ACCURACY EVALUATION OF FIXED-POINT LMS ALGORITHM

Romuald ROCHER, Daniel MENARD, Olivier SENTIEYS, Pascal SCALART

ENSSAT/IRISA
University of Rennes I
6 rue de Kérampont, BP447
22300 Lannion

ABSTRACT

The implementation of adaptive filters with fixed-point arithmetic requires to evaluate the computation quality. The accuracy may be determined by calculating the global quantization noise power in the system output. In this paper, a new model for evaluating analytically the global noise power in the LMS algorithm and in the NLMS algorithm is developed. Two existing models are presented, then the model is detailed and compared with the ones before. The accuracy of our model is analyzed by simulations.

1. INTRODUCTION

The aim of adaptive filters is to estimate a sequence of scalars from an observation sequence filtered by a system in which coefficients vary. These coefficients converge towards the optimum coefficients which minimize the mean square error (MSE) between the filtered observation signal and the desired sequence. This type of filters is used in different fields such as noise cancellation, equalization, linear prediction and channel estimation. The different algorithms existing for adaptive filtering are mainly classified in two types : Recursive Least Square (RLS) and Least Mean Square (LMS). Nevertheless, the LMS algorithm is the most used in applications because its implementation in embedded systems is more simple than the RLS algorithm. However, the use of fixed-point arithmetic is required. This type of arithmetic is less expensive in terms of cost and power consumption than the floating-point arithmetic. But, the coding of fixed-point data introduces an error called quantization noise. These different quantization noise sources are propagated in the system and lead to an output quantization noise. The power of this quantization noise is determined in order to compute the signal to quantization noise ratio (SQNR). The knowledge of the analytical expression of the SQNR allows to determine the fixed-point format of the system data for a given SQNR minimal value. Some different models have been proposed for the LMS algorithm in [2] and [4] but they are valid only for convergent rounding. So, the aim of this paper is to find an analytical expression for the noise power in the LMS algorithm for all types of quantization (convergent and non-convergent rounding, truncation). The truncation is the most common mode used in embedded systems. Indeed, its implementation requires no additional hardware.

This paper is organized as follows. The basic properties of the LMS algorithm are first recalled in section 2. Then, the existing models are detailed and their limits discussed. In section 3, the developed model is explained and the method is clarified. A model is also shown for the NLMS algorithm. Finally, in section 4, the quality of our model is evaluated through different experimentations. This allows to underline its validity and to compare its results with the others.

2. RELATED WORK

The LMS adaptive algorithm addresses the problem of estimating a sequence of scalars y_n from a sequence of length N vectors x_n [3]. The linear estimate of y_n is $w_n^t x_n$ where w_n is a length N weight vector which converges to the optimal vector w^* in the mean-square error (MSE) sense. This optimal vector is equal to $w^* = R^{-1}p$ where $R = E(x_n x_n^t)$ and $p = E(x_n y_n)$. The vector w_n is updated according to the equation

$$w_{n+1} = w_n + \mu x_n (y_n - w_n^t x_n) \qquad (1)$$

where μ is a positive constant representing the adaptation step. The maximum value of μ to ensure stability is equal to $2/\lambda_{max}$ with λ_{max} the maximum eigenvalue of R.

In the model presented in [4], the expression of the MSE in fixed-point implementation is determined. In that case, the MSE is the second order moment of the difference between the desired signal in infinite precision and the quantified computed output. Thus the MSE is given by the sum of the mean square error in infinite precision and of the noise power which is composed of three terms.

- The error due to input data quantization filtered by the coefficients.
- The input sequence filtered by the deviation of the filter coefficients from their exact values in infinite precision.
- The noise inside the filter due to fixed-point arithmetic operations.

The expression of these three terms has been determined. The two first terms are expressed as in the case of linear systems. The last term is more complex. A recurrence is determined on the deviation of the coefficients. But, few hypothesis are made to simplify the complexity of the equation. The final result is complicated (term of second order in μ^2).

The model detailed in [2] deals with the MSE like the one before but the method is different. This model also determines the MSE in the case of fixed-point implementation. Only two noises are considered. They correspond to the noise inside the filter due to arithmetic operations and the noise in the multiplication between the input signal and the error $\mu x_n(y_n - w_n^t x_n)$. A recurrence is developed on the deviation between the coefficients and their optimum value. This recurrence, calculated before in the case of infinite precision, is injected in the equation of the MSE. Then,

the MSE is determined in the case of finite precision. This expression leads to the same result as in [4] if the input noise is not considered.

In these models, the means of the quantization noises are considered as equal to zero. However, this equality is only valid in the case of quantization by convergent rounding. The mean of a noise due to the quantization of a discrete amplitude signal in the case of classic rounding is given by [5]

$$m_b = \frac{q}{2} 2^{-k} \tag{2}$$

where q is the quantization step and k the number of eliminated bits. The model proposed in the next section is developed for all types of quantization.

3. DEVELOPED MODEL

The new developed model is described in this part. The analysis of the error is done at the steady-state, once the filter coefficients have converged. Let x'_n be the input signal after quantization and y'_n the quantified desired signal.

$$\begin{aligned} x'_n &= x_n + \alpha_n \\ y'_n &= y_n + \beta_n \end{aligned} \tag{3}$$

where α_n and β_n are quantization noises with respectively means m_{α_n} and m_{β_n} and variance $\sigma^2_{\alpha_n}$ and $\sigma^2_{\beta_n}$. The filter coefficient vector is written as

$$w'_n = w_n + \rho_n \tag{4}$$

where ρ_n is the error vector of length N due to the quantization effects. This noise can not be considered as the noise due to the quantization of a signal. The error in finite precision is given by

$$e'_n = y'_n - w'^t_n x'_n - \eta_n \tag{5}$$

with η_n the global noise in the inner product $w'^t_n x'_n$. Moreover, the updated coefficients expression becomes

$$w'_{n+1} = w'_n + \mu e'_n x'_n + \gamma_n \tag{6}$$

where γ_n is the noise associated with the term $\mu e'_n x'_n$ and depends on the way the filter is computed.

So, the error is measured at the filter output. The power of the error between filter output in finite precision and in infinite precision is determined. It is composed of three terms.

$$E(b_y)^2 = E(\alpha^t_n w_n)^2 + E(\rho^t_n x_n)^2 + E(\eta^2_n) \tag{7}$$

3.1. Expression of the term $E(\alpha^t_n w_n)^2$

At the steady-state, the vector w_n can be approximated by the optimum vector w^*. So the term $E(\alpha^t_n w_n)^2$ is equal to $|w^*|^2(m^2_{\alpha_n} + \sigma^2_{\alpha_n})$ with $|w^*|^2 = \sum w^{*2}_i$. It corresponds to the input noise filtered by the optimum coefficients.

3.2. Expression of the term $E(\eta^2_n)$

The second term $E(\eta^2_n)$ depends on the specific implementation chosen for the computation of the filter output (filtered data). If the N products are computed in double precision (no multiplication bit are eliminated), this noise corresponds to a noise on the output. On the other case, it corresponds to the sum of the N product noises. These two first terms are the same as in linear systems and can be evaluated very easily as proposed in [1]

3.3. Expression of the term $E(\rho^t_n x_n)^2$

The last term $E(\rho^t_n x_n)^2$ is more complex since ρ_n is not a quantization noise. With I_N the length N identity matrix, it can be demonstrated that the noise ρ_n is given by:

$$\rho_{n+1} = F_n \rho_n + b_n \tag{8}$$

where $F_n = I_N - \mu x_n x^t_n$

$b_n = -\mu x_n w^t_n \alpha_n + \mu x_n(\beta_n - \eta_n) + \mu \alpha_n e_n + \gamma_n$

Introducing the matrice $P_n = E(\rho_n \rho^t_n)$, the equation 9 can be obtained

$$\begin{aligned} P_{n+1} &= E(b_n b^t_n) + E(F_n \rho_n b^t_n) \\ &+ E(b_n \rho^t_n F_n) + E(F_n \rho_n \rho^t_n F_n) \end{aligned} \tag{9}$$

This expression is composed by four terms which are developed in the next paragraphs.

The term $E(b_n b^t_n)$ can be estimated by approximating b_n by γ_n (γ_n is the noise associated with the term $\mu e'_n x'_n$). Indeed, b_n is composed of several terms of which γ_n is the most important since the other terms are products of weak power terms. So, the term $E(b_n b^t_n)$ is approximated by

$$E(b_n b^t_n) = E(\gamma_n \gamma^t_n) \tag{10}$$

For the term $E(F_n \rho_n b^t_n)$, b_n is replaced by γ_n.

$$E(F_n \rho_n b^t_n) = E(\rho_n) E(\gamma^t_n) - \mu E(x_n x^t_n \rho_n) E(\gamma^t_n) \tag{11}$$

However, from equation 8 at the steady-state, the term $E(x_n x^t_n \rho_n)$ is equal to

$$\mu E(x_n x^t_n \rho_n) = E(\gamma_n) \tag{12}$$

The computation of equation 11 requires the knowledge of $E(\rho_n)$. The recurrence $\rho_{n+1} = F_n F_{n-1} \rho_{n-1} + F_n b_{n-1} + b_n$ can be developed as follows

$$E(\rho_n) = (E(F_n) + E(F_n F_{n-1}) ....) E(\gamma_n) \tag{13}$$

But calculating this series is a very tedious task and can only be done by simulation. So an hypothesis is made. We suppose ρ_n and x_n non-correlated. This hypothesis will be discussed in section 4.3. Thus, with equation 12, the term $E(\rho_n)$ is equal to

$$E(\rho_n) = \frac{R^{-1} E(\gamma_n)}{\mu} \tag{14}$$

where R is the autocorrelation matrix of the input signal. Finally, the next expression is obtained

$$E(F_n \rho_n b^t_n) = \frac{R^{-1} E(\gamma_n) E(\gamma^t_n)}{\mu} - E(\gamma_n) E(\gamma^t_n) \tag{15}$$

With the same method, the term $E(b_n \rho^t_n F_n)$ can be computed with the following expression

$$E(b_n \rho^t_n F_n) = \frac{E(\gamma_n) E(\gamma^t_n) R^{-1}}{\mu} - E(\gamma_n) E(\gamma^t_n) \tag{16}$$

If the term in μ^2 is neglected, $E(F_n \rho_n \rho^t_n F_n)$ can be written as

$$E(F_n \rho_n \rho^t_n F_n) = P_n - \mu(R P_n) - \mu(P_n R) \tag{17}$$

At the steady-state, $P_{n+1} = P_n$. Thus, by introducing the expressions 10,15,16 and 17 in the equation 9 and using the trace operator, the following expression is obtained

$$\begin{aligned} 2\mu Tr(RP_n) &= 2\frac{Tr(E(\gamma_n)E(\gamma_n^t)R^{-1})}{\mu} + Tr(E(\gamma_n\gamma_n^t)) \\ &- 2Tr(E(\gamma_n)E(\gamma_n^t)) \end{aligned} \quad (18)$$

Moreover, the next equation can be written

$$Tr(RP_n) = E(\rho_n^t x_n)^2 \quad (19)$$

Thus, developing the others terms of equation 18, the term $E(\rho_n^t x_n)^2$ can be obtained from equation 20

$$E(\rho_n^t x_n)^2 = m_{\gamma_n}^2 \frac{\sum_{i=1}^{N}\sum_{k=1}^{N}(R_{ki}^{-1})}{\mu^2} + \frac{N(\sigma_{\gamma_n}^2 - m_{\gamma_n}^2)}{2\mu} \quad (20)$$

This term corresponds to the input signal filtered by the deviation on the coefficients.

3.4. The global noise power

According to the previous analyse, the global noise power Pb can be written as

$$\begin{aligned} Pb &= |w^*|^2(\sigma_{\alpha_n}^2 + m_{\alpha_n}^2) + (m_{\eta_n}^2 + \sigma_{\eta_n}^2) \\ &+ m_{\gamma_n}^2 \frac{\sum_{i=1}^{N}\sum_{k=1}^{N}(R_{ki}^{-1})}{\mu^2} + \frac{N(\sigma_{\gamma_n}^2 - m_{\gamma_n}^2)}{2\mu} \end{aligned} \quad (21)$$

This model is presented for quantization by truncation and rounding. In the case of rounding, the means of η_n and γ_n are not equal to zero since they represent the quantization of a discrete signal. From equation 21, m_{α_n} is the only term to be equal to zero in rounding quantization.

However, if the implementation is made in convergent rounding quantization, the means of η_n and γ_n are equal to zero leading to

$$Pb = |w^*|^2\sigma_{\alpha_n}^2 + \sigma_{\eta_n}^2 + \frac{N(\sigma_{\gamma_n}^2)}{2\mu} \quad (22)$$

In that case, the expression is quite similar to the model in [2] and [4] but more tractable (no term of second order in μ^2).

3.5. The NLMS algorithm

In this section the NLMS algorithm is considered. The expression of the updated equation is equal to

$$w_{n+1} = w_n + \frac{\mu}{x_n^t x_n}(y_n - w_n^t x_n)x_n \quad (23)$$

The term $\frac{1}{x_n^t x_n}$ is a normalization term which let μ be between 0 and 2. To prevent from a division in fixed-point arithmetic, the term is approximated by a power of two which greatly simplify the implementation. Thus, the division is equivalent to a shift of some bits. This method does not introduce a new noise. Here, that case of normalization is considered. The mean of $\frac{1}{x_n^t x_n}$ can be approximated by $\frac{1}{N(\sigma_x^2+m_x^2)}$ where m_x and σ_x^2 are the mean and the variance of the input signal. The noise power is given by replacing μ by $\frac{\mu}{N(\sigma_x^2+m_x^2)}$ in equation 21

$$\begin{aligned} Pb &= |w^*|^2(\sigma_{\alpha_n}^2 + m_{\alpha_n}^2) + (m_{\eta_n}^2 + \sigma_{\eta_n}^2) \\ &+ m_{\gamma_n}^2 N^2(\sigma_x^2 + m_x^2)^2 \frac{\sum_{i=1}^{N}\sum_{k=1}^{N}(R_{ki}^{-1})}{\mu^2} \\ &+ \frac{N^2(\sigma_x^2 + m_x^2)(\sigma_{\gamma_n}^2 - m_{\gamma_n}^2)}{2\mu} \end{aligned} \quad (24)$$

4. ACCURACY

In this section, simulations are made to analyse the accuracy of our model for evaluating the fixed-point noise power in a LMS algorithm. The input signal chosen is an AR(1) process given by

$$x_{n+1} = \beta x_n + u_n \quad (25)$$

where u_n is a white noise with zero mean, with variance σ_u^2 and $\beta \in [0,1[$. So, the input signal can be very correlated ($\beta \longrightarrow 1$) or not ($\beta \longrightarrow 0$). For these simulations, tests are made for quantization by truncation and rounding. The relative error between the noise power obtained with simulations and the estimated noise power with our model is computed. For these simulations, μ can vary from 0 to $0.6\mu_{max}$. Indeed, the filter coefficients convergence is ensured if $\mu < \mu_{max}$. But, in reality, to be sure that the coefficients do not diverge, a limit of $0.6\mu_{max}$ is chosen. However, as μ_{max} depends on the length filter, μ is represented by $\frac{\mu}{\mu_{max}}$ to be normalized. Moreover, the length filter N varies from 1 to 32. The input signal is fairly correlated ($\beta = 0.5$) for the two simulations.

4.1. Evaluation of the model accuracy

Figure 1 shows the relative error between the real and the estimated noise power in rounding quantization. This relative error is smaller than 25% which is a good result since it represents a difference of 1 dB between the output quantization noise power estimated by simulation and the power given by our model. So, this new developed model is valid for the case of quantization by rounding.

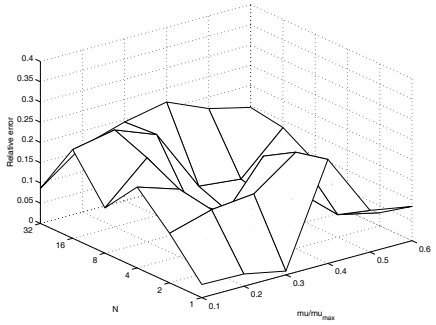

FIG. 1 –. Relative error for rounding quantization

Figure 2 represents the relative error in the case of quantization by truncation. As in the rounding quantization case, our model leads to an accurate estimation of the noise power. The relative error is smaller than 20%.

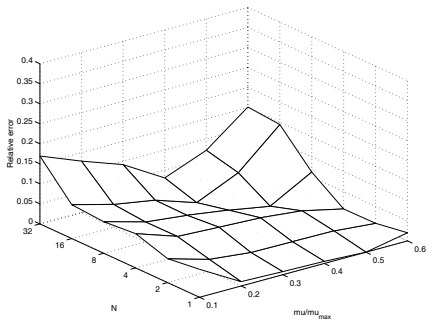

FIG. 2 –. Relative error for truncation quantization

4.2. Comparaisons with the other models

Our model is valid but we need to compare it with the two others models presented before [2, 4]. For this simulation in figure 3, N is fixed at 16 and μ varies from 0 to $0.8\mu_{max}$. This test is made in the case of quantization by rounding for which the two other proposed models are presented.

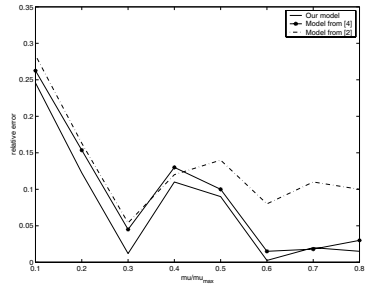

FIG. 3 –. Comparaison between the three models

The model in [2] is the less accurate because it does not integrate the input noise. Our model has better results than the model in [4]. Our simplifications are not prejudicial for the estimation quality. Moreover, the terms we have added in our model let us have a better result. In some cases, the models [2] and [4] are not accurate because they do not integrate the terms $m_{\eta_n}^2$, $m_{\gamma_n}^2 \frac{\sum_{i=1}^{N}\sum_{k=1}^{N}(R_{ki}^{-1})}{\mu^2}$ and $-\frac{Nm_{\gamma_n}^2}{2\mu}$ for a rounding quantization.

4.3. Validity of the hypothesis

In this part, the validity of the hypothesis made in section 2 is examined. It corresponds to the non-correlation between ρ_n and x_n. With this hypothesis, the following equality is obtained

$$E(\rho_n) = \frac{R^{-1}E(\gamma_n)}{\mu} \tag{26}$$

An alternative to this hypothesis is to develop the recurrence and to compute the different terms by simulation as follows

$$E(\rho_n) = (E(F_n) + E(F_nF_{n-1})....)E(\gamma_n) \tag{27}$$

In this case, this approach takes into account the correlation between ρ_n and x_n. However, the series $(E(F_n) + E(F_nF_{n-1})....)$ is very difficult to determine and is calculated by simulation here. In figure 4, the two methods to estimate $E(\rho_n)$ are compared. The relative error between the real value of $E(\rho_n)$ and its estimation by 27 and 26 is presented. The length N is 16, $\mu = 2^{-7}$ and β varies from 0 (white noise) to 0.95 (very correlated signal).

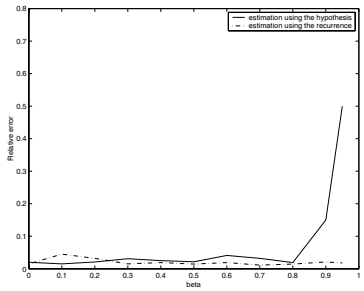

FIG. 4 –. Relative error between $E(\rho_n)$ and its estimation

This hypothesis of non-correlation between ρ_n and x_n is valid for a slightly correlated signal, but when the input signal is very correlated, the hypothesis is no more valid. On the other hand, the method using the recurrence leads to a very small relative error for all types of correlation for the input signal. However, the difficulty is to determine the order to stop the series. For our simulations, we stop the series after 200 terms to have a satisfying result. But, even if this method gives good results, it is very tedious and very expensive in computing time. So, some works need to be carried out to have a global method to estimate $E(\rho_n)$ in all cases.

5. CONCLUSION

In this paper, a new model for evaluating the noise power in a fixed-point implementation of the LMS algorithm is presented. This approach has for main advantage to be more tractable than the models [2] and [4] and to be valid for all types of quantization. This model can be improved through the determination of $E(\rho_n)$ since the two methods (equations 26 and 27) can be improved. A global model must be developed for this term. Nevertheless, further studies have to be carried out in order to develop this methodology for all types of systems and particularly, non-linear systems.

6. REFERENCES

[1] D. Ménard, O. Sentieys, "A methodology for evaluating the precision of fixed-point systems", *IEEE Conference on Acoustics, Speech and Signal Processing, vol 3, 2002.*

[2] M. Bellanger, "Digital Processing of Signals", *John Wiley and Sons, 3^{rd} edition, 2001.*

[3] S. Haykin, "Adaptive Filter Theory", *Englewood Cliffs, NJ:Prentice-Hall, 2^{nd} edition, 1991.*

[4] C. Caraiscos, B. Liu, "A Roundoff Error Analysis of the LMS Adaptive Algorithm", *IEEE Transactions Acoustic, Speech, Signal Processing, vol ASSP-32, no.1, february 1984.*

[5] G. Constantinides, P. Cheung and W. Luk "Truncation Noise in Fixed-Point SFGs", *IEE Electronic Letters, 35(23) : 2012-2014, november 1999.*