20th IMEKO TC4 International Symposium and
18th International Workshop on ADC Modelling and Testing
Research on Electric and Electronic Measurement for the Economic Upturn
Benevento, Italy, September 15-17, 2014

Roundoff Errors in the Evaluation of the Cost Function in Sine Wave Based ADC Testing

Balázs Renczes, István Kollár

Budapest University of Technology and Economics, H-1117 Magyar Tudósok körútja 2., Hungary, renczes@mit.bme.hu/kollar@mit.bme.hu

***Abstract* – In this paper sine fitting algorithms are investigated for the purpose of ADC testing. The aim is to decide whether the minimum of the cost function (CF) has been reached. For this, two different types of algorithms, the conventional Levenberg-Marquardt and the genetic-type Differential Evolution methods are investigated in order to compare their optima. It is shown that due to roundoff errors the bottom of the cost function is fairly uneven for conventional number representations for the Maximum Likelihood method, hence the minimum can only be determined with decreased precision. Finally, a band is calculated in which solutions can be considered equivalent, since their CF difference is smaller than roundoff errors.**

***Keywords* – Sine wave fit, Cost Function Evaluation, Maximum Likelihood, ADC testing[1]**

I. INTRODUCTION

Analog-digital converters transform the signals of the analogue world into values which are discrete in time and amplitude. The precision of the conversion is crucial since it determines the quality of the signal processing which follows. IEEE Standard 1241 specifies the Four-Parameter Least Squares (LS) method as the testing process. This method determines sine wave parameters (amplitudes of cosine and sine, offset and frequency) for the sinusoidal input data so that Noise and Distortion *(NAD)* is minimal [1]:

$$x[n] = A \cdot \cos(2\pi f_0 n) + B \cdot \sin(2\pi f_0 n) + C \quad (1)$$

$$NAD = \sqrt{\frac{1}{M}\sum_{x=1}^{M}(x[n] - y[n])^2}, \quad (2)$$

where *x[n]* is the n^{th} sample of the fitted sine wave, *A,B*, *C* and f_0 are the parameters of the fitted wave: amplitudes of the cosine and sine components, offset and frequency, respectively. Finally, *y[n]* is the n^{th} sample in the digitized sample set. The Effective Number of Bits (*ENOB*) of a converter, driven to its full input range can be calculated by:

$$ENOB = N - log_2 \frac{NAD}{LSB/\sqrt{12}}, \quad (3)$$

where *N* is the bit number of the ADC under test [1]. The choice of LS method seems to be justified, since it maximizes the value of *ENOB* by minimizing the value of *NAD,* thus supplying the most attractive result for the manufacturers. However, it was shown in [2] that the Maximum Likelihood (*ML*) method may yield more precise estimators than the LS method. The ML process tries to fit a sine wave on the input data so that the logarithmic likelihood function

$$\ln L(p) = \sum_{i=1}^{M} \ln\left[P\left(Y(k) = y(k)\right)\right] \quad (4)$$

is maximal, where $P(Y(k) = l)$ is the probability, that the k^{th} value of random variable vector *Y* has the digital code of *l*, *y(k)* is the k^{th} element of the (digital) sample set and *M* is the number of samples [2]. The optimum is searched in a five-dimensional parameter space (*A, B, C,* f_0, σ), with σ being the standard deviation of the modelled noise at the input of the ADC. Similarly to the LS method, amplitudes of the cosine and the sine, offset and frequency are the parameters optimized, and the standard deviation of the input noise is also added to the parameter vector.

The optimum of the ML cost function can be calculated by numerical methods. For this purpose, a MATLAB toolbox has been developed [3],[4],[5]. In this tool the cost function (which is the negative log-likelihood function) is being optimized numerically by the Levenberg-Marquardt method. However, it cannot be reliably decided whether the minimum of the function has been reached. In order to verify the results, a different minimizing algorithm is to be implemented, thus the Differential Evolution algorithm, defined in [6], has also been evaluated.

II. INVESTIGATED METHODS

A. Description

The *Levenberg-Marquardt (LM)* method is a standard minimizing algorithm that makes use of the parameters' gradient vector and of the Hessian matrix which contains the second derivatives of the parameters. This can be considered as a modified version of *Gauss-Newton*

[1] This work has been supported by the University Student Union of the Budapest University of Technology and Economics

minimizer, which finds the optimum of a second order cost function in one step. However, when the cost function is not precisely of second order, or the derivatives can only be calculated with an error, this may lead us to an incorrect result. For this reason, a scaling factor λ is also introduced. If λ equals 0, the method is equivalent with the *Gauss-Newton* method, while if λ tends to ∞, we get the *steepest descent method*.

The other method investigated is the *Differential Evolution (DE)*, which is a genetic-type, population based, stochastic function optimizer. For this algorithm it is not necessary to have any information about derivatives of the cost function. Usually it is able to find the optimum value [6].

B. Results of the algorithms

The cost function of the ML method has been evaluated with the two different algorithms for 100,000 samples. DE has found lower cost function value, the relative difference between the results was, however, in the order of magnitude 10^{-13}, which might seem negligible but it has to be stated that it is much higher than the LSB (eps) of the double floating point number representation ($\approx 2 \cdot 10^{-16}$). In order to be able to determine which parameter causes this difference, the parameter space between the two optimum vectors has been divided into 1000 parts. Taking the two solution vectors, we selected one parameter and kept the other parameters constant, while the values of the cost function have been evaluated.

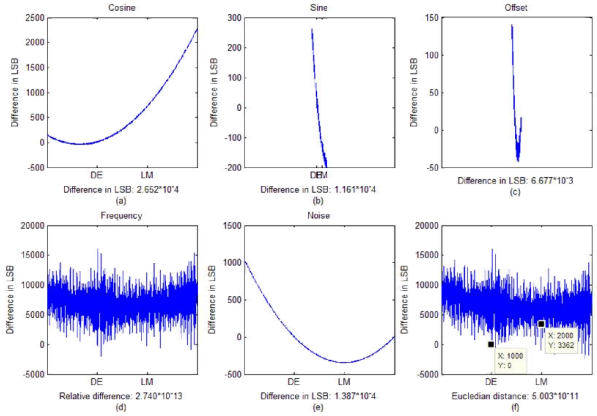

Figure 1: Cost function values in the vicinity of the Differential Evolution (DE) and the Levenberg-Marquardt (LM) optima along the parameters, holding the others constant(a-e) and along the straight line between the two parameter vectors (f)

In addition, the cost function was calculated along the straight line between the two parameter vectors. Since the difference between the two optima is fairly small, the results are plotted so that they represent the deviation from the DE optimum. The LSB on the horizontal axis is the LSB of the ADC, while on the vertical axis LSB is the resolution of the floating-point number representation (also known as *eps*) of the DE optimal cost function value. It can be seen from Fig. 1 that the difference is mostly caused by the frequency component. Furthermore, the bottom of the cost function seems to be rather noisy (Fig. 1d).

III. ERROR ANALYSIS

A. Enhanced-precision evaluation

In Eq. (1) the phase is calculated by $\varphi[n] = 2\pi f_0 n$. MATLAB guarantees that the evaluation error of sine computation is between $\pm eps$. However, if the phase itself is already calculated with roundoff error, the value of its sine will also be in error. Due to floating-point number representation, with increasing n, the roundoff error also increases, since the value of *LSB* (for double precision) is

$$LSB(x) = 2^{-52+\lfloor log_2|x| \rfloor}, \quad (5)$$

where $\lfloor x \rfloor$ is roundig x towards negative infinity. This means that for 100,000 samples the value of $LSB(\varphi)$ for the last sample is 2^{16} or 2^{17} times ($log_2\lfloor 100\,000 = 16 \rfloor$) greater than the $LSB(\varphi)$ of the first value – depending on the value of $2\pi f_0$.

To illustrate the effect of the roundoff error, values of a sine wave has been calculated in 100,000 points both with double and single precision. Since the mantissa of the double precision is much longer, than that of the single precision (53 vs. 24 bits), the result of it may be considered precise, as reference value. The error of sine calculation for single precision is shown in Fig. 2.

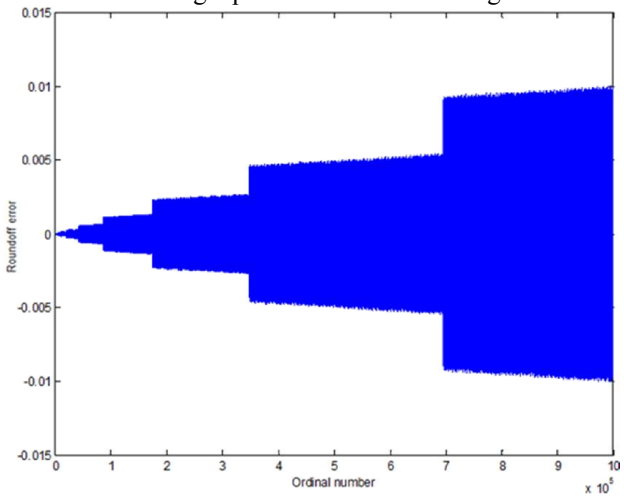

Figure 2: Illustration of roundoff error for single precision number representation

If the error, illustrated in Fig. 1, is also caused by roundoff, it can significantly be reduced by using increased-number representation precision. Therefore, for the phase evaluation, multiplication and addition have been implemented in MATLAB for 75-bit mantissa floating point numbers.

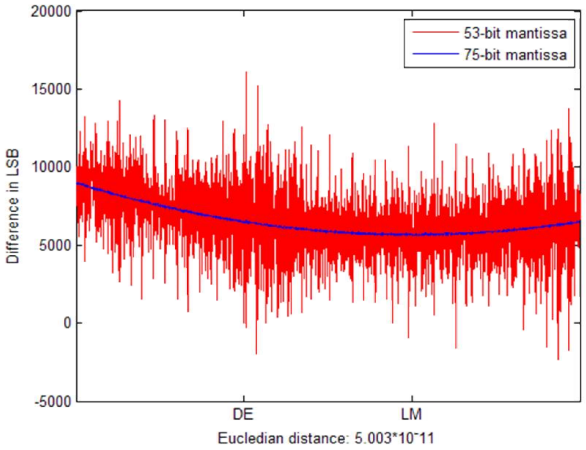

Figure 3: Cost function value changes in the vicinity of the optima with different mantissa precisions

Exploiting the periodicity of the sine, integer multiples of 2π has been subtracted from each phase. Although sine calculation itself has not been written for the extended mantissa, by evaluating the phases more precisely and mapping into $[-\pi,\pi)$, sine calculation can also be executed with the required precision. Fig. 3 clearly shows that the error (the 'noise') can be significantly reduced. What is more, the LM method is shown to yield a result closer to the real optimum. It is an interesting issue, however, how the LM could find the minimum under these conditions. In Fig. 4 it can be seen that the derivative with respect to the frequency which caused the main difference, is straight in the vicinity of the LM optimum and equals to zero only close to the optimum. Thus, the algorithm "knows" that there is a possibility to decrease, although the evaluation of the CF is in error. In this way, the LM algorithm can easier find the real optimum, while the DE did not make any use of the gradient information.

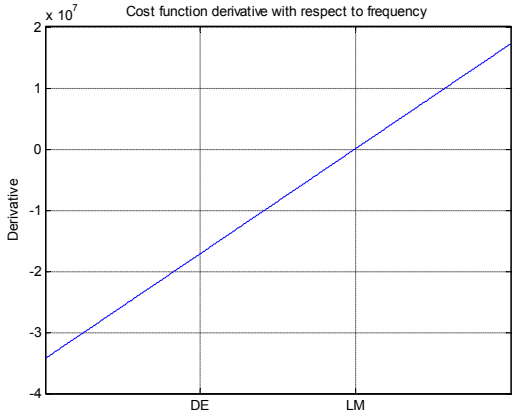

Figure 4: Cost function derivative values with respect to frequency in the vicinity of the two optima

However, it has to be mentioned that in case the LM method had reached a noisy local minimum, it would have stuck there, since the algorithm was written so that it prevents from increasing the cost function. As a consequence we can say that finding the optimum was slightly 'lucky' since it did not get stuck anywhere before reaching the minimum. Another interesting question is how we can determine if two solutions are equivalent, since it is clear from Fig. 1 (f) that for double representation solutions between the two optima found by DE and LM cannot be distinguished.

B. Worst-case analysis

Let us assume that the error of phase determination is $\Delta\varphi$. In this case the error of the instantaneous amplitude – using notations of Eq. (1) – is

$$\begin{aligned}\Delta x = A\cdot cos\varphi + B\cdot sin\varphi - A\cdot cos(\varphi+\Delta\varphi) - \\ -B\cdot sin(\varphi+\Delta\varphi) = A\cdot cos\varphi + B\cdot sin\varphi - \\ -A\cdot cos\varphi\cdot cos(\Delta\varphi) + A\cdot sin\varphi\cdot sin(\Delta\varphi) - \\ -B\cdot sin\varphi\cdot cos(\Delta\varphi) - B\cdot cos\varphi\cdot sin(\Delta\varphi)\end{aligned} \quad (6)$$

If $\Delta\varphi$ is small ($<10^{-10}$), then $cos(\Delta\varphi)\approx 1$ and $sin(\Delta\varphi)\approx\Delta\varphi$ (using double representation they cannot be represented more precisely, either), so Eq. (6) can be further transformed:

$$\begin{aligned}\Delta x = A\cdot cos\varphi + B\cdot sin\varphi - A\cdot cos\varphi\cdot 1 + \\ +A\cdot sin\varphi\cdot(\Delta\varphi) - B\cdot sin\varphi\cdot 1 - B\cdot cos\varphi\cdot(\Delta\varphi) = \\ = \Delta\varphi\cdot[A\cdot sin(\varphi) - B\cdot cos(\varphi)]\end{aligned} \quad (7)$$

$\Delta\varphi$ is the calculation error, i.e., maximum $\pm 0{,}5\cdot LSB(\varphi)$.

$$|\Delta x_{max}| = |0{,}5\cdot LSB(\varphi)\cdot[A\cdot sin(\varphi) - B\cdot cos(\varphi)]| \quad (8)$$

Maximum and minimum values for x can be given:

$$\begin{aligned}x_{max} &= x + |\Delta x_{max}| \\ x_{min} &= x - |\Delta x_{max}|\end{aligned} \quad (9)$$

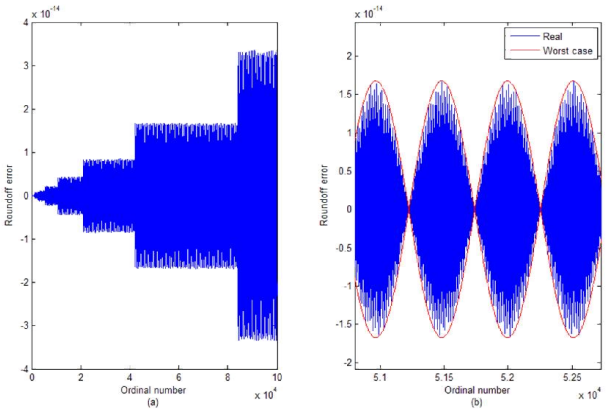

Figure 5: Roundoff error in the amplitude of the sine wave (a), and its comparison with the worst case limits (b)

The roundoff error is represented in Fig. 5. Real values were calculated using the 75-bit mantissa evaluation, which can be considered precise when compared to double precision. It can be seen that the real errors are lower, than $\pm\Delta x_{max}$. Each probability in Eq. (4) can be calculated for x, x_{max} and x_{min}. From these the maximum and minimum values can be chosen for each time instant. By this way the theoretical maximum and the minimum value of the cost function can be calculated. For the given sample set the worst case deviation band was $CF_{max} - CF_{min} = 2.31 \cdot 10^{-5}$. The maximum deviation of the 53-bit evaluation was $5.37 \cdot 10^{-7} (\sim 18\,000\ LSB)$ (see Fig. 3), which is about 45 times lower than this worst case deviation band.

C. Probability analysis

Although the maximum deviation in the measurement is in the band given by the worst case analysis, the result cannot be utilized in practice, since the tolerance band is too wide, allowing too large deviations. As Fig. 6 shows, the true computation error can be modelled as stochastic, hence it could be treated as random variable. The cost function is a result of several computation steps, whose error can be regarded as independent, so according to the Central Limit Theorem, the distribution of the cost function is approximately normal. Fig. 6 shows the histogram of the evaluation error in the vicinity of the optima (assuming that the 75-bit mantissa evaluation is precise). If a χ^2 goodness-of-fit test is executed for 700 bins, the null-hypothesis that the samples are from normal distribution cannot be rejected at 5% significance level.

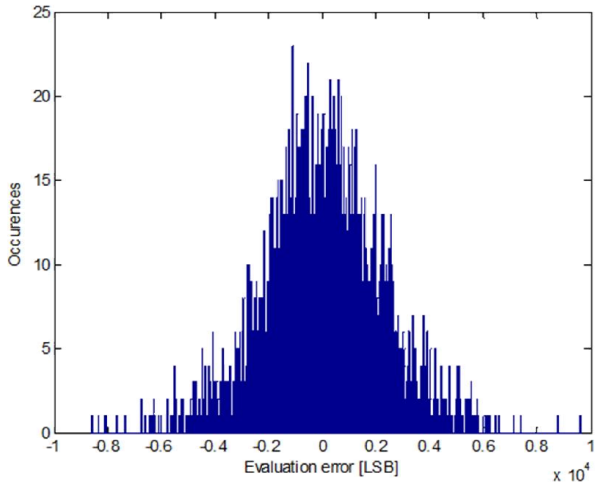

Figure 6: Histogram of the evaluation errors in the vicinity of the DE and LM optima

In Section III/A it was shown that the error was mostly caused by imprecise phase evaluation. It can be assumed that the roundoff noise of the evaluation is uniformly distributed between $\pm LSB(\varphi)/2$, according to the Pseudo Quantization Noise (PQN) model [7]. The variance of the evaluation can be calculated as

$$var\{\varphi\} = \frac{[LSB(\varphi)]^2}{12} \tag{10}$$

We can assume that errors are small, hence in further calculations operating-point linearization can be used:

$$f(x + \Delta x) = f(x) + f'(x) \cdot \Delta x, \tag{11}$$

so the error is $f'(x) \cdot \Delta x$, and the variance of it is

$$var\left\{f'(x) \cdot \Delta x\right\} = \left[f'(x)\right]^2 \cdot var\{x\}. \tag{12}$$

As an approximation, the computational errors for neighbouring samples can be treated as independent, resulting that their variances can be added. The estimator of the cost function variance is $6.98 \cdot 10^{-15}$, so the standard deviation is $\sigma_{est} = 8.35 \cdot 10^{-8}$, while the standard deviation of the errors in the vicinity of the optima is $\sigma_{real} = 6.19 \cdot 10^{-8}$. It can be seen that the estimator is really close to the real value.

For random variables of normal distribution, the probability of the event that the value of the variable is between $\pm 2\sigma$ is 95.4%, and that it is between $\pm 3\sigma$ is 99.7%. For practical measurements, the latter limit is acceptable. It can be mentioned that for the given measurement, the maximal deviation was outside the 2σ tolerance band, but inside the 3σ band. Using the latter one we can say that the optima of the LM and the DE cannot be distinguished.

IV. CONCLUSIONS

In this paper properties of sine wave fitting algorithms were investigated in the vicinity of the optima of their cost function. For Maximum Likelihood (ML) estimation the optimum was searched with the Levenberg-Marquardt (LM) and the Differential Evolution (DE) methods. Their results were close, but slightly different, indicating that the optimum could not be found with arbitrary precision due to roundoff errors. In order to see the real properties, the ML function was evaluated with enhanced precision. The evaluation showed that the true cost function is much less ragged than calculated with double precision number representation. Worst case and probability analysis were evaluated for ML cost function evaluation in order to define a tolerance band within which results can be considered equivalent. Probability analysis showed that for the given measurement, the optima of the LM and the DE algorithms are indeed equivalent.

REFERENCES

[1] Standard IEEE-1241-2010, "IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters", 2010

http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5692954

[2]Šaliga, J., Kollár, I., Michaeli, L., Buša, J., Lipták, J., Virosztek, T., "A Comparison of Least Squares and Maximum Likelihood Based Sine Fittings in ADC Testing," MEASUREMENT 46: pp. 4362-4368. (2013). DOI: 10.1016/j.measurement.2013.05.004

[3] I. Kollár, T. Virosztek, V. Pálfi, B. Renczes, "ADCTest Project", URL (Aug. 22, 2014): http://www.mit.bme.hu/projects/adctest

[4]T. Virosztek, I. Kollár, "User-Friendly Matlab Tool for Easy ADC Testing", 19th IMEKO TC4 Symposium, Barcelona, Spain, 18-19 July, 2013, paper 133. pp. 561-568. http://www.imeko.org/publications/tc4-2013/IMEKO-TC4-2013-133.pdf

[5]T. Virosztek, I. Kollár, ADC Testing in Standardized and Non-standardized Ways, Executed in a Unified Framework. In: 20th IMEKO TC4 International Symposium and 18th International Workshop on ADC Modelling and Testing. Benevento, Italy, Sep. 15-17, 2014. 6 p. Paper 232. URL: https://vm.mtmt.hu/www/index.php?mode=html&DocumentID=2720264&url_on=1&st_on=1&lang=1

[6]K. Price, R. Storn, "Differential Evolution (DE)" URL (Aug. 22, 2014): http://www1.icsi.berkeley.edu/~storn/code.html

[7]B. Widrow, I. Kollár, "Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications", Cambridge University Press, Cambridge, UK, 2008. http://www.mit.bme.hu/books/quantization/