

Variance Arithmetic*

Chengpu Wang

40 Grossman Street, Melville, NY 11747, USA

Chengpu@gmail.com

Abstract

A new deterministic uncertainty-bearing floating-point arithmetic called *variance arithmetic* is developed to track the uncertainty for arithmetic calculations statistically. It uses a novel rounding scheme to avoid the excessive rounding error propagation in conventional floating-point arithmetic. Unlike interval arithmetic, its uncertainty tracking is based on statistics and the central limit theorem, with a much tighter bounding range based on a truncated Gaussian distribution. The variance arithmetic is found to be superior to interval arithmetic in both uncertainty-tracking and uncertainty-bounding for normal usages.

Let δx be the uncertainty distribution deviation for a value x . Define $\delta x/|x|$ as precision, which represents the information content of the value with uncertainty. The variance arithmetic has precision requirements on the inputs. When the precision requirements are satisfied:

- The output values are identical to the result if there were no uncertainties in the inputs.
- The precision is preserved for unitary operations such as inversion, square and square-root.
- The variance arithmetic provides ideal approach when the analytic solution is available, to avoid the dependency problem completely.
- For a round-trip calculation without the dependency problem such as after forward and backward FFT, the original input uncertainties can be recovered.

When the precision requirements are not satisfied, the bias of the outputs can be expressed in terms of input precisions.

Keywords: computer arithmetic, error analysis, interval arithmetic, uncertainty, numerical algorithms.

AMS subject classifications: 65-00

*Submitted: May 20, 2006; Revised: November 10, 2010; July 18, 2011; September 1, 2012; Match 27, 2014;

1 Introduction

1.1 Measurement Uncertainty

Except for the simplest counting, scientific and engineering measurements never give completely precise results [1][2]. In scientific and engineering measurements, the uncertainty of a measurement x usually is characterized by either the sample deviation δx or the uncertainty range Δx [1][2].

- If $\delta x = 0$ or $\Delta x = 0$, x is a *precise value*.
- Otherwise, x is an *imprecise value*.

$P \equiv \delta x/|x|$ is defined as the *statistical precision* (or simply precision in this paper) of the measurement, in which x is the value, and δx is the uncertainty deviation. A larger precision means a coarser measurement while a smaller precision mean finer measurement. The precision of measured values ranges from an order-of-magnitude estimation of astronomical measurements to 10^{-2} to 10^{-4} of common measurements to 10^{-14} of state-of-art measurements of basic physics constants [3].

1.2 Problem of Conventional Floating-Point Arithmetic

The *conventional floating-point arithmetic* [8][9][10] assumes a constant and best-possible precision for each value all the time, and constantly generates artificial information during the calculation [11]. For example, the following calculation is carried out precisely in integer format:

$$\begin{aligned} 64919121 \times 205117922 - 159018721 \times 83739041 = \\ 13316075197586562 - 13316075197586561 = 1; \end{aligned} \quad (1.1)$$

If Formula (1.1) is carried out using conventional floating-point arithmetic:

$$\begin{aligned} 64919121 \times 205117922 - 159018721 \times 83739041 = \\ 64919121.000000000 \times 205117922.000000000 - 159018721.000000000 \times 83739041.000000000 = \\ 13316075197586562. - 13316075197586560. = 2. = 2.0000000000000000; \end{aligned} \quad (1.2)$$

1. The multiplication results exceed the maximal significance of the 64-bit IEEE floating-point representation; so they are rounded off, generating rounding errors;
2. The normalization of the subtraction result amplifies the rounding error to most significant bit (MSB) by padding zeros.

Formula (1.2) is a showcase for the problem of conventional floating-point arithmetic. Because normalization happens after each arithmetic operation [8][9][10], such generation of rounding errors happens very frequently for addition and multiplication, and such amplification of rounding errors happens very frequently for subtraction and division. The accumulation of rounding errors is an intrinsic problem of conventional floating-point arithmetic [12], and in the majority of cases such accumulation is almost uncontrollable [11]. For example, because a rounding error from lower digits quickly propagates to higher digits, the 10^{-7} precision of the 32-bit IEEE floating-point format [8][9][10] is usually not fine enough for calculations involving input data of 10^{-2} to 10^{-4} precision.

Self-censored rules are developed to avoid such rounding error propagation [12][13], such as avoiding subtracting results of large multiplication, as in Formula (1.2). However, these rules are not enforceable, and in many cases are difficult to follow, e.g., even a most carefully crafted algorithm can result in numerical instability after extensive usage. Because the propagation speed of a rounding error depends on the nature of a calculation itself, e.g., generally faster in nonlinear algorithms than linear algorithms¹ [14], propagation of rounding error in conventional floating-point arithmetic is very difficult to quantify generically [15]. Thus, it is difficult to tell if a calculation is improper or becomes excessive for a required result precision. In common practice, reasoning on an individual theoretical base is used to estimate the error and validity of calculation results, such as from the estimated transfer functions of the algorithms used in the calculation [12][16][17]. However, such analysis is both rare and generally very difficult to carry out in practice.

Today most experimental data are collected by an ADC (Analog-to-Digital Converter) [5]. The result obtained from an ADC is an integer with fixed uncertainty; thus, a smaller signal value has a coarser precision. When a waveform containing raw digitalized signals from ADC is converted into conventional floating-point representation, the information content of the digitalized waveform is distorted to favour small signals since all converted data now have the same and best possible precision. However, the effects of such distortion in signal processing are generally not clear.

What is needed is a floating-point arithmetic that tracks precision automatically. When the calculation is improper or becomes excessive, the results become insignificant. All existing uncertainty-bearing arithmetics are reviewed below.

1.3 Interval Arithmetic

Interval arithmetic [13][18][19][20][21][22] is currently a standard method to track calculation uncertainty. It ensures that the value x is absolutely bounded within its *bounding range* $[x] \equiv [\underline{x}, \bar{x}]$, in which \underline{x} and \bar{x} are lower and upper bounds for x , respectively. In this paper, interval arithmetic is simplified and tested as the following arithmetic formulas² [20]:

$$[x] + [y] = [\underline{x} + \underline{y}, \bar{x} + \bar{y}]; \quad (1.3)$$

$$[x] - [y] = [\underline{x} - \bar{y}, \bar{x} - \underline{y}]; \quad (1.4)$$

$$[x] \times [y] = [\min(\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y}), \max(\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y})]; \quad (1.5)$$

$$0 \notin [y] : [x] / [y] = [x] \times [1/\bar{y}, 1/\underline{y}]; \quad (1.6)$$

If interval arithmetic is implemented using a floating-point representation with limited resolution, its resulting bounding range is widened further [19].

A basic problem is that the bounding range used by interval arithmetic is not compatible with usual scientific and engineering measurements, which instead use the statistical mean and deviations to characterize uncertainty [1][2]. Most measured values are well approximated by a Gaussian distribution [1][2][4], which has no limited bounding range. Let *bounding leakage* be defined as the possibility of the true value to be outside a bounding range. If a bounding range is defined using a statistical rule

¹A classic example is the contrast of the uncertainty propagation in the solutions for the 2nd-order linear differential equation vs. in those of Duffing equation (which has a x^3 term in addition to the x term in a corresponding 2nd-order linear differential equation).

²For the mathematical definition of interval arithmetic, please see [22].

on bounding leakage, such as the $6\sigma - 10^{-9}$ rule for Gaussian distribution [4] (which says that the bounding leakage is about 10^{-9} for a bounding range of mean ± 6 -fold of standard deviations), there is no guarantee that the calculation result will also obey the $6\sigma - 10^{-9}$ rule using interval arithmetic, since interval arithmetic has no statistical foundation³.

Another problem is that interval arithmetic only provides the worst case of uncertainty propagation, so that it tends to over-estimate uncertainty in reality. For instance, in addition and subtraction, it gives the result when the two operands are +1 and -1 correlated respectively [24]. However, if the two operands are -1 and +1 correlated respectively instead, the actual bounding range after addition and subtraction reduces, which is called the best case in random interval arithmetic [25]. The vast overestimation of bounding ranges in these two worst cases prompts the development of affine arithmetic [24][26], which traces error sources using a first-order model. Being expensive in execution and depending on approximate modeling even for such basic operations as multiplication and division, affine arithmetic has not been widely used. In another approach, random interval arithmetic [25] reduces the uncertainty over-estimation of standard interval arithmetic by randomly choosing between the best-case and the worst-case intervals.

A third problem is that the results of interval arithmetic may depend strongly on the actual expression of an analytic function $f(x)$. For example, Formula (1.7), Formula (1.8) and Formula (1.9) are different expressions of the same $f(x)$; however, the correct result is obtained only through Formula (1.7), and uncertainty may be exaggerated in the other two forms, e.g., by 67-fold and 33-fold at input range [0.49, 0.51] using Formula (1.8) and Formula (1.9), respectively. This is called the dependence problem of interval arithmetic [21]. There is no known generic way to apply Taylor expansion in interval arithmetic, so that the dependence problem is an intrinsic problem for interval arithmetic.

$$f(x) = (x - 1/2)^2 - 1/4; \quad (1.7)$$

$$f(x) = x^2 - x; \quad (1.8)$$

$$f(x) = (x - 1)x; \quad (1.9)$$

Interval arithmetic has very coarse and algorithm-specific precision but constant zero bounding leakage. It represents the other extreme from conventional floating-point arithmetic. To meet practical needs, a better uncertainty-bearing arithmetic should be based on statistical propagation of the rounding error, while also allowing reasonable bounding leakage for normal usages, which is achieved in the variance arithmetic, as show later in this paper.

As shown previously [38], interval arithmetic grossly exaggerates calculation uncertainty, e.g., for $f^{-1}(f(x))$ the result uncertainty is much larger than the original uncertainty. On the other hand, as shown later in this paper, the variance arithmetic provides stable bounding for a given bounding leakage, e.g., ± 3.1 -fold of the result deviation from the mean for a bounding leakage of 2×10^{-7} .

³There is some attempt [23] to connect intervals in interval arithmetic to confidence interval or the equivalent so called p-box in statistics. Because this attempt seems to rely heavily on 1) specific properties of the uncertainty distribution within the interval and/or 2) specific properties of the functions upon which the interval arithmetic is used, this attempt does not seem to be generic. Anyway, this attempt seems to be outside the main course of interval arithmetic, which has no statistics in mind.

1.4 Statistical Propagation of Uncertainty

If each operand is regarded as a random variable, and the statistical correlation between the two operands is known, the resulting uncertainty is given by the *statistical propagation of uncertainty* [27][28], with the following arithmetic equations, in which σ is the deviation of a measured value x , P is its precision, and γ is the correlation between the two imprecise values:

$$(x \pm \delta x) + (y \pm \delta y) = (x + y) \quad \pm \sqrt{\delta x^2 + \delta y^2 + 2\delta x \delta y \gamma}; \quad (1.10)$$

$$(x \pm \delta x) - (y \pm \delta y) = (x - y) \quad \pm \sqrt{\delta x^2 + \delta y^2 - 2\delta x \delta y \gamma}; \quad (1.11)$$

$$(x \pm \delta x) \times (y \pm \delta y) = (x \times y) \quad \pm |x \times y| \sqrt{P_x^2 + P(y)^2 + 2P_x P(y) \gamma}; \quad (1.12)$$

$$(x \pm \delta x)/(y \pm \delta y) = (x/y) \quad \pm |x/y| \sqrt{P_x^2 + P(y)^2 - 2P_x P(y) \gamma}; \quad (1.13)$$

Tracking uncertainty propagation statistically seems a better solution. However, in practice, the correlation between two operands is generally not precisely known, so the direct use of statistical propagation of uncertainty is very limited. As show later in this paper, the variance arithmetic is based on a statistical assumption much more lenient than knowing the correlation between imprecise values.

In this paper, as a proxy for statistical propagation of uncertainty, an *independence arithmetic* always assumes that no correlation exists between any two operands, whose arithmetic equations are Formula (1.10), Formula (1.11), Formula (1.12) and Formula (1.13), where $\gamma = 0$. Independence arithmetic is actually de facto arithmetic in engineering data processing, such as in the common belief that uncertainty after averaging reduces by the square root of number of measurements [1][2], or the ubiquitous Monte Carlo method⁴ [30][29], or calculating the mean and variance of a Taylor expansion [31]. Perhaps, it is reasonable to assume the uncertainties of experimental measurements to be independent of each other, while is it not reasonable to assume the uncertainties of calculations to be independent, which is essentially the approach of variance arithmetic.

1.5 Significance Arithmetic

Significance arithmetic [32] tries to track reliable bits in an imprecise value during the calculation. In the two early attempts [33][34], the implementations of significance arithmetic are based on simple operating rules upon reliable bit counts, rather than on formal statistical approaches. They both treat the reliable bit counts as integers when applying their rules, while in reality a reliable bit count could be a fractional number [35], so they both can cause artificial quantum reduction of significance. The significance arithmetic marketed by Mathematica [35] uses a linear error model that is consistent with a first-order approximation of interval arithmetic [13][20][21], and further provides an arbitrary precision representation which is in the framework of conventional floating-point arithmetic. It is definitely not a statistical approach.

Stochastic arithmetic [15] [36], which can also be categorized as significance arithmetic, randomizes the least significant bits (LSB) of each of input floating-point values,

⁴Most but not all applications of Monte Carlo methods assume independence between any two random variables. In a minority of applications, a Monte Carlo method can be used to construct specified correlation between two random variables [29].

repeats the same calculation multiple times, and then uses statistics to seek invariant digits among the calculation results as significant digits. This approach may require too much calculation since the number of necessary repeats for each input is specific to each algorithm, especially when the algorithm contains branches. Its sampling approach may be more time-consuming and less accurate than direct statistical characterization [4], such as directly calculating the mean and deviation of the underlying distribution. It is based on modeling rounding errors in conventional floating-point arithmetic, which is quite complicated. A better approach may be to define arithmetic rules that make error tracking by probability easier.

One problem of significance arithmetic is that itself can not properly specifies the uncertainty [38]. For example, if the least significand bit of significand is used to specify uncertainty, then the representation have very coarse precision, such as $1 \pm 10^{-3} = 1024 \times 2^{-10}$ [38]. Introducing limited bits calculated inside uncertainty can not avoid this problem completely. Thus, the resolution of the conventional floating-point representation is desired. This is the reason why variance arithmetic abandoned the significance arithmetic nature of its predecessor [38].

1.6 An Overview of This Paper

In this paper, a new floating-point arithmetic called *variance arithmetic* is developed to track uncertainty during floating-point calculations of analytic functions, as described in Section 2. Compare to its predecessor [38]:

- The introduction section 1 remains almost unchanged, except it quotes the negative results from [38] for interval arithmetic and significance arithmetic.
- The theoretical foundation section 2 has the following significant differences:
 1. Both have identical uncorrelated uncertainty assumption.
 2. The variance arithmetic has different digital presentation.
 3. The variance arithmetic has similar Taylor expansion formulas, but the variance is centered around statistical means instead of the calculation results assuming no uncertainty, so variance arithmetic agrees with statistics strictly.
 4. The theoretical convergence and statistical binding leakage of the result uncertainties of the variance arithmetic are provided.
 5. The variance arithmetic has no dependency problem for analytic questions. Instead, the variance arithmetic uses statistics to trace dependency.
- Section 3 remains largely unchanged, except major improvement on the descriptions of validation methods and standards.
- Section 6 is new, to provide a statistical test of variance arithmetic in using the common math library functions.
- Section 8 reproduces the previous result for the input data with added noises. It added a new signal for test. It also evaluates the calculation errors of the library sin and tan functions.

2 The Variance Arithmetic

2.1 Foundation for the Variance Arithmetic

The statistical precision $P(x) = \delta x/|x|$ represents the reliable information content of a measurement statistically, with finer or smaller statistical precision means higher reliability and better reproducibility of the measurement [1][2]. $P(x)$ has an upper bound of 1. When $1 \leq P(x)$:

- Either δx is too coarse to determine the actual value of x ,
- Or $x \pm \delta x$ is a measurement of 0.

The inputs to variance arithmetic should obeys the *uncorrelated uncertainty assumption* [38], which means that the uncertainties of any two different inputs can be regarded as uncorrelated of each other. This assumption is consistent with the common methods in processing experimental data [1][2], such as the common knowledge or belief that precision improves with the count n as $1/\sqrt{n}$ during averaging. It is much weaker than requiring any two different inputs to be independent of each other. It can be turned into a realistic and simple statistical requirement or test between two inputs.

Taylor expansion can be used to calculate the result variance of analytic functions.

The uncorrelated uncertainty assumption only applies to imprecise inputs. When the inputs obeys the *uncorrelated uncertainty assumption*, in the intermediate steps, the variance arithmetic uses statistics to account for the correlation between uncertainties through their associated variables. Thus, the variance arithmetic has no dependency problem.

2.2 The Uncorrelated Uncertainty Assumption [38]

When there is a good estimation of the sources of uncertainty, the uncorrelated uncertainty assumption can be judged directly, e.g., if noise [1][2] is the major source of uncertainty, the uncorrelated uncertainty assumption is probably true. This criterion is necessary to ascertain repeated measurements of the same signal. Otherwise, the uncorrelated uncertainty assumption can be judged by the correlation and the respectively precision of two measurements.

The correlated parts common to different measurements are regarded as signals, which can either be desired or unwanted. Let X , Y , and Z denote three mutually independent random variables [4] with variance $V(X)$, $V(Y)$ and $V(Z)$, respectively. Let α denote a constant. Let $Cov()$ denote the covariance function. Let γ denote the correlation between $(X + Y)$ and $(\alpha X + Z)$. And let:

$$\eta_1^2 \equiv \frac{V(Y)}{V(X)}; \quad \eta_2^2 \equiv \frac{V(Z)}{V(\alpha X)} = \frac{V(Z)}{\alpha^2 V(X)}; \quad (2.1)$$

$$\gamma = \frac{Cov(X + Y, \alpha X + Z)}{\sqrt{V(X + Y)}\sqrt{V(\alpha X + Z)}} = \frac{\alpha/|\alpha|}{\sqrt{1 + \eta_1^2}\sqrt{1 + \eta_2^2}} \equiv \frac{\alpha/|\alpha|}{1 + \eta^2}; \quad (2.2)$$

Formula (2.4) gives the correlation γ between two random variables, each of which contains a completely uncorrelated part and a completely correlated part X , with η being the average ratio between these two parts. Formula (2.4) can also be interpreted reversely: if two random variables are correlated by γ , each of them can be viewed hypothetically as containing a completely uncorrelated part and a completely correlated part, with η being the average ratio between these two parts.

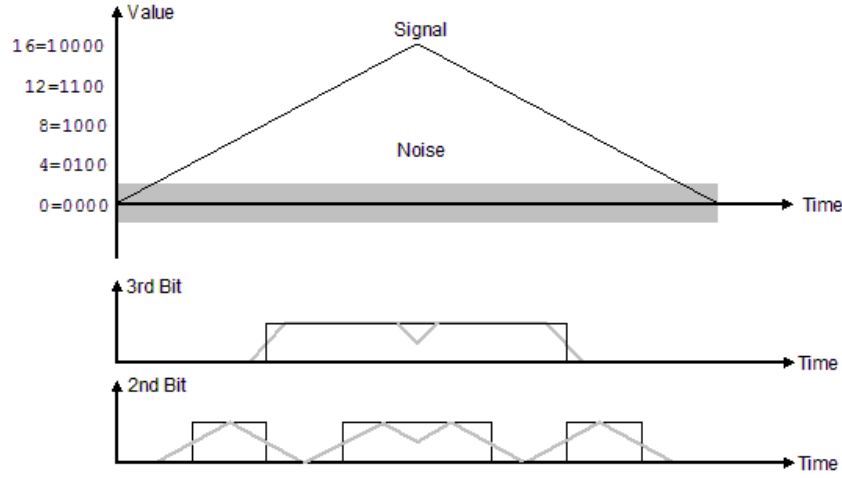


Figure 1: Effect of noise on bit values of a measured value. The triangular wave signal and the added white noise are shown at top using the thin black line and the grey area, respectively. The values are measured by a theoretical 4-bit Digital-to-Analog Converter in ideal condition, assuming LSB is the 0th bit. The measured 3rd and 2nd bits without the added noise are shown using thin black lines, while the mean values of the measured 3rd and 2nd bits with the added noise are shown using thin grey lines. This figure has been published previously [38].

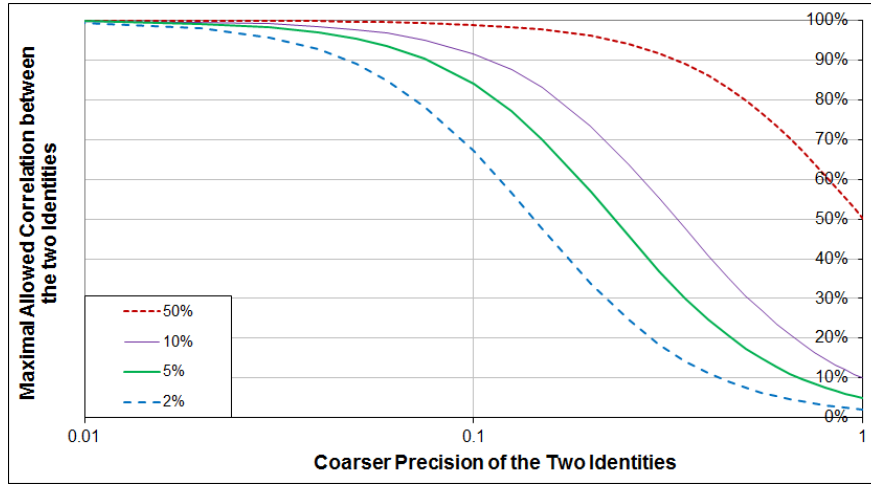


Figure 2: Allowed maximal correlation between two values vs. input precisions and independence standard (as shown in legend) for the independence uncertainty assumption of variance arithmetic to be true. This figure has been published previously [38].

One special application of Formula (2.4) is the correlation between a measured signal and its true signal, in which noise is the uncorrelated part between the two. Figure 1 shows the effect of noise on the most significant two bits of a 4-bit measured signal when $\eta = 1/4$. Its top chart shows a triangular waveform between 0 and 16 as a black line, and a white noise between -2 and +2, using the grey area. The measured signal is the sum of the triangle waveform and the noise. The middle chart of Figure 1 shows the values of the 3rd digit of the true signal as a black line, and the mean values of the 3rd bit of the measurement as a grey line. The 3rd bit is affected by the noise during its transition between 0 and 1. For example, when the signal is slightly below 8, only a small positive noise can turn the 3rd digit from 0 to 1. The bottom chart of Figure 1 shows the values of the 2nd digit of the signal and the measurement as a black line and a grey line, respectively. Figure 1 clearly shows that the correlation between the measurement and the true signal is less at the 2nd digit than at the 3rd digit. Quantitatively, according to Formula (2.4):

1. The overall measurement is 99.2% correlated to the signal with $\eta = 1/8$;
2. The 3rd digit of the measurement is 97.0% correlated to the signal with $\eta = 1/4$;
3. The 2nd digit of the measurement is 89.4% correlated to the signal with $\eta = 1/2$;
4. The 1st digit of the measurement is 70.7% correlated to the signal with $\eta = 1$;
5. The 0th digit of the measurement is 44.7% correlated to the signal with $\eta = 2$.

The above conclusion agrees with the common experiences that, below the noise level of measured signals, noises rather than true signals dominate each digit.

Similarly, while the correlated portion between two values has exactly the same value at each bit of the two values, the ratio of the uncorrelated portion to the correlated portion increases by 2-fold for each bit down from MSB of the two values. Quantitatively, let P denote the precision of an imprecise value, and let η_P denote the ratio of the uncorrelated portion to the correlated portion at level of uncertainty; then η_P increases with decreased P according to Formula (2.3). According to Formula (2.4), if two significant values are overall correlated with γ , at the level of uncertainty the correlation between the two values decreases to γ_P according to Formula (2.4).

$$\eta_P = \frac{\eta}{P}, \quad P < 1; \quad (2.3)$$

$$\frac{1}{\gamma_P} - 1 = \left(\frac{1}{\gamma} - 1 \right) \frac{1}{P^2}, \quad P < 1; \quad (2.4)$$

Figure 2 plots the relation of γ vs. P for each given γ_P in Formula (2.4). When γ_P is less than a predefined maximal threshold (e.g., 2%, 5% or 10%), the two values can be deemed virtually uncorrelated of each other at the level of uncertainty. Thus for each independence standard γ_P , there is a maximal allowed correlation between two values below which the uncorrelated uncertainty assumption of variance arithmetic holds. The maximal allowed correlation is a function of the larger precision of the two values according to Formula (2.4). Figure 2 shows that for two precisely measured values, their correlation γ is allowed to be quite high. Thus, the uncertainty assumption has much weaker statistical requirement than the assumption for independence arithmetic, which requires the two values to be independent of each other.

It is tempting to add noise to otherwise unqualified values to make them satisfy the uncertainties uncertainty assumption. As an extreme case of this approach, if two

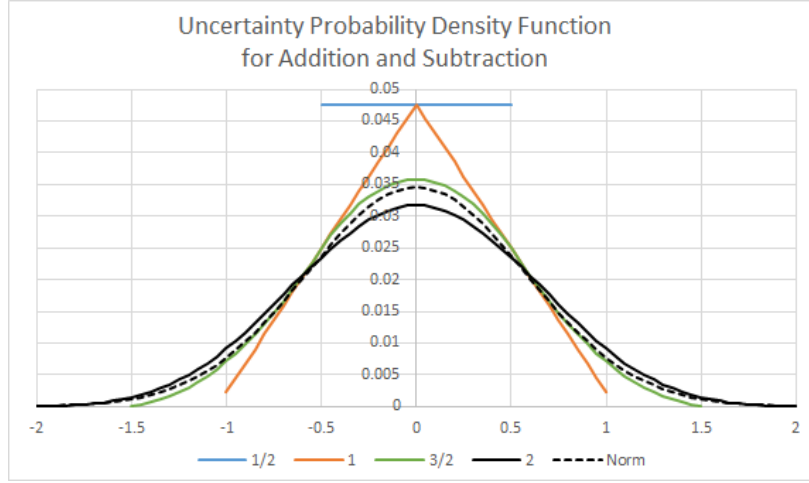


Figure 3: The probability distribution $P_{\frac{n}{2}}(x)$ of Formula (2.6), assuming $P_{\frac{1}{2}}(x)$ is uniformly distributed between $(-1/2, +1/2)$. The legend shows n . $P_{\frac{n}{2}}(x)$ is already close to the corresponding Gaussian distribution with the same mean and deviation, which is plotted in dash line of the same color. This figure has been published previously [38].

values are constructed by adding noise to the same signal, they are 50% correlated at the uncertainty level so that they will not satisfy the uncorrelated uncertainty assumption, which defines the upper bound for γ_P .⁵

2.3 Result Uncertainty For Pure Rounding [38]

Empirically, the pure rounding error due to round-off is shown to be uniformly distributed between the half bit of the least significant bit of the significand [38], whose probability density function is shown in Formula (2.5):

$$P_{\frac{1}{2}}(x) \equiv 1, \quad -1/2 \leq x \leq +1/2; \quad (2.5)$$

2.4 Result Uncertainty For Addition and Subtraction [38]

The uncorrelated uncertainty assumption suggests that the result bounding distribution density function of addition or subtraction is the convolution of the input density functions [4].

$$P_{\frac{n}{2}}(x) \equiv \int_{-\infty}^{+\infty} P_{\frac{1}{2}}(y) P_{\frac{n-1}{2}}(x-y) dy = \int_{-1/2}^{+1/2} P_{\frac{n-1}{2}}(x-y) dy, \quad n = 2, 3, 4 \dots; \quad (2.6)$$

⁵The 50% curve in Figure 2 thus defines the maximal possible correlations between any two measured signals. This other conclusion of Formula (2.4) makes sense because the measurable correlation between two measurements should be limited by the precision of their measurements..

Figure 3 shows the probability distribution $P_{\frac{n}{2}}(x)$ of Formula (2.6). $P_2(x)$ is already close to the corresponding Gaussian distribution with the same mean and deviation, which is plotted in dash line of the same color.

The Lyapunov form of the central limit theorem [4] states that if X_i is a random variable with mean μ_i and variance V_i for each i among a series of n mutually independent random variables, then with increased n , the sum $\sum_i^n X_i$ converges in distribution to the Gaussian distribution with mean $\sum_i^n \mu_i$ and variance $\sum_i^n V_i$. Applying the central limit theorem to the addition and subtraction:

- $P_{n/2}(x)$ converges in distribution to a Gaussian distribution.
- Figure 3 shows that such convergence to Gaussian distribution is very quick.
- The stable rounding error distribution is independent of any initial rounding error distribution.

Also due to the center-limit theorem, uncertainty in general is expected to be Gaussian distributed [1] [4]. The rounding error distribution is extended to describe uncertainty distribution in general.

For simplicity, define operator $\delta^2 f(x) \equiv (\delta f(x))^2$. Formula (2.7) give the result variance of addition and subtraction surrounding $x \pm y$. Formula (2.8) gives the probability density function for $x \pm \delta x$ in general, in which $N()$ is the density function of the normal distribution [4]. Formula (2.8) can be normalized as Formula (2.9).

$$\delta^2(x \pm y) = (\delta x)^2 + (\delta y)^2; \quad (2.7)$$

$$\rho(\tilde{x}, x, \delta x) = \frac{1}{\delta x} N\left(\frac{\tilde{x} - x}{\delta x}\right); \quad (2.8)$$

$$\tilde{z} \equiv \frac{\tilde{x} - x}{\delta x} : \rho(\tilde{x}, x, \delta x) d\tilde{x} = N(\tilde{z}) d\tilde{z}; \quad (2.9)$$

Formula (2.7) is different from Formula (1.10) and (1.11), because in variance arithmetic, the correlation between x and y does not contribute to the result distribution as far as the uncorrelated uncertainty assumption is satisfied. It may show that Formula (1.10) and (1.11) are wrong statistically for the result uncertainty.

2.5 A Variance Representation

A variance arithmetic representation uses a pair of conventional floating-point numbers to represent the value and the variance of an imprecise value. The square root of a variance is defined as an *uncertainty*. When the uncertainty is not specified:

- A conventional floating-point value has $1/\sqrt{3}$ -fold of its least significant value as its uncertainty, in which the *least significant value* is defined as the equivalent value of the least significant bit of the significand of the conventional floating-point value, e.g., the least significant value of 1.0 is 2^{-52} . Essentially, the floating value is regarded as containing a rounding error uniformly distributed within the least significant bit of its significand.
- An integer value within the range of $(-2^{53}, +2^{53})$ has 0 as its uncertainty.
- An integer value outside the range of $(-2^{53}, +2^{53})$ is converted to a conventional floating-point value.

2.6 Comparison

Two imprecise values can be compared statistically using their difference.

When the value difference is zero, the two imprecise values are equal. Such definition of equality deviates from statistics. In statistics, such two imprecise value has 50% possibility to be less than or greater to each other but no chance to be equal to each other.

Otherwise, the standard z-method [4] is used to judge whether they are equal to each other, or less or more than each other. For example, the difference between 1.002 ± 0.001 and 1.000 ± 0.002 is 0.002 ± 0.00224 , so that the z-value for the difference is $z = 0.002/0.00224$, and the probability for them to be not equal is $\xi(|z|) - \xi(-|z|) = 62.8\%$, in which $\xi(z)$ is the cumulative density function for normal distribution [4]. If the threshold probability of not equality is 50%, $1.000 \pm 0.002 < 1.002 \pm 0.001$. Instead of using the threshold probability, the bounding range for z can be used, such as $|z| \leq 0.67448975$ for the equivalent threshold probability of 50%.

2.7 Multiplication

To obey the precision scaling principle, the result of multiplying $x \pm \delta x$ by a precise value y is $y^2 \delta^2 x$. The result of multiplying $0 \pm \delta x$ by $0 \pm \delta y$ is a normal production distribution [4], which is centered at 0 with variance $(\delta x)^2 (\delta y)^2$. The general multiplication can be decomposed as Formula (2.10), which leads to Formula (2.11) [38] and (2.12) [38] for the result variance and precision surrounding the result xy , respectively.

$$(x \pm \delta x) \times (y \pm \delta y) = (x + (0 \pm \delta x)) \times (y + (0 \pm \delta y)); \quad (2.10)$$

$$\delta^2(xy) = x^2(\delta^2 y) + y^2(\delta^2 x) + (\delta^2 x)(\delta^2 y); \quad (2.11)$$

$$xy \neq 0 : P(xy)^2 = P(x)^2 + P(y)^2 + P(x)^2 P(y)^2; \quad (2.12)$$

Formula (2.11) is identical to Formula (1.12) for statistical propagation of uncertainty except their cross term, representing difference in their statistical requirements, respectively. Formula (2.11) is simpler and more elegant than Formula (1.5) for the interval arithmetic multiplication. The result of Formula (1.2) calculated by variance arithmetic is $2 \pm 2\sqrt{5}$. It is 2 ± 9 for interval arithmetic [19].

2.8 Uncertainty Distribution

Let $\tilde{y} = f(\tilde{x})$ be a strictly monotonic function, so that $\tilde{x} = f^{-1}(\tilde{y})$ exist. Formula (2.13) is the uncertainty distribution density function [1]. In Formula (2.13) [1], the same distribution can be expressed in either \tilde{x} or \tilde{y} or \tilde{z} , which are different representations of the same underlying random variable. Using Formula (2.13), Formula (2.14), (2.15), and (2.16) give the $\rho(\tilde{y}, y, \delta y)$ for e^x , $\ln(x)$, and x^c , respectively.

$$N(\tilde{z})d\tilde{z} = \rho(\tilde{x}, x, \delta x)d\tilde{x} = \rho(f^{-1}(\tilde{y}), x, \delta x) \frac{d\tilde{x}}{d\tilde{y}} d\tilde{y} = \rho(\tilde{y}, y, \delta y)d\tilde{y}; \quad (2.13)$$

$$y = e^x : \rho(\tilde{y}, y, \delta y) = \frac{1}{\tilde{y}} \frac{1}{\delta x} N\left(\frac{\log(\tilde{y}) - x}{\delta x}\right); \quad (2.14)$$

$$y = \ln(x) : \rho(\tilde{y}, y, \delta y) = e^{\tilde{y}} \frac{1}{\delta x} N\left(\frac{e^{\tilde{y}} - x}{\delta x}\right); \quad (2.15)$$

$$y = x^c : \rho(\tilde{y}, y, \delta y) = c\tilde{y}^{\frac{1}{c}-1} \frac{1}{\delta x} N\left(\frac{\tilde{y}^{\frac{1}{c}} - x}{\delta x}\right); \quad (2.16)$$

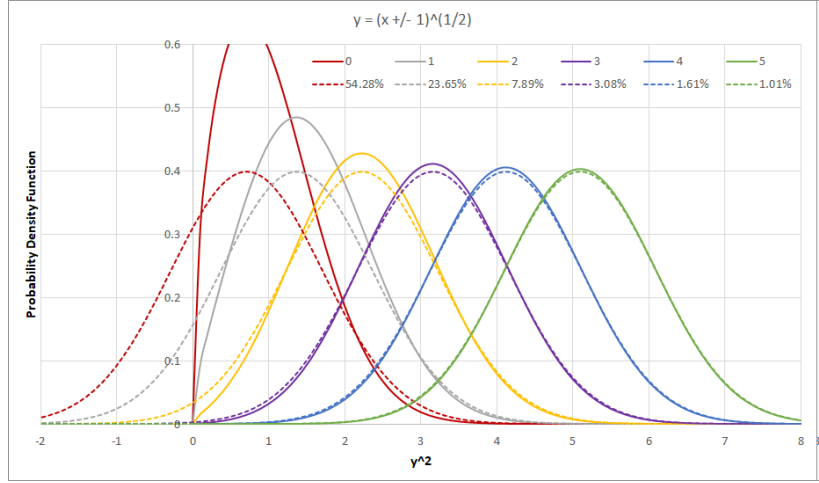


Figure 4: The probability density function for $\tilde{y} = \sqrt{x \pm 1}$, for different x as shown in the legend. The x -axis is scaled as \tilde{y}^2 . Each probability density function $\rho(\tilde{x})$ in solid line is compared with a Gaussian distribution $\varrho(\tilde{x})$ of the same mode and the same deviation in the dash line of the same color. The legend for the Gaussian distribution shows the value of Gaussian difference in percentage.

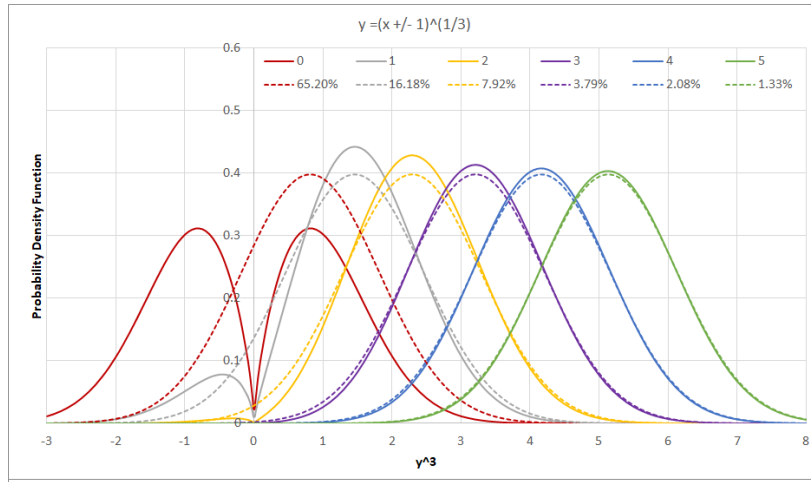


Figure 5: The probability density function for $\sqrt[3]{x \pm 1}$, for different x as shown in the legend. The y -axis is scaled as \tilde{y}^3 . Each probability density function $\rho(\tilde{x})$ in solid line is compared with a Gaussian distribution $\varrho(\tilde{x})$ of the same mode and the same deviation in the dash line of the same color. The legend for the Gaussian distribution shows the value of Gaussian difference in percentage.

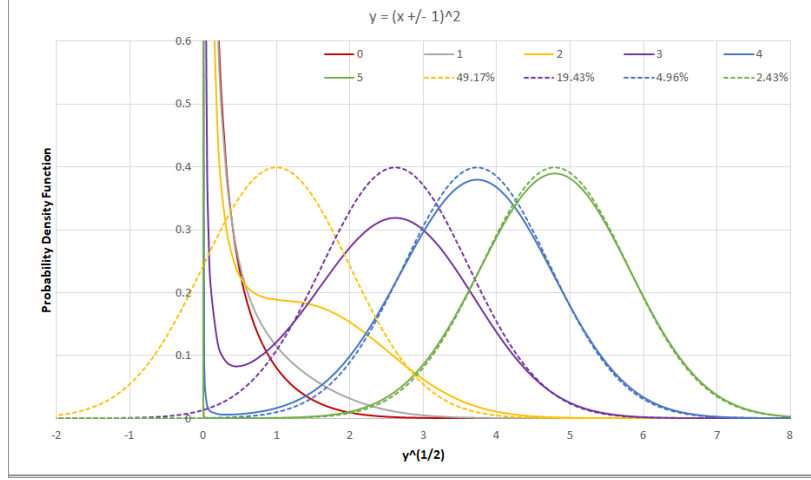


Figure 6: The probability density function for $(x \pm 1)^2$, for different x as shown in the legend. The y -axis is scaled as \sqrt{y} . Some probability density function $\rho(\tilde{x})$ in solid line is compared with a Gaussian distribution $\varrho(\tilde{x})$ of the same mode and the same deviation in the dash line of the same color. The legend for the Gaussian distribution shows the value of Gaussian difference in percentage.

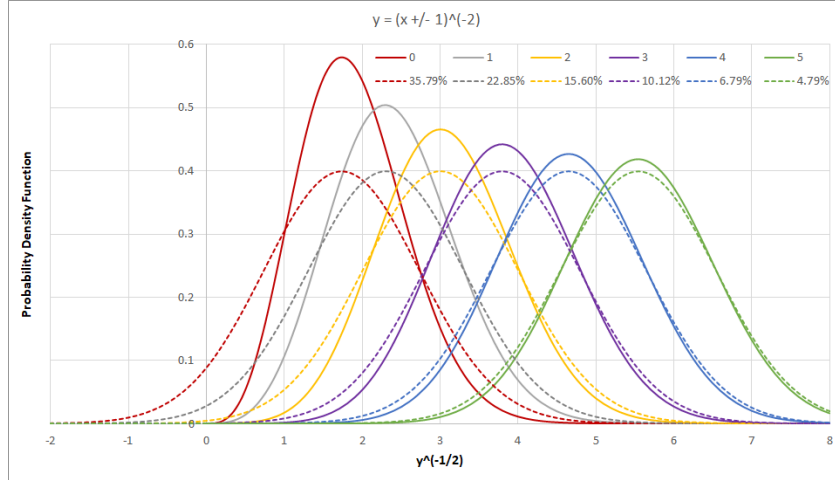


Figure 7: The probability density function for $1/(x \pm 1)^2$, for different x as shown in the legend. The y -axis is scaled as $1/\sqrt{y}$. The probability density function $\rho(\tilde{x})$ in solid line is compared with a Gaussian distribution $\varrho(\tilde{x})$ of the same mode and the same deviation in the dash line of the same color. The legend for the Gaussian distribution shows the value of Gaussian difference in percentage.

Using Formula (2.13), Formula (2.17) gives the equation for the mode of the result distribution, whose solutions $f^{-1}(\tilde{y}_m)$ for e^x , $\ln(x)$, and x^c are Formula (2.18), (2.19), and (2.20) respectively.

$$\tilde{z} = \frac{f^{-1}(\tilde{y}) - x}{\delta x} : \quad \frac{d^2 \tilde{z}}{d\tilde{y}^2} = \frac{d\tilde{z}}{d\tilde{y}}; \quad (2.17)$$

$$\tilde{y} = e^{\tilde{x}} : \quad \ln(\tilde{y}_m) = x - (\delta x)^2; \quad (2.18)$$

$$\tilde{y} = \ln(\tilde{x}) : \quad e^{\tilde{y}_m} = \frac{x + \sqrt{x^2 + 4(\delta x)^2}}{2}; \quad (2.19)$$

$$\tilde{y} = \tilde{x}^c : \quad (\tilde{y}_m)^{1/c} = \frac{x + \sqrt{x^2 - 4(c-1)(\delta x)^2}}{2}; \quad (2.20)$$

The *Gaussian difference* of $\rho(f^{-1}(\tilde{y}), x, \delta x)$ is calculated as $\int_{-\infty}^{+\infty} |\rho(f^{-1}(\tilde{y}), x, \delta x) - \varrho(\tilde{y})| d\tilde{y}$ in which $\varrho(\tilde{y})$ is the Gaussian distribution of the same mode and the same deviation δx . A Gaussian difference of 0 means a perfect Gaussian.

Viewed in the $f^{-1}(\tilde{y})$ coordinate, $\rho(\tilde{y}, y, \delta y)$ is Gaussian modulated by $\frac{d\tilde{x}}{d\tilde{y}} = 1/f_x^{(1)}$. A *zero* of the uncertainty distribution happens when $f_x^{(1)} = \infty \rightarrow \rho(\tilde{y}, y, \delta y) = 0$, while a *pole* happens when $f_x^{(1)} = 0 \rightarrow \rho(\tilde{y}, y, \delta y) = \infty$. Zeros and poles have different impacts on the properties of the uncertainty distributions.

If $y = f(x)$ is a linear transform of x , $f_x^{(1)}$ is a constant, and $\rho(\tilde{y}, y, \delta y)$ is known to be Gaussian [4]. From another prospective, a linear transformation generates neither zero nor pole according Formula (2.13).

Figure 4 shows the probability density function for $\sqrt{x \pm 1}$ according to Formula (2.16), which has a zero at $x = 0$. Each probability density function $\rho(\tilde{x})$ in solid line is compared with a normal distribution $\varrho(\tilde{x})$ of the same mode in dash line of the same color.

Figure 5 shows the probability density function for $\sqrt[3]{x \pm 1}$ according to Formula (2.16), which also has a zero at $x = 0$. Compared with $\sqrt{x \pm 1}$, $\sqrt[3]{x \pm 1}$ distributes on $\tilde{y} < 0$ also, so that the uncertainty distributions for $\sqrt[3]{0 \pm 1}$ has two equal peaks instead of one larger peak on the positive side only for $\sqrt{x \pm 1}$.

Figure 6 shows the probability density function for $(x \pm 1)^2$ according to Formula (2.16), which has a pole at $x = 0$. The uncertainty distributions for $(0 \pm 1)^2$ and $(1 \pm 1)^2$ deviate significantly from Gaussian, while the uncertainty distributions for $(4 \pm 1)^2$ and $(5 \pm 1)^2$ look Gaussian but each still with an infinitive but narrower peak at $\tilde{y} = 0$. According to Formula (2.20), $(0 \pm 1)^2$ and $(1 \pm 1)^2$ has no mode. $(0 \pm 1)^2$ is actually the χ^2 distribution [1].

Figure 7 shows the probability density function for $1/(x \pm 1)^2$ according to Formula (2.16), which has a zero at $x = 0$. Figure 7 looks more like Figure 4 than Figure 6, showing that pole or zero determines the properties of uncertainty distributions.

In all the figures, the probability density function in the \tilde{z} representation becomes more Gaussian-like when the mode of the distributions is further away from either zero or pole. Such observation leads to the following Taylor method to calculate the result variances when the distribution is nearly Gaussian in the \tilde{z} representation.

2.9 Analytic Functions

Formula (2.13) gives the uncertainty distribution of an analytic function. However, normal scientific and engineering calculations usually do not care about the result distribution, but just some simple statistics of the result, such as the result deviation

[1] [2]. Using Taylor expansion is one way to get the simple statistics when the result uncertainty distribution is nearly Gaussian.

2.9.1 Uncertainty of Taylor Expansion

An analytic function $f(x)$ can be accurately given in a range by the Taylor series as shown in Formula (2.21). Using Formula (2.9), Formula (2.23) and (2.24) [38] gives the mean $\overline{f(x)}$ and the variance $\delta^2 f(x)$, respectively, in which $\zeta(n)$ is the *variance momentum* defined by Formula (2.22), and σ is defined as the *bounding factor*.

$$\tilde{y} = f(x + \tilde{x}) = f(x + \tilde{z}\delta x) = \sum_{n=0}^{\infty} \frac{f_x^{(n)}}{n!} \tilde{z}^n (\delta x)^n; \quad (2.21)$$

$$\zeta(n) \equiv \int_{-\sigma}^{+\sigma} \tilde{z}^n N(\tilde{z}) d\tilde{z}; \quad \zeta(2n+1) = 0; \quad (2.22)$$

$$\overline{f(x)} = \int_{-\sigma}^{+\sigma} f(x + \tilde{z}\delta x) N(\tilde{z}) d\tilde{z} = f(x) + \sum_{n=1}^{\infty} (\delta x)^{2n} \frac{f_x^{(2n)} \zeta(2n)}{(2n)!}; \quad (2.23)$$

$$\begin{aligned} \delta^2 f(x) &= \overline{f(x)^2} - \overline{f(x)}^2 = \sum_{n=0}^{\infty} (\delta x)^{2n} \\ &\quad \left(\zeta(2n) \sum_{j=1}^{2n-1} \frac{f_x^{(j)}}{j!} \frac{f_x^{(2n-j)}}{(2n-j)!} - \sum_{j=1}^{n-1} \frac{f_x^{(2j)} \zeta(2j)}{(2j)!} \frac{f_x^{(2n-2j)} \zeta(2n-2j)}{(2n-2j)!} \right); \quad (2.24) \\ &= (\delta x)^2 (f_x^{(1)})^2 + (\delta x)^4 \left(f_x^{(1)} f_x^{(3)} + \frac{1}{2} (f_x^{(2)})^2 \right) \\ &\quad + (\delta x)^6 \left(\frac{1}{4} f_x^{(1)} f_x^{(5)} + \frac{1}{2} f_x^{(2)} f_x^{(4)} + \frac{5}{12} (f_x^{(3)})^2 \right) + o((\delta x)^8); \quad (2.25) \end{aligned}$$

$\overline{f(x)} - f(x)$ is defined as the *bias* which is the effect of the input uncertainty to the result. When $f(x) \neq 0$, $\frac{\delta f(x)}{|f(x)|}$ is defined as the *normalized uncertainty* of $f(x)$. It is an approximation to the precision $P(f(x)) = \frac{\delta f(x)}{|f(x)|}$.

2.9.2 Binding Factor

The choice of the bounding factor σ needs careful considerations. If $\sigma = \infty$, $\zeta(2n) = (2n-1)!!$, which may cause Formula (2.24) to diverge in most cases. When σ is limited, Formula (2.22) becomes Formula (2.26):

- For $2n < 10$ when $5 \leq \sigma$, $\zeta(2n)$ can be approximated by $\zeta(2n) = (2n-1)!!$ according to Formula (2.27).
- For large $2n$, Formula (2.26) reduces to Formula (2.28), which shows that $\zeta(2n)$ increases slower than σ^{2n} for increasing $2n$.
- For $20 < 2n \leq 200$, when $3 \leq \sigma \leq 6$, $\zeta(2n)$ is well fitted by λ^{2n} numerically, in which λ is a fitting parameter. λ is always slightly smaller than σ empirically, as expected.

$$\zeta(2n) = 2N(\sigma)\sigma^{2n} \sum_{j=1}^{\infty} \sigma^{2j-1} \frac{(2n-1)!!}{(2n-1+2j)!!}; \quad (2.26)$$

$$= (2n-1)\zeta(\sigma, 2n-2) - 2N(\sigma)\sigma^{2n-1}; \quad (2.27)$$

$$\sigma^2 \ll 2n : \quad \zeta(2n) \simeq 2N(\sigma) \frac{\sigma^{2n+1}}{2n+1}; \quad (2.28)$$

The property of $\zeta(2n)$ when $\sigma^2 \ll 2n$ is not sensitive to the underlying uncertainty distribution. If the normal distribution $N(z)$ in Formula (2.22) is replaced by a uniform distribution between $[-\sigma, +\sigma]$, Formula (2.29) is similar to Formula (2.28):

$$\zeta(2n) \equiv \int_{-\sigma}^{+\sigma} \frac{1}{2\sigma} \tilde{z}^{2n} d\tilde{z} = 2 \frac{1}{2\sigma} \frac{\sigma^{2n+1}}{2n+1}; \quad (2.29)$$

The limited range of $\tilde{x} \in (-\sigma\delta x, +\sigma\delta x)$ causes a bounding leakage ϵ according to Formula (2.30), in which $\xi(z)$ is the cumulative density function for normal distribution [4]. When $\sigma = 5$, $\epsilon = 2 \times 10^{-7}$, which is small enough for most applications. $\sigma = 5$ is also a golden standard to assert a negative result [2].

$$\epsilon = 2 - 2\xi\left(\frac{1}{2} + \sigma\right); \quad (2.30)$$

When $\sigma \geq 5$, $\zeta(2n)$ is almost identical to $(2n-1)!!$ when $2n \leq 16$, which makes Formula (2.24) quite insensitive to the actual value of σ .

Thus, $\sigma = 5$ in variance arithmetic.

2.9.3 One-dimensional Examples

Formula (2.33) gives the result variance for e^x , which always converge:

$$e^{x+\tilde{x}} = e^x \sum_{n=0}^{\infty} \frac{\tilde{x}^n}{n!}; \quad (2.31)$$

$$\overline{e^x} = e^x + e^x \sum_{n=1}^{\infty} (\delta x)^{2n} \zeta(2n) \frac{1}{(2n)!}; \quad (2.32)$$

$$\frac{\delta^2 e^x}{(e^x)^2} = \sum_{n=1}^{\infty} (\delta x)^{2n} \left(\sum_{j=1}^{2n-1} \frac{\zeta(2n)}{j! (2n-j)!} - \sum_{j=1}^{n-1} \frac{\zeta(2j)}{(2j)!} \frac{\zeta(2n-2j)}{(2n-2j)!} \right) \quad (2.33)$$

$$= (\delta x)^2 + \frac{3}{2}(\delta x)^4 + \frac{7}{6}(\delta x)^6 + \frac{5}{8}(\delta x)^8 + o((\delta x)^{10}); \quad (2.34)$$

Formula (2.37) gives the result variance for $\ln(x)$:

$$\ln(x + \tilde{x}) - \ln(x) = \ln\left(1 + \frac{\tilde{x}}{x}\right) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \frac{\tilde{x}^j}{x^j}; \quad (2.35)$$

$$\overline{\ln(x)} = \ln(x) - \sum_{n=1}^{+\infty} P(x)^{2n} \frac{\zeta(2n)}{2n}; \quad (2.36)$$

$$\delta^2 \ln(x) = \sum_{n=1}^{+\infty} P(x)^{2n} \left(\sum_{j=1}^{2n-1} \frac{\zeta(2n)}{j(2n-j)} - \sum_{j=1}^{n-1} \frac{\zeta(2j)}{2j} \frac{\zeta(2n-2j)}{2n-2j} \right) \quad (2.37)$$

$$= P(x)^2 + P(x)^4 \frac{9}{8} + P(x)^6 \frac{119}{24} + P(x)^8 \frac{991}{32} + o(P(x)^{10}); \quad (2.38)$$

Formula (2.41) gives the result variance for $\sin(x)$, which converge when $\delta x < 1$:

$$\sin(x + \tilde{x}) = \sum_{n=0}^{\infty} \sin(x)(-1)^n \frac{\tilde{x}^{2n}}{(2n)!} + \sum_{n=0}^{\infty} \cos(x)(-1)^n \frac{\tilde{x}^{2n+1}}{(2n+1)!}; \quad (2.39)$$

$$\overline{\sin(x)} = \sin(x) + \sin(x) \sum_{n=0}^{\infty} (\delta x)^{2n} (-1)^n \frac{\zeta(2n)}{(2n)!}; \quad (2.40)$$

$$\begin{aligned} \delta^2 \sin(x) &= \sum_{n=1}^{\infty} (\delta x)^{2n} (-1)^{n-1} \\ &\quad \left(\cos(x)^2 \sum_{j=0}^{n-1} \frac{\zeta(2n)}{(2j+1)!(2n-2j-1)!} - \sin(x)^2 \sum_{j=1}^{n-1} \frac{\zeta(2n) - \zeta(2j)\zeta(2n-2j)}{(2j)!(2n-2j)!} \right) \end{aligned} \quad (2.41)$$

$$= (\delta x)^2 \cos(x)^2 - (\delta x)^4 \left(\cos(x)^2 \frac{3}{2} - \frac{1}{2} \right) + (\delta x)^6 \left(\cos(x)^2 \frac{7}{6} - \frac{1}{2} \right) + o((\delta x)^8); \quad (2.42)$$

Formula (2.45) gives the result variance for x^c , in which $\binom{c}{n} = \frac{\prod_{j=0}^{n-1} (c-j)}{n!}$:

$$(x + \tilde{x})^c = x^c \left(1 + \frac{\tilde{x}}{x} \right)^c = x^c + x^c \sum_{n=1}^{\infty} \frac{\tilde{x}^n}{x^n} \binom{c}{n}; \quad (2.43)$$

$$\overline{x^c} = x^c + x^c \sum_{n=1}^{\infty} P(x)^{2n} \zeta(2n) \binom{c}{2n}; \quad (2.44)$$

$$\begin{aligned} \frac{\delta^2 x^c}{(x^c)^2} &= \sum_{n=1}^{\infty} P(x)^{2n} \left(\sum_{j=1}^{2n-1} \zeta(2n) \binom{c}{j} \binom{c}{2n-j} - \sum_{j=1}^{n-1} \zeta(2j) \binom{c}{2j} \zeta(2n-2j) \binom{c}{2n-2j} \right) \end{aligned} \quad (2.45)$$

$$= c^2 P(x)^2 + \frac{3}{2} c^2 (c-1) \left(c - \frac{5}{3} \right) P(x)^4 + \frac{7}{6} c^2 (c-1) (c-2)^2 \left(c - \frac{16}{7} \right) P(x)^6 + o(P(x)^8); \quad (2.46)$$

As the special cases for Formula (2.45), Formula (2.47) gives the result variance for x^2 , Formula (2.48) gives the result variance for \sqrt{x} , Formula (2.49) gives the result variance for $1/x$, and Formula (2.50) gives the result variance for $1/x^2$:

$$P(x^2)^2 = 4P(x)^2 + 2P(x)^4; \quad (2.47)$$

$$P(\sqrt{x})^2 = \frac{1}{4} P(x)^2 + \frac{7}{32} P(x)^4 + \frac{75}{128} P(x)^6 + o(P(x)^8); \quad (2.48)$$

$$P(1/x)^2 = P(x)^2 + 8P(x)^4 + 69P(x)^6 + o(P(x)^8); \quad (2.49)$$

$$P(1/x^2)^2 = 4P(x)^2 + 66P(x)^4 + 960P(x)^6 + o(P(x)^8); \quad (2.50)$$

2.9.4 Convergence and Termination of Taylor Expansion

If Formula (2.24) can be expressed as $\sum_{n=1}^{\infty} v(2n) \zeta(2n) P(x)^{2n}$, such as Formula (2.37) and (2.45), and if $|v(2n)| \sim \lambda^{2n}$, then Formula (2.24) converges if $P(x) < 1/(\lambda\sigma)$. λ depends on $f(x)$, e.g., it is larger for $1/x$ than for \sqrt{x} . The maximal $P(x)$ for Formula (2.24) in the $P(x)$ -form to converge is defined as the *applicable precision threshold*, which can be estimated as $1/\sigma$ according to Formula (2.28).

Formula (2.24) breaks down near a zero or pole, when the underlying uncertainty distribution deviates significantly from Gaussian, which is the case for x^c at $x = 0$, as shown in Figure 4, 5, 6, and 7. Thus, for $(x + a)^c$ in which a is another constant, $P(x) = \delta x/|x - a|$ in Formula (2.45).

The variance formula using Taylor expansion shows the nature of the calculation, such as $\delta x \rightarrow P(e^x)$, $\delta x \rightarrow \delta \sin(x)$, $P(x) \rightarrow P(x^c)$, and $P(x) \rightarrow \delta \ln(x)$, and the difference speed of variance increases of x^c depending on different c .

Both Formula (2.23) and (2.24) are subject to numerical errors and rounding errors such as from $f_x^{(j)}$, $\zeta(2j)$ and $(\delta x)^{2j}$, so both needs to be calculated using the variance arithmetic.

- The variance from Formula (2.23) can be added to the corresponding result of Formula (2.24).
- An uncertainty only needs to be correct on order of magnitude, such as with a precision of $1/\sigma = 0.2$ or finer. If the precision of the uncertainty becomes coarser than $1/\sigma$, the result should be invalidated.
- If within the range of $(x - \sigma\delta x, x + \sigma\delta x)$ an analytic function is monotonic, each term in either Formula (2.23) or (2.24) should be smaller than its corresponding previous term. Otherwise, the corresponding sum will diverge, and the result should be invalidated. On the other hand, if the new terms for both formula start to be smaller than the corresponding least significant values, the sums can be stopped.

2.9.5 Multiple Dimensional Expansion

Due to the uncorrelated uncertainty assumption, the Taylor expansion can be applied to multiple inputs, such as Formula (2.53) [38] .

$$f(x + \tilde{x}, y + \tilde{y}) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{f_{(x,y)}^{(m,n)}}{m!n!} \tilde{x}^m \tilde{y}^n; \quad (2.51)$$

$$\begin{aligned} \overline{f(x,y)} &= \int \int f(x + \tilde{x}, y + \tilde{y}) \rho(\tilde{x}, x, \delta x) \rho(\tilde{y}, y, \delta y) d\tilde{x} d\tilde{y} \\ &= \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} (\delta x)^{2m} (\delta y)^{2n} \frac{\zeta(2m)\zeta(2n)f_{(x,y)}^{(2m,2n)}}{(2m)!(2n)!}; \end{aligned} \quad (2.52)$$

$$\begin{aligned} \delta^2 f(x,y) &= \overline{f(x,y)^2} - \overline{f(x,y)}^2 = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} (\delta x)^{2m} (\delta y)^{2n} \\ &\quad \left(\zeta(2m)\zeta(2n) \sum_{i=0}^{2m} \sum_{j=0}^{2n} \frac{f_{(x,y)}^{(i,j)}}{i!j!} \frac{f_{(x,y)}^{(2m-i,2n-j)}}{(2m-i)!(2n-j)!} \right. \\ &\quad \left. - \sum_{i=0}^m \sum_{j=0}^n \frac{\zeta(2i)\zeta(2j)f_{(x,y)}^{(2i,2j)}}{(2i)!(2j)!} \frac{\zeta(2m-2i)\zeta(2n-2j)f_{(x,y)}^{(2m-2i,2n-2j)}}{(2m-2i)!(2n-2j)!} \right) \end{aligned} \quad (2.53)$$

$$\begin{aligned} &\simeq (\delta x)^2 (f_{(x,y)}^{(1,0)})^2 + (\delta y)^2 (f_{(x,y)}^{(0,1)})^2 \\ &\quad + (\delta x)^4 f_{(x,y)}^{(1,0)} \left(\frac{1}{2} f_{(x,y)}^{(2,0)} + f_{(x,y)}^{(3,0)} \right) + (\delta y)^4 \left(\frac{1}{2} f_{(x,y)}^{(0,2)} + f_{(x,y)}^{(0,1)} f_{(x,y)}^{(0,3)} \right) \\ &\quad + (\delta x)^2 (\delta y)^2 ((f_{(x,y)}^{(1,1)})^2 + f_{(x,y)}^{(0,1)} f_{(x,y)}^{(2,1)} + f_{(x,y)}^{(1,0)} f_{(x,y)}^{(1,2)}); \end{aligned} \quad (2.54)$$

Formula (2.7) and (2.11) are special cases of Formula (2.53). Although Formula (2.53) is only for 2-dimensional, it can be extended to any dimensional.

2.10 Dependency Tracing

$$\delta^2(f + g) = \delta^2 f + \delta^2 g + 2(\overline{fg} - \overline{f}\overline{g}); \quad (2.55)$$

$$\delta^2(fg) = \overline{f^2 g^2} - \overline{f} \overline{g}^2; \quad (2.56)$$

$$\delta^2 f(g) = \overline{f(g)^2} - \overline{f(g)}^2; \quad (2.57)$$

$$\delta^2(c_1 f + c_0) = c_1^2 \delta^2 f; \quad (2.58)$$

When the inputs obeys the uncorrelated uncertainty assumption, variance arithmetic uses statistics to trace dependency between uncertainties. For example:

- Formula (2.55) shows $\delta^2(f + g)$, whose dependence tracing can be demonstrated by $\delta^2(f - f) = 0$ and $\delta^2(f(x) + g(y)) = \delta^2 f + \delta^2 g$, with the latter case corresponding to Formula (2.7). Formula (2.59) shows that Formula (2.24) uses Formula (2.55) between any two terms in the Taylor expansion of Formula (2.21).

$$\begin{aligned} \delta^2 \left(\frac{f_x^{(m)} \tilde{x}^m}{m!} + \frac{f_x^{(n)} \tilde{x}^n}{n!} \right) &= (\delta x)^{2m} \left(\frac{f_x^{(m)}}{m!} \right)^2 \eta(2m) + (\delta x)^{2n} \left(\frac{f_x^{(n)}}{n!} \right)^2 \eta(2n) + \\ &2(\delta x)^{m+n} \left(\frac{f_x^{(m)} f_x^{(n)}}{m! n!} \eta(m+n) - \frac{f_x^{(m)}}{m!} \eta(m) \frac{f_x^{(n)}}{n!} \eta(n) \right); \end{aligned} \quad (2.59)$$

- Formula (2.56) shows $\delta^2(fg)$, whose dependence tracing can be demonstrated by $\delta^2(f/f) = 0$ and $\delta^2(f(x)g(y)) = \overline{f^2}(\delta^2 g) + (\delta^2 f)\overline{g^2} + (\delta^2 f)(\delta^2 g)$, with the latter case corresponding to Formula (2.11).
- Formula (2.57) shows $\delta^2 f(g(x))$, whose dependence tracing can be demonstrated by $\delta^2(f^{-1}(f)) = \delta^2 x$.
- Formula (2.58) shows the variance of linear transformation of a function, which can be applied to Formula (2.55) and (2.56) for more generic dependency tracing.

Because of dependency tracing using Taylor expansion, variance arithmetic does not suffer from the dependency problem for analytic expressions [21] which has plagued interval arithmetic.

2.11 Progressive Execution and Dependency Problem

Dependency tracing requires final analytic solution to apply Taylor expansion for the result mean and variance, such as using Formula (2.24) and (2.53). It conflicts with traditional numerical algorithms which execute progressively. For example:

- Traditionally, $x^2 - x$ can be calculated equivalently as Formula (1.7), (1.8), and (1.9). When applying Formula (2.7), (2.11), and (2.47) as separated and independent arithmetic operations, only Formula (1.7) gives the correct result, while the other two violate dependency tracing and give wrong results, as shown in Table 1. It shows that the dependency problem of interval arithmetic [21] may be due to the lack of dependency tracing.

Formula	Result difference	Wrong independence
(1.7)	0	
(1.8)	$4x(\delta x)^2$	between x^2 and x
(1.9)	$(-2x^2 + 2x)(\delta x)^2 - (\delta x)^4$	between $x - 1$ and x

Table 1: The dependency problem when applying variance arithmetic without dependency tracing for $x^2 - x$ when the input variance is $(\delta x)^2$. The correct result variance is $(2x - 1)^2(\delta x)^2 + 2(\delta x)^4$.

- In Gaussian elimination for linear equation [12], the execution is carried out progressively in the path to minimize rounding errors using conventional floating-point calculations, without any consideration for dependency tracing. The existence of different execution paths suggests that the solution itself is designed with dependency problem. Instead, in Section 4, variance arithmetic uses the matrix inversion formula by the determinants of sub-matrices, which is advised against by traditional numerical algorithm [12].
- Formula (2.60) is the single-variable version of Formula (2.57). If instead, $\delta^2 f(g(x))$ is calculated as $\delta^2 f(x)$ with $\delta^2 x = \delta^2 g(x)$, the result Formula (2.61) depends on the sequence of the composite functions, such as $P(\sqrt{x^2})^2 \simeq P(x)^2 + \frac{9}{2}P(x)^4$ v.s. $P(\sqrt{x^2})^2 \simeq P(x)^2 + \frac{9}{8}P(x)^4$.

Thus, when the dependency tracing is not applied strictly, variance arithmetic will have dependency problem. To avoid such dependency problem, all the conventional numerical algorithms need to be reevaluated for variance arithmetic.

$$\begin{aligned} \delta^2 f(g(x)) &= (\delta x)^2 (f_x^{(1)} g_x^{(1)})^2 + (\delta x)^4 ((f_x^{(1)})^2 (g_x^{(1)} g_x^{(3)} + \frac{1}{2}(g_x^{(2)})^2) + \frac{1}{2}(f_x^{(2)})^2 (g_x^{(1)})^4 \\ &\quad + f_x^{(1)} f_x^{(2)} (g_x^{(1)})^4 + 4f_x^{(1)} f_x^{(2)} (g_x^{(1)})^2 g_x^{(2)}) + o((\delta x)^6); \end{aligned} \quad (2.60)$$

$$\begin{aligned} \delta^2 f(x)|_{\delta^2 x = \delta^2 g(x)} &= (\delta x)^2 (f_x^{(1)} g_x^{(1)})^2 + (\delta x)^4 ((f_x^{(1)})^2 (g_x^{(1)} g_x^{(3)} + \frac{1}{2}(g_x^{(2)})^2) + \frac{1}{2}(f_x^{(2)})^2 (g_x^{(1)})^4 \\ &\quad + f_x^{(1)} f_x^{(3)} (g_x^{(1)})^4) + o((\delta x)^6); \end{aligned} \quad (2.61)$$

3 Verification of the Variance Arithmetic

3.1 Verification Methods and Standards

Analytic functions each with precisely known results are used to validate the result uncertainties from variance arithmetic using the following statistical properties:

- *value error*: as the difference between the result using conventional floating-point calculation and the corresponding known analytic result.
- *normalized error*: as the ratio of a value error to the corresponding uncertainty.
- *error deviation*: as the standard deviation of the normalized errors.
- *error distribution*: as the histogram of the normalized errors.
- *uncertainty mean*: as the mean deviation of the result uncertainties.
- *uncertainty deviation*: as the deviation of the result uncertainties. When the uncertainty deviation is much smaller than the corresponding uncertainty mean, the result uncertainties are almost a constant.
- *error response*: as the relation between the input uncertainties and the result uncertainties.
- *calculation response*: as the relation between the amount of calculation and the result uncertainties.

One goal of calculation involving uncertainty is to precisely account for all source of input errors, to achieve *ideal coverage*. In the case of ideal coverage:

1. The error deviation is exactly 1.
2. The central limit theorem converges the result error distribution toward normal distribution.
3. The error response fits the function, such as a linear error response is expected for a linear function.
4. The calculation response fits the function, such as more calculations generally results in larger result uncertainties.

If instead, the input uncertainty is only correct for the input errors on the order of magnitude, the *proper coverage* is achieved. In case of proper coverage, the error deviation should be close to 1: to be in the range of $[0.1, 10]$.

When the input contains uncertainties whose deviation is not precisely known, such as the numerical errors of library functions using conventional floating-point calculations, Gaussian or white noises of increasing deviations can be added to the input, until ideal coverage is reached. The needed minimal noise deviation gives a good estimation of the amount of the unspecified input uncertainties. The existence of ideal coverage is a necessary verification if the variance arithmetic correctly tracks the result uncertainty. The input noise range for the ideal coverage defines the ideal application range of input uncertainties.

Gaussian or white noises of specified deviations can also be added to the input to measure the error responses of a function.

3.2 Types of Uncertainties [38]

There are four sources of result uncertainty for a calculation [1][12]:

- input uncertainties
- rounding errors
- truncation errors
- modelling errors

As described previously, both *input uncertainties* and *rounding errors* are included in the uncertainty specification of variance arithmetic.

In many cases, because a numerical algorithm approaches its analytic counterpart only after infinite execution, a good numerical algorithm should have an estimator of the *truncation error* toward its analytic counterpart, such as the Cauchy remainder estimator for Taylor expansion [12], or the residual error for numerical integration [12]. Using conventional floating-point arithmetic, a subjective upper limit is chosen for the truncation error, to stop the numerical algorithm at limited execution [12]. However, such arbitrary upper limit may not be achievable with the amount of rounding errors accumulated during calculation, so that such upper limit may actually give a falsely small result precision. Because variance arithmetic tracks rounding errors of a calculation efficiently, it can be used to search for the optimal execution termination point for the numerical algorithm when the truncation error is no longer significant, which is named as the *truncation rule* in this paper. In other words, using variance arithmetic, the result precision of a calculation is determined by the inputs and the calculation itself. Section ?? and ?? will provide such cases of applying truncation rule to Taylor expansion and numerical integration, respectively.

The *modelling errors* arise when an approximate analytic solution is used, or when a real problem is simplified to obtain the solution. For example, Section 8 demonstrates that the discrete Fourier transformation is only an approximation for the mathematically defined Fourier transformation. Conceptually, modelling errors originate from mathematics, so they are outside the domain for variance arithmetic.

3.3 Types of Calculations to Verify [38]

Algorithms of completely different nature with each representative for its category are needed to test the generic applicability of uncertainty-bearing arithmetic. An algorithm can be categorized by comparing the amount of its input and output data:

- transformation
- generation
- reduction

A *transformation* algorithm has about equal amounts of input and output data. The information contained in the data remains about the same after transforming. The Discrete Fourier Transformation is a typical transforming algorithm, which contains exactly the same amount of input and output data, and its output data can be transformed back to the input data using essentially the same algorithm. Matrix inversion is another such reversible algorithm. For reversible transformations, a unique requirement for uncertainty-bearing arithmetic is to introduce the least amount of additional uncertainty after a *round-trip* transformation which is a *forward* followed

by a *reverse* transformation. The errors after the roundtrip transformation are 0 ideally ⁶, so both the error deviation and the error maximums should be 0. A test of uncertainty-bearing arithmetic using FFT algorithms is provided in Section 8, and a test of matrix inversion is provided in Section 4.

A *generation* algorithm has much more output data than input data. Solving differential equations numerically and generating a numerical table of a specific function are two typical generating algorithms. The generating algorithm codes mathematical knowledge into data, so there is an increase of information in the output data. From the perspective of encoding information into data, Taylor expansion is also a generating algorithm. In generating algorithms, input uncertainty should also be considered when deciding if the result is good enough so that the calculation can stop. Some generating algorithms are theoretical calculations which involve no imprecise input so that all result uncertainty is due to rounding errors. Section 7 demonstrates such an algorithm, which calculates a table of the sine function using trigonometric relations and two precise input data, $\sin(0) = 0$ and $\sin(\pi/2) = 1$. The generated data have different precision.

A *reduction* algorithm has much less output data than input data such as numerical integration and statistical characterization of a data set. Some information of the data is lost while other information is extracted during reducing. Conventional wisdom is that a reducing algorithm generally benefits from a larger input data set [4]. Such a notion needs to be re-evaluated when uncertainty accumulates during calculation. A test of uncertainty-bearing arithmetic using numerical integration is provided in Section ??.

⁶When noises are added to the original input signal, the value errors should be the results minus the overall input data. The previous approaches [38] used the original input data wrongly.

4 Matrix Inversion

4.1 Uncertainty Propagation in Matrix Determinant

Let vector $[p_1, p_2 \dots p_n]_n$ denote a permutation of the vector $(1, 2 \dots n)$ [45]. Let $\$[p_1, p_2 \dots p_n]_n$ denote the permutation sign of $[p_1, p_2 \dots p_n]_n$ [45]. For a n -by- n square matrix \mathbf{M} with the element $x_{i,j}$, $i, j = 1, 2 \dots n$, let its determinant be defined as Formula (4.1) [45] and let the sub-determinant at index (i, j) be defined as Formula (4.2) [45]. $\mathbf{M}_{i,j}$ is the matrix that results from deleting the row i and column j of \mathbf{M} [45]. Formula (4.3) holds for the arbitrary row index i or the arbitrary column index j [45].

$$|\mathbf{M}| \equiv \sum_{[p_1 \dots p_n]_n} \$[p_1 \dots p_n]_n \prod_{k=1 \dots n} x_{k,p_k}; \quad (4.1)$$

$$|\mathbf{M}|_{i,j} \equiv \sum_{\substack{p_i=j \\ [p_1 \dots p_n]_n}} \$[p_1 \dots p_n]_n \prod_{\substack{k \neq i \\ k=1 \dots n}} x_{k,p_k}; \quad (4.2)$$

$$|\mathbf{M}| = \sum_{j=1 \dots n} |\mathbf{M}_{i,j}| x_{i,j} = \sum_{i=1 \dots n} |\mathbf{M}_{i,j}| x_{i,j}; \quad (4.3)$$

Assuming $p_1, p_2 \in \{1 \dots n\}$, let $[p_1, p_2]_n$ denote the length-2 unordered permutation which satisfies $p_1 \neq p_2$, and let $\langle p_1, p_2 \rangle_n$ denote the length-2 ordered permutation which satisfies $p_1 < p_2$, and let $\langle i_1, i_2 \rangle_n$ be an arbitrary ordered permutation⁷. Formula (4.3) can be applied to $\mathbf{M}_{i,j}$, as:

$$|\mathbf{M}_{\langle i_1, i_2 \rangle_n, [j_1, j_2]_n}| \equiv \sum_{\substack{p_{i_1}=j_1, p_{i_2}=j_2 \\ [p_1 \dots p_n]_n}} \$[p_1 \dots p_n]_n \prod_{\substack{k \notin \{i_1, i_2\} \\ k=1 \dots n}} x_{k,p_k}; \quad (4.4)$$

$$|\mathbf{M}| = \sum_{j_1=1 \dots n} x_{i_1, j_1} |\mathbf{M}_{i_1, j_1}| = \sum_{j_1=1 \dots n} \sum_{\substack{j_2 \neq j_1 \\ j_2=1 \dots n}} |\mathbf{M}_{\langle i_1, i_2 \rangle_n, [j_1, j_2]_n}| x_{i_1, j_1} x_{i_2, j_2}; \quad (4.5)$$

The definition of a sub-determinant can be extended to Formula (4.6), in which $m = 1 \dots n$. Formula (4.5) can be generalized as Formula (4.7), in which $\langle i_1 \dots i_m \rangle_n$ is an arbitrary ordered permutation. The $(n-m)$ -by- $(n-m)$ matrix in $|\mathbf{M}_{\langle i_1 \dots i_m \rangle_n, [j_1 \dots j_m]_n}|$ is obtained by deleting the rows in $\{i_1 \dots i_m\}$ and the columns in $\{j_1 \dots j_m\}$. This leads to Formula (4.8).

$$|\mathbf{M}_{\langle i_1 \dots i_m \rangle_n, [j_1 \dots j_m]_n}| \equiv \sum_{[p_1 \dots p_n]_n} \$[p_1 \dots p_n]_n \prod_{\substack{k \notin \{i_1 \dots i_m\} \\ k=1 \dots n}} x_{k,p_k}; \quad (4.6)$$

$$|\mathbf{M}| = \sum_{[j_1 \dots j_m]_n} |\mathbf{M}_{\langle i_1 \dots i_m \rangle_n, [j_1 \dots j_m]_n}| \prod_{k=1}^m x_{i_k, j_k}; \quad (4.7)$$

$$||\mathbf{M}_{\langle i_1 \dots i_m \rangle_n, [j_1 \dots j_m]_n}|| = ||\mathbf{M}_{\langle i_1 \dots i_m \rangle_n, \langle j_1 \dots j_m \rangle_n}||; \quad (4.8)$$

If all the element of the matrix satisfy the uncorrelated uncertainty assumption, after applying Formula (2.52) and (2.53), Formula (4.10) and (4.11) are the result

⁷An ordered permutation is a sorted combination.

mean and variance, respectively.⁸

$$|\overline{\mathbf{M}}| = |\mathbf{M}|; \quad (4.10)$$

$$\delta^2 |\mathbf{M}| = \sum_{m=1}^n \sum_{\langle i_1 \dots i_m \rangle_n} \sum_{[j_1 \dots j_m]_n} |\mathbf{M}_{\langle i_1 \dots i_m \rangle_n, \langle j_1 \dots j_m \rangle_n}|^2 \prod_{k=1 \dots n}^{i_k \in \{i_1 \dots i_m\}} (\delta x_{i_k, j_k})^2; \quad (4.11)$$

Formula (4.10) and (4.11) assume that the uncertainties of matrix elements are independent to each other. In contrast, for the special matrix in Formula (4.12), after applying Formula (2.52) and (2.53), Formula (4.13) and (4.14) give the result mean and variance, respectively.

$$|\mathbf{M}| = \begin{vmatrix} x, z, y \\ z, y, x \\ y, x, z \end{vmatrix} = 3xyz - x^3 - y^3 - z^3; \quad (4.12)$$

$$|\overline{\mathbf{M}}| = -3x(\delta x)^2 - 3y(\delta y)^2 - 3z(\delta z)^2; \quad (4.13)$$

$$\begin{aligned} \delta^2 |\mathbf{M}| = & 3(5x^4 - 12x^2yz + 2xy^3 + 2xz^3 + 3y^2z^2)(\delta x)^2 \\ & + 3(5y^4 - 12y^2xz + 2yx^3 + 2yz^3 + 3x^2z^2)(\delta y)^2 \\ & + 3(5z^4 - 12z^2xy + 2zx^3 + 2zy^3 + 3x^2y^2)(\delta z)^2 \\ & + 9(z^2 + 2xy)(\delta x)^2(\delta y)^2 + 9(y^2 + xz)(\delta x)^2(\delta z)^2 + 9(x^2 + 2yx)(\delta y)^2(\delta z)^2 \\ & + 9(5x^2 - 2yz)(\delta x)^4 + 9(5y^2 - 2xz)(\delta y)^4 + 9(5z^2 - 2xy)(\delta z)^4 \\ & + 15(\delta x)^6 + 15(\delta y)^6 + 15(\delta z)^6; \end{aligned} \quad (4.14)$$

4.2 Adjugate Matrix

The square matrix whose element is $(-1)^{i+j}|\mathbf{M}_{j,i}|$ is defined as the *adjugate matrix* [45] \mathbf{M}^A to the original square matrix \mathbf{M} . Let \mathbf{I} be the identical matrix for \mathbf{M} . Formula (4.15) and (4.16) show the relation of \mathbf{M}^A and \mathbf{M} , which only involve addition, subtraction and multiplication.

$$\mathbf{M} \times \mathbf{M}^A = \mathbf{M}^A \times \mathbf{M} = |\mathbf{M}|\mathbf{I}; \quad (4.15)$$

$$|\mathbf{M}|\mathbf{M}^A = |\mathbf{M}^A|\mathbf{M}; \quad (4.16)$$

To test Formula (4.11):

1. A matrix \mathbf{M} is constructed using random integers uniformly distributed in the range of $[-2^8, +2^8]$. The integer arithmetic grants that $|\mathbf{M}|$, \mathbf{M}^A , and $|\mathbf{M}^A|$ are precise.
2. Gaussian noises of predefined levels are added to \mathbf{M} , to construct $\widehat{\mathbf{M}}$ which contains imprecise values. The variance arithmetic is used to calculate $|\widehat{\mathbf{M}}|$, $\widehat{\mathbf{M}}^A$, and $|\widehat{\mathbf{M}}^A|$.

⁸The previous paper [38] also contains the wrong formula:

$$\delta^2 |\mathbf{M}| = \sum_{i=1}^n \sum_{j=1}^n (\delta x_{i,j})^2 (|\mathbf{M}_{i,j}|^2 + \delta^2 |\mathbf{M}_{i,j}|); \quad (4.9)$$

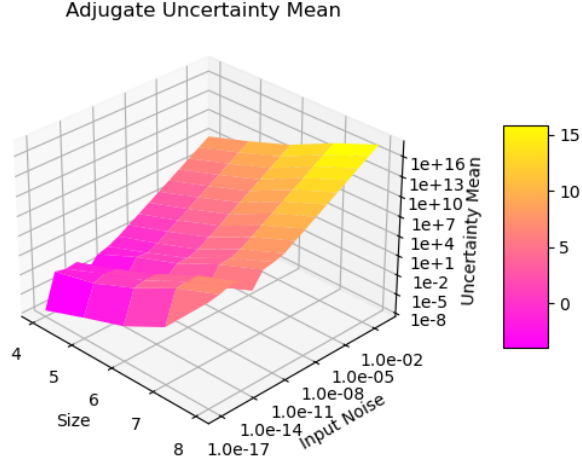


Figure 8: The result uncertainty means vs. input uncertainties and matrix size for the difference between $\widehat{\mathbf{M}}^A$ and \mathbf{M}^A . The input uncertainties run from 10^{-17} to 10^{-1} , while the matrix size runs from 4 to 8.

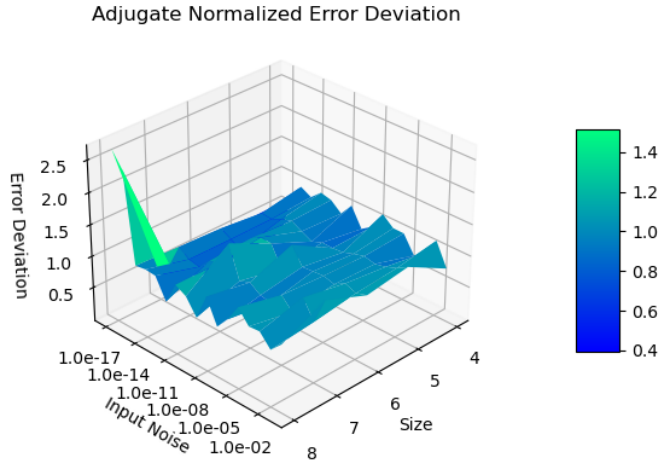


Figure 9: The result error deviations vs. input uncertainties and matrix size for the difference between $\widehat{\mathbf{M}}^A$ and \mathbf{M}^A . The input uncertainties run from 10^{-17} to 10^{-1} , while the matrix size runs from 4 to 8.

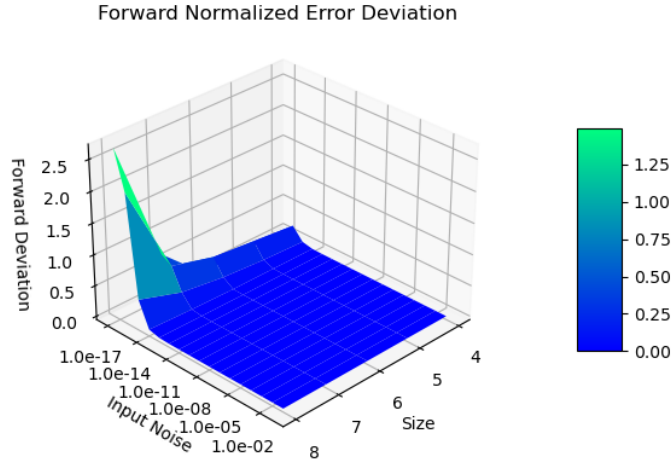


Figure 10: The result error deviations vs. input uncertainties and matrix size for the difference of Formula (4.15). The input uncertainties run from 10^{-17} to 10^{-1} , while the matrix size runs from 4 to 8.

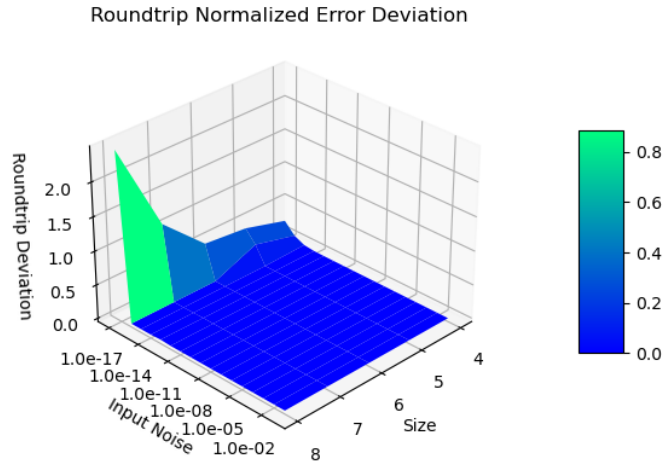


Figure 11: The result error deviations vs. input uncertainties and matrix size for the difference of Formula (??). The input uncertainties run from 10^{-17} to 10^{-1} , while the matrix size runs from 4 to 8.

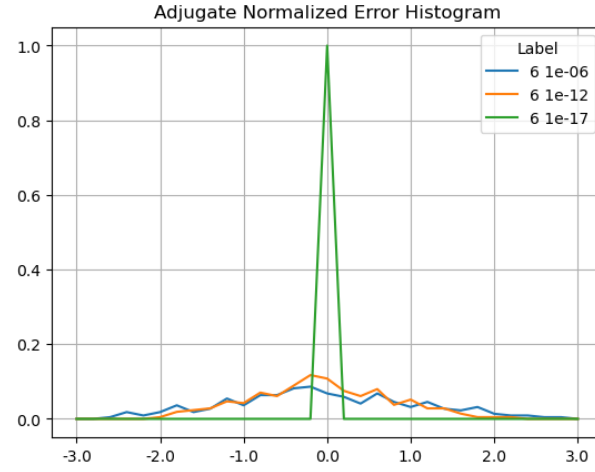


Figure 12: The result histograms for the normalized errors for the same matrix size 6 different input uncertainties as shown in the legend. \mathbf{M} is generated by random integers in the range of $[-2^8, +2^8]$.

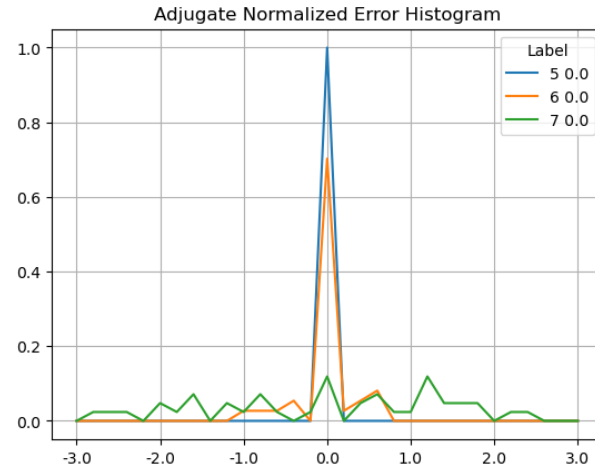


Figure 13: The result histograms for the normalized errors for different matrix size and the input uncertainty 0 as shown in the legend. \mathbf{M} is generated by random integers in the range of $[-2^{11}, +2^{11}]$.

3. The difference between $|\widehat{\mathbf{M}}|$ and \mathbf{M}^A defines the *Adjugate* Test. Figure 8 shows that the result uncertainties increase exponentially with both the input noises and the matrix size, while Figure 9 shows that the ideal coverage is achieved for the input noises greater than 10^{-16} .
4. Formula (4.15) defines the *Forward* Test. Figure 10 shows that Formula (4.15) is strictly observed for idea coverage because the difference is delta-distributed with zero distribution deviation.
5. Formula (4.16) defines the *Roundtrip* Test. Figure 11 shows that Formula (4.16) is strictly observed for idea coverage because the difference is delta-distributed with zero distribution deviation.

The significant of the conventional floating-point representation [10] has 53-bit resolution, or $\epsilon = 10^{-16}$. Adding noise at or below this level will not change the value of a conventional floating-point value, which is the lower bound for the ideal coverage. Figure 12 shows that the histograms of the normalized errors are normal distributed in ideal coverage. Below the lower bound for idea coverage, floating-point rounding errors dominate the result uncertainties. When \mathbf{M} is generated by random integers in the range of $[-2^8, +2^8]$, for the input uncertainty 10^{-17} :

- In Figure 12, the adjugate matrix calculation involves about $8 \times 5 = 40$ significant bits for a matrix size 6, so there is no rounding errors.
- The histogram for matrix size 7 is very similar to Figure 12, because the adjugate matrix calculation involves about $8 \times 6 = 48$ significant bits.
- The histogram for matrix size 8 is quite different to Figure 12, because the adjugate matrix calculation involves about $8 \times 6 = 56$ significant bits.

In Figure 13, when \mathbf{M} is generated by random integers in the range of $[-2^{11}, +2^{11}]$ without added noises:

- For matrix size 5, $11 \times 4 = 44$ significant bits are involved, so there is no rounding errors.
- For matrix size 6, $11 \times 5 = 55$ significant bits are involved, so there starts to have rounding errors but the overall distribution is still delta-like.
- For matrix size 7, $11 \times 6 = 66$ significant bits are involved, so there are significant rounding errors, and the distribution is no longer delta-like.

As shown in Figure 9, 10, and 11, the variance arithmetic can only provide proper coverage for the conventional floating-point rounding errors.

For both ideal coverages and proper coverages, the result uncertainties provided by the variance arithmetic trace the result value errors closely, even in the huge range of the value errors as suggested by Figure 8. This validates using variance arithmetic to calculate adjugate matrix.

4.3 Matrix Stability

$$\mathbf{M}^{-1} \equiv \mathbf{M}^A / |\mathbf{M}|; \quad (4.17)$$

$$||\mathbf{M}^{-1}|| |\mathbf{M}| = \frac{|\mathbf{M}^A|}{|\mathbf{M}|^{n-1}}; \quad (4.18)$$

When all values in \mathbf{M} are precise:

- The inverse matrix [45] \mathbf{M}^{-1} defined by Formula (4.17) is functionally equivalent to \mathbf{M}^A .
- Formula (4.3) contains different paths, all of which are equivalent.

When the rounding errors of the conventional floating-point arithmetic are no longer ignored, Formula (4.17) is avoided, and the different paths of Formula (4.3) is explored to calculate \mathbf{M}^{-1} in the Gaussian elimination algorithm [12]. When the input uncertainties are no longer ignored, the stability of \mathbf{M}^{-1} is measured by the conditional number [45] as defined in Formula (4.18). In contrast, the variance arithmetic provides direct quantification of uncertainty, so that the effect of $|\mathbf{M}|$ in Formula (4.17) to \mathbf{M}^{-1} can be qualified by the precision of $|\mathbf{M}|$ using Formula (2.49).

However, when $|\mathbf{M}|$ is imprecise, it introduce globally uncertainties to \mathbf{M}^{-1} , and the stability of \mathbf{M}^{-1} is measured by the conditional number [45] as defined in Formula (4.18).

When using Formula (4.3) to calculate determinant in conventional floating-point arithmetic:

- The input uncertainties and the calculation rounding errors can not be accounted for in the results.
- One path is chosen out of many possible paths, such as using different i for summing by j in Formula (4.3).
- The result is path dependent. When calculating the inverse matrix, such dependency is so strong that Formula (4.17) is avoided, and the path which is deemed to have smallest rounding error is chosen in the Gaussian elimination algorithm [12]. Even so, the result uncertainty is not clear.

In contrast, Formula (4.11) is no longer path dependent, and it accounts for all input uncertainties and calculation rounding errors.

A matrix \tilde{M} is constructed from M in two different ways:

- *rounding*: If each element of M is converted to a conventional 64-bit floating-point value to construct \tilde{M} , the result errors are due to the rounding errors of the conventional 64-bit floating-point calculations.
- *noise*: If either Gaussian noise or white noise of a particular level is added to each element of M to construct \tilde{M} , the result errors are mainly due to the input noise unless the input noise level is 10^{-16} or less. For example, if the noise-level is 10^{-3} , because the distribution deviation of the elements of M is $16384/\sqrt{3} = 9459$, the Gaussian noise is generated with the deviation of 9459×10^{-3} .
- For simplicity, a same noise-level is used for every element of M .

The calculation results using \tilde{M} can be compared with those of M in different tests:

- *determinant*: to compare $|\mathbf{M}|$ and $|\tilde{M}|$.
- *adjugate*: to compare M^A and \tilde{M}^A .
- *inverse*: to compare $M^A M = |\mathbf{M}| \mathbf{I}$ and $\tilde{M}^A \tilde{M}$, in which \mathbf{I} is the identity matrix [45] for M .

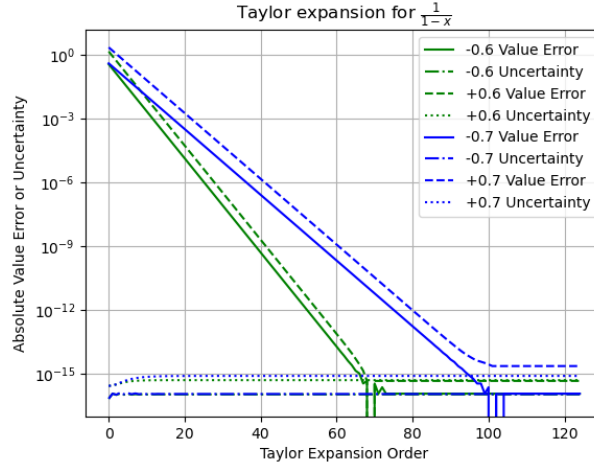


Figure 14: The Taylor expansion for $\frac{1}{1-x} = \lim_{N \rightarrow \infty} \sum_{n=0}^N x^n$, for different Taylor expansion order N as shown in the x-axis, and different x as shown in the legend. The y-axis shows either the uncertainty or the absolute reminder value error.

5 Polynomial and Taylor Expansion

For a polynomial $f(x) \equiv \sum_{j=0}^N c_j x^j$, Formula (5.1) gives the corresponding Taylor expansion:

$$f(x + \tilde{x}) = \sum_{j=0}^N c_j (x + \tilde{x})^j = \sum_{j=0}^N \tilde{x}^j \sum_{k=0}^{N-j} x^{k-j} c_{j+k} \binom{j+k}{j}; \quad (5.1)$$

When $f(x) = (ax + b)^N$, the result of applying Formula (5.1) to Formula (2.24) is Formula (2.45) with $c = N$ and $x = ax + b$. Unlike Formula (2.45), applying Formula (5.1) to Formula (2.24) is no longer limited by the applicable precision threshold, because the exponent for power is limited to positive integers.

A numerical calculation can only calculate limited terms in a Taylor expansion, and the count of terms of the Taylor expansion is needed to approximate a function $f(x)$ is defined as the *Taylor expansion order*, which can be obtained by either of the two methods [12]:

- *reminder estimation*: Theoretically, a Taylor expansion order is reached when the estimated reminder is less than the least significant value of the target result.
- *stability test*: Numerically, a Taylor expansion order is reached when the value change is less than the least significant value the current result.

Formula (2.59) suggests that a Taylor expansion is also a polynomial with $N \rightarrow \infty$ in the variance arithmetic. Using Formula (5.1), the calculated uncertainty increases with the increasing Taylor expansion order. For an evaluation with an known result, at a given Taylor expansion order, the reminder value error can be used instead of the estimated reminders. For example, for $\frac{1}{1-x}$, the reminder value error is calculated as $\sum_{n=0}^N x^n - \frac{1}{1-x}$ in which N is the Taylor expansion order. The calculated

uncertainty should be comparable to the value errors when the Taylor expansion order is reached. Figure 14 shows the reminder and the corresponding uncertainty of $\frac{1}{1-x} = \lim_{N \rightarrow \infty} \sum_{n=0}^N x^n$, in which N is the Taylor expansion order:

- The Taylor expansion order can be obtained by the stability test, such as $N = 73$ when $x = +0.6$. Except at very low Taylor expansion orders, the calculated uncertainties are quite stable, so that they can be used to affirm that the result is stable.
- In all cases, the reminder value errors stabilize at exactly the corresponding least significant values. The variance arithmetic tracks the reminder value error properly. In the worst case, when $x = -0.7$, the normalized error is 2.75.
- When x is positive, the reminder value errors approach the least significant values gradually. When x is negative, the reminder value errors oscillate near the Taylor expansion order. This means difficulty in obtaining an accurate reliable reminder estimator. On the other hand, the statistical tracking by the variance arithmetic using Formula (5.1) is straight-forward.

Using variance arithmetic for Taylor expansion, the result uncertainty tracks the reminder value error properly. The stability test method can be used for numerical calculation of Taylor expansion, as well as the calculation of the mean and the variance of the Taylor expansion. If the stability test can not be applied, the convergence of the mean and the variance can be judged numerically according to Section 6, and the reminder estimation method can be used to terminate the expansion.

6 Math Library Functions

Formula (2.33), (2.37), (2.41), and (2.45) are tested by the corresponding math library functions *exp*, *log*, *sin*, and *pow*, respectively.

At each point x for an input uncertainty δx , the result uncertainty is calculated by the variance arithmetic. The corresponding value deviation is obtained by:

1. Take 10000 samples from either Gaussian noise or uniform distribution with δx as the deviation, and construct \tilde{x} which is x plus the sampled noise.
2. For each \tilde{x} , use the corresponding library function to calculate the value error as the difference between using \tilde{x} and using x as the input.
3. The value error is divided by the result uncertainty, as the normalized error.
4. The standard deviation of the 10000 normalized errors is the error deviation.
5. In addition, for each of these tests, all value errors follow a same underlying distribution. The deviation of the value errors is defined as the *value deviation*, which the uncertainty should match.

6.1 Exponential

Figure 15 shows that the calculated uncertainties using Formula (2.33) agree very well with the measured value deviations for $e^{x+\delta x}$. As a result, the error deviations are very close to 1, even both the uncertainties and the value error deviations increase exponentially with x and δx . Such strong tracking power holds for all x of $e^{x+\delta x}$.

6.2 Logarithm

Figure 16 shows that the calculated uncertainties using Formula (2.37) agree very well with the measured value deviations for $\log(x + \delta x)$, so that the result error deviations are very close to 1, until the uncertainties suddenly diverge when the input precision is coarser than $1/5$ which is the estimated application precision threshold. After the result uncertainties diverge, the corresponding value deviations still continue on the same trends until the input data to the log function start to have negative values when the input noises are uniformly distributed. If Gaussian noises are used instead, the cutoffs for both value errors and error deviations are strictly at $1/5$. The result probability density function for $\log(x)$ has a pole at $x = 0$.

Such divergences of the result uncertainties are expected when the input precision is coarser than the estimated applicable precision threshold of $1/\sigma$, in which $\sigma = 5$ is the bonding factor of the variance arithmetic. If σ is reduced from 5 to 4, the measured applicable precision thresholds increase from $1/5$ to $1/4$, allowing more result uncertainties to be valid. The reduction of σ increases the bounding leakage ϵ from 5.7×10^{-7} to 6.3×10^{-5} according to Formula (2.30). A large leakage ϵ raises the logic questions on the validity of the result, because the leakage may mean $x \leq 0$ for $\log(x)$ in this case. In an extreme approach, in another variance arithmetic representation:

1. Each imprecise value carries its own σ or ϵ to indicate the statistical significance of its value to its uncertainty.
2. The result σ or ϵ involving two imprecise values can be found statistically.
3. The result of an analytic function always converges but at the expense of ϵ , such that when $x \rightarrow 0$, $\epsilon \rightarrow 1$ for $\log(x + \delta x)$.

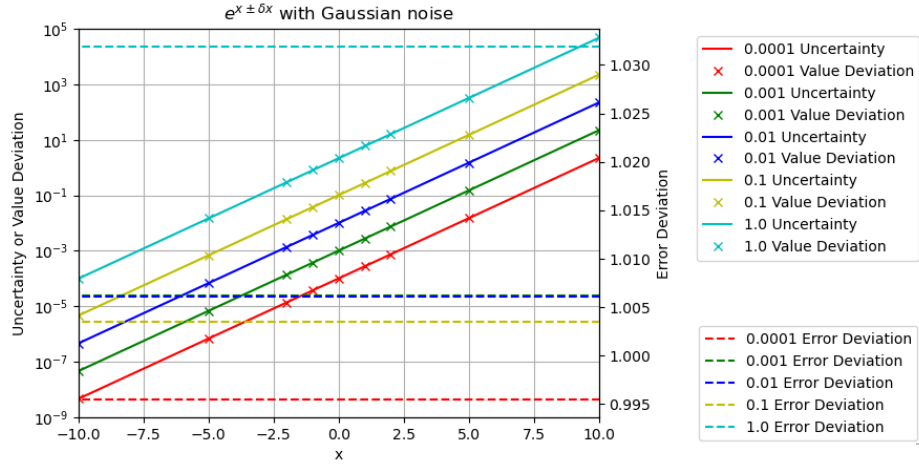


Figure 15: The calculated uncertainties vs the measured value deviations and the measured error deviations for $e^{x \pm \delta x}$, for different δx as shown in the legend. The uncertainties and the value deviations are drawn using the logarithmic y scales on the left side, while the error deviations are drawn using the linear y scales on the right side. Each color represent a δx . Gaussian noises are used to produce the input noise δx .

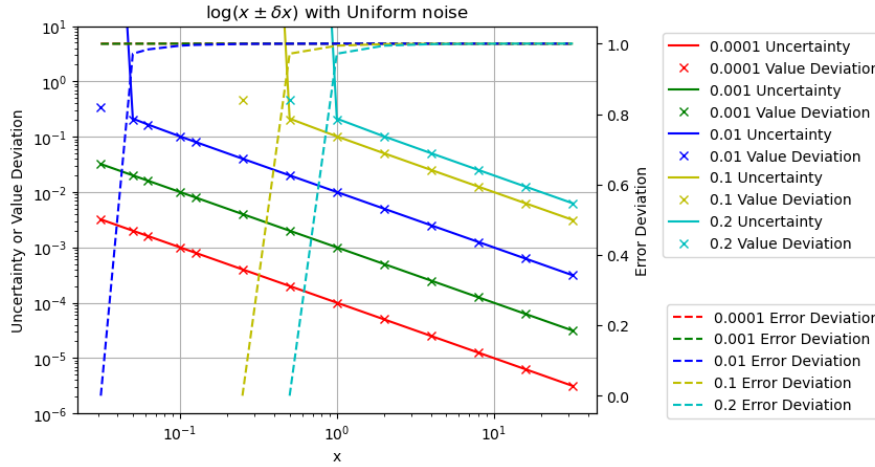


Figure 16: The calculated uncertainties vs the measured value deviations and the measured error deviations for $\log(x \pm \delta x)$, for different δx as shown in the legend. The uncertainties and the value deviations are drawn using the logarithmic y scales on the left side, while the error deviations are drawn using the linear y scales on the right side. Each color represent a δx . Gaussian noises are used to produce the input noise δx .

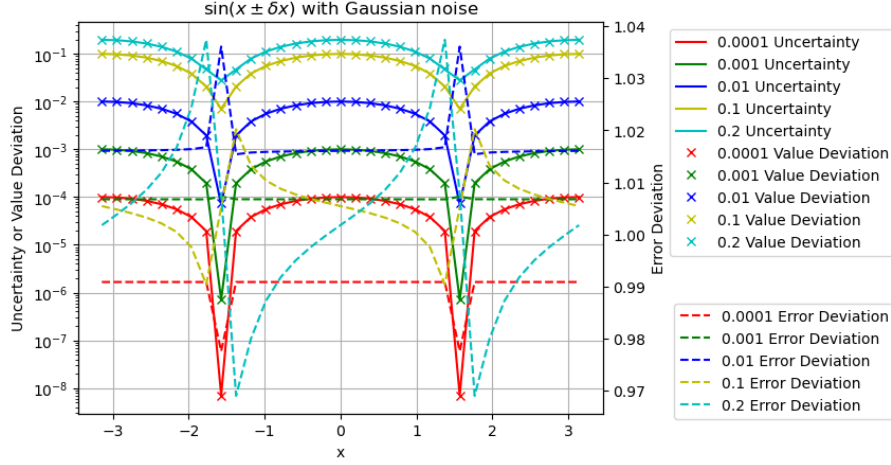


Figure 17: The calculated uncertainties vs the measured value deviations and the measured error deviations for $\sin(x \pm \delta x)$, for different δx as shown in the legend. The uncertainties and the value deviations are drawn using the logarithmic y scales on the left side, while the error deviations are drawn using the linear y scales on the right side. Each color represent a δx . Gaussian noises are used to produce the input noise δx .

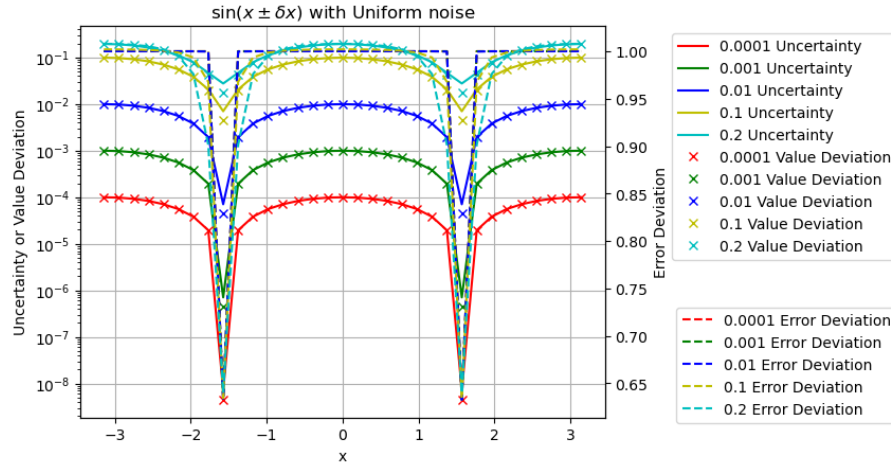


Figure 18: The calculated uncertainties vs the measured value deviations and the measured error deviations for $\sin(x \pm \delta x)$, for different δx as shown in the legend. The uncertainties and the value deviations are drawn using the logarithmic y scales on the left side, while the error deviations are drawn using the linear y scales on the right side. Each color represent a δx . Uniform noises are used to produce the input noise δx .

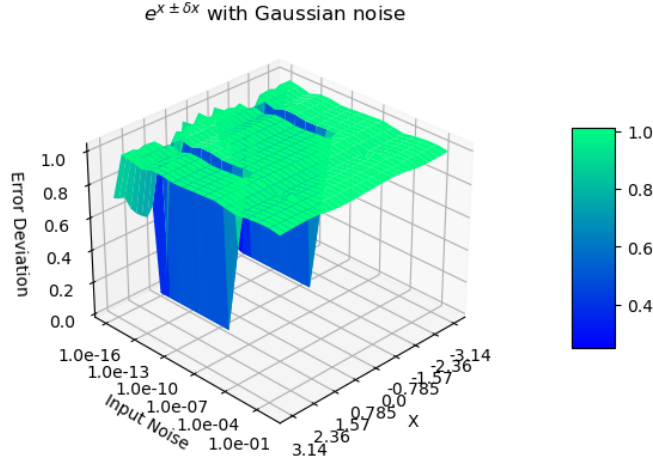


Figure 19: The error deviation for $\sin(x \pm \delta x)$, for x and δ . The x-axis is x between $-\pi$ and $+\pi$. The y-axis is δ between -10^{-16} and 1. The z-axis is the error deviation. Gaussian noises are used to produce the input noise δx .

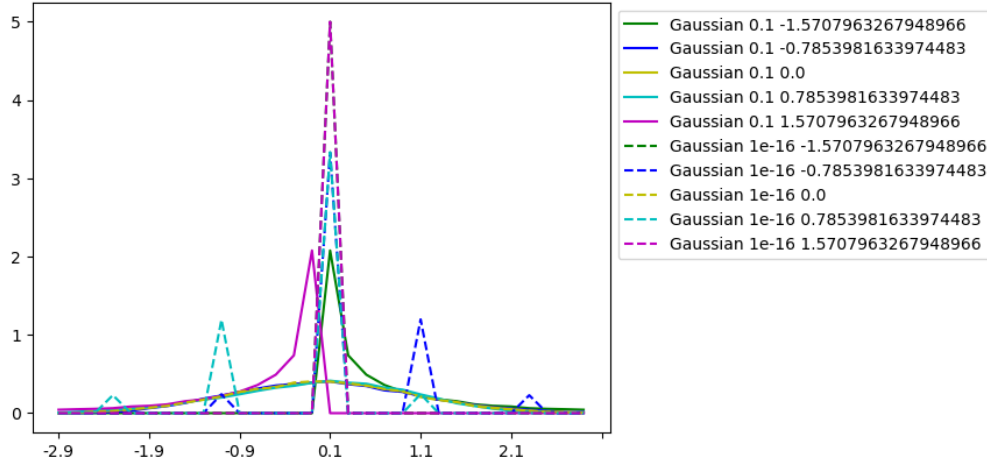


Figure 20: The histogram of the error deviation for $\sin(x \pm \delta x)$, for $x = -\pi/2, -\pi/4, 0, +\pi/4, +\pi/2$ and $\delta x = 10^{-1}, 10^{-16}$ as shown in the legend. The y-axis is the normalized count. Each color represents a different x , while each line pattern represents a different δx . Uniform noises are used to produce the input noise δx .

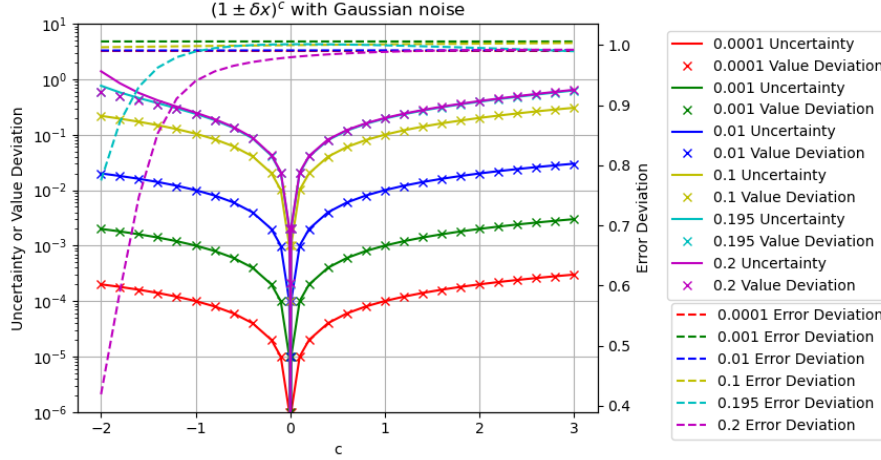


Figure 21: The calculated uncertainties vs the measured value deviations and the measured error deviations for $(1 \pm \delta x)^c$, for different δx as shown in the legend. The uncertainties and the value deviations are drawn using the logarithmic y scales on the left side, while the error deviations are drawn using the linear y scales on the right side. Each color represent a δx .

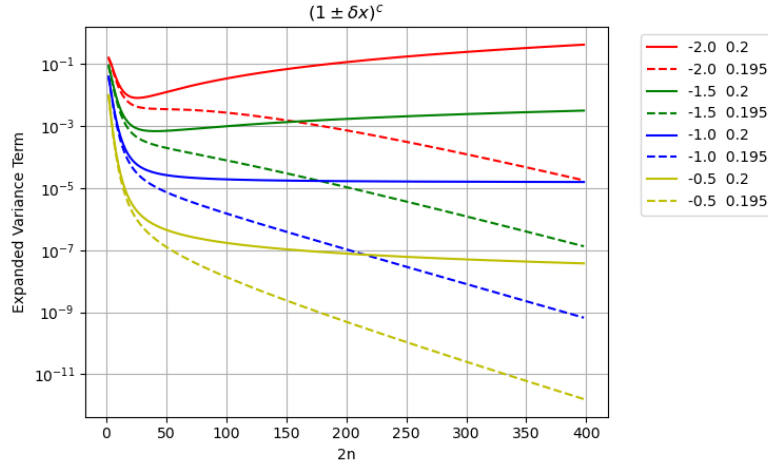


Figure 22: The individual variance term of $(1 \pm \delta x)^c$ vs the different Taylor expansion order $2n$, for different c and δx as shown in the legend. The x -axis is Taylor expansion order $2n$ from 1 to 400 in Formula (2.24). The y -axis is the contribution to $(1 \pm \delta x)^c$ at each Taylor expansion order $2n$ according to Formula (2.45). It shows that $(1 \pm 0.2)^{-2}$ and $(1 \pm 0.2)^{-1.5}$ clearly diverge, $(1 \pm 0.2)^{-1}$ seems diverges, and other cases clearly converge.

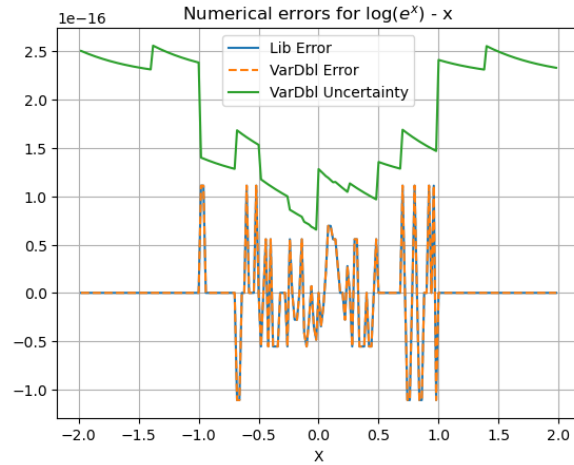


Figure 23: The values and uncertainties of $\log(e^x) - x$ vs x , as *VarDbl Error* and *VarDbl Uncertainty* in the legend. The result of the same calculation using conventional floating-point library functions is shown as *Lib Error* in the legend.

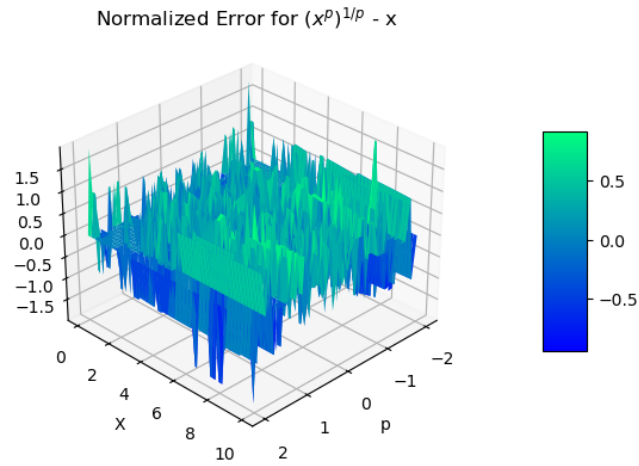


Figure 24: The normalized errors of $(x^p)^{\frac{1}{p}} - x$ vs x and p .

4. A calculation may branch based on the probability ϵ .

For the discussion simplicity, the cases of the bounding factor σ other than constant 5 are avoided in this paper.

6.3 Sine

Figure 17 shows that the calculated uncertainties using Formula (2.41) agree very well with the measured value deviations for $\sin(x + \delta x)$. Figure 17 also shows that $\delta^2 \sin(x)$ has the same periodicity as $\sin(x)$:

- When $x = 0$, $\sin(x) \simeq x$, so that $\delta^2 \sin(x) \simeq (\delta x)^2$.
- When $x = \pi/2$, $\sin(x) \simeq 1$, so that $\delta^2 \sin(x) \simeq 0$.

In Figure 17, Gaussian noises are used, while in Figure 18, ideal white noises are added for then given sample size of 10000. The difference of Figure 17 and 18 shows that the Gaussian noise is more desirable, because of its tail effect, e.g., the error deviation is much closer to 1 for $\sin(\pm\pi/2 + \delta x)$. So only Gaussian noises will be used for other analysis by default.

Figure 19 shows that the error deviation for $\sin(x + \delta x)$ is 1 except when $x = \pm\pi/4$ and $\delta x < 10^{-12}$, at where the probability density function has a zero. Figure 20 shows the histogram of the error deviation for $\sin(x + \delta x)$ when Gaussian noises are used to produce the input noise δx :

- When $\delta x = 10^{-16}$, the noise is comparable to the numerical calculation error of $\sin(x)$. The histogram is highly structured, with peaks near 0, ± 1 , and ± 2 , suggesting that the value errors are integer-folds of the least significant value.
- When $\delta x = 10^{-1}$, the noise is much large than the numerical calculation error of $\sin(x)$, so that ideal coverage is achieved. When $x = \pm\pi/2$, $\sin(x)_x^{(1)} = 0$, so that the result probability density function has a zero. The corresponding histogram is distorted from Gaussian, with the peak at $x = \pm\pi/2$, and all other values on one side of the peak only. Such distortion of the histogram results in 0 error deviations at $x = \pm\pi/2$ in Figure 19.

6.4 Power

Figure 21 shows that the calculated uncertainties of $(1 \pm \delta x)^c$ using Formula (2.45) agree very well with the measured value deviations for $(1 + \delta x)^c$ except when $\delta x = 0.2$ and $c < -1$. The reason why Formula (2.45) is not applicable exactly at the estimated applicable precision threshold $1/\sigma$ needs further discussion. Figure 22 shows the individual variance term of $(1 \pm \delta x)^c$ for the different Taylor expansion order $2n$, for different c and δx .

- Formula (2.49) and (2.50) show that $(1 \pm 0.2)^{-2}$ diverges faster than $(1 \pm 0.2)^{-1}$, which is confirmed by both Figure 21 and 22.
- When the deviation is reduced slightly to below $1/\sigma$ from 0.2 to 0.195, Figure 21 shows that the error deviations become much closer to 1. As the confirmation, Figure 22 shows reducing δx increases the convergence, e.g., $(1 \pm 0.2)^{-1.5}$ diverges while $(1 \pm 0.195)^{-2}$ converges.
- Figure 21 and 22 shows that the convergence is more sensitive to δx than c , so that the applicable precision threshold is relatively hard, which is estimated as $\delta x = 0.196$ for $c = -2$.

Figure 22 shows that the convergence of an analytic function at each input can be judged numerically and automatically for the given maximum Taylor expansion orders.

6.5 Numerical Errors for Library Functions

The combined numerical error of the library function e^x and $\log(x)$ is calculated as $\log(e^x) - x$ vs x . Figure 23 shows that using either the variance arithmetic or the conventional floating-point library functions results in the same value errors. In both cases, the value errors are 0 when $1 < |x|$. The uncertainties of the variance arithmetic bounds the value errors effectively, resulting in an error deviation about 0.409 when $|x| \leq 1$.

The numerical error of the library function x^p is calculated as $(x^p)^{1/p} - x$ vs x . Figure 24 shows that the normalized errors is not specific to either x or p , resulting in an error deviation about 0.548.

When no input noise is added, the error deviation of $\sin(x)^2 + \cos(x)^2 - 1$ is about 0.535

6.6 Summary

Formula (2.33), (2.37), (2.41), and (2.45) gives effective uncertainties for the corresponding library functions. Generally, when the input precision is above 10^{-15} , the input uncertainty achieves ideal coverage.

In ideal coverage cases, the error deviation is very close to 1 except it is 0 at where:

- when the result probability density function is zero, and δx is small enough, the uncertainty mean is finite.
- when the result probability density function is pole, and δx is large enough, the uncertainty mean is infinite.

In non ideal coverage cases, the error deviation is about 0.5 so that proper coverage is achieved.

The convergence of the result variance calculation can be judged numerically.

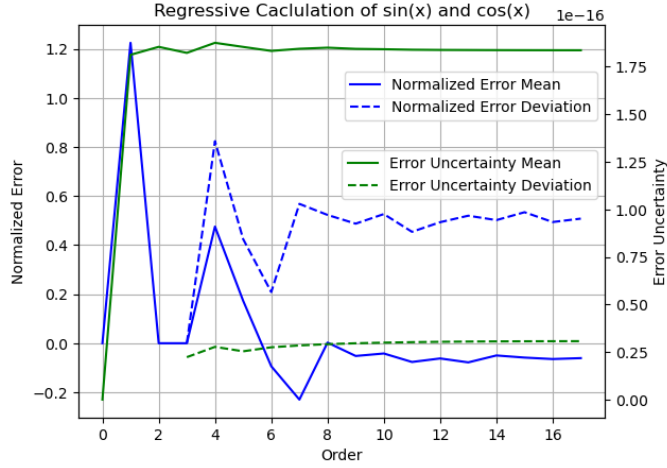


Figure 25: The error deviations and uncertainty means for $\sin(x)^2 + \cos(x)^2 - 1$, $x \in [0, \pi/4]$, for different regression order as shown in the legend.

7 Regressive Generation of Sin/Cos

Starting from Formula (7.1), Formula (7.2) and Formula (7.3) can be used recursively to calculate the sin library function.

$$\sin(0) = \cos\left(\frac{\pi}{2}\right) = 0; \quad \sin\left(\frac{\pi}{2}\right) = \cos(0) = 1; \quad (7.1)$$

$$\sin\left(\frac{\alpha + \beta}{2}\right) = \sqrt{\frac{1 - \cos(\alpha + \beta)}{2}} = \sqrt{\frac{1 - \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)}{2}}; \quad (7.2)$$

$$\cos\left(\frac{\alpha + \beta}{2}\right) = \sqrt{\frac{1 + \cos(\alpha + \beta)}{2}} = \sqrt{\frac{1 + \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)}{2}}; \quad (7.3)$$

The number of regression is defined as the order. For each order n , $2^n \sin(\frac{\pi}{2} \frac{i}{2^n})$ and $\cos(\frac{\pi}{2} \frac{i}{2^n})$ values are obtained, so that the result is more statistically stable with increasing n .

The errors of the generated $\sin(x)$ and $\cos(x)$ are checked by $\sin(x)^2 + \cos(x)^2 - 1$. The resulted value errors are comparable with but not identical to the value errors using floating-point library $\sin(x)$ and $\cos(x)$. Figure 25 shows that the result uncertainties remain almost a constant of about $1.2 \cdot 10^{-16}$, which is consist with uniformly distributed least significant values within the least significant bit of the significand. It also shows that the error deviations is close to 0.5, which is comparable to the value of 0.535 for the library $\sin(x)$ and $\cos(x)$ functions. The variance arithmetic is thus effective for this regressive generation algorithm.

Using either Taylor expansion or regressive generation of $\sin(x)$ and $\cos(x)$ are equivalent precision-wise for conventional floating-point calculations. In contrast, in the variance arithmetic, when $x \rightarrow 0$, Formula (7.2) has very coarse precision for Formula 2.48, which result in very coarse precision for $\sin(x)$. Thus, the variance arithmetic rejects this regressive generation algorithm for $\sin(x)$ and $\cos(x)$.

8 Fast Fourier Transformation

8.1 Unfaithful Frequency Response of Discrete Fourier Transformation [38]

Each testing algorithm needs to come under careful scrutiny. One important issue is whether the digital implementation of the algorithm is faithful for the original analytic algorithm. For example, the discrete Fourier transformation is only faithful for Fourier transformation at certain frequencies, and it has a different degree of faithfulness for other frequencies. This is called the *unfaithful frequency response* of the discrete Fourier transformation.

For each signal sequence $h[k]$, $k = 0, 1 \dots N-1$, in which N is a positive integer, the discrete Fourier transformation $H[n]$, $n = 0, 1 \dots N-1$ and its reverse transformation is given by Formula (8.1) and (8.2), respectively [12], in which j is the *index time* and n is the *index frequency* for the discrete Fourier transformation⁹

$$H[n] = \sum_{k=0}^{N-1} h[k] e^{i2\pi kn/N}; \quad (8.1)$$

$$h[k] = \frac{1}{N} \sum_{n=0}^{N-1} H[n] e^{-i2\pi nk/N}; \quad (8.2)$$

Formula (8.3) is the discrete forward transformation $H[n]$ of a pure sine signal $h[k] = \sin(2\pi kf/N)$ in which f is the index frequency. The continuous forward transformation of the $h[k]$ is a delta function at $n = \pm f$ with phase $\pi/2$. $H[n]$ is delta-like function at $n = \pm f$ with phase $\pi/2$ only if f is an integer. In other cases, how much the result of discrete Fourier transformation deviates from continuous Fourier transformation depends on how much f deviates from an integer, e.g., when f is exactly between two integers, the phase of the transformation is that of cosine instead of sine according to Formula (8.3). Examples of unfaithful representations of fractional frequency by the discrete Fourier transformation are shown in Figure 26. The data for Figure 26 is generated using *SciPy*, which are very reproducible using any other math libraries, including *MathLab* and *Mathematica*.

$$H[n] = \sum_{k=0}^{N-1} \sin(2\pi nk/N) e^{i2\pi nk/N} = \frac{1}{2i} \left(\sum_{k=0}^{N-1} e^{i2\pi(n+f)\frac{k}{N}} - \sum_{k=0}^{N-1} e^{i2\pi(n-f)\frac{k}{N}} \right)$$

$$= \begin{cases} iN/2, & f \text{ is integer} \\ N/\pi, & f \text{ is integer} + 1/2 \\ \frac{1}{2} \frac{\sin(2\pi f - 2\pi \frac{f}{N}) + \sin(2\pi \frac{f}{N}) - \sin(2\pi f) e^{-i2\pi \frac{n}{N}}}{\cos(2\pi \frac{n}{N}) - \cos(2\pi \frac{f}{N})} & \text{otherwise} \end{cases} \quad (8.3)$$

Due to its width, a frequency component in an unfaithful transformation may interact with other frequency components of the Discrete Fourier spectrum, thus sabotaging the whole idea of using the Fourier Transformation to decompose a signal into independent frequency components. Because the reverse discrete Fourier transformation mathematically restores the original $\{h[k]\}$ for any $\{H[n]\}$, it exaggerates

⁹The index frequency and index time are not necessarily related to time unit. The naming is just a convenient way to distinguish the two opposite domains in the Fourier transformation: the waveform domain vs the frequency domain.

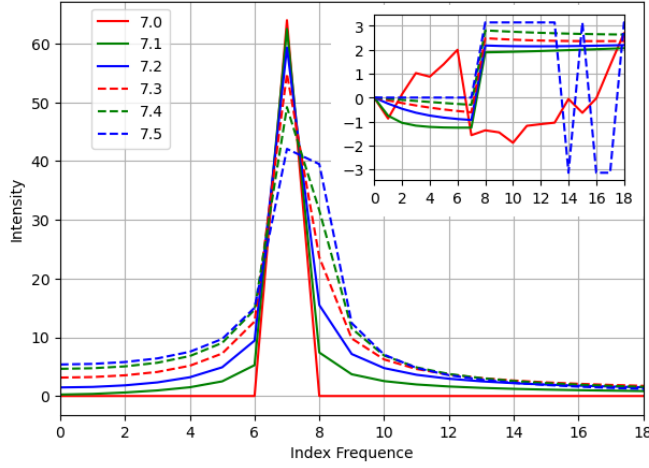


Figure 26: An unfaithful Discrete Fourier transformation is demonstrated by the spectra of a few sine signals having amplitude of 1 and slightly different frequencies as shown in legends. The x-axis shows the scale for the index frequency. The y-axis shows the scale for the intensity, while the y-axis of the embedded picture shows the scale for the phase. This figure is a practical reproduction of a previous theoretical figure [38].

and narrows all unfaithful signal components correspondingly. This means that the common method of signal processing in the Fourier space [12][15][17] may generate artifacts due to its uniform treatment of faithful and unfaithful signal components, which probably coexist in reality. Unlike aliasing [5][12][17], unfaithful representation of the discrete Fourier transformation has an equal presence in the whole frequency range so that it cannot be avoided by sampling the original signal differently.

An unfaithful representation arises from the implied assumption of the discrete Fourier transformation. The continuous Fourier transformation has an infinitive signal range so that:

$$h(t) \Leftrightarrow H(s) : h(t - \tau) \Leftrightarrow H(s)e^{i2\pi s\tau}; \quad (8.4)$$

As an analog, the discrete Fourier transformation $G[n]$ of the signal $h[k], k = 1 \dots N$ can be calculated mathematically from the discrete Fourier transformation $H[n]$ of $h[k], k = 0 \dots N - 1$:

$$G[n] = (H[n] + h[N] - h[0])e^{i2\pi n/N}; \quad (8.5)$$

Applying Formula (8.4) to Formula (8.5) results in Formula (8.6).

$$h[N] = h[0]; \quad (8.6)$$

Thus, the discrete Fourier transformation has an implied assumption that the signal $h[k]$ repeats itself outside the region of $[0, N - 1]$ [43]. For an unfaithful frequency, $h[N - 1]$ and $h[N]$ are discontinuous in regard to signal periodicity, resulting in larger peak width, lower peak height, and the wrong phase.

The unfaithfulness of the discrete Fourier transformation to the Fourier transformation is a very serious example of modeling errors, but this problem has not been addressed seriously enough previously. Because the discrete Fourier transformation has widest applications in science and engineering, this problem needs some serious attention.

8.2 FFT (Fast Fourier Transformation)

When $N = 2^L$, in which L is a positive integer, the generalized Danielson-Lanczos lemma [12] can be applied to the discrete Fourier transformation as FFT [12].

- For each output, each input is only used once, so there is no dependency problem when using Formula (2.7) and (2.11) as arithmetic operations. This simplicity avoids introducing numerical errors from Formula (2.24), because $f_x^{(j)}$ may contain numerical errors, while $(\delta x)^{2n}$ may contain rounding errors.
- When L is large, the large amount of input and output data enables high quality statistical analysis.
- The amount of calculation is L , because for each output, increasing L by 1 results in one additional step of sum of multiplication.
- Each step in the forward transformation thus increases the variance by 2-fold, so that the result uncertainty means increase with the FFT order L as $\sqrt{2}^L$. Because the reverse transformation divides the result by 2^L , the result uncertainty means decrease with the FFT order L as $\sqrt{1/2}^L$. The result uncertainty means for the roundtrip transformations is thus: $\sqrt{2}^L \times \sqrt{1/2}^L = 1$.
- The forward and reverse transformations are identical except a sign, so they are essentially the same algorithm, and their difference is purely due to input data.

A major question for the variance arithmetic is that whether the uncertainties can track the corresponding value errors effectively or not. FFT transformations provide an ideal test for this statistical property: To test if the error deviations can be close to 1 with the increasing FFT order:

- The forward transformation cancels real imprecise data of a Sin/Cos signal into a spectrum of mostly 0 values, so that both its value errors and its result uncertainties are expected to grow faster.
- In contrast, the reverse transformation spreads the spectrum of precise 0 values except at two peaks to data of a Sin/Cos signal, so that both its value errors and its result uncertainties are expected to grow slower.

8.3 Testing Signals

Only Formula (8.1) and (8.2) with integer n and k will be used.

The following signals are used for testing:

- *Sin*: $h[k] = \sin(2\pi k f / N)$, $f = 1, 2, \dots, N/2 - 1$.
- *Cos*: $h[k] = \cos(2\pi k f / N)$, $f = 1, 2, \dots, N/2 - 1$.

- *Linear*: $h[k] = k$, whose discrete Fourier transformation is Formula (8.7).

$$y \equiv i2\pi \frac{n}{N} : G(y) = \sum_{k=0}^{N-1} e^{yk} = \frac{e^{Ny} - 1}{e^y - 1} = \begin{cases} y = 0 : & N \\ y \neq 0 : & 0 \end{cases} ;$$

$$H[n] = \frac{dG}{dy} = \begin{cases} n = 0 : & \frac{N(N-1)}{2} \\ n \neq 0 : & -\frac{N}{2}(1 + i/\tan(n\frac{\pi}{N})) \end{cases} ; \quad (8.7)$$

Empirically, using the indexed sine functions:

- The results from Sin and Cos signals are statistically indistinguishable from each other.
- The results from Sin signals at different frequencies are statistically indistinguishable from each other.

So the results for Sin and Cos signals at all frequencies are pooled together for the statistical analysis, as the *Sin/Cos* signals.

8.4 Library Errors

Formula (8.1) and (8.2) limit the use of $\sin(x)$ and $\cos(x)$ to $x = 2\pi j/N$. To minimize the numerical errors of $\sin(x)$ and $\cos(x)$, *indexed sine functions* can be used instead of the library sine functions:

1. Instead of a floating-point value x for $\sin(x)$ and $\cos(x)$, the integer j is used to specify the input to $\sin(x)$ and $\cos(x)$, as $\sin(2\pi j/N)$ and $\cos(2\pi j/N)$, to avoid the floating-point rounding error of x .
2. The values of the indexed sine functions is extended from $j = 0, 1, \dots, N/8$ to the whole integer region using the periodicity of $\sin(2\pi j/N)$ and $\cos(2\pi j/N)$.

Figure 27 shows that the value errors between the library $\sin(x)$ and the indexed $\sin(x)$ increases with increasing x for FFT order 4. To confirm Figure 27, Figure 28 shows that the value errors of the library $\sin(x)$ for $\sin(j2\pi 2^4/2^6)$ increases with increasing j . The x range in Figure 27 covers all the input used in the FFT transformations of FFT order 4. The errors are also periodic in x by π , which may be amplified in the FFT calculations.

Figure 29 shows the value errors between the library $1/\tan(x)$ and the indexed $1/\tan(x)$ for $x \in (0, \pi)$. It covers all the integer inputs to the indexed cotan functions used for the Linear signal of FFT order 6 and 7. The value errors near $x = \pi$ increase with the FFT order.

Using the maximal absolute value errors, Figure 30 compares the value error strengths of the library $\sin(x)$ and $1/\tan(x)$ for increasing FFT orders. Figure 30 also shows that when checked by $\sin(x)^2 + \cos(x)^2 - 1$, the obtained indexed sine functions have proper coverage. In contrast, Figure 27, 28, and 29 all show that the library $\sin(x)$ and $1/\tan(x)$ do not have proper coverage.

It is clear that the indexed sine functions are much better than the library sine functions. Sadly, in reality, the library sine functions are used ubiquitously, so that both needs to be evaluated.

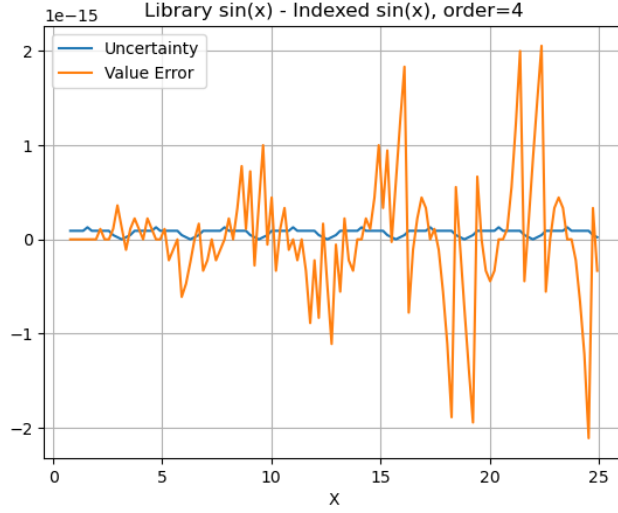


Figure 27: The difference between the library $\sin(x)$ and the indexed $\sin(x)$, for all integer input to the indexed sine functions used in the FFT transformations of FFT order 4. The uncertainty of the $\sin(x)$ values are also displayed, to mark the periodicity of π .

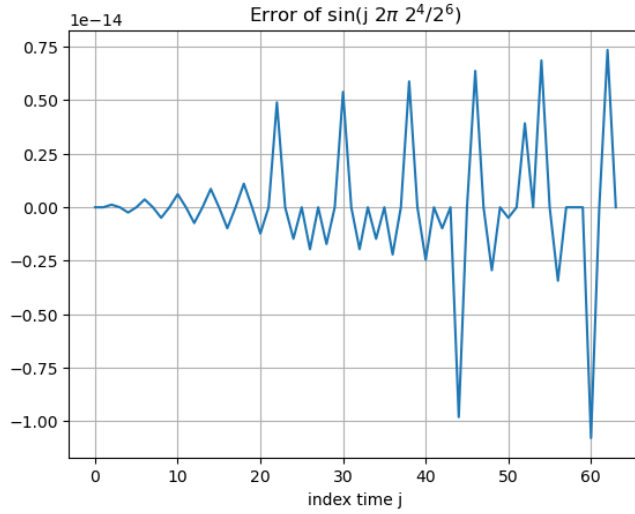


Figure 28: The difference between the library $\sin(x)$ and $\sin(j2\pi 2^4/2^6)$. $\sin(j2\pi 2^4/2^6)$ has precise values $[0, 1, 0, -1, \dots]$, so that the corresponding least significant values are $[0, 2.2 \cdot 10^{-16}, 0, 2.2 \cdot 10^{-16}, \dots]$.

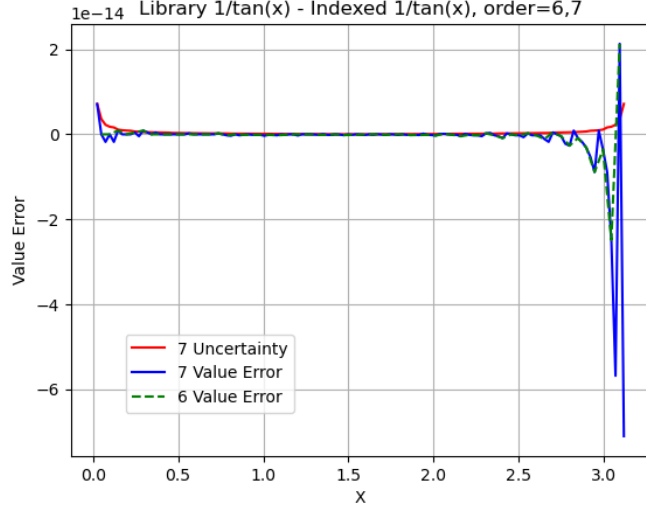


Figure 29: The difference between the library $\cos(x)/\sin(x)$ and the indexed $\cos(x)/\sin(x)$, for $x \in (0, \pi)$.

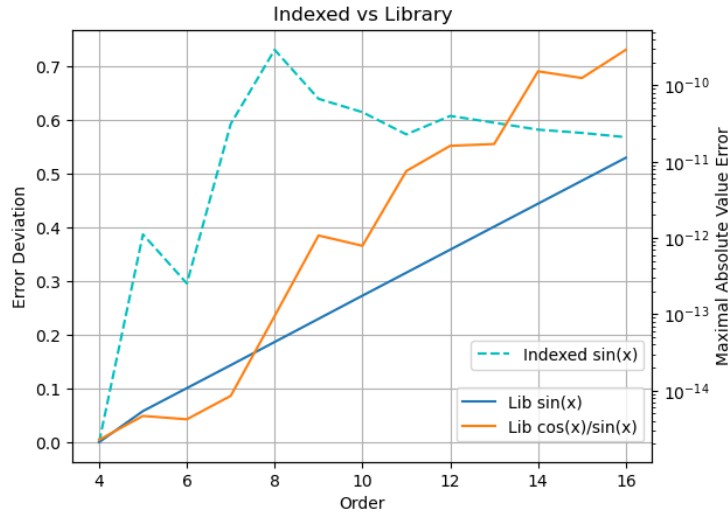


Figure 30: Comparing the indexed sine functions and the library sine functions for different FFT orders. The dashed line is the error deviations of $\sin(x)^2 + \cos(x)^2 - 1$ for the indexed sine functions. Its y-axis is on the left. The two solid lines are the maximal absolute value errors of the library $\sin(x)$ and $\cos(x)/\sin(x)$ when compared with the indexed $\sin(x)$ and $\cos(x)/\sin(x)$. Their y-axis is on the right.

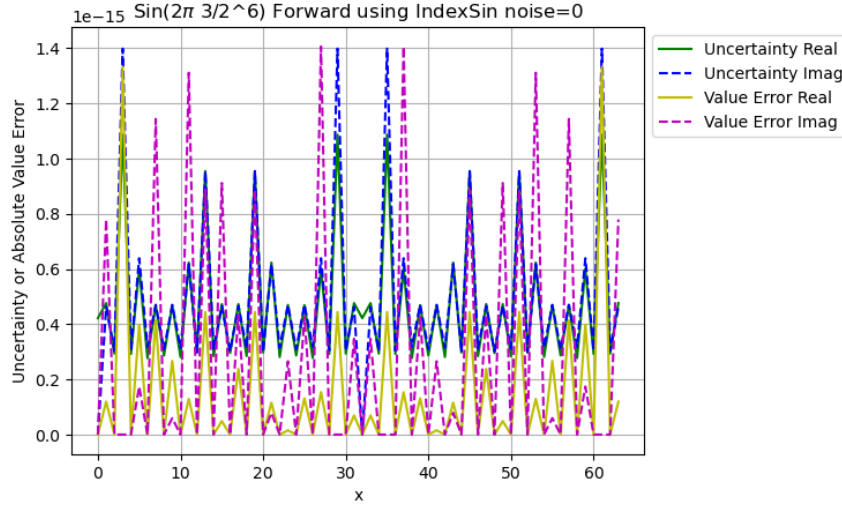


Figure 31: The FFT spectrum of $\sin(j3/2^6\pi)$ using the indexed sine functions after the forward transformation calculated by the variance arithmetic, with the uncertainty and the value errors shown in the legend. The x-axis shows the scale for index frequency. The y-axis shows the scale for uncertainty and absolute value errors.

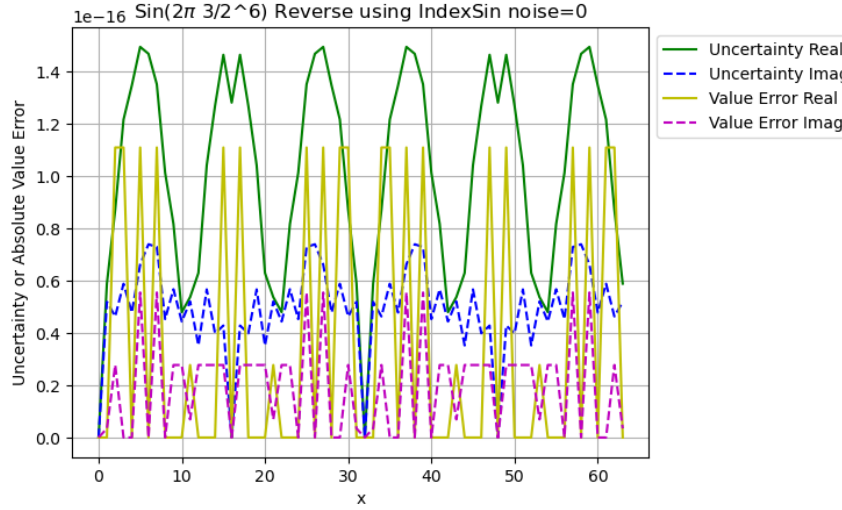


Figure 32: The FFT waveform of $\sin(j3/2^6\pi)$ using the indexed sine functions after the reverse transformation calculated by the variance arithmetic, with the uncertainty and the value errors shown in the legend. The x-axis shows the scale for index time. The y-axis shows the scale for uncertainty and absolute value errors.

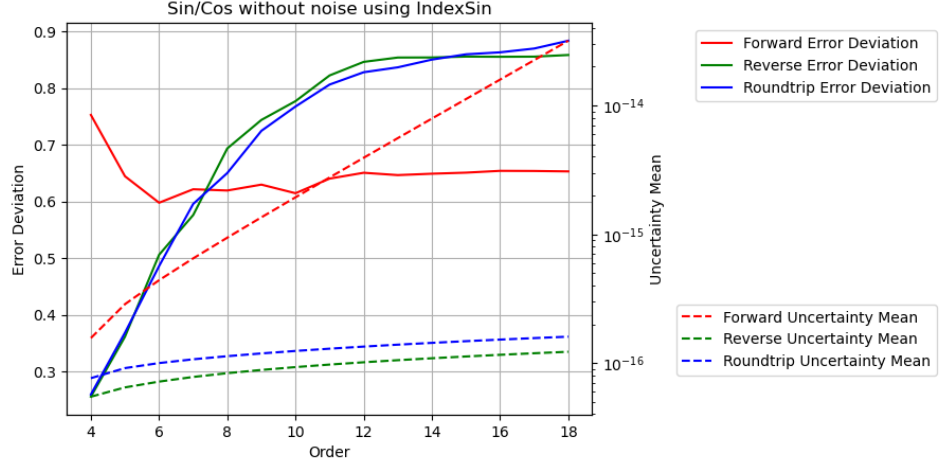


Figure 33: The result error deviations and uncertainty means of Sin/Cos signals vs. FFT order using the indexed sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. The error deviations are scaled to the linear y-axis on the left, while the uncertainty means are scaled to the logarithmic log y-axis on the right.

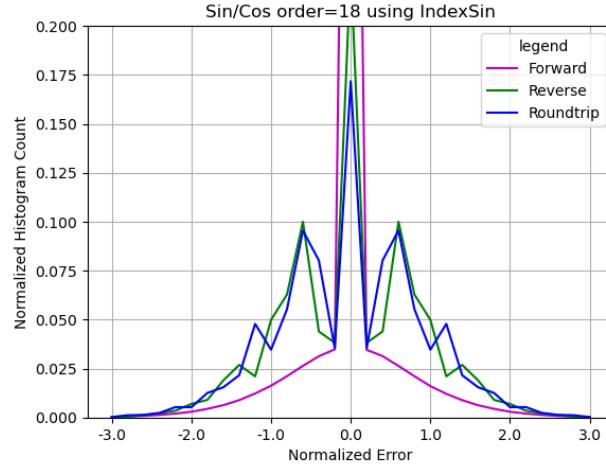


Figure 34: The histograms of the normalized errors of Sin/Cos signals without input noises using the indexed sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. The FFT order is 18.

8.5 Using the Indexed Sine Functions for Sin/Cos Signals

Using the indexed sine functions, for the waveform of $\sin(2\pi k 3/2^6)$ in which k is the index time:

- Figure 31 shows that for the forward transformation, the result value errors are comparable to the result uncertainties, with an error deviation of 0.37 for the real part, and 0.93 for the imaginary part.
- Figure 32 shows that for the reverse transformation, the result value errors are comparable to the result uncertainties, with an error deviation of 0.54 for the real part, and 0.52 for the imaginary part.

Figure 33 shows both the result uncertainty means and the error deviations vs. FFT order of FFT transformations of Sin/Cos signals using the indexed sine functions. Figure 34 shows the corresponding histograms at FFT order 17.

- As expected, the uncertainties grow much faster with the increasing FFT order for the forward transformation than those for the reverse transformation. Both are faster enough to achieve proper coverage.
- The faster growth of the uncertainties of the forward transformation in Figure 33 results in almost Gaussian distribution of the normalized errors for FFT order 18 in Figure 34. Thus, the forward transformation is expected to reach ideal coverage quicker with any added noise, and it is less sensitive to numerical calculation errors.
- The slower growth of the uncertainties of the reverse transformation in Figure 33 results in structured distribution on top of a Gaussian distribution of the normalized errors for FFT order 18 in Figure 34. Thus, the reverse transformation is expected to reach ideal coverage slower with added input noises, and it is more sensitive to numerical calculation errors.
- With the increasing FFT order, all histograms become more Gaussian like, and the error deviations for the real and the imaginary parts become more equal in value. The error deviations for the forward transformation are relatively stable. The error deviations for the reverse transformation increase with the increasing FFT orders until become stable when FFT orders is 12 or more. When the FFT order is about 12, statistical stability is reach.

Thus, proper coverage is achieved using the indexed sine functions for Sin/Cos signals.

8.6 Using the Library Sine Functions for Sin/Cos Signals

Because the least significant values are the only source of input uncertainties for variance arithmetic, the result uncertainties using either sine functions are almost identical.

The library sine functions contributes to the numerical calculation errors which is not specified by the input uncertainties of the variance arithmetic. The question is whether the variance arithmetic can track this large amount of numerical errors effectively or not.

Using the library sine functions, for the waveform of $\sin(2\pi k 3/2^6)$ in which k is the index time:

- Figure 35 shows that for the forward transformation, the result value errors are noticeably larger than the result uncertainties, with an error deviation of 7.0 for

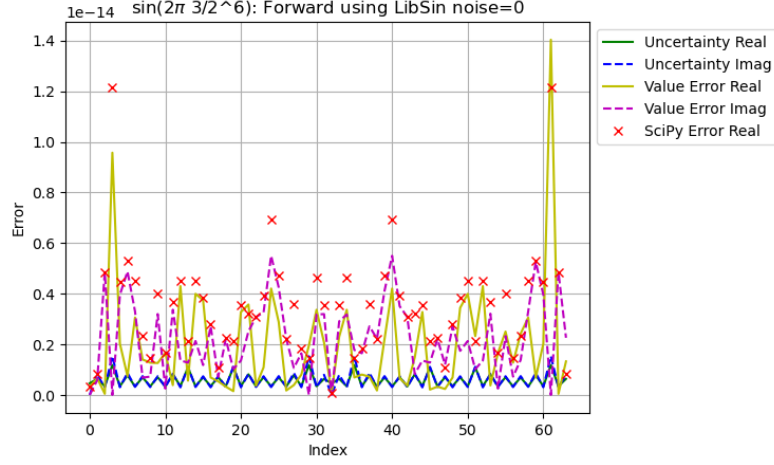


Figure 35: The FFT spectrum of $\sin(j3/2^6\pi)$ using the library sine functions after the forward transformation calculated by the variance arithmetic, with the uncertainties and the value errors shown in the legend. Also included is the corresponding result errors using *SciPy*. The x-axis shows the scale for index frequency. The y-axis shows uncertainties or absolute value errors.

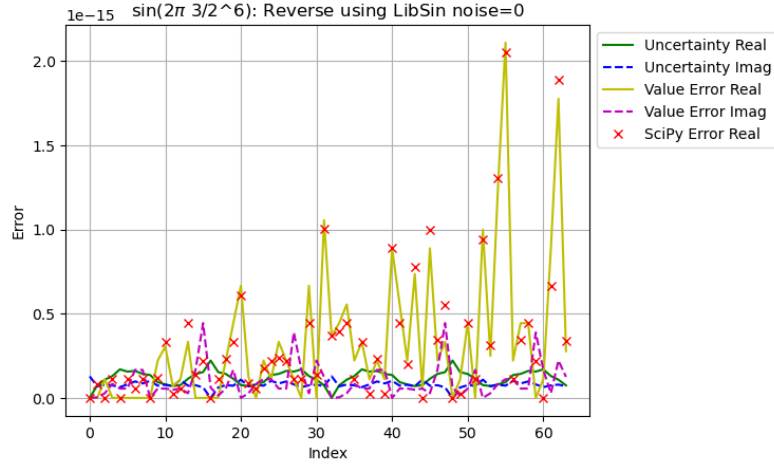


Figure 36: The FFT waveform of $\sin(j3/2^6\pi)$ using the indexed sine functions after the reverse transformation calculated by the variance arithmetic, with the uncertainties and the value errors shown in the legend. Also included is the corresponding result errors using *SciPy*. The x-axis shows the scale for index time. The y-axis shows the uncertainties or absolute value errors. No input noise is added to the Sin signal.

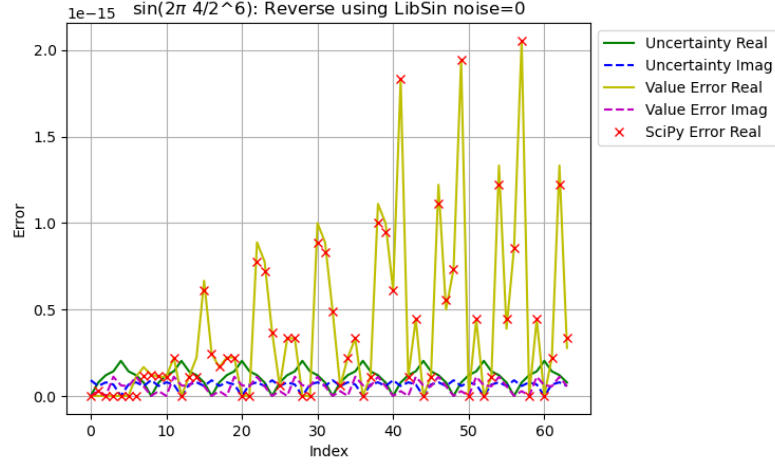


Figure 37: The FFT waveform of $\sin(j4/2^6\pi)$ using the indexed sine functions after the reverse transformation calculated by the variance arithmetic, with the uncertainties and the value errors shown in the legend. Also included is the corresponding result errors using *SciPy*. The x-axis shows the scale for index time. The y-axis shows the uncertainties or absolute value errors. No input noise is added to the Sin signal.

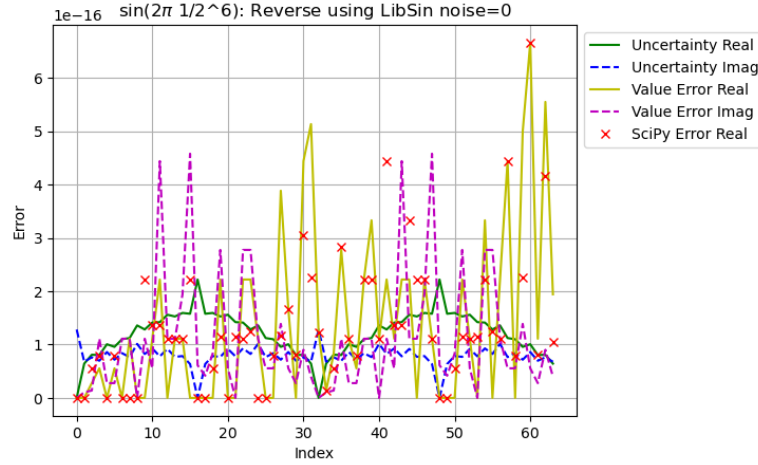


Figure 38: The FFT waveform of $\sin(j1/2^6\pi)$ using the indexed sine functions after the reverse transformation calculated by the variance arithmetic, with the uncertainties and the value errors shown in the legend. Also included is the corresponding result errors using *SciPy*. The x-axis shows the scale for index time. The y-axis shows the uncertainties or absolute value errors. No input noise is added to the Sin signal.

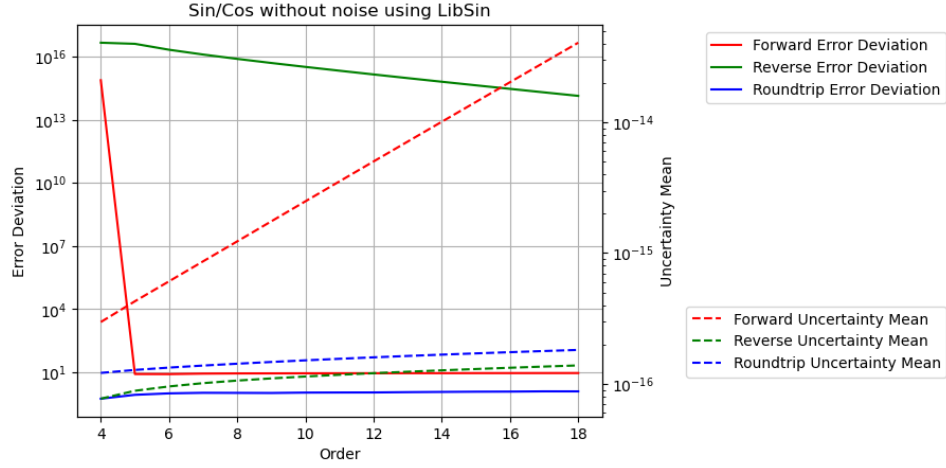


Figure 39: The result error deviations and uncertainty means of Sin/Cos signals vs. FFT order using the library sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. The error deviations are scaled to the linear y-axis on the left, while the uncertainty means are scaled to the logarithmic log y-axis on the right. No input noise is added to the Sin signal.

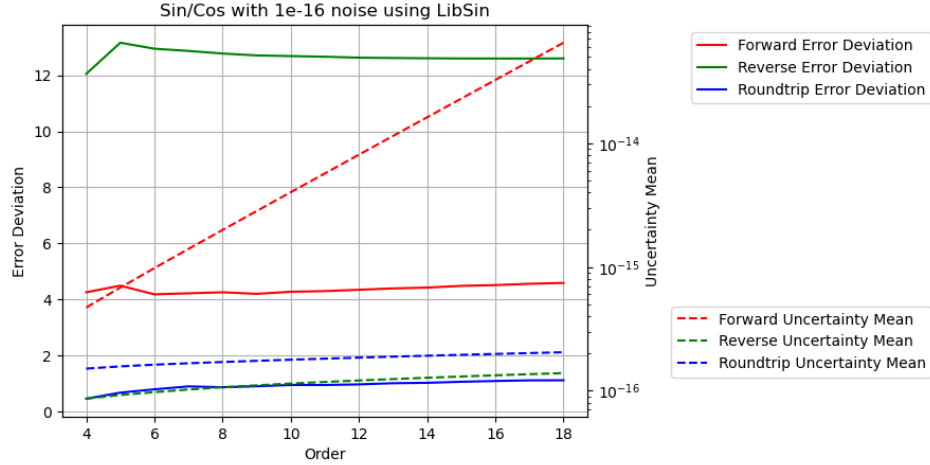


Figure 40: The result error deviations and uncertainty means of Sin/Cos signals vs. FFT order using the library sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. The error deviations are scaled to the linear y-axis on the left, while the uncertainty means are scaled to the logarithmic log y-axis on the right. 10^{-16} input noises are added to the Sin/Cos signals.

the real part, and 5.6 for the imaginary part. As expected, the error deviations are noticeably larger than their counterparts in Figure 31.

- Figure ?? shows that for the reverse transformation, the result value errors are noticeably larger than the result uncertainties, with an error deviation of 5.7 for the real part, and $0.83 \cdot 10^{15}$ for the imaginary part. The surprisingly large imaginary error deviation is caused by the small uncertainty at the index time where the value error is at a local maximum.
- As the result of the huge error deviation, Figure 39 shows that the result of the reverse transformation no longer has proper coverage for all FFT orders. When a small noise with deviation of 10^{16} is added to the input, the result uncertainty of the reverse transformation at the index time of 8 is no longer near 0, so that the result error deviations achieve proper coverage again, as shown in Figure 40.

To validate the FFT implementation in the variance arithmetic, the results are compared with the corresponding calculations using the python numerical library *SciPy* in Figure 35 and ??. The results are quite comparable, except that in *SciPy*, the data waveform is always real rather than complex. In Figure 35, the *SciPy* results have been made identical for the frequency indexes f and $-f$, and the matching to the corresponding results of the variance arithmetic is indicative. In Figure ??, the *SciPy* results match the corresponding results of the variance arithmetic well.

In Figure ??, the value errors tend to increase with the indexes, which is due to the periodic increase of the numerical errors of $\sin(x)$ with x as shown in Figure 27. When the signal frequency increases, the increase of the value errors with the index frequency in the reverse transformation becomes stronger, as shown in Figure 38, ??, and 37. On the other hand, Figure 32 shows no such increase, because it contains no library errors as described in Figure 27.

The variance arithmetic reveals the reason why to add small noises to the numerically generated input data, which is already a common practice [12]. How much noises to add to input to achieve proper coverage remains an art. When the proper tracking of the variance arithmetic is not achieved, the result may contain large amount of errors.

8.7 Linear Signal

In addition to the numerical errors of $\sin(\pi j/2^L)$, $j = 0, 1, 2 \dots 2^{L-1}$, Linear signals introduce the numerical errors of $1/\tan(\pi(j-1)/2^L)$, $j = 1, 2 \dots 2^{L-1}$ to the result. The question is whether the variance arithmetic can track this additional numerical errors or not.

Using the indexed sine functions:

- Figure 41 shows that proper coverage can be achieved for all the FFT transformations for all FFT orders.
- Figure 42 shows that for each transformation, the histogram of the Linear signal are quite similar to the counterpart of the Sin/Cos signals in Figure 34.

Using the library sine functions:

- Figure 43 shows that proper coverage can not be achieved any FFT transformation for all FFT orders, because the value errors outpace the uncertainties with the increasing FFT order. Because of this, adding noise to the input can not achieve proper coverage.

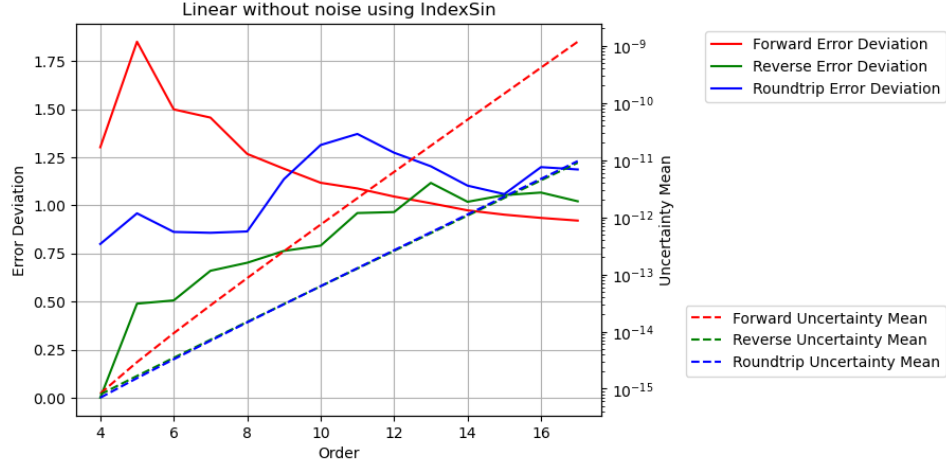


Figure 41: The result error deviations and uncertainty means of Linear signal vs. FFT order using the library sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. The error deviations are scaled to the linear y-axis on the left, while the uncertainty means are scaled to the logarithmic log y-axis on the right. No input noise is added to the Linear signal.

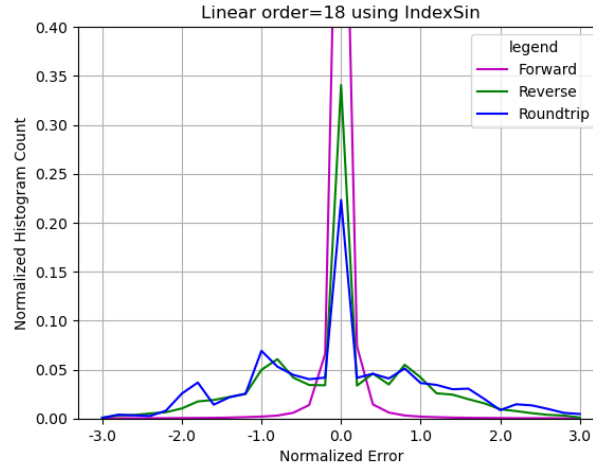


Figure 42: The histograms of the normalized errors of Sin/Cos signals without input noises using the indexed sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. No input noise is added to the Linear signal.

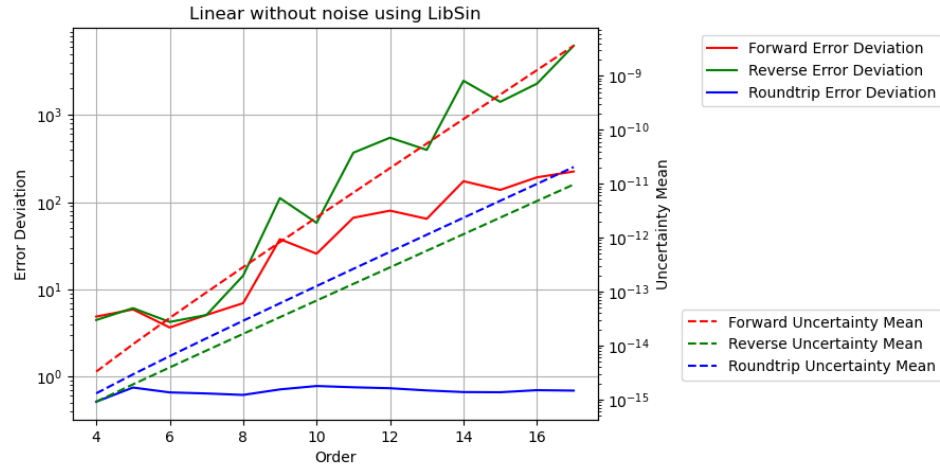


Figure 43: The result error deviations and uncertainty means of Linear signal vs. FFT order using the library sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. The error deviations are scaled to the linear y-axis on the left, while the uncertainty means are scaled to the logarithmic log y-axis on the right. No input noise is added to the Linear signal.

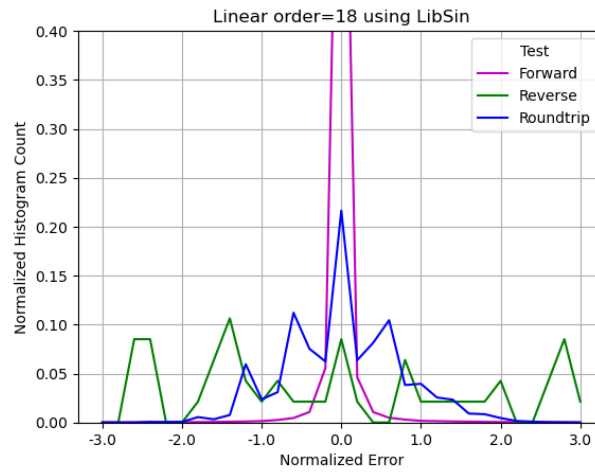


Figure 44: The histograms of the normalized errors of Sin/Cos signals without input noises using the indexed sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. No input noise is added to the Linear signal.

- Figure 44 shows that the reverse histogram contains additional peaks in addition to its counterpart in Figure 43. The additional peaks extend beyond the range of $[-3, +3]$

When too much numerical errors are not specified by the input uncertainties, proper coverage can not be achieved, which is the case when the $\sin(x)$ numerical errors in Figure 27 is combined with the $1/\tan(x)$ numerical errors in Figure 29. The only solution seems to create a new sin library using the variance arithmetic so that all numerical calculation errors are accounted for.

8.8 Ideal Coverage

Even when proper coverage can not be achieved, adding enough noise to the input can overpower the numerical calculation errors, to achieve ideal coverage.

Figure 49 shows that with 10^{-3} input noise, the result normalized errors for the forward and reverse transformations are both normal distributed, while the result normalized errors for the roundtrip transformation is delta distributed, regardless of the types of input noises.

Figure 50 shows the result error deviations and uncertainty means vs. different input uncertainties for Linear signal of order 18 using the library sin/cos functions. When the deviation of the input noise is more than 10^{-3} , the input uncertainties become ideal coverage:

- As expected, the result uncertainty means for the forward transformations increase with the FFT order L as $\sqrt{2}^L$.
- As expected, the result uncertainty means for the reverse transformations decrease with the FFT order L as $\sqrt{1/2}^L$.
- As expected, the result uncertainty means for the roundtrip transformations always equal the corresponding input uncertainties.
- As expected, the result uncertainty means for both the forward and the reverse transformations are linear to the input uncertainties, respectively, because FFT transformations are linear. As expected, the result uncertainty means for the roundtrip transformation recover the corresponding input uncertainties perfectly.
- As expected, the normalized errors for the forward and the reverse transformations are normal distributed, even when the input noise is no longer Gaussian. The normalized errors for the roundtrip transformations are delta distributed at 0, meaning the input uncertainties are perfectly recovered.
- As expected, the result error deviations for the forward and reverse transformations are constant 1, while the result error deviations for the roundtrip transformation approaches 0 exponentially with the increasing FFT order.

For Linear signals using the library sine functions, the the added noise vs FFT order regions for the ideal coverage are shown as the regions where the error deviations are 1, by Figure 51 and 52 for the forward and reverse transformations, respectively. In other regions, proper coverage is not achievable. Because the uncertainties grow slower in the reverse transformation than in the forward transformation, the reverse transformations have smaller ideal coverage region than that of the forward transformation. Because the amount of numerical errors increases with the amount of calculations, the input noise range reduces with the increasing FFT order. It is possible that ideal coverage is

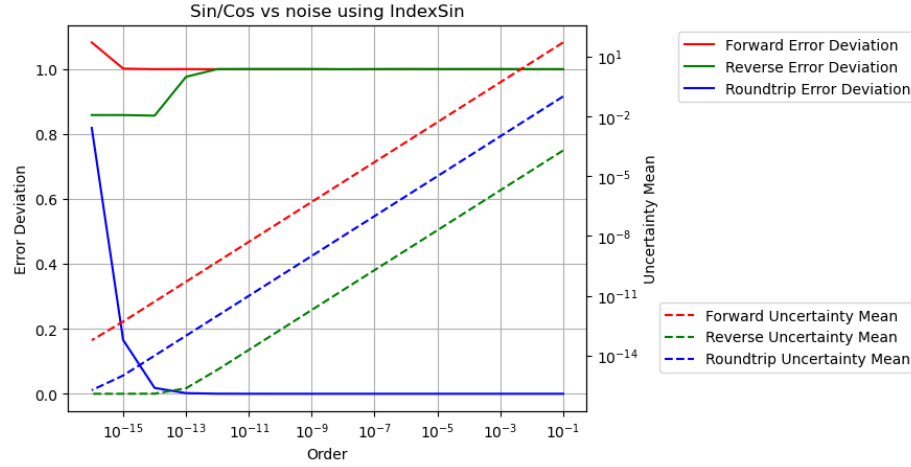


Figure 45: The result error deviations and uncertainty means for Sin/Cos signals of order 16 using indexed sin/cos functions vs. input uncertainties for forward, reverse and roundtrip transformations for FFT order 18. The error deviations are scaled to the linear y-axis on the right, while the uncertainty means are scaled to the logarithmic log y-axis on the right.

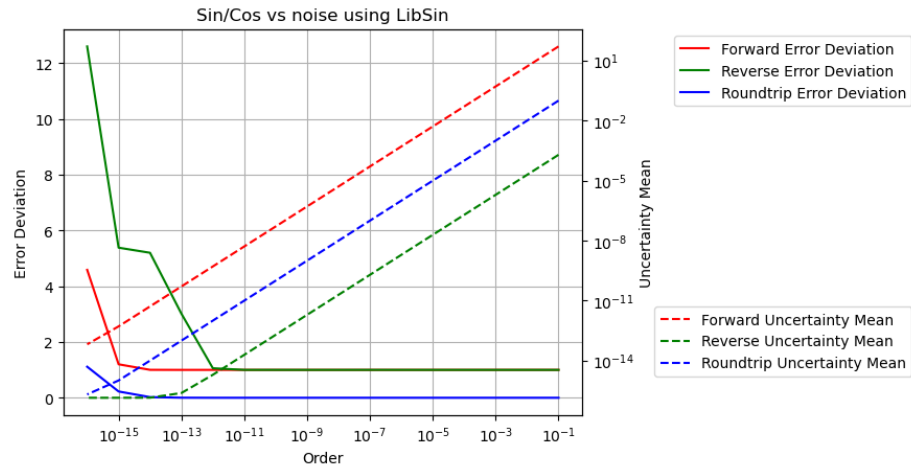


Figure 46: The result error deviations and uncertainty means for Sin/Cos signals of order 16 using library sin/cos functions vs. input uncertainties for forward, reverse and roundtrip transformations for FFT order 18. The error deviations are scaled to the linear y-axis on the right, while the uncertainty means are scaled to the logarithmic log y-axis on the right.

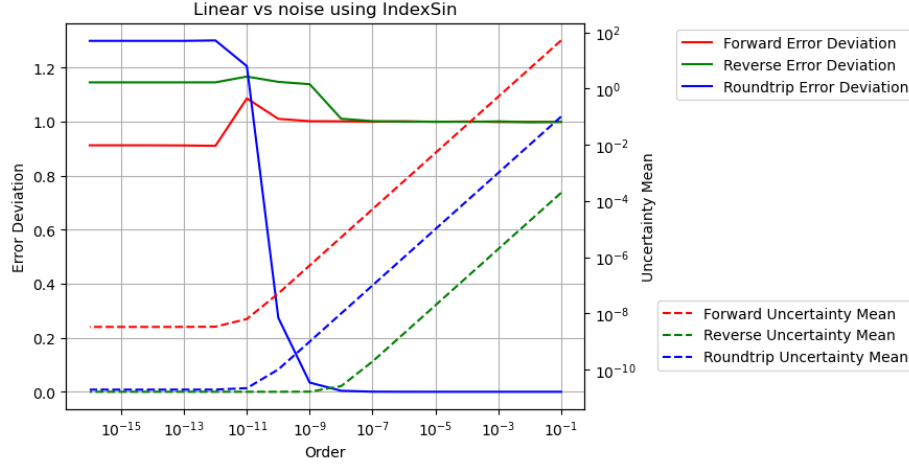


Figure 47: The result error deviations and uncertainty means using Linear signals of order 18 vs. input uncertainties for forward, reverse and roundtrip transformations for FFT order 18. The error deviations are scaled to the linear y-axis on the right, while the uncertainty means are scaled to the logarithmic log y-axis on the right.

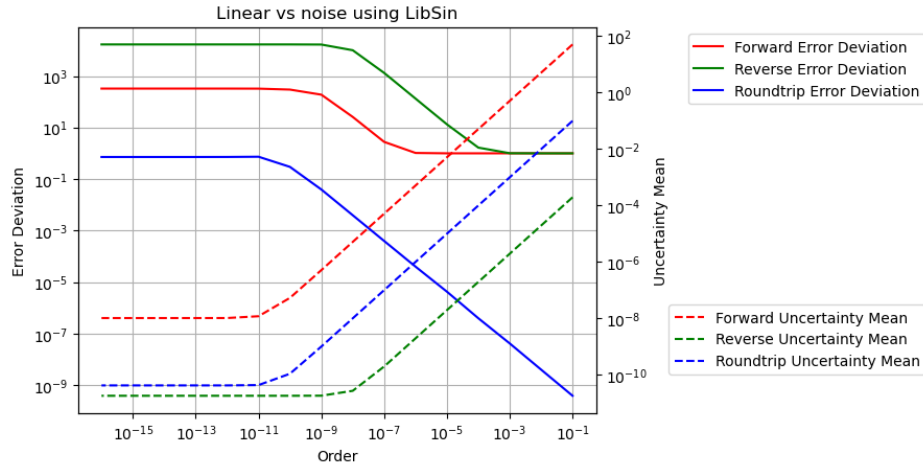


Figure 48: The result error deviations and uncertainty means using Linear signals of order 18 vs. input uncertainties for forward, reverse and roundtrip transformations for FFT order 18. The error deviations are scaled to the linear y-axis on the right, while the uncertainty means are scaled to the logarithmic log y-axis on the right.

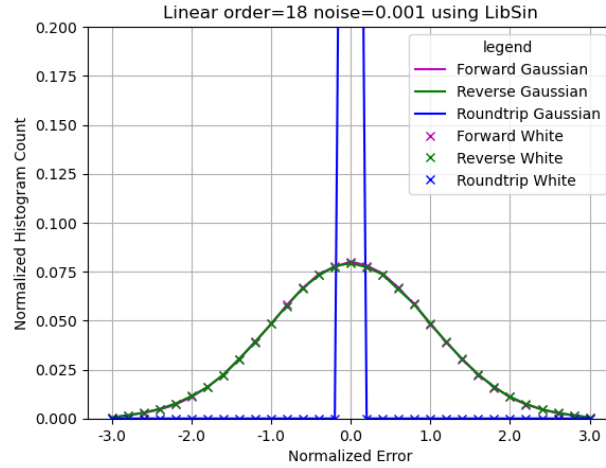


Figure 49: The histograms of the normalized errors of Linear signal without input noises using the library sine functions for forward, reverse and roundtrip FFT transformations, as shown in the legend. 10^{-3} input noises are added to the Sin/Cos signals.

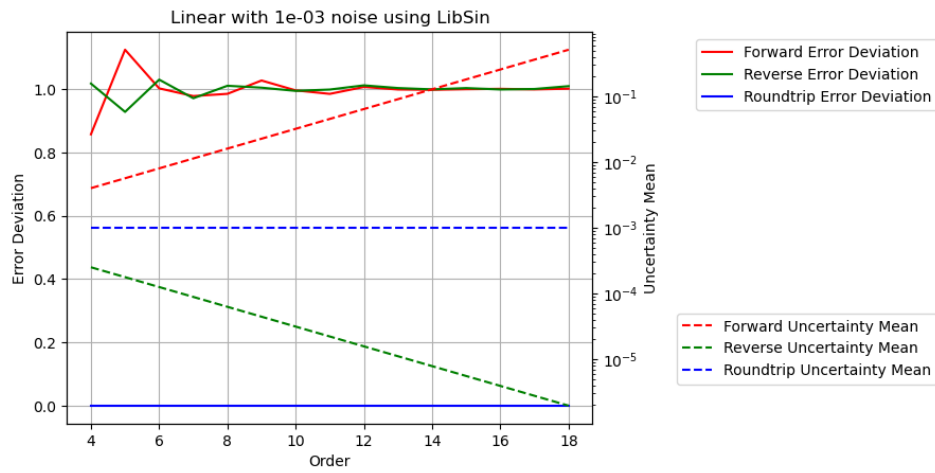


Figure 50: The result error deviations and uncertainty means using Linear signal with 10^{-3} input noises vs. FFT order for forward, reverse and roundtrip FFT transformations. The error deviations are scaled to the linear y-axis on the left, while the uncertainty means are scaled to the logarithmic log y-axis on the right.

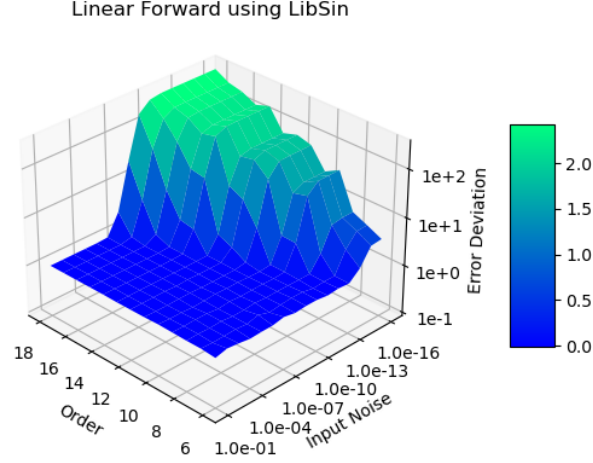


Figure 51: The result error deviations for Linear signals using library sine functions vs. input uncertainties and FFT orders for the forward transformations. The input uncertainties run from 10^{-16} to 10^{-1} , while the FFT Order runs from 6 to 18.

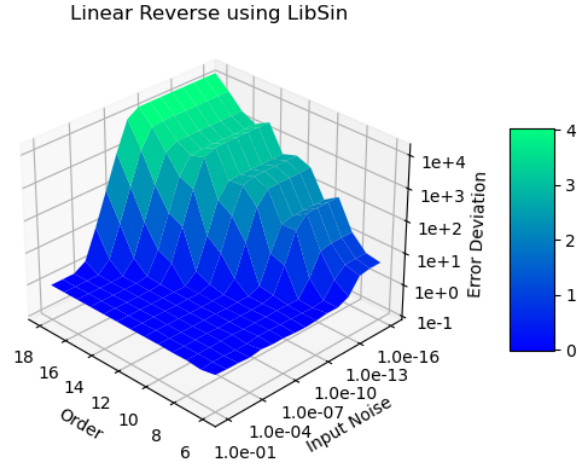


Figure 52: The result error deviations for Linear signals using library sine functions vs. input uncertainties and FFT orders for the reverse transformations. The input uncertainties run from 10^{-16} to 10^{-1} , while the FFT Order runs from 6 to 18.

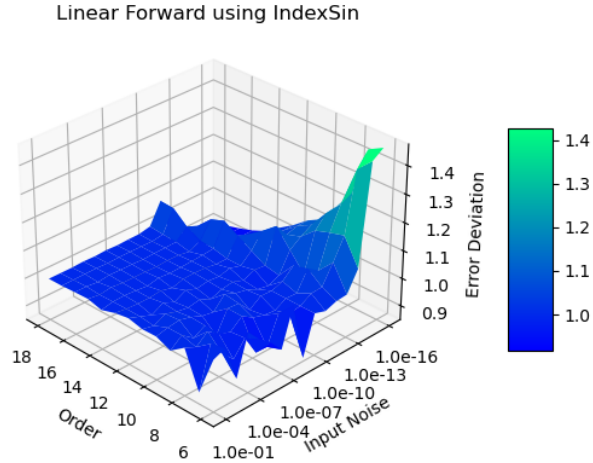


Figure 53: The result error deviations for Linear signals using indexed sine functions vs. input uncertainties and FFT orders for the forward transformations. The input uncertainties run from 10^{-16} to 10^{-1} , while the FFT Order runs from 6 to 18.

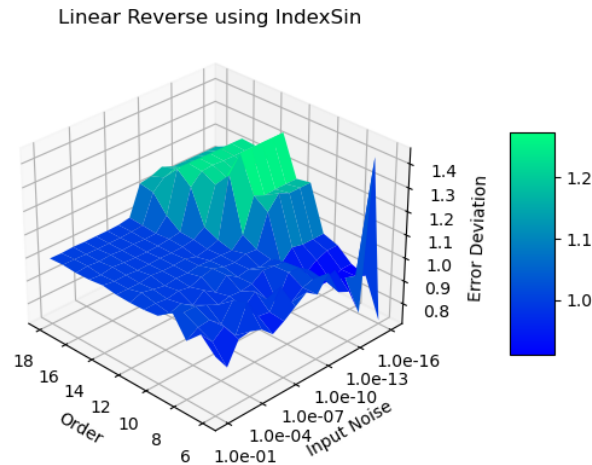


Figure 54: The result error deviations for Linear signals using indexed sine functions vs. input uncertainties and FFT orders for the reverse transformations. The input uncertainties run from 10^{-16} to 10^{-1} , while the FFT Order runs from 6 to 18.

not achievable at all, e.g., visually, when the FFT order is larger than 25 for the reverse transformation. As one of the most robust numerical algorithm that is very insensitive to input errors, FFT breaks down when the FFT order is more than 25 due to the numerical errors in the library sine functions, and such deterioration of the calculation result is not easily detectable using the conventional floating-point calculation.

In contrast, for Linear signals using the indexed sine functions, the the added noise vs FFT order regions for the ideal coverage are shown as the regions where the error deviations are 1, by Figure 53 and 54 for the forward and reverse transformations, respectively. The ideal regions are much larger. In other regions, proper coverage are achieved.

8.9 Summary

Compared to its counterpart continuous Fourier Transformation (FT), the discrete Fourier transformation (DFT) has large modeling error for its implied assumption that any waveform is periodic outside its defined time range. This modeling error has not been addressed seriously.

The library sine functions using conventional floating-point arithmetic have been shown to contain numerical errors as large as equivalently 10^{-3} of input accuracy for FFT transformations. The library $\sin(x)$ errors increase periodically with x , causing noticeable increase of the result errors with the increasing index frequency in the reverse transformation. The dependency of the result numerical errors on the amount of calculation and input data means that a small scale test can not properly qualify the result of a large scale calculation. The effect of numerical errors inside math library has not been addressed seriously.

The variance arithmetic should be used, whose values largely reproduce the corresponding results using conventional floating-point arithmetic, whose uncertainties trace all input errors, and whose result error deviations qualify the calculation quality as either ideal, or proper, or suspicious. If the library functions also have proper coverage, the calculation result will probably have proper coverage as well.

9 Comparison Using Progressive Moving-Window Linear Regression

9.1 Progressive Moving-Window Linear Regression Algorithm

Formula (9.1) gives the result of the least-square line-fit of $Y = \alpha + \beta X$ between two set of data Y_j and X_j , in which j is an integer index to identify (X, Y) pairs in the sets [12].

$$\begin{aligned}\alpha &= \frac{\sum_j Y_j}{\sum_j 1}; \\ \beta &= \frac{\sum_j X_j Y_j \sum_j 1 - \sum_j X_j \sum_j Y_j}{\sum_j X_j X_j \sum_j 1 - \sum_j X_j \sum_j X_j};\end{aligned}\tag{9.1}$$

In many applications data set Y_j is an input data stream collected with fixed rate in time, such as a data stream collected by an ADC (Analogue-to-Digital Converter) [5]. Y_j is called a time-series input, in which j indicates time. A moving window algorithm [12] is performed in a small time-window around each j . For each window of calculation, X_j can be chosen to be integers in the range of $[-H, +H]$ in which H is an integer constant specifying window's half width so that $\sum_j X_j = 0$, to reduce (9.1) into (9.2):

$$\begin{aligned}\alpha_j &= \alpha \ 2H = \sum_{X=-H+1}^H Y_{j-H+X}; \\ \beta_j &= \beta \frac{H(H+1)(2H+1)}{3} = \sum_{X=-H}^H X Y_{j-H+X};\end{aligned}\tag{9.2}$$

According to Figure ??, in which H takes an example value of 4, the calculation of (α_j, β_j) can be obtained from the previous values of $(\alpha_{j-1}, \beta_{j-1})$, to reduce the calculation of (9.2) into a progressive moving-window calculation of (9.3):

$$\begin{aligned}\beta_j &= \beta_{j-1} - \alpha_{j-1} + H(Y_{j-2H-1} + Y_j); \\ \alpha_j &= \alpha_{j-1} - Y_{j-2H-1} + Y_j;\end{aligned}\tag{9.3}$$

9.2 Dependency Problem in a Progressive Algorithm

(9.3) uses each input multiple times, so it will have dependency problem for all the three uncertainty-bearing arithmetic. The question is how the overestimation of uncertainty evolves with time.

The moving-window linear regression is done on a straight line with a constant slope of exactly $1/1024$ for each advance of time, with a full window width of 9 data points, or $H = 4$. Both average output value errors and deviations of all three arithmetic increases linearly with input deviations, and increase monotonically with time. Thus both the average output tracking ratio and the maximal output bounding ratio are largely independent of input precisions, e.g., Figure ?? shows such trend for the

average output tracking ratio using variance arithmetic. Such independence to input precision is expected for linear algorithms in general [12]. Therefore, only results with the input deviation of 10^{-3} are shown for the remaining discussions unless otherwise specified. Figure ?? shows the output deviation and the value errors vs. time while Figure ?? shows the output average tracking ratios and the maximal bounding ratios vs. time for all three arithmetics.

For interval arithmetic and independence arithmetic, the output value errors remain on a constant level, while the output deviations increase with time, so that both output average tracking ratios and maximal bounding ratios decrease with time. The stable linear increase of output deviation with time using either interval arithmetic or independence arithmetic in Figure ?? suggests that the progressive linear regression calculation has accumulated every input uncertainty, which results in the monotonic decrease of both the maximal bounding ratios and the average output tracking ratios with time using both arithmetics in Figure ??.

In contrast, while variance arithmetic has slightly larger output deviations than those of independence arithmetic, its output value errors follows its output deviations, so that both its tracking ratios and bounding ratios remain between 0.1 and 0.9. The reason for such increase of output value errors with time is due to the fact that variance arithmetic calculates only limited bits inside uncertainty, and uses larger granularity of values in calculation for larger uncertainty deviation. Such granularity of calculation is evident when comparing 2-bit or 4-bit calculation inside uncertainty using variance arithmetic in Figure ???. This mechanism of error tracking in variance arithmetic is also demonstrated in Figure ?? and Figure ??. Figure ?? shows that for fewer bits calculated inside uncertainty, the output value errors follow the output deviation closer in time, but such usage of larger granularity of values in calculation causes the result to become insignificant sooner, while for more bits calculated inside uncertainty, the average tracking ratios initially follow the result using independence arithmetic longer, and then follow the output deviation for longer duration. The similarity in patterns of the average tracking ratios for different bits calculated inside uncertainty using variance arithmetic in Figure ?? suggests that they are all driven by a same mechanism but on different time scale, which is expected when smaller granularity of error needs more time to accumulate to a same level. From the definition of tracking ratio, the granularity of error is actually measured in term of granularity of precision, e.g., Figure ?? shows that for same bits calculated inside uncertainty, smaller input uncertainty deviations results in longer tracking of the output value errors to the output deviations. The similar pattern of average tracking ratios is repeated on slower time scale for smaller input uncertainty deviations in Figure ??, revealing similar underline error-tracking mechanism in both cases. Figure ?? also shows that for the same bits calculated inside uncertainty, the average tracking ratios deviate from independence at exactly the same time. Figure ?? and Figure ?? thus demonstrate a uniform and granular error tracking mechanism of the variance arithmetic for different bits calculated inside uncertainty.

Is such increase of the value errors with the increase of uncertainty deviation using variance arithmetic desired? First, in real calculations the correct answer is not known, and the reliability of a result depends statistically on the uncertainty of the result, so that there is no reason to assume that calculating more bits inside uncertainty is any better. Conceptually, when the uncertainty of a calculation increases, the value error of the calculation is also expected to increase, which agrees with the trend shown by variance arithmetic. Second, the stability of the average output tracking ratios and the maximal bounding ratios of variance arithmetic is quite valuable in interpretation

results. For example, even the output deviation may have unexpectedly changed, as in this case if dependency problem were not known and expected, such stability still gives a good estimation of the value errors in the result using variance arithmetic. Third, such stability ensures that the result of algorithm at each window does not depend strongly on the usage history of the algorithm, which makes variance arithmetic the only practically usable uncertainty-bearing arithmetic for this progressive algorithm. To test the effect of usage history on each uncertainty-bearing arithmetic, noise is increased by 10-fold at the middle 1/3 duration of the straight line, to result in additional two test cases:

- *Changed*: In Figure ?? and ??, the input deviation is also increased by 10-fold to simulate an increase in measured uncertainty.
- *Defective*: In Figure ?? and ??, the input deviation remains the same to simulate the defect in obtaining the uncertainty deviations.

Accordingly, the original case of linear regression on a line with fixed slope is named as *Simple*.

The question is how each uncertainty-bearing arithmetic responses to this change of data in the last 1/3 duration of calculation. Using either independence or interval arithmetic, both the average output tracking ratios and the maximal output bounding ratios are decrease by about 10-fold in Figure ?? while they are not affected at all in Figure ?. They show extreme sensitivity to the usage history. Because the real input data are neither controllable nor predictable, the result uncertainty for this progressive algorithm using either interval arithmetic or independence arithmetic may no longer be interpretable. In contrast, using variance arithmetic, both the average output tracking ratios and the maximal output bounding ratios are relatively stable, while the output deviations and value errors are sensitivity to usage history, so that the result using variance arithmetic is still interpretable.

9.3 Choosing a Better Algorithm for Imprecise Inputs

Formula (9.3) has much less calculations than Formula (9.2), and it seems a highly efficient and optimized algorithm according to conventional criterion [12]. However, from the perspective of uncertainty-bearing arithmetic, Formula (9.3) is progressive while Formula (9.2) is expressive, so that Formula (9.2) should be better. Figure ?? and Figure ?? respectively show the output deviations and the value errors vs. time for using either Formula (9.3) or Formula (9.2) of a straight line with 10-fold increase of input uncertainty in the middle 1/3 duration. They show that while the progressive algorithm carries all the historical calculation uncertainty into future, the expressive algorithm is clean from any previous results. For example, at the last 1/3 duration when the moving window is already out of the area for the larger input uncertainty, the progressive algorithm still gives large result uncertainty, while the expressive algorithm gives output result only relevant to the input uncertainty within the moving window. So instead of Formula (9.3), Formula (9.2) is confirmed to be a better solution for this linear regression problem.

9.4 Modelling Dependency Problem

However, the majority algorithms used today are progressive. Most practical problems are not even mathematical and analytical in nature, so that they may have no expressive solutions. Expressive algorithms are simply just not always avoidable in practice.

With known expressive counterpart, the progressive moving-window linear regression algorithm can serve as a model for studying progressive algorithms. For example:

- The progressive moving-window linear regression shows that the dependency problem of independence and interval arithmetic can manifest as dependency on the usage history of an algorithm. Because of its stability, variance arithmetic should be used generally in progressive algorithms.
- Figure ?? shows that the result tracking ratios of the progressive linear regression is exponentially distributed, while Figure ?? shows that the result tracking ratios of the expressive linear regression is Gaussian distributed only when the uncertainty deviation is characterized correctly, e.g., the result is Gaussian distributed for the "Changed" case but not for the "noisier" case. Thus, the exponentially distributed tracking ratios does not necessarily imply dependency problem.

10 Conclusion and Discussion

10.1 Summary

The starting point of variance arithmetic is the uncorrelated uncertainty assumption, which requires input data to have decent precision for each or small overall correlation among them, as shown in Figure 2, which quantifies the statistical requirements for input data to variance arithmetic. In addition, it requires that the systematic errors is not the major source of uncertainty, and all of its input data do not have confused identities.

Due to the uncorrelated uncertainty assumption and central limit theorem, the rounding errors of variance arithmetic are shown to be bounded by a Gaussian distribution with a truncated range. The rounding error distribution is extended to describe the uncertainty distribution in general, with the uncertainty deviation of a single precision value given by Formula (??), and the result uncertainty deviation of a function given by Formula (2.24) and its multi-dimension extensions such as Formula (??).

Formula (??) is shown to describe the general uncertainty deviation propagation in variance arithmetic. The average tracking ratios and the maximal bounding ratio using variance arithmetic are shown to be independent of input precision, and stable for the amount of calculations for a few very different applications. In contrast, both average tracking ratios and the maximal bounding ratio using interval arithmetic are shown to decrease exponentially with the amount of calculations in all tests. Such stability is the major reason why variance arithmetic is better than interval arithmetic in all tests done so far.

The statistical nature of variance arithmetic provides not only quantitative explanation for the dependency problem, but also solutions to the dependency problem, which is in form of either Taylor expansion or calibration. The treatment of dependency problem is another major advantage of variance arithmetic over interval arithmetic.

variance has a central role in variance arithmetic:

- Precision is regarded as information content of a uncertainty-bearing value, which is in par with information entropy in information theory. Because of this, precision needs to be preserved when the uncertainty-bearing value is multiplied or divided by a constant, which results in the scaling principle.
- variance arithmetic itself can be deduced from the scaling principle and the uncorrelated uncertainty assumption.
- The convergence property of the result deviation using Taylor expansion method is determined by input precisions, such as for inversions and square roots.

10.2 Efficiency of Precision Arithmetic

variance arithmetic tries to solve a different mathematical problem from conventional floating-point arithmetic. For example, to calculate the determinant of a matrix:

- Conventional floating-point arithmetic may use a Laplace method [12], namely, to randomly choose a row or a column, and then to sum up the products of each element within the chosen row or the column with the corresponding sub-determinant of the element. Each sub-determinant is calculated in the same fashion. Depending on the choices of the row or the column in each stage, there are many paths to calculate the determinant of a matrix. Because conventional

floating-point arithmetic has unbounded rounding errors, each path may give a different result, and the spread of all the results depends on the stability of the matrix and each sub-matrix [46]. In this perspective, by taking a random path and assuming to get the only correct result, conventional floating-point arithmetic can be viewed as a leap-of-faith approach.

- In contrast, variance arithmetic also needs to calculate the spread of the result due to rounding error or input uncertainties, so it effectively has to cover all paths of the calculation. For example, using Formula (??), variance arithmetic starts from each elements of the matrix, and treat it as a 1x1 sub-determinant, then grow it to all possible 2x2 sub-determinants containing it, etc, until reach the determinant of the matrix. Thus, variance arithmetic takes order-of-magnitude more time than a single leap-of-faith calculation.

However, it is wrong to conclude that variance arithmetic is less efficient than conventional floating-point arithmetic, because in most cases rounding errors and input uncertainty can not be ignored. Because conventional floating-point arithmetic can not contain uncertainty in its value, it has to use another value to specify uncertainty, such as an interval of $[min, max]$ or a common statistical pair $value \pm deviation$, which may brings the following drawbacks:

- The most common way to calculate result spread using conventional floating-point arithmetic is sampling [15] [12]. Assuming the matrix size is $N \times N$, and a minimal 3-point sampling is performed on each matrix element, then the spread calculation of matrix determinant requires N^6 leap-of-faith calculations, which is still a lot. In contrast, using Formula (??), variance arithmetic only need one calculation. Thus, conventional floating-point arithmetic may be less efficient than variance arithmetic in this context.
- During to unbounded rounding errors, a conventional floating-point value losses its precision gradually and silently, so that a interval or a statistical pair itself can become unknowingly invalid. At least, it is not clear at what precision the interval or the statistical pair specifies.

10.3 Choose a better algorithm

Because variance arithmetic tries to solve a different problem than conventional floating-point arithmetic, it has completely different criterion when choosing algorithms or implementations of algorithms. For example, for matrix inversion, because conventional floating-point arithmetic has unbounded rounding errors, it will choose certain flavour of LU-decomposition over Gaussian elimination and determinant division [12]. The result difference of LU-decomposition, Gaussian elimination and determinant division shows that conventional floating-point arithmetic has strong dependency problem, which has been a way of life when using conventional floating-point arithmetic, e.g., different algorithms or different implementation of the same algorithm are expected to give different results, of which a best algorithm or implementation is always chosen for each usage context [12], even though they may be mathematically equivalent. In contrast, rounding errors are bounded in both variance arithmetic and interval arithmetic [19], so they are no longer needed to be considered. When interval arithmetic reformat a numerical question as "Given each input to be an interval, what the output would be?", it effectively states that the results for most numerical questions to be solved should be *unique* to be either one or a few intervals that tightest bounds

the results, *regardless* of the algorithm to be used, *unless* dependency problem is introduced in the implementation of an algorithm. Same concept is true for variance arithmetic, which converges all input uncertainty distribution to *ubiquitously Gaussian* at the outputs, and which further *quantifies* the source of the dependency problem. Using variance arithmetic instead of conventional floating-point arithmetic, the focus has shifted from minimizing rounding errors to minimizing dependency problem. Of the three algorithms for matrix inversion, both LU-decomposition and Gaussian elimination are progressive, which means that each input may appear multiple times in different branch at different time, whose dependency problem is difficult to quantified. On the other hand, a determinant of a $N \times N$ matrix can be treated as a N -order polynomial with N^2 variables, to be readily for the Taylor expansion, which results in Formula (??), so that the determinant division method is chosen in this paper for matrix inversion. For the same reason, in the moving-window linear regression, the worse method in conventional floating-point arithmetic, Formula (9.2), becomes the better method in variance arithmetic, and vice versa.

Due to the requirement of minimizing dependence problem, variance arithmetic has much less operational freedom than conventional arithmetic and may require extensive symbolic calculations, following practices in affine arithmetic [42]. Also, the comparison relation in conventional arithmetic needs to be re-evaluated in variance arithmetic, which brings about another reason for different algorithm selection.

10.4 Improving Precision Arithmetic

Figure 2 uses a cut-off for the test of the uncorrelated uncertainty assumption among two uncertainty-bearing values. A better approach is to associate the amount of the dependence problem with the amount of correlation between the uncertainties of the two values.

There are actually three different ways to round up $(2S + 1)4R 2^E$:

1. always round up $(2S + 1)4R 2^E$ to $(S + 1) - R 2^{E+1}$;
2. always round up $(2S + 1)4R 2^E$ to $S + R 2^{E+1}$;
3. randomly round up $(2S + 1)4R 2^E$ to either $(S + 1) - R 2^{E+1}$ or $S + R 2^{E+1}$.

The first method results in slightly slower loss of significand than the second method, while the third method changes variance arithmetic from deterministic to stochastic. Because no empirical difference has been detected among these three different rounding up methods, the first method is chosen in this paper. Further study is required to distinguish the different rounding up methods.

The objectives of variance arithmetic need to be studied further. For example, Formula (??) has rejected the effect of uncertainty on the expected value by incorporating the value shift due to uncertainty as increase of variance, such as in the case of calculating $f(x) = x^2$. The effect of such asymmetrical broadening is unclear.

The number of bits to be calculated inside uncertainty also needs to be studied further. For example, when limited bits are calculated inside uncertainty, adding insignificant higher order term of a Taylor expansion may decrease the value error while increasing the uncertainty deviation, which may call for an optimal bits to be calculated inside uncertainty for the truncation rule.

Because variance arithmetic is based on generic concepts, it is targeted to be a generic arithmetic for both uncertainty-tracking and uncertainty-bounding. However, it seems a worthwhile alternative to interval arithmetic and the de facto independence

arithmetic. Before applying it generally, variance arithmetic still needs more ground-work and testing. It should be tested further in other problems such as improper integrations, solutions to linear equations, and solutions to differential equations.

10.5 Acknowledgements

As an independent researcher, the author of this paper feels indebted to encouragements and valuable discussions with Dr. Zhong Zhong from Brookhaven National Laboratory, Prof. Hui Cao from Yale University, Dr. Anthony Begley from *Physics Reviews B*, the organizers of *AMCS 2005*, with Prof. Hamid R. Arabnia from University of Georgia in particular, and the organizers of *NKS Mathematica Forum 2007*, with Dr. Stephen Wolfram in particular. Finally, the author of this paper is very grateful for the editors and reviewers of *Reliable Computing* for their tremendous help in shaping this unusual paper from unusual source, with managing editor, Prof. Rolph Baker Kearfott in particular.

References

- [1] Sylvain Ehrenfeld and Sebastian B. Littauer. *Introduction to Statistical Methods*. McGraw-Hill, 1965.
- [2] John R. Taylor. *Introduction to Error Analysis: The Study of Output Precisions in Physical Measurements*. University Science Books, 1997.
- [3] Jurgen Bortfeldt, editor. *Fundamental Constants in Physics and Chemistry*. Springer, 1992.
- [4] Michael J. Evans and Jeffrey S. Rosenthal. *Probability and Statistics: The Science of Uncertainty*. W. H. Freeman, 2003.
- [5] Paul Horowitz and Hill Winfield. *Art of Electronics*. Cambridge Univ Press, 1995.
- [6] Fixed-point arithmetic. http://en.wikipedia.org/wiki/Fixed-point_arithmetic, 2011. wikipedia, the free encyclopedia.
- [7] Arbitrary-precision arithmetic. http://en.wikipedia.org/wiki/Arbitrary-precision_arithmetic, 2011. wikipedia, the free encyclopedia.
- [8] John P Hayes. *Computer Architecture*. McGraw-Hill, 1988.
- [9] David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, March 1991.
- [10] Institute of Electrical and Electronics Engineers. *ANSI/IEEE 754-2008 Standard for Binary Floating-Point Arithmetic*, 2008.
- [11] U. Kulish and W.M. Miranker. The arithmetic of digital computers: A new approach. *SIAM Rev.*, 28(1), 1986.
- [12] William H. Press, Saul A Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992.
- [13] Oliver Aberth. *Precise Numerical Methods Using C++*. Academic Press, 1998.
- [14] Gregory L. Baker and Jerry P. Gollub. *Chaotic Dynamics: An Introduction*. Cambridge University Press, 1990.
- [15] J. Vignes. A stochastic arithmetic for reliable scientific computation. *Mathematics and Computers in Simulation*, 35:233–261, 1993.

- [16] B. Liu and T. Kaneko. Error analysis of digital filters realized with floating-point arithmetic. *Proc. IEEE*, 57:p1735–1747, 1969.
- [17] B. D. Rao. Floating-point arithmetic and digital filters. *IEEE, Transactions on Signal Processing*, 40:85–95, 1992.
- [18] R.E. Moore. *Interval Analysis*. Prentice Hall, 1966.
- [19] W. Kramer. A prior worst case error bounds for floating-point computations. *IEEE Trans. Computers*, 47:750–756, 1998.
- [20] G. Alefeld and G. Mayer. Interval analysis: Theory and applications. *Journal of Computational and Applied Mathematics*, 121:421–464, 2000.
- [21] W. Kramer. Generalized intervals and the dependency problem. *Proceedings in Applied Mathematics and Mechanics*, 6:685–686, 2006.
- [22] A. Neumaier S.M. Rump S.P. Shary B. Kearfott, M. T. Nakao and P. Van Hentenryck. Standardized notation in interval analysis. *Computational Technologies*, 15:7–13, 2010.
- [23] W. T. Tucker and S. Ferson. *Probability bounds analysis in environmental risk assessments*. Applied Biomathematics, 100 North Country Road, Setauket, New York 11733, 2003.
- [24] J. Stolfi and L. H. de Figueiredo. An introduction to affine arithmetic. *TEMA Tend. Mat. Apl. Comput.*, 4:297–312, 2003.
- [25] R. Alt and J.-L. Lamotte. Some experiments on the evaluation of functional ranges using a random interval arithmetic. *Mathematics and Computers in Simulation*, 56:17–34, 2001.
- [26] J. Stolfi and L. H. de Figueiredo. *Self-validated numerical methods and applications*. <ftp://ftp.tecgraf.puc-rio.br/pub/lhf/doc/cbm97.ps.gz>, 1997.
- [27] Propagation of uncertainty. http://en.wikipedia.org/wiki/Propagation_of_uncertainty, 2011. wikipedia, the free encyclopedia.
- [28] S. Ferson H. M. Regan and D. Berleant. Equivalence of methods for uncertainty propagation of real-valued random variables. *International Journal of Approximate Reasoning*, 36:1–30, 2004.
- [29] C. P. Robert. *Monte Carlo Statistical Methods*. Springer, 2001.
- [30] Monte carlo method. http://en.wikipedia.org/wiki/Monte_Carlo_method, 2011. wikipedia, the free encyclopedia.
- [31] C. L. Smith. Uncertainty propagation using taylor series expansion and a spreadsheet. *Journal of the Idaho Academy of Science*, 30-2:93–105, 1994.
- [32] Significance arithmetic. http://en.wikipedia.org/wiki/Significance_arithmetic, 2011. wikipedia, the free encyclopedia.
- [33] M. Goldstein. Significance arithmetic on a digital computer. *Communications of the ACM*, 6:111–117, 1963.
- [34] R. L. Ashenurst and N. Metropolis. Unnormalized floating-point arithmetic. *Journal of the ACM*, 6:415–428, 1959.
- [35] G. Spaletta M. Sofroniou. Precise numerical computation. *The Journal of Logic and Algebraic Programming*, 65:113–134, 2005.

- [36] C. Denis N. S. Scott, F. Jezequel and J. M. Chesneaux. Numerical 'health' check for scientific codes: the cadna approach. *Computer Physics Communications*, 176(8):501–527, 2007.
- [37] C. P. Wang. Error estimation of floating-point calculations by a new floating-point type that tracks the errors. In H. R. Arabnia and I. A. Ajwa, editors, *Proceedings of the 2005 International Conference on Algorithmic Mathematics and Computer Science, AMCS 2005*, pages 84–92, 2005.
- [38] C. P. Wang. A New Uncertainty-Bearing Floating-Point Arithmetic. *Reliable Computing*, 16:308–361, 2012
- [39] A. Feldstein and R. Goodman. Convergence estimates for the distribution of trailing digits. *Journal of the ACM*, 23:287–297, 1976.
- [40] Double factorial. <http://mathworld.wolfram.com/DoubleFactorial.html>, 2014. Wolfram MathWorld.
- [41] Jagdish K. Patel; Campbell B Read Handbook of the normal distribution (2nd ed.). CRC Press. ISBN 0-8247-9342-0.
- [42] C. Pennachin, M. Looks, and João A. de Vasconcelos. Robust symbolic regression with affine arithmetic. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation (2010)*, pages 917–924, 2010.
- [43] N. Beaudoin and S. S. Beauchemin. A new numerical fourier transform in d-dimensions. *IEEE Transactions on Signal Processing*, 51-5:1422–1430, 2003.
- [44] Digital signal processor. http://en.wikipedia.org/wiki/Digital_signal_processor, 2011. wikipedia, the free encyclopedia.
- [45] J. Hefferon. Linear algebra. <http://joshua.smcvt.edu/linearalgebra/>, 2011.
- [46] Condition number. http://en.wikipedia.org/wiki/Condition_number, 2011. wikipedia, the free encyclopedia.
- [47] Hilbert matrix. http://en.wikipedia.org/wiki/Hilbert_matrix, 2011. wikipedia, the free encyclopedia.
- [48] Big o notation. http://en.wikipedia.org/wiki/Big_Oh_notation, 2011. wikipedia, the free encyclopedia.