2008 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2008)
Swissôtel Le Concorde,Bangkok,Thailand

Finite Word Length Error Analysis based on Basic Formula of Rounding Operation

Masahiro IWAHASHI
Nagaoka University of Technology
Nagaoka-shi, Niigata, 940-2188 Japan

Hitoshi KIYA
Tokyo Metropolitan University
Hino-shi, Tokyo, 191-0065 Japan

***Abstract*- This report theoretically analyzes the lossless condition on word length such that errors due to rounding of signals and coefficients are nullified by the final rounding. Our analysis is based on mathematical expression of the rounding operation and its properties. According to our formulas, the minimum word length is determined for each input signal value. Results are more accurate than conventional L2 norm analysis.**

I. INTRODUCTION

Recently, digital signal processing (DSP) has wide variety of applications due to rapidly developing advanced LSI technologies. In a DSP circuit, all the signals and coefficients are expressed as binary values with finite word length (or bit depth). In general, longer word length requires larger complexity e.g. memory and logic gates, and therefore, it is important to reduce the word length [1,2]. However, truncation of the word length (rounding) generates errors and they propagate in the circuit resulting in degradation of quality of output signals [3,4].

So far, numerous attempts have been made on analysis on errors due to the rounding [3,4,7,9]. The lifting structure of the wavelet has a unique property that the rounding does not affect on the output signal [5,6]. Therefore the 5/3 wavelet in the JPEG-2000 is utilized for lossless coding [7]. In case of the 9/7 wavelet, although it has high coding gain, it is impossible to guarantee lossless when band signals are rounded. Therefore it is utilized for lossy coding [7].

In the previous analysis, the rounding error is treated as uncorrelated additive noise with uniformly distributed probability density function [3,4]. The analysis can approximately estimate the word length in which the error becomes negligible [3]. We have analyzed sensitivity of coefficients in the 9/7 wavelet considering input signal's auto-correlation [8]. However both of them based on the L_2 norm do not provide the lossless condition such that the rounding error in the output signal becomes zero. The worst case analysis based on the L_∞ norm evaluates the maximum value of the error [4]. It leads the lossless condition, however results are too strict.

Recently, we have experimentally derived the minimum word length of signals and coefficients which guarantees lossless for DC (constant valued) input signals [9]. In this report, we theoretical analyze the lossless condition for any input signals focusing on the scaling pair which generates rounding errors in the 9/7 wavelet transform.

In **2.**, we define the word length and the rounding operation. Errors and properties of the rounding operation are mathematically described as operators. In **3.**, we derive the addition formula and the multiplication formula of the rounding operation newly introduced in this report. In **4.**, we apply these formulas to derive the mapping invariant condition and the lossless scaling pair condition. In **5.**, we confirm that the method described in this report can derive the lossless condition on the word length for each of input values, more accurately than the conventional analysis. Conclusion remarks are summarized in **6.**.

II. WORD LENGTH AND ROUNDING OPERATION

A. *Word Length of a Digital Value*

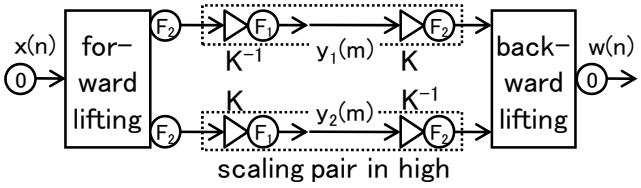
In this report, we deal with the fixed point implementation in which a digital value x is expressed by

$$x = \sum_{p=-F}^{I-1} b_p 2^p, \quad b_p \in \{0,1\} \qquad I \geq 1, F \geq 0, \quad I, F \in \mathbf{Z} \tag{1}$$

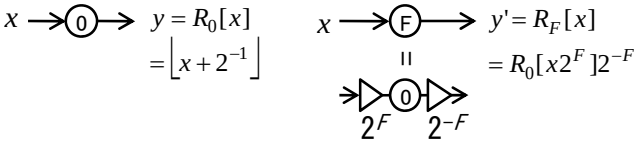
where x has I [bit] for integer part including sign and F [bit] for fraction part. Hereinafter, F is referred to as the word length. Its range is given by

$$x \in [-2^{I-1}, 2^{I-1} - 2^{-F}] \in [-2^{I-1}, 2^{I-1}) \ . \tag{2}$$

In general, the value x has more than F [bit] after multiplication with a coefficient. This report treats the case when the multiplied value is truncated into F [bit]. Similarly, a coefficient value of a multiplier is also expressed in the form of eq.(1). We treat the case when a coefficient value is truncated into W [bit]. Our purpose in this report is to analytically lead the lossless condition on F_1, F_2 and W of the scaling pair in the 9/7 wavelet illustrated in fig.1 (a) [7,9].



(a) Wavelet transform circuit.



(b) Definition of the rounding operation.

Fig.1 The minimum value of the word length F_1, F_2 of signals and W of coefficients are analyzed under the lossless condition.

978-1-4244-2565-5/08/$25.00 ©2008 IEEE

B. Notation of the Rounding Operation and its Errors

There are various types of rounding operations [10]. In this report, we deal with the rounding defined by

$$\begin{cases} R_0[x] = \lfloor x + 2^{-1} \rfloor = x' - (x' \mod 1) \\ x' = x + 2^{-1} \end{cases} \tag{3}$$

as an example. We denote the rounding error by

$$\begin{aligned} \Delta_0[x] &= x - R_0[x] \\ &= \{(x + 2^{-1}) \mod 1\} - 2^{-1} \end{aligned} . \tag{4}$$

These are expanded to a rational number with F [bit] by

$$\begin{cases} R_F[x] = R_0[x2^F]2^{-F} \\ \Delta_F[x] = x - F[x] = \Delta_0[x2^F]2^{-F} \end{cases} . \tag{5}$$

The operation $R_0[x]$ and $R_F[x]$ output an integer and a rational number, respectively as illustrated in fig.1 (b).

C. Basic Properties of the Rounding Operation

Basic properties of the rounding operation described by

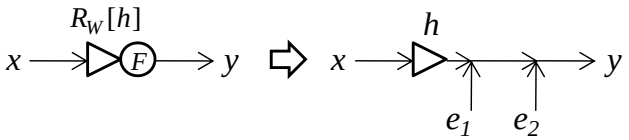
$$\begin{cases} R_0[x + y] = x + R_0[y] \\ \Delta_0[x + y] = \Delta_0[y] \end{cases} \quad x \in \mathbf{Z}, \quad y \in \mathbf{R} \tag{6}$$

can be derived from eq.(3) and (4). These equations lead the following important properties.

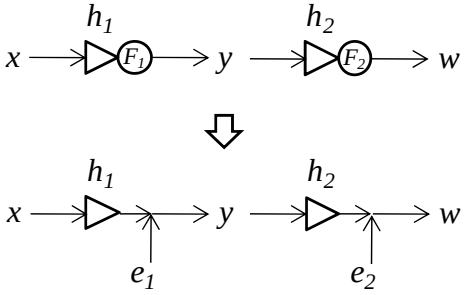
$$\begin{cases} R_0[x] = 0 \Leftrightarrow x \in [-2^{-1}, 2^{-1}) \\ R_F[x] = 0 \Leftrightarrow x \in [-2^{-1-F}, 2^{-1-F}) \end{cases} \tag{7}$$

$$\begin{cases} R_0[\Delta_0[x]] = 0 \Leftrightarrow \Delta_0[x] \in [-2^{-1}, 2^{-1}) \\ R_F[\Delta_F[x]] = 0 \Leftrightarrow \Delta_F[x] \in [-2^{-1-F}, 2^{-1-F}) \end{cases} \tag{8}$$

$$\begin{cases} \Delta_0[\Delta_0[x]] = \Delta_0[x] \\ \Delta_F[\Delta_F[x]] = \Delta_F[x] \end{cases} \tag{9}$$



(a) A scaling.



(b) A scaling pair.

Fig.2 A scaling and a scaling pair and their equivalent circuits utilized for analysis on the lossless condition.

III. Basic Formula of the Rounding Operation

The addition formula and the multiplication formula of the rounding operation are derived from the basic properties.

A. Addition formula

$$\begin{aligned} &R_F[x + y] = R_F[x] + R_F[y + \Delta_F[x]] \\ &x \in \mathbf{R}, y \in \mathbf{R} \end{aligned} \tag{10}$$

Proof:

$$\begin{aligned} &R_F[x + y] \\ &= R_F[R_F[x] + \Delta_F[x] + y] \\ &= R_0[R_0[x2^F] + \Delta_F[x]2^F + y2^F]2^{-F} \\ &= R_F[x] + R_F[\Delta_F[x] + y] \end{aligned}$$

Q.E.D.

It has following variations.

$$\begin{cases} R_F[x + y] = x + R_F[y] \\ R_F[x + y] = x + y - \Delta_F[y] \\ \Delta_F[x + y] = \Delta_F[y] \end{cases} \quad x2^F \in \mathbf{Z}, y \in \mathbf{R} \tag{11}$$

B. Multiplication Formula

$$\begin{aligned} &R_F[xy] \\ &= R_F[xR_F[y]] + R_F[x\Delta_F[y] + \Delta_F[xR_F[y]]] \\ &x \in \mathbf{R}, y \in \mathbf{R} \end{aligned} \tag{12}$$

Proof:

$$\begin{aligned} &R_F[xy] \\ &= R_F[x\Delta_F[y] + xR_F[y]] \\ &= R_F[x\Delta_F[y] + \Delta_F[xR_F[y]] + R_F[xR_F[y]]] \\ &= R_F[x\Delta_F[y] + \Delta_F[xR_F[y]]] + R_F[xR_F[y]] \end{aligned}$$

Q.E.D.

It has following variations.

$$\begin{cases} R_F[xy] = xR_F[y] + R_F[x\Delta_F[y]] \\ R_F[xy] = xy - \Delta_F[x\Delta_F[y]] \\ \Delta_F[xy] = \Delta_F[x\Delta_F[y]] \end{cases} \quad x2^F \in \mathbf{Z}, y \in \mathbf{R} \tag{13}$$

In the next section, we utilize the following formula to derive the lossless condition on word length of signals in the scaling pair in fig.2 (b).

$$\begin{aligned} &R_{F_2}[xR_{F_1}[y]] \\ &= R_{F_2}[xy] + R_{F_2}[-x\Delta_{F_1}[y] + \Delta_{F_2}[xy]] \end{aligned} \tag{14}$$

Proof:

$$\begin{aligned} &R_{F_2}[xR_{F_1}[y]] \\ &= R_0[-x\Delta_{F_1}[y]2^{F_2} + xy2^{F_2}]2^{-F_2} \\ &= R_0[-x\Delta_{F_1}[y]2^{F_2} + \Delta_0[xy2^{F_2}]]2^{-F_2} + R_{F_2}[xy] \\ &= R_{F_2}[-x\Delta_{F_1}[y] + \Delta_{F_2}[xy]] + R_{F_2}[xy] \end{aligned}$$

Q.E.D.

IV. Lossless Conditions on the Word Length

The lossless conditions of the scaling pairs in the 9/7 wavelet are derived utilizing the formulas described in III.

A. Mapping Invariant Condition

Fig.2 (a) illustrates a scaling in which both of a signal value and a coefficient value are rounded. An input value x is multiplied by a rounded value $R_W[h]$ of a given coefficient h. The result $R_W[h]x$ is rounded to $R_F[R_W[h]x]$. When it is the same as $R_F[hx]$, the mapping of x to y is invariant and effect of rounding of h is nullified. Therefore we define the mapping invariance by

$$E_m = 0 \quad for \quad \begin{cases} E_m = R_F\left[R_W[h]x\right] - R_F[hx] \\ R_W[h] = h - \Delta_W[h] \end{cases} . \tag{15}$$

From eq.(10) and (7), we lead the mapping invariant condition as

$$\left|\Delta_W[h]x - \Delta_F[hx]\right| < 2^{-1-F} . \tag{16}$$

Proof:

$$\begin{aligned} & R_F\left[R_W[h]x\right] - R_F[hx] \\ &= R_F\left[hx - \Delta_W[h]x\right] - R_F[hx] \\ &= R_F\left[\Delta_F[hx] - \Delta_W[h]x\right] + R_F[hx] - R_F[hx] \\ &= R_F\left[\Delta_F[hx] - \Delta_W[h]x\right] = 0 \\ \therefore \quad & \left|\Delta_F[hx] - \Delta_W[h]x\right| < 2^{-1-F} \end{aligned}$$

Q.E.D.

Since the upper bound of the errors and signal are given by

$$\begin{cases} \left|\Delta_W[h]\right| < 2^{-1-W} \\ \left|\Delta_F[hx]\right| < 2^{-1-F} \\ \left|x\right| < 2^{I-1} \end{cases} , \tag{17}$$

eq.(16) includes a condition:

$$W > F + I - 1 \tag{18}$$

as a special case. The mapping invariant condition gives the minimum word length W of a coefficient under a given word length F of signals and an input value x.

B. Lossless Scaling Pair Condition

Fig.2 (b) illustrates a scaling pair under the mapping invariant condition. We define its losslessness by

$$E_p = 0 \quad for \quad \begin{cases} E_p = R_{F_2}\left[h_2 R_{F_1}[h_1 x]\right] - x \\ h_1 h_2 = 1 \end{cases} . \tag{19}$$

From eq.(14), we lead the lossless scaling pair condition as

$$\left|h_2 \Delta_{F_1}[h_1 x]\right| < 2^{-1-F_2} . \tag{20}$$

This condition determines the word length F_1 and F_2 of signals for each input value x. Especially when the scaling pair is lossless for any x, eq.(20) leads

$$\left|h_2 \Delta_{F_1}[h_1 x]\right| < \left|h_2\right| 2^{-1-F_1} < 2^{-1-F_2} . \tag{21}$$

Therefore, it includes following condition as a special case.

$$F_1 > F_2 + \log_2\left|h_2\right| . \tag{22}$$

V. Simulation Results

The conditions on the word length for each input value are investigated comparing to the conventional analysis.

A. Word Length under the Mapping Invariant Condition

In fig.2(a), the output value y is described by

$$\begin{cases} y = hx + e_1 + e_2 \\ e_1 = -\Delta_W[h]x \\ e_2 = -\Delta_F\left[R_W[h]x\right] \end{cases} . \tag{23}$$

When the error e_1 is negligible to the error e_2, the conventional L_2 norm analysis gives the equation:

$$\sigma_{e_1}^2 = \frac{1}{12}\left(2^{-1-W} 2^{-1+I} \cdot 2\right)^2 << \sigma_{e_2}^2 = \frac{1}{12}\left(2^{-1-F} \cdot 2\right)^2 . \tag{24}$$

This leads the condition :

$$W > F + I - 1 . \tag{25}$$

The conventional L_∞ norm analysis is based on

$$\max\left|e_1\right| = 2^{-1-W} 2^{-1+I} < \max\left|e_2\right| = 2^{-1-F} . \tag{26}$$

It also leads the condition in eq.(25).

Fig.3 illustrates the absolute value of E_m in eq. (15) at the minimum word length W determined by eq.(25). It is confirmed that the condition does not guarantee E_m=0. In our analysis, the minimum word length W_{min} is defined by

$$W_{\min} = \{W \in \mathbf{Z} \mid \ E_m = 0, \ \forall W \geq W_{\min}\} . \tag{27}$$

It is determined using eq.(16) for each input value x.

Fig.4 indicates the minimum W for given h and F. In case of h=K, F=0, the condition by the conventional analysis is W >8 [bit]. The condition by our analysis is less than or greater that 8 [bit] depending on x. For example, $W_{\min}$=13 at x=202 and $W_{\min}$=-1 (the lowest bit of integer part of h is not necessary) at x=1.

The conventional analysis methods treat the input values statistically, and therefore, it is impossible to estimate $W_{\min}$ for each input value x. On the contrary, our analysis in eq.(16) gives $W_{\min}$ for each x and it includes the conventional result in eq.(25) as a special case.

B. Word Length under the Lossless Scaling Pair Condition

Similarly, in fig.2 (b), the output w is described by

$$\begin{cases} w = h_2 y + e_2 \\ y = h_1 x + e_1 \end{cases}, \quad \begin{cases} e_1 = -\Delta_{F_1}[h_1 x] \\ e_2 = -\Delta_{F_2}\left[h_2 y\right] \end{cases} . \tag{28}$$

The L_2 norm analysis is based on the inequality:

$$\sigma_{h_2 e_1}^2 = \frac{1}{12}\left(h_2 2^{-1+F_1} \cdot 2\right)^2 << \sigma_{e_2}^2 = \frac{1}{12}\left(2^{-1-F_2} \cdot 2\right)^2 . \tag{29}$$

It leads the lossless condition as

$$F_1 > F_2 + \log_2\left|h_2\right| . \tag{30}$$

The L_∞ norm analysis based on

$$\max|h_2 e_1| = |h_2| 2^{-1+F_1} < \max|e_2| = 2^{-1-F_2} \qquad (31)$$

also leads the same condition in eq.(30).

This condition agrees with eq.(22) derived from our analysis, and therefore the losslessness is guaranteed. Fig.5 indicates the minimum F_1 for given F_2, h_1, h_2 and x under the mapping invariant condition. In case of h_1=K, this is the scaling pair in high in fig.1, the conventional analysis leads F_1> -0.30 and F_1>0.70 for F_2=0 and F_2=1 respectively. In our analysis, the minimum word length is defined by

$$F_{1,\min} = \{F_1 \in \mathbf{Z} \mid E_p = 0, \ \forall F_1 \geq F_{1,\min}\} \quad . \qquad (32)$$

It is determined using eq.(20) for each input value x. For example, $F_{1,\min}$=-1 at x=1 and $F_{1,\min}$=0 at x=2 for F_2=0 and F_2=1 respectively in case of h_1=K. All of $F_{1,\min}$ determined by our analysis are less than the conventional results. This implies that the conventional analysis always gives redundant values of $F_{1,\min}$.

It should be noticed that, in the scaling pair in fig.1, input values are limited to a specific set of values (see [9]).

VI. Conclusion Remarks

In this report, we mathematically derived formulas of the rounding operation to analyze the lossless condition on word length of signals and coefficients. It was confirmed that our equations can determine the minimum word length for each input value more precisely than the conventional analysis.

References

[1] A. Descampe, F. O. Devaux, G. Rouvroy, J. D. Legat, J. J. Quisquater, B. Macq, "A Flexible Hardware JPEG 2000 Decoder for Digital Cinema", IEEE Trans. Circuits and Systems for VT, vol. 16, issue 11, pp.1397 - 1410, Nov. 2006

[2] Bing Fei Wu, Chung Fu Lin, "Memory-efficient Architecture for JPEG 2000 Coprocessor with Large Tile Image, IEEE Trans. Circuits and Systems II, vol.53, issue 4, pp.304-308, April 2006.

[3] A. M. Reza, Lian Zhu, "Analysis of Error in The Fixed-point Implementation of Two-dimensional Discrete Wavelet Transforms", IEEE Trans. Circuits and Systems I, vol.52, issue 3, pp.641-655, March 2005.

[4] M. Primbs, "Worst-case Error Analysis of Lifting-based Fast DCT-algorithms", IEEE Trans. on Signal Processing, vol. 53, pp.3211-3218, 2005.

[5] W. Sweldens, "The Lifting Scheme: A Custom-design Construction of Biorthogonal Wavelets", Technical Report 1994:7, Industrial Mathematics Initiative, Department of Mathematics, University of South Carolina, 1994.

[6] H. Kiya, M. Yae, M. Iwahashi, "Linear Phase Two Channel Filter Bank allowing Perfect Reconstruction", IEEE International Symposium on Circuits and Systems, no.2, pp.951-954, May 1992.

[7] ISO/IEC FCD15444-1, "JPEG2000 Image Coding System", March 2000.

[8] Y. Tonomura, S. Chokchaitam, M. Iwahashi, "Minimum Hardware Implementation of Multipliers of the Lifting Wavelet Transform", IEEE International Conference on Image Processing, WA-L4, pp.2499-2502, Oct. 2004.

[9] H. Kiya , M. Iwahashi, O. Watanabe, "A New Class of Lifting Wavelet Transform for Guaranteeing Losslessness of Specific Signals", IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.3273-3276, March 2008.

[10] IEEE Standard 754-1985, "IEEE Standard for Binary Floating-Point Arithmetic", 1985.

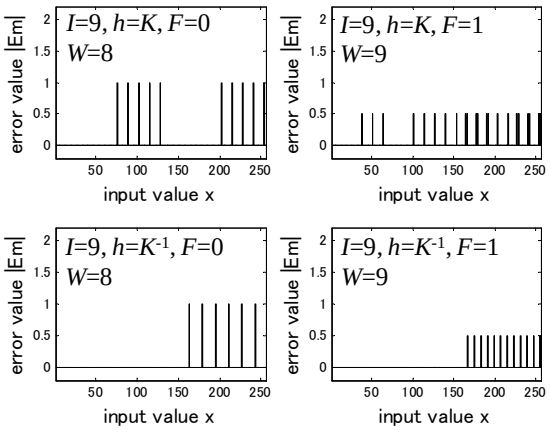


Fig.3 The error $|E_m|$ at the minimum word length W determined by the conventional analysis in eq.(25).

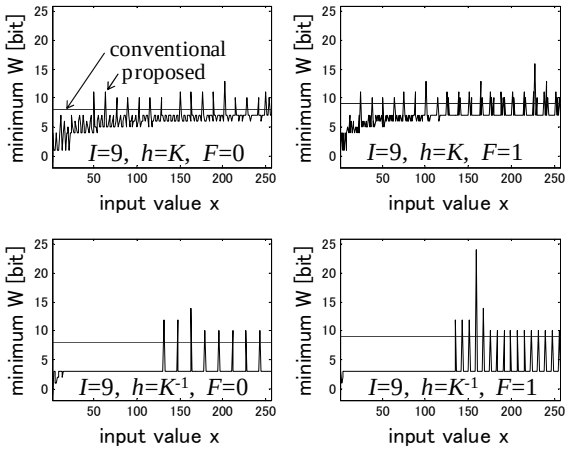


Fig.4 The minimum word length W determined by our analysis in eq.(16) and the conventional analysis in eq.(25).

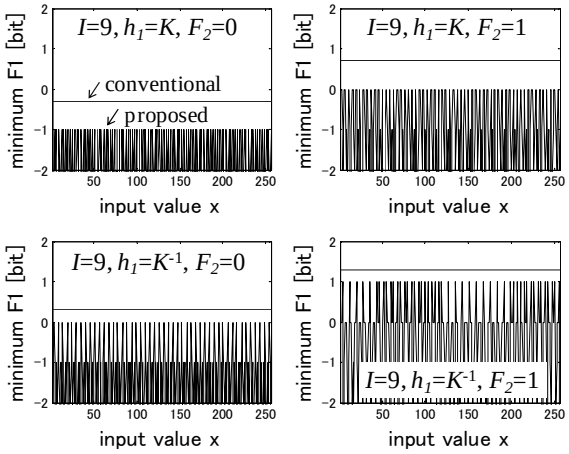


Fig.5 The minimum word length F_1 determined by our analysis in eq.(20) and the conventional analysis in eq.(30).