Please cite as:

B. Renczes, I. Kollár, A. Moschitta, P. Carbone, „Numerical Optimization Problems of Sine-Wave Fitting Algorithms in the Presence of Roundoff Errors“, *IEEE Trans. Instrum. Meas., vol. 65, No. 8, pp. 1785-1795,* 2016

Digital Object Identifier of the paper: 10.1109/TIM.2016.2562218

This is the accepted version of the paper. The final version is available on the IEEE Xplore.

Numerical Optimization Problems of Sine Wave Fitting Algorithms in the Presence of Roundoff Errors[1]

Balázs Renczes*, István Kollár*, Antonio Moschitta‡, Paolo Carbone‡

*Budapest University of Technology and Economics, Department of Measurement and Information Systems, Budapest, Hungary

‡University of Perugia, Department of Engineering, Perugia, Italy

***Abstract*—In this paper the effect of roundoff errors on sine wave fitting algorithms is investigated. It is shown that the standard calculation of sine wave parameters may result in unexpectedly large errors, even with floating point number representation. The three- and four-parameter Least Squares, the Maximum Likelihood and the Quantile-Based Estimator methods are investigated. It is pointed out that imprecise phase storage and summation affect almost every sine wave fitting algorithm. Furthermore, the necessary solution of sometimes ill-conditioned systems and the imprecisely evaluated distribution of the observation noise, as additional error sources are also shown to influence a part of the fitting methods. Besides error descriptions, compensation techniques are also suggested in order to mitigate the effect of the error sources. The enhancement of precision and robustness due to these suggestions is demonstrated, while keeping the given limited precision number representation platform. The Quantile-Based estimator is shown to overcome roundoff error problems when its applicability conditions are fulfilled. In addition, its performance over the Least Squares estimator is highlighted. Finally, it is pointed out that the investigated methods show similar sensitivity to the inaccurate knowledge of the frequency of the sine wave.**

Keywords*—*Numerical stability, Sine fitting, Roundoff errors, Least Squares methods, Maximum Likelihood estimation, Quantization, Quantile-based estimator

I. INTRODUCTION

Computers offer a fast and efficient way to process large data sets. In many cases, the properties of the data set can be described with a few parameters. For the description of a sine wave with DC, four parameters are needed. A sampled sine wave with arbitrary phase and with offset can be described by:

$$y_n = A \cdot \cos\left(2\pi \frac{f_0}{f_s} n\right) + B \cdot \sin\left(2\pi \frac{f_0}{f_s} n\right) + \mathrm{C}, \tag{1}$$

where A and B denote the amplitudes of the cosine and sine components, respectively, C is the offset, f_0 is the frequency of the sine wave, f_s is the sampling frequency and n is the ordinal number of the sample. Two cases can be distinguished: in the four-parameter fit all the four parameters (i.e., A, B, C and f_0/f_s) are estimated, while in the three-parameter fit frequency ratio f_0/f_s is assumed to be known exactly [1].

[1] Please cite as: B. Renczes, I. Kollár, A. Moschitta, P. Carbone, „Numerical Optimization Problems of Sine-Wave Fitting Algorithms in the Presence of Roundoff Errors“, *IEEE Trans. Instrum. Meas., vol. 65, No. 8, pp. 1785-1795,* 2016

Performing of the fit with low error is necessary for two reasons. On the one hand, the accuracy of parameter estimation depends on the quality of the fitting. On the other hand, these results can be used to test analog-to-digital converters (ADCs) [1], and the quality of the test depends on the accuracy of the estimated values of the parameters.

Different methods are available to estimate sine parameters. The most widespread one is the Least Squares Estimator (LSE). This method is obtained by minimizing the sum of the squared errors. The error sequence is the difference between the acquired data sequence and (1). The performance and accuracy of this method have been widely investigated [2]. Meanwhile, competitive methods have been proposed, including the Maximum Likelihood Estimator (MLE) [3] and the Quantile-Based Estimator (QBE) [4].

The MLE can iteratively estimate three or four parameters of the signal. It estimates the parameter set that maximizes the probability of observing the measured data [3]. The QBE in its current form is restricted to coherent sampling. It assumes that frequency ratio f_0/f_s is known exactly, i.e., performs a three-parameter fitting, based on the Gauss-Markov theorem. This method can determine parameters in one step. When its applicability conditions are fulfilled, the QBE can be shown to approximate the MLE [4].

The results of sine wave fitting can be used to characterize ADCs. In this regard, one of the mostly used indicators is the Effective Number of Bits (ENOB) of the converter, given by:

$$ENOB = b - log_2 \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2}}{\Delta/\sqrt{12}}, \tag{2}$$

where **x** is the measured data vector, **y** denotes the fitted signal, *N* is the record length, *b* is the nominal ADC resolution, and Δ is the quantization step of the converter. The ENOB value describes the 'true' resolution of the converter, i.e., with increasing fit error the ENOB decreases.

There are several error sources that can distort the result of parameter estimation and also reduce the ENOB, including nonlinear distortions, time base inaccuracies [5], and failing to fulfil the hypotheses on the quantization error [6]. Additional errors may be caused by representing data and executing the required operations with limited precision.

Due to the wide number range it can represent, floating point (FP) number representation is used during the execution of the parameter estimation methods. FP numbers are given in a normalized form $M \cdot 2^E$, where *M* denotes the mantissa and *E* is the exponent. In FP number representation the average roundoff error is roughly proportional to the represented number. Thus, larger numbers are affected by larger absolute roundoff errors. In Personal Computers (*PCs*) IEEE double precision is used, with a 53-bit mantissa and a relative error $eps_d=2.22\cdot10^{-16}$, while in Digital Signal Processors (*DSPs*) single precision is widely used, with $eps_s=1.19\cdot10^{-7}$. A detailed discussion of *eps* can be found in [7].

In particular, the effect of using single precision is considered here for two reasons. First, this representation suffers from much higher roundoff errors compared to double precision. Thus, double precision fitting results can be used as reference. On the other hand, for many practical applications, involving real time operations performed by DSPs, single precision can be advantageous, because it requires a reduced amount of processing power and memory [8]. This also applies to embedded systems equipped with small processing devices, such as microcontrollers or Field Programmable Gate Arrays (FPGAs), where power consumption and chip area are critical

figures of merit. Recent results are mentioned in the literature, discussing efficient FPGA architectures, including multiple-precision operations [9][10], as well as single precision implementations of commonly used algorithms and mathematical functions [11][12][13].

Building on these premises, this paper focuses on the sensitivity of the LSE, the MLE, and the QBE to roundoff errors, extending the ideas presented in [14]. Sections II to V follow a decreasing level of generality. Sections II and III focus on the sensitivity of the phase storage and the summation to limited precision. These errors influence both the LSE and the MLE methods. In Section IV it is pointed out that roundoff errors are also connected to the condition number of the equation system to be solved. This phenomenon may distort the result of the four-parameter LSE and the MLE. In Section V the evaluation of the Cumulative Distribution Function (CDF) is investigated. This error source has an effect specially on the MLE. Besides error descriptions possible techniques are also introduced in order to improve performance. In Section VI the Quantile Based Estimator is shown to overcome roundoff problems, when its applicability conditions are fulfilled. Its performance over the LSE is also demonstrated. Finally, the sensitivity of the LSE, MLE, and QBE to the inaccurate knowledge of the involved frequencies is compared in Section VII.

II. PHASE STORAGE ERROR

A. Error analysis

To illustrate the effect of roundoff errors on sine wave fitting algorithms, let us assume that we have a sampled sine wave:

$$x_n = A_0 \cdot \sin \varphi_n , \qquad n = 1, \dots, N \tag{3}$$

$$\varphi_n = 2\pi \frac{f_0}{f_s} n, \qquad n = 1, \dots, N \tag{4}$$

where A_0 is the peak amplitude of the sine wave, and $\varphi_n = n\varphi_1$ is the phase of the n^{th} sample. Let us assume a 12-bit bipolar ADC with Full Scale FS=1, i.e., the range of the converter is $]-FS/2\,;FS/2]$. Let the amplitude of the sinusoidal excitation be $A_0 = 0.49$ (almost fully driven ADC), $N = 50000$, and $\frac{f_0}{f_s} = \frac{1001}{50000}$. This setting ensures coherent sampling. Moreover, it fulfils the relative prime condition of [1]. Using single precision, there are several problems to face with. First, relative frequency f_0/f_s cannot be stored precisely. The relative error of the storage of this number is $-3.2 \cdot 10^{-8}$. This inaccuracy in the frequency storage results in a drift phenomenon. The error of the signal model increases with increasing *n*, see Section VII. For the last sample, the roundoff error grows to $1.6 \cdot 10^{-3}$. As the resolution of a bipolar 12-bit ADC with FS=1 is $2.44 \cdot 10^{-4}$, the error of frequency storage cannot be neglected. A possible solution to this problem is described in Section II-D. The effect of inaccurate frequency knowledge for different estimators is investigated in Section VII.

Henceforth we assume that the phase information φ_n is not distorted by the error of frequency storage, i.e., it can be calculated precisely, but then the result is rounded to the nearest representable single precision value. The roundoff error of φ_n leads to the evaluation error of samples in y_n, i.e., in the fitted sine wave [15]. This is illustrated in Fig. 1.

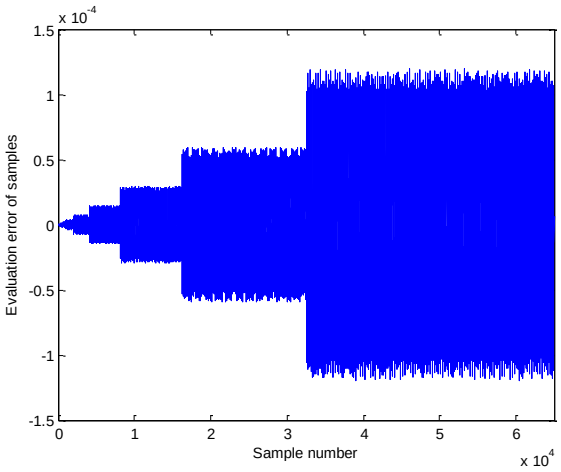

Fig. 1. Evaluation error of samples in the investigated sine wave

It can be observed that the error increases with the ordinal number of the sample. The effect of the error of phase storage on the samples can be determined with linearization:

$$f(z + \Delta z) \cong f(z) + \Delta z \cdot f'(z) \tag{5}$$

that is:

$$A_0 \sin\{\varphi_n + (\Delta\varphi)_n\} \cong A_0 \sin\varphi[n] + A_0 \cos\varphi_n \cdot (\Delta\varphi)_n, \tag{6}$$

where $(\Delta\varphi)_n$ is the storage error associated to the phase of the n^{th} element in the record. It is clear from (6) that the higher the error in the phase storage, the higher the evaluation error of samples. The latter has a cosinusoidal envelope over the data set. Depending on the instantaneous phase, this error can be much larger than the single precision eps_s.

It is important to notice that the error sequence in Fig. 1 originates from phase storage error, rather than from measurement noise. This effect may be modeled as an additional 'noise' source injected into the system. Since the absolute value of the roundoff error for φ_n increases with n, the maximum value assumed by the error sequence grows with increasing record length N. For the considered signal, we have:

$$LSB(\varphi_1) = 1.49 \cdot 10^{-8} \text{ and } LSB(\varphi_N) = 4.88 \cdot 10^{-4}, \tag{7}$$

where LSB is the resolution of the number representation, and $LSB(1)=eps_s$. Thus, for the last sample the representation error is by 4 orders of magnitude larger than for the first one. According to (6), the evaluation error of samples at the end of the data set is also in the order of magnitude 10^{-4}. This error is added to the fitted sine wave y_n.

To understand the problem more in detail, let us take the cost function (CF) of the LS estimator:

$$CF_{\text{LS}} = \sum_{n=1}^{N} \left(x_n - (y_n + e_{phase,n})\right)^2 = \sum_{n=1}^{N} \left(e_n - e_{phase,n}\right)^2 \tag{8}$$

where $e_n = x_n - y_n$ is the error sequence without the evaluation error of samples, and $e_{phase,n}$ contains the evaluation error of samples, i.e. from (6):

$$e_{phase,n} \approx A_0 \cos\varphi_n \cdot (\Delta\varphi)_n. \tag{9}$$

Thus, the cost function can be written as:

$$CF_{\mathrm{LS}} = \sum_{n=1}^{N} \left(e_n^2 - 2e_n e_{phase,n} + e_{phase,n}^2\right) \tag{10}$$

Since the expected value of each second term is 0, their sum will be neglected in the further calculations. Notice that the cost function is a random variable in a sense that for different sampling sets it assumes different values. Introducing notations

$$\varepsilon_{\mathrm{est}} = \sum_{n=1}^{N} e_n^2 \quad \text{and} \quad \varepsilon_{\mathrm{phase}} = \sum_{n=1}^{N} e_{phase,n}^2 \tag{11}$$

we have

$$CF_{LS} \approx \varepsilon_{est} + \varepsilon_{\mathrm{phase}}. \tag{12}$$

Here, ε_{est} is of Gaussian distribution, according to the central limit theorem.

The deviation of CF_{LS} from $\mathrm{E}\{\varepsilon_{\mathrm{est}}\}$ is represented by $\mathrm{var}\{\varepsilon_{est}\}$, $\mathrm{E}\{\varepsilon_{\mathrm{phase}}\}$, and $\mathrm{var}\{\varepsilon_{\mathrm{phase}}\}$. The variance of ε_{est} depends on the actual distribution of e_n. A rough guess can be obtained assuming that it is uniformly distributed with zero mean. Then:

$$\mathrm{var}\{e_n^2\} = \frac{\Delta^4}{80} - \left(\frac{\Delta^2}{12}\right)^2 = \frac{\Delta^4}{180} \quad \text{and} \quad \mathrm{var}\{\varepsilon_{est}\} = N\frac{\Delta^4}{180}, \tag{13}$$

where Δ is the quantization step of the converter, [7].

For $\varepsilon_{\mathrm{phase}}$ the expected value and the variance grows in portion with N. Thus, the standard deviation of $\varepsilon_{\mathrm{phase}}$ is proportional to $\sqrt{N}$. Consequently, the expected value of $\varepsilon_{\mathrm{phase}}$ dominates over its standard deviation.

Let us calculate the expected value of $\varepsilon_{\mathrm{phase}}$. The expected value of a single error term, using (9) is:

$$\mathrm{E}\{e_{\mathrm{phase},n}^2\} = A_0^2 \cos^2\varphi_n \cdot var\{\varphi_n\} \tag{14}$$

As an approximation, the variance of the phase storage is approximated to be proportional to the absolute value of the phase, i.e.,

$$var\{\varphi_n\} = \frac{LSB^2(\varphi_n)}{12} \approx \varphi_n^2 \cdot \frac{eps^2}{12}, \tag{15}$$

also taking into consideration that the roundoff error is uniformly distributed between $\pm LSB(\varphi_n)$. Substituting this result to (14) and considering that the variances of the roundoff errors of the phase storage for adjacent samples can be regarded as independent, we get:

$$E\{\varepsilon_{\mathrm{phase}}\} \approx \sum_{n=1}^{N} A_0^2 \cos^2(\varphi_n) \cdot \varphi_n^2 \cdot \frac{eps^2}{12} = \sum_{n=1}^{N} A_0^2 \cos^2(\varphi_n) \cdot (n \cdot \varphi_1)^2 \cdot \frac{eps^2}{12}, \tag{16}$$

where $\varphi_1 = 2\pi\frac{f_0}{f_s}$. The effect of function $\cos^2$ on average can be regarded as 0.5. Furthermore, identity $\sum_{n=1}^{N} n^2 \approx N^3/3$ is utilized. With these approximations, the following equation can be obtained:

$$E\{\varepsilon_{\text{phase}}\} \approx \frac{1}{2} A_0^2 (\varphi_1)^2 \cdot \frac{eps^2}{12} \cdot \frac{N^3}{3} \tag{17}$$

Now the standard deviation of ε_{est} and the expected value of $\varepsilon_{\text{phase}}$ should be compared. If the latter assumes higher values, the evaluation error of samples dominates over the uncertainty of the cost function:

$$\frac{E\{\varepsilon_{\text{phase}}\}}{\sqrt{var\{\varepsilon_{\text{est}}\}}} \approx \frac{\frac{1}{2} A_0^2 (\varphi_1)^2 \cdot \frac{eps^2}{12} \cdot \frac{N^3}{3}}{\frac{\Delta^2}{\sqrt{180}} \sqrt{N}} \tag{18}$$

Considering that $A_0 \approx FS/2$ and $\Delta = FS/2^b$, we have:

$$\frac{E\{\varepsilon_{\text{phase}}\}}{\sqrt{var\{\varepsilon_{\text{est}}\}}} \approx \frac{2^{2(b-1)}\sqrt{180}}{72} \cdot (\varphi_1)^2 \cdot eps^2 \cdot N^{5/2} \tag{19}$$

If the example given by (3) is investigated, i.e.,

$$b = 12, \quad \varphi_1 = 0.1258, \quad eps = eps_s, \quad N = 50000 \tag{20}$$

we have:

$$\frac{\mathrm{E}\{\varepsilon_{\text{phase}}\}}{\sqrt{var\{\varepsilon_{\text{est}}\}}} \approx 197 \tag{21}$$

which means that $\mathrm{E}\{\varepsilon_{\text{phase}}\} \gg \sqrt{var\{\varepsilon_{\text{est}}\}}$, that is, $\mathrm{E}\{\varepsilon_{\text{phase}}\}$ is important.

These derivations are valid for the LS cost function. The cost function of the MLE behaves similarly (recall that for additive normal observation noise the MLE and the LS estimates coincide, thus they can be expected to behave similarly).

B. Proposed solution to decrease the phase storage error

As described in Section II-A, the phase storage error originates from the inaccurate storage of the phase information in (4). More precisely, the absolute roundoff error increases with increasing φ_{n}, as described in Section I. However, the precise phase information can be extracted from the terms of φ_{n} in (4), using the following method.

Since sine and cosine in (1) are periodic functions, the fractional part of $\frac{f_0}{f_s} n$ contains the information that is needed to calculate their values for a given *n*. The proposed method is that the phase information should be calculated by $2\pi \left\langle \frac{f_0}{f_s} n \right\rangle$, where $\langle \cdot \rangle$ is the fractional part operator after rounding to the nearest integer value:

$$\left\langle \frac{f_0}{f_s} n \right\rangle = \frac{f_0}{f_s} n - \text{round}\left(\frac{f_0}{f_s} n \right). \tag{22}$$

For example, $\langle 2.3 \rangle = 0.3$ and $\langle 2.6 \rangle = -0.4$. The calculation of the fractional part does not inject roundoff error to the system, since an integer number can be subtracted precisely from a floating-point number. While the fractional part is still represented in single precision, its magnitude is mapped to $]-0.5; 0.5]$, also limiting the error in (6) to a predictable value. In addition, the imprecise storage of π has much lower effect on the result, than in the case it is multiplied by growing $\frac{f_0}{f_s} n$.

However, the calculation of the fractional part cannot be performed in the standard way. If $\frac{f_0}{f_s}n$ were calculated using single precision, it would also be distorted by roundoff error. Consequently, $\left\langle \frac{f_0}{f_s}n \right\rangle$ would also be imprecise. Thus, $\frac{f_0}{f_s}n$ has to be calculated with enhanced precision. A possible solution for the problem is described in detail in Section II-D.

Summarizing the proposed method, the periodic property can be exploited to perform the operation more accurately with the following steps:

a) calculate $\frac{f_0}{f_s}n$ with enhanced precision

b) calculate $\left\langle \frac{f_0}{f_s}n \right\rangle$

c) cast the result back to single

d) multiple only this fractional part by 2π.

Since the resulting phase information is in range $]-\pi;\pi]$, the absolute phase storage error is always lower than $2.4 \cdot 10^{-7}$. This is a major improvement compared to the second term of (7).

The described algorithm is significantly different from the precise sine evaluation of [16]. The latter method focuses on the precise evaluation of the sine function, assuming that the sine wave argument (i.e., the phase) is accurately known. Here, phase is not accurately known, since the imprecise storage introduces a phase roundoff error. Hence, the accurate value of the sine function cannot be determined, since the precise information was lost at the storage of the phase. From practical point of view, the proposed method evaluates the phase information with enhanced precision instead of single precision, maps the result to a limited range and only after this limitation does the rounding – resulting in a much lower absolute error.

During the enhancement method, only the phase storage in (4) was calculated with increased precision. This approach could be extended to other portions of the fitting algorithm, at a price of increased processing time. However, as a general rule, only the critical parts of the algorithm should be improved, finding a balance between run time and accuracy.

C. The effect of phase storage error

Phase storage error has an effect on most sine wave fitting algorithms (an exception is the QBE). This is shown for the case of the three-parameter LS fitting. For this algorithm, the parameters can be calculated by solving the following linear equation system in Least Squares sense [1]:

$$\mathbf{D_0 s_0} = \mathbf{x} \tag{23}$$

where

$$\mathbf{D_0} = \begin{bmatrix} \cos(2\pi f_0 t_1) & \sin(2\pi f_0 t_1) & 1 \\ \cos(2\pi f_0 t_2) & \sin(2\pi f_0 t_2) & 1 \\ & \vdots & \\ \cos(2\pi f_0 t_N) & \sin(2\pi f_0 t_N) & 1 \end{bmatrix}, \tag{24}$$

$\mathbf{s_0}$ contains the estimated in-phase and quadrature components of the signal model, and the offset i.e., *A*, *B* and *C* in (1), respectively, and $t_n = n/f_s$ [1].

The algorithm was evaluated for the sine wave given by (3), using singular value decomposition (SVD). The ADC under test is bipolar, it has a nominal bit number of b=12, FS=1 and is assumed to be unaffected by noise and non-linear distortions. The excitation signal is purely sinusoidal with an amplitude of 0.49, as in Section II-A. To overcome the error of frequency storage, relative frequency $f_0/f_s = 1/64$ was set so that it can be stored precisely even with single precision. Thus, only the effect of imprecise phase storage can be observed.

Results were obtained using different methods. First, single precision arithmetic was used to show the ENOB loss caused by roundoff errors. Besides, the method proposed in Section II-B was performed to improve results. Finally, double precision evaluation was executed as reference. TABLE I. shows the ENOB values for different record lengths.

TABLE I. ENOB OF A 12-BIT ADC FOR THE DIFFERENT METHODS

Record length	**ENOB value**		
	Single precision without enhancement	***Single precision with enhancement***	***Double precision***
10000	11.97	11.97	11.97
20000	11.96	11.97	11.97
50000	11.87	11.97	11.97
100000	11.84	11.97	11.97
200000	11.59	11.97	11.97

It can be observed that for a record length of 10000 samples there is no significant difference between the algorithms. However, with increasing record length, the roundoff errors also increase. For N=50000, the ENOB loss for the single precision evaluation without the proposed enhancement is about 0.1 bit which is not negligible. Contrarily, the enhanced single precision evaluation eliminates this problem, as shown in the second column of TABLE I. Notice again that the ENOB loss is only due to imprecise phase storage, and it is injected as an additional error source, regardless of the quality of the ADC.

D. Enhanced precision phase evaluation on a given platform

In Section II-A imprecise phase storage was shown to distort the result of sine fitting. The suggested enhancement method was to evaluate the fractional part of $\frac{f_0}{f_s} n$ and then multiply it by 2π. The crucial part is the evaluation of the fractional part with increased precision. This implies that the operation of multiplication has to be implemented more accurately.

In order to do that, first we have to find a way for increased bit number representation, while not slowing down the computer too much. The problem of floating point representation is that it has finite mantissa length. Thus, operations cannot be performed with arbitrary precision. To overcome this problem, each number can be split into more parts [14]. E.g., in single representation they can be split into 3 parts, so that each part contains 11 significant bits. Thus, the difference between the exponents is 11. The original number can be calculated as the sum of these three parts. The benefit of splitting is that the product of two splits contains at most 11+11=22 significant bits (and single representation has 23-bit mantissa). This means that the product can be stored precisely.

If operation $\frac{f_0}{f_s}n$ is to be performed, $\frac{f_0}{f_s}$ should be split into three parts. Record length *N* can be assumed to be less than $2^{22} \approx 4$ million. In this case, each *n* value from 1 to *N* can be represented with two splits precisely. After splitting the terms into parts, the multiplication can be calculated based on convolution:

$$\left(\frac{f_0}{f_s}n\right)_{enh} = [f_1, f_2, f_3]*[n_1, n_2] = [f_1 \cdot n_1,\ f_1 \cdot n_2 + f_2 \cdot n_1,\ f_2 \cdot n_2 + f_3 \cdot n_1,\ f_3 \cdot n_2] \tag{25}$$

where f_1 and n_1 denote the first split of the parts from $\frac{f_0}{f_s}$ and *n*, respectively. Notice again that the resulting vector elements are not distorted by roundoff errors, since their 23-bit mantissa contains at most 22 significant bits. After the multiplication the fraction part of the splits can be calculated. The result contains 4 parts, the sum of which is $\left\langle \frac{f_0}{f_s}n \right\rangle$. Using this approach, the roundoff error is reduced significantly as shown in Section II-C.

This method can also solve the problem of imprecise frequency storage. If needed, even the roundoff error of the stored frequency can be eliminated. The real frequency can be given as the sum of the nearest single value and a correction factor, i.e., the roundoff error:

$$\left(\frac{f_0}{f_s}\right)_{\text{real}} = \left(\frac{f_0}{f_s}\right)_{\text{single}} + \left(\frac{f_0}{f_s}\right)_{\text{corr.}}. \tag{26}$$

The multiplication of the precisely stored frequency with *n* can be performed similarly to (25). It is important to mention again that only the critical part of sine wave fitting algorithms (i.e., the phase evaluation) is calculated with enhanced precision. Theoretically arbitrary precision could be achieved on any platform, e.g., double precision could be implemented on a single precision DSP. However, due to the increased time consumption the algorithm would become practically useless.

III. Summation error

A. Error analysis

Besides phase evaluation, other error sources also influence the result of sine wave fitting. Fitting algorithms typically minimize a cost function (CF). The evaluation usually requires the calculation of a summation. If the record length is increased, the error of summation also increases, since the value of the sum is accumulating. For instance, the LSE minimizes the sum of the squared errors:

$$CF_{\text{LS}} = \sum_{n=1}^{N} (x_n - y_n)^2. \tag{27}$$

In the following we assume that the evaluation error of samples has been eliminated. In the computer, the calculated CF_{LS} is also corrupted by roundoff errors. A roundoff error is on average proportional to the magnitude of the number to be stored. By assuming standard summation, i.e., the value of the sum is accumulating, the variance of the result is [17]:

$$var\{\varepsilon_{\text{sum}}\} \approx \sum_{n=1}^{N} (n \cdot E\{e_n^2\})^2 \frac{eps^2}{12} = E^2\{e_n^2\} \frac{N^3}{3} \frac{eps^2}{12}, \tag{28}$$

where ε_{sum} is the error term, introduced by summation. This variance should be compared to the variance of the estimated CF, evaluated in (13):

$$\frac{\text{var}\{\varepsilon_{\text{sum}}\}}{\text{var}\{\varepsilon_{est}\}} \approx \frac{180}{5184} \cdot eps^2 \cdot N^2. \tag{29}$$

Consequently, using single precision, the summation error will be greater than the uncertainty of the CF estimation only if $N > 4.5 \cdot 10^7$. On the one hand, this is not a practical case: N is usually much smaller. On the other hand, while the summation error in (29) may seem negligible, it was shown in [15] that these roundoff errors cause a ragged CF, i.e., a CF featuring local minima, worsening the accuracy of iterative solutions (four-parameter LSE, MLE). Therefore, it makes sense to strive for smaller summation error.

Summation problems may be mitigated by representing data with increased precision, similarly to the method presented in Section II. This would improve the accuracy of the summation at a price of a considerably increased processing time. However, advanced summation techniques may also be applied, as shown in the following.

B. Proposed solutions to decrease the summation error

As described in Section III-A, the standard deviation of the summation can become much larger than the LSB of the number representation. The phenomenon originates from the naive approach of summation. This accumulates the result, adding small numbers to a growing sum. Possible solutions for decreasing the summation error in floating point is subtracting the mean value beforehand, using Kahan's compensated summation [18], or executing pairwise summation. For the pairwise summation a short analysis is given here.

In this method groups and subgroups can be built from the numbers and then they can be added gradually, as shown in Fig. 2. The groups contain values that assume approximately the same order of magnitude, resulting in lower roundoff errors at the summation steps. This technique is similar to the calculation of the FFT. It achieves higher accuracy, with low extra computational costs.

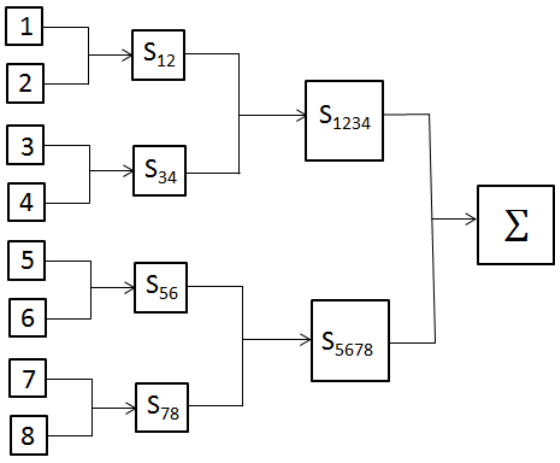

Fig. 2. Illustration of pairwise summation for 8 values

Accuracy enhancement can be demonstrated, considering the following. During pairwise summation, the sum is growing from left to right in Fig. 2. The expected value of every element to be summed is e_{RMS}^2. The relative variance of the roundoff errors of the elements can be given as $LSB^2/12$, considering the relative roundoff error is uniformly distributed between $-LSB/2$ and $LSB/2$. The variances at the addition steps can be regarded as independent. Thus, the variance of the summation can be calculated as the sum of these variances. For S_{12} the expected value is $2e_{RMS}^2$, while for S_{1234} it is $4e_{RMS}^2$. Although the relative variances are still $\frac{LSB^2}{12}$, the absolute variances grow, due to the growing sum. For the sake of simplicity, let us assume that the number of values to be summed is a power of 2. The variance of the whole summation can be calculated as follows:

$$\begin{aligned} var\{\varepsilon_{sum,pairwise}\} &\approx \frac{LSB^2}{12}\left(N(e_{\mathrm{RMS}}^2)^2 + \frac{N}{2}(2\cdot e_{\mathrm{RMS}}^2)^2 + \cdots + \frac{N}{N}(N\cdot e_{\mathrm{RMS}}^2)^2\right) \\ &= \frac{LSB^2}{12} N e_{\mathrm{RMS}}^4 \sum_{n=1}^{log_2(N)} 2^n = \frac{LSB^2}{6} N^2 e_{\mathrm{RMS}}^4 \end{aligned} \tag{30}$$

The result can be compared to the uncertainty of the CF:

$$\frac{var\{\varepsilon_{sum,pairwise}\}}{var\{\varepsilon_{est}\}} \approx 30\cdot eps^2\cdot N \tag{31}$$

The result compared to (29) shows that the summation error significantly decreases using the pairwise summation technique. Furthermore, the observed ratio increases in portion with N instead of N^2. With pairwise summation, N should be greater than $2.3\cdot 10^{12}$ so that the summation error is greater than the uncertainty of the CF.

During the proof it was assumed that N is a power of 2, but this method can improve the result of the summation for an arbitrary record length [18].

IV. Conditioning of the Equation System

A. Error analysis

In practical cases, the frequency of the measured signal is usually unknown. The LSE and MLE can also estimate this parameter iteratively, refining the frequency estimate at each iteration step. Therefore, these algorithms need a good initial estimate of the frequency of the sine wave. For this purpose, the interpolated FFT algorithm can be used, for example, [19].

The solution of the four-parameter LS sine fitting problem is similar to that of the three parameter sine fitting described in (23). However, in this case frequency correction $\Delta\frac{f}{f_s}$ is one of the parameters to be estimated. The estimation of this fourth parameter introduces additional error sources, because the four-parameter LSE requires adding an extra column to system matrix **D,** [1]:

$$\mathbf{D} = \begin{bmatrix} \cos\varphi_1 & \sin\varphi_1 & 1 & 2\pi t_1(-A\sin\varphi_1 + B\cos\varphi_1) \\ \cos\varphi_2 & \sin\varphi_2 & 1 & 2\pi t_2(-A\sin\varphi_2 + B\cos\varphi_2) \\ \vdots & \vdots & \vdots & \vdots \\ \cos\varphi_N & \sin\varphi_N & 1 & 2\pi t_N(-A\sin\varphi_N + B\cos\varphi_N) \end{bmatrix} \tag{32}$$

$$\varphi_n = 2\pi\frac{f_i}{f_s}n,$$

where f_i is the frequency estimate in iteration step *i*. Observe that the maximum value in the fourth column in (32) grows with *N*, since t_N grows. This may lead to the solution of an ill-conditioned equation system.

In fact, in numerical calculations the condition number (i.e., the ratio between the highest and lowest singular values) of a matrix is a measure of robustness. From practical point of view, it can be treated as sensitivity: the higher the condition number, the higher the sensitivity of the result to perturbations. For example, during the solution of a general matrix equation for **v**:

$$\mathbf{L}\mathbf{v} = \mathbf{w}, \tag{33}$$

where **L** is a matrix, and **v** and **w** are vectors, the condition number of **L** determines the sensitivity of the solution for **v** to numerical inaccuracies of **L** and **w**.

Let us consider a 12-bit ADC with $\Delta = 1$ and $N = 50000$. The excitation signal has the following parameters:

$$A = 1200,\ B = 1650,\ C = 2048\,,\ \ \frac{f}{f_s} = \frac{1}{64}, \tag{34}$$

so that 99.6% of the entire ADC range is excited by the stimulus. With these parameters the condition number of **D** is $1.2 \cdot 10^8$, larger than $1/eps_s$=$8.39 \cdot 10^6$. Hence, the matrix is very much ill-conditioned.

The needed frequency correction is usually small. This ill-conditioning may result in calculation errors that may distort this small value considerably. Consequently, it is not ensured that after the iteration the system gets closer to the optimum frequency. Thus, also convergence problems may occur, see Section IV-B. Notice that the effect of ill-conditioning is illustrated for the four-parameter LSE, but it also affects the MLE method.

B. Proposed solution to decrease ill-conditioning

Ill-conditioning is caused by the fourth column of **D**, since it may contain much larger values compared to the first three columns [8]. The problem can be solved by dividing the elements of the fourth column with a proper number, i.e. by proper scaling. If the fourth column were scaled properly, the condition number could be improved considerably [20]. The scaled matrix can be given as:

$$\mathbf{D_{scaled}} = \begin{bmatrix} \cos\varphi_1 & \sin\varphi_1 & 1 & 2\pi t_1\{-A\sin\varphi_1 + B\cos\varphi_1\}/\lambda_{scaling} \\ \cos\varphi_2 & \sin\varphi_2 & 1 & 2\pi t_2\{-A\sin\varphi_2 + B\cos\varphi_2\}/\lambda_{scaling} \\ \vdots & \vdots & \vdots & \vdots \\ \cos\varphi_N & \sin\varphi_N & 1 & 2\pi t_N\{-A\sin\varphi_N + B\cos\varphi_N\}/\lambda_{scaling} \end{bmatrix} \tag{35}$$

where $\lambda_{scaling}$ denotes the scaling factor. In order to solve the same equation system, the fourth parameter to be calculated after scaling is $\left(\Delta\frac{f}{f_s}\right)\cdot\lambda_{scaling}$.

In [21] the conditioning of the four-parameter LS problem was investigated. It was shown that for large record length, the condition number can be decreased to approx. 3.74. A scaling factor for the problem was suggested in [8]. However, it does not contain the amplitude of the input signal. Observations for different parameter settings showed that the optimal scaling factor is approximately:

$$\lambda_{\text{scaling,opt.}} \approx K \cdot N \cdot \pi, \qquad K = \sqrt{A^2 + B^2}, \tag{36}$$

i.e., the scaling factor is proportional to aggregated amplitude *K* and to record length *N*. However, it is independent of the ratio between signal frequency and sampling frequency. Using the given scaling factor, the condition number of **D** with the parameters in (34) decreases to 3.78. This value is approximately the optimal condition number

determined in [21]. Notice that during the solution the relative frequency, i.e., f/f_s was determined. The physical frequency of the signal can be calculated by multiplying this result with the sampling frequency.

To show the gained performance enhancement, the four parameter LS fitting algorithm was evaluated using single precision, with the parameters given in (34), both with and without the scaling factor given in (36). Let the ADC under test be unipolar, with *b=12*, again. However, the range of the ADC is now [0;4095]. Fine tuning of the frequency $\Delta\frac{f}{f_s}$ in each iteration step is shown in Fig. 3. It can be observed that without scaling the algorithm failed to converge. However, scaling the fourth column of **D** solved the convergence problem.

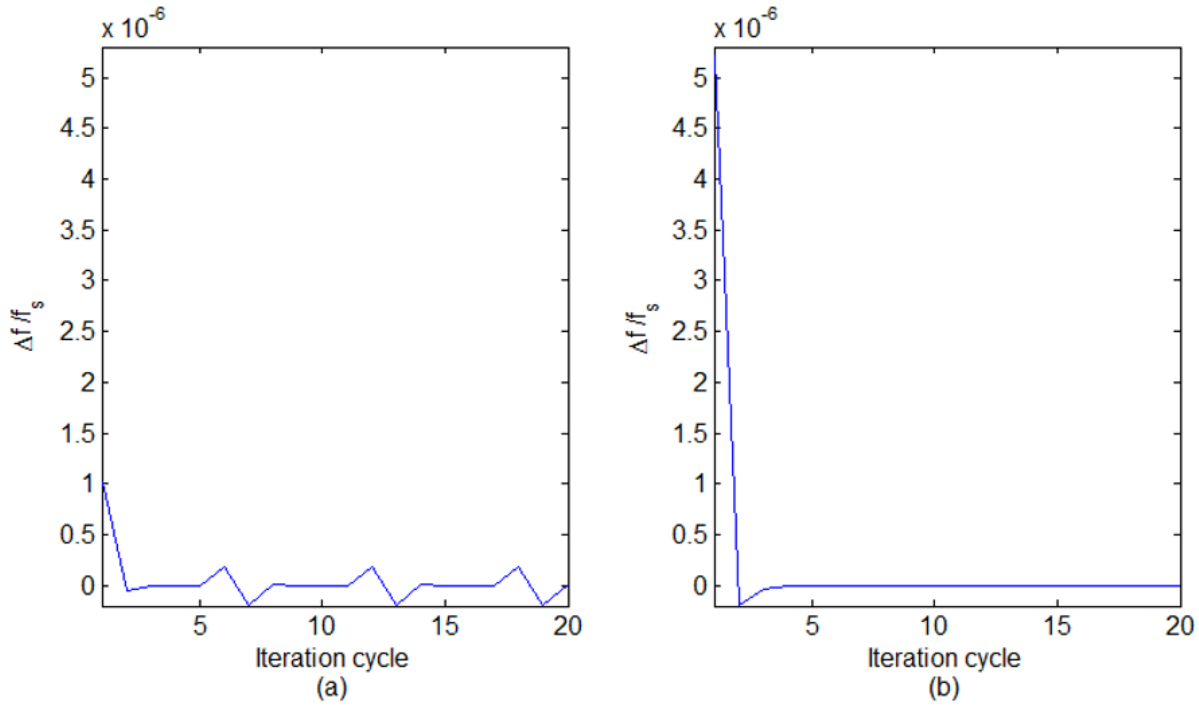

Fig. 3. Frequency fine tuning without scaling (a) and with scaling (b)

V. Noise CDF of the Maximum Likelihood Estimator (MLE)

A. Error analysis

Error sources up to this point affect both the LSE and the MLE methods. This section investigates the effect of the observation noise model on the parameter estimation. Since LSE does not utilize any noise model, this error is special for the MLE.

The MLE method [3] tries to find model parameters corresponding to the highest probability of observing the collected data set, assuming a known distribution for (random) errors. In particular, if the acquired signal is affected by an additive white noise, the complexity of the MLE can be reduced by using the log-likelihood function:

$$\ln L(p) = \sum_{n=1}^{N} ln[\mathrm{P}(X_n = x_n)], \tag{37}$$

where $\mathrm{P}(X_n = m)$ is the probability that the n^{th} sample of **X** equals to a digital code *m*, X_n is a possible output code and x_n, with n=1,…,*N*, denotes the observed sequence of ADC equivalent output levels [3].

In this method the additional noise is also taken into consideration, resulting in the signal model:

$$x_n = Q(y_n + \xi_n) \qquad n = 1 \dots N \tag{38}$$

where Q denotes the operation of quantization, y_n is given by (1) and ξ_n is the additional noise. The operation of quantization is based on the transition levels of the quantizer. If the transition levels are unknown, they can be determined, for instance, by the histogram test [23].

The noise model of $\boldsymbol{\xi}$ determines the probabilities in (37). The noise is usually assumed to be additive white Gaussian noise (AWGN), with zero mean and standard deviation σ. The probabilities in (37) under the AWGN assumption are usually computed by using the erf function. Thus, the probability that the n^{th} sample is in the l^{th} code bin can be calculated as:

$$P(X_n = l) = \frac{1}{2}\left[\text{erf}\left(\frac{T_l - y_n}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{T_{l-1} - y_n}{\sigma\sqrt{2}}\right)\right], \tag{39}$$

where T_l is the l^{th} transition level [22], and

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2} dz. \tag{40}$$

From numerical point of view, this calculation method is a potential error source, considering the following. The Gaussian distribution can describe the noise model satisfactorily for low noise values, such that $|\xi_n| < 3\sigma$. However, for larger values of $|\xi_n|$, the Probability Density Function (PDF) of the Gaussian distribution converges quickly to zero, and its Cumulative Distribution Function (CDF) converges quickly either to 0 (when $\xi_n < -3\sigma$) or to 1 (when $\xi_n > 3\sigma$). The argument values of the erf function may assume large values for two reasons. First, during real measurements an outlier, i.e., a sample affected by large noise value ξ_n, may actually appear, despite its low probability of occurrence. On the other hand, after initialization, model parameters may significantly differ from the optimal values. At run-time, this indicates that the model optimization needs additional iterations. The problem is that value of the erf function at 4 is $1 - 1.58 \cdot 10^{-8}$. The deviation of this value from 1 is so small that it cannot be represented using single precision.

For large erf arguments (when $\xi_n > 3\sigma$), the result of (39) can be given in a form:

$$P(X_n = l) = \frac{1}{2}[(1 + \delta_1) - (1 + \delta_2)], \tag{41}$$

where δ denotes a small value. If both $\frac{T_l - y_n}{\sqrt{2}\sigma} > 4$ and $\frac{T_{l-1} - y_n}{\sqrt{2}\sigma} > 4$ hold, both single precision erf evaluations would yield a result of exactly 1. This is caused by the fact that δ_1 and δ_2 cannot be represented beside 1. The difference in (39) is in this case 0, leading to singularity when evaluating the log-likelihood function. Thus, if for only one sample the arguments of the erf functions are higher than 4, (37) cannot be determined. Consequently, the algorithm in this form is numerically unusable.

B. Proposed solution to evaluate the CDF with small error

The proposed solution for the problem is that for the samples at which the noise level is high, the complementary error function (erfc) should be used instead of the erf function. This function analytically equals to $1 - \text{erf}$, but due

to the storage of the deviation from 1, it can represent function values for larger operand values much more accurately than the erf function.

The calculation of (39) can be given using erfc:

$$P(X_n = l) = \frac{1}{2}\left[\mathrm{erfc}\left(\frac{T_{l-1} - y_n}{\sigma\sqrt{2}}\right) - \mathrm{erfc}\left(\frac{T_l - y_n}{\sigma\sqrt{2}}\right)\right] = \frac{1}{2}(\delta_1 - \delta_2), \tag{42}$$

It can be observed that instead of calculating the small difference of relatively large numbers in (41), the difference can be calculated directly.

However, this approach cannot handle arguments, for which $\frac{T_l - y_n}{\sigma\sqrt{2}} < -3$. In this case, erf is close to -1 and erfc is close to 2. Thus, the same problem occurs, as for high erf argument values: a small number should be represented beside 2. The problem can be solved, using identity $\mathrm{erfc}(-z) = 2 - \mathrm{erfc}(z)$. The calculation of (39) in this case can be given as:

$$P(X_n = l) = \frac{1}{2}\left[\mathrm{erfc}\left(-\frac{T_l - y_n}{\sigma\sqrt{2}}\right) - \mathrm{erfc}\left(-\frac{T_{l-1} - y_n}{\sigma\sqrt{2}}\right)\right]. \tag{43}$$

Depending on argument values $\frac{T_l - y_n}{\sigma\sqrt{2}}$ and $\frac{T_{l-1} - y_n}{\sigma\sqrt{2}}$, the following rule should be followed:

$$\text{used evaluation: } \begin{cases} (39), \text{if} & -0.477 \leq \text{argument values} \leq 0.477 \\ (42), \text{if} & 0.477 < \textit{argument values} \\ (43), \text{if} & \text{argument values} < -0.477 \end{cases}, \tag{44}$$

With this technique, the representable range is widened to $\left|\frac{T_l - y_n}{\sigma\sqrt{2}}\right| < 26$. This is a major improvement, compared to the original range $\left|\frac{T_l - y_n}{\sigma\sqrt{2}}\right| < 4$.

It should be noted that if $\frac{T_l - y_n}{\sigma\sqrt{2}}$ and $\frac{T_{l-1} - y_n}{\sigma\sqrt{2}}$ are in different ranges of (44), the importance of using different evaluations becomes less important. In this case, the difference between the erf values is never zero. Thus, (39) can be used without singularity issues.

VI. Quantile Based Estimator (QBE)

In this section the Quantile Based Estimator (QBE) is considered. It is shown that this estimator can be very robust to roundoff error sources described in Sections II and III, at least when its applicability conditions are fulfilled.

The QBE can be derived based on the Gauss-Markov Theorem [4]. The measured signal is assumed to be digitized by a quantizer with known transition levels. The transition levels can be determined, for instance by the histogram test [23], similarly to the MLE. QBE can easily be given for the estimation of a constant signal affected by AWGN, by linearizing the effect of the noise on the CDF of the samples. This approximates the Best Linear Unbiased Estimator (BLUE) [24].

The QBE can be extended to the three-parameter sine wave fitting case. In this case, the sampling has to be synchronized so that the sine wave is sampled at the same phase positions *m* times. In order to ensure this property, the sampling should be coherent and

$$\frac{f_0}{f_s} = \frac{J}{R} \tag{45}$$

should hold, *J* and *R* being relative prime integers. In fact, the instantaneous phase of the sampled sinusoidal stimulus can assume only *R* distinct values [25]. The record length can be given by $N = mR$. Thus, by grouping the collected samples on a "per-phase" basis, the three-parameter sine fitting can be transformed into an equivalent set of *R* DC estimation problems. Each set is based on the usage of a constant stimulus at the ADC input and contains *m* samples. In absence of noise that is:

$$d_r = A\cos\varphi_r + B\sin\varphi_r + C \quad r = 1, \dots, R, \tag{46}$$

where φ_r is the phase of the r^{th} set. Having these data sets, the conventional Gauss-Markov estimate can be determined.

The QBE takes quantizer non-linearities into consideration, providing almost unbiased results for the sine wave parameters [4]. In fact, the QBE approximates the MLE that is asymptotically unbiased for large data records, regardless of the ADC resolution and signal dynamics. The QBE is also faster than the MLE, since the Gauss-Markov estimate can be determined in one step. However, in its present form it can only estimate *A*, *B* and *C*, but not the frequency.

The QBE takes advantage of restriction that a coherently sampled sine wave is repeated *m* times. The advantage is that in this case the phases of the repeated samples are exactly the same as in the first period. Formally, $\varphi_{r+kR} = \varphi_r, k \in Z^+$. This implies that for high phase values, φ_{r+kR} can be substituted with φ_r, similarly to Section II.B. Consequently, the phase evaluation and storage problem is avoided inherently. In addition, the samples are divided into subgroups. This division decreases the roundoff error caused by the summation error, similarly to Section III.B. Thus, roundoff error problems described in Sections II and III can be reduced considerably using the QBE.

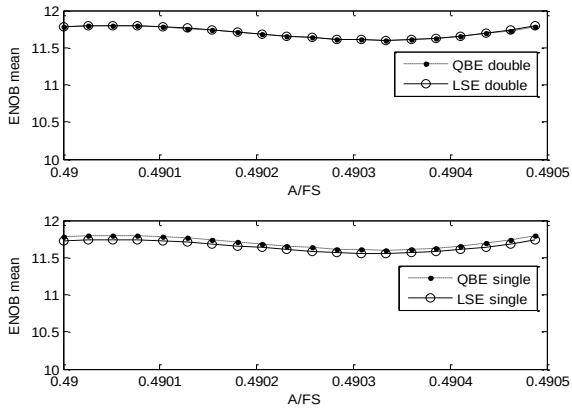

Fig. 4(a). Mean ENOB, for a record length of 10^5 samples, performing phase evaluation in double precision (upper plot) and single precision (lower plot)

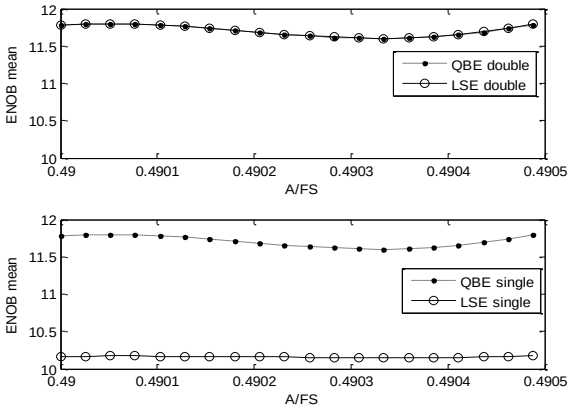

Fig. 4(b). Mean ENOB, for a record length of 10^6 samples, performing phase evaluation in double precision (upper plot) and single precision (lower plot)

The robustness of the QBE to the usage of single precision representation was verified through simulations and was compared to that of the LSE, for different values of ADC resolution, record length, and for different INL (integral non-linearity) levels. The results show that the QBE is increasingly advantageous over the LSE, when the record length or the ADC resolution is increased. For instance, Fig. 4(a), obtained for a 12-bit ADC unaffected by INL, shows the mean ENOB as a function of the sinewave amplitude A, assuming values in $[0.49FS; 0.49FS+\Delta]$, with $f_0/f_s=1/64$, $\sigma=0.2\Delta$, and $N=10^5$. When double precision is used, the LSE and the QBE behave identically (see upper plot). However, when single precision is used (see lower plot), the QBE ENOB is almost unchanged, while the LSE ENOB is decreased by about 0.05 bits. This phenomenon is magnified when record length N is increased. In fact, when N is increased to 10^6, Fig. 4(b) shows that the performance of the QBE is unaffected by the usage of single precision, while the ENOB of the LSE drops by about 1.6 bits.

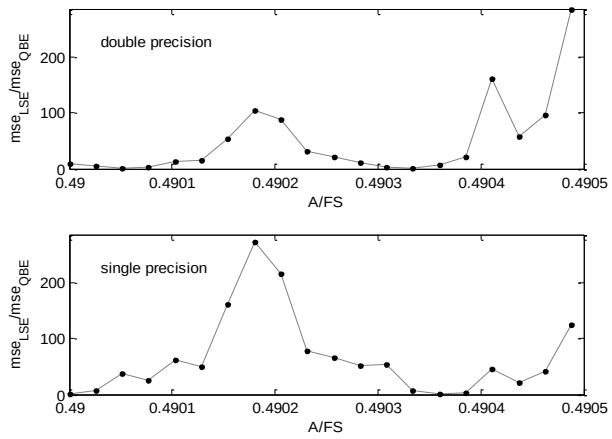

Fig. 5(a). Ratio between the mse of the LSE and the QBE, for the estimation of sinewave amplitude A, under the same conditions as Fig. 4(a)

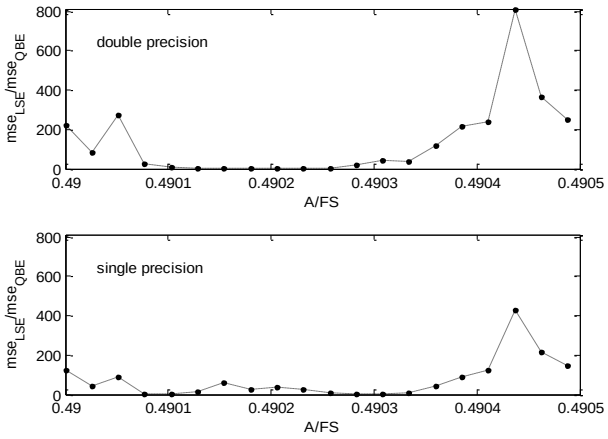

Fig. 5(b). Ratio between the mse of the LSE and the QBE, for the estimation of sinewave amplitude A, when the ADC is affected by INL

The QBE consents a more accurate estimation of the sinewave parameters, even if the calculated ENOB is similar to that of the LSE. For instance, Fig. 5(a), obtained under the same conditions as Fig. 4(a), shows the ratio between

the mean squared error (mse) of the LSE and the QBE, when both estimators are used to estimate sinewave amplitude *A*. It can be observed that, even if the QBE and the LSE yield similar ENOB, the QBE has a much lower mse (mean squared error of the estimated amplitude of the sine wave with respect to the true one), because it is almost unbiased. Notice that in this case the exact value of *A* is known, since the results are simulated.

The above described effect is increased when the ADC is affected by INL. Fig. 5(b) was derived under the same conditions as Fig. 5a. The only exception is that the 12-bit ADC is affected by INL, uniformly distributed in [-0.3Δ;0.3Δ]. As expected, the mse ratio increases with respect to Fig. 5(a), because the QBE uses information about quantizer thresholds, while the LSE does not.

The drawback of the presented QBE is that it is only applicable when the relative frequency satisfies (45). Consequently, the frequency of the sine wave has to be set precisely for several digits on a generator. This means that the QBE is potentially sensitive to synchronization inaccuracies (also due to roundoff errors in the representation of f_0/f_s), leading to estimation errors in practical scenarios.

VII. The effect of inaccurate knowledge of the frequency of the sine wave

In this section the sensitivity of different methods to the inaccurate knowledge of the frequency is investigated. Note that in the previous sections the frequency was assumed to be known precisely. The frequency error may originate from imprecise estimation and imprecise storage. The latter problem can be solved by the method described in Section II.D. In this case, the original assumption of precise frequency knowledge is met. However, if the frequency is estimated imprecisely or the imprecise storage is not compensated, an additional error is injected to the system. In this case, instead of (4) we have:

$$2\pi\frac{f_0+\Delta f_0}{f_s}n=\varphi_n+(\Delta\varphi)_{n,\mathrm{freq.}},\qquad n=1,\dots,N \tag{47}$$

and the phase error is:

$$(\Delta\varphi)_{n,\mathrm{freq.}}=2\pi\frac{\Delta f_0}{f_s}n,\qquad n=1,\dots,N, \tag{48}$$

where $(\Delta\varphi)_{n,\mathrm{freq.}}$ denotes the phase error due to imprecise frequency information. It should be noticed that contrarily to the error of phase storage, this error is systematic, i.e., its sign and amplitude can be given exactly. Furthermore, the amplitude of this error grows with increasing record length.

To gain further insight and compare the sensitivity of the LSE, MLE, and QBE algorithms to frequency inaccuracy, a Monte Carlo analysis was carried out. To this aim, each estimator's signal model assumed a frequency ratio $f_0/f_s=1/20$. A relative error $\nu=10^{-6}$ on the frequency of the sinewave was assumed, simulating the acquisition of a sinewave with A≅0.2501036, C≅0.2500010 and a ratio $f_0/f_s=(1+\nu)\cdot 1/20$. The simulation also assumed the signal to be digitized by a bipolar uniform ADC with *FS*=1, unaffected by integral nonlinearities (INL). ADC resolutions of 8, 10, and 12 bits were considered. An AWGN was assumed, again with standard deviation $\sigma=0.4\Delta$ for each considered resolution. Thus, the three estimators were tested against a three-parameter sine fitting scenario, for several values of record length *N*. The simulator generated a stimulus, digitized it, and run each estimator 5 times, evaluating the ENOB.

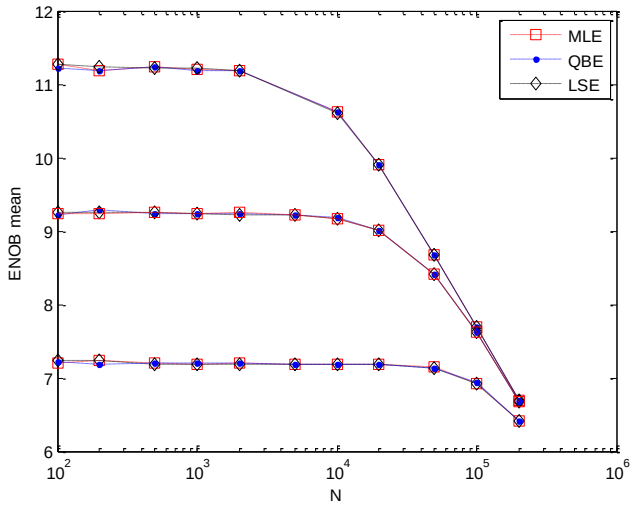

Fig. 6. Mean ENOB for the three-parameter estimation of a sine wave using different estimators (b=8, 10, 12)

Fig. 6 shows the mean values of the ENOB as a function of record length N. It can be observed that, when N is increased beyond a critical value, the calculated ENOB decreases considerably. On the other hand, the three considered estimators show similar sensitivity to frequency inaccuracy. Such behavior can be explained by observing that when the sample index is large, the instantaneous phase of a collected sample may significantly differ from the instantaneous phase assumed by the estimators' signal models. Fig. 6 also shows that the requirements on frequency accuracy become less stringent if the record length or the ADC resolution is decreased. Note that on the other hand, sensitivity to frequency error also means for the four-parameter LSE and the MLE (that is, when the frequency also is estimated) that minimization vs. frequency is effective, since frequency changes noticeably change the cost function. Thus, the frequency is well determined by the minimization of the CF.

VIII. Conclusions

In this paper numerical optimization problems of sine wave fitting algorithms were investigated. It was pointed out that the roundoff errors of numerical calculations may distort the result of the fitting algorithms considerably.

Roundoff errors due to imprecise phase storage and imprecise summation were shown to affect both the Least Squares estimator and the Maximum Likelihood estimator. To decrease the influence of the phase storage error, it was shown that dealing with only the fractional part (mod 2π) of the phase information helps, exploiting the periodic property of the sine wave. Pairwise summation was proven to decrease the summation error.

Ill-conditioned matrix equations were pointed out to cause numerical instability in the four-parameter Least Squares and in the Maximum Likelihood methods. An illustrative example was given to demonstrate that even divergence issues may occur due to ill-conditioning. Finally, a good scaling factor was given for the Least Squares estimator in order to ensure well-conditioning.

Numerical evaluation of the distribution of the utilized observation noise was also considered. The Maximum Likelihood estimator is influenced by this error. By evaluating the cumulative distribution function of the Gaussian distribution using the erfc function, the numerical instability of the method was be significantly decreased.

The Quantile Based Estimator was pointed out to be robust to the phase storage error and to summation errors, if it is provided that the input signal is coherently sampled. Furthermore, its performance over the Least Squares estimator was demonstrated.

Finally, all the investigated estimators were shown to have similar sensitivity to the inaccurate knowledge of the ratio of the frequency of the sinewave and the sampling frequency.

Source files of the described algorithms in MATLAB are available at [26].

ACKNOWLEDGEMENT

This work was partially supported by the Hungarian Research Fund – OTKA 115820, the ARTEMIS JU and the Hungarian National Research, Development and Innovation Fund in the frame of the R5-COP (Reconfigurable ROS-based Resilient Reasoning Robotic Cooperating Systems) project.

The authors thank the reviewers for their suggestions that helped to improve the paper.

REFERENCES

[1] Standard IEEE-1241-2010, "IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters", 2011, DOI: 10.1109/IEEESTD.2011.5692956

[2] P. Handel, "Properties of the IEEE-STD-1057 four-parameter sine wave fit algorithm," *IEEE Trans. Instrum. Meas.*, vol. 49, no. 6, pp. 1189–1193, Dec. 2000. DOI: 10.1109/19.893254

[3] J. Šaliga, I. Kollár, L. Michaeli, J. Buša, J. Lipták and T. Virosztek, "A Comparison of Least Squares and Maximum Likelihood Based Sine Fittings in ADC Testing," *Measurement*, no. 46, pp. 4362-4368. (2013). DOI: 10.1016/j.measurement.2013.05.004

[4] A. Moschitta, J. Schoukens, P. Carbone, "Parametric System Identification Using Quantized Data", *IEEE Trans. Instrum. Meas*., vol. 64, no. 8, pp. 2312-2322, Aug. 2015. DOI: 10.1109/TIM.2015.2390833

[5] G. Vandersteen, Y. Rolain, and J. Schoukens, "An identification technique for data acquisition characterization in the presence of nonlinear distortions and time base distortions," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 5, pp. 1355–1363, Oct. 2001. DOI: 10.1109/19.963210

[6] I. Kollár and J. J. Blair, "Improved determination of the best fitting sine wave in ADC testing," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 5, pp. 1978–1983, Oct. 2005. DOI: 10.1109/TIM.2005.855082

[7] B. Widrow, I. Kollár, "Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications", *Cambridge University Press*, Cambridge, UK, 2008. DOI: 10.1017/CBO9780511754661

[8] P. Ramos, T. Radil, F. Janeiro, "Implementation of sine-fitting algorithms in systems with 32-bit floating point representation", *Measurement*, no. 45, pp. 155-163, 2012. DOI: 10.1016/j.measurement.2011.05.011

[9] M.K.Aiswal, R.C.C.Cheung, M. Balakrishnan, K. Paul, "Unified Architecture for Double/Two-Parallel Single Precision Floating Point Adder," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol.61, no.7, pp.521-525, July 2014. DOI: 10.1109/TCSII.2014.2327314

[10] K.Paldurai, K. Hariharan, "FPGA implementation of delay optimized single precision floating point multiplier", *2015 IEEE International Conference on Advanced Computing and Communication Systems*, pp.1-5, 5-7 Jan. 2015. DOI: 10.1109/ICACCS.2015.7324094

[11] Fengbo Ren, R. Dorrace, Wenyao Xu, D.Markovic, "A single-precision compressive sensing signal reconstruction engine on FPGAs", *2013 IEEE 23rd International Conference on Field Programmable Logic and Applications (FPL)*, pp.1-4, 2-4 Sept. 2013. DOI: 10.1109/FPL.2013.6645574

[12] Chen Dong, Chen He, Sun Xing, Pang Long, "Implementation of Single-Precision Floating-Point Trigonometric Functions with Small Area", *2012 IEEE International Conference on Control Engineering and Communication Technology (ICCECT*), pp.589-592, 7-9 Dec. 2012. DOI: 10.1109/ICCECT.2012.186

[13] Yangming Li; Yajuan He; Yanming He; Ziji Zhang; Shaowei Zhen; Ping Luo, "Design of a power optimized 1024-point 32-bit single precision FFT processor", *2014 12th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, pp.1-3, 28-31 Oct. 2014. DOI: 10.1109/ICSICT.2014.7021572

[14] B. Renczes, I. Kollár, P. Carbone, A. Moschitta, V. Pálfi, T. Virosztek, "Analyzing Numerical Optimization Problems of Finite Resolution Sine Wave Fitting Algorithms", *Proceedings of IEEE International Instrum. Meas. Technology Conference*, pp. 1662-1667, Pisa, Italy 2015. DOI: 10.1109/I2MTC.2015.7151529

[15] B. Renczes, I. Kollár, "Roundoff Errors in the Evaluation of the Cost Function in Sine Wave Based ADC Testing", *20th IMEKO TC4 International Symposium and 18th International Workshop on ADC Modelling and Testing*, Benevento, Italy, Sep. 15-17, 2014. pp. 248-252, Paper 214.

[16] S. Gal, B. Bachelis, "An Accurate Elementary mathematical library for the IEEE floating point standard", *ACM Transactions on Mathematical Software*, vol. 17, Issue 1, Mar 1991. DOI: 10.1145/103147.103151

[17] T. G. Robertazzi, S. C. Schwartz, "Best Ordering for Floating-Point Addition", *ACM Transactions on Mathematical Software*, vol. 14, Issue 1, 1988. DOI: 10.1145/42288.42343

[18] N. J. Higham, "The accuracy of floating point summation", *SIAM Journal on Scientific Computing* 14 (4): pp. 783–799, 1993. DOI: 10.1137/0914050

[19] J. Schoukens, R. Pintelon, and H. Van Hamme, "The interpolated fast Fourier transform: A comparative study," *IEEE Trans. Instrum. Meas.*, vol. 41, no. 2, pp. 226–232, Apr. 1992. DOI: 10.1109/19.137352

[20] G. Golub, C. van Loan, "Matrix Computations", *The John Hopkins University Press*, Baltimore, USA, 1983

[21] K. Chen, Y. Xue, "Improving four-parameter sine wave fitting by normalization", *Computer Standards and Interfaces*, vol. 29, pp. 184-190, 2007. DOI: 10.1016/j.csi.2006.05.005

[22] L. Balogh, I. Kollár, A. Sárhegyi, "Maximum Likelihood Estimation of ADC Parameters," *Proceedings of IEEE International Instrum. Meas. Technology Conference*, pp. 24-29, Austin, USA 2010. DOI: 10.1109/IMTC.2010.5488286

[23] J. Blair. "Histogram measurement of ADC nonlinearities using sine waves", IEEE Trans. Instrum. Meas., vol. 43, no. 3, pp.:373-383, 1994. DOI: 10.1109/19.293454

[24] L. Y. Wang, G. G. Yin, J. Zhang, Y. Zhao, "System Identification with Quantized Observations", *Springer Science*, 2010. DOI: 10.1007/978-0-8176-4956-2

[25] A. Moschitta, P. Carbone, "Testing Data Converters when Sampling is Incoherent," *13th Workshop on ADC Modelling and Testing IWADC 2008*, Florence, Italy, Sept. 22-24, 2008. pp. 22-24. http://www.imeko.org/publications/iwadc-2008/IMEKO-IWADC-2008-194.pdf

[26] B. Renczes, I. Kollár "Source files for Numerical Optimization Problems of Sine Wave Fitting Algorithms in the Presence of Roundoff Errors", 2016. URL: https://github.com/renczes/TIM2016_Num_Opt