Journal of Universal Computer Science, vol. 1, no. 7 (1995), 469-483
submitted: 15/12/94, accepted: 26/6/95, appeared: 28/7/95©Springer Pub. Co.

Round-off error propagation in the solution of the heat equation by finite differences

Fabienne Jézéquel
(Université Pierre et Marie Curie, France
Fabienne.Jezequel@masi.ibp.fr)

Abstract: The effect of round-off errors on the numerical solution of the heat equation by finite differences can be theoretically determined by computing the mean error at each time step. The floating point error propagation is then theoretically time linear. The experimental simulations agree with this result for the towards zero rounding arithmetic. However the results are not so good for the rounding to the nearest artihmetic. The theoretical formulas provide an approximation of the experimental round-off errors. In these formulas the mean value of the assignment operator is used, and consequently, their reliability depends on the arithmetic used.

Key Words: Floating point arithmetic, numerical error propagation, partial differential equations, finite difference methods
Category: G.1.8

1 Introduction

In the computational solution of partial differential equations, two types of errors are generated : the method error due to approximations inherent in the numerical method and the round-off error due to the floating point arithmetic of the computer used. This paper presents an analysis of the round-off error propagation in the solution of the heat equation by finite differences. For each point of the mesh, the solution is approximated by a scalar product with three terms. Therefore previous studies concerning round-off errors in arithmetical operations and in scalar products are presented. This analysis has been carried out for the towards zero rounding arithmetic and for the rounding to the nearest arithmetic, these two rounding modes respecting the 754-IEEE standard. Different cases have been considered depending on whether initial data and finite difference scheme coefficients are exactly represented or not in the computer. To conclude, the main results obtained are finally represented.

2 Previous results concerning round-off errors

2.1 Assignment error

Let x be a real number and X its floating point representation. The relative assignment error on X is $\alpha = \frac{(X-x)}{X}$. Let **P** be the set of all the possible relative assignment errors α. The mean value $\bar{\alpha}$ and the standard deviation σ^2 of **P** can be computed according to the rounding mode, the base and the number of bits in the mantissa in the floating point representation [see Alt 76, Alt 78, Hamming 70, Knuth 69, La Porte, Vignes, 74a and Vignes 93]. Let b be the base (usually b is 2 or 16) and p the number of digits in the mantissa in the standard floating point representation,

for the towards zero rounding arithmetic :

$$\bar{\alpha} = b^{-p} \frac{(1-b)}{2 \log b}$$

$$\sigma^2 = b^{-2p} [\frac{(b^2-1)}{6 \log b} - \frac{(b-1)^2}{(2 \log b)^2}]$$

and for the rounding to the nearest arithmetic :

$$\bar{\alpha} = 0$$

$$\sigma^2 = b^{-2p} \frac{(b^2-1)}{24 \log b}.$$

2.2 Error due to arithmetical operators

Let $+$, $-$, $\times$, $/$ be the exact operators on real numbers and $\oplus$, $\ominus$, $\otimes$, $\oslash$ the corresponding floating point operators (addition, subtraction, multiplication and division) on $\mathbf{F}$, which is the set of all the values representable in the machine. The following formulas have been obtained by considering only first-order approximations in b^{-p} :

Let x and y be real numbers, X and Y their representations in $\mathbf{F}$:

$X \approx x(1+\alpha)$ and $Y \approx y(1+\beta)$

$$X \oplus Y \approx x + y + \alpha x + \beta y + \mu(x+y)$$

$$X \ominus Y \approx x - y + \alpha x - \beta y + \mu(x-y)$$

$$X \otimes Y \approx (x \times y)(1 + \alpha + \beta + \mu')$$

$$X \oslash Y \approx (x/y)(1 + \alpha - \beta + \mu')$$

with α, β, μ , μ' being elements of $\mathbf{P}$.

2.3 Error in the computation of scalar products

Let us consider a scalar product $r = \sum_{i=1}^{n} x_i y_i$, with x_i and y_i real numbers. If it is computed using this cumulative method :

$R := 0$; FOR $I = 1$ TO N DO $R := R \oplus X[I] \otimes Y[I]$,

the absolute error ρ on the exact scalar product r is defined by : $\rho = R - r$.

If $X_i = x_i(1+\lambda_i)$ and $Y_i = y_i(1+\mu_i)$ the error ρ can be estimated by the following formula :

$$\begin{aligned} \rho \approx & \sum_{i=1}^{n} x_i y_i (\lambda_i + \mu_i + \beta_i) \\ & + \alpha_1 (x_1 y_1 + x_2 y_2) \\ & + \alpha_2 (x_1 y_1 + x_2 y_2 + x_3 y_3) \\ & + \ldots \\ & + \alpha_{n-1} (x_1 y_1 + x_2 y_2 + \ldots + x_n y_n) \end{aligned}$$

with α_i, β_i, λ_i, μ_i being elements of $\mathbf{P}$.

We assume that the errors α_i, β_i, λ_i, μ_i are independent and have the same mean value $\bar{\alpha}$ and the same standard deviation σ^2. Under these hypotheses, the mean values of ρ and ρ^2 are given by :

$$\bar{\rho} = \bar{\alpha} r \frac{(n^2+7n-2)}{2n}$$

$$\bar{\rho^2} = (\bar{\alpha})^2[(\frac{3n^3+41n^2+134n-72}{12n})r^2+(\frac{(n-2)(n^2+3n-6)}{12n})s^2]+\sigma^2[(\frac{n+1}{3})r^2+(\frac{n^2+19n-6}{6n})s^2]$$

with $s^2 = \sum_{i=1}^{n}(x_i y_i)^2$ [see Alt 78 and La Porte, Vignes 74b].

3 Error propagation in the solution of the heat equation

3.1 Finite difference scheme

The one-dimensional heat equation describes the heat propagation in a linear bar. Let $U(x,t)$ be the temperature on this bar at point x and time t. The heat equation in $[a,b] \times [t_0, +\infty[$ is described by the following system :

$$\frac{\partial U(x,t)}{\partial t} - K\frac{\partial^2 U(x,t)}{\partial x^2} = 0 \text{ , with } K > 0$$
$$\forall t \geq t_0 \text{ , } U(a,t) = U_a(t) \text{ and } U(b,t) = U_b(t)$$
$$\forall x \in [a,b] \text{ , } U(x,t_0) = U^0(x)$$

The constant K represents the material thermic diffusivity.

The domain is discretized with space step Δx and time step Δt :

$$x_i = a + i\Delta x, \ i = 0, 1, ...n$$
$$x_0 = a \text{ and } x_n = b$$
$$t_j = t_0 + j\Delta t, j = 0, 1, ...$$

Let U_i^j be the solution at point x_i and time t_j. The explicit finite difference method is used :

$$\frac{U_i^{j+1} - U_i^j}{\Delta t} = K(\frac{U_{i-1}^j - 2U_i^j + U_{i+1}^j}{(\Delta x)^2}) \quad \text{for } i = 1, ..., n-1 \text{ , } j = 0, 1, ...$$

then: $U_i^{j+1} = \frac{K\Delta t}{(\Delta x)^2}U_{i-1}^j + (1 - 2\frac{K\Delta t}{(\Delta x)^2})U_i^j + \frac{K\Delta t}{(\Delta x)^2}U_{i+1}^j$

To ensure the stability of this scheme, the relation $\frac{K\Delta t}{(\Delta x)^2} < \frac{1}{2}$ must be satisfied.

If $c_1 = \frac{K\Delta t}{(\Delta x)^2}$ and $c_2 = 1 - 2c_1$, the finite difference scheme is :

$$U_i^{j+1} = c_1 U_{i-1}^j + c_2 U_i^j + c_1 U_{i+1}^j \quad \text{for } i = 1, ..., n-1 \text{ , } j = 0, 1, ...$$

3.2 Theoretical round-off error

3.2.1 Relative round-off error

To estimate the round-off error in the computation of U_i^j with the finite difference scheme previously proposed, several notations are necessary. Let U_i^j, c_1, c_2 be the algebraic values and $\tilde{U_i^j}$, $\tilde{c_1}$, $\tilde{c_2}$ the computed values.

Then for $i = 1, ..., n-1$, $j = 0, 1, ...,$

$$\tilde{U}_i^{j+1} = ((\tilde{c_1} \otimes \tilde{U}_{i-1}^j) \oplus (\tilde{c_2} \otimes \tilde{U_i^j})) \oplus (\tilde{c_1} \otimes \tilde{U}_{i+1}^j),$$

this formula being neither commutative nor associative.

Let μ_i^j be the relative error on U_i^j due to the cumulation of assignment errors and round-off errors generated in previous iterations : $\tilde{U_i^j} = U_i^j(1 + \mu_i^j)$.

μ_i^0 merely represents the assignment error on U_i^0 and the mean value $\bar{\mu}^0$ is equal to the mean value of the assignment operator $\bar{\alpha}$.

Let λ_1 and λ_2 be the relative errors on c_1 and c_2 : $\tilde{c_1} = c_1\ (1 + \lambda_1)$ and $\tilde{c_2} = c_2\ (1 + \lambda_2)$.

If c_1 and c_2 are not results of computations or are computed in infinite precision, then λ_1 and λ_2 are merely assignment errors.

U_i^{j+1} is computed by a scalar product with three terms. Therefore the formula providing the round-off error in the computation of scalar products can be applied :

$$\begin{aligned}
\mu_i^{j+1} = \tfrac{1}{U_i^{j+1}}\ \Big[\ & c_1 U_{i-1}^j(\lambda_1 + \mu_{i-1}^j + \beta_1) \\
& + c_2 U_i^j(\lambda_2 + \mu_i^j + \beta_2) \\
& + c_1 U_{i+1}^j(\lambda_1 + \mu_{i+1}^j + \beta_3) \\
& + \alpha_1(c_1 U_{i-1}^j + c_2 U_i^j) \\
& + \alpha_2(c_1 U_{i-1}^j + c_2 U_i^j + c_1 U_{i+1}^j)\ \Big]
\end{aligned}$$

with α_i, β_i being elements of $\mathbf{P}$.

The errors α_1 and α_2 are due to additions $\oplus$, β_1, β_2 and β_3 to multiplications $\otimes$. Assignment errors α_i, β_i are assumed to be independent and consequently :

$$\bar{\alpha_i} = \bar{\beta_i} = \bar{\alpha}$$

$$\overline{\alpha_i{}^2} = \overline{\beta_i{}^2} = (\bar{\alpha})^2 + \sigma^2$$

$$\overline{\alpha_i \beta_k} = (\bar{\alpha})^2 \text{ etc ...}$$

3.2.2 First moment

As assignment errors are assumed to be independent and have the same mean value $\bar{\alpha}$, the first moment is, for $j = 0, 1, \dots$ and $i = 1, \dots, n-1$:

$$\overline{\mu_i^{j+1}} = \frac{c_1 U_{i-1}^j}{U_i^{j+1}} \left(\overline{\mu_{i-1}^j} + 3\bar{\alpha} + \lambda_1\right) + \frac{c_2 U_i^j}{U_i^{j+1}} \left(\overline{\mu_i^j} + 3\bar{\alpha} + \lambda_2\right) + \frac{c_1 U_{i+1}^j}{U_i^{j+1}} \left(\overline{\mu_{i+1}^j} + 2\bar{\alpha} + \lambda_1\right).$$

The space step and the time step are assumed to be low enough to allow the following approximation :

$$\forall j \geq 0, \quad \forall i = 1, \dots, n-1, \; \frac{U_{i-1}^j}{U_i^{j+1}} \approx \frac{U_i^j}{U_i^{j+1}} \approx \frac{U_{i+1}^j}{U_i^{j+1}} \approx 1.$$

Therefore at a fixed iteration, all relative errors μ_i^j have the same mean value :

$$\forall i = 1, \dots n-1, \; \overline{\mu_i^j} = \overline{\mu^j}.$$

then

$$\forall j \geq 0, \; \overline{\mu^{j+1}} = \overline{\mu^j} + (3 - c_1)\bar{\alpha} + 2c_1\lambda_1 + c_2\lambda_2$$

therefore :

$$\forall j \geq 0, \; \overline{\mu^j} = \overline{\mu^0} + ((3 - c_1)\bar{\alpha} + 2c_1\lambda_1 + c_2\lambda_2)j.$$

The round-off error propagation in the solution of the heat equation by finite differences is theoretically time linear. The general formula above is simplified if the coefficients c_1 and c_2 and the initial data U_i^0 are exactly represented. For the rounding to the nearest arithmetic, the mean value of the assignment errors, $\bar{\alpha}$ is theoretically zero. In this case, if the coefficients c_1 and c_2 and the initial data U_i^0 are exactly represented, $\lambda_1 = \lambda_2 = \overline{\mu^0} = 0$, and the first moment remains theoretically zero. Thus it is necessary to estimate the second moment.

3.2.3 Second moment

The estimation of the second moment has been carried out for the rounding to the nearest arithmetic under the following assumptions :

c_1 and c_2 are exactly represented : $\lambda_1 = \lambda_2 = 0$,

initial data are exactly represented : $\forall i = 1, \dots, n-1, \; \mu_i^0 = 0$.

As for the estimation of the first moment, it is assumed that :

$$\forall j \geq 1, \quad \forall i = 1, \dots, n-1 \;, \quad \frac{U_{i-1}^{j-1}}{U_i^j} \approx \frac{U_i^{j-1}}{U_i^j} \approx \frac{U_{i+1}^{j-1}}{U_i^j} \approx 1.$$

The estimation of $\overline{(\mu_i^{j+1})^2}$ induces the emergence of terms $\overline{\gamma\mu_i^j}$, γ being a relative assignment error.

It is assumed that $\overline{\gamma\mu_{i-1}^j} = \overline{\gamma\mu_{i+1}^j} = \overline{\gamma\mu_i^j}$.

Therefore, $\overline{(\mu_i^{j+1})^2} = (2c_1^2 + c_2^2)\ \overline{(\mu_i^j)^2} + (7c_1^2 + 3c_2)\ \sigma^2$
$$+2\big((1-c_1)\overline{\alpha_1\mu_i^j} + \overline{\alpha_2\mu_i^j} + c_1\overline{\beta_1\mu_i^j} + c_2\overline{\beta_2\mu_i^j} + c_1\overline{\beta_3\mu_i^j}\big).$$

As $\forall i = 1, ..., n-1, \quad \mu_i^0 = 0, \quad \overline{(\mu^1)^2} = (7c_1^2 + 3c_2)\ \sigma^2.$

$\overline{\gamma\mu_i^j}$ are estimated, γ being a relative assignment error :

$$\overline{\alpha_1\mu_i^j} = (1-c_1)\sigma^2 + \overline{\alpha_1\mu_i^{j-1}}$$
$$\overline{\alpha_2\mu_i^j} = \sigma^2 + \overline{\alpha_2\mu_i^{j-1}}$$
$$\overline{\beta_1\mu_i^j} = c_1\sigma^2 + \overline{\beta_1\mu_i^{j-1}}$$
$$\overline{\beta_2\mu_i^j} = c_2\sigma^2 + \overline{\beta_2\mu_i^{j-1}}$$
$$\overline{\beta_3\mu_i^j} = c_1\sigma^2 + \overline{\beta_3\mu_i^{j-1}}$$

As $\overline{\gamma\mu_i^0} = 0$, finally :

$$\overline{\alpha_1\mu_i^j} = (1-c_1)j\sigma^2$$
$$\overline{\alpha_2\mu_i^j} = j\sigma^2$$
$$\overline{\beta_1\mu_i^j} = \overline{\beta_3\mu_i^j} = c_1 j\sigma^2$$
$$\overline{\beta_2\mu_i^j} = c_2 j\sigma^2.$$

Therefore : $\forall j \geq 1, \ \overline{(\mu_i^{j+1})^2} = (2c_1^2 + c_2^2)\ \overline{(\mu_i^j)^2} + (1+2j)\ (7c_1^2 + 3c_2)\ \sigma^2.$

Then
$$\overline{(\mu^0)^2} = 0$$
and $\forall j \geq 1, \ \overline{(\mu^j)^2} = (7c_1^2 + 3c_2)\ \sigma^2\ \sum_{k=0}^{j-1} (1+2k)\ (2c_1^2 + c_2^2)^{j-k-1}.$

The evolution of the second moment is thus of degree 2. This result remains coherent with the linear evolution of the first moment.

3.3 Experimental round-off error

3.4 First moment

The experimental first moment $\overline{\mu^j}$ is computed according to the following formula :

$$\overline{\mu^j} = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{\tilde{U_i^j} - U_i^j}{\tilde{U_i^j}}\right)$$

where U_i^j represents the algebraic value and $\tilde{U}_i^j$ the computed value. The number of significant digits lost in computations does not depend on the precision of the floating point arithmetic [see Chesneaux 88 and Chesneaux 90]. Therefore U_i^j, theoretically computed in infinite precision, can be computed in double precision. Then $\tilde{U}_i^j$ is the result of the same computation carried out in single precision [see Hull, Swenson 66].

The theoretical expression of the first moment has been validated for the towards zero rounding arithmetic and the rounding to the nearest arithmetic, on a computer using base 2 with $p = 24$ and respecting the 754-IEEE standard. Four cases can occur depending on whether the coefficients c_1 et c_2 and the initial data U_i^0 are exactly represented or not. Each case has been studied for the rounding to the nearest arithmetic, where the mean value of the assignment errors $\bar{\alpha}$ is zero, and for the towards zero rounding arithmetic, where $\bar{\alpha}$ is not zero. The number of space steps n is set to 100, the number of time steps is set to 1000.

1st case :

If the coefficients c_1, c_2 and the initial data U_i^0 are both not exactly represented, the first moment is :

$$\forall j \geq 0, \quad \overline{\mu^j} = \overline{\mu^0} + ((3 - c_1)\bar{\alpha} + 2c_1\lambda_1 + c_2\lambda_2)j$$

The experimental moment is time linear as well. However the theoretical moment is in absolute value slightly greater than the experimental moment. This difference may be due to an overvaluation of the theoretical mean value $\bar{\alpha}$.
Results concerning the following example are presented in the appendix :

$$c_1 = \tfrac{1}{6}, \quad c_2 = \tfrac{2}{3}$$

$$\forall i = 0, 1, \cdots, n, \quad U_i^0 = \sin(\tfrac{i\pi}{n}) + \log 2$$

2nd case :

If the coefficients c_1 and c_2 are exactly represented, but the initial data U_i^0 are not exactly represented, the first moment is :

$$\forall j \geq 0, \quad \overline{\mu^j} = \overline{\mu^0} + (3 - c_1)\bar{\alpha}j$$

In the towards zero rounding arithmetic, the theoretical moment and the experimental one are both linear. The theoretical moment overestimates again slightly in absolute value the experimental moment.
In the rounding to the nearest arithmetic, as the mean value $\bar{\alpha}$ is zero, the first moment remains theoretically equal to the mean error on data $\overline{\mu^0}$. In opposition to the theoretical moment, the experimental moment is not constant. However its order of magnitude (10^{-7}) is very satisfying for single precision results.
Graphical results for the following example are presented in the appendix :

$$c_1 = \tfrac{3}{16}, \quad c_2 = \tfrac{5}{8}$$

$$\forall i = 0, 1, \cdots, n, \quad U_i^0 = \sin(\tfrac{i\pi}{n}) + \log 2$$

3rd case :

If the initial data U_i^0 are exactly represented, but the coefficients c_1 and c_2 are not exactly represented, the first moment is :

$$\forall j \geq 0, \quad \overline{\mu^j} = ((3-c_1)\bar{\alpha} + 2c_1\lambda_1 + c_2\lambda_2)j$$

The choice of initial data which are exactly represented is more problematic than the choice of exactly represented coefficients. For instance, if initial data are of the form :

$$\forall i = 0, 1, \cdots, n, \; U_i^0 = i2^r,$$

with r being a relative integer, the scheme does not perform evolutions in time :

$$\forall j \geq 0, \; U_i^j = U_i^0.$$

The following example is presented in the appendix :

$$c_1 = \tfrac{1}{6}, \quad c_2 = \tfrac{2}{3}$$

$$\forall i = 0, 1, \cdots, n, \; \begin{array}{l} \text{if } i \text{ is odd, } U_i^0 = i/16 \\ \text{if } i \text{ is even,} U_i^0 = i \end{array}$$

The theoretical moment is linear and remains greater than the experimental moment in absolute value. In the towards zero rounding arithmetic, the experimental moment remains linear for all time intervals considered. In the rounding to the nearest arithmetic, the experimental moment is not perfectly linear (see graphical results in the appendix).

4th case :

If the initial data U_i^0 and the coefficients c_1 and c_2 are exactly represented, the first moment is :

$$\forall j \geq 0, \quad \overline{\mu_j} = (3-c_1)\bar{\alpha}j$$

In this case, the experimental moment is compared with the theoretical moment only in the towards zero rounding arithmetic, because in the rounding to the nearest arithmetic the mean value $\bar{\alpha}$ is theoretically zero.

In the appendix, the following example is presented :

$$c_1 = \tfrac{3}{16}, \quad c_2 = \tfrac{5}{8}$$

$$\forall i = 0, 1, \cdots, n, \; \begin{array}{l} \text{if } i \text{ is odd, } U_i^0 = i/16 \\ \text{if } i \text{ is even,} U_i^0 = i \end{array}$$

The experimental moment, as well as the theoretical one, is linear. The theoretical moment remains slightly greater than the experimental one in absolute value because of the overvaluation of the mean value of the assignment errors $\bar{\alpha}$.

3.4.1 Second moment

The second moment, $\overline{(\mu^j)^2}$, can be experimentally computed according to the following formula :

$$\overline{(\mu^j)^2} = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{\tilde{U}_i^j - U_i^j}{\tilde{U}_i^j}\right)^2$$

where U_i^j is the algebraic value and $\tilde{U}_i^j$ the computed one. As for the first moment, U_i^j, theoretically computed in infinite precision, is computed in double precision and $\tilde{U}_i^j$ is computed in single precision.

The second moment has been computed using the rounding to the nearest arithmetic, when the initial data U_i^0 and the coefficients c_1 and c_2 are exactly represented. In the appendix, the evolution of the ratio of the experimental moment by the theoretical one $\overline{(\mu^j)^2}^{\frac{1}{2}}{}_{exp}/\overline{(\mu^j)^2}^{\frac{1}{2}}{}_{theo}$ is presented. The example considered is the same as for the study of the first moment, when both initial data and coefficients are exactly represented.

4 Conclusion

In the towards zero rounding arithmetic, the round-off error generated in the solution of the heat equation is correctly modelled for all finite difference scheme coefficients and initial data. The round-off error propagation is then time linear. The theoretical error depends strongly on the mean value of the assignment errors and, while providing the order of magnitude of the experimental error, overestimates it slightly.

In the rounding to the nearest arithmetic, the round-off error generated is always smaller than in the towards zero rounding arithmetic. In the case where neither the coefficients nor the initial data are exactly represented, the round-off error is linear and is correctly described by the theoretical formula. However if the finite difference scheme coefficients or the initial data are exactly represented, theoretical formulas are not verified by the experimental study. The round-off error modelling is rather difficult in the rounding to the nearest arithmetic, where the mean value of the assignment error, which is theoretically zero, is never practically zero.

Theoretical formulas are much more robust in the towards zero rounding arithmetic than in the rounding to the nearest arithmetic. However from this study it seems obvious that the round-off error generated in the solution of the heat equation is usually linear.

Appendix : graphical results

1st case : $c_1 = \frac{1}{6},\ c_2 = \frac{2}{3},\ n = 100,\ \forall i = 0, 1, \cdots, n,\ U_i^0 = \sin(\frac{i\pi}{n}) + \log 2$

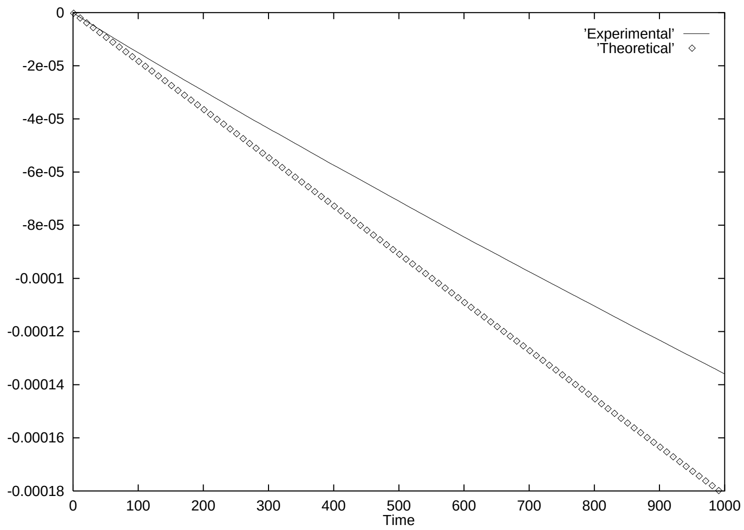

Figure 1: First moment, towards zero rounding arithmetic

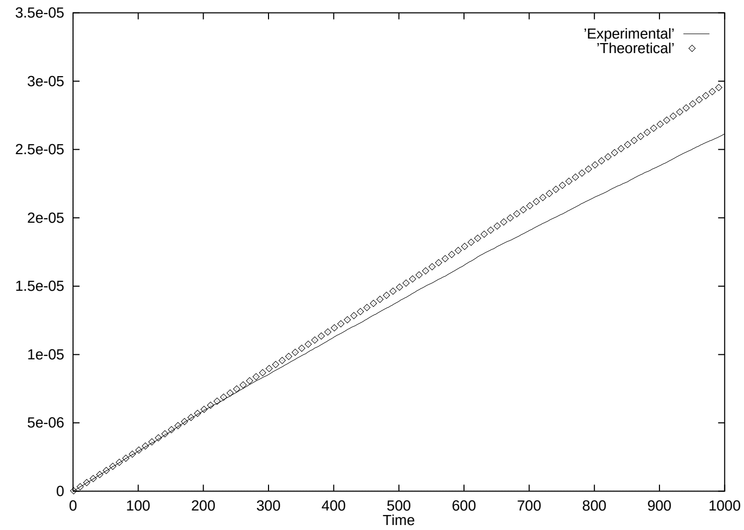

Figure 2: First moment, rounding to the nearest arithmetic

2nd case : $c_1 = \frac{3}{16},\ c_2 = \frac{5}{8},\ n = 100,\ \forall i = 0, 1, \cdots, n,\ U_i^0 = \sin(\frac{i\pi}{n}) + \log 2$

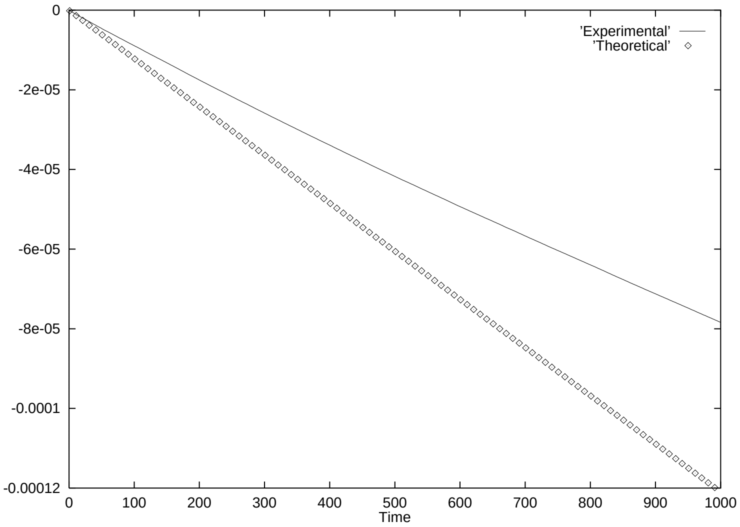

Figure 3: First moment, towards zero rounding arithmetic

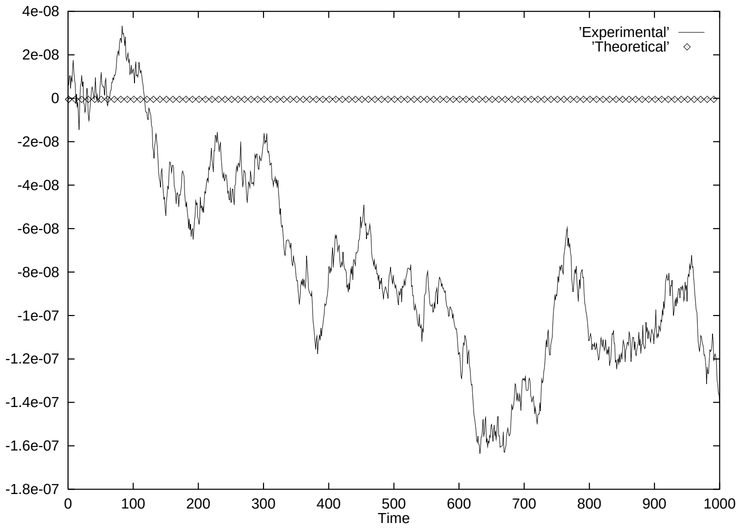

Figure 4: First moment, rounding to the nearest arithmetic

3rd case : $c_1 = \frac{1}{6},\ \ c_2 = \frac{2}{3},\ \ n = 100,\ \forall i = 0, 1, \cdots, n$, if i is odd, $U_i^0 = i/16$
if i is even, $U_i^0 = i$

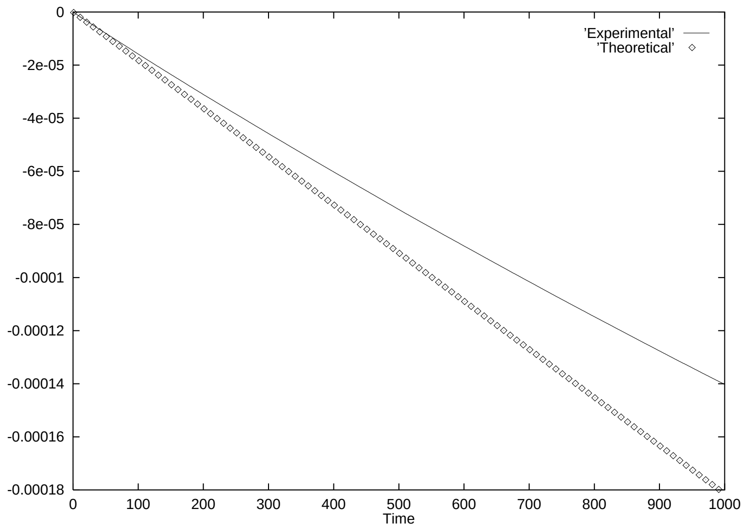

Figure 5: First moment, towards zero rounding arithmetic

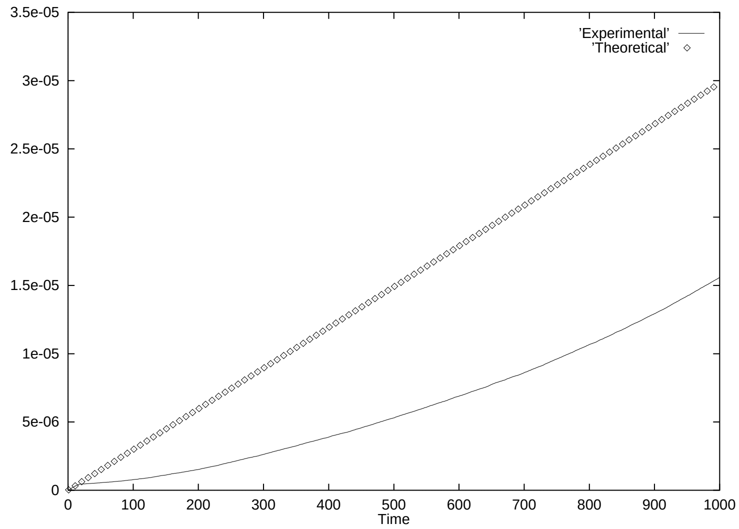

Figure 6: First moment, rounding to the nearest arithmetic

4th case : $c_1 = \frac{3}{16}$, $c_2 = \frac{5}{8}$, $n = 100$, $\forall i = 0, 1, \cdots, n$, if i is odd, $U_i^0 = i/16$
if i is even,$U_i^0 = i$

towards zero rounding arithmetic

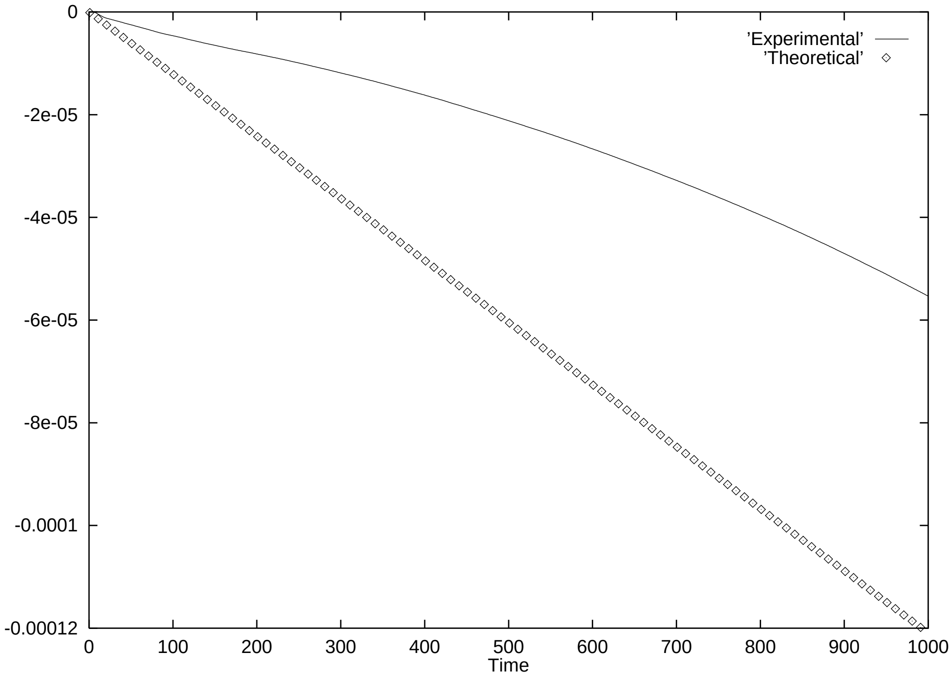

Figure 7: First moment

4th case : $c_1 = \frac{3}{16}$, $c_2 = \frac{5}{8}$, $n = 100$, $\forall i = 0, 1, \cdots, n$, if i is odd, $U_i^0 = i/16$
if i is even, $U_i^0 = i$

rounding to the nearest arithmetic

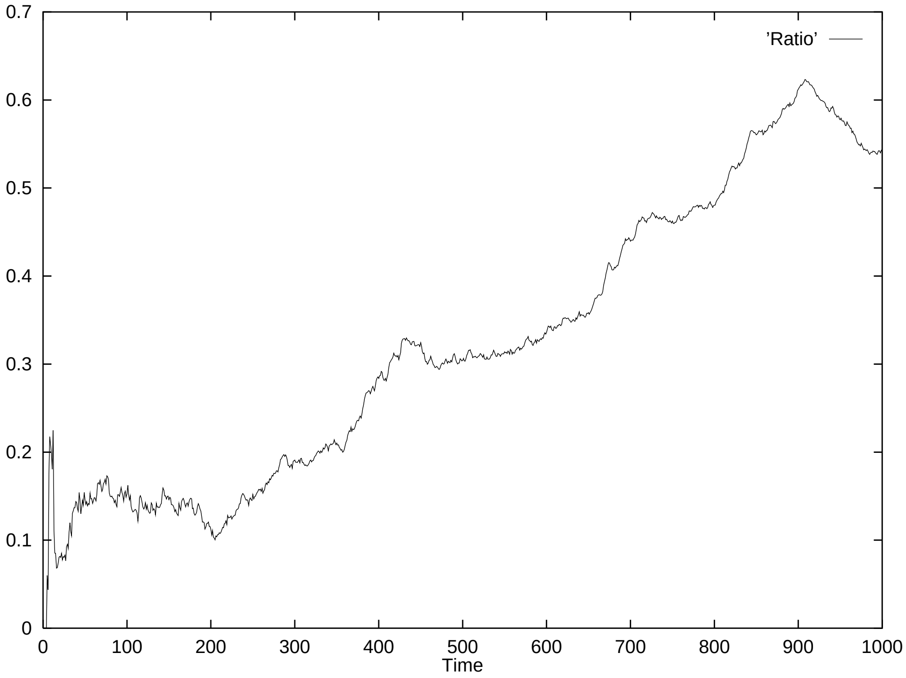

Figure 8: Ratio of the experimental 2nd moment by the theoretical 2nd moment

References

[Alt 76] Alt, R.: "Etude statistique de l'erreur numérique d'affectation sur un ordinateur en base quelconque. Application à l'erreur commise dans le calcul d'une somme de produits de nombres"; IFP Report 76-5 (1976).

[Alt 78] Alt, R.: "Error propagation in Fourier Transforms"; Mathematics and Computers in Simulation, 20 (1978), 37-43.

[Chesneaux 88] Chesneaux, J.-M.: "Etude théorique et implémentation en ADA de la méthode CESTAC"; Doctoral Thesis, Univ. Paris VI, (1988).

[Chesneaux 90] Chesneaux, J.-M.: "Study of the computing accuracy by using probabilistic approach. Contribution to computer arithmetic and self-validating numerical methods"; ed. C. Ulrich, J.C. Baltzer (1990), 19-30.

[Hamming 70] Hamming, R. W.: "On the distribution of numbers"; Bell System Techn. J. 49, 8 (1970), 1609-1625.

[Hull, Swenson 66] Hull, T. E., Swenson, J. R.: "Test of probalistic model for propagation of round-off errors"; A.C.M. 9, 2 (1966) 108-111.

[Knuth 69] Knuth, D. E.: "The art of computer programming"; Add. Wesley. (1969).

[La Porte, Vignes 74a] La Porte, M., Vignes, J.: "Etude statistique des erreurs dans l'arithmétique des ordinateurs; application au contrôle des résultats d'algorithmes numériques"; Numer. Math. 23 (1974) 63-72.

[La Porte, Vignes 74b] La Porte, M., Vignes, J.: "Algorithmes numériques, analyse et mise en œuvre"; Vol. 1, Editions Technip, Paris (1974).

[Vignes 93] Vignes, J.: "A stochastic arithmetic for reliable computation"; Mathematics and Computers in Simulation, 35 (1993), 233-261.