Numerical Accuracy Evaluation for Polynomial Computation

Naud Jean-Charles, Daniel Menard

Numerical Accuracy Evaluation for Polynomial Computation

Jean-Charles Naud, Daniel Ménard

RESEARCH REPORT
N° 7878
February 2012
Project-Teams CAIRN

ISSN 0249-6399 ISRN INRIA/RR--7878--FR+ENG

Numerical Accuracy Evaluation for Polynomial Computation

Jean-Charles Naud[*†], Daniel Ménard[‡†]

Project-Teams CAIRN

Research Report n° 7878 — February 2012 — 17 pages

Abstract: Fixed-point conversion requires fast analytical methods to evaluate the accuracy degradation due to quantization noises. Usually, analytical methods do not consider the correlation between quantization noises. Correlation between quantization noises occurs when a data is quantized several times. This report explained, through an example, the methodology used in the ID.Fix tool to support correlation. To decrease the complexity, a method to group together several quantization noises inside a same noise source is described. The maximal relative estimation error obtained with the proposed approach is less than 2%.

Key-words: fixed-point, quantization, analytic approach, noise, correlation

This is a note

This is a second note

* University of Rennes, INRIA, ENSSAT 6 rue Kerampont, 22300 Lannion. e-mail: jean-charles.naud@irisa.fr

† Shared foot note

‡ IRISA/INRIA, ENSSAT 6 rue Kerampont, 22300 Lannion. e-mail: menard@irisa.fr

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Évaluation de la précision numérique pour calcul polynomial

Résumé : La conversion en virgule fixe nécessite des méthodes analytiques rapides pour évaluer la dégradation de la précision liée aux bruits de quantification. Actuellement, les méthodes analytiques ne considèrent pas la corrélation entre les bruits de quantification. Cette corrélation est due à la quantification d'une même donnée plusieurs fois. Ce rapport explique, au travers d'un exemple, la méthodologie utilisée dans l'outils ID.Fix pour prendre en compte la corrélation. Pour diminuer la complexité, une méthode de regroupement des bruits de quantification dans une source de bruit est décrite. La valeur maximale de l'erreur d'estimation relative obtenue avec l'approche proposée est inférieuree à 2%.

Mots-clés : virgule fixe, quantification, approche analytique, correlation

Contents

1 Introduction

Fixed-point arithmetic is widely used in embedded systems to reduce implementation costs like execution time, area and power consumption. Fixed-point conversion is composed of two main steps corresponding to dynamic range evaluation and word-length (WL) optimization. The aim of WL optimization is to minimize the implementation cost as long as the effects of finite precision are acceptable. This optimization process is based on an iterative procedure where the numerical accuracy is evaluated a great number of times. Thus, efficient methods are required to evaluate this numerical accuracy to limit the optimization time.

To evaluate numerical accuracy, approaches based on fixed-point simulations are generic, but they also lead to long execution times. Thus, the search space is drastically limited and sub-optimal solutions are obtained. Analytic methods reduce significantly the evaluation time by providing the mathematical expression of a metric equivalent to the numerical accuracy. The output quantization noise power is widely used as a relevant metric for evaluating the numerical accuracy.

Usually, analytical methods do not consider the correlation between quantization noises. Correlation between quantization noises occurs when a data is quantized several times. In this report, the expressions of the correlation and the covariance, considering the number of eliminated bits, are presented for truncation and rounding. Then, this report explained, through an example, the methodology used in the ID.Fix tool to support correlation.

2 Methodology description

Problem description

Correlation between QNSs occurs when a data x_0 is quantized several times. Figure 1 shows such an example where x_i is the data after each quantization $\mathcal{Q}_i$ with i equal to 1 or 2. $\mathcal{Q}_i$ leads to an unavoidable quantization error e_i between the values of the data x_i and x_0. e_i can be assimilated to a noise source and we denote E_i as the random variable corresponding to this error. Let w_i denote the fractional part word-length of the data x_i and q_i the quantization step associated to x_i. $q_i = 2^{-w_i}$ with w_i the weight of the least significant bit. The number of bits eliminated during the quantization process $\mathcal{Q}_i$ is defined as k_i. The relation between the quantization step q_i and q_0 is $q_i = 2^{k_i} q_0$, with i equal to 1 or 2. If x_0 is in infinite precision q_0 is equal to zero and k_1 and k_2 tend to infinity. Let X_i denote the set containing all the values that can be represented in the fixed-point format after the quantization $\mathcal{Q}_i$.

Given that k_2 bits are common between the QNSs e_1 and e_2, these QNSs are correlated. Let y denote the output of the global targeted system. Let H_i denote the system having e_i as input and y as output. As the QNS e_i propagates through the system H_i, correlation between different QNSs e_i obviously influences the output noise power and has therefore to be considered for a precise numerical accuracy evaluation.

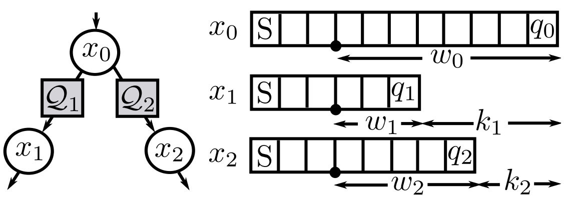

Figure 1: Quantized data representation

2.1 Quantization Mode Description

In this section, the probability density function and the statistical moments of the QNSs generated during a quantization process are presented for the truncation and rounding quantization modes in the case of two's complement coding. The quantization process $\mathcal{Q}_1$ presented in Figure 1 is under consideration. The quantization error e_1 resulting from the quantization process $\mathcal{Q}_1$ is defined as

$$e_1 = x_0 - x_1. \tag{1}$$

By using Widrow's model [1, 2], e_1 can be assimilated to an additive white noise, uniformly distributed, which is uncorrelated to the signal.

2.1.1 Truncation

In the case of truncation, the data x_0 is always rounded towards the lower value available in the set X_1 and becomes

$$x_1 = \lfloor x_0 \cdot q_1^{-1} \rfloor \cdot q_1 = \quad t \cdot q_1 \quad \forall x_0 \in [t \cdot q_1, (t+1) \cdot q_1[\tag{2}$$

with $\lfloor \cdot \rfloor$ the floor function defined as $\lfloor x_0 \rfloor = \max\left(n \in \mathbb{Z} | \, n \leq x_0\right)$ and with q_1 the quantization step. The probability density function (PDF) of the QNS $p_{E_1}(e_1)$ is given by (3) with δ the Kronecker delta.

$$p_{E_1}(e_1) = \frac{1}{2^{k_1}} \sum_{j=0}^{2^{k_1}-1} \delta(e_1 - j \cdot q_0) \tag{3}$$

2.1.2 Rounding

Rounding quantization mode rounds the value x_0 to the nearest value available in the set X_1 as

$$x_1 = \left\lfloor \left(x_0 + \frac{1}{2} q_1 \right) \cdot q_1^{-1} \right\rfloor \cdot q_1. \tag{4}$$

The midpoint $q_m = (t + \frac{1}{2}) \cdot q_1$ between $t \cdot q_1$ and $(t+1) \cdot q_1$ is always rounded up to the higher value $(t+1) \cdot q_1$. For this quantization mode, the PDF $p_{E_1}(e_1)$ is given by

$$p_{E_1}(e_1) = \frac{1}{2^{k_1}} \sum_{j=-2^{k_1-1}}^{2^{k_1-1}-1} \delta(e_1 - j \cdot q_0.). \tag{5}$$

From [3] and [4], mean and variance expressions are given in Table 1 for each quantization mode Q_1. If x_0 has a continuous amplitude, as in analog-to-digital conversion, k_1 is considered as $+\infty$.

2.2 Correlation and covariance expressions

In this section, the expressions of the correlation and the covariance between two QNSs e_1 and e_2, resulting from the quantization of one unique data x_0 as presented in Figure 1, are determined. The covariance is used in the expression of the global output quantization noise power to improve the quality of the noise power estimation. The reasoning to determine the correlation and covariance expressions is detailed in the following with $k_1 \geq k_2$ and for the case

Table 1: Mean and variance for the two quantization modes

Quantization mode Q_1	Mean	Variance
Truncation	$\frac{q_1}{2} \cdot (1 - 2^{-k_1})$	$\frac{q_1^2}{12} \cdot (1 - 2^{-2k_1})$
Rounding	$-\frac{q_1}{2} \cdot (2^{-k_1})$	$\frac{q_1^2}{12} \cdot (1 - 2^{-2k_1})$

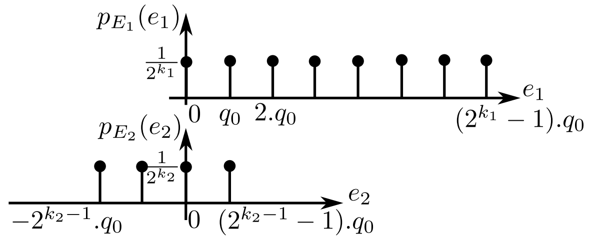

Figure 2: PDF of the QNSs e_1 and e_2 for $\mathcal{Q}_1 = T$, $\mathcal{Q}_2 = R$, $k_1 = 3$ and $k_2 = 2$. q_0 is the quantization step.

where $\mathcal{Q}_1$ is a truncation (T) and $\mathcal{Q}_2$ a rounding (R). The two discrete PDFs p_{E_1} and p_{E_2} of respectively the QNSs e_1 and e_2 are presented in Figure 2 for the case of $k_1 = 3$ and $k_2 = 2$.

Let E_1 and E_2 denote the discrete random variables corresponding to the QNSs. The correlation $\mathrm{E}[E_1 \cdot E_2]$ is determined from p_{E_1,E_2} the joint probability density function between E_1 and E_2 as

$$\mathrm{E}[E_1 \cdot E_2] = \sum_i \sum_j i.q_0.j.q_0.p_{E_1,E_2}(e_1 = i.q_0, e_2 = j.q_0) \tag{6}$$

where i and j enumerate the possible events of e_1 and e_2.

The joint probability density function p_{E_1,E_2} is obtained from $p_{E_1|E_2}$ the conditional probability of E_1 given E_2 as

$$p_{E_1,E_2}(e_1, e_2) \quad = \quad p_{E_1|E_2}(e_1|e_2) \cdot p_{E_2}(e_2), \tag{7}$$

For each value of e_2, $2^{k_1-k_2}$ values are obtained for e_1, thus

$$\begin{aligned} p_{E_1|E_2} \quad = \quad & \frac{1}{2^{k_1-k_2}} \cdot \\ & \left[\sum_{j=0}^{2^{k_2-1}-1} \delta(e_2 - jq_0) \sum_{t=0}^{2^{k_1-k_2}-1} \delta\left(e_1 - e_2 - t \cdot q_2\right)\right) \\ + & \sum_{j=-2^{k_2-1}}^{-1} \delta(e_2 - jq_0) \sum_{t=0}^{2^{k_1-k_2}-1} \delta\left(e_1 - e_2 - (t+1) \cdot q_2\right) \right] \end{aligned} \tag{8}$$

To illustrate equation 8, the example with $k_1 = 3$ and $k_2 = 2$ is considered. Thus, the different cases for the quantization error e_1 and e_2 are provided in Table 2.

The columns bit 2, bit 1 and bit 0 specify the three LSB of the quantized data (x_0). The weight of the bit 0 is q_0. In our case, e_2 is due to the elimination of 2 bits with rounding quantization mode. The figure 3 presents the different values of the error e_2.

bit 2	bit 1	bit 0	e_1	e_2
0	0	0	0	0
0	0	1	q_0	q_0
0	1	0	$2q_0$	$-2q_0$
0	1	1	$3q_0$	$-q_0$
1	0	0	$4q_0$	0
1	0	1	$5q_0$	q_0
1	1	0	$6q_0$	$-2q_0$
1	1	1	$7q_0$	$-q_0$

Table 2: Values of the eliminated bits during the quantization and associated error e_1 and e_2.

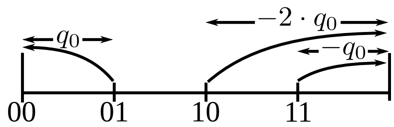

Figure 3: Different values of the error e_2 for the rounding quantization with $k_2 = 2$

The result is shown in Figure 4. Red Dirac and green Dirac correspond respectively to the first part (two first summations) and the second part (two last summations) of the previous expression. The amplitude of the different Dirac are the same and equal to $\frac{1}{2^{k_1-k_2}}$.

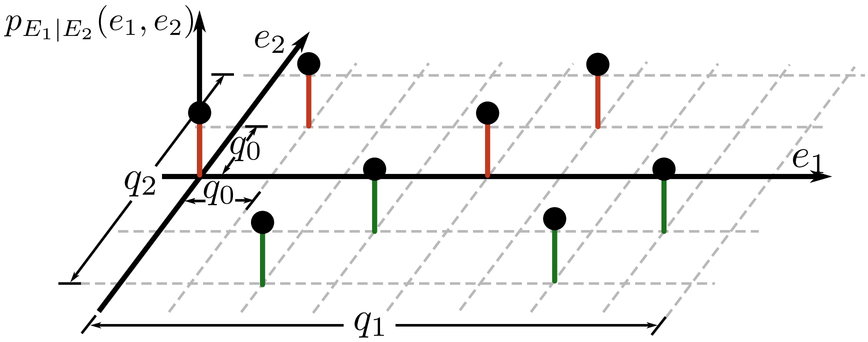

Figure 4: Joint probability $p_{E_1|E_2(e_1,e_2)}$

From eq. 7 and 6, the correlation between E_1 and E_2 becomes

$$\begin{aligned} \mathrm{E}\left[E_1.E_2\right] &= \frac{q_0^2}{2^{k_1}} \sum_{t=0}^{2^{k_1-k_2}-1} \sum_{j=0}^{2^{k_2-1}-1} j(j+t.2^{k_2}) \\ &+ (j-2^{k_2-1})(j+2^{k_2-1}+t.2^{k_2}) \\ &= -\frac{q_2^2}{24} + \frac{q_0^2}{6} - \frac{q_0 \cdot q_1}{4}. \end{aligned} \quad (9)$$

The covariance $\mathrm{cov}(E_1, E_2)$ is determined from the correlation term $\mathrm{E}\left[E_1.E_2\right]$ as

$$\mathrm{cov}(E_1, E_2) = \mathrm{E}\left[E_1 \cdot E_2\right] - \mathrm{E}\left[E_1\right] \cdot \mathrm{E}\left[E_2\right]. \quad (10)$$

Table 3: Correlation and covariance expressions for the different quantization modes of truncation (T) or rounding (R), and for different conditions on k_1 and k_2

Q_1	Q_2	Condition	$\mathrm{E}[E_1.E_2]$	$\mathrm{cov}(E_1, E_2)$
T	T	$k_1 \geq k_2$	$\frac{q_2^2}{12} + \frac{q_0^2}{6} - \frac{q_0.q_2}{4} - \frac{q_0.q_1}{4} + \frac{q_1.q_2}{4}$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
T	R	$k_1 \geq k_2$	$-\frac{q_2^2}{24} + \frac{q_0^2}{6} - \frac{q_0.q_1}{4}$	$-\frac{q_2^2}{24} - \frac{q_0^2}{12}$
R	T	$k_1 > k_2$	$\frac{q_2^2}{12} + \frac{q_0^2}{6} - \frac{q_0.q_2}{4}$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
R	R	$k_1 = k_2$	$\frac{q_2^2}{12} + \frac{q_0^2}{6}$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
R	R	$k_1 > k_2$	$-\frac{q_2^2}{24} + \frac{q_0^2}{6}$	$-\frac{q_2^2}{24} - \frac{q_0^2}{12}$

The term $\mathrm{E}[E_1].\mathrm{E}[E_2]$ can be computed from the equations given in Table 1 and is equal to

$$\begin{aligned} \mathrm{E}[E_1].\mathrm{E}[E_2] &= \frac{q_1}{2}\left(1 - 2^{-k_1}\right)\frac{q_2}{2}(-2^{-k_2}) \\ &= \frac{q_0^2 - q_0.q_1}{4} \end{aligned} \quad (11)$$

Thus, from eq. 9 and 11, the expression of the covariance becomes

$$\mathrm{cov}(E_1, E_2) = -\frac{q_2^2}{24} - \frac{q_0^2}{12} \quad (12)$$

Given that $k_1 \geq k_2$, the term q_1 is eliminated because just the k_2 least significant bits are common between e_1 and e_2. The correlation and covariance expressions are given in Table 4 for the different quantization modes Q_i corresponding to truncation (T) or rounding (R) and for different conditions on k_1 and k_2. If x_0 is in infinite precision, q_0 is equal to 0.

2.3 Output Quantization Noise Power

Different models have been proposed to estimate the power of the quantization noise at the output of a system [5, 6, 7, 8]. These approaches do not consider the correlation between quantization noises, but they can be easily extended to integrate this correlation to improve the estimation quality.

From [6], the output quantization noise e_y is the sum of the contributions of the N QNSs e_i

$$e_y(n) = \sum_{i=1}^{N}\sum_{t=0}^{\infty} h_i(t, n) \cdot e_i(n - t) \quad (13)$$

where h_i corresponds to the time-varying impulse response of the system H_i between e_i and the output y. The term t represents the delay and n the time.

The power P_{e_y} of the output quantization noise is obtained by determining the second order moment of e_y with a similar derivation as in [6]. From (14), the expression of P_{e_y} is

$$P_{e_y} = \mathrm{E}\left[E_y^2\right] = E\left[\left(\sum_{i=1}^{N}\sum_{t=0}^{\infty} h_i(t, n) \cdot e_i(n - t)\right)^2\right]$$

$$\begin{aligned}
\mathrm{E}\left[E_y^2\right] &= \mathrm{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=0}^{\infty}\sum_{v=0}^{\infty} h_i(t,n)\cdot h_j(v,n)\cdot e_i(n-t)\cdot e_j(n-v)\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{N}\sum_{t=0}^{\infty} h_i^2(t,n)\cdot e_i^2(n-t)\right. \\
&\quad + \sum_{i=1}^{N}\sum_{t=0}^{\infty}\sum_{\substack{v=0\\v\neq t}}^{\infty} h_i(t,n)\cdot h_i(v,n)\cdot e_i(n-t)\cdot e_i(n-v) \\
&\quad + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\sum_{t=0}^{\infty} h_i(t,n)\cdot h_j(t,n)\cdot e_i(n-t)\cdot e_j(n-t) \\
&\quad \left. + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\sum_{t=0}^{\infty}\sum_{\substack{v=0\\v\neq t}}^{\infty} h_i(t,n)\cdot h_j(v,n)\cdot e_i(n-t)\cdot e_j(n-v)\right] \qquad (14)
\end{aligned}$$

The different quantization noises e_i are assumed to be white so the autocorrelation is equal to

$$\mathrm{E}[e_i(n-t)e_i(n-v)] = \sigma_i^2\delta(t-v)+\mu_i^2 \qquad (15)$$

and the intercorrelation

$$\mathrm{E}[e_i(n-t)e_j(n-v)] = \mathrm{cov}(E_i,E_j)\delta(t-v)+\mu_i\mu_j \qquad (16)$$

$$\begin{aligned}
\mathrm{E}\left[E_y^2\right] &= \sum_{i=1}^{N}\sum_{t=0}^{\infty} E\left[h_i^2(t,n)\right]\cdot(\sigma_i^2+\mu_i^2) \\
&\quad + \sum_{i=1}^{N}\sum_{t=0}^{\infty}\sum_{\substack{v=0\\v\neq t}}^{\infty} \mathrm{E}\left[h_i(t,n)\cdot h_i(v,n)\right]\cdot\mu_i^2 \\
&\quad + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\sum_{t=0}^{\infty} \mathrm{E}\left[h_i(t,n)\cdot h_j(t,n)\right]\cdot(\mu_i\cdot\mu_j+\mathrm{cov}(E_i,E_j)) \\
&\quad + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\sum_{t=0}^{\infty}\sum_{\substack{v=0\\v\neq t}}^{\infty} E\left[h_i(t,n)\cdot h_j(v,n)\right]\cdot\mu_i\cdot\mu_j \qquad (17)
\end{aligned}$$

$$\begin{aligned}
P_{e_y} = \mathrm{E}\left[E_y^2\right] &= \sum_{i=1}^{N} K_i\cdot\sigma_i^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} L_{ij}\cdot\mu_i\cdot\mu_j \\
&\quad + \sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N} M_{ij}\cdot\mathrm{cov}(E_i,E_j)
\end{aligned}$$

with

$$K_i = \sum_{t=0}^{\infty} \mathrm{E}\left[h_i^2(t,n)\right],$$
$$L_{ij} = \sum_{t=0}^{\infty}\sum_{v=0}^{\infty} \mathrm{E}\left[h_i(t,n)h_j(v,n)\right],$$
$$M_{ij} = \sum_{t=0}^{\infty} \mathrm{E}\left[h_i(t,n)h_j(t,n)\right], \tag{18}$$

and where K_i, L_{ij} and M_{ij} are constant terms depending only of the system in infinite precision, which can thus be determined only once. The variance σ_i^2, the mean μ_i and the covariance $\mathrm{cov}(E_i, E_j)$ between QNSs depend on the quantization modes, the number of bit w_i for the fractional part and the number of bits eliminated k_i for each data x_i.

3 Application description

To illustrate the proposed approach, the computation of a 3^{rd} order polynomial using the Horner scheme is presented. The Signal Flow Graph is presented in figure 5. The expression of the output y is

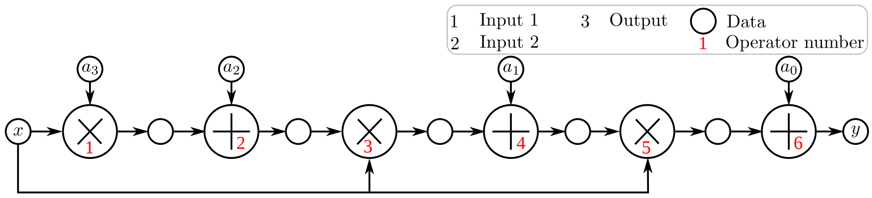

Figure 5: 3^{rd} order polynomial using the Horner scheme

$$y(n) = a_0 + x(n).(a_1 + x(n).(a_2 + x(n).(a_3 + x(n)))) \tag{19}$$

3.1 Quantizations noise modelling

A quantization noise is generated when the fractional word length (w_{FP}) of a data is reduced. Each quantization error is modelled as a noise with the mean μ and the variance σ. A quantization noise can be modelled with the quantization step q, the quantization mode (t_Q) (truncation (T) or rounding (R)) and the number of eliminated bits k between the old and the new format.

3.2 Noise source

The noise sources b allow grouping together different quantization noises e in the signal flow graph. Given that the complexity of the method depends on the number of quantization noise, the quantization noises are grouped together as much as possible to decrease the number of noise sources proceeded by the method.

The figure 6 shows, in the general case, the potential quantization noises (I, II, III, IV) associated with the noise source b. The w_{FP} of inputs and output of operation P_1 are specified with $w_{FP_{P1,0}}$, $w_{FP_{P1,1}}$ and $w_{FP_{P1,2}}$.

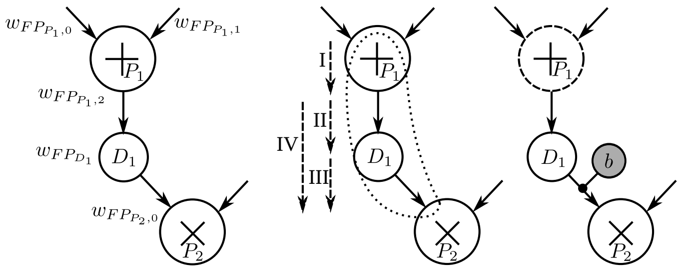

Figure 6: Quantization noise in an noise source

P_i and D_i represent respectively the number associated to the operations and the data. Two cases are identified. In the first case, there is a word-length constraint $w_{FP_{D1}}$ applied on data D_1. The quantization noises are I,II and III. In the second case, there is no restriction on $w_{FP_{D1}}$. In this case, only quantization noise I and IV are considered.

For the example of polynomial computation, the noise sources are inserted as show in figure 7.

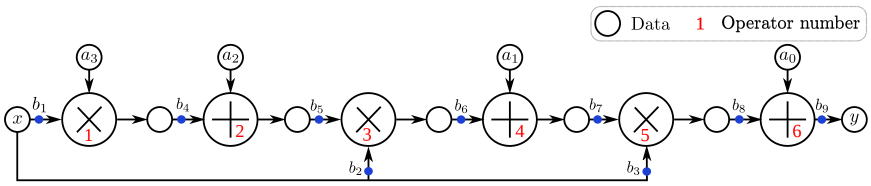

Figure 7: Noise sources insertion

3.3 Noise model propagation

The noise model propagation of each operator is inserted. An operation with two inputs x and y and the output z is considered. Let e_x, e_y and e_z denotes the quantization noises associated with the input and the output. For the addition/subtraction, the output noise expression is :

$$e_z = e_x \pm e_y$$

For the multiplication, the output noise expression is :

$$e_z = e_x y \pm e_y x$$

Constant value are not considered to generate a quantization noise. By using previous expression, the noise data flow graph is generated as show in figure 8.

3.4 Determination of the impulse response of the system

Let H_i denote the system having the noise source b_i as input and y as output. Let $h_i(t, n)$ denote the impulse response of H_i. The term t represents the delay and n the time. In our case, there is no delay, so $h_i = 0$, $\forall t \neq 0$.

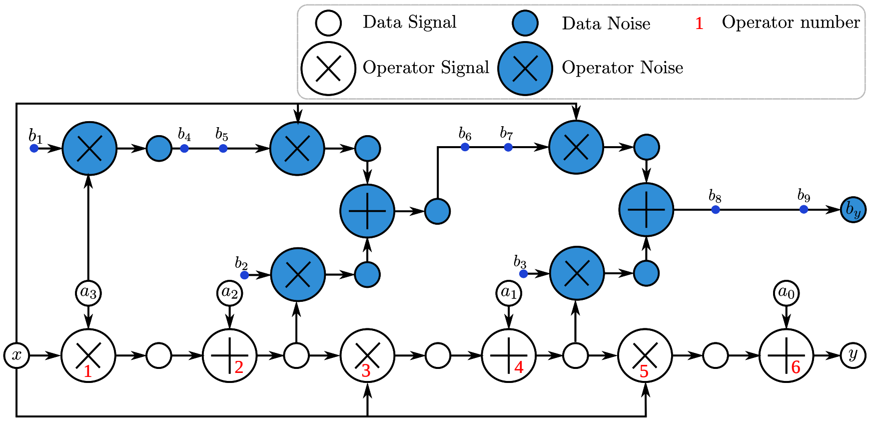

Figure 8: Noise models insertion

The expressions of the different impulse response h_i are :

$$\begin{aligned}
h_1(0,n) &= a_3 * x^2(n) \\
h_2(0,n) &= (a_3 * x(n) + a_2) \cdot x(n) \\
h_3(0,n) &= (a_3 * x(n) + a_2) \cdot x(n) + a_1 \\
h_4(0,n) &= x^2(n) \\
h_5(0,n) &= x^2(n) \\
h_6(0,n) &= x(n) \\
h_7(0,n) &= x(n) \\
h_8(0,n) &= 1 \\
h_9(0,n) &= 1
\end{aligned} \tag{20}$$

The matrices K, L et M are computed only one time with the followers equations.

$$K_i = \sum_{t=0}^{\infty} \mathrm{E}\left[h_i^2(t,n)\right] \tag{21}$$

$$L_{ij} = \sum_{t=0}^{\infty} \sum_{v=0}^{\infty} \mathrm{E}\left[h_i(t,n) h_j(v,n)\right] \tag{22}$$

$$M_{ij} = \sum_{t=0}^{\infty} \mathrm{E}\left[h_i(t,n) h_j(t,n)\right] \tag{23}$$

The N-length vector K is equal to

$$\begin{pmatrix} \mathrm{E}\left[(a_3 * x^2(n))^2\right] \\ \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n))^2\right] \\ \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n) + a_1)^2\right] \\ \mathrm{E}\left[x^4(n)\right] \\ \mathrm{E}\left[x^4(n)\right] \\ \mathrm{E}\left[x^2(n)\right] \\ \mathrm{E}\left[x^2(n)\right] \\ 1 \\ 1 \end{pmatrix}$$

The different element of the $N \times N$ matrix L are

$$
\begin{aligned}
L_{11} &= \mathrm{E}\left[(a_3 * x^2(n))^2\right] \\
L_{12} = L_{21} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot ((a_3 * x(n) + a_2) \cdot x(n))\right] \\
L_{13} = L_{31} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot ((a_3 * x(n) + a_2) \cdot x(n) + a_1)\right] \\
L_{14} = L_{41} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot x^2(n))\right] \\
L_{15} = L_{51} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot x^2(n))\right] \\
L_{16} = L_{61} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot x(n))\right] \\
L_{17} = L_{71} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot x(n))\right] \\
L_{18} = L_{81} &= \mathrm{E}\left[a_3 * x^2(n)\right] \\
L_{19} = L_{91} &= \mathrm{E}\left[a_3 * x^2(n)\right]
\end{aligned}
$$

$$
\begin{aligned}
L_{22} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n))^2\right] \\
L_{23} = L_{32} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n)) \cdot ((a_3 * x + a_2) \cdot x(n) + a_1)]\right] \\
L_{24} = L_{42} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n)) \cdot x^2(n)\right] \\
L_{25} = L_{52} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n)) \cdot x^2(n)\right] \\
L_{26} = L_{62} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n)) \cdot x(n)\right] \\
L_{27} = L_{72} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n)) \cdot x(n)\right] \\
L_{28} = L_{82} &= \mathrm{E}\left[(a_3 * x(n) + a_2) \cdot x(n)\right] \\
L_{29} = L_{92} &= \mathrm{E}\left[(a_3 * x(n) + a_2) \cdot x(n)\right]
\end{aligned}
$$

$$
\begin{aligned}
L_{33} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n) + a_1)^2\right] \\
L_{34} = L_{43} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n) + a_1) \cdot x^2(n)\right] \\
L_{35} = L_{53} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n) + a_1) \cdot x^2(n)\right] \\
L_{36} = L_{63} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n) + a_1) \cdot x(n)\right] \\
L_{37} = L_{73} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n) + a_1) \cdot x(n)\right] \\
L_{38} = L_{83} &= \mathrm{E}\left[(a_3 * x(n) + a_2) \cdot x(n) + a_1\right] \\
L_{39} = L_{93} &= \mathrm{E}\left[(a_3 * x(n) + a_2) \cdot x(n) + a_1\right]
\end{aligned}
$$

$$
\begin{aligned}
L_{44} &= \mathrm{E}\left[(x^2(n))^2\right] \\
L_{45} = L_{54} &= \mathrm{E}\left[(x^2(n))^2\right] \\
L_{46} = L_{64} &= \mathrm{E}\left[(x^2(n)) \cdot x(n)\right] \\
L_{47} = L_{74} &= \mathrm{E}\left[(x^2(n)) \cdot x(n)\right] \\
L_{48} = L_{84} &= \mathrm{E}\left[x^2(n)\right] \\
L_{49} = L_{94} &= \mathrm{E}\left[x^2(n)\right]
\end{aligned}
$$

$$
\begin{aligned}
L_{55} &= \mathrm{E}\left[(x^2(n))^2\right] \\
L_{56} = L_{54} &= \mathrm{E}\left[(x^2(n)) \cdot x(n)\right] \\
L_{57} = L_{75} &= \mathrm{E}\left[(x^2(n)) \cdot x(n)\right] \\
L_{58} = L_{85} &= \mathrm{E}\left[x^2(n)\right] \\
L_{59} = L_{95} &= \mathrm{E}\left[x^2(n)\right]
\end{aligned}
$$

$$\begin{aligned}
L_{66} &= \mathrm{E}\left[x^2(n)\right] \\
L_{67} = L_{76} &= \mathrm{E}\left[x^2(n)\right] \\
L_{68} = L_{86} &= \mathrm{E}\left[x(n)\right] \\
L_{69} = L_{96} &= \mathrm{E}\left[x(n)\right] \\
\\
L_{77} &= \mathrm{E}\left[x^2(n)\right] \\
L_{78} = L_{87} &= \mathrm{E}\left[x(n)\right] \\
L_{79} = L_{97} &= \mathrm{E}\left[x(n)\right] \\
\\
L_{88} &= 1 \\
L_{89} = L_{98} &= 1 \\
\\
L_{99} &= 1
\end{aligned}$$

The different element of the $N \times N$ matrix M are

$$\begin{aligned}
M_{12} = M_{21} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot ((a_3 * x(n) + a_2) \cdot x(n))\right] \\
M_{13} = M_{31} &= \mathrm{E}\left[(a_3 * x^2(n)) \cdot ((a_3 * x(n) + a_2) \cdot x(n) + a_1)\right] \\
M_{23} = M_{32} &= \mathrm{E}\left[((a_3 * x(n) + a_2) \cdot x(n)) \cdot ((a_3 * x + a_2) \cdot x(n) + a_1)\right] \\
others &= 0
\end{aligned} \tag{24}$$

To decrease the computation time, the symmetry of L and M is used. Moreover, the M_{ij} terms are not computed if there is no correlation between b_i and b_j. When there is a correlation between b_i and b_j, each noise source b_i or b_j is a single quantization noise. In our case, for the matrix M, only the terms M_{12} , M_{13} and M_{23} have to be computed (M_{21} , M_{31} and M_{32} are deduced by symmetry).

4 Noise power expression

The global noise power expression is

$$P_{b_y} = \sum_{i=1}^{N} K_i.\sigma_i^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} L_{ij} \cdot \mu_i.\mu_j + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} M_{ij} \cdot cov(B_i, B_j) \tag{25}$$

The variable of this expression are σ_i^2, μ_i and $cov(B_i, B_j)$ which depends on the data word-length and the quantization modes. The $cov(B_i, B_j)$ term are compute with the table 4.

Let q_0 is the quantization step of the initial data (x in this example). Let q_1 and q_2 are the quantization step of data quantized and k_1 and k_2 the number of eliminated bits. If the initial data x is in infinite precision q_0 is equal to zero and k_1 and k_2 tend to infinity.

Q_1	Q_2	Condition	$cov(B_1, B_2)$
T	T	$k_1 \geq k_2$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
T	R	$k_1 \geq k_2$	$-\frac{q_2^2}{24} - \frac{q_0^2}{12}$
R	T	$k_1 > k_2$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
R	R	$k_1 = k_2$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
R	R	$k_1 > k_2$	$-\frac{q_2^2}{24} - \frac{q_0^2}{12}$

Table 4: Covariance expressions for the different quantization modes of truncation (T) or rounding (R), and for different conditions on k_1 and k_2

5 Conclusion

In the context of numerical accuracy evaluation of fixed-point systems, the expressions of the correlation and the covariance between QNSs resulting from the quantization of one unique data has been proposed in this report. The expression of the global output quantization noise integrates correlation between QNSs, which improves the quality of the estimation of the output quantization noise compared to existing approaches. The noise power in the case of a third-order polynomial computation has been described in this report.

References

[1] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge, UK: Cambridge University Press, 2008. 5

[2] A. Sripad and D. L. Snyder, "A Necessary and Sufficient Condition for Quantization Error to be Uniform and White," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 5, pp. 442–448, Oct. 1977. 5

[3] G. Constantinides, P. Cheung, and W. Luk, "Truncation Noise in Fixed-Point SFGs," *IEE Electronics Letters*, vol. 35, no. 23, pp. 2012–2014, Nov. 1999. 5

[4] D. Menard, D. Novo, R. Rocher, F. Catthoor, and O. Sentieys, "Quantization Mode Opportunities in Fixed-Point System Design," in *Proc. European Signal Processing Conference (EUSIPCO)*, Aalborg, Aug. 2010, pp. 542–546. 5

[5] C. Shi and R. Brodersen, "A perturbation theory on statistical quantization effects in fixed-point DSP with non-stationary inputs," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, Vancouver, May. 2004, pp. 373–376. 8

[6] R. Rocher, D. Menard, P. Scalart, and O. Sentieys, "Analytical accuracy evaluation of Fixed-Point Systems," in *Proc. European Signal Processing Conference (EUSIPCO)*, Poznan, Sep. 2007. 8

[7] P. Fiore, "Efficient Approximate Wordlength Optimization," *IEEE Transactions on Computers*, vol. 57, no. 11, pp. 1561 –1570, Nov 2008. 8

[8] G. Caffarena, J. López, A. Fernandez, and C. Carreras, "SQNR Estimation of Fixed-Point DSP Algorithms," *EURASIP Journal on Advance Signal Processing*, vol. 2010, 2010. 8

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399