NUMERICAL ACCURACY OF A CERTAIN CLASS OF ITERATIVE METHODS FOR SOLVING LINEAR SYSTEM

DAVOD KHOJASTEH SALKUYEH AND FAEZEH TOUTOUNIAN

Abstract. One of the most important problems for solving the linear system $Ax = b$, by using the iterative methods, is to use a good stopping criterion and to determine the common significant digits between each corresponding components of computed solution and exact solution. In this paper, for a certain class of iterative methods, we propose a way to determine the number of common significant digits of $x_m$ and $x$, where $x_m$ and $x$ are the computed solution at iteration $m$ and exact solution, respectively. By using the CADNA library which allows us to estimate the round-off error effect on any computed result, we also propose a good stopping criterion which is able to stop the process as soon as a satisfactory informatical solution is obtained. Numerical examples are used to show the good numerical properties.

AMS Mathematics Subject Classification : 65F10, 65G50.
Key words and phrases: Iterative methods, significant digits, FOM algorithm, error propagation, CESTAC method, stochastic arithmetic, CADNA library

1. Introduction

Consider the linear system

$$Ax = b,$$

where $A$ is a real nonsingular matrix of order $N$. For solving Eq. (1) by an iterative method, one can use the common strategy for stopping the iterations. For example, for a given tolerance $\epsilon > 0$, we can use the following stopping criterions

1. $\|x_n - x_{n-1}\| < \epsilon,$
2. $\|b - Ax_n\| < \epsilon,$

where $\| . \|$ is a vector norm and $\{x_n\}$ is a sequence of vectors such that

$$\lim_{n \to \infty} x_n = x = A^{-1}b.$$ 

If one of the above stopping criterions is used then the number of signficants digits that are common to corresponding entries of $x_n$ and $x$ can not be specified.
Another problem is to choose the value \( \epsilon \). When \( \epsilon \) is chosen too large, then the iterative process is stopped too soon, and consequently the approximate solution has a poor accuracy. On the contrary, when \( \epsilon \) is chosen too small, it is possible, due to the numerical instabilities, that many useless iterations are performed without improving the accuracy of the solution. The aim of this paper is to solve these problems.

This paper is organized as follows. In section 2, two theorems are proved in order to derive a lower bound for the common significant digits of each corresponding components of \( x_m \) and \( x \). In section 3, a brief description of stochastic round-off analysis, the CESTAC method and the CADNA software is described. In section 4, by using the theorems of section 2 and the CADNA library which allows us to estimate the round-off error effect on any computed result, we propose a good stopping criterion which is able to stop the process as soon as a satisfactory informatical solution is obtained and to estimate the number of significant digits of its components. In section 4, some numerical results are also given to show the good numerical properties.

2. Theoretical results

2.1. Preliminaries

As in [11], we recall the following definition

**Definition 1.** Let \( p \) and \( q \) be two real numbers, the number of significant digits that are common to \( p \) and \( q \) can be defined in \((−\infty, +\infty)\) by

1. for \( p \neq q \),
   \[
   C_{p,q} = \log_{10} \left| \frac{p + q}{2(p - q)} \right| .
   \]
2. \( \forall p \in \mathbb{R}, C_{p,p} = +\infty \).

For two vectors \( a, b \in \mathbb{R}^n \) we define the following number

1. for \( a \neq b \),
   \[
   C_{a,b} = \log_{10} \frac{\| a + b \|_2}{2\sqrt{n} \| a - b \|_2} .
   \] (3)
2. \( \forall a \in \mathbb{R}^n, C_{a,a} = +\infty \).

Now, in the following lemma, we shall develop the important property of the number \( C_{a,b} \).

**Lemma 1.** If \( a = (a_i) \) and \( b = (b_i) \) are two vectors in \( \mathbb{R}^n \) and

\[
\frac{|a_i + b_i|}{\| a + b \|_\infty} = \alpha_i \times 10^{-\beta_i} ,
\] (4)
where \(0.1 \leq \alpha_i < 1\) and \(\beta_i \geq -1\), for \(i = 1, \ldots, n\), then
\[C_{a_i, b_i} \geq C_{a, b} - (\beta_i + 1).\]

**Proof.** By using (3), we can write
\[|a_i - b_i| \leq \|a - b\|_2 = \frac{\|a + b\|_2}{2\sqrt{n}} 10^{-C_{a, b}} \leq \frac{\|a + b\|_\infty}{2} 10^{-C_{a, b}},\]
in which, for the latter inequality we used the inequality
\[\|x\|_2 \leq \sqrt{n} \|x\|_\infty, \quad \forall x \in \mathbb{R}^n.\]

By using the hypothesis (4), relation (5) can be written as follows
\[|a_i - b_i| \leq \frac{|a_i + b_i|}{2\alpha_i} 10^{-C_{a, b} + \beta_i} \leq \frac{|a_i + b_i|}{2} 10^{-(C_{a, b} - (\beta_i + 1))}.\]

Therefore, we have
\[C_{a_i, b_i} = \log_{10} \frac{|a_i + b_i|}{2(a_i - b_i)} \geq C_{a, b} - (\beta_i + 1).\]

**Remark 1.** If we define \(\beta_{\text{max}} = \max_{1 \leq i \leq n} \beta_i\), which corresponds to the component with smallest magnitude, then
\[C_{a_i, b_i} \geq C_{a, b} - (\beta_i + 1) \geq C_{a, b} - (\beta_{\text{max}} + 1), \quad \text{for} \ i = 1, \ldots, n.\]

This shows that all the corresponding components of \(a\) and \(b\) have at least \(C^* = C_{a, b} - (\beta_{\text{max}} + 1)\) significant digits common to each other.

### 2.2. Numerical accuracy of a certain class of iterative methods for solving linear systems

Now, we begin this subsection by a theorem which will be useful for computing the common significant digits of each corresponding component of the computed solution and exact solution for a linear system using an iterative method.

**Theorem 1.** Suppose that \(x_{m+1} = Mx_m + c\) is a convergence iterative method to the exact solution \(x\) of the system (1) with \(c \neq 0\). Then, for sufficiently large value of \(m\), we have
\[C_{x_m, x} - C_{x_m, x_{m+1}} = z_m,\]
where
\[\log_{10} \left|1 - \|M\|_2 \right| \lesssim z_m \lesssim \log_{10} (1 + \|M\|_2).\]
Proof. Let $e_m = x_m - x$. Then one can see

$$x_m - x = M^m(x_0 - x) = M^m e_0,$$

and

$$x_m - x_{m+1} = M^m(I - M)e_0.$$  \hspace{1cm} (7)

On the other hand, we observe that

$$x_m - x_m + 1 = M^m(x_0 + x_1) + 2c.$$ \hspace{1cm} (8)

By induction, we have

$$x_m + x_{m+1} = M^m e_0 + T_m,$$ \hspace{1cm} (9)

where

$$T_m = M^m(x_0 + x) + 2(I - M)^{-1}(I - M^m)c.$$  

From the hypotheses, we can conclude that

$$\lim_{m \to \infty} \| T_m \|_2 = \| 2(I - M)^{-1}c \|_2 \neq 0.$$ \hspace{1cm} (10)

In the same way, we can show that

$$x_m + x = T_m.$$ \hspace{1cm} (11)

Now from Eqs. (7)-(9) and Eq. (11) we conclude

$$C_{x_m, x}C_{x_m, x_{m+1}} = \frac{\log_{10} \| T_m \|_2}{2\sqrt{n} \| M^m e_0 \|_2} - \frac{\log_{10} \| M^m e_0 + T_m \|_2}{2\sqrt{n} \| M^m e_0 + T_m \|_2} = \frac{\log_{10} \| T_m + M^m e_0 \|_2}{\| T_m + M^m e_0 \|_2} + \log_{10} \frac{\| (I - M)M^m e_0 \|_2}{\| M^m e_0 \|_2},$$ \hspace{1cm} (12)

For the first term of this equation, we have

$$\lim_{m \to \infty} \log_{10} \frac{\| T_m \|_2}{\| T_m + M^m e_0 \|_2} = 0,$$ \hspace{1cm} (13)

and for the second term we can write

$$\| (I - M)M^m e_0 \|_2 \leq \| I - M \|_2 \leq 1 + \| M \|_2,$$

and also

$$\frac{\| (I - M)M^m e_0 \|_2}{\| M^m e_0 \|_2} \geq \frac{\| M^m e_0 \|_2 - \| M \|_2 \| M^m e_0 \|_2}{\| M^m e_0 \|_2} = 1 - \| M \|_2.$$
Therefore

$$\log_{10}|1-\|M\|_2| \leq \log_{10}\frac{\| (I-M)M^m e_0 \|_2}{\| M^m e_0 \|_2} \leq \log_{10}(1+\|M\|_2). \quad (14)$$

Finally from Eqs. (12)-(14) the desired relation is obtained. □

2.3. Numerical accuracy of FOM for solving linear systems

In this section we consider the Full Orthogonalization Method (FOM) (see [16]), which is an iterative method for solving the linear systems. As previous section, we establish a theorem which provides a similar result for the iterative method FOM algorithm. First we recall some fundamental properties of FOM algorithm. In the FOM algorithm the Arnoldi method is used for the construction of an orthogonal basis \{v_1, \ldots, v_m\} for the Krylov subspace \mathcal{K}^m(A; r_0) which is defined by

$$\mathcal{K}^m(A; r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}.$$ 

The Arnoldi-Modified Gram-Schmidt algorithm can be described as follows [16].

Algorithm 1. Arnoldi-Modified Gram-Schmidt

1. Choose \(x_0\) and compute \(r_0 = b - Ax_0\) and \(v_1 = r_0 / \| r_0 \|_2\).
2. For \(j=1, 2, \ldots, m\) Do:
3. Compute \(w_j := Av_j\)
4. For \(i=1, \ldots, j\) Do:
5. \(h_{ij} := (w_j, v_i)\)
6. \(w_j := w_j - h_{i,j} v_i\)
7. EndDo
8. \(h_{j+1} := \| w_j \|_2\). If \(h_{j+1} = 0\), then Stop
9. \(v_{j+1} := \frac{w_j}{h_{j+1}}\)
10. EndDo

Let \(V_m = [v_1, v_2, \ldots, v_m]\) and \(H_m\) be the \(m \times m\) Hessenberg matrix whose nonzero entries \(h_{ij}\) are defined by Algorithm 1, then the following relation holds

$$V_m^T AV_m = H_m.$$

The idea of FOM algorithm is to construct an approximate solution of the form \(x_m = x_0 + z_m\), where \(z_m\) is an element of \(\mathcal{K}^m(A; r_0)\) with the property \(r_m(= b - Ax_m) \perp \mathcal{K}^m(A; r_0)\). This gives the following algorithm.

Algorithm 2. Restarted FOM (FOM(m))

1. Compute \(r_0 = b - Ax_0\), \(\beta = \| r_0 \|_2\) and \(v_1 = r_0 / \beta\)
2. Generate the Arnoldi basis and matrix \(H_m\) using the Arnoldi algorithm starting with \(v_1\)
3. Compute \( y_m = H_m^{-1} (\beta e_1) \) and \( x_m = x_0 + V_m y_m \).
   If convergence test satisfied then stop.
4. Set \( x_0 = x_m \) and go to 1.

In order to be able to insert a good stopping criterion in the code of the above algorithm and to determine the number of common significant digits between corresponding components of computed solution and exact solution, we propose the following algorithm which differs a little from the restarted FOM algorithm.

**Algorithm 3. FOM-like method**

1. Select an initial guess \( x_0 \) and set \( m = 1 \)
2. Compute \( r_{m-1} = b - Ax_{m-1} \), \( \beta = \| r_{m-1} \|_2 \) and \( v_1 = r_{m-1}/\beta \)
3. Generate the Arnoldi basis and matrix \( H_m \) using the Arnoldi algorithm starting with \( v_1 \)
4. Compute \( y_m = H_m^{-1} (\beta e_1) \) and \( x_m = x_{m-1} + V_m y_m \).
   If convergence test satisfied then stop.
5. Set \( m = m+1 \) and go to 2.

Now, we state and prove the following theorem

**Theorem 2.** Let \( A \) be a real symmetric positive definite (SPD) matrix. If \( x_m \) for \( m \geq 1 \), be the computed solution at iteration \( m \) using the algorithm 3, then for sufficiently small residual norm at iteration \( m \), we have

\[
C_{x_m, x_{m-1}} - C_{x_{m-1}, x} = z_m
\]

where

\[
-\log_{10}(\text{cond}(A)) \lesssim z_m \lesssim \log_{10}(\text{cond}(A)).
\]

**Proof.** From step 4 of Algorithm 3 and relation \( x_m - x = -A^{-1} r_m \), it can be easily seen that

\[
C_{x_m, x_{m-1}} - C_{x_{m-1}, x} = \log_{10} \frac{\| 2x_{m-1} + V_m y_m \|_2}{\sqrt{n} \| V_m y_m \|_2} - \log_{10} \frac{\| 2x - A^{-1} r_m \|_2}{\sqrt{n} \| A^{-1} r_m \|_2}
\]

\[
= \log_{10} \frac{\| 2x_{m-1} + V_m y_m \|_2}{\| 2x - A^{-1} r_m \|_2} + \log_{10} \frac{\| A^{-1} r_m \|_2}{\| V_m y_m \|_2},
\]

(15)

Suppose that the eigenvalues of \( A \) are labelled in decreasing order, i.e.,

\[
0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n,
\]

and the Ritz values in step \( m \) which are the eigenvalues of \( H_m \) are labelled as follows

\[
\theta^{(m)}_1 \leq \theta^{(m)}_2 \leq \cdots \leq \theta^{(m)}_m,
\]

then the Ritz values satisfy the relations (see [17])
Numerical Accuracy of a Certain Class of Iterative Methods

\[ \theta_{i}^{(m)} > \lambda_{i}, \quad \text{for} \quad i = 1, \ldots, m, \quad (16) \]

and

\[ \theta_{m+1-i}^{(m)} < \lambda_{n+1-i}, \quad \text{for} \quad i = 1, \ldots, m, \quad (17) \]

From these relations we can conclude that \( H_{m} \) is an SPD matrix. So, by using Eq. (16) and the definition of \( \beta \) we have

\[ \| V_{m}y_{m} \|_{2} = \| y_{m} \|_{2} = \| H_{m}^{-1}(\beta e_{1}) \|_{2} = \| r_{m} \|_{2} = \| H_{m}^{-1}e_{1} \|_{2} = \| H_{m}^{-1}e_{1} \| \]

\[ \leq \| r_{m} \|_{2} \lambda_{\max}(H_{m}^{-1}) \leq \| r_{m} \|_{2} / \lambda_{1}, \]

We observe that when the residual norm is sufficiently small the first term of the Eq. (15) is negligible.

On the other hand, for the second term of Eq. (15), we have

\[ \| A^{-1}r_{m} \|_{2} = \| A^{-1}r_{m} \|_{2} \| e_{1} \|_{2} \| H_{m}^{-1}e_{1} \|_{2} \]

\[ \leq \| r_{m} \|_{2} \lambda_{\max}(H_{m}^{-1}) \leq \| r_{m} \|_{2} / \lambda_{1}, \]

(18)

with

\[ 0 < \frac{1}{\lambda_{n}} \leq \| A^{-1}r_{m} \|_{2} \leq \frac{1}{\lambda_{1}}, \quad (19) \]

because \( A \) is an SPD matrix. From Eqs. (16) and (17), it follows that

\[ 0 < \lambda_{1} \leq \| e_{1} \|_{2} / \| H_{m}^{-1}e_{1} \|_{2} \leq \lambda_{n}, \quad (20) \]

Finally, Eqs. (18)-(20) yield the desired result. \( \square \)

3. The CESTAC method and the CADNA library

3.1. Introduction

The probabilistic approach to round-off errors [3,10-13] consists in identifying each round-off error, due to the affection of a variable or an elementary arithmetic operation, to a random variable. Moreover, these different variables are supposed to be independent and identically distributed.

The aim of the CESTAC method [18-20,22], based on this probabilistic approach, is to estimate the effect of propagation of the round-off errors on every computed result obtained with the floating point arithmetic. It consists in making the round-off errors propagate in different ways in order to distinguish between a stable part of mantissa (considered as the significant one) and an unstable part (non-significant).

The different propagations are obtained by changing randomly the last bit of the mantisa of each intermediate computed result. In this way, a random arithmetic is generated. Then, by running the program several times in parallel,
a sample of the different values for each intermediate result is obtained. The mean value defines the computed value and Student’s test estimates its accuracy.

It has been proved [2-4,8] that, under certain regularity conditions, every computed result \( R \) obtained with the CESTAC method can be modelled by

\[
R = r + \sum_{i=1}^{n} u_i(d).2^{-p}.z_i,
\]

where \( u_i(d) \) are constants depending only on the data \( d \), \( p \) is the number of bits of the mantissa and the \( z_i \)s are independent identically distributed and centered random variables. The number of arithmetical operations is \( n \) and \( r \) is the exact mathematical result.

Consequently, each computed result can be modelled by a Gaussian random variable centered on the exact mathematical result. Its mean value is estimated from a sample using Student’s test. So, in practice, the use of the CESTAC method consists in

(i) Running in parallel \( N \) times ( \( N = 2 \) or \( 3 \) ) the program with this new arithmetic. Consequently, for each result \( R \) of any floating-point arithmetic operation, a set of \( N \) computed results \( R_i, i = 1,...,N \), is obtained.
(ii) Taking the mean value \( \bar{R} = \frac{1}{N} \sum_{i=1}^{N} R_i \) of the \( R_i \) as the computed result.
(iii) Using the Student distribution to estimate a confidence interval for \( R \), and then compute the number \( C_{\bar{R}} \) of significant digits of \( R \), defined by

\[
C_{\bar{R}} = \log_{10}(\frac{N^2R}{s^2}) \quad \text{with} \quad s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (R_i - \bar{R})^2,
\]

where \( \tau_\beta \) is the value of Student distribution for \( N-1 \) degrees of freedom and a probability level \( 1 - \beta \). If \( R_i = 0, i = 1,...,N \), or if \( C_{\bar{R}} \leq 0 \), then \( \bar{R} \) is an informatical zero denoted 0. This concept of informatical zero has been introduced by Vignes [21].

3.2. Stochastic arithmetic

By using the CESTAC method so that the \( N \) runs of the computer program take place in parallel, the \( N \) results of each arithmetic operation can be considered as realisations of a Gaussian random variable centered on the exact result. We can therefore define a new number, called stochastic number, and a new arithmetic, called stochastic arithmetic, applied to these numbers. We present below the main definitions and properties of this arithmetic. For more details see [9].

**Definition 1.** We define the set \( S \) of stochastic numbers as the set of Gaussian random variables. We denote an element \( X \in S \) by \( X = (\mu, \sigma^2) \), where \( \mu \) is the mean value of \( X \) and \( \sigma \) its standard deviation. If \( X \in S \) and \( X = (\mu, \sigma^2) \), there exists \( \lambda_\beta \), depending only on \( \beta \), such that

\[
P(X \in [\mu - \lambda_\beta \sigma, \mu + \lambda_\beta \sigma]) = 1 - \beta
\]
Numerical Accuracy of a Certain Class of Iterative Methods

$I_{\beta,x} = [\mu - \lambda_\beta \sigma, \mu + \lambda_\beta \sigma]$ is a confidence interval of $\mu$ at $(1 - \beta)$. An upper bound to the number of significant digits common to $\mu$ and each element of $I_{\beta,x}$ is

$$C_{\beta,X} = \log_{10}(\frac{|\mu|}{\lambda_\beta \sigma}).$$

The following definition is the modelling of the concept of informatical zero proposed in [21].

**Definition 2.** $X \in S$ is a stochastic zero if and only if

$$C_{\beta,X} \leq 0 \quad \text{or} \quad X = (0,0).$$

The stochastic arithmetic described above can be used in scientific codes to serve

1. during the run of a scientific code, to estimate the accuracy of a numerical result, to detect the numerical instabilities, and to check the branching;
2. to eliminate the programming expedients that are absolutely unfounded, such as those used, for example, in the termination criteria of iterative methods, and replace them by criteria that directly reflect the mathematical condition that must be satisfied at the solution.

### 3.3. The CADNA library

CADNA (Control of Accuracy and Debugging for Numerical Application) is a library for programs written in FORTRAN 77, FORTRAN 90, or in ADA which allows the computation using stochastic arithmetic by automatically implementing the CESTAC method [5,6]. CADNA is able to estimate the accuracy of the computed results, and to detect numerical instabilities occurring during the run. To use the CADNA library, it suffices to place the instruction USE CADNA at the top of the initial FORTRAN or ADA source code and to replace the declarations of the real type by the stochastic type and to change some statements such as printing statements.

During the run, as soon as a numerical anomaly (for example, appearance of the informatical zero in a computation or a criterion) occurs, a message is written in a special file called `Cadna-stability-f90.lst`. The user must consult this file after the program has run. If it is empty, this means the program has been run without any problem, that it has accordingly been validated, and that the results have been given with their associated accuracy. If it contains messages, the user, using the debugger associated with the compiler, will find the instructions that are the cause of these numerical anomalies, and must reflect in order to correct them if necessary. The program execution time using the CADNA library is only multiplied by a factor 3, which is perfectly acceptable in view of the major advantage offered, i.e., the validation of programs. CADNA
is also able to estimate the influence of data errors on the result provided by the computer.

4. Numerical results

As we mentioned in section 1, in the implementation of the iterative methods the following problems arise:

1. How can the iterative process be stopped correctly?
2. What is the accuracy of the computed solution given by computer?

Now, we show that the use of the theorems of section 2 and CADNA library allows us to solve these problems. Let us first consider the results of theorem 1 and lemma 1. From these, for a convergence iterative method $x_{m+1} = Mx_m + c$ and for sufficiently large value of $m$, we have

$$C_{x_{m},x_{m}}^{(i)} \geq C_{x_{m},x_{m+1}}^{(i)} - (\beta_i + 1)$$

from Lemma 1

$$= C_{x_{m},x_{m+1}} + z_m - (\beta_i + 1)$$

from Theorem 1

$$\geq C_{x_{m},x_{m+1}} + \log_{10} |1 - \|M\|_2| - (\beta_i + 1),$$

(21)

where $x_{m}^{(i)}$, $x^{(i)}$ are the $i$th components of the computed solution $x_m$ and exact solution $x$, respectively. Hence the number

$$C_{i}^{*} = C_{x_{m},x_{m+1}} + \log_{10} |1 - \|M\|_2| - (\beta_i + 1), \quad i = 1,\ldots,n,$$

is almost a lower bound for the number of common significant digits of $x_{m}^{(i)}$ and $x^{(i)}$. Note that these numbers are readily available, because the norm $\|M\|_2$ can be estimated by power method [23]. In addition, if we define

$$C^{*} = C_{x_{m},x_{m+1}} + \log_{10} |1 - \|M\|_2| - (\beta_{\max} + 1),$$

according to the definition of $\beta_{\max}$, we have

$$C_{i}^{*} \geq C^{*}, \quad \text{for } i = 1,2,\ldots,n,$$

in which the equality holds for the components with smallest magnitude. Therefore, in the floating arithmetic by using the stoping criterion

$$\text{if } C^{*} \geq k \quad \text{and } \quad \|r_m\|_2 \leq \epsilon,$$

(22)

one should expect each component of computed solution to have at least $k$ significant digits. Note that, in this stopping criterion the criterion $\|r_m\|_2 \leq \epsilon$ is used only to insure that the parameter $z_m$ in Eq. (6) is a good approximation of $C_{x_{m},x} - C_{x_{m},x_{m+1}}$. So, it is not necessary that the arbitrary value $\epsilon$ is chosen too small. It has been observed in experiments that $\epsilon = 10^{-2}$ is a good value and this tolerance furnishes the values $C_{i}^{*}, i = 1,2,\ldots,n$, which are the good
lower bounds for the number of significant digits of components of computed solution $x_m$. However, for choosing the integer $k$ a new problem arises. When $k$ is chosen too small ($k = 3$) the process is broken off too early and consequently the solution obtained has poor accuracy. On the other hand, when $k$ is chosen too large ($k = 10$), it is possible, due to numerical instabilities or/and stationarity, that the criterion $C^* \geq k$ is never satisfied. So, we need to use additional criterion for the cases:

(i) the algorithm is stationary and can not converge.

(ii) the computer is not able to improve the accuracy of computed solution, because of the round-off error propagation.

As explained in [4], the stochastic arithmetic allows the development of a criterion for these cases. In stochastic arithmetic, when the iterative process becomes stationary (before the stopping criterion (22) is satisfied ), that is, the difference between two iterates is insignificant, the values $x_{m+1}^{(i)} - x_m^{(i)}$, $i = 1, 2, \ldots, n$ are stochastic zero. So, with the CADNA library which automatically implements the CESTAC method, and using the stopping criterion

$$|x_{m+1}^{(l)} - x_m^{(l)}| = 0, \quad \text{and} \quad \|r_m\|_2 \leq \epsilon$$

with $|x_m^{(l)}| = \min_i |x_m^{(l)}|$ and $\epsilon$ large (for example $\epsilon = 10^{-2}$), it is possible to stop the iterative process as it becomes stationary and furnish the components of the computed solution $x_m$ with at least $C^*$ significant digits. It is clear that, by computing the numbers $C_i^*$, $i = 1, 2, \ldots, n$, we will have the accuracy of each components.

Similarly, for the FOM-like algorithm, if we define

$$C_i^* = C_{x_m, x_{m+1}} - \log_{10}(\text{cond}(A)) - (\beta_i + 1),$$

and

$$C^* = C_{x_m, x_{m+1}} - \log_{10}(\text{cond}(A)) - (\beta_{\max} + 1),$$

then, the lemma 1 and theorem 2 allow us to use the stopping criterion (23) to stop correctly the code of FOM-like algorithm and to furnish the accuracy of the computed solution given by computer. In the above relations we can replace $\log_{10}(\text{cond}(A))$ by $\log_{10}(s)$ where $s$ is an appropriate upper bound for $\text{cond}(A)$ and can be readily computed [14,17].

Let us now present the examples and the results which we obtained by the FORTRAN code of iterative methods SOR, FOM(m)-like with the CADNA library and the test (23). The computation has been performed on a PC computer in double precision with the stochastic arithmetic using the CADNA library.

**Example 1.** We consider the following second-order elliptic equation

$$-\nabla (a(x,y)\nabla u) = f(x,y), \quad (x,y) \in \Omega = (0,1)^2$$

and
\[ u \mid_{x=0}, \quad u \mid_{y=0} \text{ given} \]

\[ \frac{\partial u}{\partial x} \mid_{x=1}, \quad \frac{\partial u}{\partial y} \mid_{y=1} \text{ given} \]

where

\[ a(x, y) = \begin{cases} 
  1, & x < \frac{1}{2} \quad \text{or} \quad y < \frac{1}{2} \\
  10^4, & x > \frac{1}{2} \quad \text{and} \quad y > \frac{1}{2} 
\end{cases} \]

and which was described in [1]. Discritizing the problem by finite element method and piecewise linear basis functions on isosceles right triangles with meshsize \( h = 0.25 \), leads to a linear system with a SPD matrix of dimension \( n = 18 \). In our test, the right hand side is chosen once the discrete equations are formed. It is selected so that the solution to the discrete system is as follows.

\[
x_i = \begin{cases} 
  2^{-6}, & \text{if } (i \mod 6) = 0, \\
  2^{-3}, & \text{if } (i \mod 6) = 1, 2, 3, \\
  2^3, & \text{if } (i \mod 6) = 4, \\
  2^6, & \text{otherwise}
\end{cases}
\]

This allows an easy verification of the results. With \( \epsilon = 10^{-2} \) and \( x_0 = [0, 0, \ldots, 0]^T \) the results obtained by iterative method SOR with relaxation parameter \( \omega = 1.55 \) are presented in Table 1. Table 1 contains the computed solution, exact solution and \( C_i^* \) for each component.

**Table 1 : Numerical results for Example 1**

<table>
<thead>
<tr>
<th>( x(i) )</th>
<th>Approximate solution</th>
<th>Exact solution</th>
<th>( C_i^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>x(1)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(2)</td>
<td>0.12499999999996E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(3)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(4)</td>
<td>0.79999999999995E+001</td>
<td>0.80000000000000E+001</td>
<td>13</td>
</tr>
<tr>
<td>x(5)</td>
<td>0.63999999999999E+002</td>
<td>0.64000000000000E+002</td>
<td>14</td>
</tr>
<tr>
<td>x(6)</td>
<td>0.15624999999999E+001</td>
<td>0.15625000000000E+001</td>
<td>10</td>
</tr>
<tr>
<td>x(7)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(8)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(9)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(10)</td>
<td>0.79999999999998E+001</td>
<td>0.80000000000000E+001</td>
<td>13</td>
</tr>
<tr>
<td>x(11)</td>
<td>0.63999999999999E+002</td>
<td>0.64000000000000E+002</td>
<td>13</td>
</tr>
<tr>
<td>x(12)</td>
<td>0.15624999999999E+001</td>
<td>0.15625000000000E+001</td>
<td>10</td>
</tr>
<tr>
<td>x(13)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(14)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(15)</td>
<td>0.12499999999999E+000</td>
<td>0.12500000000000E+000</td>
<td>11</td>
</tr>
<tr>
<td>x(16)</td>
<td>0.79999999999998E+001</td>
<td>0.80000000000000E+001</td>
<td>13</td>
</tr>
<tr>
<td>x(17)</td>
<td>0.63999999999999E+002</td>
<td>0.64000000000000E+002</td>
<td>13</td>
</tr>
<tr>
<td>x(18)</td>
<td>0.15624999999997E-002</td>
<td>0.15625000000000E-002</td>
<td>10</td>
</tr>
</tbody>
</table>
It must be noted that the process was stopped by the stopping criterion (23) at 276th iteration, and the norm of the residual for the computed solution was \( \| \mathbf{r}_{276} \| = 0.3E - 12 \). It is necessary to mention that the norm of the residual was not significantly reduced after iteration 276. The results presented in Table 1 show that each computed component has at least \( C_i^* \) exact significant digits. We observe that the computed components \( x_6, x_{12} \) and \( x_{18} \) with smallest magnitude have 12 exact significant digits. For this components we have the lower bounds \( C_i^* = C_i = 10 \). The other components have at least 13 significant digits and their obtained smallest lower bound is \( C_i^* = 11 \). This example shows that with the CADNA library and an appropriate stopping criterion can not only obtain a satisfactory solution, but also furnish a good lower bound for the number of exact digits of each component of computed solution.

Example 2. In the second example, which is taken from [15], we consider the constant coefficient equation

\[
-\Delta u + 100u = f,
\]

in the region \((1, 2) \times (1, 3)\) with Dirichlet boundary conditions. Discretizing (25) on an \( n_x \times n_y \) grid, by using the second order centered differences for the Laplacian, gives rise to a sparse SPD linear system of equations of order \( n = n_x \times n_y \) (see [3]). In our test, we take \( n_x = 19 \) and \( n_y = 39 \). This yields a matrix of order \( n = 741 \) with moderate condition number 29.2719. As two preceding examples, 2 and 3, the tolerance \( \epsilon \) and initial vector are \( \epsilon = 10^{-2} \) and \( x_0 = [0, 0, ..., 0]^T \). The right hand side is chosen once the equations are formed. It is selected so that the solution to discrete system is \( u_{ij} = i^2 + j^2 \). The results of solving of this linear systems by FOM-like algorithm at iteration 12 is shown in table 2.

<table>
<thead>
<tr>
<th>( x(i) )</th>
<th>approx. solution</th>
<th>exact solution</th>
<th>( C_i^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>x(1)</td>
<td>0.664076728566626E+001</td>
<td>0.664076728566626E+001</td>
<td>9</td>
</tr>
<tr>
<td>x(2)</td>
<td>0.733992918590374E+001</td>
<td>0.733992918590374E+001</td>
<td>9</td>
</tr>
<tr>
<td>x(3)</td>
<td>0.8111895221060E+001</td>
<td>0.8111895221060E+001</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>x(350)</td>
<td>0.88355765332654E+002</td>
<td>0.883557653326541E+002</td>
<td>10</td>
</tr>
<tr>
<td>x(351)</td>
<td>0.99813105855615E+002</td>
<td>0.998131058556147E+002</td>
<td>10</td>
</tr>
<tr>
<td>x(352)</td>
<td>0.11278365864575E+003</td>
<td>0.112783658645752E+003</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>x(739)</td>
<td>0.284343909061207E+004</td>
<td>0.284343909061207E+004</td>
<td>11</td>
</tr>
<tr>
<td>x(740)</td>
<td>0.334631533448911E+004</td>
<td>0.334631533448911E+004</td>
<td>11</td>
</tr>
<tr>
<td>x(741)</td>
<td>0.393878933966242E+004</td>
<td>0.393878933966242E+004</td>
<td>12</td>
</tr>
</tbody>
</table>
In this test, the process was stopped at the 10th iteration, and the norm of the residual for the computed solution was $\| r_{10} \|_2 = 0.2E - 10$. The run of the code without stopping criterion showed that the norm of the residual was not significantly reduce after the 16th iteration. The results presented in Table 2 show that, for this problem with moderate condition number, every computed component has at least 14 exact significant digits which is greater than corresponding numbers computed $C^*_i$. So we observe that the numbers $C^*_i$, $i = 1, \ldots, n$ give the good information about the number of significant digits of each computed components. Consequently, by FOM-like method performed with the CADNA library, it is possible to obtain a satisfactory informatical solution for the SPD linear system and determine the accuracy of each component.

Example 3. We consider the linear system with $A = B^T B$, where $B$ is a matrix of dimension 50 and defined by

$$b_{ij} = \begin{cases} 0.25 & \text{if } i = j \\ 1 & \text{otherwise.} \end{cases}$$

The condition number of $A$ is $\text{cond}(A)=4312.1$ and the right hand side is determined so that the solution $x$ is as follows,

$$x_i = 1 + i, \quad i = 1, 2, \ldots, n.$$ 

As two preceding examples 2 and 3, the tolerance and initial vector are $\epsilon = 10^{-2}$ and $x_0 = [0, 0, \ldots, 0]$. The solution furnished by the code of the FOM-like algorithm and $C^*_i$ are presented in Table 3. The iterative process was stopped by the stopping criterion (23) at 3rd iteration and the norm of the residual was $\| r_3 \| = 0$, which means that the computed solution is a satisfactory informatical solution and computer is unable to distinguish the vector $r_3$ from the null vector and to improve the computed solution. The results presented in Table 3, show that, for this problem with large condition number, every computed component has at least 12 exact significant digits. The computed lower bounds $C^*_i$, $i = 1, \ldots, n$ give the good information about the number of significant digits of each computed components. As we observe that, for the problem with large condition number, the optimal test (23) furnishes the good results.

<table>
<thead>
<tr>
<th>$x(i)$</th>
<th>approx. solution</th>
<th>exact solution</th>
<th>$C^*_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x(1)$</td>
<td>0.200000000000E+001</td>
<td>0.200000000000000E+001</td>
<td>9</td>
</tr>
<tr>
<td>$x(2)$</td>
<td>0.299999999999E+001</td>
<td>0.300000000000000E+001</td>
<td>9</td>
</tr>
<tr>
<td>$x(3)$</td>
<td>0.4000000000003E+001</td>
<td>0.400000000000000E+001</td>
<td>9</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$x(48)$</td>
<td>0.48999999999999E+002</td>
<td>0.490000000000000E+002</td>
<td>10</td>
</tr>
<tr>
<td>$x(49)$</td>
<td>0.499999999999999E+002</td>
<td>0.500000000000000E+002</td>
<td>10</td>
</tr>
<tr>
<td>$x(50)$</td>
<td>0.510000000000000E+002</td>
<td>0.510000000000000E+002</td>
<td>11</td>
</tr>
</tbody>
</table>
5. Conclusion

In this paper we developed a technique for estimating the number of exact significant digits of components of computed solution. We have shown that the use of CADNA library allows us to furnish an optimal criterion which uses informational zero and is able to stop correctly the iterative process and to save computer time, because many useless operations and iterations are not performed. As we observe the advantages of the given optimal stopping criterion are:

(i) the moderate tolerance, for example $\epsilon = 10^{-2}$, always implies the good results.

(ii) the process stops as soon as a satisfactory informational solution is reached.

In addition, the numerical tests show that each components $x^{(i)}$ of computed solution $x_m$ has at least $C^*_i$ exact significant digits. Consequently, we observe that by simple calculation it is possible to determine the accuracy of the computed solution and the use of the CADNA library and appropriate test allows to stop correctly the process.

6. Acknowledgements

We would like to thank Prof. J. Vignes and Prof. J. M. Chesneaux for advise on many aspects of this work.

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 Mohaghegh Ardabili University of Ardabil, Iran. His research interests are mainly iterative methods for sparse linear systems and finite element method.

Department of Mathematics, Mohaghegh Ardabili University, P. O. Box. 56199-11367, Ardabil, Iran.

e-mail: khojaste@uma.ac.ir

Faezeh Toutounian received her B.Sc in Mathematics from Ferdowsi University of Mashhad, Iran, two degree of M.Sc in Mathematical statistics and applied computer and her Ph.D in Mathematics from Paris VI University, France. She spent two sabbatical years in 1985 and 1996 at Paris VI University. She is currently a professor of Mathematics at Ferdowsi University of Mashhad. Her research interests are mainly numerical linear algebra, iterative methods and error analysis.

Department of Mathematics, School of Mathematical Sciences, Ferdowsi University of Mashhad, P.O. Box 1159-91775, Mashhad, Iran.

e-mail: toutouni@math.um.ac.ir