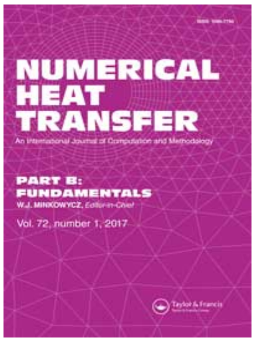
NUMERICAL HEAT TRANSFER
An International Journal of Computation and Methodology
PART B: FUNDAMENTALS
W.J. MINKOWYCZ, Editor-in-Chief
Vol. 72, number 1, 2017
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Numerical Heat Transfer, Part B: Fundamentals

An International Journal of Computation and Methodology

ISSN: 1040-7790 (Print) 1521-0626 (Online) Journal homepage: http://www.tandfonline.com/loi/unhb20

An example for the effect of round-off errors on numerical heat transfer

Shan-Cong Mou, Yu-Xuan Luan, Wen-Tao Ji, Jian-Fei Zhang & Wen-Quan Tao

To cite this article: Shan-Cong Mou, Yu-Xuan Luan, Wen-Tao Ji, Jian-Fei Zhang & Wen-Quan Tao (2017) An example for the effect of round-off errors on numerical heat transfer, Numerical Heat Transfer, Part B: Fundamentals, 72:1, 21-32

To link to this article: http://dx.doi.org/10.1080/10407790.2017.1338096

Published online: 11 Jul 2017.

NUMERICAL HEAT TRANSFER, PART B
2017, VOL. 72, NO. 1, 21–32
https://doi.org/10.1080/10407790.2017.1338096

Taylor & Francis
Taylor & Francis Group

An example for the effect of round-off errors on numerical heat transfer

Shan-Cong Mou, Yu-Xuan Luan, Wen-Tao Ji, Jian-Fei Zhang, and Wen-Quan Tao

Key Laboratory of Thermo-Fluid Science and Engineering of MOE, Xi'an Jiaotong University, Xi'an, China

ABSTRACT
The effect of round-off errors on the solution of numerical heat transfer is illustrated by a simple example both analytically and numerically. It is found that the upper bound of the round-off error under both conditions with or without an inner heat source is proportional to the square of grid number—n^2. Increase in grid number might lead to larger round-off errors. The magnitude of relative round-off error is also determined by the specific problem. Proper treatment of the computation procedure can reduce the round-off error obviously. The precision can be improved with this method without occupation of additional computational resources.

ARTICLE HISTORY
Received 27 March 2017
Accepted 16 May 2017

1. Introduction

When solving nonlinear partial differential equations (PDEs) of heat transfer and fluid-flow problems with a numerical approach, the errors would arise in three different types: round-off errors, iterative convergence errors, and truncation errors, respectively [1–3].

Truncation error is defined as the difference between PDE and the corresponding finite-difference equations (FDEs) [2]. Iteration error is the difference between the exact and iterative solutions of discrete equation [4]. There have been numerous works done by investigators on verification [2] and reducing these errors. For example, one has to meet the condition of consistency that describes the extent to which FDEs approximate the PDEs, then use higher order scheme [3] or refine the grids [4] to reduce truncation error. As for the iteration error, it can be reduced by increasing the number of iterations. Currently, two widely accepted criteria for interrupting an iterative process have been developed [5].

Round-off error is caused by the representation of real numbers by a finite number of significant digits in computers [1, 6]. Relative round-off error of addition, subtraction, multiplication, and division denoted by op can be expressed as follows [7, 8]:

$$\mathrm{fl}(x_1 \text{ op } x_2) = (x_1 \text{ op } x_2)(1+\varepsilon) \quad \varepsilon < \text{ò} \tag{1}$$

where $\mathrm{fl}(x_1 \text{ op } x_2)$ is the result of floating-point calculation. Machine precision ò is the relative round-off error which is always bounded by $\text{ò} = 2^{-p}$. The procedure of floating-point arithmetic is introduced according to IEEE754 standard [9] in Section 2.

The research of round-off error can date back to 1940s (Goldstine and Neumann [10] and Turing [11]). Later in the 1960s, Moore [12] developed the interval arithmetic to analyze the round-off error and Wilkinson [7] discussed the round-off error in the algebra process in detail. Their works are the basis of the following investigations on round-off error.

Recently, there have been numerous works on the upper bound of the round-off error. Some tools are developed to calculate the upper bound (Gappa [13], Fluctuat [14], based on interval arithmetic

CONTACT W.-T. Ji ✉ wentaoji@xjtu.edu.cn Key Laboratory of Thermo-Fluid Science and Engineering of MOE, Xi'an Jiaotong University, Xi'an 710049, China.

Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/unhb.

Nomenclature

A, B, C, P, Q	coefficients in TDMA
a, b	coefficients
Bi	Biot number
E	round-off error
E_r	relative round-off error
e	exponent; relative error in basic operations
F	set of floating-point number
F	cross-sectional area, m^2
f	arbitrary function
fl	floating point
h	convective heat transfer coefficient, W/m^2
k	overall heat transfer coefficients, W/m^2/K
L	length of the slab, m
M_1	grid number of the rightmost grid point
m	significant
op	operation
p	precision
R	set of real number
r_o	round function
S	inner heat source, W/m^3
s	sign
T	temperature, °C
$\boldsymbol{x}$	arbitrary vector
x	variable; coordinate along the slab, m
δ	maximal absolute error for numbers close to zero
δx	distance between grid points, m
ε	relative error in basic operations
ò	machine precision
λ	thermal conductivity, W/m/K
ϕ	solution of the PDE
Subscripts	
1	left side of the slab
2	right side of the slab
E	east
f	fluid
i	TDMA calculation step; grid number
n	total grid number
P	current grid point
up	upper bound
W	west

and symbolic Taylor expansion). The analysis of the round-off error was conducted in many fields including chemical kinemics [15], molecular dynamics [16], and astronomy [17].

To the best knowledge of the authors, the discussions on round-off errors in computational fluid dynamics (CFD) and numerical heat transfer (NHT) were mostly limited in that the magnitude of round-off error is proportional to the grid number [2] and that the levels of machine precision are simply increased to reduce the round-off error [1].

Neglecting the iteration error, the error between the exact solution of the PDE and the computer solution to the FDEs can be expressed as follows [2, 3]:

$$\phi(i,n) - \tilde{\phi}_i^n = \underbrace{\phi(i,n) - \phi_i^n}_{\text{discretization error}} + \underbrace{\phi_i^n - \tilde{\phi}_i^n}_{\text{round - off error}} \tag{2}$$

It is obvious that a contradiction exists between refining the grid to reduce the discretization error and reducing the grid number to limit the accumulation of round-off errors since the machine precision is always finite. All those previous discussions are mostly qualitative, relatively few investigations were being performed on quantitative analysis of round-off errors in NHT. This paper uses the method of Taylor expansions [18] to identify the effect of round-off errors on a one-dimensional heat transfer problem. Besides, other factors including the grid number and Bi number are also analyzed.

In the following presentation, it is divided into five sections: floating-point arithmetic is introduced in Section 2; Section 3 presents the mathematical model and numerical method of the example; in Section 4, the rigorous upper bound of round-off errors is calculated using the method of symbolic Taylor expansion [18] and a numerical experiment is performed to examine it; finally, some conclusions are made in Section 5.

2. Floating-point arithmetic

2.1. Floating-point representation

This section serves to provide some information of floating-point arithmetic, based on which the analysis of the simple example is presented in Section 3.

Table 1. Value of machine precision.

Precision (bits)	ò	δ
Single (32)	2^{-24}	2^{-150}
Double (64)	2^{-53}	2^{-1075}
Quadruple (128)	2^{-113}	2^{-16495}

As is defined in IEEE754 standard [9], a binary floating-point number has the form:

$$(-1)^{s}\times 2^{e}\times m \tag{3}$$

where m, e, s are significant, exponent, and sign, respectively. And the standard describes three formats: single (32 bits), double (64 bits), and quad (128 bits) (see [9, Section 3] for detail).

The set of floating-point numbers is denoted as F which is a subset of the real number set R. And r_{o}: R → F is a rounding operator which returns the closest floating-point number of a given real number [9]. The following formula gives the model of rounding [19]:

$$r_{\text{o}}(x) = x(1+\varepsilon_i) + d \tag{4}$$

where $|\varepsilon_i| \leq$ ò, $|d| \leq \delta$ and $\varepsilon_i \times d = 0$. The values of ò and δ of different formats are given in Table 1. Since δ is negligible compared to ò, the term δ could be neglected in the following sections [20].

2.2. Floating-point operation

In the standard, several floating-point arithmetic operations are defined. Suppose an operation op: $\mathrm{R}^n \to \mathrm{R}$ and $\mathrm{op}_{\mathrm{fl}}$ is its corresponding floating-point operation.

If

$$\mathrm{op}_{\mathrm{fl}}(\mathbf{x}) = r_{\text{o}}(\mathrm{op}(\mathbf{x})) \tag{5}$$

holds for all $\mathbf{x}$ in R^n, then the operation is exactly rounded. According to the IEEE754 standard, the following basic operations are exactly rounded: +, −, ×, / [18].

Finally the model for floating-point arithmetic of those exactly rounded operations could be expressed as follows:

$$\mathrm{op}_{\mathrm{fl}}(\mathbf{x}) = \mathrm{op}(\mathbf{x})(1+\varepsilon_i) \tag{6}$$

In the following section, the model of the example is presented in detail.

3. Mathematical model and numerical solution

3.1. Mathematical model

Consider the simple problem of an infinite vertical plate with the third kind of boundary conditions at two faces, which is one-dimensional and steady-state heat conduction, see Figure 1 [3]:

$$\begin{cases} \frac{1}{F(x)}\frac{\mathrm{d}}{\mathrm{d}x}\left[\lambda F(x)\frac{\mathrm{d}T}{\mathrm{d}x}\right] + S = 0 \\ h_{f1}(T_{f1} - T|_{x=0}) + \lambda\frac{\mathrm{d}T}{\mathrm{d}x}|_{x=0} = 0 \\ h_{f2}(T|_{x=L} - T_{f2}) + \lambda\frac{\mathrm{d}T}{\mathrm{d}x}|_{x=L} = 0 \end{cases} \tag{7}$$

where h_f is convective heat transfer coefficient, λ is thermal conductivity, and L is plate thickness.

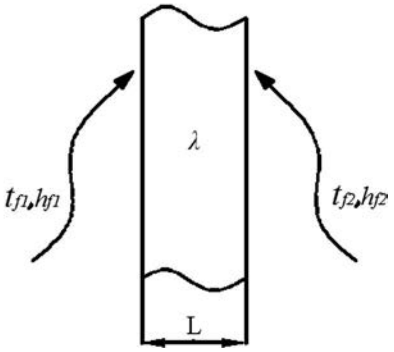

Figure 1. Physical model.

At first, consider the condition in which the inner heat source $S = 0$. The analytical solution to this problem can be written as:

$$T = \frac{T_{f2} - T_{f1}}{L + \left(\frac{1}{h_{f1}} + \frac{1}{h_{f2}}\right)\lambda} x + \frac{T_{f1}L + \lambda\left(\frac{T_{f1}}{h_{f2}} + \frac{T_{f2}}{h_{f1}}\right)}{L + \left(\frac{1}{h_{f1}} + \frac{1}{h_{f2}}\right)\lambda} \tag{8}$$

Note that the solution is linear, and this is the main reason that the rather simple problem is served as the example—the nonexistence of discretization errors and iteration errors, which means the only error existed between the numerical solution and the analytical solution to this problem is round-off error. These unique features would provide great convenience to identify the effect of round-off errors.

3.2. Numerical solution

In this section, the numerical method is introduced to solve this problem. And the basic information for round-off error analysis is provided.

The control volume integration method [3] is applied to discretize the governing equation and uniform grid with cell central scheme [21] is used to discretize the computational domain, see Figure 2. Discretization equation of boundary is obtained from energy balance for each control volume. The FDEs have the form [3]:

$$a_P T_P = a_E T_E + a_W T_W + b \tag{9}$$

where

$$\begin{gathered} a_E = \frac{F_e k_e}{\delta x_e}, \quad a_W = \frac{F_w k_w}{\delta x_w} \\ a_P = a_E + a_W - S_P F_P \Delta x \\ b = S_C F_P \Delta x \end{gathered} \tag{10}$$

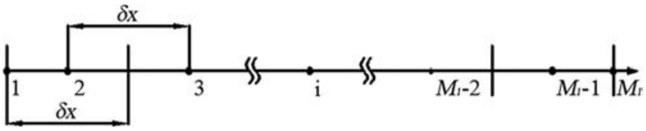

Figure 2. Discretization of the computational domain.

Tridiagonal matrix algorithm (TDMA) [22] is applied to solve the FDEs. For a system of equations that has a tridiagonal form:

$$A_iT_i = B_iT_{i+1} + C_iT_{i+1} + D_i \tag{11}$$

After forward elimination, Eq. (11) can be written as:

$$T_{i-1} = P_{i-1}T_i + Q_{i-1} \tag{12}$$

where

$$P_i = \frac{B_i}{A_i - C_iP_{i-1}} \tag{13}$$

$$Q_i = \frac{D_i + C_iQ_{i-1}}{A_i - C_iP_{i-1}} \tag{14}$$

With back-substitution, the temperature on every node could be solved, for example:

$$T_{M_1} = Q_{M_1} \tag{15}$$

where M_1 is the number of the rightmost point which equals the total grid number, n.

In this case:

$$P_1 = \frac{2\frac{\lambda}{\delta x}}{2\frac{\lambda}{\delta x} + h_{f1}} \tag{16}$$

$$P_2 = \frac{\frac{\lambda}{\delta x}}{3\frac{\lambda}{\delta x} - 2P_1\frac{\lambda}{\delta x}} \tag{17}$$

$$P_i = \frac{\frac{\lambda}{\delta x}}{2\frac{\lambda}{\delta x} - P_{i-1}\frac{\lambda}{\delta x}} \quad i > 2 \quad \text{and} \quad i < M_1 - 1 \tag{18}$$

$$P_{M_1-1} = \frac{2\frac{\lambda}{\delta x}}{3\frac{\lambda}{\delta x} - P_{M_1-2}\frac{\lambda}{\delta x}} \tag{19}$$

$$Q_1 = \frac{h_{f1}T_{f1}}{h_{f1} + \frac{2\lambda}{\delta x}} \tag{20}$$

$$Q_2 = \frac{2\frac{\lambda}{\delta x}}{3\frac{\lambda}{\delta x} - 2P_1\frac{2\lambda}{\delta x}} \tag{21}$$

$$Q_i = \frac{\frac{\lambda}{\delta x}Q_{i-1}}{2\frac{\lambda}{\delta x} - P_{i-1}\frac{\lambda}{\delta x}} \quad i > 2 \quad \text{and} \quad i < M_1 - 1 \tag{22}$$

$$Q_{M_1-1} = \frac{\frac{\lambda}{\delta x}Q_{M_1-2}}{3\frac{\lambda}{\delta x} - P_{M_1-2}\frac{\lambda}{\delta x}} \tag{23}$$

$$Q_{M_1} = \frac{h_{f2}T_{f2} + 2\frac{\lambda}{\delta x}Q_{M_1-1}}{h_{f2} + 2\frac{\lambda}{\delta x} - 2\frac{\lambda}{\delta x}P_{M_1-1}} \tag{24}$$

Tridiagonal matrix algorithm is a direct method for one-dimensional situations, which is widely used in a line-by-line form in programs to solve multidimensional CFD and NHT problems (say, ADI [3]). Therefore, it is important to analyze the accumulation of round-off errors in TDMA and identify its effect on the precision of the numerical solution. Analysis of round-off error would be presented in the next section.

4. Round-off error analysis

4.1. Round-off error calculation

This section will primarily focus on the accumulation of round-off errors in TDMA and neglect the errors induced by floating-point representation of domain discretization which is relatively small.

As mentioned in Section 2, the model for floating-point arithmetic of those basic operations (say, addition, subtraction, multiplication, and division) is:

$$\mathrm{op}_{\mathrm{fl}}(\mathbf{x}) = \mathrm{op}(\mathbf{x})(1+\varepsilon_i) \tag{25}$$

Given a function *f*: $\mathrm{R}^k \to \mathrm{R}$, it is calculated by a computer in the form of fl(*f*): $\mathrm{R}^k \to \mathrm{F}$, with all of the operations in *f* replaced by the corresponding floating-point ones as well as variables and constants, unless they are already floating-point numbers. Substituting Eq. (25) into fl(*f*) and denote it as $\hat{f}(\mathbf{x},\varepsilon)$. The round-off error when computing *f* can be expressed as [18]:

$$E(f) = \hat{f}(\mathbf{x},\varepsilon) - f(\mathbf{x}) \tag{26}$$

Applying Taylor expansion, the round-off errors of *f*(**x**) for all the principal variables are:

$$E(f) = \sum_{i=1}^{k} \left.\frac{\partial \hat{f}}{\partial \varepsilon_i}\right|_{(\mathbf{x},0)} \varepsilon_i + O(\grave{o}^2) \tag{27}$$

Note that $|\varepsilon_i| \le \grave{o}$ is relatively small (Table 1), the term $O(\grave{o}^2)$ is neglected. Define the upper bound of round-off error as:

$$E_{\mathrm{up}}(f) = \grave{o}\sum_{i=1}^{k} \left|\left.\frac{\partial \hat{f}}{\partial \varepsilon_i}\right|_{(\mathbf{x},0)}\right| \tag{28}$$

Our goal is to calculate the upper bound for the accumulation of round-off error in the example. Now, consider Eq. (13), the error between the true value P_{i-1} and the computed value $\hat{P}_{i-1}$ is:

$$E(P_{i-1}) = \hat{P}_{i-1} - P_{i-1} \tag{29}$$

And $\hat{P}_i$ can be written as:

$$\hat{P}_i = \mathrm{fl}\left(\frac{\hat{B}_i}{\hat{A}_i - \hat{C}_i\hat{P}_{i-1}}\right) \tag{30}$$

A_i, B_i, C_i are coefficients that have been calculated outside of TDMA. In other words, they are already floating-point numbers. Since they are calculated differently in different codes, for sake of conciseness, here the round-off errors induced during calculating those coefficients outside TDMA would be neglected, and they have minor influence on the results.

Then

$$\hat{P}_i = \frac{B_i}{[A_i - C_i(P_{i-1} + E(P_{i-1}))(1+\varepsilon_1)](1+\varepsilon_2)}(1+\varepsilon_3) \tag{31}$$

Substitute Eq. (31) into Eq. (26), applying Taylor expansion and neglecting higher order terms, we get the recurrence relation of the error sequence:

$$E(P_i) = (1+\varepsilon_1+\varepsilon_2+\varepsilon_3)P_i^2 E(P_{i-1}) + P_i^2 P_{i-1}\varepsilon_1 - P_i\varepsilon_2 + P_i\varepsilon_3 \tag{32}$$

Since $\varepsilon_i \ll 1$, neglecting ε_i in the coefficient of $E(P_{i-1})$, we have:

$$|E(P_i)| \leq P_i^2|E(P_{i-1})| + \left|P_i^2P_{i-1}\varepsilon_1 - P_i\varepsilon_2 + P_i\varepsilon_3\right| \leq P_i^2|E(P_{i-1})| + (2P_i + P_i^2P_{i-1})\grave{o} \tag{33}$$

Thus, the recurrence relation of the upper bound error sequence can be written as:

$$E_{\mathrm{up}}(P_i) = (2P_i + P_i^2P_{i-1})\grave{o} + P_i^2E_{\mathrm{sup}}(P_{i-1}) \tag{34}$$

From Eq. (18), the equation of P_i could be written as:

$$P_i = \frac{1}{2 - P_{i-1}} \tag{35}$$

The general term can be calculated as follows by Fix-point method, where P_2 is the first term given by Eq. (17):

$$P_i = \frac{1}{2 - P_{i-1}} = 1 + \frac{1}{\frac{1}{P_2 - 1} - (i-2)} \tag{36}$$

$$P_{M_1-1} = \frac{1}{1 + \frac{1}{2}\frac{1}{\frac{n}{Bi_1} + n - 2.5}} \approx 1 - \frac{1}{2}\frac{1}{\frac{n}{Bi_1} + n - 2.5} \tag{37}$$

where Bi is Biot number. It is obvious that $P_i < 1$ and P_i is monotonically increasing with respect to i.

Substituting Eq. (36) into Eq. (34), we can calculate the general term of the upper bound:

$$E_{\mathrm{up}}(P_i) \approx i\grave{o} \quad 1 << i < M_1 - 1 \tag{38}$$

Substituting Eq. (38) into Eq. (19) and applying Taylor expansion again, we have:

$$E_{\mathrm{up}}(P_{M_1-1}) = \frac{1}{2}n\grave{o} \tag{39}$$

Similarly, the upper bound of round-off error when computing Q_i can also be calculated, and with some reasonable approximation, the final bound is:

$$E_{\mathrm{up}}(Q_{M_1-1}) = \frac{1}{2}T_{f1}n\grave{o} \tag{40}$$

And in the process of calculating $E_{\mathrm{sup}}(Q_{M_1-1})$, to make sure that the accumulation of round-off errors to be convergent, grid number n has to satisfy the following inequality (see Appendix A for detail):

$$n < \sqrt{\frac{Bi_1}{Bi_1 + 1}\frac{1}{2\grave{o}}} \tag{41}$$

Finally substitute Eqs. (39) and (40) into Eq. (24) and use Eq. (28) again. We get the expression of the maximum relative error $E_{\mathrm{r\,max}}$ for Q_{M_1} (since $T_{M_1} = Q_{M_1}$, it is also the relative error $E_{\mathrm{r\,max}}$ for T_{M_1}):

$$E_{\mathrm{r\,max}} = \frac{E_{\mathrm{up}}(T_{M_1})}{T_{M_1}} = \left(\frac{1}{\frac{T_{f2}}{T_{f1}}Bi_2 + 1} + \frac{1}{Bi_2 + \frac{Bi_1}{Bi_1+1}}\right)n^2\grave{o} \tag{42}$$

$Bi_1 = \frac{h_{f1}L}{\lambda}$, $\quad Bi_2 = \frac{h_{f2}L}{\lambda}$, respectively.

So far, the expression of the maximum relative round-off error $E_{\mathrm{r\,max}}$ between the numerical solution and the exact solution has been derived.

4.2. Effect of the round-off error and numerical experiment

In this section, discussions on the maximum relative round-off error expression are given combining with a numerical experiment. Based on those discussions, a method on reducing the error without occupying additional computational resource is put forward.

Define the relative error E_r between numerical solution $\mathrm{fl}(T_{M_1})$ and exact solution T_{M_1} as:

$$E_\mathrm{r} = \frac{|\mathrm{fl}(T_{M_1}) - T_{M_1}|}{T_{M_1}} \tag{43}$$

The experiment is performed on the relative round-off error E_r. Since the upper bound of E_r cannot always be reached in computing, the correspondence between E_r and E_rmax has to be demonstrated:

1. Let $Bi_1 = Bi_2 = Bi$, vary the value of Bi from 0.005 to 5 [Figure 3(*a*)], the magnitude of relative round-off error decreases as Bi increases, which agrees well with our theoretical model.

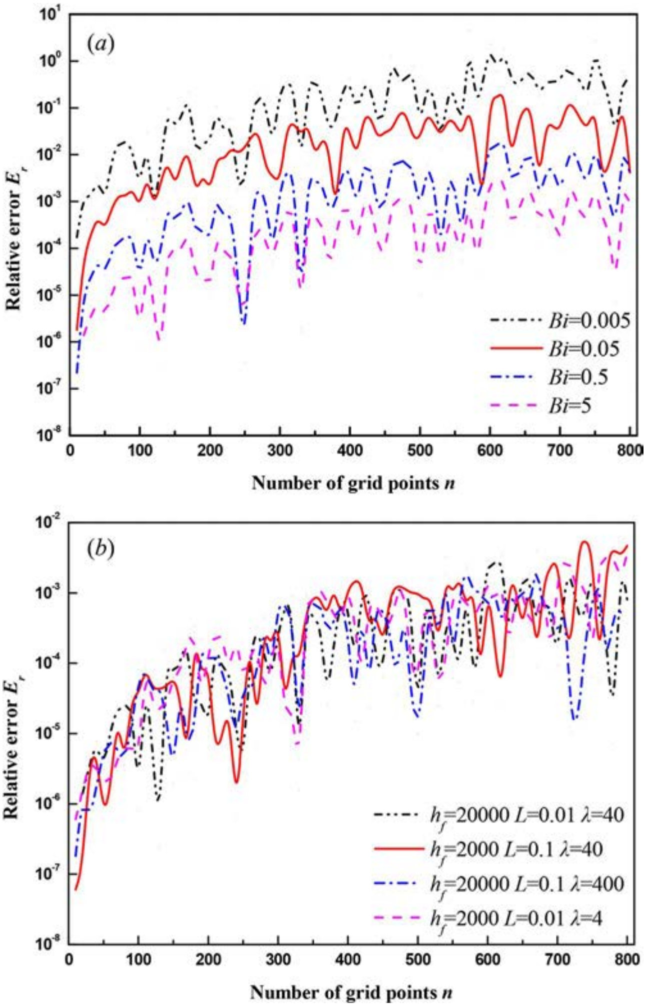

Figure 3. Relative round-off error *versus* grid number when $Bi_1 = Bi_2 = Bi$. (*a*) *Bi* varies from 0.005 to 5. (*b*) *Bi* keeps unchanged while varying other parameters.

Table 2. Solution to the example: T_{M_1} (parameters: $L = 0.01$ m, $T_{f1} = 100$°C, $T_{f1} = 20$°C, $h_{f1} = 2000$ W/m^2, $h_{f2} = 20$ W/m^2).

		Numerical solution				
Exact solution		$M_1 = 3$	$M_1 = 10$	$M_1 = 20$	$M_1 = 40$	$M_1 = 80$
99.21182	Single (32 bit)	99.21181	99.21188	99.20975	99.22114	99.18736
	Double (64 bit)	99.21181	99.21190	99.21159	99.21210	99.21248

2. Keeping the value of Bi unchanged by varying the geometry parameter L and the convective heat transfer coefficient h_f at the same time. According to the model, the magnitude of relative round-off error should remain the same [Figure 3(b)].

So far, it has been demonstrated that E_{rmax} can be a valid representation of the actual round-off error $\mathrm{E_r}$. Thus, all the following discussions are based on the expression of E_{rmax}.

1. The maximum relative round-off error is proportional to the machine precision— ε, thus using high level of machine precision is a way to reduce the round-off error, which agrees well with the result of numerical experiment (Table 2) as well as the arguments in [1, 2]. However, huge amount of the computational resource would be required.
2. It should be noted that the magnitude of both the maximum round-off error and the maximum relative round-off error is proportional to the square of grid number—n^2, which is much larger than the proposed number n [2, Chapter 3]. Thus the magnitude of the round-off error can increase very fast when using more grids to reduce the discretization error, especially when n or the number of iteration steps is relatively large.
3. Normally, in a given number of grids, increasing of machine precision would reduce the effect of round-off error. However, it will use more of the computing resources. Considering the contradictions between computational resources and machine precision, discretization error and round-off error, to reduce the round-off error without the occupation of more computational resources or the increase in discretization error, the specific problem has to be conscientiously considered. Here, the influence of B_i number is identified and a method to improve the precision is put forward:

Note that the influence of Bi_1 and Bi_2 on the value of the relative round-off error is different [Eq. (42)]. It is found that the precision of computation result changes by varying the Bi number: The convective heat transfer coefficient is relatively large on one side but rather small on the other side (It is common in engineering applications, for example, the forced convection heat transfer between air and water). The different arrangement of the computation may have a significant effect on the precision of numerical results. Take an instance, when keeping $h_{f1} = 20\,\mathrm{W/m^2}$ unchanged ($L = 0.01$ m, $\lambda = 1$ W/m, $Bi_1 = 0.2$) while varying h_{f2} from $20\,\mathrm{W/m^2}$ to $20,000\,\mathrm{W/m^2}$ ($L = 0.01\,\mathrm{m}$, $\lambda = 1\,\mathrm{W/m}$, $Bi_2 = 0.2 - 200$) the magnitude of relative error reduced significantly with the increase in h_{f2} [Figure 4(a)]. However, when keeping $h_{f2} = 20\,\mathrm{W/m^2}$ unchanged ($L = 0.01\,\mathrm{m}$, $\lambda = 1\,\mathrm{W/m}$, $Bi_2 = 0.2$) while varying h_{f1} from $20\,\mathrm{W/m^2}$ to $20,000\,\mathrm{W/m^2}$ ($L = 0.01\,\mathrm{m}$, $\lambda = 1\,\mathrm{W/m}$, $Bi_1 = 0.2 - 200$), it has little influence on the magnitude of the relative error [Figure 4(b)]. This can be explained by Eq. (42): the influence of Bi_1 is limited in the term $\frac{Bi_1}{Bi_1+1}$, its value is less than 1 no matter how large Bi_1 is. However, if Bi_2 is chosen as the larger one, the magnitude of relative error can be reduced significantly (when the convective heat transfer coefficients at the two faces are $20\,\mathrm{W/m^2}$ and $2,000\,\mathrm{W/m^2}$, respectively, if h_{f1} is chosen to be $2,000\,\mathrm{W/m^2}$ and h_{f2} to be $20\,\mathrm{W/m^2}$, the relative round-off error can be 16 times larger than the other arrangement that $h_{f1} = 20\,\mathrm{W/m^2}$ and $h_{f2} = 2,000\,\mathrm{W/m^2}$). It indicated that proper choice of Bi at two surfaces can have a significant effect on the precision of the numerical solution. With proper treatment of the computation procedure, the round-off error can be reduced significantly while keeping the same machine precision level. In this case, change the position of Bi_1 and Bi_2, make the larger Bi as Bi_2 can systematically minimize the round-off errors without occupying additional computational resources.

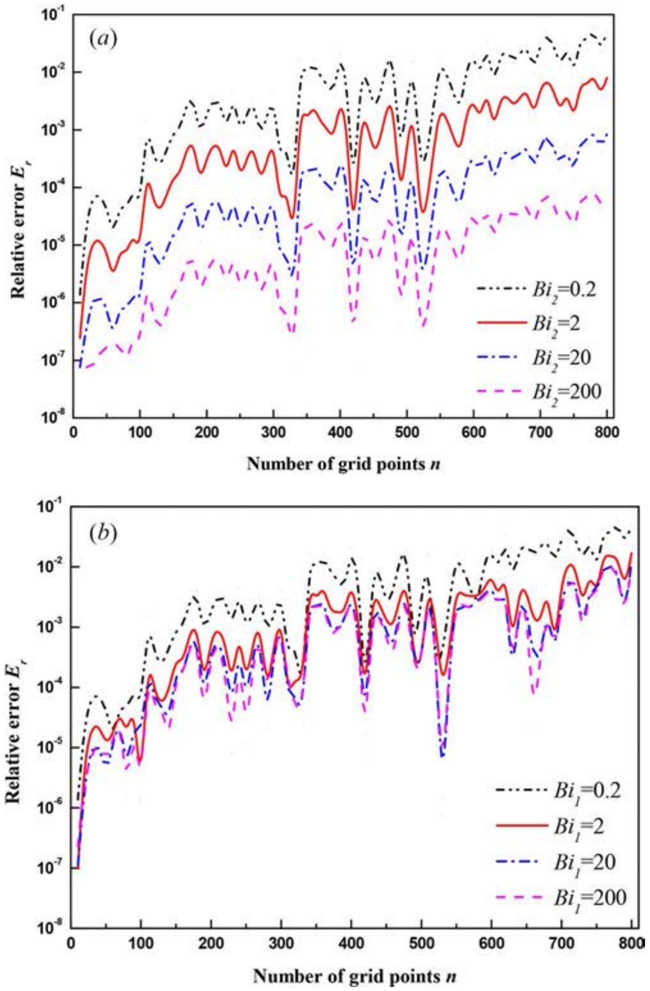

Figure 4. Relative round-off error *versus* grid number when $Bi_1 \neq Bi_2$. (*a*) Bi_2 varies from 0.2 to 200 while $Bi_1 = 0.2$. (*b*) Bi_1 varies from 0.2 to 200 while $Bi_2 = 0.2$.

4.3. Effect of the source term

The calculation is quite similar as discussed before. The expression of the maximum relative error E_{rmax} is derived when $S = S_c$, where S_c is the intensity of constant inner heat resource:

$$E = \left(\frac{\frac{S_c L^2}{12\lambda T_{f1}} + 1}{\frac{T_{f2}}{T_{f1}} Bi_2 + \frac{S_c L^2}{2\lambda T_{f1}} + 1} + \frac{1}{Bi_2 + \frac{Bi_1}{Bi_1+1}} \right) n^2 \grave{o} \tag{44}$$

It is found that the magnitude of relative round-off error is still proportional to the square of grid number. Bi_2 also has great effect on the magnitude of error, since the second term in square is the same as before. While, if the magnitude of inner heat source is moderate, it has less a significant influence on the round-off error because S_c is both included in numerator and denominator.

5. Conclusion

In this paper, the effect of round-off errors on NHT is identified by a simple heat conduction example, where iterative and truncation errors do not exist, and increase in grid number might result in greater error. The method of computing the upper bound of round-off error is introduced. During the process of computing the error, a recommended constraint of grid number is given. Finally, combining with the numerical experiment, the expression of relative round-off error is analyzed in detail, based on which a method on reducing the error without occupying additional computational resource is put forward. The following conclusion can be drawn.

1. The upper bound of the round-off error is proportional to ò, where ò is machine precision, the error could be reduced by improving the level of machine precision.
2. As can be seen from the expression with or without inner heat source, the magnitude of accumulated relative round-off errors is proportional to the square of the grid number n^2, much larger than the proposed number n [2].
3. As mentioned before, former researchers suggested to refine the grid to reduce the discretization error, but no criteria of the upper limit of the grid number have been put forward so far. To make sure the convergence of the round-off error, a constraint of the grid number n to be chosen in the example is recommended.
4. The effect of round-off errors on the numerical solution is not simply determined by the machine precision and grid number. However it is also related to the specific problem and dominated by the nondimensional parameters, such as *Bi* number in this example. With proper choice of *Bi* at two surfaces, the round-off error can be reduced significantly without occupying additional computational resources.
5. Since TDMA have been widely used in NHT and CFD programs, when the calculation is not convergent and the number of computer bits are limited, increasing of grid number might lead to larger round-off errors. Consideration of the effect of round-off error is highly recommended.

Acknowledgment

This work has been financially supported by the National Key Research and Development Program — China (2016YFB0601201).

References

[1] K. H. Versteeg and W. Malalasekera, *An Introduction to Computational Fluid Dynamics: The Finite Volume Method*, chap. 7, Pearson Education, London, UK, 2007.

[2] J. C. Tannehill, D. A. Anderson, and R. H. Pletcher, *Computational Fluid Mechanics and Heat Transfer*, 2nd ed., chap. 3, Taylor & Francis, Washington, DC, 1997.

[3] W. Q. Tao, *Numerical Heat Transfer*, 2nd ed., chap. 3, Xi'an Jiaotong University Press, Xi'an, China, 2001.

[4] C. H. Marchi and A. F. C. Silva, Unidimensional Numerical Solution Error Estimation for Convergent Apparent Order, *Numer. Heat Transfer B Fund.*, vol. 42, no. 2, pp. 167–188, 2002.

[5] M. A. Martins and C. H. Marchi, Estimate of Iteration Errors in Computational Fluid Dynamics, *Numer. Heat Transfer B Fund.*, vol. 53, no. 3, pp. 234–245, 2008.

[6] D. Goldberg, What Every Computer Scientist Should Know About Floating-point Arithmetic, *ACM Comput. Surv. (CSUR)*, vol. 23, no. 1, pp. 5–48, 1991.

[7] J. H. Wilkinson, *Rounding Error in Algebraic Process*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[8] N. J. Higham, Accuracy and Stability of Numerical Algorithms, *J. Am. Stat. Assoc.*, vol. 16, no. 94, pp. 285–289, 2002.

[9] D. Zuras, et al. *IEEE Standard for Floating-point Arithmetic*, vol. 754, pp. 1–70, IEEE Std, 2008.

[10] H. H. Goldstine and J. V. Neumann, Numerical Inverting of Matrices of High Order, *Bull. Am. Math. Soc.*, vol. 53, no. 11, pp. 1021–1099, 1947.

[11] A. M. Turing, Rounding-off Errors in Matrix Processes, *Quart. J. Mech. Appl. Math.*, vol. 1, no. 1, pp. 287–308, 1948.

[12] R. E. Moore, *Interval Analysis*, vol. 4, chap. 1, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[13] M. Daumas and G. Melquiond, Certification of Bounds on Expressions Involving Rounded Operators, *ACM Trans. Math. Software*, vol. 37, no. 1, pp. 495–507, 2007.

[14] D. Delmas, E. Goubault, S. Putot, J. Souyris, K. Tekkal, and F. V'Edrine. Towards an Industrial Use of FLUCTUAT on Safety-critical Avionics Software, Formal Methods for Industrial Critical Systems, *International Workshop*, FMICS, Eindhoven, The Netherlands, November, vol. 5825, pp. 53–69, 2009.

[15] A. Lukassen and M. Kiehl, Reduction of Round-off Errors in Chemical Kinetics, *Combust. Theor. Model.*, vol. 21, pp. 183–204, 2016.

[16] M. Taufer, O. Padron, P. Saponaro, and S. Patel. Improving Numerical Reproducibility and Stability in Large-scale Numerical Simulations on GPUs, Parallel & Distributed Processing (IPDPS), *IEEE International Symposium on Parallel and Distributed Processing*, pp. 1–9, Atlanta, GA, 2010.

[17] T. Quinn and S. Tremaine, Roundoff Error in Long-term Planetary Orbit Integrations, *Astronom. J.*, vol. 99, pp. 1016–1023, 1990.

[18] A. Solovyev, C. Jacobsen, Z. Rakamarić, et al. Rigorous Estimation of Floating-point Round-off errors with Symbolic Taylor Expansions, *International Symposium on Formal Methods*, vol. 9109, pp. 532–550, 2015.

[19] F. Goualard, How Do You Compute the Midpoint of an Interval?, *ACM Trans. Math. Software (TOMS)*, vol. 40, no. 2, pp. 11, 2014.

[20] A. Rocca, V. Magron, and T. Dang, Certified Roundoff Error Bounds Using Bernstein Expansions and Sparse Krivine-Stengle Representations, arXiv preprint arXiv:1610.07038, 2016.

[21] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*, chap. 3, CRC Press, 1980.

[22] A. Khalid and S. Antonin, *Petroleum Reservoir Simulation*, Chapman & Hall, London, UK, 1979.

Appendix A

Recall the process of calculating $E(P_\mathrm{i})$, we have:

$$\widehat{Q}_i = \frac{[D_i + C_i Q_{i-1}(1+\varepsilon_4)](1+\varepsilon_5)}{[A_i - C_i(P_{i-1} + E(P_{i-1}))(1+\varepsilon_1)](1+\varepsilon_2)}(1+\varepsilon_6) \tag{45}$$

Substitute into Eq. (26), applying Taylor expansion, and neglecting higher order terms, we get the recurrence relation of the error sequence:

$$E(Q_i) = (1+N)(1+\varepsilon_4)P_i E(Q_{i-1}) + NQ_i + (1+N)\varepsilon_4 P_i Q_{i-1} \tag{46}$$

where

$$N = \varepsilon_5 + \varepsilon_6 - \frac{A_i}{C_i}P_i\varepsilon_2 + P_i P_{i-1}(\varepsilon_1 + \varepsilon_2) + P_i(1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_5 + \varepsilon_6)E(P_{i-1}) \tag{47}$$

For the convergence of the error, the absolute value of the coefficient must satisfy:

$$(1+N)(1+\varepsilon_4)P_i < 1 \tag{48}$$

When $i \gg 1$, the value of $P_i \to 1$. Since $E(P_{i-1}) \sim i\varepsilon$, the value of N is determined by the term:

$$P_i(1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_5 + \varepsilon_6)E(P_{i-1}) \approx iP_i\varepsilon \tag{49}$$

Combining the expression of P_i [Eq. (36)] we have:

$$i < \sqrt{\frac{Bi_1}{Bi_1 + 1}\frac{1}{2\grave{o}}} \tag{50}$$