Abstract. The article presents cases of lack of convergence of transitional problems solutions both for the linear and non-linear examples because of using IEEE-754 standard. In a linear case, electric circuit solution using state variables was presented. As a non-linear case, ferroresonance system solved by various numerical procedures was shown. To solve the proposed problems, the 64-bit Mathcad Prime 3.0 environment was used.

Introduction. Nowadays, very often to solve technical matters described by mathematical equations, ready applications such as Matlab, Mathcad, or Mathematica are used. The programs contain numerous preset procedures allowing to change various parameters which also have influence on accuracy. It is obvious that the value of many of these parameters has its limitations. Such a limitation is the maximum number of digits representing the number in these environments (17digits). It is caused by a still used internal way of representing floating point numbers using the BFP record (binary floating point). In order to standardize rules on floating point operations the IEEE-754 standard [4] was developed which defines the BFP formats available for binary system that is binary_32 called single precision, binary_64 - double precision and binary_128 - quadruple precision. However, in most of the packages in available for binary system that is binary_32 called single precision, binary_64 - double precision and binary_128 - quadruple precision. However, in most of the packages in both numeric (as Matlab) and universal as (Mathcad) they use numbers record in double precision. Their record is composed of 64 bits, where 11 bits fall for the exponent and 52 bits for the mantissa. Such a method of recording the number gives a range from about ±2.2 10^{308} to ±1.8 10^{308}. For example, the expression 2K^3 / K^3 = 2 will not be calculated for K> 170 because the program will attempt to determine separately the value of a counter and afterwards a denominator thus exceeding the limit values understood by the computer. In addition to limiting the representation of numbers due to the length of computer words it is worth determining separately the value of a counter and afterwards.

Additionally, one should note that for many rational decimals there is no exact binary representation for irrational numbers.

Example of linear circuit analysis

To solve the linear circuit (Fig.1) the equation of state variables was applied (1). In turn, to determine the transition matrix the Sylvester method, and further the method of Taylor series developing were used (5). When using a computer it should be emphasized that the method of Taylor is offered in references as effective for any non-singular matrix A, and for any time [1].

Fig. 1. Tested linear circuit, e(t)=Eo sin(ωt)

Then multiplying equation (2) by \( H^{-1} \) we obtain:

\[
\begin{align*}
\mathbf{x} &= \mathbf{H}^{-1} \mathbf{C x} + \mathbf{H}^{-1} \mathbf{D e}(t) = \mathbf{A x} + \mathbf{B e}(t)
\end{align*}
\]

In the considered example matrices A and B in Mathcad are determined as follows:

\[
\begin{align*}
H^{-1} C &= \begin{pmatrix}
-R_C C_1 & -R_C C_2 & -L_1 & 0 \\
-R_C C_1 & -R_C C_2 & L_1 & -L_2 \\
0 & -R_C C_2 & 0 & L_2 \\
C_1 & 0 & 0 & 0
\end{pmatrix}^{-1} = \begin{pmatrix}
1 & 0 & 0 & R_1 + R_1 \\
-1 & 1 & 0 & -R_2 \\
-1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix} \\
H^{-1} D &= \begin{pmatrix}
-R_C C_1 & -R_C C_2 & -L_1 & 0 \\
-R_C C_1 & -R_C C_2 & L_1 & -L_2 \\
0 & -R_C C_2 & 0 & L_2 \\
C_1 & 0 & 0 & 0
\end{pmatrix}^{-1} = \begin{pmatrix}
-1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 \\
C_1 & 0 & 0 & 0
\end{pmatrix}
\end{align*}
\]
Because the solution of the inhomogeneous equation (1) can be represented as a formula (4):

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}B(e)\,d\tau$$

to be determined by a closed Sylvester formula [7]:

$$e^{At} = A_{\text{Sylvester}}(t) = \sum_k \frac{(A \cdot t)^k}{K!}$$

When the matrix $A$ of $n$-th order has $n$ different eigenvalues, the transition matrix (the matrix exponential function) may be determined by a closed Sylvester formula [7]:

$$e^{At} = A_{\text{Sylvester}}(t) = \sum_{\lambda_i} \exp(\lambda_i t) \prod_{\lambda_j \neq \lambda_i} \left| \lambda_j - \lambda_i \right|$$

Using formula (6) an analytic solution of considered circuit was obtained in Mathcad. Additionally to validate solutions the symbolic processor of this environment was used allowing the use of Laplace method. Examples of voltage and current waveforms showing their compatibility are presented in Fig.2,3.

It so happens that the matrix $A$ in the equation of state (1) has multiple roots of the characteristic equation and then you can not apply the Sylvester method to determine transitional matrix. Additionally in the literature an alternative calculation of this matrix by the formula (5) is reported [1, 2, 3, 6, 7, 9]. Therefore, it has been attempted to solve the same example using Taylor method to develop a transitional matrix. In the Fig. 4,5 there are examples of solutions of the same example but with use of (5). Comparing these solutions, it was found that after a short time, the bifurcation of these solutions was started, although in Taylor series as many as $K=100$ components were used. In addition, it must be admitted that the time of the solutions of inhomogeneous state equation using the formula (4) by using (5) increased by several times compared with the use of (6). Therefore, in further simulations a superposition of states using a simple symbolic solution of steady state for sine extortion was employed. The results coincided with the results using (6).

Fig.4,5 is obvious if one compares the transitional matrices determined on the basis of Taylor series development even for $K=170$ that is the maximum value interpreted in Mathcad for the expression $K!$. The results for this case are shown in Fig.6 and Fig.7.

The situation in this case is improved (but still for an unsatisfactory range of less than 15 ms) by an increased number of components of the series (5) to $K = 170$ that is the maximum value interpreted in Mathcad for the expression $K!$ . The results for this case are shown in Fig.6 and Fig.7.

It turns out that in environments such as Mathcad one can not obtain satisfactory accuracy of transitional matrix determined on the basis of Taylor series development even using a large number of components. In this case, for $K=170$ Mathcad generates a message: Found a number with a magnitude greater than $10^9307$ while trying to evaluate this expression.
The voltages waveforms for the solution of the inhomogeneous equation by Sylvester and Taylor method for K=170

In [2] pattern (7) is formulated which determines the maximum error made in calculating the matrix function according to the formula (5). This ensures that the rest of the series (above K components) is convergent, if:

$$\|AT\| \leq 1$$

(7)

The following is a fragment of simulation in Mathcad environment which computes the norms of $AT$ for the $T1$ and $T2$ as well as the maximum error based on (7):

$$K \gg 100 \quad r_{ij} := \frac{\text{norm}(AT)}{\text{norm}(AT)}$$

As you can see from the above simulation the norm of $AT$ assumes a value of 1 already for the time 15 $\mu$s while the error for $T1$ (where the norm amounts to 1000) is enormous and rises with increasing $K$. On the other hand, below is shown a simulation of a transitional matrix for the time $T1$ to which waveforms of the determined quantities were still overlapping Fig.6,7.

$$A_{\text{Taylor}}(0.01) = \begin{bmatrix} 0.79 & 0.018 & 0.01 & 33.261 \\ 6.775 \times 10^{-3} & -5.767 \times 10^{-3} & -1.595 & -4.107 \\ -5.093 \times 10^{-3} & 2.127 \times 10^{-3} & 0.792 & -2.233 \times 10^{-3} \\ -3.326 \times 10^{-3} & 1.095 \times 10^{-3} & -4.466 \times 10^{-3} & 0.779 \end{bmatrix}$$

$$A_{\text{Sylvester}}(0.01) = \begin{bmatrix} 0.79 & 0.018 & 0.01 & 33.261 \\ 6.775 \times 10^{-3} & -5.881 \times 10^{-3} & -1.595 & -4.107 \\ -5.093 \times 10^{-3} & 2.127 \times 10^{-3} & 0.792 & -2.233 \times 10^{-3} \\ -3.326 \times 10^{-3} & 1.095 \times 10^{-3} & -4.466 \times 10^{-3} & 0.779 \end{bmatrix}$$

Assuming that the values of the elements of the matrix from Sylvester method are accurate a matrix of errors was also defined:

$$A_{\text{Sylvester}}(0.01) - A_{\text{Taylor}}(0.01) = \begin{bmatrix} 6.135 \times 10^{-8} & -2.084 \times 10^{-6} & -6.356 \times 10^{-6} & 5.204 \times 10^{-6} \\ 7.555 \times 10^{-8} & -1.149 \times 10^{-4} & -1.92 \times 10^{-4} & 2.145 \times 10^{-4} \\ 1.875 \times 10^{-11} & 2.093 \times 10^{-8} & 4.967 \times 10^{-8} & -3.196 \times 10^{-8} \\ -1.6 \times 10^{-10} & 4.518 \times 10^{-8} & 3.323 \times 10^{-7} & -1.96 \times 10^{-7} \end{bmatrix}$$

As you can see the values of matrix elements are indeed lower than the error estimated on the basis of (7), but the estimate does not constitute reliable information about the error of method. Of course, in the case of application of Taylor method one can use the discrete (iterative) method [2] where the matrix $\exp(AT)$ is determined once for the satisfying low-value $T$.

**Example of nonlinear circuit analysis**

The second considered example is a circuit with the highly non-linear coil (Fig.8). The circuit was a simplified model of the system power transformer [8]. Characteristics $i(t)$ was approximated by a polynomial of 11-th degree (8).

$$i(t) = a + b \frac{d^2i(t)}{dt^2}$$

(8)

In turn, reduction of above system in the flux function leads to the well-known second-order equation called nonlinear damped oscillator:

$$\frac{d^2\psi}{dt^2} + \frac{1}{LC} \frac{d\psi}{dt} + \frac{a}{C} \psi + \frac{b}{C} \psi^3 = \omega E_m \cos(\omega t)$$

In the first approach to the solution of the system (9) the Runge-Kutta procedure (in Mathcad – rkfixed) was used, wherein to determine the step leading to the convergence, iterative simulations were carried out reducing the step by half in each iteration. Figure 9 shows the course of the flux for a pre-100-fold lower voltage $E_{in} = 135V$ where convergence was achieved very quickly even for as wide a time range as more than 20 seconds.

Further simulations were carried out already for a given voltage $E_m = 13.5kV$ (Tab. 2). This time by increasing the number of steps to a maximum value accepted by the Mathcad environment (89x10^5) one failed to achieve the expected convergence. As you can see in Fig 10, at a borderline small step one has managed to retain convergence only for a small time range, i.e. a little over 0.15 seconds. Since the matrix $A$ of the system (9) is not best conditioned, also the procedure dedicated for badly conditioned systems (in Mathcad - Stiffr procedure) was applied. Unfortunately one also failed to achieve convergence for any step. In Fig.11 flux waveforms obtained by the Stiffr and Runge-Kutta procedure for the
same small step were compared. As expected waveforms coincided only for a small time range. It should be emphasized that the simulations were carried out for the case of both a dimensionless and dimensional form but unfortunately it did not improve the convergence. The Fig. 12 shows the waveforms of relative errors for both of them in the functions of iterations which means step decreasing by half.

Using all family of procedures of Mathcad environment one did not get convergence for any of the desired time ranges. What is worse, Mathcad should display a message: not converging, and in the considered case it does not appear [5].

Using all family of procedures of Mathcad environment one did not get convergence for any of the desired time ranges. What is worse, Mathcad should display a message: not converging, and in the considered case it does not appear [5].

**Conclusion**

The presented results lead to the fundamental conclusion that for certain parameters known mathematical formulas (despite dedicating them to computing [1,2,3,6]) may not be effective in many currently popular numeric or multi-tasking programs, such as Mathcad or Matlab due to the binding IEEE-754 standard. At the same time whereas in the first example, the inaccuracy of the determination of transitional matrix based on the Taylor series was clearly visible then in the case of the nonlinear example the user can assume that if the program does not alert of the lack of convergence or exceeding limit values it means that it has obtained a reliable solution. Of course, there are specialized environments based on the size of binary128 (quadruple precision), which would undoubtedly improve the results and it seems that it is a matter of time when popular calculation tools mentioned in the article increase the precision of number representation.

**REFERENCES**