Regularization of algorithms for estimation of errors of differential equations approximate solutions

A N Rogalev and N A Feodorova

1 Institute of Computational Modeling, 660036, Krasnoyarsk, Akademgorodok, house. 50, building 44, ICM SB RAS, Russia
2 Institute of Space and Information Technologies, Siberian Federal University, 660074, Krasnoyarsk, st. Kirensky 26, bldg. ULK, Russia
E-mail: rogalyov@icm.krasn.ru

Abstract. When estimating the error of numerical solutions of ordinary differential equations, it is proposed to use the regularization of error estimation algorithms for approximate solutions. Regularization is considered as the use of additional a priori information about the numerical solution. A priori information is estimated by the method of the inverse error analysis, which implements the transformation of the statement of the problem of evaluating the accuracy of the numerical solution. In this case, the problem is selected so that the approximate solution is the exact solution of the transformed problem, close to the original problem.

It is also possible to add some additional restrictions to the conditions of the problem. This allows us to solve the incorrectly posed problems of estimating errors in numerical solutions, which in most cases are unstable. It is known that earlier in the numerical solution of linear algebra problems the inverse error analysis was used by Wilkinson J. and also Voevodin V.V., application of the inverse error analysis in other areas of numerical analysis can be noted. Using direct and inverse error analysis Voevodin V.V. effectively calculated majorants of rounding errors in the most important methods of linear algebra and significantly developed the method of inverse error analysis. However, the regularization process associated with the reverse analysis of errors has not been previously studied.

The article discusses the characteristics of regularization in the inverse analysis of errors in the numerical solution of ordinary differential equations with initial data. The regularization of error estimation algorithms has advantages in the numerical solution of ordinary differential equations of real models of physics and technology.

1. The algorithm for estimating global error actually using regularization

The main purpose of the inverse error analysis is to compute the accuracy of the numerical solution of a problem, considering the numerical solution as an exact solution of the problem. In this case, the problem should be close to the original problem and obtained using the original problem as a starting point. Initially, the inverse analysis method was used for the problems of linear numerical algebra [1, 2, 3], later this method was applied to the numerical solution of ordinary differential equations (ODE) and partial differential equations.

The reason for the widespread use of the inverse error analysis is a large increase of the values for error estimates of numerical solutions obtained for variety of error estimation algorithms. Such an increase in error estimates does not depend on how the error of the numerical solution actually behaves. This can be explained by the fact that the algorithms for estimating the errors
of numerical solutions are ill-posed problems; where the stability condition is violated. Stability violation and the exponential growth of error estimates results in the fact “direct error analysis is almost never applicable, except for the simplest applications of the discretization method, and usually you have to rely heavily on the information obtained in the process of computing numerical solutions” [4].

Confirmation of the instability of error estimation methods can be also found in [5], where the author shows that “influential authors in computational mathematics (probably Richtmayer) noticed around 1950 that all difference schemes were incorrect in a certain sense: as $h \to 0$ makes it necessary to increase the mesh dimension unlimitedly.”

To remove the increase in error estimates, it is possible to regularize the problem [6, 7]. This means using additional information that the exact solution belongs to compact set. It is important because it means, when solving incorrect problems, it is necessary to reduce the class of possible solutions. One of the ways to obtain additional information in the reverse analysis of errors is that the numerical solution found is identified with the exact solution to the problem approximating the original problem [1, 2, 3].

This regularization is performed using methods of the reverse analysis of errors, but it is not explicitly explained and, to the best of our knowledge. Using the backward error analysis, fairly accurate error estimates were obtained for solving systems of linear algebraic equations (SLAE) [8, 9], and for other problems [10, 11].

Consider the system of ordinary differential equations (ODEs)

$$\frac{dy}{dt} = f(t, y),$$  \hspace{1cm} (1)

with initial data $y(t_0) = y_0$, where $f(t, y)$, $y$- elements of some Banach space.

We apply the numerical method to (1) and obtain a numerical solution that approximates the projection of the exact solution onto the difference grid in the region of the function argument. The uniform grid built for the ODE system is defined as $t_n = t_0 + nh$, $n = 0, \ldots, N$ $N$- is a positive integer. To compare exact and approximate solutions, it is necessary to project all the solutions into one functional space (discrete or continuous one).

To simplify the notation for the problem (1) we use the Euler method

$$\varphi(y_h)_i := \begin{cases} 
-y_{h,0} + \alpha, & i = 0, \\
\frac{(y_{h,i} - y_{h,i-1})}{h} + f(t_{i-1}, y_{h,i-1}), & i = 1, \ldots, N. 
\end{cases}$$  \hspace{1cm} (2)

The global error can be expressed as equal to the difference between the numerical and exact solutions, expanded according to the Taylor formula

$$y(t_{j+1}) - y_{h,j} = y(t_j) - y_{h,j} + h_j [f(t_j, y(t_j)) - f(t_j, y_{h,j})] + h_j^2 y''(\xi_j)/2.$$  \hspace{1cm} (3)

To explain the need for regularization, we construct the following example. The global error in the node $t_{j+1}$ is taken on the left side of this equality. It should be noted that it has two sources of the global error.

(I) A local error or an error introduced at the step from $t_j$ to $t_{j+1}$ under the assumption that all values for $t_j$ are computed exactly, coincides with the remainder term $h^2 y''(\xi)/2$ in Taylor’s formula.

(II) The error transported along the trajectory will be computed using the formula

$$h_j [f(t_j, y(t_j)) - f(t_j, y_{h,j})] = h J(\xi)(y(t_j) - y_j),$$

where the Jacobian is computed in a preliminary intermediate point $J(\xi)$. 

So, simplifying the notation, we can consider the global error computed at point $t_{j+1}$ equal to the next formula $= (1 + h_j J) \cdot (the \ global \ error \ at \ t_j) + the \ local \ error \ at \ t_{j+1}$. The global error $y(t_j) - y_j$ at the node $t_j$ is multiplied by the value $(1 + h_j J)$, called the transition factor.

When regularizing the global error estimation method, the deferred correction approach [10] may be useful. Deferred correction is increasing the accuracy of the numerical solution $y_h$, that is performed by adding a correction term to the numerical solution found earlier.

When computed $y_h$, it is necessary to select the numerical method $\phi$, for example, recalling the method by which the numerical solution $y_h$ was obtained earlier. This will save computational costs. It is also possible to use a higher order accuracy method.

To implement deferred correction, you can apply the local error $\lambda := \phi(\Delta y)$. For the Euler method

$$\lambda_i = \begin{cases} 0, & i = 0 \\ -\frac{h y''(\tau_i)}{2}, & i = 1, \ldots, N \end{cases}$$

in this formula, $\tau$ is an intermediate point for the $i-$ interval.

Obviously, if the local error $\lambda$ is known with a sufficient degree of accuracy, then the exact solution at the nodes of the grid $\Delta y$ is easy to find by solving the equation $\phi(\Delta y) = \lambda$. The main idea of the deferred correction method is to obtain an estimate of the local error using the operator $\psi$, an estimate of the local error of the numerical solution $y_h$. Since we used $\eta = \Delta y + O(h^p)$ to estimate the value of the order of $h^p$, it would seem that the estimate would satisfy $\psi(\eta) = \lambda + O(h^{r+p})$.

Thus, deferred correction evaluates the global error using the local error estimate. Moreover, it predicts that the relation $\psi(\Delta y) = \lambda + O(h^{r+p})$ holds for the operator $\psi$.

As a result, the deferred correction involves three components: the numerical solution $y_h$, the effective numerical method $\phi$, and the estimated local error function $\psi$. The refined solution $\overline{y}_h$ is calculated from the equation $\phi(\overline{y}_h) = \psi(y_h)$ and satisfies the equality $\overline{y}_h = \Delta y + O(h^{r+p})$ if for an arbitrary function $y_h = \Delta y + O(h^p)$, $\psi(\Delta y) = \varphi(\Delta y) + O(h^{r+p})$ and $\psi(\Delta z) = O(h^p)$.

2. Estimation of an approximate solution and regularization

Consider the operator equation

$$Az = u, z \in Z, u \in U,$$

where $Z, U$ - are some metric spaces.

An estimate of the error of the approximate solution $z_h$ of the equation (5) is an arbitrary domain $\Omega_h \in Z$, in which $z_h$ lies and in which the equation (5) has at least one solution $z$. Let $M_z$- be the set of all solutions of the equation (5). Then the residual (defect) of the approximate (numerical) $z_h$ solution in the norm of the solution space is the quantity

$$\delta(z_h) = \inf_{z_h \in M_z} \| Az_h - u \|,$$

where $\| \cdot \| -$ is some norm in the solution space.

Provided that the equation has no solutions, it can be assumed that $\delta(z_h) = \infty$. It is not possible to calculate the error of the numerical solution, since in the general case the exact solution is not known. If the set of values $\| z - z_h \| \leq \delta_0$ has at least one solution to the equation (5), then the set of values (sphere) $\| z - z_h \| \leq \delta_0$ is error estimation.

Problem (5) is called correctly posed if the following conditions are satisfied: 1) the operator equation (5) is solvable for any right-hand side $f \in U$; 2) the solution of the operator equation (4) is stable under perturbation of the right-hand side of the equation (5), this means the continuity of the inverse operator $A^{-1}$, defined over the entire space $U$; 3) the solution of operator equation
(5) is unique. If at least one of these conditions is not fulfilled, then the problem is called incorrectly posed [12].

In accordance with the definition of the problem [12], the following properties of the error estimation problem can be noted: 1) the operator equation (5) is solvable for any right-hand side of the system of ordinary differential equations \( f \); 2) the inverse operator \( A^{-1} \) is not continuous throughout the entire solution space; 3) the solution of the operator equation (5) is not unique.

Suppose that the operator equation (5) is being solved with the right-hand side \( f \), having a unique solution \( z \in Z \). As a rule, the right-hand side of \( f \) is unknown, we can use its approximate value \( f_\delta \in U \), for which the inequality \( \rho(u_\delta, f_\delta) \leq \delta \) holds in the metric of the space \( U \). In this inequality, the value \( \delta \) is called the error of the element \( u_\delta \).

The following approach is possible for constructing a regularizing algorithm for an ill-posed problem. Suppose that there exists an a priori given compact correctness set \( M \). We also assume that the mapping \( A \) onto the set \( A_M \) is one-to-one. In this case, the inverse map \( A^{-1} \), defined on \( A_M \), is continuous, then for an approximate solution we can take an arbitrary element \( z_\delta \in M \), such that \( \rho(Az_\delta, u_\delta) \leq \delta \).

It is known that [12],[6] that the element \( z_\delta \in M \), minimizes the defect \( \rho(Az, u_\delta) \) on the set \( M \) is called the quasisolution of the equation (5):

\[
\rho(Az_\delta, f_\delta) = \inf_{z \in M} \lim_{\delta \to 0} \rho(Az, f_\delta).
\] (7)

This scheme uses additional information about the desired solution, sometimes it is enough to choose a regularization parameter. The generalized defect principle is often used as the main algorithm for choosing the regularization parameter.

Let us analyze the regularization of the error estimation operator for numerical solutions of systems of differential equations (1), where \( y \in \mathbb{R}^n, t \in \mathbb{R} \) and \( f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \). The first type of error analysis is defect analysis (2), which works by changing the equation and imposing fixed conditions on it. This is the most natural form of reverse analysis for ODE. A numerical method is used to solve an equation whose exact solution is known. This equation differs from the original one by introducing a defect, considered as a non-autonomous perturbation of the ODE (1) [14, 15, 16].

The defect can be considered as an autonomous perturbation of the ODE due to an increase in the dimension of the system, the standard method of converting a non-autonomous ODE system to an equivalent autonomous system is used, we set \( y_{n+1} = t \), and a new equation of the system is introduced, respectively. A defect is most often used in the context of defect management problems in which the value of the defect norm for a suitable norm (most often, the maximum norm) is controlled by a numerical method. Numerical methods that use defect control most often contain an interpolant formed on the basis of a numerical solution of the numerical method for estimating or calculating a defect, then it is used to control the step size of the numerical method. The differential equation under study should be considered as an approximate equation in any case, and therefore we will go on to a modified problem for which we can find the exact solution.

Rewriting the defect equation as

\[
\frac{du}{dt} = f(t, u) + \varepsilon v(t),
\] (8)

we see that provided that \( \|v(t)\| < 1 \) for some acceptable norm in the proposed problem, the numerical method constructs the exact solution of the \( \varepsilon \)-close problem. Depending on the problem, it may be more appropriate to consider the relative defect

\[
\frac{du}{dt} = f(t, u) (1 + \varepsilon v(t)),
\]
or a defect related to \( u \) (in this case \( \delta = \delta(t, u) \))

\[
\frac{du}{dt} = f(t, u) + \varepsilon v(t)u.
\]

For example, for a piecewise cubic Hermitian interpolant on each segment \([t_n, t_{n+1}]\)

\[
u_n(t) = (\theta - 1)^2(2\theta + 1)y_n + \theta(\theta - 1)^2 h_n f(t_n, y_n) + \theta^2(\theta - 1)^2 h_n f(t_{n+1}, y_{n+1}),
\]
in the local coordinates \( \theta = \frac{t - t_n}{h_n} \), it is possible to calculate the derivative of the interpolant.

Using this function, you can find a defect.

3. Results of numerical experiments

A) For the oscillation equation (written in the form of a system of two differential equations of the first order) on the interval \([0, 10000]\), the values of the numerical solution were computed using the one-step Runge-Kutta method of the fourth order and the multi-step Adams method of the fourth order.

An error estimate (global error) was obtained for the system. We used deferred difference correction and Richardson extrapolation \([17]\), as well as a reverse error analysis.

Estimates computed using inverse error analysis:

\[
\begin{align*}
err(10) &= 0.456148217 \cdot 10^{-6}, & err(100) &= 0.64657767 \cdot 10^{-5} \\
err(1000) &= 0.785725623 \cdot 10^{-4}, & err(10000) &= 0.187345329 \cdot 10^{-3} 
\end{align*}
\]

Global error estimates computed by delayed differential correction or Richardson extrapolation (the minimum value was selected from the pair of values computed by these methods):

\[
\begin{align*}
err(10) &= 0.5641641709 \cdot 10^{-4}, & err(100) &= 0.238305021 \cdot 10^{-3} \\
err(1000) &= 0.162085872 \cdot 10^{-2}, & err(10000) &= 0.415155384 \cdot 10^{-1} 
\end{align*}
\]

B) Let us solve the ODE system

\[
\begin{align*}
\frac{dy_1(t)}{dt} &= y_2(t), & \frac{dy_2(t)}{dt} &= y_3(t), & \frac{dy_3(t)}{dt} &= -y_2(t) - 0.5y_1^2(t) + 1
\end{align*}
\]

with initial data \( y_1(0) = 0, y_2(0) = 0, y_3(0) = 1 \) using the one-step Runge-Kutta method of the fourth order and the multi-step Adams method of the fourth order.

These error estimates are computed using delayed difference correction or Richardson extrapolation at \( t = 1 \):

\[
\begin{align*}
err(y_1)(1) &= 0.617423026019743194 - 0.617423026020436306 \approx 10^{-16}, \\
err(y_2)(2) &= 0.129539072722494274 - 0.129539072723032510 \approx 10^{-16}, \\
err(y_3)(3) &= 1.34403247053568098 - 1.34403247053760988 \approx 10^{-12}
\end{align*}
\]

These error estimates are computed using delayed differential correction or Richardson extrapolation at the point \( t = 7.5 \):

\[
\begin{align*}
err(y_1)(7.5) &\approx 10^{-4}, err(y_2)(7.5) \approx 10^{-4}, err(y_3)(7.5) \approx 10^{-3},
\end{align*}
\]
These error estimates are computed by inverse error analysis at the point $t = 7.5$

$$\text{err}(y_1)(7.5) \approx 10^{-6}, \text{err}(y_2)(7.5) \approx 10^{-6}, \text{err}(y_3)(7/5) \approx 10^{-5},$$

C) The numerical error was estimated when solving the Röessler system [18]:

$$\frac{dy_1}{dt} = -y_2 - y_3, \quad \frac{dy_2}{dt} = y_1 + ay_2, \quad \frac{dy_3}{dt} = b + y_3 (y_1 - c)$$

where $a, b, c$– are positive constants.

For the parameters $a = b = 0.2$ and $c < 4.2$, the Röessler equations have a stable limit cycle. At these values of the parameters, the period and the shape of the limit cycle complete the sequence of doubling the period. Immediately after the point $c = 4.2$, the phenomenon of the so-called chaotic attractor arises. The clearly defined lines of the limit cycles blur and fill the phase space with an infinite countable set of trajectories with fractal properties.

These error estimates are obtained using delayed difference correction or Richardson extrapolation for Röessler system at $c = 4$ :

$$\text{erry}_1(4) \approx 10^{-3}, \text{erry}_2(4) \approx 10^{-2}, \text{erry}_3(4) \approx 10^{-3}. \quad (12)$$

These error estimates are obtained using inverse error analysis for Röessler system at $c = 4$ :

$$\text{err}(y_1)(4) \approx 10^{-5}, \text{err}(y_2)(4) \approx 10^{-4}, \text{err}(y_3)(4) \approx 10^{-4}. \quad (13)$$

4. Conclusion

The results of applying the inverse error analysis of many approximate solutions of ODE systems indicate the effectiveness of the inverse error analysis. Error analysis becomes more resistant to the influence of errors when computing the boundaries of the set of exact solutions of the original system (while the introduced errors cannot be large). The incorrectness of the problem is reduced by numerically assessing the deviation of the numerical solution from the boundary of this set. The regularization of the operator of the problem being solved, that is, obtaining additional information about the solutions may consist in the transition from the original system to systems that take into account the properties of the computed solution. Regularization leads to more accurate estimates of errors in numerical solutions.

Acknowledgments

The work was supported by the program of Russian Foundation for Basic Research, project 19-01-00038

References

[1] Wilkinson J 1963 Rounding errors in algebraic processes (London, Her Majesty’s Stationary Office;Englewood Cliffs, N.Y.: Prentice Hall)

[2] Voevodin V V 1967 On the asymptotic distribution of rounding errors in linear transformations J. Vychisl. Matem. Matem. Phys. 7 965-976 (in Russ.)

[3] Voevodin V V 1969 Rounding errors and stability in direct methods of linear algebra ( Moskva:MGU Publ.) p 153 (in Russ)

[4] Stetter H 1973 Analysis of discretization methods for ordinary differential equations (Springer Verlag: Berlin, Heidelberg, New York)
[5] Kalitkin N N, Yukhno L F and Kuzmina L V 2011 A quantitative criterion for conditioning systems of linear algebraic equations Mathematical Modeling 23 3-6 (In Russ.)
[6] Dorofeev K, Titarenko V and Yagola A L 2003 Algorithms for constructing posterior errors of solutions for ill-posed problems J. Vychsl. Matem. Matem. Phys 43 (1) 1225 (in Russ.)
[7] Morozov V A 1987 Regular solution methods of incorrectly assigned problems (Moscow: Science) (In Russ.)
[8] Rogalev A N, Doronin S V and Moskvichev V V 2018 Estimation of the accuracy of the numerical analysis of the deformed state of the power structures of technical objects J. Computat. Technol 23 88-101 (In Russ.)
[9] Rogalev A N 2019 Constructing a regularizing operator with reverse analysis of errors Scientific readings named after Marchuk.-2019 Proceedings of the International Conference. pp 406-412. (In Russ.)
[10] Skeel R 1981 A theoretical framework for proving accuracy results for deferred corrections SIAM J. Numer.Anal. 19 171 -196
[11] Rogalev A N and Rogalev A A 2020 Estimates of the accuracy of numerical solutions using regularization J. Phys.: Conf. Series 1441 012165
[12] Tikhonov A N, Goncharsky A V, Stepanov V V and Yagola A G 1983 Regularization algorithms and a priori information (Moscow: Science) p 200 (In Russ.)
[13] Hayes W and Jackson K 2005 A survey of shadowing methods for numerical solutions of ordinary differential equations. Applied Numerical Mathematics, 53 (2-4) 299–321
[14] Reich S 1999 Backwards error analysis for numerical integrators SIAM J. Numer. Anal. 36 1549-70
[15] Corless R 1992 Defect-controlled numerical methods and shadowing for chaotic differential equations Physica D: Nonlinear Phenomena 60 323 -334
[16] Corless R 1994 What good are numerical simulations of chaotic dynamical systems Computers & Mathematics with Applications 28 107 -121
[17] Skeel R 1986 Thirteen ways to estimate global error SIAM J. Numer.Anal. 48 1 -20
[18] Peitgen H, Jurgens H and Saupe D 2004 The Rössler Attractor Chaos and Fractals: New Frontiers of Science, Springer pp 636-646