Abstract—An important task in solving second order linear ordinary differential equations by the finite difference method is to choose a suitable stepsize $h$. In this paper, by using the stochastic arithmetic, the CESTAC method and the CADNA library we present a procedure to estimate the optimal stepsize $h_{opt}$, the stepsize which minimizes the global error consisting of truncation and round-off error.

Keywords—ordinary differential equations, optimal stepsize, error, stochastic arithmetic, CESTAC, CADNA.

I. INTRODUCTION

NUMERICAL algorithms which include a stepsize are affected by a global error, which consists of both a truncation error and a round-off error. In these algorithms as the stepsize decreases, the truncation error also decreases, but the round-off error may increase. The problem is now to find the stepsize which minimizes the global error. In general, it is difficult to estimate the optimal stepsize in an algorithm. In [4], Chesneaux and Jézéquel showed that by using the CESTAC (Contrôle et Estimation Stochastique des Arrondis de Calcul) method [19], [20] and the CADNA library [2], [3], [10], [15], one can estimate the optimal stepsize for the numerical computation of integrals using the trapezoidal and Simpson’s rules. Then, Abbasbandy and Araghi [1] developed this method to general closed Newton-Cotes integration rules. The development of the method to multiple integrals can be found in [11]. In [14], Salkuyeh et al. proposed a procedure with stepsize control for solving $n$ one-dimensional initial value problems.

In this paper, we present a strategy to control the stepsize in the finite difference method for solving the following linear two-point boundary value problem

$$\begin{cases}
y''(x) + f(x)y'(x) + g(x)y(x) = r(x), & x \in (a, b), \\
y(a) = \alpha, & y(b) = \beta.
\end{cases}$$

(1)

This paper is organized as follows. In section II, a brief description of the finite difference method for solving (1) and our main results are given. Section III is devoted to a review of the stochastic round-off error analysis, the CESTAC method and the CADNA library. In section IV, some numerical experiments are given. Section V is devoted to some concluding remarks.

D. K. Salkuyeh is with the Department of Mathematics, University of Mohaghegh Ardabili, P. O. Box. 179, Ardabil, Iran, e-mail: khojaste@uma.ac.ir, salkuyeh@yahoo.com.

II. MAIN RESULTS

In order to solve the problem (1) by the finite difference method, we subdivide the interval $[a, b]$ into $n$ equal subintervals by the grid points $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$ where $h = (b-a)/n$. At the internal grid points $x_i$, $i = 1, 2, \ldots, n - 1$, we replace $y''(x_i)$ and $y'(x_i)$ in the differential equation (1) by the difference quotients

$$y''(x_i) = \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} + O(h^2),$$

and

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_{i-1})}{2h} + O(h^2).$$

Then, by neglecting $O(h^2)$ we obtain the system of linear equations

$$\begin{align*}
(1 - \frac{h}{2} f_i) y_{i-1} + \left(-2 + g_i h^2\right) y_i &+ \left(1 + \frac{h}{2}\right) y_{i+1} = h^2 r_i, \\
i &\in 1, 2, \ldots, n - 1.
\end{align*}$$

(2)

where $f_i = f(x_i)$, $g_i = g(x_i)$, $r_i = r(x_i)$ and $y_i \approx y(x_i)$. This system has to be complemented by the two boundary conditions $y_0 = \alpha$ and $y_n = \beta$. The matrix of system (2) is tridiagonal and one can solve it by the known Thomas algorithm [13]. By solving system (2), the approximate values $y_i$ to the exact solution $y(x_i)$ are obtained. We recall the following theorem concerning the convergence order of the method.

Theorem 1. ([12]) Suppose that the problem (1) has a unique solution $y$,

$$h|p(x)| < 2, \quad x \in [a, b],$$

and

$$q(x) \leq \gamma, \quad x \in [a, b],$$

for some $\gamma < 0$. Moreover, assume that $y$ is four times continuously differentiable. Then

$$y_i - y(x_i) = ch^2 + O(h^3),$$

(3)

where $c$ is a constant independent of $h$.

Now, we recall the following definition [4], [9].

Definition 1. Let $r$ and $s$ be two real numbers. The number of exact significant digits that are common to $r$ and $s$ can be defined in $(-\infty, +\infty)$ by
1. for \( r \neq s \),
\[
C_{r,s} = \log_{10} \frac{r + s}{2(r - s)}.
\]
2. \( C_{r,r} = +\infty \).

Next, we state and prove the following theorem.

**Theorem 2.** Let all the assumptions of Theorem 1 hold. Moreover, assume that \( y_i(h) \) and \( y_i(h/2) \) are the computed solutions obtained by the finite difference method at grid point \( x_i \) with stepsizes \( h \) and \( h/2 \), respectively. Then

\[
C_{y_i(h), y_i(h/2)} = C_{y_i(h), y(x_i)} + \log_{10} \frac{4}{3} + O(h).
\]

**Proof.** From (3), we have

\[
y_i(h) + y(x_i) = 2y(x_i) + ch^2 + O(h^3), \tag{4}
\]

\[
y_i(h) - y_i(h/2) = \frac{3}{4} ch^2 + O(h^3), \tag{5}
\]

\[
y_i(h) + y_i(h/2) = 2y(x_i) + \frac{5}{4} ch^2 + O(h^3). \tag{6}
\]

Then, from Definition 1, we obtain

\[
C_{y_i(h), y_i(h/2)} - C_{y_i(h), y(x_i)} = \log_{10} \left| \frac{y_i(h) + y_i(h/2)}{2(y_i(h) - y_i(h/2))} \right| - \log_{10} \left| \frac{y_i(h) + y(x_i)}{2(y_i(h) - y(x_i))} \right|
\]

\[
= \log_{10} \left| \frac{y_i(h) + y_i(h/2)}{y_i(h) - y_i(h/2)} \right| + \log_{10} \left| \frac{y_i(h) + y(x_i)}{y_i(h) - y(x_i)} \right| - \log_{10} \frac{2}{3} - \log_{10} \frac{2}{3}.
\]

From (3), we have

\[
\log_{10} \left| \frac{y_i(h) + y_i(h/2)}{y_i(h) - y_i(h/2)} \right| = \log_{10} \left| \frac{2y(x_i) + \frac{3}{4} ch^2 + O(h^3)}{2y(x_i) + \frac{5}{4} ch^2 + O(h^3)} \right| = O(h^2). \tag{7}
\]

On the other hand, we have

\[
\log_{10} \left| \frac{y_i(h) + y_i(h/2)}{y_i(h) - y_i(h/2)} \right| = \log_{10} \left| \frac{c + O(h)}{c + \frac{3}{4} O(h)} \right| = \log_{10} \frac{4}{3} + \log_{10} \frac{4}{3} + O(h). \tag{9}
\]

Substituting (8) and (9) in (7) the desired relation is obtained.

This theorem shows, if \( h \) is small enough, then the number of common significant digits between \( y_i(h) \) and \( y_i(h/2) \) is the same as the number of the common significant digits between \( y_i(h) \) and \( y(x_i) \) up to less than one digit. It is necessary to mention that \( \log_{10} \frac{4}{3} \approx 0.1249 \). Theorem 2, has been stated by taking into account only the truncation error on two approximate solutions \( y_i(h) \) and \( y_i(h/2) \). But, as we know the computed results are also affected by round-off error propagation. Next, we describe how round-off errors can be estimated with a probabilistic approach in order to determine the exact significant digits of any computed result.

### III. The CESTAC Method

Each result provided by a numerical algorithm which is performed on a computer always contains an error resulting from round-off error propagation. The Discrete Stochastic Arithmetic (DSA) [16] is a probabilistic approach for round-off error propagation. With the DSA, which is the joint use of the synchronous implementation of CESTAC method and the stochastic order relations, it is possible to estimate the accuracy of the results provided by a computer, to detect the numerical instabilities occurring during the run of a scientific code, and to check the branchings that exist in the code. In this section, we briefly review the CESTAC method and DSA.

#### A. Brief description of the CESTAC method and its implementation

The main idea of the CESTAC method is defined in [18], [20] and consists of:

- synchronously performing the same code \( N \) times with a different round-off error propagation for each run,
- estimating the common part of these results and to consider that this part is representative of the exact result.

In practice, these different round-off error propagations are obtained in using random rounding mode.

Indeed, each result \( r \) of a floating-point operation which is not an exact floating-point value is always bounded by two floating-points values \( R^- \) and \( R^+ \), each of them being so representative of the exact result.

The random rounding consists at the level of each floating-point operation or assignment to choose as result randomly with an equal probability either \( R^- \) or \( R^+ \). Then when the same code is executed \( N \) times with a computer using this random rounding, \( N \) results \( R_k \), \( k = 1, \ldots, N \) are obtained. It has been proved in [5], [6] that, under some hypotheses, these \( N \) results belong to a quasi-Gaussian distribution centered on the exact result \( r \). So, in practice, by considering the mean value \( \overline{R} \) of the \( R_k \) as the computed result, and using Student’s test, it is possible to obtain a confidence interval of \( \overline{R} \) with a probability \((1 - \beta)\) and then to estimate the number of exact significant digits of \( \overline{R} \) by the formula (10)

\[
C_{\overline{R}} = \log_{10}(\sqrt{N}\overline{R}/\tau_r \sigma), \tag{10}
\]

with \( \overline{R} = (1/N) \sum_{i=1}^{N} R_i \) and \( \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (R_i - \overline{R})^2 \). \( \tau_r \) is the value of the Student distribution for \( N - 1 \) degrees of freedom and a probability level \( 1 - \beta \). In practice \( N = 3, \beta = 0.05 \) and then \( \tau_r = 4.343003 \).

The result provided by equation (10) is reliable if the hypotheses underlying the method hold in practice. It has been proved that [5], [6], [21], these hypotheses hold when:

1) The operands of any multiplication are both significant.
2) The divisor of any division is significant.

It is then absolutely necessary during the run of a code to control the points 1) and 2). This control is done with the concept of computational zero also named computational zero or computed zero [22].
Definition 2. Each result provided by CESTAC method is a computational zero denoted by \( \oplus.0 \) if and only if one of the two conditions holds:

1) \( \forall i, i = 1, \ldots, N, R_i = 0 \).
2) \( C_{\text{R}} \leq 0 \), \( (C_{\text{R}} \text{ obtained with equation (10)}) \).

When \( C_{\text{R}} \leq 0 \), then \( \text{R} \) is an insignificant value (\( \text{R} \) has no significant digit). From the concept of \( \oplus.0 \), discrete stochastic relations (DSR) have been defined (equality and order relations).

Definition 3. Let \( X \) and \( Y \) be \( N \)-samples provided by CESTAC method, discrete stochastic equality denoted by \( s = \) is defined as:

\[
X_s = Y \quad \text{if} \quad X - Y = \oplus.0.
\]

Definition 4. Let \( X \) and \( Y \) be \( N \)-samples provided by CESTAC method, discrete stochastic inequalities denoted by \( s > \) and \( s \geq \) are defined as:

\[
X_s > Y \quad \text{if} \quad X > Y \quad \text{and} \quad X - Y \neq \oplus.0.
\]

\[
X_s \geq Y \quad \text{if} \quad \exists \text{a real} \quad \text{Xs} \quad \text{such that} \quad X - Y = \oplus.0.
\]

The DSA is the association of the CESTAC method, the concept of computational zero and the discrete stochastic relations (see [7], [8], [21]). With this DSA it is possible to control the run of a scientific code, to detect the numerical instabilities and the violation of the hypotheses underlying the method. But in practice how to implement this?

As we observed, the two main specificities denoted by the CESTAC method are:

- The random rounding, which consists in creating \( R^- \) and \( R^+ \) and in choosing randomly one or the other.
- The manner to perform the \( N \) runs of a code.

With IEEE arithmetic and the possibilities of ADA, C++, and Fortran to create new structures and to overload the operators it is easy to implement the CESTAC method.

The random rounding uses the IEEE rounding toward \(+\infty\) and toward \(-\infty\). These roundings occur whenever an arithmetic operation has a result that is not exact. Then no artificial round-off error is introduced in the computation. The choice of the rounding is at random with an equal probability for the \((N-1)\) first samples and the choice of the last one is the opposite of the choice of the \((N-1)\)th sample. With this random rounding the theorems on exact rounding are respected.

We have seen previously that it is absolutely necessary to detect, during the run of a code, the emergence of \( \oplus.0 \) for controlling the validity of the CESTAC method. To achieve this it suffices to use the synchronous implementation which consists in performing each arithmetic operation \( N \) times with the random rounding before performing the next. Thus for each numerical result we have \( N \) samples, from which with equation (10) the number of exact significant digits of the mean value, considered as the computed result, is estimated.

With this implementation the stochastic order relations defined above may also be easily created. Then during the run of a code a dynamic control may be done.

B. The CADNA library

The CADNA software [2], [3] is a library which implements automatically the DSA in any code written in Fortran. Using the CADNA library (Control of Accuracy and Debugging for Numerical Application), each standard FP types have their corresponding stochastic types. Every intrinsic function and operator are overloaded for those types. When a stochastic variable is printed, only its significant digits are displayed to point out its accuracy. If a number has no significant digit (i.e., a computed zero), the symbol \( \oplus.0 \) is displayed.

The modifications that the user has to do in his Fortran source are mainly to change the declaration statements of real type by stochastic type, and the input-output statement (see [3]). Thus, when a modified Fortran source combined with the CADNA library is run, it is as \((N = 3)\) identical codes were simultaneously run on \( N \) synchronized computers each of them using the random rounding mode. So round-off error propagation can be analyzed step by step and then any numerical anomaly can be dynamically detected. This leads to the self validation of the method and a numerical debugging scientific code.

In the next section, we describe how the use of the CADNA library allows us to control the stepsize in the method of the finite difference for solving (1).

IV. NUMERICAL EXPERIMENTS

According to the previous section and Theorem 2, by using the CESTAC method and the CADNA library, we propose a procedure for computing an approximate optimal stepsize in solving (1) by the finite difference method as follows. Let

\[
h_m = \frac{b - a}{2^m}, \quad m = 1, 2, 3, \ldots,
\]

and

\[
x_i = x + ih, \quad i = 0, 1, \ldots, 2^m.
\]

Let also

\[
Y^{(m)} = (y_1(h_m), y_2(h_m), \ldots, y_{2^m-1}(h_m)),
\]

be the approximate solution computed from (2) with \( n = 2^m \).

In this case, the following algorithm, AOSFDM (for approximate optimal stepsize in the finite difference method), is proposed.

Algorithm 1. AOSFDM

1. \( m := 1 \).
2. Compute \( Y^{(m)} \) and \( Y^{(m+1)} \).
3. Set \( Y^{(m)} = (Y_2^{(m+1)}, Y_4^{(m+1)}, \ldots, Y_{2(m+1)-1}^{(m+1)}) \).
4. If \( \|Y^{(m)} - Y^{(m+1)}\|_\infty = \oplus.0 \), then stop (\( h_{m+1} \) is an approximate optimal stepsize).
5. Else \( m := m + 1 \) and goto 2.

Step 4 of this algorithm means that if

\[
|Y_i^{(m)} - Y_i^{(m+1)}| = \oplus.0, \quad i = 1, 2, \ldots, 2^m - 1,
\]

then the process is stopped. In this case, the number of common significant digits between \( Y_i^{(m)} \) and \( Y_i^{(m+1)} \) is the same.
as the number of the common significant digits between $Y^{(m)}_i$ and $y(x)_i$, up to less than one digit. It is necessary to mention that if

$$|Y_i^{(m)} - \tilde{Y}_i^m| = 0, i = 1, 2, \ldots, 2^m - 1,$$

then the transformation from $Y^{(m)}_i$ into $Y^{(m+1)}_i$ is only due to the round-off errors and further reduction in the stepsize would be useless and $h_{m+1}$ can be considered as an approximate optimal stepsize.

Now we present the examples and the results which we obtained by the Fortran code of the finite difference methods for solving (1) combined with the CADNA library, version BETA [15]. All the numerical experiments were computed in double precision. We consider three examples and for each example, the numerical results are given in two separate tables. In the first table of each example the values of $\|Y^{(m)} - Y^{(m)}\|_\infty$ and $\|Y^{(m)} - Y^{(m)}_e\|_\infty$ for different values of $m$ are given, where

$$Y^{(m)}_i = (y(x)_i, y(x)_2, \ldots, y(x)_{2^m-1}),$$

and in the second table, the approximate solution obtained with the approximate optimal stepsize and the exact solution in some grid points are compared.

### Example 1.

In this example, we consider

$$\begin{align*}
y'' + (x + 1)y' - 3y &= 3x^2 + 4x + 1, \\
y(1) &= 2, \quad y(2) = 10.
\end{align*}$$

The exact solution of this problem is $y(x) = x^3 + x$. Numerical results are given in TABLE I and TABLE II. TABLE I shows that the approximate optimal stepsize is

$$h_{13} = \frac{2 - 1}{2^13} = 1.2207e - 4.$$

The approximate solution computed with $h_{13}$ together with the exact solution in some grid points are given in TABLE II. As we observe, all of the digits of $y^{(14)}_i$, up to one, coincide with that of the $y(x)_i$. Moreover, we see that by the CADNA library one can represent the computed solution with their exact decimal figures.

### Example 2.

This example is devoted to

$$\begin{align*}
y'' + xy' - (1 + x^2)y &= x \cos x - (2 + x^2) \sin x, \\
y(0) &= 0, \quad y(\frac{\pi}{2}) = 1.
\end{align*}$$

### Example 3.

In this example, we consider

$$\begin{align*}
y'' + e^{-x}y' - y &= e + 2e^x, \\
y(0) &= -1, \quad y(1) = 0.
\end{align*}$$

The exact solution of this problem is $y(x) = (x - 1)e^x$. Numerical results are given in TABLE V and TABLE VI. TABLE V gives

<table>
<thead>
<tr>
<th>$i$</th>
<th>$y(x)_i$</th>
<th>$y^{(13)}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2009767413284E+001</td>
<td>0.2009767413E+001</td>
</tr>
<tr>
<td>500</td>
<td>0.2534801123805E+000</td>
<td>0.2534801127E+000</td>
</tr>
<tr>
<td>1000</td>
<td>0.3169928349543E+001</td>
<td>0.316992835E+001</td>
</tr>
<tr>
<td>1500</td>
<td>0.391628728206E+001</td>
<td>0.3916287281E+001</td>
</tr>
<tr>
<td>2000</td>
<td>0.4784969598602E+001</td>
<td>0.4784969600E+001</td>
</tr>
<tr>
<td>2500</td>
<td>0.5786367014802E+001</td>
<td>0.578636701E+001</td>
</tr>
</tbody>
</table>

The approximate solution computed with $h_{14}$ together with the exact solution in some grid points are given in TABLE IV. As we observe, all of the digits of $y^{(14)}_i$, up to one, coincide with that of the $y(x)_i$.
The approximate optimal stepsize as
\[ h_{14} = \frac{2 - 1}{2^{14}} = 6.1035e - 5. \]

The approximate solution computed with \( h_{14} \) together with the exact solution in some grid points are given in TABLE V. We observe, all of the digits of \( y_i^{(14)} \), up to one, coincide with that of the \( y(x_i) \). We also see that the CADNA library has furnished the computed solutions with their exact digits.

\[ \text{TABLE V} \]

<table>
<thead>
<tr>
<th>( m )</th>
<th>( |Y^{(m)} - Y^{(m)}|_{\infty} )</th>
<th>( |Y^{(m)} - Y^{(m)}|_{\infty} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.39586380275902E+00</td>
<td>0.40242751944880E+00</td>
</tr>
<tr>
<td>2</td>
<td>0.4905161723882E+00</td>
<td>0.6558972435498E+00</td>
</tr>
<tr>
<td>3</td>
<td>0.1239470101500E+00</td>
<td>0.1653810711577E+00</td>
</tr>
<tr>
<td>4</td>
<td>0.3109982505E+00</td>
<td>0.3521826598E+00</td>
</tr>
<tr>
<td>5</td>
<td>0.777271095E+00</td>
<td>0.10293215E+00</td>
</tr>
<tr>
<td>6</td>
<td>0.1944888E+00</td>
<td>0.24811727E+00</td>
</tr>
<tr>
<td>7</td>
<td>0.4862426E-05</td>
<td>0.6161044E-05</td>
</tr>
<tr>
<td>8</td>
<td>0.12156E-004</td>
<td>0.151709E-005</td>
</tr>
<tr>
<td>9</td>
<td>0.3039E-006</td>
<td>0.3393E-006</td>
</tr>
<tr>
<td>10</td>
<td>0.795E-007</td>
<td>0.8311E-007</td>
</tr>
<tr>
<td>11</td>
<td>0.19E-007</td>
<td>0.18E-007</td>
</tr>
<tr>
<td>12</td>
<td>0.4E-008</td>
<td>0.5E-008</td>
</tr>
<tr>
<td>13</td>
<td>@.0</td>
<td>@.0</td>
</tr>
</tbody>
</table>

\[ \text{TABLE VI} \]

<table>
<thead>
<tr>
<th>( i )</th>
<th>( y(x_i) )</th>
<th>( y_i^{(14)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.999999999254812E+00</td>
<td>-0.99999999254E+00</td>
</tr>
<tr>
<td>1000</td>
<td>-0.991914407570705E+00</td>
<td>-0.99191440E+00</td>
</tr>
<tr>
<td>2000</td>
<td>-0.964872503584451E+00</td>
<td>-0.96487250E+00</td>
</tr>
<tr>
<td>3000</td>
<td>-0.9140885656060E+00</td>
<td>-0.91408852E+00</td>
</tr>
<tr>
<td>4000</td>
<td>-0.833852399520346E+00</td>
<td>-0.83385239E+00</td>
</tr>
<tr>
<td>5000</td>
<td>-0.717337376111629E+00</td>
<td>-0.71733737E+00</td>
</tr>
<tr>
<td>6000</td>
<td>-0.556592544770840E+00</td>
<td>-0.55659254E+00</td>
</tr>
<tr>
<td>7000</td>
<td>-0.34196964192728E+00</td>
<td>-0.34196964E+00</td>
</tr>
<tr>
<td>8000</td>
<td>-0.62233896040408E-01</td>
<td>-0.62233896E-01</td>
</tr>
<tr>
<td>8191</td>
<td>-0.331781009179E-03</td>
<td>-0.33178100E-03</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this paper, a theorem has been stated to provide a stopping criterion to control the stepsize in the finite difference method for solving the linear two-point boundary value problems. We observed that the use of the CESTAC method and the CADNA library allows us to estimate the optimal stepsize. Numerical examples show that the proposed method is effective.

REFERENCES


Davod Khojasteh Salkuyeh received his B.Sc from Sharif University of Technology, Tehran, Iran and his M.Sc from Ferdowsi University of Mashhad, Mashhad, Iran. He received his Ph.D degree under supervision of professor Faezeh Toutounian at Ferdowsi University of Mashhad in 2003. He is currently an assistant professor of Mathematics at University of Mohaghegh Ardabili, Ardabil, Iran. His research interests are mainly iterative methods for large sparse linear systems of equations and error analysis.