Numerical stability of biorthogonal wavelet transforms

Gerlind Plonka[1] and Hagen Schumacher[2], Manfred Tasche[2]

[1] Department of Mathematics, University of Duisburg-Essen, Campus Duisburg, 47048 Duisburg, Germany
plonka@math.uni-duisburg.de

[2] Institute of Mathematics, University of Rostock, 18055 Rostock, Germany
tasche@mathematik.uni-rostock.de, hagenschumacher@web.de

October 25, 2005

Short titel. Stability of wavelet transforms

Abstract

Biorthogonal wavelets are essential tools for numerous practical applications. It is very important that wavelet transforms work numerically stable in floating point arithmetic. This paper presents new results on the worst-case analysis of roundoff errors occurring in floating point computation of periodic biorthogonal wavelet transforms, i.e. multilevel wavelet decompositions and reconstructions. Both of these wavelet algorithms can be realized by matrix-vector products with sparse structured matrices. It is shown that under certain conditions the wavelet algorithms can be remarkably stable. Numerous tests demonstrate the performance of the results.

Mathematics Subject Classification 2000. 65T60, 65G50.

Key words. Biorthogonal wavelet transform, low-pass filter, high-pass filter, periodic wavelet transform, wavelet decomposition, wavelet reconstruction, wavelet decomposition-reconstruction, numerical stability.

1 Introduction

Biorthogonal wavelet transforms have found important applications especially in signal and image processing. In particular, biorthogonal wavelets with compact supports, leading to filter banks with FIR filters, are used very frequently [4, 5, 6, 11]. Essential reasons for the great success of the discrete wavelet transform are the low arithmetic cost and the simple implementation. However, besides the arithmetic cost also numerical stability should be taken into consideration. The numerical stability characterizes the influence of roundoff errors caused by arithmetic operations and by precomputation of transform matrices in a binary floating point arithmetic (i.e. IEEE standard), where the real input data are machine numbers and every intermediate result of the algorithm is rounded to the next machine number. The main goal of the worst-case error analysis is the qualitative investigation of the occurring roundoff errors.
For valuation of numerical stability of an algorithm, we use the concept of the backward error [21]. Here the main idea is that the roundoff error is interpreted to be obtained by application of the exact algorithm to noisy input data. This concept permits a careful analysis of the used numerical method in finite precision arithmetic.
The periodic biorthogonal wavelet transforms can be realized by matrix-vector products with sparse, structured matrices.

Let now $\mathbf{A} \in \mathbb{R}^{n \times n}$ be an invertible matrix and let $\mathbb{F}$ be the set of machine numbers. For every input vector $\mathbf{x} \in \mathbb{F}^n$ let $\mathbf{y} := \mathbf{A}\mathbf{x}$ be the exact output vector. Let $\hat{\mathbf{y}} = \mathrm{fl}(\hat{\mathbf{A}}\mathbf{x})$ be the numerical realization of the matrix-vector product by a given algorithm, where $\hat{\mathbf{A}} \in \mathbb{F}^{n \times n}$ consists of precomputed entries. The unit roundoff of the underlying floating point arithmetic is denoted by u. Then the algorithm for computing of $\mathbf{A}\mathbf{x}$ is called *normwise forward stable* (see [8], p. 142), if there exists a constant $\kappa = \kappa_{\mathbf{A}} > 0$ with $\kappa_{\mathbf{A}} u \ll 1$ such that for all $\mathbf{x} \in \mathbb{F}^n$

$$\|\hat{\mathbf{y}} - \mathbf{y}\|_2 \leq (\kappa_{\mathbf{A}} u + \mathcal{O}(u^2)) \, \|\mathbf{x}\|_2 \, .$$

Here $\| \cdot \|_2$ denotes the Euclidean norm. The numerical algorithm is said to be *normwise backward stable*, if for $\Delta \mathbf{x} := \mathbf{A}^{-1}(\hat{\mathbf{y}} - \mathbf{y})$ there exists a constant $k = k_{\mathbf{A}} > 0$ with $k_{\mathbf{A}} \, u \ll 1$ such that for all $\mathbf{x} \in \mathbb{F}^n$

$$\|\Delta \mathbf{x}\|_2 \leq (k_{\mathbf{A}} \, u + \mathcal{O}(u^2)) \, \|\mathbf{x}\|_2 \, .$$

The constant $k_{\mathbf{A}}$ is called *worst case* (*backward*) *stability constant.* Recent investigations of the roundoff errors for matrix-vector multiplications (see [8, 14, 15, 16, 19]) show the following:

1. Arithmetic cost and numerical stability are two different properties of a numerical algorithm being not directly related. In particular, comparing two numerical algorithms for the same problem, the algorithm with lower arithmetic cost does not automatically have better numerical behavior in floating point arithmetic.
2. Rounding errors in precomputed entries of the transform matrix can have essential influence on the numerical stability of the algorithm.

The numerical stability of discrete wavelet transform has been only rarely discussed in the literature. In [1], a necessary and sufficient condition for Riesz stability of biorthogonal wavelet bases in $L^2(\mathbb{R})$ has been presented. This condition is especially satisfied for the CDF wavelets in [3]. However, the Riesz stability property is not longer given for biorthogonal wavelet packets (see [2]).

A first estimate of the forward error for the discrete periodic biorthogonal wavelet transform can be found in [10]. It has been shown by Keinert [10] that some biorthogonal wavelet and wavelet packet transforms are forward stable for a small number of decomposition and reconstruction steps while for others a considerable roundoff error is accumulated. For periodic biorthogonal spline wavelets, the condition of the transform matrices grows exponentially with the spline order (see [18]).
In [14], a detailed analysis of the backward error for orthogonal wavelet transforms shows "perfect stability" in this case provided that the orthogonal filters have small filter lengths. For longer (finite) filter lengths, the arithmetic cost as well as the numerical stability of the wavelet transform matrix can be improved using a suitable orthogonal matrix factorization. First estimates of the roundoff errors for biorthogonal wavelet transforms in floating point arithmetic can be found in [16], where the worst case as well as the average case was considered.

The goal of this paper is an exact worst-case analysis of roundoff errors occurring in floating point computation of biorthogonal wavelet transforms. The paper is organized as follows. In Section 2 we introduce the periodic biorthogonal wavelet transform and show how it can be realized by matrix-vector products. We will provide numerous examples of special biorthogonal filters which are used for numerical tests. In Section 3 we give a short introduction to matrix-vector products in floating point arithmetic. Sections 4, 5, and 6 contain the new results on the numerical stability of wavelet decomposition, wavelet reconstruction, and of the total error, if the decomposition and the reconstruction are successively applied. We provide worst-case stability constants k_p which depend on the number of decomposition/reconstruction levels p, the lengths of the biorthogonal FIR filters and the spectral norms of some structured matrices which are given by the filter coefficients. In particular, these constants k_p are independent of the length of the period of the wavelet transform. Thus these constants are also correct for decomposition and reconstruction algorithms with non-periodic biorthogonal wavelets. Numerous tests demonstrate the performance of the results.

2 Periodic biorthogonal wavelet transforms

Let the real sequences $h = (h_k)_{k=-\infty}^{\infty}$ and $\tilde{h} = (\tilde{h}_k)_{k=-\infty}^{\infty}$ be real *biorthogonal low-pass filters* with finite filter lengths l_h and $l_{\tilde{h}}$, i.e., the two filters have a *finite impulse response.* Note that the *filter length* of h is explained by

$$l_h := \max\{|k-l|+1 : \, k, \, l \in \mathbb{Z} \text{ with } h_k h_l \neq 0\}.$$

The *support* of h is given by $\operatorname{supp} h := \{k \in \mathbb{Z} \, : \, h_k \neq 0\}$. In this paper, we assume that $l_h \geq 2$, $l_{\tilde{h}} \geq 2$ and put

$$\operatorname{diam}(\operatorname{supp} h) := l_h - 1, \quad \operatorname{diam}(\operatorname{supp} \tilde{h}) := l_{\tilde{h}} - 1.$$

Further, let $g = (g_k)_{k=-\infty}^{\infty}$ and $\tilde{g} = (\tilde{g}_k)_{k=-\infty}^{\infty}$ with $g_k := (-1)^k \, \tilde{h}_{1-k}$ and $\tilde{g}_k := (-1)^k \, h_{1-k}$ $(k \in \mathbb{Z})$ be the corresponding *high-pass filters.* Note that $l_g = l_{\tilde{h}}$ and $l_{\tilde{g}} = l_h$. It is known (see [20], pp. 156–158; [5], pp. 74–82; [11], pp. 21–25) that these filters possess the following properties:

$$\begin{aligned}
\sum_{n=-\infty}^{\infty} h_n\,\tilde{h}_{n-2k} = \sum_{n=-\infty}^{\infty} g_n\,\tilde{g}_{n-2k} &= \delta(k) \quad \text{(duality)},\\
\sum_{n=-\infty}^{\infty} h_n\,\tilde{g}_{n-2k} = \sum_{n=-\infty}^{\infty} g_n\,\tilde{h}_{n-2k} &= 0 \quad \text{(independence)},\\
\sum_{k=-\infty}^{\infty} (h_{m-2k}\,\tilde{h}_{n-2k} + g_{m-2k}\,\tilde{g}_{n-2k}) &= \delta(n-m) \quad \text{(perfect reconstruction)},\\
\sum_{n=-\infty}^{\infty} h_n = \sum_{n=-\infty}^{\infty} \tilde{h}_n = \sqrt{2},\ \sum_{n=-\infty}^{\infty} g_n &= \sum_{n=-\infty}^{\infty} \tilde{g}_n = 0 \quad \text{(normalization)}.
\end{aligned}$$

Here k, m, n are arbitrary integers and δ denotes the Kronecker symbol. Note that h, g, $\tilde{h}$, and $\tilde{g}$ form a filter bank of perfect reconstruction.
For $j \in \mathbb{N}$, let

$$n_j := 2^j.$$

The n_j-periodic filter coefficients are given by

$$\begin{aligned}
h_{j,k} &:= \sum_{m=-\infty}^{\infty} h_{k+n_j m}, \quad \tilde{h}_{j,k} := \sum_{m=-\infty}^{\infty} \tilde{h}_{k+n_j m},\\
g_{j,k} &:= \sum_{m=-\infty}^{\infty} g_{k+n_j m}, \quad \tilde{g}_{j,k} := \sum_{m=-\infty}^{\infty} \tilde{g}_{k+n_j m}.
\end{aligned}$$

Observe that for $n_j \geq l := \max\{l_h,\, l_g\}$, these four series contain only one nonzero term. Now we consider the matrices

$$\begin{aligned}
\mathbf{H}_j &:= (h_{j,r-2k})_{r,k=0}^{n_j,n_{j-1}}, \quad \tilde{\mathbf{H}}_j := (\tilde{h}_{j,r-2k})_{r,k=0}^{n_j,n_{j-1}},\\
\mathbf{G}_j &:= (g_{j,r-2k})_{r,k=0}^{n_j,n_{j-1}}, \quad \tilde{\mathbf{G}}_j := (\tilde{g}_{j,r-2k})_{r,k=0}^{n_j,n_{j-1}}.
\end{aligned}$$

The above properties of the filter coefficients yield the conditions of duality, independence, perfect reconstruction and normalization for these non-quadratic matrices:

$$\begin{aligned}
\mathbf{H}_j^T \tilde{\mathbf{H}}_j = \mathbf{G}_j^T \tilde{\mathbf{G}}_j &= \mathbf{I}_{j-1},\\
\mathbf{G}_j^T \tilde{\mathbf{H}}_j = \mathbf{H}_j^T \tilde{\mathbf{G}}_j &= \mathbf{O}_{j-1},\\
\mathbf{H}_j \tilde{\mathbf{H}}_j^T + \mathbf{G}_j \tilde{\mathbf{G}}_j^T &= \mathbf{I}_j,\\
\mathbf{H}_j^T \mathbf{1}_j = \tilde{\mathbf{H}}_j^T \mathbf{1}_j &= \sqrt{2}\,\mathbf{1}_{j-1},\\
\mathbf{G}_j^T \mathbf{1}_j = \tilde{\mathbf{G}}_j^T \mathbf{1}_j &= \mathbf{o}_{j-1}.
\end{aligned}$$

Here we use the notations $\mathbf{1}_j := (1,\, 1,\, \ldots\, ,\, 1)^T \in \mathbb{R}^{n_j}$, $\mathbf{o}_j := (0,\, 0,\, \ldots\, ,\, 0)^T \in \mathbb{R}^{n_j}$, $\mathbf{I}_j$ for the identity matrix of order n_j and $\mathbf{O}_j$ for the quadratic zero matrix of order n_j.
The 1-*level wavelet decomposition* or *discrete periodic biorthogonal wavelet transform* of a vector $\mathbf{c}_j = (c_{j,k})_{k=0}^{n_j-1} \in \mathbb{R}^{n_j}$ can be written in the form

$$\mathbf{c}_{j-1} = \tilde{\mathbf{H}}_j^T \mathbf{c}_j, \qquad \mathbf{d}_{j-1} = \tilde{\mathbf{G}}_j^T \mathbf{c}_j,$$

where $\mathbf{c}_{j-1} = (c_{j-1,r})_{r=0}^{n_{j-1}-1}$ is the corresponding low-pass part and $\mathbf{d}_{j-1} = (d_{j-1,r})_{r=0}^{n_{j-1}-1}$ is the high-pass part of the input signal $\mathbf{c}_j$. Equivalently, we have for the components

$$c_{j-1,r} = \sum_{k=0}^{n_j-1} \tilde{h}_{j,k-2r}\, c_{j,k}, \quad d_{j-1,r} = \sum_{k=0}^{n_j-1} \tilde{g}_{j,k-2r}\, c_{j,k}\,.$$

The 1-*level wavelet reconstruction* or *inverse discrete periodic biorthogonal wavelet transform* is given by

$$\mathbf{c}_j = \mathbf{H}_j \mathbf{c}_{j-1} + \mathbf{G}_j \mathbf{d}_{j-1},$$

i.e., for a given low-pass signal $\mathbf{c}_{j-1} \in \mathbb{R}^{n_{j-1}}$ and a given high-pass signal $\mathbf{d}_{j-1} \in \mathbb{R}^{n_{j-1}}$ the original signal $\mathbf{c}_j \in \mathbb{R}^{n_j}$ can be reconstructed. For the components it follows

$$c_{j,r} = \sum_{k=0}^{n_{j-1}-1} (h_{j,r-2k}\, c_{j-1,k} + g_{j,r-2k}\, d_{j-1,k})$$

(see e.g. [6, 10, 11, 12]). The decomposition and reconstruction will be iteratively applied. Let $j,\, p \in \mathbb{N}$ with $p < j$ and $l \leq n_{j-p+1}$ be given. Generally, the p-*level wavelet decomposition* of $\mathbf{c}_j \in \mathbb{R}^{n_j}$ is defined as the block vector

$$(\mathbf{c}_{j-p}^T,\, \mathbf{d}_{j-p}^T,\, \ldots\, ,\, \mathbf{d}_{j-1}^T)^T \in \mathbb{R}^{n_j}\,,$$

where $\mathbf{c}_\nu,\, \mathbf{d}_\nu \in \mathbb{R}^{n_\nu}$ are recursively computed by

$$\mathbf{c}_{j-\nu-1} = \tilde{\mathbf{H}}_{j-\nu}^T \mathbf{c}_{j-\nu}, \quad \mathbf{d}_{j-\nu-1} = \tilde{\mathbf{G}}_{j-\nu}^T \mathbf{c}_{j-\nu}, \quad (\nu = 0, \ldots\, , p-1),$$

such that

$$\begin{aligned} \mathbf{c}_{j-p} &= \tilde{\mathbf{H}}_{j-p+1}^T \ldots \tilde{\mathbf{H}}_j^T \mathbf{c}_j\,, \\ \mathbf{d}_{j-p} &= \tilde{\mathbf{G}}_{j-p+1}^T \tilde{\mathbf{H}}_{j-p+2}^T \ldots \tilde{\mathbf{H}}_j^T \mathbf{c}_j\,, \\ &\vdots \\ \mathbf{d}_{j-1} &= \tilde{\mathbf{G}}_j^T \mathbf{c}_j\,. \end{aligned} \tag{2.1}$$

Conversely, the p-*level wavelet reconstruction* of the block vector $(\mathbf{c}_{j-p}^T,\, \mathbf{d}_{j-p}^T,\, \ldots\, ,\, \mathbf{d}_{j-1}^T)^T$ with $\mathbf{c}_\nu,\, \mathbf{d}_\nu \in \mathbb{R}^{n_\nu}$ is given by the recursive computation of the vector $\mathbf{c}_j \in \mathbb{R}^{n_j}$ with

$$\mathbf{c}_\nu = \mathbf{H}_\nu\, \mathbf{c}_{\nu-1} + \mathbf{G}_\nu\, \mathbf{d}_{\nu-1} \quad (\nu = j-p+1, \ldots\, , j).$$

Finally, $\mathbf{c}_j \in \mathbb{R}^{n_j}$ is determined in the following form

$$\begin{aligned} \mathbf{c}_j \;=\; & \mathbf{H}_j \ldots \mathbf{H}_{j-p+1}\, \mathbf{c}_{j-p} + \mathbf{H}_j \ldots \mathbf{H}_{j-p+2} \mathbf{G}_{j-p+1}\, \mathbf{d}_{j-p} + \ldots \\ & + \mathbf{H}_j \mathbf{G}_{j-1}\, \mathbf{d}_{j-2} + \mathbf{G}_j\, \mathbf{d}_{j-1}\,. \end{aligned} \tag{2.2}$$

It can be simply observed that the p-level wavelet decomposition is an invertible endomorphism on $\mathbb{R}^{n_j}$ mapping $\mathbf{c}_j$ to $(\mathbf{c}_{j-p}^T,\, \mathbf{d}_{j-p}^T,\, \ldots\, ,\, \mathbf{d}_{j-1}^T)^T$, and the p-level wavelet reconstruction is the inverse mapping.

We present numerous examples of special biorthogonal low-pass filters in Table 1 with the convention that $h_k = 0$ and $\tilde{h}_k = 0$, respectively, for every $k \in \mathbb{Z}$ not occurring in Table 1. Our tests in Sections 4, 5, and 6 are based on these filters. Note that the

biorthogonal wavelets are classified by the number of vanishing moments. The CDF$(\tilde{m}, m)$ filters (see [3]) possess $\tilde{m}$ vanishing moments for decomposition and m vanishing moments for reconstruction. The binomial-m filters (see [9, 10]) possess m vanishing moments for both decomposition and reconstruction. The Barlaud filter with $m = 2$ can be found in [6], p. 281.
In the literature (see [6], pp. 259–285; [20], pp. 455–462; [12], pp. 271–272; [17]), one can find many other biorthogonal low-pass filters which generate biorthogonal wavelets with compact supports.

wavelet	k	-4	-3	-2	-1	0	1	2	3	4	5	6
CDF(3, 1)	$2\sqrt{2}\,h_k$					1	1					
	$8\sqrt{2}\,\tilde{h}_k$			-1	1	8	8	1	-1			
CDF(5, 1)	$\sqrt{2}\,h_k$					1	1					
	$128\sqrt{2}\,\tilde{h}_k$	3	-3	-22	22	128	128	22	-22	-3	3	
CDF(2, 2)	$2\sqrt{2}\,h_k$				1	2	1					
	$4\sqrt{2}\,\tilde{h}_k$			-1	2	6	2	-1				
CDF(4, 2)	$2\sqrt{2}\,h_k$				1	2	1					
	$64\sqrt{2}\,\tilde{h}_k$	3	-6	-16	38	90	38	-16	-6	3		
CDF(1, 3)	$4\sqrt{2}\,h_k$				1	3	3	1				
	$2\sqrt{2}\,\tilde{h}_k$				-1	3	3	-1				
CDF(3, 3)	$4\sqrt{2}\,h_k$				1	3	3	1				
	$32\sqrt{2}\,\tilde{h}_k$		3	-9	-7	45	45	-7	-9	3		
binomial-2	$2\sqrt{2}\,h_k$					3	2	-1				
	$2\sqrt{2}\,\tilde{h}_k$				1	2	1					
binomial-3	$2\sqrt{2}\,h_k$					-1	3	3	-1			
	$4\sqrt{2}\,\tilde{h}_k$					1	3	3	1			
binomial-4	$8\sqrt{2}\,h_k$						-5	20	10	-12	3	
	$8\sqrt{2}\,\tilde{h}_k$					1	4	6	4	1		
binomial-5	$8\sqrt{2}\,h_k$					3	-15	20	20	-15	3	
	$16\sqrt{2}\,\tilde{h}_k$					1	5	10	10	5	1	
binomial-6	$16\sqrt{2}\,h_k$					7	-42	77	28	-63	30	-5
	$32\sqrt{2}\,\tilde{h}_k$				1	6	15	20	15	6	1	
Barlaud	$10\sqrt{2}\,h_k$				-1	5	12	5	-1			
	$140\sqrt{2}\,\tilde{h}_k$			-3	-15	73	170	73	-15	-3		

Table 1: Low-pass filter coefficients for biorthogonal filters

3 Matrix-vector products in floating point arithmetic

We consider a binary floating point number system $\mathbb{F}$ which is characterized by the precision t and the exponent range $e_{\min} \le e \le e_{\max}$. A nonzero element of $\mathbb{F}$ can be expressed in the form

$$b = \pm 2^e \, (2^{-1} + b_2 \, 2^{-2} + \ldots + b_t \, 2^{-t})$$

with $b_i \in \{0, 1\}$ $(i = 2, \dots, t)$. Then b lies in the *range of* $\mathbb{F}$, i.e.

$$2^{e_{\min}-1} \le |b| \le 2^{e_{\max}}\,(1 - 2^{-t})\,.$$

Each $x \in \mathbb{R} \setminus \{0\}$ in the range of $\mathbb{F}$ can be approximated by a floating point number $\mathrm{fl}(x) \in \mathbb{F}$ with $|\mathrm{fl}(x) - x| \le u\,|x|$, where $u := 2^{-t}$ is the unit roundoff.
In order to carry out a rounding error analysis of a wavelet algorithm, we assume that the following *standard model of floating point arithmetic* by Wilkinson (see [21] or [8], pp. 40–45) is true:
For arbitrary floating point numbers x, $y \in \mathbb{F}$ and any basic arithmetical operation $\circ \in \{+, -, \times, /\}$, the exact value $x \circ y \in \mathbb{R}$ and the computed value $\mathrm{fl}(x \circ y) \in \mathbb{F}$ are related by

$$\mathrm{fl}(x \circ y) = (x \circ y)\,(1 + \epsilon) \qquad (|\epsilon| \le u). \tag{3.1}$$

It is usual to assume that (3.1) holds also for the square root operation, i.e., for all positive x in the range of $\mathbb{F}$ we suppose that

$$\mathrm{fl}(\sqrt{x}) = \sqrt{x}\,(1 + \epsilon) \qquad (|\epsilon| \le u).$$

In this model, we disregard underflow and overflow.

The above model is valid for most computers, in particular it holds for IEEE standard arithmetic. For the IEEE arithmetic of single precision, we have $t = 24$, $e_{\min} = -125$, $e_{\max} = 128$, and $u = 2^{-24} \approx 5.96 \times 10^{-8}$. For the IEEE arithmetic of double precision, we have $t = 53$, $e_{\min} = -1021$, $e_{\max} = 1024$, and $u = 2^{-53} \approx 1.11 \times 10^{-16}$ (see [8], p. 41).
We are especially interested in a rounding analysis for matrix-vector products. First, we consider inner products. With the unit roundoff u let now

$$\gamma_n := \tfrac{nu}{1-nu} \qquad (n \in \mathbb{N},\ nu < 1).$$

Further, for vectors $\mathbf{a} = (a_i)_{i=0}^{n-1} \in \mathbb{R}^n$ and matrices $\mathbf{A} = (a_{i,k})_{i,k=0}^{m-1,n-1} \in \mathbb{R}^{m\times n}$ let $|\mathbf{a}| := (|a_i|)_{i=0}^{n-1}$ and $|\mathbf{A}| := (|a_{i,k}|)_{i,k=0}^{m-1,n-1}$ be the corresponding vectors and matrices of absolute values. Then we have

Lemma 3.1 *Let $n \in \mathbb{N}$ be given with $nu < 1$. Then for a recursive computation of the inner product of arbitrary vectors $\mathbf{a}$, $\mathbf{b} \in \mathbb{F}^n$, we have*

$$|\mathrm{fl}(\mathbf{a}^T\mathbf{b}) - \mathbf{a}^T\mathbf{b}| \le \gamma_n\,|\mathbf{a}|^T|\mathbf{b}| = (nu + \mathcal{O}(u^2))\,|\mathbf{a}|^T|\mathbf{b}|.$$

Proof. The proof follows immediately by induction over n (see e.g. [8], p. 68–69). □

If the vector $\mathbf{a} \in \mathbb{F}^n$ possesses at most $l \le n$ nonzero entries, then we obtain as a trivial consequence of Lemma 3.1 that for arbitrary $\mathbf{b} \in \mathbb{F}^n$

$$|\mathrm{fl}(\mathbf{a}^T\mathbf{b}) - \mathbf{a}^T\mathbf{b}| \le \gamma_l\,|\mathbf{a}|^T|\mathbf{b}|.$$

Now we consider matrix-vector products. For a matrix $\mathbf{A} = (a_{i,k})_{i,k=0}^{m-1,n-1} \in \mathbb{R}^{m\times n}$ let $\operatorname{sign}\mathbf{A} := (\operatorname{sign} a_{i,k})_{i,k=0}^{m-1,n-1}$ be the corresponding sign matrix, where for $a \in \mathbb{R}$, $\operatorname{sign} a := a/|a|$ for $a \ne 0$ and $\operatorname{sign} 0 := 0$. Further, for two vectors $\mathbf{a} = (a_i)_{i=0}^{n-1}$, $\mathbf{b} = (b_i)_{i=0}^{n-1} \in \mathbb{R}^n$ we write $\mathbf{a} \le \mathbf{b}$, if $a_i \le b_i$ for all $i = 0, \dots, n-1$. Analogously, we write $\mathbf{A} \le \mathbf{B}$ for two matrices $\mathbf{A}$, $\mathbf{B}$ of the same size, if this inequality is true for each element. Then we obtain (see also [14])

Lemma 3.2 *Let m, n, $l \in \mathbb{N}$ with $2 \le l < n$ and $lu < 1$ be given. Let $\mathbf{A} = (a_{i,k})_{i,k=0}^{m-1,n-1} \in \mathbb{R}^{m \times n}$ be a matrix containing at most l nonzero entries in each row. Further let each $a_{i,k}$ lie in the range of $\mathbb{F}$. Assume that the nonzero entries $a_{i,k}$ are precomputed by $\hat{a}_{i,k} \in \mathbb{F}$, where*

$$|\hat{a}_{jk} - a_{jk}| \le \eta\, u \tag{3.2}$$

with some constant $\eta > 0$, and set $\hat{a}_{i,k} := 0$ if $a_{i,k} = 0$. Let $\hat{\mathbf{A}} := (\hat{a}_{i,k})_{i,k=0}^{m-1,n-1}$.
Then for arbitrary $\mathbf{x} \in \mathbb{F}^n$, the error $\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \mathbf{A}\mathbf{x}$ satisfies the elementwise estimate

$$|\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \mathbf{A}\mathbf{x}| \le \gamma_l\, |\mathbf{A}||\mathbf{x}| + (\eta\, u + \gamma_l\, \eta\, u)\, |\mathrm{sign}\, \mathbf{A}|\, |\mathbf{x}|$$

in the case of recursive summation.

Proof. The assumption (3.2) implies that

$$|\hat{\mathbf{A}} - \mathbf{A}| \le \eta\, u\, |\mathrm{sign}\, \mathbf{A}|.$$

Hence the error vector $\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \mathbf{A}\mathbf{x}$ can be estimated as follows

$$\begin{aligned} |\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \mathbf{A}\mathbf{x}| &\le |\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \hat{\mathbf{A}}\mathbf{x}| + |(\hat{\mathbf{A}} - \mathbf{A})\mathbf{x}| \\ &\le |\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \hat{\mathbf{A}}\mathbf{x}| + \eta\, u\, |\,\mathrm{sign}\, \mathbf{A}|\, |\mathbf{x}|. \end{aligned}$$

For the first term we obtain by Lemma 3.1

$$\begin{aligned} |\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \hat{\mathbf{A}}\mathbf{x}| &\le \gamma_l\, |\hat{\mathbf{A}}|\, |\mathbf{x}| \\ &\le \gamma_l\, |\mathbf{A}|\, |\mathbf{x}| + \gamma_l\, |\hat{\mathbf{A}} - \mathbf{A}|\, |\mathbf{x}| \\ &\le \gamma_l\, |\mathbf{A}|\, |\mathbf{x}| + \gamma_l\, \eta\, u\, |\mathrm{sign}\, \mathbf{A}|\, |\mathbf{x}|, \end{aligned}$$

where we have used that each row contains at most l nonzero entries. □

Using the spectral norm of the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, given by

$$\|\mathbf{A}\|_2 := \sqrt{\rho(\mathbf{A}^T \mathbf{A})},$$

where $\rho(\mathbf{A}^T\mathbf{A})$ denotes the spectral radius of $\mathbf{A}^T\mathbf{A}$, we finally obtain an error estimate in the Euclidian norm

$$\|\mathrm{fl}(\hat{\mathbf{A}}\mathbf{x}) - \mathbf{A}\mathbf{x}\|_2 \le \gamma_l\, \|\, |\mathbf{A}|\, \|_2\, \|\mathbf{x}\|_2 + (1 + \gamma_l)\, \eta\, u\, \|\, |\mathrm{sign}\, \mathbf{A}|\, \|_2\, \|\mathbf{x}\|_2.$$

4 Numerical stability of wavelet decomposition

First we introduce the vectors $\mathbf{h}_j := (h_{j,k})_{k=0}^{n_j-1}$, $\tilde{\mathbf{h}}_j := (\tilde{h}_{j,k})_{k=0}^{n_j-1}$, $\mathbf{g}_j := (g_{j,k})_{k=0}^{n_j-1}$, and $\tilde{\mathbf{g}}_j := (\tilde{g}_{j,k})_{k=0}^{n_j-1}$. The *forward shift matrix* $\mathbf{S}_j$ of order n_j is defined for arbitrary $\mathbf{x} = (x_k)_{k=0}^{n_j-1} \in \mathbb{R}^{n_j}$ by

$$\mathbf{S}_j\, \mathbf{x} := (x_{n_j-1},\, x_0,\, x_1,\, \ldots,\, x_{n_j-2})^T.$$

Then the matrices $\mathbf{H}_j$, $\tilde{\mathbf{H}}_j$, $\mathbf{G}_j$, and $\tilde{\mathbf{G}}_j$ in Section 2 can be expressed by shifts of the first columns:

$$\begin{aligned}
\mathbf{H}_j &:= \big(\mathbf{h}_j,\, \mathbf{S}_j^2\mathbf{h}_j,\, \mathbf{S}_j^4\mathbf{h}_j,\, \ldots,\, \mathbf{S}_j^{n_j-2}\mathbf{h}_j\big),\\
\tilde{\mathbf{H}}_j &:= \big(\tilde{\mathbf{h}}_j,\, \mathbf{S}_j^2\tilde{\mathbf{h}}_j,\, \mathbf{S}_j^4\tilde{\mathbf{h}}_j,\, \ldots,\, \mathbf{S}_j^{n_j-2}\tilde{\mathbf{h}}_j\big),\\
\mathbf{G}_j &:= \big(\mathbf{g}_j,\, \mathbf{S}_j^2\mathbf{g}_j,\, \mathbf{S}_j^4\mathbf{g}_j,\, \ldots,\, \mathbf{S}_j^{n_j-2}\mathbf{g}_j\big),\\
\tilde{\mathbf{G}}_j &:= \big(\tilde{\mathbf{g}}_j,\, \mathbf{S}_j^2\tilde{\mathbf{g}}_j,\, \mathbf{S}_j^4\tilde{\mathbf{g}}_j,\, \ldots,\, \mathbf{S}_j^{n_j-2}\tilde{\mathbf{g}}_j\big),
\end{aligned}$$

such that

$$\mathbf{H}_j^T\mathbf{H}_j = \tilde{\mathbf{G}}_j^T\tilde{\mathbf{G}}_j = \operatorname{circ}\mathbf{u}_{j-1}, \quad \tilde{\mathbf{H}}_j^T\tilde{\mathbf{H}}_j = \mathbf{G}_j^T\mathbf{G}_j = \operatorname{circ}\tilde{\mathbf{u}}_{j-1} \tag{4.1}$$

are circulant matrices of order n_{j-1} with the characteristic vectors

$$\begin{aligned}
\mathbf{u}_{j-1} &:= \big((\mathbf{h}_j)^T\mathbf{h}_j,\, (\mathbf{S}_j^2\mathbf{h}_j)^T\mathbf{h}_j,\, \ldots,\, (\mathbf{S}_j^{n_j-2}\mathbf{h}_j)^T\mathbf{h}_j\big)^T,\\
\tilde{\mathbf{u}}_{j-1} &:= \big((\tilde{\mathbf{h}}_j)^T\tilde{\mathbf{h}}_j,\, (\mathbf{S}_j^2\tilde{\mathbf{h}}_j)^T\tilde{\mathbf{h}}_j,\, \ldots,\, (\mathbf{S}_j^{n_j-2}\tilde{\mathbf{h}}_j)^T\tilde{\mathbf{h}}_j\big)^T.
\end{aligned}$$

Lemma 4.1 *If a filter $h = (h_k)_{k=-\infty}^{\infty}$ has a finite length l_h and if $\operatorname{supp} h = \{a,\, a+1,\, \ldots,\, a+l_h-1\}$ $(a \in \mathbb{Z})$, then for all $j \in \mathbb{N}$ with $n_j \geq l_h$, we have*

$$\|\mathbf{H}_j^T\mathbf{H}_j\|_1 = \mu_h^2, \quad \|\mathbf{H}_j\|_2 \leq \mu_h$$

with the constant

$$\begin{aligned}
\mu_h^2 &:= \sum_{m=0}^{n_{j-1}-1} \big|\sum_{k=0}^{n_j-1} h_{j,k}\, h_{j,k-2m}\big|\\
&= \sum_{k=a}^{a+l_h-1} h_k^2 \;+\; 2 \sum_{m=1}^{\lfloor (l_h-1)/2\rfloor} \big|\sum_{k=a+2m}^{a+l_h-1} h_k\, h_{k-2m}\big|.
\end{aligned}$$

If $h_k \geq 0$ for all $k \in \mathbb{Z}$, then $\|\mathbf{H}_j\|_2 = \mu_h$.

Proof. (i) From (4.1) it follows immediately that

$$\begin{aligned}
\|\mathbf{H}_j^T\mathbf{H}_j\|_1 &= \|\operatorname{circ}\mathbf{u}_{j-1}\|_1 = \|\mathbf{u}_{j-1}\|_1\\
&= \sum_{m=0}^{n_{j-1}-1} \big|\sum_{k=0}^{n_j-1} h_{j,k}\, h_{j,k-2m}\big| = \mu_h^2.
\end{aligned}$$

Without loss of generality we can assume that $a = 0$. Further, $n_j \geq l_h$. Then we have $h_{j,k} = h_k \neq 0$ for $k \in \{0,\, \ldots,\, l_h-1\}$ and $h_{j,k} = 0$ for $k \in \{l_h,\, \ldots,\, n_j-1\}$. Note that $(h_{j,k})_{k=-\infty}^{\infty}$ is n_j-periodic. Hence we obtain that

$$\begin{aligned}
\sum_{m=0}^{n_{j-1}-1} \big|\sum_{k=0}^{n_j-1} h_{j,k}\, h_{j,k-2m}\big| &= \sum_{m=0}^{n_{j-1}-1} \big|\sum_{k=0}^{l_h-1} h_k\, h_{j,k-2m}\big|\\
&= \sum_{k=0}^{l_h-1} h_k^2 \;+\; \sum_{m=1}^{n_{j-1}-1} \big|\sum_{k=0}^{l_h-1} h_k\, h_{j,k-2m}\big|\\
&= \sum_{k=0}^{l_h-1} h_k^2 \;+\; 2 \sum_{m=1}^{\lfloor (l_h-1)/2\rfloor} \big|\sum_{k=2m}^{l_h-1} h_k\, h_{k-2m}\big|.
\end{aligned}$$

(ii) Let $\mathbf{F}_{j-1} := (w^{mn})_{m,n=0}^{n_{j-1}-1}$ be the Fourier matrix of order n_{j-1}, where $w := \exp(\frac{-2\pi \mathrm{i}}{n_{j-1}})$. Since the circulant matrix $\mathbf{H}_j^T \mathbf{H}_j = \mathrm{circ}\, \mathbf{u}_{j-1}$ can be diagonalized by $\mathbf{F}_{j-1}$ [7], i.e.

$$\mathbf{F}_{j-1}(\mathbf{H}_j^T \mathbf{H}_j)\mathbf{F}_{j-1}^{-1} = \mathrm{diag}\,(\mathbf{F}_{j-1}\mathbf{u}_{j-1}),$$

we see that the eigenvalues of $\mathbf{H}_j^T \mathbf{H}_j$ are the components of the vector $\mathbf{F}_{j-1}\mathbf{u}_{j-1}$. Hence we obtain that

$$\begin{aligned} \|\mathbf{H}_j^T \mathbf{H}_j\|_2 &= \|\mathbf{H}_j\|_2^2 = \max\{|(\mathbf{F}_{j-1}\mathbf{u}_{j-1})_k| \,:\, k = 0, \ldots, n_{j-1}-1\} \\ &= \|\mathbf{F}_{j-1}\mathbf{u}_{j-1}\|_\infty \le \|\mathbf{u}_{j-1}\|_1 = \mu_h^2 \end{aligned}$$

such that $\|\mathbf{H}_j\|_2 \le \mu_h$. If all filter coefficients h_k are non-negative, then $\mathbf{u}_{j-1} \ge \mathbf{o}_{j-1}$ such that $\|\mathbf{F}_{j-1}\mathbf{u}_{j-1}\|_\infty = \|\mathbf{u}_{j-1}\|_1$ and $\|\mathbf{H}_j\|_2 = \mu_h$. □

Corollary 4.2 *For $l_h \le n_{j-1}$ we have*

$$\|\,|\mathrm{sign}\,\mathbf{H}_j|\,\|_2 \le \begin{cases} \frac{1}{2}\sqrt{2}\, l_h & \text{if } l_h \text{ even}, \\ \frac{1}{2}\sqrt{2}\,\sqrt{l_h^2+1} & \text{if } l_h \text{ odd}. \end{cases}$$

Proof. (i) For even l_h, we obtain

$$(l_h,\, l_h-2,\, l_h-4,\, \ldots,\, 2,\, 0,\, \ldots,\, 0,\, 2,\, \ldots,\, l_h-4,\, l_h-2)^T \in \mathbb{R}^{n_{j-1}}$$

as characteristic vector of the circulant matrix $|\mathrm{sign}\,\mathbf{H}_j|^T|\mathrm{sign}\,\mathbf{H}_j|$. By Lemma 4.1 we see that

$$\|\,|\mathrm{sign}\,\mathbf{H}_j|\,\|_2^2 = 2\left(2+4+\ldots+(l_h-2)\right) + l_h = \tfrac{1}{2}\, l_h^2.$$

(ii) For odd l_h, the characteristic vector of the circulant matrix $|\mathrm{sign}\,\mathbf{H}_j|^T|\mathrm{sign}\,\mathbf{H}_j|$ reads as follows

$$(l_h,\, l_h-2,\, l_h-4,\, \ldots,\, 1,\, 0,\, \ldots,\, 0,\, 1,\, \ldots,\, l_h-4,\, l_h-2)^T \in \mathbb{R}^{n_{j-1}}.$$

Then by Lemma 4.1 it follows that

$$\|\,|\mathrm{sign}\,\mathbf{H}_j|\,\|_2^2 = 2\left(1+3+\ldots+(l_h-2)\right) + l_h = \tfrac{1}{2}\,(l_h^2+1).$$

This completes the proof. □

Now we consider biorthogonal low-pass filters h and $\tilde{h}$ with finite lengths. The corresponding high-pass filters are denoted by g and $\tilde{g}$. Let $l = \max\{l_h, l_g\} \le n_j$. By definition of μ_h in Lemma 4.1 we see immediately that $\mu_{\tilde{g}} = \mu_h$ and $\mu_g = \mu_{\tilde{h}}$. Further we obtain

$$\begin{aligned} \|\mathbf{H}_j\|_2 &= \|\tilde{\mathbf{G}}_j\|_2\,, \qquad \|\mathbf{G}_j\|_2 = \|\tilde{\mathbf{H}}_j\|_2\,, \\ \|\,|\mathbf{H}_j|\,\|_2 &= \|\,|\tilde{\mathbf{G}}_j|\,\|_2\,, \qquad \|\,|\mathbf{G}_j|\,\|_2 = \|\,|\tilde{\mathbf{H}}_j|\,\|_2\,, \\ \|\,|\mathrm{sign}\,\mathbf{H}_j|\,\|_2 &= \|\,|\mathrm{sign}\,\tilde{\mathbf{G}}_j|\,\|_2\,, \qquad \|\,|\mathrm{sign}\,\mathbf{G}_j|\,\|_2 = \|\,|\mathrm{sign}\,\tilde{\mathbf{H}}_j|\,\|_2\,. \end{aligned}$$

Thus we conclude that

$$\|\mathbf{H}_j\|_2 = \|\tilde{\mathbf{G}}_j\|_2 \le \mu_h\,, \qquad \|\mathbf{G}_j\|_2 = \|\tilde{\mathbf{H}}_j\|_2 \le \mu_g. \tag{4.2}$$

We introduce the following notations

$$\begin{aligned} \mu_{|h|} &:= \|\,|\mathbf{H}_j|\,\|_2, \qquad \mu_{|\mathrm{sign}\ h|} := \|\,|\mathrm{sign}\,\mathbf{H}_j|\,\|_2, \\ \mu_{|g|} &:= \|\,|\mathbf{G}_j|\,\|_2, \qquad \mu_{|\mathrm{sign}\ g|} := \|\,|\mathrm{sign}\,\mathbf{G}_j|\,\|_2. \end{aligned}$$

In Table 2 we present the corresponding constants for the special biorthogonal filters of Table 1.

wavelet	μ_h	μ_g	$\mu_{\lvert h\rvert}$	$\mu_{\lvert g\rvert}$	$\mu_{\lvert\text{sign } h\rvert}$	$\mu_{\lvert\text{sign } g\rvert}$
CDF(3,1)	1	$\sqrt{17}/4$	1	5/4	$\sqrt{2}$	$3\sqrt{2}$
CDF(5,1)	1	$\sqrt{4721}/64$	1	89/64	$\sqrt{2}$	$5\sqrt{2}$
CDF(2,2)	1	$\sqrt{2}$	1	$\sqrt{10}/2$	2	$\sqrt{13}$
CDF(4,2)	1	$\sqrt{2}$	1	$\sqrt{754}/16$	$\sqrt{5}$	$\sqrt{41}$
CDF(1,3)	1	2	1	2	$2\sqrt{2}$	$2\sqrt{2}$
CDF(3,3)	1	2	1	2	$2\sqrt{2}$	$4\sqrt{2}$
binomial-2	$\sqrt{10}/2$	1	$\sqrt{10}/2$	1	2	2
binomial-3	2	1	2	1	$2\sqrt{2}$	$2\sqrt{2}$
binomial-4	$\sqrt{614}/8$	1	$\sqrt{674}/8$	1	$\sqrt{13}$	$\sqrt{13}$
binomial-5	4	1	19/4	1	$3\sqrt{2}$	$3\sqrt{2}$
binomial-6	$\sqrt{3291}/8$	1	$\sqrt{4138}/8$	1	5	5
Barlaud	1	1459/1400	37/25	1972/1225	$\sqrt{13}$	5

Table 2: Constants for biorthogonal filters in Table 1

Let $j,\, p \in \mathbb{N}$ with $p < j$ be given. Assume that $l = \max\{l_h,\, l_g\} \le n_{j-p+1}$. In Section 2, the p-level wavelet decomposition of an input vector $\mathbf{c}_j \in \mathbb{F}^{n_j}$ has been given by

$$\mathbf{c}_{j-\nu} = \tilde{\mathbf{H}}_{j-\nu+1}^T \mathbf{c}_{j-\nu+1}, \quad \mathbf{d}_{j-\nu} = \tilde{\mathbf{G}}_{j-\nu+1}^T \mathbf{c}_{j-\nu+1} \quad (\nu = 1,\, \ldots\, ,\, p). \tag{4.3}$$

Thus the vector $\mathbf{c}_j$ is decomposed into the vector $\left(\mathbf{c}_{j-p}^T,\, \mathbf{d}_{j-p}^T,\, \ldots\, ,\, \mathbf{d}_{j-1}^T\right)^T \in \mathbb{R}^{n_j}$. The corresponding computed vectors have the form

$$\hat{\mathbf{c}}_{j-\nu} := \mathrm{fl}\big(\hat{\tilde{\mathbf{H}}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\big), \quad \hat{\mathbf{d}}_{j-\nu} = \mathrm{fl}\big(\hat{\tilde{\mathbf{G}}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\big) \quad (\nu = 1,\, \ldots\, ,\, p)$$

with $\hat{\mathbf{c}}_j = \mathbf{c}_j$. First we observe the following estimate of the forward error for 1-level wavelet decomposition and partial reconstruction.

Lemma 4.3 (i) *Let $\tilde{h} = (\tilde{h}_k)_{k=-\infty}^{\infty}$ be a filter of finite length $l_{\tilde{h}} \le n_j$. Assume that all the filter coefficients $\tilde{h}_k \neq 0$ are precomputed by $\hat{\tilde{h}}_k \in \mathbb{F}$, where*

$$|\hat{\tilde{h}}_k - h_k| < \eta\, u$$

with some constant $\eta > 0$, and set $\hat{\tilde{h}}_k := 0$ if $\tilde{h}_k = 0$. Let $\hat{\tilde{\mathbf{H}}}_j := (\hat{\tilde{h}}_{j,r-2k})_{r,k=0}^{n_j-1,n_{j-1}-1}$ be the precomputed matrix of $\tilde{\mathbf{H}}_j$. Then for arbitrary $\mathbf{c}_j \in \mathbb{F}^{n_j}$, the error $\hat{\mathbf{c}}_{j-1} - \mathbf{c}_{j-1}$ can be estimated by

$$\|\hat{\mathbf{c}}_{j-1} - \mathbf{c}_{j-1}\|_2 = \|\mathrm{fl}(\hat{\tilde{\mathbf{H}}}_j^T \mathbf{c}_j) - \tilde{\mathbf{H}}_j^T \mathbf{c}_j\|_2 \le \Big(l_{\tilde{h}}\, \mu_{|\tilde{h}|}\, u + \eta\, \mu_{|\text{sign }\ \tilde{h}|}\, u + \mathcal{O}(u^2)\Big)\, \|\mathbf{c}_j\|_2.$$

(ii) *Let $h = (h_k)_{k=-\infty}^{\infty}$ be a filter of finite length $l_h \le n_{j-1}$. Assume that all the filter coefficients $h_k \neq 0$ are precomputed by $\hat{h}_k \in \mathbb{F}$, where*

$$|\hat{h}_k - h_k| < \eta\, u$$

with some constant $\eta > 0$, and set $\hat{h}_k := 0$ if $h_k = 0$. Let $\hat{\mathbf{H}}_j := (\hat{h}_{j,r-2k})_{r,k=0}^{n_j-1,n_{j-1}-1}$ be the precomputed matrix of $\mathbf{H}_j$. Then for arbitrary $\mathbf{c}_{j-1} \in \mathbb{F}^{n_{j-1}}$, the error $\mathrm{fl}(\hat{\mathbf{H}}_j \mathbf{c}_{j-1}) - \mathbf{H}_j \mathbf{c}_{j-1}$ can be estimated by

$$\|\mathrm{fl}(\hat{\mathbf{H}}_j \mathbf{c}_{j-1}) - \mathbf{H}_j \mathbf{c}_{j-1}\|_2 \le \Big(\lceil l_h/2 \rceil\, \mu_{|h|}\, u + \eta\, \mu_{|\text{sign }\ h|}\, u + \mathcal{O}(u^2)\Big)\, \|\mathbf{c}_{j-1}\|_2.$$

Proof. This follows immediately from Lemma 3.2. □

Iterated application of Lemma 4.3 leads to the following estimates for the forward error of ν-level wavelet decomposition ($\nu = 1, \dots, p$).

Theorem 4.4 *Let $h = (h_k)_{k=-\infty}^{\infty}$ and $\tilde{h} = (\tilde{h}_k)_{k=-\infty}^{\infty}$ be biorthogonal low-pass filters with finite lengths. Assume that for some $\eta > 0$, the precomputed filter coefficients in $\mathbb{F}$ satisfy the conditions*

$$\begin{aligned} |\hat{h}_k - h_k| &\le \eta\, u \quad (k \in \operatorname{supp} h), \qquad \hat{h}_k = 0 \quad (k \in \mathbb{Z} \setminus \operatorname{supp} h), \\ |\hat{\tilde{h}}_k - \tilde{h}_k| &\le \eta\, u \quad (k \in \operatorname{supp} \tilde{h}), \qquad \hat{\tilde{h}}_k = 0 \quad (k \in \mathbb{Z} \setminus \operatorname{supp} \tilde{h}). \end{aligned}$$

Let $j,\, p \in \mathbb{N}$ with $p \le j$ and $l = \max\{l_h,\, l_g\} \le n_{j-p+1}$ be given. Then for $\nu = 1, \dots, p$ we have

$$\begin{aligned} \|\hat{\mathbf{c}}_{j-\nu} - \mathbf{c}_{j-\nu}\|_2 &\le \nu \left(l_g\, \mu_{|g|}\, u + \eta\, \mu_{|\mathrm{sign}\ g|}\, u + \mathcal{O}(u^2) \right) \mu_g^{\nu-1}\, \|\mathbf{c}_j\|_2\,, \\ \|\hat{\mathbf{d}}_{j-\nu} - \mathbf{d}_{j-\nu}\|_2 &\le \left(l_h\, \mu_{|h|}\, u + \eta\, \mu_{|\mathrm{sign}\ h|}\, u + \mathcal{O}(u^2) \right) \mu_g^{\nu-1}\, \|\mathbf{c}_j\|_2 \\ &\quad + (\nu - 1)\, \mu_h \left(l_g\, \mu_{|g|}\, u + \eta\, \mu_{|\mathrm{sign}\ g|}\, u + \mathcal{O}(u^2) \right) \mu_g^{\nu-2}\, \|\mathbf{c}_j\|_2\,. \end{aligned}$$

Proof. We know by (4.3) that for $\nu = 1, \dots, p$

$$\hat{\mathbf{c}}_{j-\nu} - \mathbf{c}_{j-\nu} = \left(\hat{\mathbf{c}}_{j-\nu} - \tilde{\mathbf{H}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1} \right) + \tilde{\mathbf{H}}_{j-\nu+1}^T (\hat{\mathbf{c}}_{j-\nu+1} - \mathbf{c}_{j-\nu+1}).$$

By Lemma 4.3 (i), we can estimate

$$\begin{aligned} \|\hat{\mathbf{c}}_{j-\nu} - \tilde{\mathbf{H}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\|_2 &\le \left(l_g\, \mu_{|g|}\, u + \eta\, \mu_{|\mathrm{sign}\ g|}\, u + \mathcal{O}(u^2) \right) \|\hat{\mathbf{c}}_{j-\nu+1}\|_2\,, \\ \|\hat{\mathbf{d}}_{j-\nu} - \tilde{\mathbf{G}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\|_2 &\le \left(l_h\, \mu_{|h|}\, u + \eta\, \mu_{|\mathrm{sign}\ h|}\, u + \mathcal{O}(u^2) \right) \|\hat{\mathbf{c}}_{j-\nu+1}\|_2\,. \end{aligned}$$

Further we obtain by Lemma 4.3 (i), (4.3), and (4.2) that for $\nu = 1, \dots, p-1$

$$\begin{aligned} \|\hat{\mathbf{c}}_{j-\nu}\|_2 &\le \|\tilde{\mathbf{H}}_{j-\nu+1}^T \mathbf{c}_{j-\nu+1}\|_2 + \|\hat{\mathbf{c}}_{j-\nu} - \mathbf{c}_{j-\nu}\|_2 \\ &\le \left(\|\tilde{\mathbf{H}}_{j-\nu+1}\|_2 + \mathcal{O}(u) \right) \|\mathbf{c}_{j-\nu+1}\|_2 \le \left(\mu_g + \mathcal{O}(u) \right) \|\mathbf{c}_{j-\nu+1}\|_2 \\ &\le \left(\mu_g^{\nu} + \mathcal{O}(u) \right) \|\mathbf{c}_j\|_2\,. \end{aligned}$$

Hence we get

$$\begin{aligned} \|\hat{\mathbf{c}}_{j-\nu} - \tilde{\mathbf{H}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\|_2 &\le \left(l_g\, \mu_{|g|}\, u + \eta\, \mu_{|\mathrm{sign}\ g|}\, u + \mathcal{O}(u^2) \right) \mu_g^{\nu-1}\, \|\mathbf{c}_j\|_2\,, \\ \|\hat{\mathbf{d}}_{j-\nu} - \tilde{\mathbf{G}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\|_2 &\le \left(l_h\, \mu_{|h|}\, u + \eta\, \mu_{|\mathrm{sign}\ h|}\, u + \mathcal{O}(u^2) \right) \mu_g^{\nu-1}\, \|\mathbf{c}_j\|_2 \end{aligned}$$

such that the results follow immediately by iterative application of this procedure. □

Now we discuss the backward error of the p-level wavelet decomposition. Therefore we introduce the *backward error vector* $\Delta^{(p)}\mathbf{c}_j$ which is defined by

$$\begin{aligned} \mathbf{c}_j + \Delta^{(p)}\mathbf{c}_j &= \mathbf{H}_j \dots \mathbf{H}_{j-p+1} \hat{\mathbf{c}}_{j-p} + \mathbf{H}_j \dots \mathbf{H}_{j-p+2} \mathbf{G}_{j-p+1} \hat{\mathbf{d}}_{j-p} + \dots \\ &\quad + \mathbf{H}_j \mathbf{H}_{j-1} \mathbf{G}_{j-2} \hat{\mathbf{d}}_{j-3} + \mathbf{H}_j \mathbf{G}_{j-1} \hat{\mathbf{d}}_{j-2} + \mathbf{G}_j \hat{\mathbf{d}}_{j-1}\,. \end{aligned}$$

Thus the backward error vector $\Delta^{(p)}\mathbf{c}_j$ of the p-level wavelet decomposition is explained by the p-level wavelet reconstruction of the numerically decomposed vectors. A p-level wavelet decomposition is called *numerically backward stable*, if there is a positive constant k_p with $k_p\,u \ll 1$ such that

$$\|\Delta^{(p)}\mathbf{c}_j\|_2 \le \big(k_p\,u + \mathcal{O}(u^2)\big)\,\|\mathbf{c}_j\|_2$$

for all input vectors $\mathbf{c}_j \in \mathbb{F}^{n_j}$. The constant k_p measures the numerical backward stability of the p-level wavelet decomposition.

Theorem 4.5 *Under the assumptions of Theorem* 4.4, *the p-level wavelet decomposition is numerically backward stable with the constant*

$$\begin{aligned} k_p &= p\,\mu_h^p\,\mu_g^{p-1}\,(l_g\,\mu_{|g|} + \eta\,\mu_{|\mathrm{sign}\ g|}) \\ &\quad + \sum_{k=0}^{p-1}\Big(k\,\mu_h^{k+1}\,\mu_g^k\,(l_g\,\mu_{|g|} + \eta\,\mu_{|\mathrm{sign}\ g|}) + \mu_h^k\,\mu_g^{k+1}\,(l_h\,\mu_{|h|} + \eta\,\mu_{|\mathrm{sign}\ g|})\Big). \end{aligned}$$

Proof. The p-level wavelet reconstruction of $\mathbf{c}_j$ in (2.2) and the definition of $\Delta^{(p)}\mathbf{c}_j$ yield

$$\begin{aligned} \Delta^{(p)}\mathbf{c}_j &= \mathbf{H}_j\ldots\mathbf{H}_{j-p+1}\,(\hat{\mathbf{c}}_{j-p} - \mathbf{c}_{j-p}) + \mathbf{H}_j\ldots\mathbf{H}_{j-p+2}\mathbf{G}_{j-p+1}\,(\hat{\mathbf{d}}_{j-p} - \mathbf{d}_{j-p}) + \ldots \\ &\quad + \mathbf{H}_j\mathbf{H}_{j-1}\mathbf{G}_{j-2}\,(\hat{\mathbf{d}}_{j-3} - \mathbf{d}_{j-3}) + \mathbf{H}_j\mathbf{G}_{j-1}\,(\hat{\mathbf{d}}_{j-2} - \mathbf{d}_{j-2}) + \mathbf{G}_j\,(\hat{\mathbf{d}}_{j-1} - \mathbf{d}_{j-1})\,. \end{aligned}$$

From

$$\begin{aligned} \|\Delta^{(p)}\mathbf{c}_j\|_2 &\le \mu_h^p\,\|\hat{\mathbf{c}}_{j-p} - \mathbf{c}_{j-p}\|_2 + \mu_h^{p-1}\,\mu_g\,\|\hat{\mathbf{d}}_{j-p} - \mathbf{d}_{j-p}\|_2 + \ldots \\ &\quad + \mu_h^2\,\mu_g\,\|\hat{\mathbf{d}}_{j-3} - \mathbf{d}_{j-3}\|_2 + \mu_h\,\mu_g\,\|\hat{\mathbf{d}}_{j-2} - \mathbf{d}_{j-2}\|_2 + \mu_g\,\|\hat{\mathbf{d}}_{j-1} - \mathbf{d}_{j-1}\|_2 \end{aligned}$$

and Theorem 4.4, the assertion follows immediately. □

Biorthogonal low-pass filters of CDF wavelets (see Table 1) possess the property

$$\|\mathbf{H}_j\|_2 = \|\tilde{\mathbf{G}}_j\|_2 = \mu_h = 1.$$

Binomial wavelets (see Table 1) satisfy the condition

$$\|\mathbf{G}_j\|_2 = \|\tilde{\mathbf{H}}_j\|_2 = \mu_g = 1.$$

In these cases, Theorem 4.5 can be simplified.

Corollary 4.6 *Under the assumptions of Theorem* 4.4, *we have in the case* $\mu_h = 1$ *and* $\mu_g > 1$ *that*

$$k_p = \frac{(2p-1)\mu_g^{p+1} - 3p\mu_g^p + p\mu_g^{p-1} + \mu_g}{(\mu_g-1)^2}\,(l_g\,\mu_{|g|} + \eta\,\mu_{|\mathrm{sign}\ g|}) + \frac{\mu_g^{p+1} - \mu_g}{\mu_g - 1}\,(l_h\,\mu_{|h|} + \eta\,\mu_{|\mathrm{sign}\ h|}).$$

If $\mu_g = 1$ *and* $\mu_h > 1$, *then*

$$k_p = \frac{(2p-1)\mu_h^{p+2} - 3p\mu_h^{p+1} + p\mu_h^p + \mu_h^2}{(\mu_h-1)^2}\,(l_g\,\mu_{|g|} + \eta\,\mu_{|\mathrm{sign}\ g|}) + \frac{\mu_h^p - 1}{\mu_h - 1}\,(l_h\,\mu_{|h|} + \eta\,\mu_{|\mathrm{sign}\ h|}).$$

Proof. In the case $\mu_h = 1$ and $\mu_g > 1$, Theorem 4.5 implies that

$$\begin{aligned} k_p &= p\,\mu_g^{p-1}\,(l_g\,\mu_{|g|} + \eta\,\mu_{|\text{sign }g|}) + \sum_{k=1}^{p-1} k\,\mu_g^k\,(l_g\,\mu_{|g|} + \eta\,\mu_{|\text{sign }g|}) \\ &\quad + \sum_{k=0}^{p-1} \mu_g^{k+1}\,(l_h\,\mu_{|h|} + \eta\,\mu_{|\text{sign }h|}). \end{aligned}$$

By

$$\sum_{k=0}^{p-1} \mu_g^{k+1} = \tfrac{\mu_g^{p+1}-\mu_g}{\mu_g-1}\,, \quad \sum_{k=1}^{p-1} k\,\mu_g^k = \tfrac{(p-1)\mu_g^{p+1}-p\mu_g^p+\mu_g}{(\mu_g-1)^2}$$

we obtain the above stability constant k_p. The case $\mu_g = 1$ and $\mu_h > 1$ can be handled quite similarly such that this proof is omitted. □

Example 4.7 For the biorthogonal low-pass filters of the CDF(1,3) wavelets (see Tables 1 and 2) we obtain the stability constant

$$k_p = 2^p\,(12\,p + 3\sqrt{2}\,\eta\,p - 8) + 8.$$

For the biorthogonal low-pass filters of the binomial-3 wavelets (see Tables 1 and 2) we find

$$k_p = 2^p\left(12\,p + (6p-4)\,\sqrt{2}\,\eta - 8\right) + 8 + 6\,\sqrt{2}\,\eta.$$

Note that the stability constants k_p depend on the level p, but they do not depend on the length n_j of the input vectors. Hence, these constants are also true for the wavelet decomposition with non-periodic biorthogonal wavelets.

Now we illustrate the results of Corollary 4.6 by numerical tests. We choose 50 random vectors $\mathbf{c}_{10} \in \mathbb{F}^{1024}$. Every component of these vectors is a random number being uniformly distributed in [0, 1]. For the determination of the backward error vector, we compute the p–level wavelet decomposition in IEEE arithmetic of single precision and the p–level wavelet reconstruction in IEEE arithmetic of double precision with $p \in \{1, \ldots, 8\}$. To illustrate the results in the Figures 1 and 2, we plot the relative roundoff error norms for different levels p. The solid line indicates the worst case error bound $k_p\,u$ from Corollary 4.6. Each "+" corresponds to the numerical error $\|\Delta^{(p)}\mathbf{c}_{10}\|_2/\|\mathbf{c}_{10}\|_2$ for one of the 50 simulations. Figures 1 and 2 show the numerical stability of the p–level wavelet decomposition with CDF(3,1) and binomial–6 wavelets, respectively. We see that the numerical stability for CDF(3,1) wavelets is much better than for binomial–6 wavelets. In [10], this behavior of binomial–6 wavelets is called "instable".
Observe that the obtained estimates for the worst case error exceed the indeed errors by about a factor 10. This can be seen as a rule of thumb. But the qualitative error is reflected suitably.

5 Numerical stability of wavelet reconstruction

Let h and $\tilde{h}$ be given biorthogonal low-pass filters with finite lengths. Further let g and $\tilde{g}$ be the corresponding high-pass filters. Let $j,\, p \in \mathbb{N}$ with $p < j$ and $l = \max\{l_h,\, l_g\} \le n_{j-p+1}$.

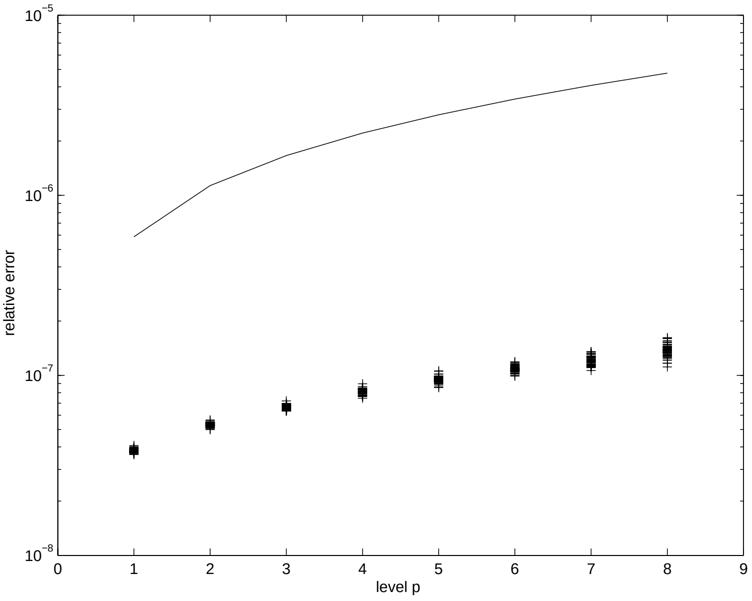

Figure 1: Relative backward error of the p-level decomposition with CDF(3,1) wavelets

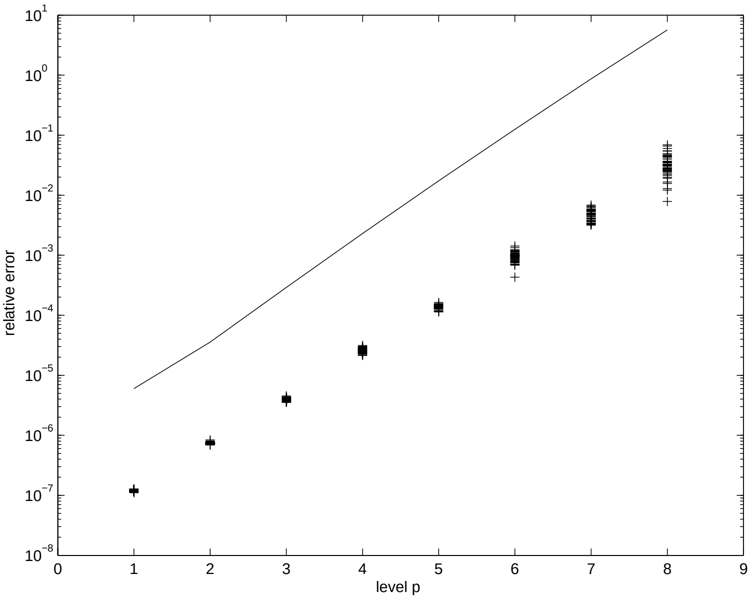

Figure 2: Relative backward error of the p-level decomposition with binomial-6 wavelets

Now we consider the numerical stability of the p-level wavelet reconstruction which reads as follows

$$\mathbf{c}_\nu = \mathbf{H}_k\, \mathbf{c}_{\nu-1} + \mathbf{G}_\nu\, \mathbf{d}_{\nu-1} \quad (\nu = j-p+1, \ldots, j). \tag{5.1}$$

Starting with the block vector

$$(\mathbf{c}_{j-p}^T, \mathbf{d}_{j-p}^T, \ldots, \mathbf{d}_{j-1}^T)^T \in \mathbb{F}^{n_j} \quad (\mathbf{c}_\nu, \mathbf{d}_\nu \in \mathbb{F}^{n_\nu}),$$

we reconstruct the vector $\mathbf{c}_j \in \mathbb{R}^{n_j}$. Assume that for some $\eta > 0$, the precomputed filter coefficients $\hat{h}_k$ and $\hat{g}_k$ satisfy the conditions

$$\begin{aligned} |\hat{h}_k - h_k| &\le \eta\, u \quad (k \in \operatorname{supp} h), \qquad \hat{h}_k = 0 \quad (k \in \mathbb{Z} \setminus \operatorname{supp} h)\,, \\ |\hat{\tilde{h}}_k - \tilde{h}_k| &\le \eta\, u \quad (k \in \operatorname{supp} \tilde{h}), \qquad \hat{\tilde{h}}_k = 0 \quad (k \in \mathbb{Z} \setminus \operatorname{supp} \tilde{h})\,. \end{aligned}$$

By $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{G}}_k$ we denote the matrices with precomputed entries corresponding to $\mathbf{H}_k$ and $\mathbf{G}_k$. Instead of $\mathbf{c}_j \in \mathbb{R}^{n_j}$, we compute the vector $\mathbf{c}_j^{(p)} \in \mathbb{F}^{n_j}$ by the following recursive procedure:

$$\begin{aligned} \mathbf{c}_{j-p+1}^{(1)} &:= \mathrm{fl}\Big(\mathrm{fl}(\hat{\mathbf{H}}_{j-p+1}\, \mathbf{c}_{j-p}) + \mathrm{fl}(\hat{\mathbf{G}}_{j-p+1}\, \mathbf{d}_{j-p})\Big)\,, \\ \mathbf{c}_{j-p+2}^{(2)} &:= \mathrm{fl}\Big(\mathrm{fl}(\hat{\mathbf{H}}_{j-p+2}\, \mathbf{c}_{j-p+1}^{(1)}) + \mathrm{fl}(\hat{\mathbf{G}}_{j-p+2}\, \mathbf{d}_{j-p+1})\Big)\,, \\ &\ \vdots \\ \mathbf{c}_{j}^{(p)} &:= \mathrm{fl}\Big(\mathrm{fl}(\hat{\mathbf{H}}_{j}\, \mathbf{c}_{j-1}^{(p-1)}) + \mathrm{fl}(\hat{\mathbf{G}}_{j}\, \mathbf{d}_{j-1})\Big)\,. \end{aligned}$$

First we consider the forward error of the reconstruction algorithm.

Theorem 5.1 *Let $h = (h_k)_{k=-\infty}^{\infty}$ and $\tilde{h} = (\tilde{h}_k)_{k=-\infty}^{\infty}$ be biorthogonal low-pass filters with finite lengths. Assume that for some $\eta > 0$, the precomputed filter coefficients in $\mathbb{F}$ satisfy the conditions*

$$\begin{aligned} |\hat{h}_k - h_k| &\le \eta\, u \quad (k \in \operatorname{supp} h), \qquad \hat{h}_k = 0 \quad (k \in \mathbb{Z} \setminus \operatorname{supp} h)\,, \\ |\hat{\tilde{h}}_k - \tilde{h}_k| &\le \eta\, u \quad (k \in \operatorname{supp} \tilde{h}), \qquad \hat{\tilde{h}}_k = 0 \quad (k \in \mathbb{Z} \setminus \operatorname{supp} \tilde{h})\,. \end{aligned}$$

Let $j,\, p \in \mathbb{N}$ with $p \le j$ and $l = \max\{l_h,\, l_g\} \le n_{j-p}$ be given.
Then the forward error of the p-level wavelet reconstruction can be estimated by

$$\begin{aligned} &\|\mathbf{c}_j^{(p)} - \mathbf{c}_j\|_2 \le (p\, \mu_h^{p-1}\, e_h\, u + \mathcal{O}(u^2))\, \|\mathbf{c}_{j-p}\|_2 \\ &+ \sum_{n=1}^{p} \Big((n-1)\, \mu_h^{n-2} \mu_g\, e_h\, u + \mu_h^{n-1}\, e_g\, u + \mathcal{O}(u^2)\Big)\, \|\mathbf{d}_{j-n}\|_2 \end{aligned} \tag{5.2}$$

with

$$e_h := \lceil l_h/2 \rceil\, \mu_{|h|} + \eta\, \mu_{|\mathrm{sign}\ h|} + \mu_h\,, \quad e_g := \lceil l_g/2 \rceil\, \mu_{|g|} + \eta\, \mu_{|\mathrm{sign}\ g|} + \mu_g\,.$$

Proof. We apply induction over p.
(i) Let $p = 1$. With

$$\mathbf{c}_j^{(1)} := \mathrm{fl}\Big(\mathrm{fl}(\hat{\mathbf{H}}_j \mathbf{c}_{j-1}) + \mathrm{fl}(\hat{\mathbf{G}}_j \mathbf{d}_{j-1})\Big)$$

we denote the computed vector of $\mathbf{c}_j$ of the 1-level wavelet reconstruction. Using (5.1) and triangle inequality, we can estimate

$$\begin{aligned} \|\mathbf{c}_j^{(1)} - \mathbf{c}_j\|_2 \le\ & \|\mathbf{c}_j^{(1)} - (\mathrm{fl}(\hat{\mathbf{H}}_j \mathbf{c}_{j-1}) + \mathrm{fl}(\hat{\mathbf{G}}_j \mathbf{d}_{j-1}))\|_2 \\ & + \|(\mathrm{fl}(\hat{\mathbf{H}}_j \mathbf{c}_{j-1}) - \mathbf{H}_j \mathbf{c}_{j-1}\|_2 + \|(\mathrm{fl}(\hat{\mathbf{G}}_j \mathbf{d}_{j-1}) - \mathbf{G}_j \mathbf{d}_{j-1}\|_2\,. \end{aligned} \tag{5.3}$$

From Lemma 4.3 (ii), it follows that

$$\begin{aligned}\|\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1}) - \mathbf{H}_j\mathbf{c}_{j-1}\|_2 &\le ((e_h - \mu_h)u + \mathcal{O}(u^2))\,\|\mathbf{c}_{j-1}\|_2\,, &(5.4)\\ \|\mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1}) - \mathbf{G}_j\mathbf{d}_{j-1}\|_2 &\le ((e_g - \mu_g)\,u + \mathcal{O}(u^2))\,\|\mathbf{d}_{j-1}\|_2\,. &(5.5)\end{aligned}$$

By assumption (3.1) of Wilkinson, we obtain the componentwise estimate

$$|\mathbf{c}_j^{(1)} - (\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1}) + \mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1}))| \le |\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1}) + \mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1})|\,u$$

and therefore

$$\|\mathbf{c}_j^{(1)} - (\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1}) + \mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1}))\|_2 \le \|\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1})\|_2\,u + \|\mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1})\|_2\,u\,.$$

Using (5.4), (5.5), and Lemma 4.1, we find

$$\begin{aligned}\|\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1})\|_2 &\le \|\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1}) - \mathbf{H}_j\mathbf{c}_{j-1}\|_2 + \|\mathbf{H}_j\mathbf{c}_{j-1}\|_2 \le (\mu_h + \mathcal{O}(u))\,\|\mathbf{c}_{j-1}\|_2\,,\\ \|\mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1})\|_2 &\le \|\mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1}) - \mathbf{G}_j\mathbf{d}_{j-1}\|_2 + \|\mathbf{G}_j\mathbf{d}_{j-1}\|_2 \le (\mu_g + \mathcal{O}(u))\,\|\mathbf{d}_{j-1}\|_2\,.\end{aligned}$$

Thus we see that

$$\|\mathbf{c}_j^{(1)} - (\mathrm{fl}(\hat{\mathbf{H}}_j\mathbf{c}_{j-1}) + \mathrm{fl}(\hat{\mathbf{G}}_j\mathbf{d}_{j-1}))\|_2 \le (\mu_h\,u + \mathcal{O}(u^2))\,\|\mathbf{c}_{j-1}\|_2 + (\mu_g\,u + \mathcal{O}(u^2))\,\|\mathbf{d}_{j-1}\|_2\,. \tag{5.6}$$

Finally, (5.3) – (5.6) yield the wanted estimate

$$\|\mathbf{c}_j^{(1)} - \mathbf{c}_j\|_2 \le (e_h\,u + \mathcal{O}(u^2))\,\|\mathbf{c}_{j-1}\|_2 + (e_g\,u + \mathcal{O}(u^2))\,\|\mathbf{d}_{j-1}\|_2\,. \tag{5.7}$$

(ii) Let $j,\,p \in \mathbb{N}$ with $p+1 \le j$ and $l \le n_{j-p-1}$ be given. Assume that the estimate (5.2) is true for p. The $(p+1)$-level wavelet reconstruction starts with the given block vector

$$(\mathbf{c}_{j-p-1}^T,\,\mathbf{d}_{j-p-1}^T,\,\ldots\,,\,\mathbf{d}_{j-1}^T)^T \in \mathbb{F}^{n_j}$$

and reads as follows

$$\mathbf{c}_\nu = \mathbf{H}_\nu\,\mathbf{c}_{\nu-1} + \mathbf{G}_\nu\,\mathbf{d}_{\nu-1} \quad (\nu = j-p,\ldots,j)$$

such that for $\nu = j-1$ we have

$$\mathbf{c}_{j-1} = \mathbf{H}_{j-1}\ldots\mathbf{H}_{j-p}\mathbf{c}_{j-p-1} + \mathbf{H}_{j-1}\ldots\mathbf{H}_{j-p+1}\mathbf{G}_{j-p}\mathbf{c}_{j-p-1} + \ldots + \mathbf{G}_{j-1}\mathbf{d}_{j-2}\,.$$

Hence by Lemma 4.1 it follows that

$$\|\mathbf{c}_{j-1}\|_2 \le \mu_h^p\,\|\mathbf{c}_{j-p-1}\|_2 + \sum_{n=2}^{p+1}\mu_h^{n-2}\,\mu_g\,\|\mathbf{d}_{j-n}\|_2\,. \tag{5.8}$$

By our assumption of induction we have

$$\begin{aligned}\|\mathbf{c}_{j-1}^{(p)} - \mathbf{c}_{j-1}\|_2 \le\; &(p\,\mu_h^{p-1}\,e_h\,u + \mathcal{O}(u^2))\,\|\mathbf{c}_{j-p-1}\|_2\\ &+ \sum_{n=1}^{p}\Big((n-1)\,\mu_h^{n-2}\mu_g\,e_h\,u + \mu_h^{n-1}\,e_g\,u + \mathcal{O}(u^2)\Big)\,\|\mathbf{d}_{j-n-1}\|_2\,. \quad (5.9)\end{aligned}$$

Introducing the auxiliary vector $\tilde{\mathbf{c}}_j := \mathbf{H}_j \mathbf{c}_{j-1}^{(p)} + \mathbf{G}_j \mathbf{d}_{j-1}$, we can estimate that

$$\|\mathbf{c}_j^{(p+1)} - \mathbf{c}_j\|_2 \leq \|\mathbf{c}_j^{(p+1)} - \tilde{\mathbf{c}}_j\|_2 + \|\tilde{\mathbf{c}}_j - \mathbf{c}_j\|_2 \,. \tag{5.10}$$

From (5.7) it follows immediately that

$$\|\mathbf{c}_j^{(p+1)} - \tilde{\mathbf{c}}_j\|_2 \leq (e_h\, u + \mathcal{O}(u^2))\, \|\mathbf{c}_{j-1}^{(p)}\|_2 + (e_g\, u + \mathcal{O}(u^2))\, \|\mathbf{d}_{j-1}\|_2 \tag{5.11}$$

where

$$\|\mathbf{c}_{j-1}^{(p)}\|_2 \leq \|\mathbf{c}_{j-1}^{(p)} - \mathbf{c}_{j-1}\|_2 + \|\mathbf{c}_{j-1}\|_2 \,. \tag{5.12}$$

By (5.1) and Lemma 4.1, we obtain that

$$\|\tilde{\mathbf{c}}_j - \mathbf{c}_j\|_2 = \|\mathbf{H}_j(\mathbf{c}_{j-1}^{(p)} - \mathbf{c}_{j-1})\|_2 \leq \mu_h\, \|\mathbf{c}_{j-1}^{(p)} - \mathbf{c}_{j-1}\|_2 \,. \tag{5.13}$$

Using (5.10) – (5.13), we conclude that

$$\begin{aligned} \|\mathbf{c}_j^{(p+1)} - \mathbf{c}_j\|_2 \quad \leq \quad & (e_h\, u + \mathcal{O}(u^2))\, (\|\mathbf{c}_{j-1}^{(p)} - \mathbf{c}_{j-1}\|_2 + \|\mathbf{c}_{j-1}\|_2) + (e_g\, u + \mathcal{O}(u^2))\, \|\mathbf{d}_{j-1}\|_2 \\ & + \mu_h\, \|\mathbf{c}_{j-1}^{(p)} - \mathbf{c}_{j-1}\|_2 \end{aligned}$$

such that by (5.8) and (5.9) we get the estimate (5.2) in the case $p+1$. □

With this result, we are able to describe the backward error for the p-level wavelet reconstruction. Let $\mathbf{c}_j^{(p)}$ be the numerically reconstructed vector of the p-level wavelet reconstruction of the input block vector

$$\mathbf{b} := (\mathbf{c}_{j-p}^T,\, \mathbf{d}_{j-p}^T,\, \ldots\, ,\, \mathbf{d}_{j-1}^T)^T \in \mathbb{F}^{n_j} \,. \tag{5.14}$$

The error vector $\Delta^{(p)}$ of the p-level wavelet reconstruction is explained by the exact p-level decomposition of the numerically reconstructed vector $\mathbf{c}_j^{(p)}$, i.e., we have the backward error block vector

$$\Delta^{(p)} := ((\Delta \mathbf{c}_{j-p})^T,\, (\Delta \mathbf{d}_{j-p})^T,\, \ldots\, ,\, (\Delta \mathbf{d}_{j-1})^T\,)^T$$

with

$$\begin{aligned} \mathbf{c}_{j-p} + \Delta \mathbf{c}_{j-p} \quad &= \quad \tilde{\mathbf{H}}_{j-p+1}^T \ldots \tilde{\mathbf{H}}_j^T \mathbf{c}_j^{(p)} \,, \\ \mathbf{d}_{j-p} + \Delta \mathbf{d}_{j-p} \quad &= \quad \tilde{\mathbf{G}}_{j-p+1}^T \tilde{\mathbf{H}}_{j-p+2}^T \ldots \tilde{\mathbf{H}}_j^T \mathbf{c}_j^{(p)} \,, \\ &\ \ \vdots \\ \mathbf{d}_{j-1} + \Delta \mathbf{d}_{j-1} \quad &= \quad \tilde{\mathbf{G}}_j^T \mathbf{c}_j^{(p)}. \end{aligned} \tag{5.15}$$

We say that a p-level wavelet reconstruction is *numerically backward stable*, if there exists a positive constant $\tilde{k}_p$ with $\tilde{k}_p\, u \ll 1$ such that

$$\|\Delta^{(p)}\|_2 \leq (\tilde{k}_p\, u + \mathcal{O}(u^2))\, \|\mathbf{b}\|_2$$

for all input block vectors (5.14). Hence the constant $\tilde{k}_p$ measures the numerical backward stability of the p-level wavelet reconstruction.

Theorem 5.2 *Under the assumptions of Theorem* 5.1, *the p-level wavelet reconstruction is numerically backward stable with the constant* $\tilde{k}_p$, *where*

$$\tilde{k}_p^2 = (\mu_g^{2p} + \mu_h^2 \sum_{m=0}^{p-1} \mu_g^{2m}) \Big[p^2\, \mu_h^{2p-2}\, e_h^2 \,+ \sum_{n=1}^{p} \Big((n-1)\, e_h\, \mu_h^{n-2}\, \mu_g + \mu_h^{n-1}\, e_g\Big)^2\Big]\,.$$

Proof. By (2.1) and (5.15) we get

$$\begin{aligned}
\Delta\mathbf{c}_{j-p} &= \tilde{\mathbf{H}}_{j-p+1}^T \ldots \tilde{\mathbf{H}}_j^T (\mathbf{c}_j^{(p)} - \mathbf{c}_j)\,,\\
\Delta\mathbf{d}_{j-p} &= \tilde{\mathbf{G}}_{j-p+1}^T \tilde{\mathbf{H}}_{j-p+2}^T \ldots \tilde{\mathbf{H}}_j^T (\mathbf{c}_j^{(p)} - \mathbf{c}_j)\,,\\
&\vdots\\
\Delta\mathbf{d}_{j-1} &= \tilde{\mathbf{G}}_j^T (\mathbf{c}_j^{(p)} - \mathbf{c}_j)
\end{aligned}$$

such that by (4.2) the backward error block vector $\Delta^{(p)}$ can be estimated as follows

$$\|\Delta^{(p)}\|_2^2 \le \Big(\mu_g^{2p} + \mu_h^2 \sum_{m=0}^{p-1} \mu_g^{2m}\Big)\, \|\mathbf{c}_j^{(p)} - \mathbf{c}_j\|_2^2\,.$$

Further by Theorem 5.1 and Cauchy-Schwarz inequality, we see that

$$\|\mathbf{c}_j^{(p)} - \mathbf{c}_j\|_2^2 \le \Big[p^2 \mu_h^{2p-2} e_h^2 + \sum_{n=1}^{p} \Big((n-1)\mu_h^{n-2} \mu_g e_h + \mu_h^{n-1} e_g\Big)^2 + \mathcal{O}(u)\Big]\, u^2\, \|\mathbf{b}\|_2^2\,,$$

where $\mathbf{b}$ is defined in (5.14). This completes the proof. □

Example 5.3 For the biorthogonal low-pass filters of the CDF(1,3) wavelets (see Tables 1 and 2), we obtain $e_h = 3 + 2\sqrt{2}\eta$ and $e_g = 6 + 2\sqrt{2}\eta$. Thus by Theorem 5.2 it follows that

$$\tilde{k}_p^2 = \tfrac{4^{p+1}-1}{9}\, [(36 + 48\sqrt{2}\eta + 32\eta^2)\, p^3 + (81 + 72\sqrt{2}\eta + 24\eta^2)\, p^2 + (18 - 12\sqrt{2}\eta - 8\eta^2)\, p]\,.$$

Now we illustrate the results of Theorem 5.2 by numerical tests. We choose 50 random block vectors $\mathbf{b} \in \mathbb{F}^{1024}$ of the form (5.14). Every component of these vectors is a random number being uniformly distributed in [0, 1]. For the determination of the backward error vector $\Delta^{(p)}$, we compute the p–level wavelet reconstruction in IEEE arithmetic of single precision and the p–level wavelet decomposition in IEEE arithmetic of double precision with $p \in \{1, \ldots, 8\}$. To illustrate the results in the Figures 3 and 4, we plot the relative roundoff error norms for different levels p. The solid line indicates the worst case error bound $\tilde{k}_p\, u$ found in Theorem 5.2. For each of the 50 simulations the computed relative error $\|\Delta^{(p)}\|_2/\|\mathbf{b}\|_2$ is indicated by "+". Figures 3 and 4 show the numerical stability behavior of the p–level wavelet reconstruction with CDF(3,1) and binomial–6 wavelets, respectively.

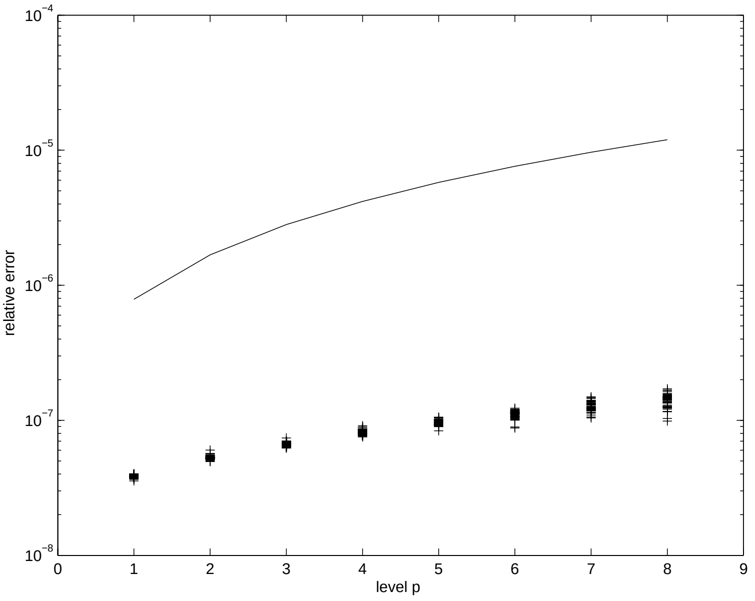

Figure 3: Relative backward error of the p-level reconstruction with CDF(3,1) wavelets

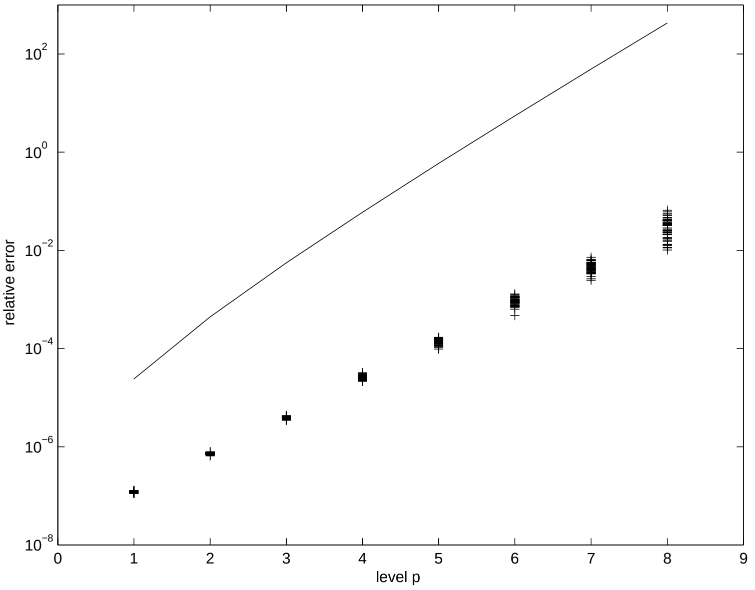

Figure 4: Relative backward error of the p-level reconstruction with binomial-6 wavelets

6 Wavelet decomposition-reconstruction

Finally we consider the worst case error of the p-level wavelet decomposition with enclosed reconstruction, called *p-level wavelet decomposition-reconstruction.* This case appears in many applications of the wavelet theory.

Let h and $\tilde{h}$ be given biorthogonal low-pass filters with finite lengths. Further let g and $\tilde{g}$ be the corresponding high-pass filters. Let $j,\, p \in \mathbb{N}$ with $p < j$ and $l = \max\{l_h,\, l_g\} \leq n_{j-p+1}$. Then by computation of the p-level wavelet decomposition of an arbitrary input vector

$\mathbf{c}_j \in \mathbb{F}^{n_j}$, we obtain the vectors

$$\hat{\mathbf{c}}_{j-\nu} := \mathrm{fl}\big(\hat{\tilde{\mathbf{H}}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\big), \quad \hat{\mathbf{d}}_{j-\nu} = \mathrm{fl}\big(\hat{\tilde{\mathbf{G}}}_{j-\nu+1}^T \hat{\mathbf{c}}_{j-\nu+1}\big) \quad (\nu = 1, \ldots, p)$$

with $\hat{\mathbf{c}}_j = \mathbf{c}_j$. Starting with the block vector

$$\big(\hat{\mathbf{c}}_{j-p}^T, \hat{\mathbf{d}}_{j-p}^T, \ldots, \hat{\mathbf{d}}_{j-1}^T\big)^T \in \mathbb{F}^{n_j}$$

we realize the p-level wavelet reconstruction. Then we obtain the resulting vector $\hat{\mathbf{c}}_j^{(p)} \in \mathbb{F}^{n_j}$ computed recursively by

$$\begin{aligned}
\hat{\mathbf{c}}_{j-p+1}^{(1)} &:= \mathrm{fl}\Big(\mathrm{fl}(\hat{\mathbf{H}}_{j-p+1}\,\hat{\mathbf{c}}_{j-p}) + \mathrm{fl}(\hat{\mathbf{G}}_{j-p+1}\,\hat{\mathbf{d}}_{j-p})\Big)\,, \\
\hat{\mathbf{c}}_{j-p+2}^{(2)} &:= \mathrm{fl}\Big(\mathrm{fl}(\hat{\mathbf{H}}_{j-p+2}\,\hat{\mathbf{c}}_{j-p+1}^{(1)}) + \mathrm{fl}(\hat{\mathbf{G}}_{j-p+2}\,\hat{\mathbf{d}}_{j-p+1})\Big)\,, \\
&\;\vdots \\
\hat{\mathbf{c}}_{j}^{(p)} &:= \mathrm{fl}\Big(\mathrm{fl}(\hat{\mathbf{H}}_{j}\,\hat{\mathbf{c}}_{j-1}^{(p-1)}) + \mathrm{fl}(\hat{\mathbf{G}}_{j}\,\hat{\mathbf{d}}_{j-1})\Big)\,.
\end{aligned}$$

By the perfect reconstruction property, the p-level wavelet decomposition-reconstruction coincides with the identity such that $\hat{\mathbf{c}}_j^{(p)}$ should be an approximation of $\mathbf{c}_j$. We say that a p-level wavelet decomposition with enclosed reconstruction is *numerically backward stable*, if there exists a positive constant $\hat{k}_p$ with $\hat{k}_p u \ll 1$ such that

$$\|\hat{\mathbf{c}}_j^{(p)} - \mathbf{c}_j\|_2 \le (\hat{k}_p\, u + \mathcal{O}(u))\, \|\mathbf{c}_j\|_2$$

for all input vectors $\mathbf{c}_j \in \mathbb{F}^{n_j}$. Hence the constant $\hat{k}_p$ measures the numerical backward stability of the p-level decomposition-reconstruction.

Theorem 6.1 *Under the assumptions of Theorem* 5.1*, the p-level wavelet decomposition-reconstruction is numerically stable with the constant*

$$\hat{k}_p = p\, e_h\, \mu_h^{p-1}\, \mu_g^p + \sum_{\nu=1}^{p} \Big((\nu-1)\, e_h\, \mu_h^{\nu-1}\, \mu_g^{\nu} + e_g\, \mu_h^{\nu}\, \mu_g^{\nu-1}\Big)\,.$$

Proof. Using Theorem 5.1, we can estimate that

$$\begin{aligned}
\|\hat{\mathbf{c}}_j^{(p)} - \mathbf{c}_j\|_2 \;\le\; & (p\, e_h\, \mu_h^{p-1}\, u + \mathcal{O}(u^2))\, \|\hat{\mathbf{c}}_{j-p}\|_2 \\
& + \sum_{\nu=1}^{p} \Big((\nu-1)\, \mu_h^{\nu-2}\, \mu_g\, e_h\, u + \mu_h^{\nu-1}\, e_g\, u + \mathcal{O}(u^2)\Big)\, \|\hat{\mathbf{d}}_{j-\nu}\|_2\,.
\end{aligned}$$

By Theorem 4.4, (4.2), and (2.1) we conclude that

$$\begin{aligned}
\|\hat{\mathbf{c}}_{j-p}\|_2 &\le \|\hat{\mathbf{c}}_{j-p} - \mathbf{c}_{j-p}\|_2 + \|\mathbf{c}_{j-p}\|_2 \le \|\hat{\mathbf{c}}_{j-p} - \mathbf{c}_{j-p}\|_2 + \mu_g^p\, \|\mathbf{c}_j\|_2 \\
&\le (\mu_g^p + \mathcal{O}(u))\, \|\mathbf{c}_j\|_2\,.
\end{aligned}$$

Analogously we see that for $\nu = 1, \ldots, p$

$$\|\hat{\mathbf{d}}_{j-\nu}\|_2 \le (\mu_h\, \mu_g^{\nu-1} + \mathcal{O}(u))\, \|\mathbf{c}_j\|_2\,.$$

This completes the proof. □

In the cases $\mu_h = 1$ and $\mu_g = 1$, respectively, Theorem 6.1 can be simplified:

Corollary 6.2 *Under the assumptions of Theorem* 5.1*, we have in the case* $\mu_h = 1$ *and* $\mu_g > 1$ *that*

$$\hat{k}_p = \frac{(2p-1)\mu_g^{p+2} - 3p\mu_g^{p+1} + p\mu_g^p + \mu_g^2}{(\mu_g-1)^2}\, e_h + \frac{\mu_g^p - 1}{\mu_g - 1}\, e_g\,.$$

If $\mu_g = 1$ *and* $\mu_h > 1$*, then*

$$\hat{k}_p = \frac{(2p-1)\mu_h^{p+1} - 3p\mu_h^p + p\mu_h^{p-1} + \mu_h}{(\mu_h-1)^2}\, e_h + \frac{\mu_h^{p+1} - \mu_h}{\mu_h - 1}\, e_g\,.$$

The proof follows immediately from Theorem 6.1 and is omitted here.

Example 6.3 For the biorthogonal low-pass filters of the CDF(1,3) wavelets (see Tables 1 and 2), we obtain $e_h = 3 + 2\sqrt{2}\eta$ and $e_g = 6 + 2\sqrt{2}\eta$. Thus by Corollary 6.2 it follows that

$$\begin{aligned} \hat{k}_p &= (3\,p\,2^p - 2^{p+2} + 4)\, e_h + (2^p - 1)\, e_g \\ &= 2^p\,(9\,p + 6\,p\,\eta\sqrt{2} - 6\,\eta\sqrt{2} - 6) + 6\,\eta\sqrt{2} + 6\,. \end{aligned}$$

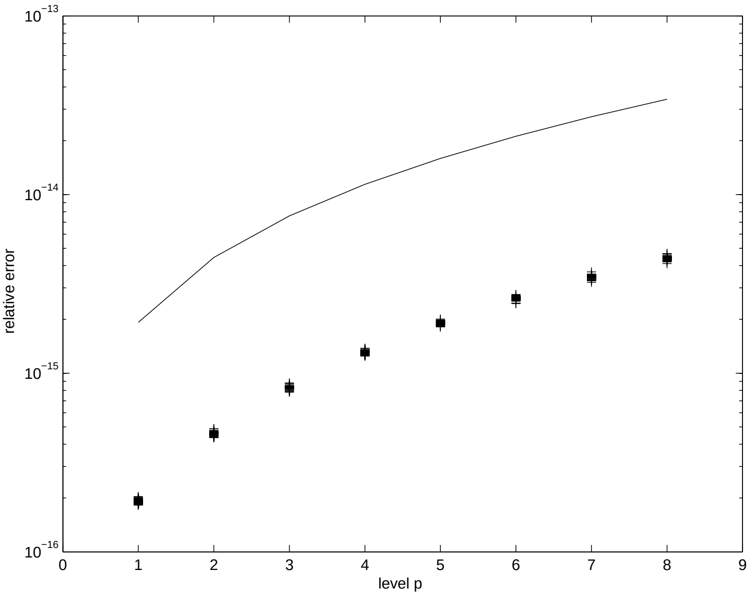

Figure 5: Relative backward error of the p-level decomposition-reconstruction with CDF(3,1) wavelets

Finally we illustrate the results of Corollary 6.2 by numerical tests. We choose 50 random vectors $\mathbf{c}_{10} \in \mathbb{F}^{1024}$. Every component of these vectors is uniformly distributed in [0, 1]. For the determination of the backward error vector, we compute the p–level wavelet decomposition-reconstruction in IEEE arithmetic of double precision with $p \in \{1, \ldots, 8\}$. To illustrate the results in the Figures 5 and 6, we plot the relative roundoff error norms for different levels p. The solid line indicates the worst case error bound $\hat{k}_p\, u$ found in Corollary 6.2. For each of the 50 simulations the computed relative error $\|\hat{\mathbf{c}}_{10}^{(p)} - \mathbf{c}_{10}\|_2 / \|\mathbf{c}_{10}\|_2$ is denoted by "+". Figures 5 and 6 show the numerical stability of the p–level decomposition-reconstruction with CDF(3,1) and binomial–6 wavelets, respectively. We see again that CDF(3,1) wavelets have a better numerical stability than binomial–6 wavelets.

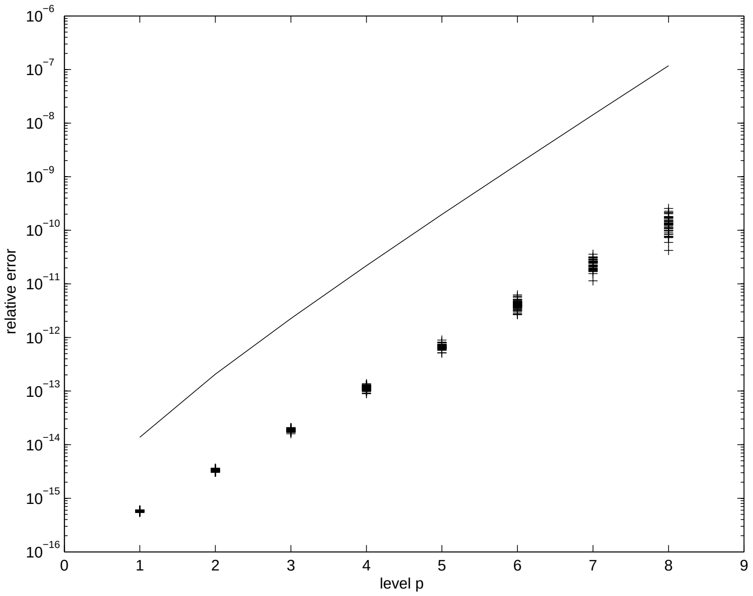

Figure 6: Relative backward error of the p-level decomposition-reconstruction with binomial-6 wavelets

References

[1] A. Cohen & I. Daubechies, A stability criterion for biorthogonal wavelet bases and their related subband coding scheme, Duke Math. J. **68** (1992), 313-335.

[2] A. Cohen & I. Daubechies, On the instability of arbitrary biorthogonal wavelet packets, SIAM J. Math. Anal. **24** (1993), 1340-1354.

[3] A. Cohen, I. Daubechies & J.C. Feauveau, Biorthogonal bases of compactly supported wavelets, Commun. Pure Appl. Math. **45** (1992), 485-560.

[4] A. Cohen, Biorthogonal wavelets, in: *Wavelets - A Tutorial in Theory and Applications*, (C.K. Chui, ed.), Academic Press, Boston, 1992, 123-152.

[5] A. Cohen, *Numerical Analysis of Wavelet Methods*, Elsevier, Amsterdam, 2003.

[6] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[7] P.J. Davis, *Circulant Matrices*, J. Wiley & Sons, New York, 1979.

[8] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[9] F. Keinert, Biorthogonal wavelets for fast matrix computations, Appl. Comput. Harmon. Anal. **1** (1994), 147-156.

[10] F. Keinert, Numerical stability of biorthogonal wavelet transforms, Adv. Comput. Math. **4** (1995), 1-26.

[11] F. Keinert, *Wavelets and Multiwavelets*, Chapman and Hall/CRC, Boca Raton, 2004.

[12] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1999.

[13] G. Plonka & M. Tasche, A unified approach to periodic wavelets, in: *Wavelets: Theory, Algorithms and Applications* (C. K. Chui, L. Montefusco & L. Puccio, eds.), Academic Press, San Diego, 1994, 137-151.

[14] G. Plonka & M. Tasche, Numerical stability of fast trigonometric and orthogonal wavelet transforms, in: *Advances in Constructive Approximation*, (M. Neamtu & E.B. Saff, eds.), Nashboro Press, Brentwood, 2004, 393-419.

[15] G. Plonka & M. Tasche, Fast and numerically stable algorithms for discrete cosine transforms, Linear Algebra Appl. **394** (2005), 309-345.

[16] H. Schumacher, Numerical stability of wavelet algorithms (in German), thesis, Univ. Rostock, 2003.

[17] W. Sweldens, The lifting scheme: a custom–design construction of biorthogonal wavelets, Appl. Comput. Harmon. Anal. **3** (1996), 186-200.

[18] R. Turcajová, Numerical condition of discrete wavelet transforms, SIAM J. Matrix Anal. Appl. **18** (1997), 981-999.

[19] M. Tasche & H. Zeuner, Worst and average case roundoff error analysis for FFT, BIT **41** (2001), 563-581.

[20] M.V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, AK Peters, Wellesley, 1993.

[21] J.H. Wilkinson, Error analysis of floating point computation, Numer. Math. **2** (1960), 319-343.