.

For each level n and layer j, we want to choose the vector $\boldsymbol{y}_n^{(j)}$ to obtain a signal representation which is as sparse as possible. This is the topic of Section 3.

2.4 Importance of the proportion of small scale vectors for an AMO basis

For AMO bases, the P_{ssv} satisfies $0 \le P_{\text{ssv}} \le 1$. Our motivation for introducing a multilayer basis is to obtain bases having a significantly higher P_{ssv} than the previously proposed single-layer design [15]. The next section shows that bases constructed with the multilayer design satisfy the property that P_{ssv} converges to its upper bound as $n_{\max}$ goes to infinity. To explain why a high P_{ssv} is advantageous in signal processing applications, we consider in this section a simple noise reduction application.

Let $\boldsymbol{S} = \boldsymbol{p}_k + \boldsymbol{W}$ denote a signal where $\boldsymbol{p}_k \in \mathbb{R}^N$ contains the samples of an unknown degree-k polynomial and $\boldsymbol{W} \in \mathbb{R}^N$ is an unknown zero-mean white noise of variance σ^2. Our objective is to remove as much noise as possible from $\boldsymbol{S}$ to recover $\boldsymbol{p}_k$. Expressing $\boldsymbol{S}$ in the AMO basis B yields $\boldsymbol{S} = \sum_{n=1}^{n_{\max}} \sum_{j=1}^{J_n} \sum_{m=1}^{M_n} (\boldsymbol{S}^T\boldsymbol{x}_{n,m}^{(j)})\,\boldsymbol{x}_{n,m}^{(j)}$. Now, consider the situation where the P_{ssv} is high, and where all small scale vectors are orthogonal to polynomials of degree k or less, i.e. $\boldsymbol{p}_k^T\boldsymbol{x}_{n,m}^{(j)} = 0\ \forall n < n_{\max}$. In other words, B provides a sparse representation for these polynomials. It follows that $\boldsymbol{S}$ simplifies to

$$\boldsymbol{S} = \sum_{j=1}^{J_{n_{\max}}} (\boldsymbol{S}^T\boldsymbol{x}_{n_{\max},1}^{(j)})\,\boldsymbol{x}_{n_{\max},1}^{(j)} + \sum_{n=1}^{n_{\max}-1} \sum_{j=1}^{J_n} \sum_{m=1}^{M_n} (\boldsymbol{W}^T\boldsymbol{x}_{n,m}^{(j)})\,\boldsymbol{x}_{n,m}^{(j)}. \tag{6}$$

We observe from (6) that the small scale components of $\boldsymbol{S}$, i.e. the coefficients $\boldsymbol{W}^T\boldsymbol{x}_{n,m}^{(j)}$ in the second summation, are completely determined by the noise $\boldsymbol{W}$. This suggests that an estimator $\hat{\boldsymbol{p}}_k$ of $\boldsymbol{p}_k$ can be obtained

by setting to zero all the small scale components of $\boldsymbol{S}$, i.e. $\hat{\boldsymbol{p}}_k := \sum_{j=1}^{J_{n_{\max}}} (\boldsymbol{S}^T \boldsymbol{x}_{n_{\max},1}^{(j)})\, \boldsymbol{x}_{n_{\max},1}^{(j)}$.Combining this with the fact that

$$\boldsymbol{p}_k = \sum_{j=1}^{J_{n_{\max}}} (\boldsymbol{p}_k^T \boldsymbol{x}_{n_{\max},1}^{(j)})\, \boldsymbol{x}_{n_{\max},1}^{(j)}, \tag{7}$$

and with the fact that the basis vectors are orthonormal implies that $\| \hat{\boldsymbol{p}}_k - \boldsymbol{p}_k \|^2 = \sum_{j=1}^{J_{n_{\max}}} (\boldsymbol{W}^T \boldsymbol{x}_{n_{\max},1}^{(j)})^2$, where $\| \cdot \|$ denotes the Euclidean norm. Taking the expectation value yields

$$\mathbb{E}\left[\| \hat{\boldsymbol{p}}_k - \boldsymbol{p}_k \|^2 \right] = \sigma^2 J_{n_{\max}} = \sigma^2 N (1 - P_{\text{ssv}}), \tag{8}$$

because $\boldsymbol{W}$ is a white noise and $J_{n_{\max}} = N - N_{\text{ssv}}$, where $N_{\text{ssv}} := P_{\text{ssv}} \times N$ denotes the number of small scale vectors. The result (8) shows that a P_{ssv} which is closer to one leads to a better estimator $\hat{\boldsymbol{p}}_k$, i.e. to a smaller value of the mean square error.

The identity (7) indicates that any polynomial of degree k or less can be represented exactly with $J_{n_{\max}} = N(1 - P_{\text{ssv}})$ non-zero coefficients, hence a larger P_{ssv} leads to a sparser representation for these polynomials. If the noise is *white* and all small scale vectors have zero coordinates for polynomials of degree k or less, a higher P_{ssv} is advantageous for both noise reduction and representation sparsity.

2.5 Proportion of small scale vectors for an AMO basis

In this section we calculate the P_{ssv} of an AMO basis by counting all the small scale vectors. We start by observing that the number of vectors at level n and layer j, where $n \in \{1, 2, \ldots, n_{\max} - 1\}$ and $j \geq 1$, is given by $N(n, j) = 2^{n_{\max} - n}$. The total number of small scale vectors depends on the number of layers J_n at each level. For level $n = 1$ and layer j, $j \geq 1$, the number of degrees of freedom is $n_{\text{df}}(1, j) = \ell_1 - j$. The constraint on the minimal number of degrees of freedom is $n_{\text{df}}(1, j) \geq \alpha$, which implies that $j \leq \ell_1 - \alpha$ and therefore $J_1 = \ell_1 - \alpha$. For level $n \in \{2, 3, \ldots, n_{\max} - 1\}$ and layer j, $j \geq 1$, the number of layers is constant and given by $J_n = \alpha$.

The number N_{ssv} of small scale vectors can be written as a sum of level one vectors and vectors from the levels 2 to $n_{\max} - 1$, i.e.

$$N_{\text{ssv}} \;=\; \sum_{n=1}^{n_{\max}-1} N(n,1)\, J_n \;\;=\;\; 2^{n_{\max}-1}\,(\ell_1 - \alpha) + \sum_{n=2}^{n_{\max}-1} 2^{n_{\max}-n}\,\alpha \;=\; 2^{n_{\max}-1}\,\ell_1 - 2\,\alpha. \tag{9}$$

The P_{ssv} is the ratio of N_{ssv} to N. Note that the relationship (1) implies that $\ell_n = 2^{n-1}\ell_1$ for $n \in \{1, 2, \ldots, n_{\max}\}$. Since $N := \ell_{n_{\max}}$, it follows that $N = 2^{n_{\max}-1}\,\ell_1$. Using the latter expression of N and (9), we get

$$P_{\text{ssv}} = \frac{N_{\text{ssv}}}{N} = \frac{2^{n_{\max}-1}\,\ell_1 - 2\alpha}{2^{n_{\max}-1}\,\ell_1} = 1 - \frac{\alpha}{2^{n_{\max}-2}\,\ell_1}, \tag{10}$$

which converges to 1 as $n_{\max} \to \infty$. This is a significant improvement on our previous basis construction method for MPC [15], using only one layer per level, for which the P_{ssv} could not exceed $\frac{2}{3}$. For the example of Section 2.2, where $n_{\max} = 4$, $\ell_1 = 4$ and $\alpha = 2$, we get $N_{\text{ssv}} = 4 \times 2^{4-1} - 2 \times 2 = 28$ and $P_{\text{ssv}} = 1 - \frac{2}{2^{4-2} \times 4} = \frac{7}{8}$.

According to (10), if α and $n_{\max}$ are fixed, then decreasing ℓ_1 increases the P_{ssv}. It follows that the minimal value of ℓ_1, denoted by $\ell_1^{(\min)}$ in the following, yields the maximal P_{ssv} for fixed values of α and $n_{\max}$. Since the vector $\boldsymbol{x}_1^{(1)} \in \mathbb{R}^{\ell_1}$ from level one and layer one has $\ell_1 - 1$ degrees of freedom, then $\alpha \leq \ell_1 - 1 \Rightarrow \ell_1 \geq 1 + \alpha$, i.e.

$$\ell_1^{(\min)} = 1 + \alpha. \tag{11}$$

For α and $n_{\max}$ fixed, the choice $\ell_1 = 1 + \alpha$ yields the maximal P_{ssv}. We emphasize that the choice $\ell_1 > 1 + \alpha$ is also a relevant possibility because a decrease of the P_{ssv} due to an increase of ℓ_1 can always be compensated by an increase of $n_{\max}$. In the following, an AMO basis such that $\ell_1 = 1 + \alpha$ will be called a *minimal-ℓ_1 basis*.

3 Producing a sparse representation with an AMO basis

3.1 Quantification of sparsity

At level n and layer j, the components of a signal $\boldsymbol{s}$ are given by $\{\boldsymbol{s}^T\boldsymbol{x}_{n,m}^{(j)} : m = 1, 2, \ldots, M_n.\}$. Since $x_{n,m}^{(j)}[k] = x_n^{(j)}[k]\,\forall k \in \mathcal{S}_{n,m}$, the components $\boldsymbol{s}^T\boldsymbol{x}_{n,m}^{(j)}$ can always be written in the form $\{\boldsymbol{s}_i^T\boldsymbol{x}_n^{(j)} : i \in I\}$, where

$$\boldsymbol{s}_i := (s[i], s[i+1], \ldots, s[i+\ell_n - 1])^T \in \mathbb{R}^{\ell_n} \tag{12}$$

and I is a finite set of indices. To take into account both the normalization constraint $\| \boldsymbol{x}_{n,m}^{(j)} \| = 1$ and the orthogonality of $\boldsymbol{x}_{n,m}^{(j)}$ to all vectors in $X_{n,m}^{(j)}$, we showed in Section 2.3 that it suffices to apply the change of variable $\boldsymbol{x}_n^{(j)} = P_{n,j}^T\,\boldsymbol{y}_n^{(j)}$, where $\boldsymbol{y}_n^{(j)} \in \mathbb{R}^{d_n}$ is a unit-norm vector. It follows that the components of the signal $\boldsymbol{s}$ at level n and layer j may be written as

$$\boldsymbol{s}_i^T P_{n,j}^T\,\boldsymbol{y}_n^{(j)} = (P_{n,j}\,\boldsymbol{s}_i)^T\boldsymbol{y}_n^{(j)}.$$

Using the notation $\boldsymbol{\sigma}_i := P_{n,j}\,\boldsymbol{s}_i \in \mathbb{R}^{d_n}$, the signal coordinates at level n and layer j take the form

$$\{\boldsymbol{\sigma}_i^T\boldsymbol{y}_n^{(j)} \;:\; i \in I, \| \boldsymbol{y}_n^{(j)} \| = 1\}.$$

Our objective is to construct an AMO basis that provides a sparse representation for the signal $\boldsymbol{s}$, i.e. a representation that maximizes the number of zero coordinates. Considering only the signal coordinates from level n, layer j, a naive approach would be to minimize the number of non-zero coordinates, as measured by the function

$$g(\boldsymbol{y}_n^{(j)}) = \sum_{i \in I} u(|\boldsymbol{\sigma}_i^T\boldsymbol{y}_n^{(j)}|),$$

where u denotes the Heaviside unit-step function ($u(x) = 1$ if $x \geq 0$ and $u(x) = 0$ if $x < 0$). However, the function g lacks subtlety as it does not distinguish cases for which $|\boldsymbol{\sigma}_i^T\boldsymbol{y}_n^{(j)}| = 1$ from cases where $|\boldsymbol{\sigma}_i^T\boldsymbol{y}_n^{(j)}| = 10^{-14}$: both cases will be counted as non-zero values. For this reason, it seems more appropriate to replace the Heaviside function by a *continuous approximation*. For this purpose, we propose the function u_p defined as

$$u_p(x) := \begin{cases} x^p & \text{if } x > 0 \\ 0 & \text{if } x \leq 0, \end{cases} \tag{13}$$

where p is a parameter chosen in the interval $]0, 1]$. For values of p close to 0, the function u_p is an approximation of the Heaviside function u in the sense that $\lim_{p \to 0^+} u_p = u$.

This leads us to quantify sparsity with the *sparsity approximation function*

$$f(\boldsymbol{y}_n^{(j)}) := \sum_{i \in I} |\boldsymbol{\sigma}_i^T\boldsymbol{y}_n^{(j)}|^p, \tag{14}$$

where $p \in]0, 1]$. Unlike the function g, smaller values of $|\boldsymbol{\sigma}_i^T\boldsymbol{y}_n^{(j)}|$ produce smaller contributions to f, as intended. In addition, we observe that decreasing the parameter p increases the degree of penalization of the non-zero values of $\boldsymbol{\sigma}_i^T\boldsymbol{y}_n^{(j)}$. It follows that the parameter p can be used to construct AMO bases having different degrees of sparsity.

To obtain a sparse representation of the signal, we need to find one of the unit-norm vectors $\boldsymbol{y}_n^{(j)}$ which minimizes the sparsity function f. The solution of this optimization problem is the topic of the next section.

3.2 Optimizing the sparsity approximation function

Consider a finite set of vectors $S = \{\boldsymbol{\sigma}_i : i \in I\} \subset \mathbb{R}^d$ where I is a finite set of indices. In order to simplify the presentation of the results presented in this section, we have removed as many indices as possible. For a given value of $p \in]0, 1]$, define the optimization p-norm minimization problem over the Euclidean unit sphere

$$\min\{f(\boldsymbol{y}) \;:\; \|\boldsymbol{y}\|^2 = 1,\ \boldsymbol{y} \in \mathbb{R}^d\}, \qquad \text{where } f(\boldsymbol{y}) = \sum_{i \in I} |\boldsymbol{\sigma}_i^T\boldsymbol{y}|^p \tag{15}$$

and let $\hat{\boldsymbol{y}}$ denote an optimal solution. Such a solution exists since the objective function is continuous and the domain is compact. Furthermore, it is trivial to see that if S does not span $\mathbb{R}^d$, then $\hat{\boldsymbol{y}}$ is orthogonal to each of the vectors in S and consequently the optimal objective function value achieves its trivial lower bound: $f(\hat{\boldsymbol{y}}) = 0$.

The case in which S spans $\mathbb{R}^d$ is not trivial. We have seen that as p converges to 0^+, the function f returns the number of non-zero values of the scalar products $\boldsymbol{\sigma}_i^T \boldsymbol{y}$. Minimizing f is then equivalent to maximizing the number of zero values of the scalar products $\boldsymbol{\sigma}_i^T \boldsymbol{y}$. For any $p \in]0, 1]$, the following theorem proves that an optimal solution $\hat{\boldsymbol{y}}$ of problem (15) has at least $d-1$ zero-valued scalar products $\boldsymbol{\sigma}_i^T \boldsymbol{y}$.

Theorem 3.1 *If S spans $\mathbb{R}^d$, then any optimal solution $\hat{\boldsymbol{y}}$ of problem (15) is orthogonal to at least $d-1$ linearly independent vectors $\boldsymbol{\sigma}_i \in S$, ensuring a minimum of $d-1$ zero values of the scalar products $\boldsymbol{\sigma}_i^T \hat{\boldsymbol{y}}$.*

Proof. The proof is done by showing that if S spans $\mathbb{R}^d$, then the set $\{\hat{\boldsymbol{y}}\} \cup \{\boldsymbol{\sigma}_i : \boldsymbol{\sigma}_i^T \hat{\boldsymbol{y}} = 0\}$ also spans $\mathbb{R}^d$, which implies that $\hat{\boldsymbol{y}}$ is orthogonal to $d-1$ linearly independent vectors of S. Let us partition the set of indices I as follows:

$$I^- := \{i \in I : \boldsymbol{\sigma}_i^T \hat{\boldsymbol{y}} < 0\},\ I^+ := \{i \in I : \boldsymbol{\sigma}_i^T \hat{\boldsymbol{y}} > 0\} \text{ and } I^0 := \{i \in I : \boldsymbol{\sigma}_i^T \hat{\boldsymbol{y}} = 0\}.$$

Suppose, by contradiction, that the set $\{\hat{\boldsymbol{y}}\} \cup \{\boldsymbol{\sigma}_i : i \in I^0\}$ does not span $\mathbb{R}^d$. Then, there exists a normalized vector $\boldsymbol{v} \in \mathbb{R}^d$ such that $\|\boldsymbol{v}\| = 1, \boldsymbol{v}^T \hat{\boldsymbol{y}} = 0$ and $\boldsymbol{v}^T \boldsymbol{\sigma}_i = 0, \forall\, i \in I^0$. Figure 2 illustrates a cross-section of R^d of the hyperplane containing $\hat{\boldsymbol{y}}$, $\boldsymbol{v}$ as well as the origin.

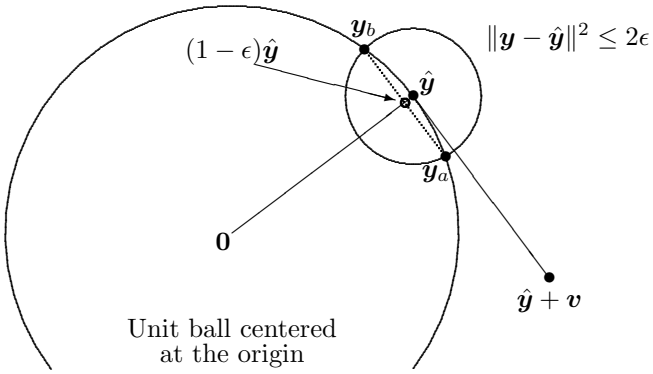

Figure 2: The point $(1-\epsilon)\hat{\boldsymbol{y}}$ belongs to the segment joining $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$.

It also follows that there exists an $\epsilon > 0$ such that

$$f(\boldsymbol{y}) = \sum_{i \in I^- \cup I^+} (\boldsymbol{c}_i^T \boldsymbol{y})^p + \sum_{i \in I^0} |\boldsymbol{\sigma}_i^T \boldsymbol{y}|^p, \text{ where } \boldsymbol{c}_i = \begin{cases} -\boldsymbol{\sigma}_i & \text{if } i \in I^- \\ \boldsymbol{\sigma}_i & \text{if } i \in I^+. \end{cases}$$

whenever $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 \leq 2\epsilon$. We define a parametrization of a line segment by $\boldsymbol{y}(t) = (1-\epsilon)\,\hat{\boldsymbol{y}} + t\,\boldsymbol{v}$, where $|t| \leq \sqrt{2\epsilon - \epsilon^2}$. This line segment is represented by a dotted line in Figure 2. The bounds on the variable t are chosen so that $\boldsymbol{y}(t)$ satisfies the constraint $\|\boldsymbol{y}(t) - \hat{\boldsymbol{y}}\|^2 \leq 2\epsilon$. Denote the endpoints of the line segment by $\boldsymbol{y}_a := (1-\epsilon)\ \hat{\boldsymbol{y}} + \sqrt{2\epsilon - \epsilon^2}\ \boldsymbol{v}$ and $\boldsymbol{y}_b := (1-\epsilon)\ \hat{\boldsymbol{y}} - \sqrt{2\epsilon - \epsilon^2}\ \boldsymbol{v}$. By construction, these points satisfy $\|\boldsymbol{y}_a\|^2 = \|\boldsymbol{y}_b\|^2 = 1$ and $\|\boldsymbol{y}_a - \hat{\boldsymbol{y}}\|^2 = \|\boldsymbol{y}_b - \hat{\boldsymbol{y}}\|^2 = 2\epsilon$. On this segment, consider the function of a single variable $g(t) := f(\boldsymbol{y}(t)) = f((1-\epsilon)\hat{\boldsymbol{y}} + t\,\boldsymbol{v})$ for $|t| \leq \sqrt{2\epsilon - \epsilon^2}$. The following properties of the function g are used below in the proof:

$$g(0) = f((1-\epsilon)\hat{\boldsymbol{y}}) = (1-\epsilon)^p f(\hat{\boldsymbol{y}}) < f(\hat{\boldsymbol{y}}) \quad \text{and} \tag{16}$$

$$g(t) = \sum_{i \in I^- \cup I^+} \left((1-\epsilon)\boldsymbol{\sigma}_i^T \hat{\boldsymbol{y}} + t\,\boldsymbol{\sigma}_i^T \boldsymbol{v}\right)^p \quad \text{when} \quad |t| \leq \sqrt{2\epsilon - \epsilon^2} \tag{17}$$

since $\boldsymbol{\sigma}_i^T \boldsymbol{v} = 0$ and $\boldsymbol{\sigma}_i^T \hat{\boldsymbol{y}} = 0$ for every $i \in I^0$. The function g is concave when $|t| \leq \sqrt{2\epsilon - \epsilon^2}$ since it is the sum of finitely many concave functions. Combining this last observation with $f(\boldsymbol{y}_a) = g(\sqrt{2\epsilon - \epsilon^2})$ and $f(\boldsymbol{y}_b) = g(-\sqrt{2\epsilon - \epsilon^2})$ ensures that

$$g(0) \quad \geq \quad \frac{1}{2}\left(g(\sqrt{2\epsilon - \epsilon^2}) + g(-\sqrt{2\epsilon - \epsilon^2})\right) \quad = \quad \frac{1}{2}\left(f(\boldsymbol{y}_a) + f(\boldsymbol{y}_b)\right). \tag{18}$$

Recall that $\hat{\boldsymbol{y}}$ is an optimal solution and that $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$ are feasible for Problem (15) because they have unit norm, and therefore $\frac{1}{2}\left(f(\boldsymbol{y}_a) + f(\boldsymbol{y}_b)\right) \geq f(\hat{\boldsymbol{y}})$. Combining this with equation (18) leads to $g(0) \geq f(\hat{\boldsymbol{y}})$, which contradicts Equation (16). □

The following corollary gives a practicable way of finding an optimal solution of (15).

Corollary 3.2 *For every J in the finite set*

$$L \quad = \quad \{J \subset I : |J| = d - 1, \text{ and } \{\boldsymbol{\sigma}_i\}_{i \in J} \text{ are linearly independent}\},$$

let $\boldsymbol{y}_J \in \mathbb{R}^d$ be a normalized vector orthogonal to every $\boldsymbol{\sigma}_i$, for $i \in J$. The vector $\boldsymbol{y}_J \in \mathbb{R}^d$ that has the least value of $f(\boldsymbol{y}_J)$ over all $J \in L$ is an optimal solution of Problem (15).

Proof. The proof follows directly from Theorem 3.1. Note that the optimal solution is not unique since $f(\boldsymbol{y}_J) = f(-\boldsymbol{y}_J)$. □

In practice, all elements J of the finite set L need to be enumerated. The vector $\boldsymbol{y}_J$ of Corollary 3.2 is obtained as the kernel of a matrix having dimensions $(d-1, d)$ whose lines are formed of the coordinates of the vectors $\boldsymbol{\sigma}_i$, where $i \in J$.

4 Algorithmic considerations

4.1 Reference signal and sparsity ratio SR

Our basis construction is iterative and proceeds from small to large scales, layer by layer. For level n and layer j, $\boldsymbol{y}_n^{(j)}$ is a normalized vector that minimizes the sparsity approximation function $f(\boldsymbol{y}_n^{(j)}) = \sum_{i \in I} |\boldsymbol{\sigma}_i^T \boldsymbol{y}_n^{(j)}|^p$, where $\boldsymbol{\sigma}_i = P_{n,j}\, \boldsymbol{s}_i$ and $\boldsymbol{s}_i \in \mathbb{R}^{\ell_n}$ is defined by (12). The signals $\boldsymbol{s}_i \in \mathbb{R}^{\ell_n}$ are contained in a signal $\boldsymbol{s}$ that we call the *reference signal*. For an AMO basis of $\mathbb{R}^N$, the size N_s of the reference signal must satisfy $N_s \geq N$. In the examples presented in this paper, we use $N_s = 1.5N$.

We introduced the sparsity approximation function f as a continuous approximation of the number of non-zero coefficients $\boldsymbol{\sigma}_i^T \boldsymbol{y}_n^{(j)}$ for a vector $\boldsymbol{y}_n^{(j)}$ and a signal $\boldsymbol{s}$. The sparsity approximation function is used in the optimization process to determine the optimal vector $\boldsymbol{y}_n^{(j)}$. In order to compare the sparsity of different bases, we return to a more direct measure of sparsity by introducing the *sparsity ratio* (SR) as percentage of near zero coefficients defined as $\text{SR} = \frac{100}{N} \times |\{k \in \{1, 2, \ldots, N\} : |c_k| < \tau\}| \in [0, 100]$, where τ is a small positive real number. For an AMO basis, the coefficients c_k of a signal $\boldsymbol{s}$ are the coordinates $\boldsymbol{s}^T \boldsymbol{x}_{n,m}^{(j)}$. For a wavelet basis, they are the discrete wavelet transform (DWT) coefficients, computed with the *Mathematica* package *Wavelets Explorer*. In both cases, we use $\tau = 10^{-12}$ to reduce the sensitivity to rounding errors.

4.2 Tackling the combinatorial explosion problem

The vector $\boldsymbol{y}_n^{(j)}$ must be orthogonal to $\ell_n - 1$ sub-signals $\boldsymbol{\sigma}_i$, where $i \in I$. The number $N^\dagger$ of subsets of $\ell_n - 1$ such sub-signals is given by

$$N^\dagger = \begin{pmatrix} |I| \\ \ell_n - 1 \end{pmatrix} = \frac{|I|!}{(\ell_n - 1)!\,(|I| - \ell_n + 1)!}.$$

The quantity $N^\dagger$ increases rapidly as $|I|$ increases and complete enumeration may be computationally intractable. For this reason, we compute the optimal solution associated to a representative subset of all available signals $\boldsymbol{s}_i$. More precisely, instead of using the set of indices I we use a *reduced set of indices*

$I_k := \{1, 1+k, 1+2k, \ldots\} \cap I$ where the *step size* $k \geq 1$ is an integer less than N. It is emphasized that the optimal vector $\boldsymbol{y}_n^{(j)}$ truly maximizes the sparsity approximation function for the representative subset of signals $\boldsymbol{s}_i$ considered, but the optimal vectors $\boldsymbol{y}_n^{(j)}$ obtained for I_{k_1} and I_{k_2} can differ if $k_1 < k_2$.

The computing time required to build the AMO basis depends on the basis size N, the reference signal size N_s, the smallest scale ℓ_1 and the minimal number of degrees of freedom α. Fortunately, we observed for the signals considered in Section 5 that highly sparse solutions could be obtained for $N \approx 1000$ and $N_s = 1.5N$ in a fairly short time (between 15 seconds and 13 minutes) using relatively large step sizes, e.g. $50 \leq k \leq 200$ (computations were made with the version 7 of the scientific computing system *Mathematica* on a MacPro 3.1 computer). Computing the AMO basis several times using a series of decreasing step sizes, we observed that the SR typically increases until it reaches a plateau for sufficiently small values of k, e.g. $k = 80$ for $N = 1280$. For $N \approx 1000$, an excessive computing time prevented us from verifying that this plateau actually extends to the unit step size of $k = 1$. In that sense, we are not certain that the maximum possible SR was obtained, which is not necessarily a drawback if the SR obtained for $k > 1$ is very high.

4.3 Choice of the number of degrees of freedom α and of the minimum scale ℓ_1

We choose the values of the minimal number of degrees of freedom α and of the minimum scale $\ell_1 \geq 1 + \alpha$ to maximize the AMO basis sparsity ratio. For a signal having an unknown structure, the optimal value of the pair (α, ℓ_1) can be determined empirically by comparing the AMO basis sparsity for different values of (α, ℓ_1). In this context, we found that the following approach was effective to estimate a good value for ℓ_1. The layers from level one have the largest number of vectors. To obtain a highly sparse signal representation, it is therefore necessary that the magnitude of most coefficients $\boldsymbol{s}_i^T \boldsymbol{x}_1^{(1)}$ be less than $\tau = 10^{-12}$. From this standpoint, one way to estimate ℓ_1 consists in building only the vector $\boldsymbol{x}_1^{(1)}$, corresponding to level one and layer one, and then in computing the percentage of negligible coefficients in $\{\boldsymbol{s}_i^T \boldsymbol{x}_1^{(1)}, i \in I\}$ as a function of ℓ_1. The value of ℓ_1 should be high enough for this percentage to be large, assuming that it is possible for the signal considered. The value of ℓ_1 selected in this way provides a lower bound on α for the AMO basis construction, i.e. $\alpha \geq \ell_1 - 1$.

The empirical approach described above is not needed for some signals having a known structure because it is possible to predict a good value for α. For example, a sparse representation of piecewise linear signals can usually be obtained by making each small scale vector $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ orthogonal to both the constant vector $\boldsymbol{P}_0$ defined by $\boldsymbol{P}_0[i] = 1$, and the linear vector $\boldsymbol{P}_1$ defined by $\boldsymbol{P}_1[i] = i$, for $i \in \{0, 1, \ldots, \ell_n - 1\}$. However, this orthogonality is possible only if $\alpha \geq 2$ for the AMO basis. In this case, we observed in our numerical experiments that AMO bases with $\alpha \geq 2$ produce highly sparse signal representations. More generally, if the signal is such that many vanishing signal components can be obtained by having each small scale vector orthogonal to q fixed vectors, where $q \geq 1$ is an integer, then we observed that AMO bases with $\alpha \geq q$ produce highly sparse signal representations. We will use this approach to choose ℓ_1 in the examples of Section 5. Also, q will be called the number of *orthogonality vectors*.

4.4 Automated selection of the exponent p defining the sparsity approximation

The AMO basis obtained for a given signal depends on the value of the exponent $p \in]0, 1]$. We observed that the values of p that maximize the SR depend on the signal family. Consequently, the optimal range used for p must be determined for each signal family and for each optimization problem. This situation led us to perform an adaptive choice of p for each layer of the AMO basis.

Our method for the determination of p is the following. In the first step, we compute the optimal solutions $\boldsymbol{x}_n^{(j)}$ for several values of p, i.e. $p = 1, 1/2, 1/2^2, \ldots, 1/2^{10}$. The optimal solutions are computed from the reduced set of vectors $\{\boldsymbol{s}_i, i \in I_k\}$. In the second step, we choose among these solutions the solution $\boldsymbol{x}_n^{(j)}$ for which the proportion of coefficients smaller than 10^{-12} is the largest, as computed from the list of coefficients $\{\boldsymbol{s}_i^T \boldsymbol{x}_n^{(j)}, i \in I\}$, i.e. using the whole reference signal. Note that most of the computational effort is concentrated in the enumeration of the elements of the set L, and that solving for several values of p does not significantly increase the computational time.

During the construction of an AMO basis, we observe that the optimal value of p varies from one layer to the next. This adaptation of the value of p for each vector layer can result in a significant increase of the SR of our AMO bases, e.g. 5% to 15%. Note that for the examples considered in this paper, the optimal values of p were observed to be always significantly larger than $1/2^{10}$.

5 Comparing sparsity for AMO bases and Daubechies wavelet bases

5.1 Comparison method

We compare the sparsity ratios for seven families of random signals. For each family, we generate a *reference signal* which is a realization of size $N_s = 1.5\,N$ of the random signal. For this reference signal, we compute a minimal-ℓ_1 AMO basis of size N. For each signal family, we also generate 100 *test-signals* of size N by splitting a new realization of size $100\,N$ into 100 adjacent and disjoint pieces of size N. The resulting test-signals are statistically independent of the reference signal.

To compare the sparsity of the bases, we compute one SR for each of the 100 signals (for both the AMO basis and the wavelets bases) and then we obtain the corresponding minimum and median SR. This approach has two advantages. First, it ensures a representative SR for the signal family. Second, it shows that optimization of the AMO basis on the reference signal is sufficient to obtain a high SR for other independent signals from the same family.

We chose the Daubechies wavelets because they have compact support, as do AMO bases. In the construction of AMO bases, both the minimal scale ℓ_1 and the values of p are adjusted to produce the highest possible SR for each signal family. In the same spirit, and to enhance the performance of Daubechies wavelets, we select for each test-signal the order $n \in \{1, 2, \ldots, 40\}$ producing the highest SR. More precisely, our procedure to compute the SR for Daubechies wavelets is the following. First, for all the 100 test-signals, i.e. for $i \in \{1, 2, ..., 100\}$, we compute the SR for all orders $n \in \{1, 2, \ldots, 40\}$ and we determine the value $n_D(i)$ of n which gives the largest value $\mathrm{SR}(i)$ of the SR for the i^{th} test-signal. Second, we compute the median and standard deviation of the orders $\{n_D(1), n_D(2), ..., n_D(100)\}$. Finally, we compute the minimal and median values of the sparsity ratios $\{\mathrm{SR}(1), \mathrm{SR}(2), ..., \mathrm{SR}(100)\}$.

For each signal family, we give in Table 1 the minimal and median values of the SR for both AMO bases and Daubechies wavelet bases, and we also give the median and standard deviation of the order of the best performing wavelets.

5.2 Signals construction

We consider piecewise smooth signals with randomly located discontinuities. The signals are constructed by juxtaposing a random number of signal pieces. The i^{th} signal piece of size N_i is composed of the samples of a continuous function h_i, i.e. it is of the form $(h_i(0), \ldots, h_i(N_i - 1))$. The N_i's are uniformly distributed integer random variables on the interval $[100, 200]$. The choice of the function h_i for the i^{th} signal piece is independent of the previously chosen signal pieces.

A signal of size N_s is constructed in three steps. First, k signal pieces are generated, where k is the smallest integer that satisfies $\sum_{i=1}^{k} N_i \geq N_s$. Second, the signal pieces are juxtaposed so that the i^{th} piece comes to the right of the $(i-1)^{\text{th}}$ piece for each $i \in \{1, 2, \ldots, k\}$. Finally, the first N_s coordinates of the resulting signal are selected. The reference signal is a realization of size $N_s \geq N$ of the random signal.

For the construction of the signal pieces, we use three types of functions h. The first type is a polynomial of degree less than or equal to 3 for which the coefficients are mutually independent random variables chosen independently for each signal piece, i.e. $h(t) = \sum_{k=0}^{3} C_k\, t^k$, where $C_k \sim \mathrm{U}[-1/N^k, 1/N^k]$ for each k. The second function type is a sinusoid of the form $h(t) = A_0 + A\cos(2\pi f t) + B\sin(2\pi f t)$, where $f = 0.02$ is constant and the coefficients A_0, A and B are random variables which are chosen independently for each signal piece, i.e. $A_0 \sim \mathrm{U}[-5, 5]$, $A = R\cos(\phi)$ and $B = R\sin(\phi)$, where A_0, $R \sim \mathrm{U}[0, 1]$ and $\phi \sim \mathrm{U}]\pi, \pi]$

are independent random variables. The third function type is an exponential of the form $h(t) = D\, e^{-t/200}$, where the coefficient $D \sim \mathrm{U}[-1, 1]$ is chosen independently for each signal piece.

Seven signal families are considered. Signals from the first three families, labeled by P, S and E in the first column of Table 1, are composed of a juxtaposition of signal pieces constructed with the same function type, i.e. polynomials, sinusoids or exponentials respectively. Signals from the next three families, labeled by P-S, P-E and S-E in Table 1, are composed of a juxtaposition of signal pieces constructed with two distinct function types. For example, the family P-E contains both polynomial signal pieces and exponential signal pieces. During the signal construction, the choice of the signal type (polynomial or exponential) for a signal piece is made randomly, giving equal probabilities to the two possibilities. Signals from the last family, labeled by P-S-E in Table 1, are composed of a juxtaposition of signal pieces constructed with the three function types. The three types of signal pieces are chosen with equal probability. A signal from the P-S-E family is displayed in Figure 3.

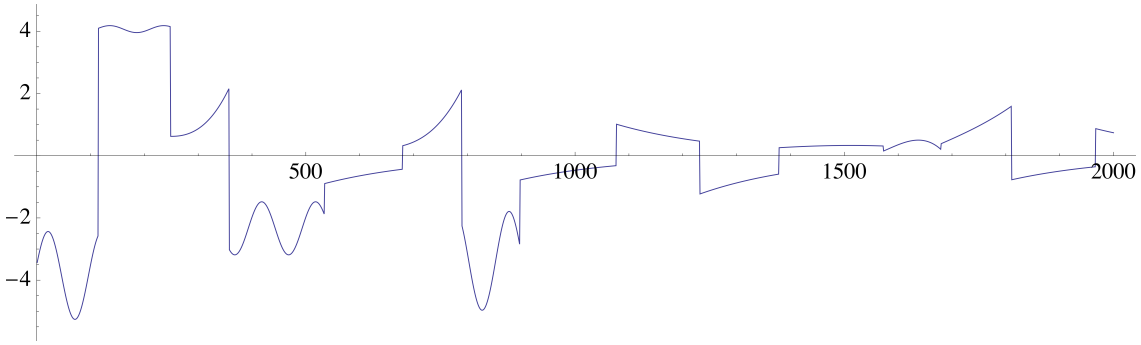

Figure 3: This is a realization from the signal family for which signals are composed of a juxtaposition of polynomials, sinusoids and exponentials.

We emphasize that the reference signal selected for the AMO basis construction should be *representative*, i.e. it should contain at least one piece of each of the signal types which are present in the signal family of interest.

5.3 Results and interpretation

The minimum and median SR obtained for AMO and Daubechies wavelet bases are given in Table 1. These percentages may vary slightly when different realizations are used for both the reference signal and the test-signals, and are written in the form $p \pm s$, where s is the standard deviation of the percentage p.

Table 1: Comparison of the sparsity ratio for AMO bases and Daubechies wavelet bases.

	AMO					Daubechies wavelets		
Signal	$n_{\max}$	ℓ_1	min SR	median SR	time mns	n_D	min SR	median SR
P	9	5	82.1 ± 0.1	84.5 ± 0.1	15	4 ± 0	75.9 ± 0.4	78.7 ± 0.2
S	10	4	85.2 ± 0.1	87.8 ± 0.3	2	28 ± 0.3	47.5 ± 1.0	52.5 ± 0.3
E	10	2	94.0 ± 0.1	95.1 ± 0.1	0.26	7 ± 0	63.9 ± 0.2	67.3 ± 0.3
P-S	8	7	73.0 ± 1.6	75.7 ± 1.0	3	28 ± 11	46.6 ± 0.9	53.3 ± 0.2
P-E	8	6	77.2 ± 0.1	80.8 ± 1.5	5.8	6 ± 0.7	62.9 ± 0.2	67.7 ± 0.04
S-E	9	6	82.4 ± 0.3	84.7 ± 0.1	3.1	28 ± 5	48.8 ± 1.0	53.4 ± 0.04
P-S-E	8	8	68.0 ± 1.0	72.0 ± 1.0	12.7	28 ± 10.0	46.5 ± 0.2	53.7 ± 0.3

5.3.1 Piecewise polynomial signals

For these signals, a sparse representation can be obtained if each small scale vector $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ is orthogonal to the four polynomial vectors $\boldsymbol{P}_k := \left(0^k, 1^k, 2^k, \ldots, (\ell_n - 1)^k\right)^T$, for $k = 0, 1, 2, 3$ (with $0^0 = 1$). These are

the $q = 4$ orthogonality vectors, as described in Section 4.3. Therefore we use $\alpha = 4$ with $\ell_1 = \alpha + 1 = 5$ because we choose the minimal-ℓ_1 basis.

Table 1 shows that both the AMO basis and the Daubechies wavelet bases have high minimal sparsity ratios, i.e. 82.1% versus 75.9% respectively, which is expected since Daubechies wavelets are designed to be orthogonal to polynomials. In this case the AMO basis has a small but significant 6% improvement in SR over the Daubechies wavelet basis. Note that the variability from one experiment to the next (changing both the reference signal and the test-signals) is small as the standard deviation is less than 0.1%. We observed that all AMO vectors from the levels 1 to 6 and some of the vectors from level 7 are orthogonal (to the numerical precision) to the polynomial signals $\boldsymbol{P}_k$, $k = 0, 1, 2, 3$. We emphasize that this orthogonality is a consequence of the optimization, as it is not built into the construction algorithm.

The AMO basis vectors are displayed in Figure 4. We observe that vectors $\boldsymbol{x}_n^{(j)}$ from many layers have a size $\ell(\boldsymbol{x}_n^{(j)})$ which is *de facto* smaller than $2^{n-1}\ell_1$. For example, vectors from the third layer of level 2 (Figure 4, left side, 4th vector from the top) and from the first three layers of level 3 (Figure 4, left side, 6th to 8th vector from the top) have nearly vanishing coordinates on about half of their support, which effectively reduces their size. Closer examination of the figure reveals that a similar situation occurs for the vectors of level 2 and layer 3, the first three layers of the levels 3, 4 and 5, the first four layers of level 6 as well as for the first three layers of level 7. This adaptive size reduction of the basis vector support size improves the ability of the AMO basis to increase its SR in the presence of discontinuities. For the piecewise polynomial reference signal, the average distance between two consecutive discontinuities is 150, hence it is advantageous to have vector sizes smaller than 150. The vectors from the four layers of level 6 (right side, top four vectors) have dimension 150 but their size has effectively been reduced to about 100 or less, which contributes to an increase of the SR. The phenomenon of adaptive size reduction is observed to various degrees in all the examples.

5.3.2 Piecewise sinusoidal signals

For these signals, a sparse representation can be obtained if each small scale vector $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ is orthogonal to the two sinusoidal vectors $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$ defined by $\boldsymbol{\alpha}_n := (\sin(2\pi f\,(0), \sin(2\pi f\,(1), \ldots, \sin(2\pi f\,(\ell_n - 1)))^T$ and $\boldsymbol{\beta}_n := (\cos(2\pi f\,(0), \cos(2\pi f\,(1), \ \ldots, \cos(2\pi f\,(\ell_n - 1)))^T$ and to the constant vector $\boldsymbol{P}_0$. There are $q = 3$ orthogonality vectors, and we use $\alpha = 3$ with $\ell_1 = 4$ (for the minimal-ℓ_1 basis). Table 1 shows that a minimal SR of 85.2% is obtained with the AMO basis, which is much higher than the minimal SR of 47.5% obtained using Daubechies wavelets. Most AMO small scale vectors $\boldsymbol{x}_n^{(j)}$ are exactly orthogonal to the sinusoidal signals $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$ and to the constant vector $\boldsymbol{P}_0$. In contrast, Daubechies wavelets can only approximate this orthogonality at the price of using a high order wavelet, i.e. $n_D = 28 \pm 0.3$. Using higher order Daubechies wavelets does not increase the SR due to the presence of signal discontinuities.

5.3.3 Piecewise exponential signals

For these signals, a sparse representation can be obtained if each $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ is orthogonal to the exponential vector $\boldsymbol{e}_n$ defined by $\boldsymbol{e}_n := \left(e^{-0/L}, e^{-1/L}, \ldots, e^{-(\ell_n-1)/L}\right)^T$. Since $q = 1$, we used $\alpha = 1$ with $\ell_1 = 2$. Table 1 displays an excellent minimal SR of 94.0% obtained with the AMO basis, which is significantly better than the 63.9% minimal SR obtained with the Daubechies wavelet. Most small scale vectors $\boldsymbol{x}_n^{(j)}$ are exactly orthogonal to the exponential signals $\boldsymbol{e}_n$, whereas Daubechies wavelets can only approximate this orthogonality with a higher order wavelet ($n_D = 7 \pm 0$). In this case, there is only one layer per level and the AMO basis is very close to the Haar basis at small scale, e.g. $\boldsymbol{x}_1^{(1)} = (0.705337, -0.708872)^T$ and $\boldsymbol{x}_2^{(1)} = (0.498736, 0.496249, -0.503748, -0.501236)^T$. This piecewise exponential signal is compressed as effectively by the AMO basis as a piecewise constant signal by a Haar basis.

5.3.4 Piecewise polynomial-sinusoidal signals

For these signals, a sparse representation can be obtained if each $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ is orthogonal to the four polynomial signals $\boldsymbol{P}_k$, $k = 0, 1, 2, 3$ and to the two sinusoidal signals $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$. Since $q = 4 + 2$, we used $\alpha = 6$ with $\ell_1 = 7$. We observe in Table 1 that a good minimal SR of 73.0% is obtained with the AMO

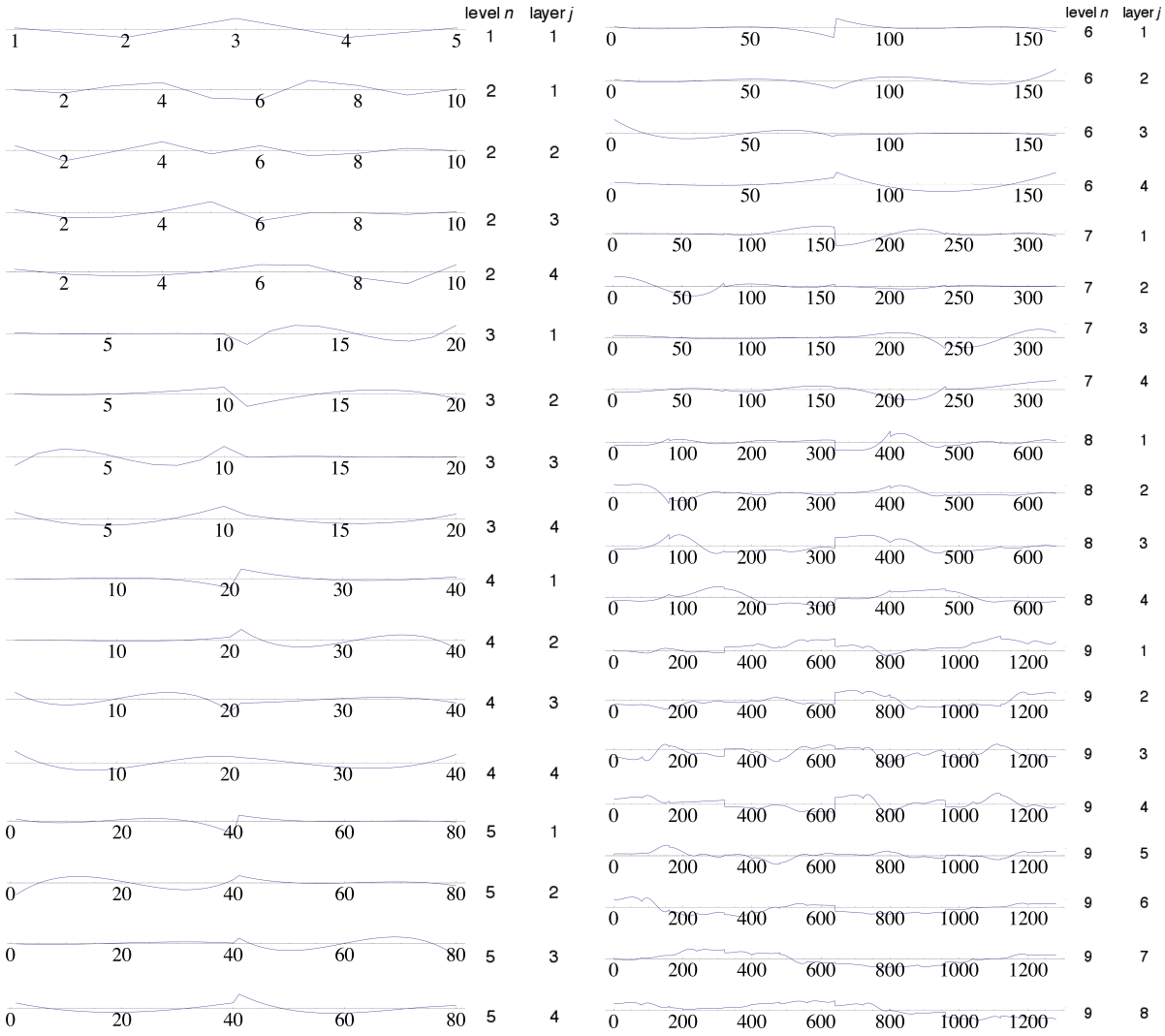

Figure 4: Basis vectors $\boldsymbol{x}_n^{(j)}$ obtained for piecewise polynomial signals, displayed from top to bottom in their construction order. The level and layer indices are indicated on the right side of each vector. Consecutive vector points are joined by linear segments.

basis, which is significantly better than the 46.6% minimal SR obtained with the Daubechies wavelets of order 28 ± 11 respectively. The high 73% SR obtained by the AMO basis is made possible by the fact that most small scale AMO basis vectors are exactly orthogonal to the four vectors $\boldsymbol{P}_k$, $k = 0, 1, 2, 3$ and the two vectors $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$. The frequency of the sinusoid is too high for Daubechies wavelets to produce better results in the presence of discontinuities.

5.3.5 Piecewise polynomial-exponential signals

For these signals, a sparse representation can be obtained if each $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ is orthogonal to the four polynomial signals $\boldsymbol{P}_k$, $k = 0, 1, 2, 3$ and to the exponential signals $\boldsymbol{e}_n$. Since $q = 4 + 1$, we used $\alpha = 5$ with $\ell_1 = 6$. We observe in Table 1 that a very good minimal SR of 77.2% is obtained with the AMO basis, which is significantly better than the 62.9% minimal SR obtained with the Daubechies wavelet of order 6 ± 0.7 respectively. In this case, the advantage of the AMO basis on the wavelet bases comes from the ability of the AMO small scale vectors to be exactly orthogonal to both polynomial and exponential signals. On the average, half of the signal pieces are polynomial and the exponential signal pieces vary fairly slowly since $T = 200$, hence wavelets naturally produce a fairly high SR in this case.

5.3.6 Piecewise sinusoidal-exponential signals

For these signals, a sparse representation can be obtained if each $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ is orthogonal to the two sinusoidal signals $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$, to the constant signal $\boldsymbol{P}_0$ and to the exponential signal $\boldsymbol{e}_n$. Since $q = 3 + 1$, we used $\alpha = 4$ with $\ell_1 = 5$. We observe in Table 1 that the AMO basis achieves a very good minimal SR of 82.4%, whereas the wavelets of order 28 ± 5 get a minimal SR of 48.8%. In this case, the Daubechies wavelets of order 28 ± 5 achieve approximate orthogonality to the exponential signals and poor orthogonality to the sinusoidal signals, whereas small scale AMO basis vectors achieve perfect orthogonality to both.

5.3.7 Piecewise polynomial-sinusoidal-exponential signals

For these signals, a sparse representation can be obtained if each $\boldsymbol{x}_n^{(j)} \in \mathbb{R}^{\ell_n}$ is orthogonal to the four polynomial signals $\boldsymbol{P}_k$, $k = 0, 1, 2, 3$, the two sinusoidal signals $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_n$ and the exponential signal $\boldsymbol{e}_n$. Since $q = 4 + 2 + 1$, we used $\alpha = 7$ with $\ell_1 = 8$. We observe in Table 1 that a good minimal SR of 68% is obtained with the AMO basis, which is significantly better than the 46.5% minimal SR obtained with the Daubechies wavelets of order 28 ± 10. The Daubechies wavelets perform fairly well because they achieve perfect orthogonality to the polynomial pieces and approximate orthogonality to the slowly varying exponential piece. In contrast, most AMO small scale vectors achieve perfect orthogonality to all three signal types. The minimal SR achieved by the AMO basis is slightly lower than for the other signals because ℓ_1 is larger and consequently the P_{ssv} is lower.

6 Concluding remarks

We presented a new design of multiscale orthonormal basis for which the proportion of small scale vectors is high, which is advantageous for both data compression and noise reduction applications. We proposed a measure of sparsity based on the p-norm and solved the optimal sparsity problem in the context of an AMO basis, i.e. maximization of sparsity from small scale to large scales.

We compared the sparsity ratios obtained with AMO bases and Daubechies wavelet bases for seven families of sampled piecewise smooth signals for which the discontinuities were located randomly. The piecewise smooth signals were composed of polynomials, sinusoids and exponentials signal pieces. We showed that AMO bases provide highly sparse representations (i.e. SR $> 68\%$) for all these signals, whereas discrete wavelets perform really well (i.e. SR $= 75.9\%$) for piecewise polynomial signals only. In all cases, the AMO basis produced a significantly higher SR than the Daubechies wavelet bases, i.e. with an increase in SR ranging from 6% to 37%. Moreover, the minimal SR produced by AMO bases always exceeded the wavelets median SR by at least 3%.

For piecewise polynomial signals, the AMO basis minimal SR of 82.1% exceeded by 6% the 75.9% minimal SR produced by Daubechies wavelets. The higher SR of the AMO basis is explained by the adaptive support size reduction of small scale vectors, which is advantageous in the presence of discontinuities. For the signal families that contain sinusoidal and/or exponential pieces, the higher SR of AMO bases is also explained by their ability to achieve perfect orthogonality to a larger family of signals, i.e. linear combinations of polynomials, sinusoids and exponentials.

AMO bases have three main advantages over discrete wavelet bases. First, they are found automatically as the solution of a sequence of optimization problems, which eliminates the problem of selecting a wavelet for a given signal. Second, they provide a possibly sparser representation. Finally, they have the ability to produce zero coefficients for a larger family of piecewise smooth signals that includes linear combinations of polynomials, sinusoids and exponentials. It follows from these advantages that significantly higher sparsity ratios (SR) can be obtained, with corresponding gains in compression ratios or noise reduction.

AMO bases also have drawbacks compared to discrete wavelet bases. First, the computation of an AMO basis for a given reference signal can be fairly intensive (e.g. minutes for $N \approx 1000$). In the context of discrete wavelet bases, the corresponding work is the choice of the basis itself. Second, the basis vectors must be stored, requiring an amount of space proportional to N^2. Finally, no fast transform is known, i.e. the

computation of the signal coordinates via scalar products in $\mathbb{R}^N$ is a computation of order N^2, compared to N for the fast wavelet transform.

In the context where signals from the same family are processed sequentially on fixed size windows, the AMO basis can be computed only once and therefore the computational cost of the AMO basis itself becomes irrelevant. It may well be worth paying this cost to get a higher sparsity ratio and corresponding gains in signal processing performance.

If one uses a value of ℓ_1 which is larger than necessary to obtain a sparse representation (i.e. $\ell_1 > q+1$), then the additional degrees of freedom available to small scale vectors make room for further optimization that could involve objectives other than sparsity, e.g. maximizing correlation with a structured noise that we want to filter. This avenue is a promising topic for future research on AMO bases.

References

[1] S. Akkarakaran and P. P. Vaidyanathan. Results on principal component filter banks: colored noise suppression and existence issues. *IEEE Transactions on Information Theory*, 47(3):1003–1020, March 2001.

[2] B.R. Bakshi. Multiscale pca with application to multivariate statistical process monitoring. *AICHE J.*, 44:1596–1610, 1998.

[3] J. O. Chapa and R. M. Rao. Algorithms for designing wavelets to match a specified signal. *IEEE Transactions on Signal Processing*, 48(12):3395–3406, 2000.

[4] R. R. Coifman and N. Saito. The local karhunen-loève bases. In *Proc.IEEE International Symposium on Time-Frequency and Time-Scale Analysis*, pages 129–132. IEEE Signal Processing Society, 1996.

[5] D.A. Donald, Y.L. Everingham, L.W. McKinna, and D. Coomans. Feature selection in the wavelet domain: Adaptive wavelets. In *Comprehensive Chemometrics*, pages 647–679. Elsevier, Oxford, 2009.

[6] J. S. Geronimo, D. P. Hardin, and P. R. Massopust. Fractal functions and wavelet expansions based on several scaling functions. *Journal of Approximation Theory*, 78(3):373–401, 1994.

[7] A. Gupta, S. D. Joshi, and S. Prasad. A new approach for estimation of statistically matched wavelet. *IEEE Transactions on Signal Processing*, 53(5):1778–1793, 2005.

[8] O. S. Jahromi, B. A. Francis, and R. H. Kwong. Algebraic theory of optimal filter banks. *IEEE Transactions on Signal Processing*, 51(2):442–457, February 2003.

[9] R. Kakarala and P. Ogunbona. Signal analysis using a multiresolution form of the singular value decomposition. *IEEE Transactions on Image Processing*, 10(5):724–735, 2001.

[10] R. E. Learned and A. S. Willsky. A wavelet packet approach to transient signal classification. *Appl. Comput. Harmonic Anal.*, 2(3):265–278, 1995.

[11] J. M. Lilly and J. Park. Multiwavelet spectral and polarization analyses of seismic records. *Geophys. J. Int.*, 122(3):1001–1021, 1995.

[12] S. Mallat and Z. Zhang. Singularity detection and processing with wavelets. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[13] Stéphane Mallat. *A wavelet tour of signal processing*. Academic Press, 525 B Street, Suite 1900, San Diego, CA 92101-4495, USA, 1998. (a) section 4.3. (b) section 7.2.1. (c) section 10.1.1. (d) section 9.1.3.

[14] Y. Mallet, D. Coomans, J. Kautsky, and O. De Vel. Classification using adaptive wavelets for feature extraction. *IEEE transactions on pattern analysis and machine intelligence*, 19(10):1058–1066, 1997.

[15] A. Saucier. Construction of data-adaptive orthogonal wavelet bases with an extension of principal component analysis. *Applied and computational harmonic analysis*, 18:300–328, 2005.

[16] A. H. Twek, D. Sinha, and P. Jorgensen. On the optimal choice of a wavelet for signal representation. *IEEE Transactions on Information Theory*, 38(2):747–765, 1992.