A Discrete Model for Correlation Between Quantization Noises

Jean-Charles Naud, *Student Member, IEEE*, Daniel Ménard, *Member, IEEE*, Gabriel Caffarena, *Senior Member, IEEE*, and Olivier Sentieys, *Member, IEEE*

***Abstract*—The automation of fixed-point conversion requires fast methods to evaluate the numerical accuracy of the system. As an alternative to a simulation-based approach, most of the analytical methods use perturbation theory to provide the expression of the quantization noise at the output of a system. Most existing analytical methods do not consider a correlation between noise sources. This assumption is no longer valid when a unique datum is quantized several times. This brief proposes to study the correlation between quantization noises with different quantization modes (truncation and rounding) and considering the number of eliminated bits. Then, the expression of the power of the output quantization noise is provided when the correlation between the noise sources is considered. The proposed approach allows improving significantly the estimation of the output quantization noise power compared to the classical approach, with a slight increase of the computation time. In our experiment, the maximal relative estimation error obtained with the proposed approach is less than 2% compared to 84% when a correlation is not taken into account.**

***Index Terms*—Correlation, fixed-point arithmetic, numerical accuracy evaluation, quantization noise, round-off noise.**

I. 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 time. Thus, the search space is drastically limited, and suboptimal 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.

Most existing methods do not take into account the correlation between the different quantization noise sources (QNSs), which is due to multiple quantizations of the same data. Consequently, the quality of the estimation can be degraded when multiple quantizations occur and the solution obtained from the WL optimization process can lead to an implementation cost overhead. In [1], graph transformations are applied to handle the correlation. This technique can become complex and is valid only for truncation. In [2], the expression of the correlation is provided but for a very specific case.

In this brief, the expressions of the correlation and the covariance, considering the number of eliminated bits, are proposed for truncation and rounding. Then, to take correlation into account, the expression of the output quantization noise power is provided. This model extends existing methods by integrating the covariance between the noise sources and leads to a slight increase in complexity.

This brief is organized as follows. Existing works handling correlation between QNSs are presented in Section II. In Section III, quantization modes corresponding to truncation and rounding are defined. The expressions of the correlation and the covariance between QNSs are detailed in Section IV. The expression of the global output quantization noise considering the correlation is provided in Section V. The quality of the proposed models is analyzed through experiments in Section VI, while Section VII concludes this brief.

II. Related Work

Correlation between QNSs occurs when a datum x_0 is quantized several times. Fig. 1 shows such an example where x_i is the datum after each quantization $\mathcal{Q}_i$ with i equal to one or two. $\mathcal{Q}_i$ 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 WL of the datum x_i and q_i denote the quantization step associated to x_i. $q_i = 2^{-w_i}$ with w_i as 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 steps q_i and q_0 is $q_i = 2^{k_i} q_0$, with i equal to one or two. 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 $\mathcal{Q}_i$.

Manuscript received November 9, 2011; revised February 10, 2012; accepted October 1, 2012. Date of publication November 16, 2012; date of current version January 4, 2013. This work was supported by the Nano 2012 R&D research program in collaboration with STmicroelectronics. This brief was recommended by Associate Editor O. Gustafsson.

J.-C. Naud, D. Ménard, and O. Sentieys are with Inria, Irisa, University of Rennes, 22305 Lannion, France (e-mail: jean-charles.naud@irisa.fr; daniel.menard@irisa.fr; olivier.sentieys@irisa.fr).

G. Caffarena is with Universidad CEU San Pablo, 28668 Boadilla del Monte, Spain (e-mail: gabriel.caffarenafernandez@ceu.es).

Color versions of one or more of the figures in this brief are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSII.2012.2222838

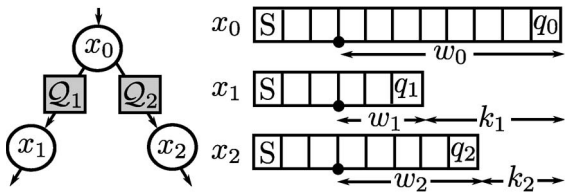

Fig. 1. Quantized data representation.

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.

In [1], the problem of correlation between different QNSs is handled through signal flow graph transformations. If $k_1 > k_2$, the quantization processes $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are transformed into two quantization processes $\mathcal{Q}'_1$ and $\mathcal{Q}'_0$. The quantization process $\mathcal{Q}'_0$, corresponding to the elimination of k_2 bits, is inserted just after the datum x_0. The associated noise source e'_0 propagates through the system corresponding to the combination of the systems H_1 and H_2. The modified quantization process $\mathcal{Q}'_1$ corresponds to the elimination of $k_1 - k_2$ bits. The associated noise source e'_1 propagates through the system H_1. As the quantization process $\mathcal{Q}_2$ disappears, the noise sources e'_1 and e'_0 are considered not correlated, and classical techniques can be used to estimate the output quantization noise. First, this technique is only valid for truncation. Indeed, a quantization with rounding cannot be decomposed into two independent quantization processes. Second, the system considered for noise propagation depends on the values of k_2 and k_1. Therefore, to have a generic model, all the combinations between the systems H_i have to be considered. If one unique datum is quantized P times, the number of systems to consider is $P! + 1$, so growing up exponentially and making more complex the expression of the output quantization noise.

In [2], the correlation between two quantization noises of one datum after multiplication by constants is studied. The correlation between the quantization noise after the multiplication of x_1 by a constant A_1 and the quantization noise after the multiplication of x_2 by constant A_2 is analyzed. Only the case of rounding and the case of linear time-invariant systems are considered. Moreover, the two quantization steps are identical, and only A_1 and A_2 can be different.

In this brief, a correlation model considering different quantization modes (truncation and rounding), different quantization steps, and different kinds of input (x_0 is in infinite or finite precision) is proposed. The proposed approach modifies the expression of the output quantization noise by adding one more term but with a slight increase in the complexity to calculate this expression.

III. Quantization Mode Description

In this section, the probability density function (pdf) 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 $\mathcal{Q}_1$, shown in Fig. 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 the model by Widrow [3], [4], e_1 can be assimilated to an additive white noise, uniformly distributed, which is uncorrelated to the signal.

1) Truncation: In the case of truncation, the datum x_0 is always rounded toward the lower value available in the set X_1 and becomes

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

with $\lfloor \cdot \rfloor$ as the floor function defined as $\lfloor x_0 \rfloor = \max(n \in \mathbb{Z} | n \leq x_0)$ and with q_1 as the quantization step. The pdf of the QNS $p_{E_1}(e_1)$ is given by (3) with δ as 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) Rounding: Rounding the 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 $q_m = (t + (1/2)) \cdot q_1$ 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

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

TABLE I
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	$-\frac{q_1}{2} \cdot (2^{-k_1})$	$\frac{q_1^2}{12} \cdot (1 - 2^{-2k_1})$

From [5] and [6], mean and variance expressions are given in Table I for each quantization mode Q_1. If x_0 has a continuous amplitude, as in analog-to-digital conversion, k_1 is considered as $+\infty$.

IV. 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 datum x_0 as shown in Fig. 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. When the datum x_0 is quantized P times, $P(P-1)/2$ correlations, between the P QNSs, have to be studied. The reasoning to determine the correlation and covariance expressions is detailed in the following with $k_1 \geq k_2$ and for the cases where $\mathcal{Q}_1$ is a truncation (T) and $\mathcal{Q}_2$ is a rounding (R). The two discrete pdfs p_{E_1} and p_{E_2} of the QNSs

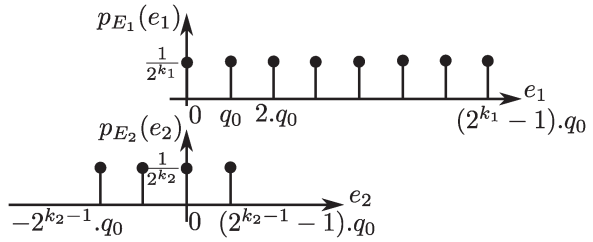

Fig. 2. A pdf of the QNSs e_1 and e_2 for $\mathcal{Q}_1 = T$, $\mathcal{Q}_2 = R$, $k_1 = 3$, and $k_2 = 2$. q_0 is the quantization step.

e_1 and e_2, respectively, are shown in Fig. 2 for the cases 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 $E[E_1 \cdot E_2]$ is determined from p_{E_1,E_2} the joint pdf between E_1 and E_2 as

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

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

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

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

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

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

From (7) and (6), the correlation between E_1 and E_2 becomes

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

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

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

The term $E[E_1] \cdot E[E_2]$ can be computed from the equations given in Table I and is equal to

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

TABLE II
CORRELATION AND COVARIANCE EXPRESSIONS FOR THE DIFFERENT QUANTIZATION MODES OF (T) TRUNCATION OR (R) ROUNDING 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	$k_1 \geq k_2$	$\frac{q_2^2}{12} + \frac{q_0^2}{6} - \frac{q_0 \cdot q_2}{4} - \frac{q_0 \cdot q_1}{4} + \frac{q_1 \cdot q_2}{4}$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
T	R	$k_1 \geq k_2$	$-\frac{q_2^2}{24} + \frac{q_0^2}{6} - \frac{q_0 \cdot q_1}{4}$	$-\frac{q_2^2}{24} - \frac{q_0^2}{12}$
R	T	$k_1 > k_2$	$\frac{q_2^2}{12} + \frac{q_0^2}{6} - \frac{q_0 \cdot q_2}{4}$	$\frac{q_2^2}{12} - \frac{q_0^2}{12}$
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	$k_1 > k_2$	$-\frac{q_2^2}{24} + \frac{q_0^2}{6}$	$-\frac{q_2^2}{24} - \frac{q_0^2}{12}$

Thus, from (9) and (11), the expression of the covariance becomes

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

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 II 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 zero.

V. OUTPUT QUANTIZATION NOISE POWER

Different models have been proposed to estimate the power of the quantization noise at the output of a system [7]–[10]. 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 [8], the output quantization noise e_y is the sum of the contributions of the N QNSs e_i

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

where h_i corresponds to the time-varying impulse response of the system H_i between e_i and the output y.

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 [8]. From (13), the expression of P_{e_y} is

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

with

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

$$L_{ij} = \sum_{t=0}^{\infty}\sum_{v=0}^{\infty} E\left[h_i(t,n)h_j(v,n)\right]$$
$$M_{ij} = \sum_{t=0}^{\infty} E\left[h_i(t,n)h_j(t,n)\right] \quad (14)$$

and where K_i, L_{ij}, and M_{ij} are constant terms depending only on 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 datum x_i. The quantization of a unique datum P times leads to P QNSs and to $P(P-1)$ nonnull terms M_{ij}.

Compared to existing methods, a new term $M_{ij} \cdot \mathrm{cov}(E_i, E_j)$ is introduced. The covariance $\mathrm{cov}(E_i, E_j)$ is not equal to zero when both quantization processes $\mathcal{Q}_i$ and $\mathcal{Q}_j$ quantize the same data and when the delay inside the systems H_i and H_j is the same because the QNSs are white. Otherwise, $\mathrm{cov}(E_i, E_j)$ is equal to zero. The terms M_{ij} are not computed when $\mathrm{cov}(E_i, E_j)$ is always equal to zero. Thus, the increase in time due to the computation of M_{ij} depends on the number of QNSs, which can be correlated. As $M_{ij} \cdot \mathrm{cov}(E_i, E_j)$ can be positive or negative, if the quantization correlation is not considered, the quantization noise power can be overestimated or underestimated.

This model is valid for systems made up of smooth operations and if the input signals satisfy the conditions associated to the model by Widrow and Kollár [3]. An operation is smooth if its associated function is differentiable.

VI. Experiments

In this section, the estimations obtained with the proposed models are compared with the results obtained by the equivalent simulations, which are considered as the reference. To obtain relevant statistics, 10^7 samples are used to simulate the systems. The system is simulated in fixed point and in double-precision floating point. The error e_i associated with a datum x_i is obtained by subtracting the floating-point values and the fixed-point values. The floating-point simulation is considered as the reference as it is a reliable approximation of the infinite precision. Indeed, the error due to floating-point arithmetic can be neglected compared to the error due to the fixed point when the data WLs are small.

A. Correlation Between QNSs

In this section, the quality of the proposed model to estimate the covariance given in (10) and in Table II is evaluated. For the sake of clarity, ρ, the correlation coefficient between the QNSs e_1 and e_2 resulting from the quantization of x_0 as in Fig. 1, is analyzed. ρ is determined from the covariance $\mathrm{cov}(E_1, E_2)$ and the standard deviations σ_1 and σ_2 of noises e_1 and e_2, and it is defined as

$$\rho = \begin{cases} \frac{\mathrm{cov}(E_1,E_2)}{\sigma_1 \cdot \sigma_2}, & \text{if } \sigma_1 \neq 0 \text{ and } \sigma_2 \neq 0 \\ 0, & \text{else.} \end{cases} \quad (15)$$

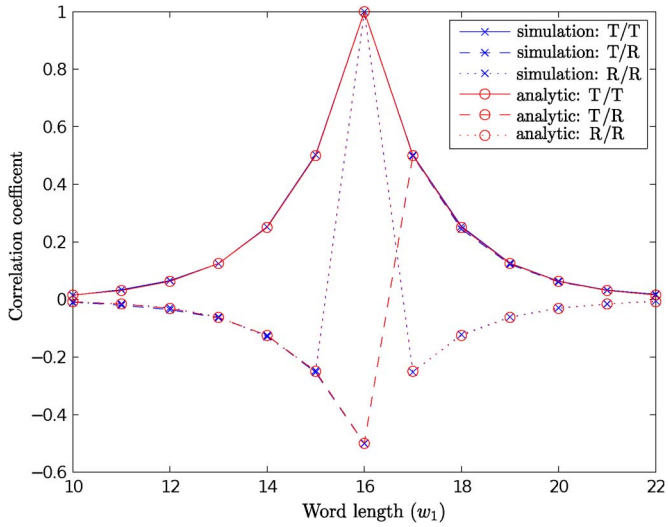

Fig. 3. Correlation coefficient ρ obtained with the proposed model and by simulations for different quantization modes and for different values of the fractional part WLs w_1.

The correlation coefficient ρ is evaluated for three combinations of rounding modes and for different fractional part WLs w_1 and w_2 of the data x_1 and x_2, respectively. w_2 is set to 16 bits, w_1 varies from 10 to 22 bits, and x_0 is considered as a continuous amplitude datum (i.e., $w_0 = +\infty$). In the first case, Q_1 and Q_2 are both truncations (T/T). In the second case, Q_1 is a truncation, and Q_2 is a rounding (T/R). In the third case, Q_1 and Q_2 are rounding (R/R).

The values of the correlation coefficients obtained with the proposed model and by simulations are shown in Fig. 3 for different values of the fractional part WLs w_1. The results show that the difference between the proposed model and the reference is very small. The maximal relative error of the proposed estimation is below 1.5%.

The amplitude of the correlation coefficient is maximal when the two WLs w_1 and w_2 are equal. In this case, $|\rho|$ is equal to one when the two quantization modes are identical, and $|\rho|$ is equal to 1/2 when the two quantization modes are different. For $Q_1 = T$ and $Q_2 = R$, $|\rho|$ is equal to 1/2 when $w_1 = w_2$ or $w_1 = w_2 - 1$. The correlation coefficients reduce when the difference between the WLs w_1 and w_2 increases.

B. Estimation of the Output Quantization Noise Power

In this section, the quality of the proposed approach to estimate the output quantization noise with (14) is evaluated and compared with the quality of a classical approach, which does not consider the correlation [in this case, $\mathrm{cov}(E_i, E_j) = 0 \forall i, j$ in (14)]. To illustrate the benefit of the proposed approach to improve the estimation quality of P_{e_y} compared to a classical approach, the computation of a third-order polynomial is considered. This polynomial is $p(x) = 0.99998 - 0.99765 \cdot x + 0.949615 \cdot x^2 - 0.63603 \cdot x^3$. The expression of the output quantization noise can be found in [11]. For data having multiple quantizations, half of the quantization is carried out with a WL set to 30 bits and the remaining quantizations are carried out with a WL w_T varying from 25 to 35 bits. To analyze

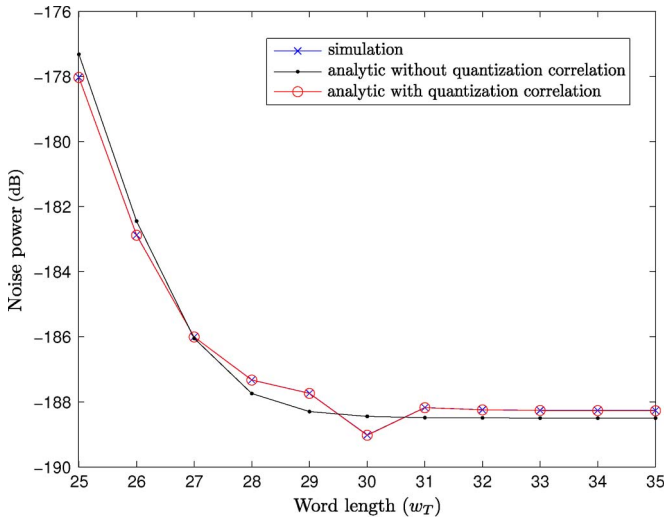

Fig. 4. Accuracy degradation for the polynomial $p(x)$ computation.

correlation effects, the WL of the other data is set such as no bit is eliminated. Quantization modes are selected randomly.

The real values of P_{e_y}, obtained by simulation, and the estimation of P_{e_y}, obtained with the proposed approach and with a classical approach, are shown in Fig. 4 for different WLs w_T. The results show that the estimations obtained with the proposed approach and the reference are very close. In this example, the maximal value of the relative estimation error (REE) is lower than 1% with the proposed approach and can reach up to 12% for a classical approach.

The REE is evaluated on different benchmarks. Let R_w denote the REE of the proposed approach and R_{wo} denote the REE of a classical approach. For the different benchmarks, a digital signal processor-like architecture is considered. The WL of the multiplication input is set to 16 bits. The WL of the multiplication output and the addition input and output is set to 32 bits. To evaluate R_w and R_{wo}, N_{cq} different combinations of quantization modes are tested. The number of correlated QNSs N_c is given and can be compared to N, the number of QNSs in the application. The execution time to obtain the expression of the output quantization noise power is measured for the proposed approach and for the classical approach. Let $I_{w/wo}$ denote the increase in time of the proposed approach compared to the classical approach. The mean and maximal values of R_w and R_{wo} are presented in Table III for six applications, which are a third-order polynomial computation with the direct form (Poly. 3 direct) and the Horner scheme (Poly. 3 Horner), a fifth-order polynomial computation with Horner scheme (Poly. 5 Horner), an $L2$-norm of an l-length vector $(l = 4)$, a second-order Voltera filter, and a function approximation. The function $1/(1 + x)$ is approximated in the interval [0, 1] with four third-order polynomials as described in [12].

Compared to the classical approach, the proposed estimation leads to an accurate estimation. The maximal value of R_w is 1.9% instead of 84% for R_{wo}. For the polynomial computation, when the correlation is not taken into account, the overestimation leads to an energy consumption overhead between 10% and 18% depending on the accuracy constraint. As shown by the results, the increase in time is limited and is less than 8.9% for the different tested applications. Indeed, the number of terms

TABLE III
REE FOR THE OUTPUT QUANTIZATION

	N_{cq}	N	N_c	R_{wo} mean	R_{wo} max	R_w mean	R_w max	$I_{w/wo}$
Poly. 3 direct	64	14	6	9.3%	40.1%	0.4%	1.3%	+6.4%
Poly. 3 Horner	8	9	3	12.7%	30.6%	0.1%	0.2%	+7.4%
Poly. 5 Horner	32	15	5	18.1%	79.9%	0.1%	1.4%	+8.9%
Function approx.	256	56	24	22.4%	84.1%	1.3%	1.9%	+1.3%
Vector L2-norm	256	15	8	7.1%	49.8%	0.23%	0.9%	+3.2%
Volterra filter	256	42	20	14.3%	63.7%	0.3%	1.4%	+0.7%

M_{ij} to compute depends on N_c and not on N. Moreover, each M_{ij} term is less complex than the term L_{ij}. The proposed approach allows improving significantly the estimation of the accuracy (P_{e_y}) compared to the classical approach, with a negligible increase in the computation time.

VII. 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 datum have been proposed in this brief. The model considers the number of eliminated bits and the quantization mode. A maximal relative error of the proposed estimation below 2% is reported. This model extends existing analytical methods with a negligible increase in the complexity of the evaluation. 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.

REFERENCES

[1] G. Constantinides, P. Cheung, and W. Luk, "Wordlength optimization for linear digital signal processing," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 22, no. 10, pp. 1432–1442, Oct. 2003.
[2] S. Parker and P. Girard, "Correlated noise due to roundoff in fixed point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, no. 4, pp. 204–211, Apr. 1976.
[3] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[4] A. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization error to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 5, pp. 442–448, Oct. 1977.
[5] G. Constantinides, P. Cheung, and W. Luk, "Truncation noise in fixed-point SFGs," *Electron. Lett.*, vol. 35, no. 23, pp. 2012–2014, Nov. 1999.
[6] D. Menard, D. Novo, R. Rocher, F. Catthoor, and O. Sentieys, "Quantization mode opportunities in fixed-point system design," in *Proc. EUSIPCO*, Aalborg, Denmark, Aug. 2010, pp. 542–546.
[7] C. Shi and R. Brodersen, "A perturbation theory on statistical quantization effects in fixed-point DSP with non-stationary inputs," in *Proc. IEEE ISCAS*, Vancouver, BC, Canada, May 2004, pp. 373–376.
[8] R. Rocher, D. Menard, P. Scalart, and O. Sentieys, "Analytical accuracy evaluation of fixed-point systems," in *Proc. EUSIPCO*, Poznan, Poland, Sep. 2007, pp. 999–1003.
[9] P. Fiore, "Efficient approximate wordlength optimization," *IEEE Trans. Comput.*, vol. 57, no. 11, pp. 1561–1570, Nov. 2008.
[10] G. Caffarena, J. López, A. Fernandez, and C. Carreras, "SQNR estimation of fixed-point DSP algorithms," *EURASIP J. Adv. Signal Process.*, vol. 2010, pp. 1–13, Feb. 2010.
[11] J. Naud, "Numerical accuracy evaluation for polynomial computation," INIRA, Tech. Rep. 7878, Feb. 2012.
[12] N. Brisebarre, J.-M. Muller, and A. Tisserand, "Computing machine-efficient polynomial approximations," *ACM Trans. Math. Softw.*, vol. 32, no. 2, pp. 236–256, Jun. 2006.