Electronic Transactions on Numerical Analysis.
Volume 17, pp. 168-180, 2004.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

MULTIDIMENSIONAL SMOOTHING USING HYPERBOLIC INTERPOLATORY WAVELETS [*]

MARKUS HEGLAND[†], OLE M. NIELSEN[‡], AND ZUOWEI SHEN[§]

Abstract. We propose the application of hyperbolic interpolatory wavelets for large-scale d-dimensional data fitting. In particular, we show how wavelets can be used as a highly efficient tool for multidimensional smoothing. The grid underlying these wavelets is a sparse grid. The hyperbolic interpolatory wavelet space of level j uses $O(j^{d-1}2^j)$ basis functions and it is shown that under sufficient smoothness an approximation error of order $O\left(\binom{j+d-1}{d-1}2^{-2j}\right)$ can be achieved. The implementation uses the fast wavelet transform and an efficient indexing method to access the wavelet coefficients. A practical example demonstrates the efficiency of the approach.

Key words. sparse grids, predictive modelling, wavelets, smoothing, data mining.

AMS subject classifications. 65C60, 65D10, 65T60.

1. Introduction. Predictive modelling aims to recover functional relations from observed data. More concisely, given a sequence of values of a response variable $y^{(1)}, \ldots, y^{(n)}$ and a sequence of values of a predictor variable (which is typically a vector), $(x_1^{(1)}, \ldots, x_d^{(n)})$, …, $(x_1^{(1)}, \ldots, x_d^{(n)})$ predictive modelling recovers a function f such that

$$f(x_1^{(i)}, \ldots, x_d^{(i)}) \approx y^{(i)}.$$

In the cases considered here the response and the components of the predictors are real numbers.

In a smoothing approach to predictive modelling, the function is obtained by minimising a cost functional

$$J_\alpha(f) = \sum_{i=1}^{n} (f(x^{(i)}) - y^{(i)})^2 + \alpha(\mathcal{L}f, \mathcal{L}f), \tag{1.1}$$

where α is the smoothing parameter which can be found, e.g., by cross-validation [13] and the operator $\mathcal{L}$ is densely defined in $L_2(\mathbb{R}^d)$, and typically is a differential operator. Finding a good function class which both allows the efficient approximation of the predictive function f and uses algorithms which are scalable in the number d of predictor variables $x_1, \ldots, x_d$ is a major challenge. Function classes which have been used in the past for this problem include radial basis functions [18], artificial neural nets, multivariate adaptive regression splines [13], and generalised additive models [13]. Here we present a new method for predictive modelling based on hyperbolic interpolatory wavelets. A related approach which applies sparse grids to classification is discussed in [9].

First, the function space for f should provide a good approximation of a large class of functions under reasonable smoothness assumptions. Second, the evaluation of the function

[*]Received August 27, 2002. Accepted for publication April 13, 2004. Recommended by Martin Gutknecht. The research presented here was partly funded by the Australian Advanced Computational Systems CRC, Australian Partnership for Advanced Computations (APAC), the Danish Research Council and the Academic Research Fund RP3981647. This research was partially done while the first author was visiting the Institute for Mathematical Sciences, National University of Singapore in 2003. The visit was supported by the Institute.
[†]Mathematical Sciences Institute,Australian National University, Canberra ACT 0200, Australia. E-mail: markus.hegland@anu.edu.au
[‡]Urban Risk Research Group, Geoscience Australia, Canberra ACT, Australia.
[§]Department of Mathematics, National University of Singapore, Singapore.

should be fast and the function itself should be represented by a limited number of nonzero coefficients. As we will see, hyperbolic interpolatory wavelets satisfy both conditions. If the function space for f has a relatively low dimension, then nonlinear approximation theory provides a feasible technique for the determination of an approximation with few terms [5]. This M-term approximation proceeds by first determining all the coefficients of the expansion of f in the space and then selecting the M most important ones, e.g., by using the threshold method [8]. This approach is very successful in the case of small d, e.g., in the case of image processing where $d = 2$ (see [6]). For problems of higher dimension, this approach is not feasible as the dimension of the underlying function space becomes prohibitively large. In this case, the M-term approximation technique needs to be preceded by a basis selection procedure which finds a smaller basis. This step makes the problem "seriously" nonlinear and typically greedy algorithms are used [13].

Under sufficient smoothness conditions (which are related to the operator $\mathcal{L}$), hyperbolic interpolatory wavelet spaces provide a good trade-off between function space complexity and approximation error. Hyperbolic orthogonal and bi-orthogonal wavelets were introduced in [7, 10, 11] where error analyses were also given. Further, hyperbolic bi-orthogonal wavelets have been shown to be effective for the solution of high dimensional elliptic PDEs and IEs ([15, 11]). Here we consider the case of interpolatory wavelets where the interpolation points form a sparse grid [19]. At the same time as we developed our hyperbolic interpolatory wavelet approach a related approach was developed which uses sparse grid approximations for the classification problem [9].

In Section 2 we introduce hyperbolic interpolatory wavelets and discuss their approximation properties. In Section 3 we discuss the implementation and give an application.

2. Approximation by hyperbolic interpolatory wavelets.

Data mining applications require access to function values on the grid points for further processing. For interpolatory wavelets the function values on the grid are directly related to the wavelet coefficients and can be easily retrieved. This is in contrast to orthogonal and biorthogonal wavelets for which more elaborate summations are required to retrieve the function values on the grid points. The approximation properties for hyperbolic biorthogonal wavelets have been studied in [7]. In the following we will show that similar estimates can be given for interpolatory wavelets using a different approach.

In the next subsection some basic properties of one-dimensional interpolatory wavelets are reviewed. The reader familiar with the literature on this topic can skip this subsection and move straight to the following subsection which introduces hyperbolic interpolatory wavelets.

2.1. One dimensional interpolatory wavelets.

Compactly supported interpolatory wavelets are defined by *compactly supported, interpolatory refinable functions* φ. A compactly supported function φ is *refinable* if there is a finitely supported sequence a_k, such that the function φ satisfies the following equation:

$$\varphi(x) = 2\sum_{k\in\mathbb{Z}} a_k \varphi(2x-k).$$

The sequence $\{a_k\}_{k\in\mathbb{Z}}$ is called the refinement mask for $\varphi(x)$. A continuous refinable function φ is *interpolatory* if

$$\varphi(l) = \delta_l := \begin{cases} 1, & l = 0, \\ 0, & l \in \mathbb{Z}\backslash\{0\}. \end{cases} \qquad (2.1)$$

A simple example of an interpolatory refineable function φ is the hat function given in the following example.

ETNA
Kent State University
etna@mcs.kent.edu

EXAMPLE 2.1. *Let*

$$\varphi(x) = \begin{cases} 1+x, & x \in [-1,0], \\ 1-x, & x \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

$\varphi(x)$ *is an interpolatory refinable function and its refinement mask is*

$$a_k = \begin{cases} 1/4, & k = -1, 1, \\ 1/2, & k = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Other examples are given in [3] in terms of their Fourier transforms $\widehat{\varphi}(\omega)$ by

$$\widehat{\varphi}(\omega) = \prod_{j=1}^{\infty} \widehat{a}_N(\omega/2^j), \quad \omega \in \mathbb{R}, \tag{2.2}$$

and the *symbols* are

$$\widehat{a}_N(\omega) := \cos^N(\omega/2) \left(\sum_{k=0}^{N/2-1} \binom{N/2-1+k}{k} \sin^{2k}(\omega/2) \right), \tag{2.3}$$

where N is an even number. More details on interpolatory refinable functions can be found in [14].

We will now recall the wavelet decomposition using interpolatory refinable functions. First φ defines a shift-invariant subspace V_0 of $L_\infty(\mathbb{R})$ as

$$V_0 := \overline{\operatorname{span}\{\varphi(\cdot - l) : l \in \mathbb{Z}\}}.$$

The dilations of the space V_0 are

$$V_j := \{f(2^j \cdot) : f \in V_0\}.$$

Since φ is refinable, $V_j \subset V_{j+1}$, $j \in \mathbb{Z}$. For $f \in C(\mathbb{R})$, the interpolation operator at the jth dyadic level is

$$\mathcal{P}_{V_j} f := \sum_{l \in \mathbb{Z}} f(l/2^j) \varphi_{j,l}.$$

The order of the interpolation error is known for smooth functions [4]. The smoothness can be described using Sobolev spaces $W_\infty^r(\mathbb{R})$ [1]. One also requires that the refinable function φ satisfies a Strang-Fix condition of order r which is given in terms of the Fourier transform $\widehat{\varphi}$ as

$$\widehat{\varphi}(0) \neq 0, \quad D^j \widehat{\varphi}(2\pi k) = 0, \quad j = 0, 1, \ldots, r-1, \quad k \in \mathbb{Z} \backslash \{0\},$$

where $D = \frac{d}{dx}$. Now, if f is a compactly supported function in the *Sobolev space* $W_\infty^{r+1}(\mathbb{R})$ with *Sobolev semi-norm* $|f|_{r+1}^S$ then the interpolation error $f - \mathcal{P}_{V_j} f$ satisfies

$$\left\| f - \mathcal{P}_{V_j} f \right\| \leq \text{const}\, |f|_{r+1}^S 2^{-jr}. \tag{2.4}$$

The interpolation $\mathcal{P}_{V_j} f$ interpolates f on the lattice $(\frac{1}{2^j})\mathbb{Z}$, i.e.,

$$(\mathcal{P}_{V_j} f)(l/2^j) = f(l/2^j), \quad l \in \mathbb{Z}.$$

ETNA
Kent State University
etna@mcs.kent.edu

As $V_{j-1} \subset V_j$ one can now introduce an algebraic complement W_{j-1} such that

$$V_j = V_{j-1} \oplus W_{j-1}.$$

In particular, W_j is a dilation of W_0, where W_0 is the shift-invariant subspace of $L_\infty(\mathbb{R})$ generated by

$$\psi := \varphi(2 \cdot -1).$$

It follows that $V_1 = V_0 \oplus W_0$ [8, 16], with $W_0 = \overline{\text{span}\{\psi(\cdot - l) : l \in \mathbb{Z}\}}$, and

$$W_j := \{f(2^j \cdot) : f \in W_0\},$$

and one finds that

$$\mathcal{P}_{V_j} f = \mathcal{P}_{V_{j-1}} f + \mathcal{P}_{W_{j-1}}(f - \mathcal{P}_{V_{j-1}} f).$$

The spaces V_j satisfy the decomposition

$$V_j = \bigoplus_{i=0}^{j} S_i, \qquad S_j = \begin{cases} V_0, & j = 0, \\ W_{j-1}, & j > 0. \end{cases} \tag{2.5}$$

Since $\{\varphi_{0,l}\}_l$ is a basis for V_0 and $\{\psi_{j,l}\}_l$ is a basis for W_j it follows that S_j has the basis $\{\sigma_{j,l}\}_l$ given as

$$\sigma_{j,l} = \begin{cases} \varphi_{0,l}, & j = 0, \\ \psi_{j-1,l}, & j > 0. \end{cases} \tag{2.6}$$

In general, let f be a compactly supported continuous function. Then, the sequence $\mathcal{P}_{V_j} f$ converges to f in $\|\cdot\|$ as j goes to infinity. In particular, as j goes to infinity (and $j' = 0$) f has the expansion

$$f = \sum_{j=0}^{\infty} \sum_{l \in \mathbb{Z}} s_{j,l} \sigma_{j,l},$$

where for each fixed j, the sequence $\{s_{j,l}\}_{l \in \mathbb{Z}}$ is the sequence of wavelet coefficients related to the space S_j.

The *Besov space* $B^{r'}_{p,q}(\mathbb{R})$ can be defined in many ways. One definition uses the wavelet coefficients. Let $\varphi \in C^r(\mathbb{R})$ satisfying the Strang-Fix condition of order r, with $r > r'$. Then,

$$f = \sum_{j=0}^{\infty} \sum_{l \in \mathbb{Z}} s_{j,l} \sigma_{j,l}$$

is in $B^{r'}_{p,q}(\mathbb{R})$ if and only if

$$\|s_{0,l}\|_p + (\sum_{j=1}^{\infty} (2^{jr''} (\sum_{l \in \mathbb{Z}} |s_{j,l}|^p)^{1/p})^q)^{1/q} < \infty,$$

where $r'' = r' + 1/2 - 1/p$. Furthermore, the Besov norm $|f|^B_{r'}$ is equivalent to the above number. We use in this paper the space $B^{r+1/2}_{1,1}(\mathbb{R})$. Its norm is:

$$\sum_{j=0}^{\infty} \sum_{l \in \mathbb{Z}} 2^{jr} |s_{j,l}|.$$

ETNA
Kent State University
etna@mcs.kent.edu

It follows that the wavelet coefficients satisfy

$$\sum_{l\in\mathbb{Z}} |s_{j,l}| = o(2^{-jr}).$$

As we consider compactly supported wavelets and functions, only $O(2^j)$ of the coefficients are nonzero. Thus, the average over all (nonzero) coefficients at level l is of order $o(2^{-j(r+1)})$.

If f is a compactly supported function in $C^{r+1}(\mathbb{R})$, then f is in both the *Sobolev space* $W^{r+1}_{\infty}(\mathbb{R})$ and the *Besov space* $B^{r+1/2}_{1,1}(\mathbb{R})$ (it is, in fact, in $B^{r+1}_{1,1}(\mathbb{R})$). Further facts of Sobolev and Besov spaces can be found in [5] and [8].

2.2. Hyperbolic interpolatory wavelets. The hyperbolic wavelet basis [7] is a subset of the *rectangular wavelet basis* [5]:

$$\boldsymbol{\sigma}_{\boldsymbol{j},\boldsymbol{l}} = \bigotimes_{s=1}^{d} \sigma_{j_s,l_s}, \quad l_1\ldots,l_d\in\mathbb{Z}, \quad 0\le j_1,\ldots,j_d\le j, \tag{2.7}$$

where $\boldsymbol{j} = (j_1,\ldots,j_d)$ and $\boldsymbol{l}$ is as above. The *hyperbolic interpolatory wavelet basis* is obtained from (2.7) by removing all elements for which $j_1+\cdots+j_d > j$ leaving

$$\boldsymbol{\sigma}_{\boldsymbol{j},\boldsymbol{l}} = \bigotimes_{s=1}^{d} \sigma_{j_s,l_s}, \quad l_1\ldots,l_d\in\mathbb{Z}, \quad 0\le j_1,\ldots,j_d\le j, \quad j_1+\cdots+j_d\le j, \tag{2.8}$$

where $\bigotimes_{s=1}^{d}\sigma_{j_s,l_s}(x_1,\ldots,x_d) = \sigma_{j_1,l_1}(x_1)\cdots\sigma_{j_d,l_d}(x_d)$.

These basis functions span function spaces T_j which alternatively are defined recursively as

$$T_{j+1} = T_j \bigoplus \Big(\bigoplus_{j_1+\ldots+j_d=j+1} \bigotimes_{s=1}^{d} S_{j_s} \Big), \tag{2.9}$$

where S_j are the spaces of univariate wavelets defined in (2.5) and

$$T_0 = S_0\otimes\ldots\otimes S_0.$$

Note that these function spaces are constructed using dilations and shifts of a refinable function φ. With

$$Z_j = \Big(\bigoplus_{j_1+\ldots+j_d=j+1} \bigotimes_{s=1}^{d} S_{j_s} \Big),$$

one has $T_{j+1} = T_j\oplus Z_j$. The basis functions $\sigma_{j_1,l_1}(x_1)\cdots\sigma_{j_d,l_d}(x_d)$, with $\sum_{s=1}^{d} j_s = j+1$, form a Lagrange basis for the spaces Z_j, with the grid points

$$\mathcal{I}_j := \left\{ \left(\frac{l_1}{2^{j_1}},\ldots,\frac{l_d}{2^{j_d}}\right) : l_1\ldots,l_d\in\mathbb{Z}, 0\le j_1,\ldots,j_d;\, j_1+\cdots+j_d = j \right\},$$

and thus they are 1 at $(\frac{l_1}{2^{j_1}},\ldots,\frac{l_d}{2^{j_d}})$ and 0 at all other points of $\mathcal{I}_j$. Thus one can introduce an interpolation operator on T_j by

$$P_{T_j} = P_{T_{j-1}} \oplus P_{Z_{j-1}}(I - P_{T_{j-1}}).$$

ETNA
Kent State University
etna@mcs.kent.edu

This operator provides the interpolant on the grid $\mathcal{I} = \bigcup \mathcal{I}_j$ and the grid $\mathcal{I}$ as the grid known as *sparse grid* in the literature [19, 17, 12]. Thus P_{T_j} provides a direct way to compute the sparse grid interpolant for which the combination formula can be used [11]. A bounded subset of $\mathbb{R}^d$ contains $O(j^{d-1}2^j)$ sparse grid points which is substantially less than the $O(2^{jd})$ regular grid points in the same subset. It can be seen that for compactly supported interpolatory wavelets the number of wavelets contributing to the function values for arguments in a bounded subset is $O(2^{d-1}2^j)$.

Next we estimate the interpolation error $f - \mathcal{P}_{T_j} f$. For this we need the following lemma:

LEMMA 2.1. *The cardinality of the set* $R = \{(j_1, \dots, j_d) : 0 \le j_s \le j;\ j_1 + \cdots + j_d = j\}$ *is* $\#R = \binom{j+d-1}{d-1}$.

Proof. Let $\mathcal{C}(j,d)$ denote the cardinality of R. If $j_1 = 0$ the cardinality of R is $\mathcal{C}(j, d-1)$, since there are only $d-1$ degrees of freedom. In the remaining cases where $j_1 \ge 1$, we have that $j \ge 1$ and the cardinality of R is $\mathcal{C}(j-1, d)$. Hence we have the recursion $\mathcal{C}(j,d) = \mathcal{C}(j, d-1) + \mathcal{C}(j-1, d)$. We use the fact that $\mathcal{C}(1,1) = 1$ to start an induction and the induction step over j and d to see that

$$\begin{aligned}\mathcal{C}(j,d) &= \mathcal{C}(j,d-1) + \mathcal{C}(j-1,d)\\ &= \binom{j+d-2}{d-2} + \binom{j+d-2}{d-1}\\ &= \binom{j+d-1}{d-1},\end{aligned}$$

based on the standard recursion formula for binomial coefficients. □

We are now ready to state the following theorem:

THEOREM 2.2. *Let T_j be the hyperbolic interpolatory wavelet space as defined in equation (2.9). Suppose that the underlying refinable function φ satisfies the Strang-Fix condition of order r. Let $\mathcal{P}_{T_j}$ be the interpolation operator from $C(\mathbb{R}^d)$ onto T_j. Then for an arbitrary compactly supported function f in $C^{r'}(\mathbb{R}^d)$ with $r' \ge r + d/2$,*

$$\left\| f - \mathcal{P}_{T_j} f \right\| \le \text{const } |f|^B_{r+d/2} \binom{j+d-1}{d-1} 2^{-jr},$$

where $|f|^B_{r+d/2}$ is the Besov norm of f in $B^{r+d/2}_{1,1}$. The constant is independent of f and j.

Proof. We first discuss the case that $g \in C^{r+1}(\mathbb{R}^d)$ is a compactly supported tensor product function, i.e.,

$$g = \bigotimes_{s=1}^{d} g_s,$$

so that we have $g(\boldsymbol{x}) = g_1(x_1) \cdots g_d(x_d)$.

Let g_s have the 1D expansion

$$g_s = \sum_{j_s=0}^{\infty} \sum_{l_s \in \mathbb{Z}} h_{j_s,l_s} \sigma_{j_s,l_s}, \quad s = 1, \dots, d.$$

ETNA
Kent State University
etna@mcs.kent.edu

Then, subtracting the wavelet expansion for $\mathcal{P}_{T_j} g$ from the wavelet expansion for g, we get

$$\begin{aligned} g - \mathcal{P}_{T_j} g &= \sum_{\substack{j_1+\ldots+j_d>j \\ 0\le j_1\ldots j_d}} \sum_{\boldsymbol{l}\in\mathbb{Z}^d} \bigotimes_{s=1}^{d} h_{j_s,l_s}\sigma_{j_s,l_s} \\ &= \sum_{\substack{j_1+\ldots+j_d>j \\ 0\le j_1\ldots j_d}} \bigotimes_{s=1}^{d} \sum_{\boldsymbol{l}\in\mathbb{Z}^d} h_{j_s,l_s}\sigma_{j_s,l_s}. \end{aligned}$$

Since

$$\left\| \bigotimes_{s=1}^{d} \sum_{\boldsymbol{l}\in\mathbb{Z}^d} h_{j_s,l_s}\sigma_{j_s,l_s} \right\| \le \mathrm{const}\ |g|_r^S\ 2^{-(j_1+\cdots+j_d)r},$$

we get

$$\begin{aligned} \left\| g - \mathcal{P}_{T_j} g \right\| &\le \sum_{\substack{j_1+\ldots+j_d>j \\ 0\le j_1\ldots j_d}} \left\| \bigotimes_{s=1}^{d} \sum_{\boldsymbol{l}\in\mathbb{Z}^d} h_{j_s,l_s}\sigma_{j_s,l_s} \right\| \\ &\le \mathrm{const}\ |g|_r^S \sum_{\substack{j_1+\ldots+j_d>j \\ 0\le j_1\ldots j_d}} 2^{-(j_1+\cdots+j_d)r} \\ &= \mathrm{const}\ |g|_r^S \sum_{i=j+1}^{\infty} \binom{i+d-1}{d-1} 2^{-ir}. \end{aligned}$$

Here, we have used Lemma 2.1. Although the constant 'const' may vary throughout the proof, we use the generic name 'const' for items independent of j. Rewriting the above, we get

$$\left\| g - \mathcal{P}_{T_j} g \right\| \le \mathrm{const}\ |g|_r^S \binom{j+d-1}{d-1} 2^{-jr} \left(\sum_{i=j+1}^{\infty} \binom{i+d-1}{d-1} \Big/ \binom{j+d-1}{d-1} 2^{-(i-j)r} \right).$$

We now let $k = i - j$ and observe that

$$\begin{aligned} \binom{k+j+d-1}{d-1} \Big/ \binom{j+d-1}{d-1} &= \frac{(k+j+d-1)(k+j+d-2)\cdots(k+j+1)}{(j+d-1)(j+d-2)\cdots(j+1)} \\ &= \prod_{l=1}^{d-1} \frac{k+j+l}{j+l} \\ &= \prod_{l=1}^{d-1} \left(\frac{k}{j+l} + 1 \right). \end{aligned}$$

We use this to simplify the series

$$\sum_{k=1}^{\infty} \binom{k+j+d-1}{d-1} \Big/ \binom{j+d-1}{d-1}\ 2^{-kr}$$

ETNA
Kent State University
etna@mcs.kent.edu

to

$$\sum_{k=1}^{\infty}\prod_{l=1}^{d-1}\left(\frac{k}{j+l}+1\right)2^{-kr} \;\le\; \sum_{k=1}^{\infty}(k+1)^{d-1}\,2^{-kr},$$

which is convergent by the quotient rule:

$$\lim_{k\to\infty}\left(\frac{k+2}{k+1}\right)^{d-1} 2^{-r} < 1$$

for $r > 0$. This is a constant independent of j so we have the bound

$$\|g - \mathcal{P}_{T_j} g\| \le \text{const } |g|_r^S \binom{j+d-1}{d-1} 2^{-jr} \le \mathcal{O}(j^{d-1})2^{-jr}. \tag{2.10}$$

Thus we get convergence in j for fixed d. While the bound grows exponentially in d, however, as j grows very slowly with the grid size, this dependence is close to constant.

For the general case, we first choose an interpolatory refinable function $\varphi \in C^{r'}(\mathbb{R})$, with $r' > r + d/2$, that satisfies the Strang-fix condition r'. Such a function φ can be constructed by choosing N in (2.3) sufficiently large (see, e.g., [2] and [14]). Let Φ be the tensor product of φ, i.e.,

$$\Phi(\boldsymbol{x}) := \varphi(x_1)\cdots\varphi(x_d).$$

Thus, $\Phi \in C^{r'}(\mathbb{R}^d)$ is a d-variable interpolatory refinable function. Then we can define the d-dimensional tensor product interpolatory wavelets as (see, e.g., [8])

$$\Psi_{\boldsymbol{\nu}} := \Phi(2\cdot - \boldsymbol{\nu}), \quad \boldsymbol{\nu} \in \{0,1\}^d \backslash \{\boldsymbol{0}\} =: E_0,$$

where $\{0,1\}^d$ is the set of all d dimensional vectors with entries either 0 or 1.

For an arbitrary compactly supported function $f \in C^{r'}(\mathbb{R}^d)$, with $r' > r + d/2$, expanding f in terms of this basis, one obtains that

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{l}\in\mathbb{Z}^d} c_{0,\boldsymbol{l}}\Phi(\boldsymbol{x}-\boldsymbol{l}) + \sum_{i=0}^{\infty}\sum_{\boldsymbol{l}\in\mathbb{Z}^d}\sum_{\boldsymbol{\nu}\in E_0} d_{i,\boldsymbol{l}}^{\boldsymbol{\nu}}\Psi_{\boldsymbol{\nu}}(2^i\boldsymbol{x}-\boldsymbol{l}).$$

Note that this is a different decomposition than what is used elsewhere in this paper because there are no mixed scales. This is done in order to use the bound obtained for the tensor product function (2.10) for each scale in the general proof. Therefore,

$$\begin{aligned}\|f - \mathcal{P}_{T_j} f\| \le& \sum_{\boldsymbol{l}\in\mathbb{Z}^d} |c_{0,\boldsymbol{l}}|\, \|\Phi(\cdot - \boldsymbol{l}) - \mathcal{P}_{T_j}\Phi(\cdot-\boldsymbol{l})\| \\ &+ \sum_{i=0}^{\infty}\sum_{\boldsymbol{l}\in\mathbb{Z}^d}\sum_{\boldsymbol{\nu}\in E_0} |d_{i,\boldsymbol{l}}^{\boldsymbol{\nu}}|\, \|\Psi_{\boldsymbol{\nu}}(2^i\cdot - \boldsymbol{l}) - \mathcal{P}_{T_j}\Psi_{\boldsymbol{\nu}}(2^i\cdot-\boldsymbol{l})\|.\end{aligned}$$

Since Φ and $\Psi_{\boldsymbol{\nu}}(2^i\cdot)$ are tensor product functions, we use the bound given in (2.10) to obtain

$$\|\Phi(\cdot-\boldsymbol{l}) - \mathcal{P}_{T_j}\Phi(\cdot-\boldsymbol{l})\| \le \text{const}\binom{j+d-1}{d-1}|\Phi|_r^S\, 2^{-jr},$$

and

$$\|\Psi_{\boldsymbol{\nu}}(2^i\cdot-\boldsymbol{l}) - \mathcal{P}_{T_j}\Psi_{\boldsymbol{\nu}}(2^i\cdot-\boldsymbol{l})\| \le \text{const}\binom{j+d-1}{d-1}|\Psi_{\boldsymbol{\nu}}(2^i\cdot)|_r^S\, 2^{-jr}, \quad i\in\mathbb{Z}, \boldsymbol{\nu}\in E_0.$$

ETNA
Kent State University
etna@mcs.kent.edu

Since f is in $C^{r'}(\mathbb{R}^d)$, it is in the Besov space $B^{r+d/2}_{1,1}$. Furthermore, its Besov norm $|f|^B_{r+d/2}$ is equivalent to

$$\sum_{\boldsymbol{l}\in\mathbb{Z}^d} |c_{0,l}| + \left(\sum_{i=0}^{\infty} 2^{ir}\left(\sum_{\boldsymbol{l}\in\mathbb{Z}^d}\sum_{\boldsymbol{\nu}\in E_0} |d^{\boldsymbol{\nu}}_{i,\boldsymbol{l}}|\right)\right), \tag{2.11}$$

(see [8], Theorem 2.7), and, since

$$\left|\Psi_{\boldsymbol{\nu}}(2^i\boldsymbol{x})\right|^S_r = 2^{ir}\left|\Psi_{\boldsymbol{\nu}}(\boldsymbol{x})\right|^S_r,$$

we have

$$\begin{aligned}\left\|f - \mathcal{P}_{T_j} f\right\| &\le \text{const}\binom{j+d-1}{d-1}2^{-jr}\left(\sum_{\boldsymbol{l}\in\mathbb{Z}^d}|c_{0,\boldsymbol{l}}| + \sum_{i=0}^{\infty}\sum_{\boldsymbol{l}\in\mathbb{Z}^d}\sum_{\boldsymbol{\nu}\in E_0}\left|d^{\boldsymbol{\nu}}_{i,\boldsymbol{l}}\right| 2^{ir}\right)\\ &\le \text{const}\,|f|^B_{r+d/2}\binom{j+d-1}{d-1}2^{-jr}.\end{aligned}$$

The last inequality follows from (2.11). □

REMARK 2.1. *The accuracy of the compressed (orthogonal or biorthogonal) wavelet expansion has been discussed in [7, 10, 15], where it was called the hyperbolic wavelet basis. However, the proof there, which depends on the vanishing moments of the wavelet functions, cannot be applied to the interpolatory wavelet expansion used here, since they do not have the required vanishing moments.*

The choice of interpolatory wavelets is motivated by their practical advantages. In particular, the wavelet coefficients of hyperbolic interpolatory wavelets are closely related to the function values on a sparse grid which can be retrieved with little extra computation. These function values can then be used to determine many local properties of the functions like slope, curvature, and interactions.

REMARK 2.2. *The Besov spaces provide some information about the size of the wavelet coefficients $d^{\boldsymbol{\nu}}_{i,\boldsymbol{l}}$. In fact, in the case of $f \in B^{r+d/2}_{1,1}$, one has*

$$\sum_{\boldsymbol{l}\in\mathbb{Z}^d}\sum_{\boldsymbol{\nu}\in E_0}\left|d^{\boldsymbol{\nu}}_{j,\boldsymbol{l}}\right| = o(2^{-jr}).$$

As E_0 contains $O(2^d)$ elements and for compactly supported wavelets and functions $O(2^{dj})$ of the coefficients are non-zero the average of all coefficients at level j is of order $o(2^{-j(r+d)-d})$.

3. Implementation and Application.

3.1. The smoothing problem in T_j. The hyperbolic interpolatory wavelets are now used to solve the following smoothing problem. Given the data set: $(\boldsymbol{x}^{(i)}, y^{(i)})$, $i = 1,\dots,n$, $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$, we wish to minimise the functional

$$J_\alpha(f) = \sum_{i=1}^{n}(f(\boldsymbol{x}^{(i)}) - y^{(i)})^2 + \alpha\int_{\mathbb{R}^d}|\mathcal{L}f(\boldsymbol{x})|^2\, d\boldsymbol{x},$$

where n is the number of data points and f is limited to T_j. This leads to the following matrix problem for the vector of wavelet coefficients $\boldsymbol{d}$:

$$J_\alpha(\boldsymbol{d}) = \|\boldsymbol{M}\boldsymbol{d} - \boldsymbol{y}\|^2 + \alpha\boldsymbol{d}^T\boldsymbol{L}\boldsymbol{d},$$

ETNA
Kent State University
etna@mcs.kent.edu

where

$$[\boldsymbol{L}]_{\boldsymbol{k},\boldsymbol{l}} = \int_{\mathbb{R}^d} \mathcal{L}\boldsymbol{\sigma}_{j,\boldsymbol{k}} \mathcal{L}\boldsymbol{\sigma}_{j,\boldsymbol{l}}\, d\boldsymbol{x},$$
$$[\boldsymbol{M}]_{i,\boldsymbol{l}} = \boldsymbol{\sigma}_{j,\boldsymbol{l}}(\boldsymbol{x}^{(i)}), \quad i = 1, \ldots, n.$$

This problem has the normal equations:

$$\left(\boldsymbol{M}^T\boldsymbol{M} + \alpha \boldsymbol{L}\right)\boldsymbol{d} = \boldsymbol{M}^T\boldsymbol{y}. \tag{3.1}$$

The computations of the matrices use the wavelet basis functions $\boldsymbol{\sigma}_{j,\boldsymbol{l}}$. It turns out that the computations could also be done using products of shifted and translated refinable functions φ. For this one first observes that $S_j \subset V_j$ and thus

$$\bigotimes_{s=1}^{d} S_{j_s} \subset \bigotimes_{s=1}^{d} V_{j_s}.$$

The basis of the right-hand side are just products of shifts and fixed dilations of the refinable function φ. After computing the corresponding matrices to L and M in this case the fast wavelet transform is used to determine the components relating to the wavelet spaces. This does not require the determination of the matrices for the "full" spaces $\bigotimes_{s=1}^{d} V_{j_s}$ but only for a selection of much smaller spaces. In a sense this approach is akin to the combination technique of sparse grids which also requires the solution of problems on smaller but regular spaces. In contrast to the combination method the method proposed here, however, is the exact solution in the wavelet space and not only an approximation. Moreover, the determination of the matrices in the wavelet space can be computed very efficiently using the fast wavelet transform.

3.2. Example: Predictive modelling of forest cover type. In this section we demonstrate how the method developed in the previous sections can be used in data mining for multidimensional predictive modelling. We will take a (Bayesian) classification problem as our example. The data description and the problem statement are available at the web page http://kdd.ics.uci.edu/databases/covertype/covertype.html.

The aim is to find a model for the forest cover type as a function of the following 10 cartographic parameters:

Variable	**Description**
ELEVATION	Altitude above sea level
ASPECT	Azimuth
SLOPE	Inclination
HORIZ_HYDRO	Horizontal distance to water
VERTI_HYDRO	Vertical distance to water
HORIZ_ROAD	Horizontal distance to roadways
HILL_SHADE_9	Hill shade at 9am
HILL_SHADE_12	Hill shade at noon
HILL_SHADE_15	Hill shade at 3pm
HORIZ_FIRE	Horizontal distance to fire points

All values are averaged over 30x30 meter cells and are observed on a regular grid with 30 meter spacing in both directions.

The response variable takes one of seven values (1-7) corresponding to the different possible types of forest cover. They are Spruce fir, Lodgepole pine, Ponderosa pine, Cottonwood/Willow, Aspen, Douglas fir, and Krummholz, respectively. However, for this study we will use the data to train a predictive model predicting only the presence or absence of one type of forest cover at a time. In particular, we will determine predictors for the existence of Ponderosa pine. Predictors for the other types can be obtained in similar ways. The response variable is thus

$$y = \begin{cases} 1, & \text{cover = Ponderosa pine,} \\ 0, & \text{otherwise.} \end{cases} \quad , \qquad k = 1, \ldots, 7.$$

The predictor is the vector $\boldsymbol{x}$ of cartographic measurements.

Each model found using the smoothing methodology of Section 3.1 is a continuous function f. The classifier is obtained by thresholding this function where the class is predicted to be Ponderosa pine if $f(\boldsymbol{x}) > 0.5$ and is not Ponderosa pine otherwise. On any test set the performance of the function f is estimated by the *misclassification rate*. This is the number of times the classifier predicts Ponderosa pine when in fact it is something else for the test set plus the number of times that the classifier predicts something else but the actual data is indeed Ponderosa pine.

The smoothing method requires the choice of two parameters, the smoothing parameter α and the maximal level j. If α is chosen too small and j too large then one would get *overfitting* of large misclassification rates. In order to determine these parameters the data is randomly separated into three distinct subsets of approximately equal size. The first part, the training set is used to compute functions for a variety of different parameters α and j. The second part, the test set is used to estimate the misclassification rate for these functions which have been determined from the training set. The parameters are selected to make this error estimate minimal. Finally, the third part of the data is used to estimate the misclassification rate of the model with these "optimal" parameters. For an in-depth discussion of the properties of this method, see [13].

The above problem is ten-dimensional at a first glance. However, some variables are likely to be more important than others, so we have conducted an initial 1D study predicting y as a function of each of the independent variables separately. The most important variables are then selected for predictions of high dimension.

Figure 3.1 shows approximating functions f which depend on one variable x_i only. The y-axis corresponds to the function values $f(x_i)$ and the x-axis to the values of x_i. The name of the particular variable x_i together with the overall classification rate is provided as label to each plot. It can be seen that "elevation" is the most important predictor variable with a classification rate of 0.80 followed by "distance to roads" with a rate of 0.59 and "distance to fire points" with a rate of 0.55. The rates and the best values of α and j are listed in Table 3.1

The important question at this point is how much the maximal rate of the 1D predictions can be improved by simultaneously taking more variables into account. This depends on the problem at hand and can be verified by progressively increasing the dimensionality including variables according to their importance as assessed in the 1D study. Table 3.2 shows the classification rates obtained from multivariate models. In this case j was fixed to be 3 but α was found using the test set as described above. It is seen that the best classification rate was increased from 0.80 in the 1D case to 0.87 in the six dimensional case. The third column provides a reference prediction using a generic surface where possible. The generic surface is obtained by using the spaces $V_j \otimes V_j$ instead of the hyperbolic interpolatory wavelet spaces. It is seen that the predictive power of the compressed surface is almost as good as that of a full surface. In this particular example, the gain in classification rate from increased dimensionality is modest. Other examples, where a multivariate model has much more predictive power, the advantage of using the compressed system will be greater.

ETNA
Kent State University
etna@mcs.kent.edu

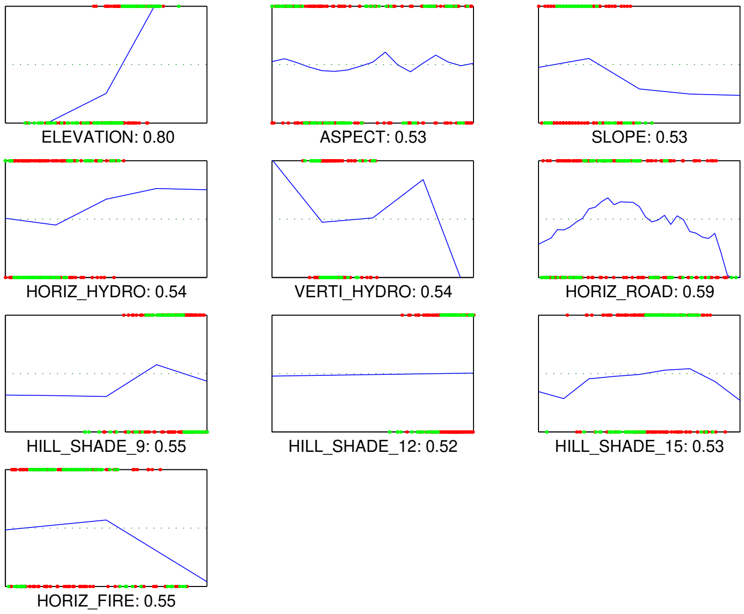

FIG. 3.1. *One dimensional predictions of Ponderosa pine. y-axis: predictions* $f(x_i)$*, x-axis: predictor* x_i *used, label: name of predictor* x_i *and classification rate.*

Variable	α	j	**Rate**
ELEVATION	100	1	0.80
ASPECT	0.1	4	0.53
SLOPE	10	2	0.53
HORIZ_HYDRO	0.01	2	0.54
VERTI_HYDRO	0.1	2	0.54
HORIZ_ROAD	1	5	0.59
HILL_SHADE_9	100	2	0.55
HILL_SHADE_12	10000	0	0.52
HILL_SHADE_15	10	2	0.52
HORIZ_FIRE	0.1	1	0.55

TABLE 3.1

Results of the 1D predictions corresponding to Figure 3.1. The chosen value of α *and* j *are given together with the classification rate for each of the ten variables.*

	Compressed system		Generic system	
Dimensions	**Rate**	**# terms**	**Rate**	**# terms**
2	0.8216	37	0.8260	81
3	0.8359	123	0.8483	729
4	0.8487	368	0.8631	6561
5	0.8587	1032	N/A	59049
6	0.8697	2768	N/A	531441

TABLE 3.2

The best multivariate predictions obtained from the compressed system and the generic system where possible. It is seen that the compressed system predicts almost as well as the generic system but it requires much less storage. The CPU time required to compute the generic system increased rapidly with increased dimensionality and it was not possible to compute the generic system for dimensions higher than 4 because of excessive storage requirements.

4. Conclusion. We have introduced hyperbolic interpolatory wavelets and demonstrated how they can be used for data mining applications, in particular predictive modelling based on smoothing. Compared to related approaches using (bi-)orthogonal wavelets the approach suggested here has advantages for data mining as function values on the grid points are readily accessible and could be used to determine properties of the functions like derivatives, and, possibly more importantly, have an immediate meaning for the application which is not the case for the wavelet coefficients of (bi-)orthogonal wavelets. The method performs well for up to around 10 dimensions. We are now generalising this approach for higher dimensions. Further information and related software can be found at our web page
http://datamining.anu.edu.au.

REFERENCES

[1] R. A. ADAMS, *Sobolev spaces*, Academic press, New York, 1975.
[2] I. DAUBECHIES, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., XLI (1988), pp. 909–996.
[3] ———, *Ten Lectures on Wavelets*, SIAM, 1992.
[4] C. DE BOOR, K. HÖLLIG, AND S. RIEMENSCHNEIDER, *Box splines*, Springer Verlag, New York, 1993.
[5] R. DEVORE, *Nonlinear approximation*, Acta Numerica, 7 (1998), pp. 51–150.
[6] R. DEVORE, B. JAWERTH, AND B. LUCIER, *Surface compression*, Comput. Aided Geom. Design, 9 (1992), pp. 219–239.
[7] R. A. DEVORE, S. V. KONYAGIN, AND V. N. TEMLYAKOV, *Hyperbolic wavelet approximation*, Constr. Approx., 14 (1998), pp. 1–26.
[8] D. DONOHO, *Interpolating wavelet transforms*. Available at the web page (1992): http://www-stat.stanford.edu/~donoho/Reports.
[9] J. GARCKE, M. GRIEBEL, AND M. THESS, *Data mining with sparse grids*, Computing, 67 (2001), pp. 225–253.
[10] M. GRIEBEL AND S. KNAPEK, *Optimized tensor-product approximation spaces*, Constr. Approx., 16 (2000), pp. 525–540.
[11] M. GRIEBEL AND P. OSWALD, *Tensor product type subspace splittings and multilevel iterative methods for anisotropic problems*, Adv. Comput. Math., 4 (1995), pp. 171–206.
[12] M. GRIEBEL AND G. ZUMBUSCH, *Adaptive sparse grids for hyperbolic conservation laws*, in Proceedings of the 7th International Conference on Hyperbolic Problems, Theory, Numerics, Applications, Birkhäuser, 1998.
[13] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer Verlag, 2001.
[14] J. HUI, S. D. RIEMENSCHNEIDER, AND Z. SHEN, *Multivariate compactly supported fundamental refinable functions, duals and biorthogonal wavelets*, Stud. Appl. Math., 102 (1999), pp. 173–204.
[15] S. KNAPEK AND F. KOSTER, *Integral operators on sparse grids*. Preprint.
[16] S. D. RIEMENSCHNEIDER AND Z. W. SHEN, *Interpolatory wavelet packets*, Appl. Comput. Harmon. Anal., 8 (2000), pp. 320–324.
[17] F. SPRENGEL, *Interpolation and wavelets on sparse Gauss-Chebyshev grids*, in Multivariate Approximation, Recent Trends and Results, W. Haussmann, K.Jetter, and M. Reimer, eds., vol. 101 of Mathematical Research, Akademie-Verlag, Berlin, 1997, pp. 269–286.
[18] G. WAHBA, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. in Appl. Math., vol. 59, SIAM, 1990.
[19] C. ZENGER, *Sparse grids*, Technical Report, Technische Universität München, 1990.