Estimates of the accuracy of numerical solutions using regularization

A N Rogalev[1], **A A Rogalev**[2]

[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: rogalyovn@icm.krasn.ru

Abstract. This article explores the accuracy of numerical solutions, and suggests methods for analyzing accuracy depending on the properties of the problem. Numerical studies of complex expensive objects of technology and physics require that the computational results be obtained with guaranteed accuracy. It also depends on the fact that in the work of technical objects there are large intervals of operation time not observed experimentally. Therefore, there is a need to describe the location of the observed and calculated values, as well as the accuracy with which they are calculated. The effect of strong growth in estimates of error bounds is manifested for a large number of methods used to estimate the error of a numerical solution. This means the lack of correctness of algorithms for evaluating the accuracy of numerical solutions due to the failure of the stability conditions with respect to perturbations of the right-hand side. For many problems, among all the algorithms, the backward analysis of errors turned out to be the most effective method for assessing the accuracy of numerical solutions. The backward analysis of errors consists in the fact that when assessing the accuracy (error) of a numerical solution, the numerical solution is considered as an exact solution to a problem close to the original problem. A backward error analysis was proposed and developed in the algorithms of J Wilkinson in the context of the numerical solution of problems of linear algebra, and in the algorithms of V V Voevodin, who widely distributed it to many areas of numerical analysis. In the framework of the backward error analysis, the regularization of the algorithm for estimating the error of a numerical solution is reasonably applied. This article explores methods for the backward analysis of errors of numerical solutions.

1. Introduction

In order to evaluate the accuracy of a numerical solution, the backward error analysis algorithm considers the numerical solution as an exact solution of a problem close to the original problem. This distinguishes the backward analysis from the direct error analysis, which estimates the accuracy of the errors in the calculation of numerical solutions. It is known that in the methods of direct error analysis, a large increase in the bounds of estimates is noted, which in most cases greatly exceeds the values of the numerical errors themselves. In the monograph [1], Stetter argued that " 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." This thesis is confirmed by the article [2], in which it is written that some classics of computational mathematics (it seems Richtmayer)

noticed around 1950 that all difference schemes are incorrect in a certain sense: as $h \to 0$ it necessary to increase the mesh dimension unlimitedly. In fact, for many numerical algorithms, the boundary of the error of the numerical solution (which can be constructed) is growing rapidly, although the error values themselves often turn out to be much smaller. This was the basis for the development of backward analysis problems of the theory of errors: given the error of the function, it is required to determine the errors of its arguments. An unknown-dependent equation with many solutions is constructed. In order to single out one solution to the problem, it is necessary to put an additional condition (or several conditions). Hence, when assessing the accuracy of the obtained numerical solution, this solution is believed to be an exact solution to a problem approximating the original problem [3], [4], [5]. Such a method was developed by Wilkinson J. [3] in problems for the numerical solution of linear algebra problems, and extended to other areas of numerical analysis. Using direct and backward error analysis [4], [5] Voevodin V V effectively computed majorants of rounding errors in the most important methods of linear algebra, and significantly developed the results. If we consider the rounding errors of the results of intermediate calculations as functions that depend on random input data, for direct problems it was proved that the rounding errors asymptotically behave as independent, uniformly distributed random variables (in terms of the number of digits of the representation of numbers). Studies were conducted on the influence of small perturbations of the input data on the solution of many problems of linear algebra, including incorrectly posed ones. Using the backward error analysis, fairly accurate error estimates were obtained for solutions of systems of linear algebraic equations (SLAE) [6], [7], as well as for many other problems.

2. The global error estimation algorithm for ODE systems in the operator form

Let a system of ordinary differential equations (ODEs) with initial data

$$F(y) := \begin{pmatrix} -y(t_0) + y_0 \\ -y'(t) + f(t, y(t)) \end{pmatrix}, t_0 \le t \le T \tag{1}$$

be solved. This system is written in the form of a nonlinear operator equation $F(y) = 0$, in which all additional conditions are included in the operator that maps some function spaces. The requirements are imposed to ensure the existence and uniqueness of the solution. We apply the numerical method and obtain the approximate solution, which approximates the projection of the exact solution onto the difference grid of the domain of the function argument. The uniform grid constructed for the ODE system is given as $t_n = a + nh$, $n = 0, ..., N \quad N-$ is a positive integer number. To compare the exact and approximate solutions, one has to project all the solutions into one functional space (discrete or continuous). Definition 1. The global error of the interpolant of the numerical solution of the ODE system (1) is the difference between the exact solution of the system (1) and the interpolant of the numerical solution. In some cases, the term global error is also used to denote the norm of the difference of these values.

Definition 2. A defect, or residual, is a quantity

$$\delta(t) := \frac{du}{dt}(t) - f(t, u). \tag{2}$$

The backward analysis of numerical solutions errors can be performed by applying a defect to compute the bound of global error $\| y(t) - u(t) \|$, where $y(t)-$ is an exact solution of the ODE system (1) and $u(t)$ is an interpolant constructed by a numerical solution y_h. Since this means that $\frac{du}{dt} = f(t, u) + \delta(t)$, the defect is equal to the quantity measuring the extent to which the numerical solution does not satisfy the differential equation (1). The concept of a defect (residual) can also be compared with the difference between the original equation (1) and the equation whose exact solution is a numerical solution. The global error norm $\varepsilon := \|y_h - \Delta y\|$.

In the general case, for a sufficiently smooth operator F , the error $\varepsilon = O(h^r)$, where h is the grid step, r is a positive integer characterizing the accuracy order of the numerical solution [1], [2].

In evaluating the global error, the deferred correction approach [8] is of great importance, consisting in computing a more accurate numerical solution $\overline{y_h}$, obtained by adding the correction term to the numerical solution found earlier. Computation $\overline{y_h}$ requires the use of some numerical method ϕ. As ϕ, you can use the method by which a numerical solution y_h was obtained, which can be done with cost savings. As a variant of the method ϕ , a more efficient method order $p \geq 1$ can also be used. For example, let the Euler method be chosen for problem (1)

$$\varphi(y_h)_i := \begin{cases} -y_{h,0} + \alpha, & \text{i} = 0, \\ -\frac{(y_{h,i} - y_{h,i-1})}{h} + f(t_{i-1}, y_{h,i-1}), & \text{i} = 1, ..., N \end{cases} . \tag{3}$$

To implement deferred correction, a local error $\lambda := \phi(\Delta y)$ is used, which characterizes the quality of approximation by an exact solution of the problem of solving the difference equation $\phi(y_h)$.

For the Euler method

$$\lambda_i = \begin{cases} 0, & \text{i} = 0 \\ -\frac{hy''(\tau_i)}{2}, & \text{i} = 1, ..., N \end{cases} ,$$

in this formula, τ is some intermediate point on the $i-$ interval. Obviously, if the magnitude of the local error λ is known with a sufficient degree of accuracy, then it is easy to find the exact solution at the grid nodes Δy 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 some operator ψ, an estimator of the local error of the numerical solution y_h. Since we used $\eta = \Delta y + O(h^r)$ to estimate the magnitude of the order h^p, it would seem that at best the estimate would satisfy $\psi(\eta) = \lambda + O(h^{r+p})$. The more precise solution $\overline{\eta}$ is found as a solution of system of equations $\phi(\overline{\eta}) = \psi(\eta)$; if $\psi(\eta) = \phi(\Delta y) + O(h^{r+p})$, it can be expected, for a stable method ϕ, $\overline{\eta} = \Delta y + O(h^p)$ which gives an estimate of the global error $\eta - \overline{\eta}$ accurate to $O(h^{r+p})$. So, deferred correction reduces the global error estimation problem to the local error estimation problem. It can be assumed that the operator ψ must satisfy the relation $\psi(\Delta y) = \lambda + O(h^{r+p})$ However, this is not enough, sinceϕ it satisfies this condition precisely and one cannot be sure what $\varphi(y_h)$ is a suitable estimate, especially if y_h is found when solving the equation $\varphi(y_h) = 0$.

It is necessary to put the second condition on psi so that the error $O(h^r)$ when calculating y_h is reduced by $O(h^p)$. Many numerical experiments have shown that this condition is usually satisfied if $\psi(\Delta z) = O(h^p)$ for an arbitrary sufficiently smooth function z

As an example of the estimated function of the local error of the Euler method, we consider

$$\psi(y_h)_i := \begin{cases} 0, & \text{i} = 0 \\ -\left(\frac{h}{2}\right) \cdot \left(\frac{f(t_i, y_{h,i}) - f(t_{i-1}, y_{h,i-1})}{h}\right), & \text{i} = 1, ..., N. \end{cases} \tag{4}$$

For an arbitrary function z

$$\psi(\Delta z)_n = -\left(\frac{h}{2}\right)\left(f_t(\tau_n^{'}, z(\tau_n^{'})) + f_z(\tau_n^{'}, z(\tau_n^{'}))z'(\tau_n^{'})\right) O(h)$$

and with $z = y$

$$\psi(\Delta y)_n = -\left(\frac{h}{2}\right) y''(\tau_n^{'}) = \lambda_n + O(h^2).$$

As a result, delayed correction has three components: a numerical solution y_h, an effective numerical method ϕ, and an estimated local error function ψ An improved solution $\overline{y_h}$ is calculated on the basis of the equation $\varphi(\overline{y_h}) = \psi(y_h)$ and satisfies the relation $\overline{y_h} = \Delta y + O(h^{r+p})$, provided that for an arbitrary function $y_h = \Delta y + O(h^r)$, $\psi(\Delta y) = \varphi(\Delta y) + O(h^{r+p})$ and $\psi(\Delta z) = O(h^p)$

3. Regularization in constructing an estimate of an approximate solution using a defect change

We define the operator equation

$$Az = u, z \in Z, u \in U, \tag{5}$$

where $Z, U-$ are some metric spaces. Problem (5) is called correctly posed if the following conditions are satisfied: 1) the operator equation (5) is solvable for any right-hand side $u \in U$; 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 [9].

The class of ill-posed problems is very wide; these include the problems of summing Fourier series with coefficients given with errors, some optimal control problems, some linear algebra problems, and minimizing functionals. The experience of using direct and backward error analysis methods gives reason to believe that these methods are also incorrect problems; for the backward analysis, the use of regularization of the problem is justified.

Suppose that the operator equation (5) is being solved with the right-hand side $\overline{u}$, having a unique solution $\overline{z} \in Z$. As a rule, the right-hand side of $\overline{u}$ is unknown , we can use its approximate value $u_\delta \in U$, for which the inequality $\rho(u_\delta, \overline{u}) \leq \delta$. holds in the metric of the space U . In this inequality, the value δ is called the error of the element u_δ. For a regularized problem, we can construct a family R_δ of such operators depending on the parameter δ, such that $z_\delta = R_\delta u_\delta \overset{Z}{\to} \overline{z}$ at $\delta \to 0$. 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 map A onto the set A_M is one-to-one. In this case, the backward mapping A^{-1}, defined on A_M, is continuous, then for the approximate solution we can take an arbitrary element $z_\delta \in M$, such that $\rho(Az_\delta, u_\delta) \leq \delta$.

It is known [9],[10] that the element $z_\delta \in M$, minimizing the defect $\rho(Az, u_\delta)$ on the set M is called the quasisolution of the equation (4):

$$\rho(Az_\delta, u_\delta) = \inf_{z \in M} \rho(Az, u_\delta). \tag{6}$$

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

Let us analyze the regularization of the error estimation operator for numerical solutions of systems of differential equations (1), where $y \in R^n$, $t \in R$ and $f : R^n \times R \to R^n$. The first type of error analysis is defect analysis (2), which operates by modifying the equation and imposing fixed conditions on the equation. This is the most natural form of backward analysis for ODEs. 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 (2), considered as a non-autonomous perturbation of ODEs (1) [11],[12], [13]. 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

$$\frac{du}{dt} = f(t,u) + \varepsilon v(t), \tag{7}$$

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 ε -close problem. Depending on the problem, it may be more appropriate to consider the relative defect

$$\frac{du}{dt} = f(t,u)\left(1 + \varepsilon v(t)\right),$$

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.

4. 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 calculated using the one-step Runge- Kutta method of the fourth order and the multi-step Adams method of the fourth order. Given that the exact solution to the system is known and does not necessitate the use of numerical methods aimed at compensating for complex areas of solutions (for example, singularities or stiffnesses), an error estimate (global error) was obtained for the system. We used deferred difference correction and Richardson extrapolation [14], as well as a reverse error analysis. Estimates calculated using reverse error analysis look like this:

$$err(10) = 0.456148217 \cdot 10^{-6}, \qquad err(100) = 0.6465776799 \cdot 10^{-5}$$
$$err(1000) = 0.78572562324 \cdot 10^{-4}, \qquad err(10000) = 0.187345329827 \cdot 10^{-3}$$

B) Let us solve the ODE system

$$\frac{dy_1(t)}{dt} = y_2(t),\ \frac{dy_2(t)}{dt} = y_3(t),\ \frac{dy_3(t)}{dt} = -y_2(t) - 0.5y_1^2(t) + 1$$

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.

Error estimates calculated using delayed difference correction or Richardson extrapolation are equal at $t = 1$

$$err_{y1}(1) = .617423026019743194 - .617423026020436306 \approx 10^{-16},$$
$$err_{y2}(1) = .1.29539072722494274 - .1.29539072723032510 \approx 10^{-16},$$
$$err_{y3}(1) = 1.34603247053568098 - 1.34603247053760988 \approx 10^{-12}$$

Error estimates calculated using delayed difference correction or Richardson extrapolation are equal at $t = 7.5$

$$err_{y1}(7.5) \approx 10^{-5}, err_{y1}(7.5) \approx 10^{-4}, err_{y1}(7.5) \approx 10^{-3}.$$

Error estimates calculated using the backward error analysis are equal at $t = 7.5$

$$err_{y1}(7.5) \approx 10^{-6}, err_{y1}(7.5) \approx 10^{-6}, err_{y1}(7.5) \approx 10^{-5}.$$

5. Conclusion

The experience of estimating errors in numerical solutions confirms that one of the advantages of the inverse error analysis is that applying it to a well-defined problem with a small inverse error leads to a small direct error (global error for numerical solutions to ODEs). Finding a certain set of finite diameter to which the exact solution belongs (at least for sufficiently small errors) and estimating the distance from the approximate solution to its boundary allows us to reduce the influence of the incorrectness of the inverse error analysis, which manifests itself in the absence of stability. We can characterize the analysis of errors in numerical solutions as an incorrectly posed problem. The regularization of the operator of the problem to be solved, which consists in replacing the original system with systems of a simpler form, helps to obtain more accurate error estimates.

References

[1] Stetter H 1973 *Analysis of discretization methods for ordinary differential equations* (Springer Verlag: Berlin, Heidelberg, New York)
[2] Kalitkin N N, Yukhno L F, Kuzmina L V 2011 A quantitative criterion for conditioning systems of linear algebraic equations *Mathematical Modeling* V 23 pp 3-6 (In Russ.)
[3] Wilkinson J 1963 *Rounding errors in algebraic processes* (London, Her Majesty's Stationary Office)
[4] Voevodin V V 1967 *On the asymptotic distribution of rounding errors in linear transformations* J. Vychisl. Matem. Matem. Phys. V 7 pp. 965-976 (in Russ.)
[5] Voevodin V V 1969 *Rounding errors and stability in direct methods of linear algebra* (Moskva:MGU Publ.) 153 (in Russ)
[6] Rogalev A N 2017 Backward problems of error estimation of numerical solutions *Proc. International Conference on Computational and Applied Mathematics "VPM17"* http://conf.nsc.ru/cam17/ru/proceedings pp. 739-743
[7] Rogalev A N, Doronin S V, 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* V 23 pp 88-101 (In Russ.)
[8] Skeel R 1981 A theoretical framework for proving accuracy results for deferred corrections *SIAM J. Numer.Anal.* v 19 pp 171 -196
[9] Tikhonov A N, Goncharsky AV, Stepanov VV, Yagola A G 1983 *Regularization algorithms and a priori information* (Moscow: Science) 200 p (In Russ.)
[10] Korolev Yu M, Yagola A G 2012 Error estimation in linear inverse problems in the presence of a priori information *Comput. Meth. and Program.* V 13 pp 14-18 (In Russ.)
[11] Reich S 1999 Backwards error analysis for numerical integrators *SIAM J. Numer. Anal.* V 36 pp 1549-70
[12] Corless R 1992 Defect-controlled numerical methods and shadowing for chaotic differential equations *Physica D: Nonlinear Phenomena* V 60 pp 323 -334
[13] Corless R 1994 What good are nimerical simulations of chaotic dynamical systems *Computers & Mathematics with Applications* V 28 pp 107 -121
[14] Skeel R 1986 Thirteen ways to estimate global error *SIAM J. Numer.Anal.* v 48 pp 1 -20