Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2023, 31 (3) 218–227
http://journals.rudn.ru/miph

Research article

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-3-218-227
EDN: YENIDI

Numerical integration of the Cauchy problem with non-singular special points

Aleksandr A. Belov[1, 2], **Igor V. Gorbov**[1]

[1] *M. V. Lomonosov Moscow State University,*
1, bld. 2, Leninskie Gory, Moscow, 119991, Russian Federation
[2] *Peoples' Friendship University of Russia (RUDN University),*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation

(received: July 28, 2023; revised: August 28, 2023; accepted: September 8, 2023)

Abstract. Solutions of many applied Cauchy problems for ordinary differential equations have one or more multiple zeros on the integration segment. Examples are the equations of special functions of mathematical physics. The presence of multiples of zeros significantly complicates the numerical calculation, since such problems are ill-conditioned. Round-off errors may corrupt all decimal digits of the solution. Therefore, multiple zeros should be treated as special points of the differential equations. In the present paper, a local solution transformation is proposed, which converts the multiple zero into a simple one. The calculation of the latter is not difficult. This makes it possible to dramatically improve the accuracy and reliability of the calculation. Illustrative examples have been carried out, which confirm the advantages of the proposed method.

Key words and phrases: ordinary differential equations, Cauchy problem, multiple zero, solution transformation

1. Introduction

Consider the Cauchy problem for an ordinary differential equation (ODE)

$$du/dt = f(u,t), \quad u(0) = u^0. \tag{1}$$

The solution of many such problems has one or more multiples of zeros inside the integration segment. Examples are special functions: elliptic Weierstrass functions [1], θ-function [2], derivatives of cylindrical functions [3], and a number of others.

To calculate them, power series, Fourier series or direct numerical integration of the original equation [2] are used. The latter method seems to be the most versatile. However, the numerical calculation of such problems faces a typical difficulty. If the grid node exactly coincides with the position of the solution zero and the order of accuracy of the scheme does not exceed the multiplicity of the zero, then the further numerical solution identically equals and calculation becomes impossible. If the grid node does not coincide with the soluton zero, but is close to it, then the numerical solution becomes so small in absolute magnitude that it turns out to be comparable to unit round-off errors.

After passing a multiple of zero, the integral curves diverge rapidly, so the contribution of rounding errors increases by many orders of magnitude. Thus, passing a multiple of zero "removes" several significant digits from the solution. The more multiples of zeros fall on the integration segment, the greater the loss of accuracy. Such tasks are called ill-conditioned [4].

Therefore, we propose to consider multiple zeros in the solution of differential equations as special points along with poles and root singularities. We call them non-singular special points.

In the present paper, a new method for calculating problems with non-singular features is proposed. It consists of two stages:

1) numerical detection of the nearest zero, calculation of its position and multiplicity;
2) local transformation of the solution, which converts a multiple zero into a simple one. The calculation of such a solution is not difficult.

The method is generalized to ODE systems. Examples illustrating the advantages of the proposed approach are given.

2. Detection of the nearest zero

Let the nearest zero of the solution $u(t)$ be located at the point T and has a multiplicity q. The values of q and T are unknown in advance. Let us introduce the grid t_n, $0 \leqslant n \leqslant N$, $h = t_{n+1} - t_n$ for the independent variable. Let the calculation be carried out according to some difference scheme. The numerical solution is denoted by u_n. Obviously, the algorithm for investigating the nearest zero can use only those values of u_n for which $t_n < T$. Otherwise, the accuracy of such a study deteriorates dramatically.

Earlier in [5, 6], an algorithm for numerical detection of the nearest pole in the solution of the ODE was proposed. A zero can be considered as a pole of negative order. Therefore, we can apply this technique to the study of zeros. Let us describe the corresponding procedure. Near zero, the representation is valid

$$u = C_q(T-t)^q + C_{q+1}(T-t)^{q+1} + \dots . \tag{2}$$

Let us neglect the second and subsequent terms and differentiate this equality. Taking into account (1), we get

$$f = -\frac{qu}{T-t}. \tag{3}$$

Let us write (3) in nodes n and $n+1$. We obtain a system of equations with respect to the quantities q and T. Its solution has the form

$$q_n = \frac{t_n - t_{n+1}}{u_n/f_n - u_{n+1}/f_{n+1}}, \quad T_n = q\frac{u_n}{f_n} + t_n. \tag{4}$$

Although the exact value of q is an integer, the calculated q_n turns out to be a float-point number.

The formulas (4) are actually a difference scheme for q and T. Its error consists of two factors: the error of the original difference scheme for the problem (1) and the error introduced by discarding the second and subsequent terms in (2). The first factor can be reduced by conducting a global thickening of the grid $h \to 0$. The second factor decreases with the tendency of $t_n \to T$ even if the grid step is fixed.

It is not difficult to show that if the calculated values of q_n and T_n tend to be constant when the number of the current node n increases, then the detected singular point is a multiple of zero. The justification of this statement reproduces almost verbatim the proof of Theorem 1 from [7].

3. Transformation of the solution

w-transformation. Suppose, during the calculation using the procedure described above, a multiple zero of the solution $u(t)$ is detected. This means that for some n, the next change in the calculated q_n and T_n is quite small: $|q_n - q_{n-1}| < \varepsilon$, $|T_n - T_{n-1}| < \varepsilon$, where ε is some small number. The number of the node where this condition is met is denoted by n_*.

Round q_{n_*} to an integer and introduce a new unknown function

$$w = \operatorname{sign}(u)|u|^{1/q}. \tag{5}$$

It is not difficult to make sure that $w(t)$ satisfies the problem

$$\frac{dw}{dt} = \frac{w^{1-q}}{q} f(w^q, t), \quad w(t_{n_*}) = \operatorname{sign}(u_{n_*})|u_{n_*}|^{1/q}. \tag{6}$$

The function w has a simple zero at the point T. Numerical calculation of such a solution is not difficult.

Starting from the moment t_{n_*} we solve the problem (6) according to the same scheme as the original problem. Simultaneously, at each step, we calculate the solution $u_n = (w_n)^q$ both before and after the zero. After passing w through zero, we return to the calculation of the original problem (1). Similarly, the calculation of the second and subsequent zeros is carried out.

τ-transformation. The geometric interpretation of the transformation described above is that the multiple zero of the function u becomes a simple zero of the function w. The same result can be achieved by introducing a transformation of the independent variable instead of the solution.

Let us calculate q (rounded to an integer) and T. Let us introduce a new argument $\tau = (T-t)^q$. The solution $u(\tau)$ has a simple zero at the point T.

In the new argument, the equation (1) takes the form

$$\frac{du}{d\tau} = -\frac{1}{q}\tau^{1/q-1} f(u, T-\tau^{1/q}). \tag{7}$$

The calculation is carried out in the same way as described in the previous paragraph.

4. Generalizations

ODE systems. It is easy to generalize the described approach to the case of an ODE system of the order of J

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}, t), \quad \mathbf{u}(0) = \mathbf{u}^0, \tag{8}$$

where $\mathbf{u} = \{u^1, u^2, \dots, u^J\}$, $\mathbf{f} = \{f^1, f^2, \dots, f^J\}$.

Let several components of the solution contain multiple zeros located in the general case at different points. Then a representation similar to (2) is valid for each of these components. For each component of the solution, we conduct the study described in section 2. Let the nearest zero be located in the component u^k; it corresponds to the moment of time T^k and has the order q^k. Let us introduce a replacement (5) for the k-th component without changing other components. The resulting system takes the form

$$\begin{gathered}\frac{dw^k}{dt} = \frac{1}{q^k}[w^k]^{1-q^k} f^k(u^1, u^2, \dots, u^{k-1}, [w^k]^{q_k}, u^{k+1}, \dots, u^J), \\ \frac{du^j}{dt} = f^j(u^1, u^2, \dots, u^{k-1}, [w^k]^{q^k}, u^{k+1}, \dots, u^J), \quad 1 \leqslant j \leqslant J, \quad j \neq k.\end{gathered} \tag{9}$$

Let us calculate the system (9) until the component w^k passes through zero. Simultaneously with w^k at each step we calculate $u^k = [w^k]^{q^k}$. Then we return to the original system (8) and integrate it, simultaneously conducting a numerical study of zeros in each component. When the nearest multiple zero of one of the components is detected, we introduce a system similar to (9), etc.

Multiple constant. In addition to multiple zeros, similar difficulties are presented by points where the solution itself is different from zero, and several first derivatives are zero. Such features are natural to denote as multiple constants. In the vicinity of such a point, the solution is represented as

$$u(t) = A + C_q(T-t)^q + \dots, \tag{10}$$

where $A \neq 0$. The proposed approach can be applied directly to such problems if the value of A is known exactly. To do this, it is enough to make a transformation

$$w = A + \operatorname{sign}(u)|u|^{1/q}. \tag{11}$$

The case when A is unknown in advance is particularly difficult. We have attempted to construct various difference schemes for calculating A by analogy with 2. However, the accuracy of this calculation was insufficient to construct a transformation of the form (11). Therefore, we leave the case of the unknown A outside the scope of this work.

5. Validation

Test example. As test examples, it is advisable to choose problems with a known exact solution, which is expressed in elementary functions. This allows a particularly thorough verification of the numerical method.

Let us set the exact solution

$$u_{\mathrm{ex}}(t) = \cos^q(\pi t + \pi/4). \tag{12}$$

It has zeros of multiplicity q at points $T_k = 1/4 + k$, $k = 1, 2, ...$ Let us construct a differential equation for it. There are different ways to do this. However, an equation with the right-hand side depending only on t is of no interest, since it is solved by quadrature. On the other hand, the right-hand side, which depends only on u, also appears to be a special case. Therefore, we consider a non-autonomous equation

$$\frac{du}{dt} = -q\pi|u|^{1-1/q}\sin(\pi t - \pi/4). \tag{13}$$

The initial condition is set according to (12). The integration segment $0 < t < t_{\max}$ is selected so that it contains a specified number of multiples of zeros.

Figure 1 shows the field of integral curves for this problem. This graph illustrates what is said in section 1. The rapid divergence of the integral curves after each multiple of zero is clearly visible. It is also seen that even a relatively small change in the initial condition significantly changes the integral curve.

Along with the equation (13) in the argument “time” t, the corresponding system was considered in the argument “arc length of the integral curve” l [8, 9]. Recall the formulas for the transition to this argument

$$\frac{du}{dl} = \frac{f}{\sqrt{1+f^2}}, \quad \frac{dt}{dl} = \frac{1}{\sqrt{1+f^2}}. \tag{14}$$

It is easy to see that in this argument the vector of the right parts has unit length. It is also known [8] that parameterization through the arc length provides the best conditionality of the problem (in a global sense, i.e. over the entire segment $0 < t < t_{\max}$).

Testing methodology. The calculation of the task (13) or (14) is carried out until the specified time point $t_{\max}$ is reached. Each calculation was carried out on a set of thickening grids: the first grid contained N intervals of length h, the second had $2N$ intervals of length $h/2$, etc. The error of the numerical solution was calculated on each grid as the difference between numerical and exact solutions

$$\delta_n = u_n - u_{\mathrm{ex}}(t_n). \tag{15}$$

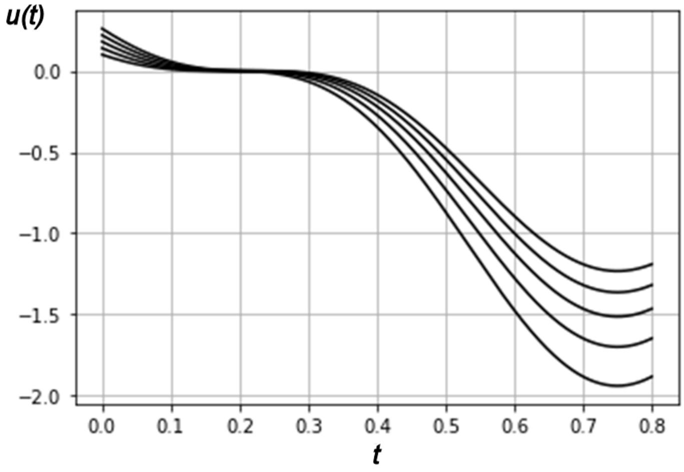

Figure 1. The field of integral curves of the equation (13)

For the problem (14), the exact solution as a function of the arc length is unknown, so we consider the error according to (15), substituting the calculated time points t_n into the exact solution (12).

Method choice. Let us put $t_{\max} = 3\pi/2 \approx 4.7$. Then the segment $0 < t < t_{\max}$ contains 5 zeros of the solution. Let $q = 3$. Let us calculate the problem (13) using an explicit four-stage Runge–Kutta scheme (ERK4) [10] using the proposed approach.

Figure 2 shows the error of the obtained solution depending on the number of grid nodes on a double logarithmic scale. Power convergence $\delta_N \sim N^{-p}$ corresponds to a straight line with a slope of $-p$.

Visually, the error curve decreases and tends to a straight line with a slope of -4. This corresponds to the theoretical 4th order of accuracy of this scheme. On excessively detailed grids, the error reaches the value $\sim 10^{-14}$ and ceases to decrease. This corresponds to the background of rounding errors. It can be seen that they are only 100 times larger than the unit rounding error. This shows the high reliability of the proposed approach.

For comparison, we performed calculations of this problem without using the proposed approach. Various schemes were used: the explicit ERK4 scheme, the explicit-implicit one-step Rosenbrock scheme with complex coefficient CROS [11], implicit optimal backward Runge–Kutta scheme BORK4 [12, 13] and the explicit Dorman–Prince method with automatic step selection DoPri5 [14, 15]. The error obtained in these calculations is also shown in figure 2. It can be seen that the ERK4, CROS and BORK4 schemes without replacement give approximately the same errors. The rate of their descending roughly corresponds to the first order of accuracy, which is sharply different from their theoretical orders of accuracy. The convergence of the DoPri5 method turns out to be faster, but the accuracy cannot be obtained better than 10^{-3}.

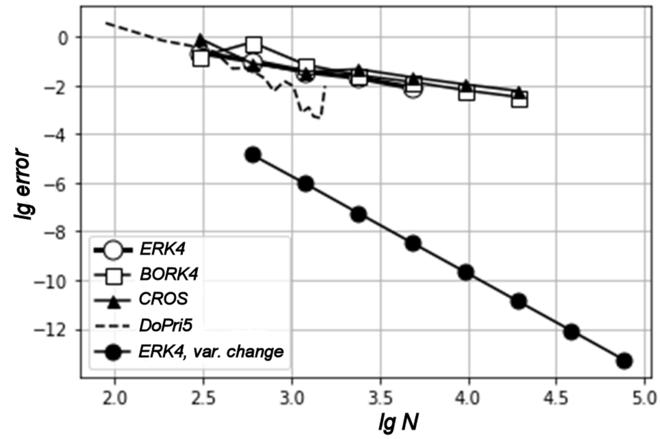

Figure 2. Errors in the test (13)

Thereby, from figure 2 it can be seen that the proposed approach dramatically increases the accuracy and reliability of the calculation. The problem under consideration presents a significant difficulty for classical schemes. However, the proposed approach allows calculations to be carried out even according to explicit schemes and to obtain an accuracy not much higher than the errors of unit round-off error.

Figure 3 shows similar calculations of the problem (14). It is clearly seen that the ERK4 scheme with the proposed replacement implements the theoretical order of accuracy and provides excellent accuracy up to $\sim 10^{-14}$. In calculations without the proposed replacement, all schemes give significantly worse accuracy and do not implement the theoretical order of convergence.

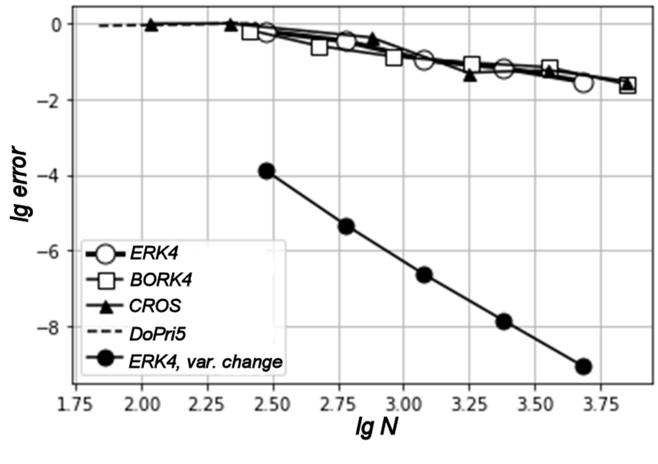

Figure 3. Errors in the test (14)

6. Conclusion

The test calculations have shown that the proposed approach for numerical solution of the Cauchy problems with multiple zeros on the integration segment provides high accuracy and reliability of calculation for a wide class of problems. At the same time, standard approaches demonstrate unsatisfactory accuracy. The simplicity of implementation, the possibility of generalization and use with a large set of numerical schemes make the method convenient for application to applied problems.

Acknowledgments

This work did not receive specific funding.

References

[1] E. Janke, F. Emde, and F. Losch, *Tafeln Horer Functionen.* Stutgart: B. G. Teubner Verlagsgesellschaft, 1960.

[2] *NIST Digital Library of Mathematical Functions*, https://dlmf.nist.gov.

[3] M. K. Kerimov, "Studies on the zeros of Bessel functions and methods for their computation," *Computational Mathematics and Mathematical Physics*, vol. 54, pp. 1337–1388, 2014. DOI: 10.1134/S0965542514090073.

[4] N. N. Kalitkin and P. V. Koryakin, *Numerical methods. Vol.2: Methods of mathematical physics [Chislennye Metody. T.2: Metody matematicheskoi fiziki].* Moscow: Akademiya, 2013, in Russian.

[5] A. A. Belov, "Numerical detection and study of singularities in solutions of differential equations," *Doklady Mathematics*, vol. 93, no. 3, pp. 334–338, 2016. DOI: 10.1134/S1064562416020010.

[6] A. A. Belov, "Numerical blow-up diagnostics for differential equation solutions," *Computational Mathematics and Mathematical Physics*, vol. 57, no. 1, pp. 122–132, 2017. DOI: 10.1134/S0965542517010031.

[7] A. A. Belov and N. N. Kalitkin, "Numerical integration of a Cauchy problem whose solution has integer-order poles on the real axis," *Differential equations*, vol. 58, pp. 810–833, 2022. DOI: 10.1134/S0012266122060088.

[8] V. I. Shalashilin and E. B. Kuznetsov, *The method of continuation by parameter and the best parametrization [Metod prodolzheniia po parametru i nailuchshaia parametrizatsiia].* Moscow: Editorial URSS, 1999, in Russian.

[9] E. B. Kuznetsov and S. S. Leonov, "Parametrization of the Cauchy problem for systems of ordinary differential equations with limiting singular points," *Computational Mathematics and Mathematical Physics*, vol. 57, pp. 931–952, 2017. DOI: 10.1134/S0965542517060094.

[10] E. A. Alshina, E. M. Zaks, and N. N. Kalitkin, “Optimal parameters of explicit Runge–Kutta schemes of low orders [Optimalnye parametry iavnykh skhem Runge–Kutty nizkikh poriadkov],” *Math. modeling*, vol. 18, no. 2, pp. 61–71, 2006, in Russian.

[11] H. H. Rosenbrock, “Some general implicit processes for the numerical solution of differential equations,” *The Computer Journal*, vol. 5, no. 4, pp. 329–330, 1963. DOI: 10.1093/comjnl/5.4.329.

[12] N. N. Kalitkin and I. P. Poshivaylo, “Inverse Ls-stable Runge–Kutta schemes,” *Doklady Mathematics*, vol. 85, pp. 139–143, 2012. DOI: 10.1134/S1064562412010103.

[13] N. N. Kalitkin and I. P. Poshivaylo, “Computations with inverse Runge–Kutta schemes,” *Mathematical Models and Computer Simulations*, vol. 6, pp. 272–285, 2014. DOI: 10.1134/S2070048214030077.

[14] E. Hairer and G. Wanner, *Solving ordinary differential equations. II. Stiff and differential-algebraic problems.* Berlin, New York: Springer-Verlag, 1996.

[15] L. F. Shampine and M. W. Reichelt, “The Matlab ODE suite,” *SIAM Journal on Scientific Computing*, vol. 18, no. 1, pp. 1–22, 1997. DOI: 10.1137/S1064827594276424.

For citation:

A. A. Belov, I. V. Gorbov, Numerical integration of the Cauchy problem with non-singular special points, Discrete and Continuous Models and Applied Computational Science 31 (3) (2023) 218–227. DOI: 10.22363/2658-4670-2023-31-3-218-227.

Information about the authors:

Belov, Aleksandr A. — Candidate of Physical and Mathematical Sciences, Assistant professor of Department of Computational Mathematics and Artificial Intelligence of Peoples’ Friendship University of Russia (RUDN University); Researcher of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: aa.belov@physics.msu.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-0918-9263)

Gorbov, Igor V. — Master’s degree student of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: garri-g@bk.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0009-0005-5335-6179)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-3-218-227
EDN: YENIDI

Численное интегрирование задач Коши с несингулярными особыми точками

А. А. Белов[1,2], **И. В. Горбов**[1]

[1] *Московский государственный университет им. М. В. Ломоносова,*
Ленинские горы, д. 1, стр. 2, Москва, 119991, Россия
[2] *Российский университет дружбы народов,*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

Аннотация. Решения многих прикладных задач Коши для обыкновенных дифференциальных уравнений имеют один или несколько кратных нулей на отрезке интегрирования. Примерами являются уравнения специальных функций математической физики. Наличие кратных нулей существенно затрудняет численный расчёт, поскольку такие задачи являются плохо обусловленными. Из-за ошибок округления в решении может не остаться ни одного верного знака. Поэтому кратные нули следует отнести к особым точкам ОДУ. В данной работе предложена локальная замена искомой функции, которая преобразует кратный нуль решения в простой. Расчёт последнего не представляет трудностей. Это позволяет кардинально повысить точность и надёжность расчёта. Проведены иллюстративные примеры, которые подтверждают преимущества предлагаемого метода.

Ключевые слова: обыкновенные дифференциальные уравнения, задача Коши, кратные нули, преобразование решения