ISSN 2070-0482, Mathematical Models and Computer Simulations, 2015, Vol. 7, No. 1, pp. 24–35. © Pleiades Publishing, Ltd., 2015.
Original Russian Text © N.N. Kalitkin, I.P. Poshivaylo, 2014, published in Matematicheskoe Modelirovanie, 2014, Vol. 26, No. 7, pp. 3–18.

Solving the Cauchy Problem with Guaranteed Accuracy for Stiff Systems by the Arc Length Method

N. N. Kalitkin and I. P. Poshivaylo

Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Miusskaya pl. 4, Moscow, 125047 Russia
e-mail: kalitkin@imamod.ru, ilya.poshivaylo@gmail.com

Received June 24, 2013

Abstract—The arc length method is an efficient way of solving the Cauchy problem for systems of ordinary differential equations that have areas with large right-hand sides (stiff and ill-conditioned problems). It is shown how to get a posteriori asymptotically exact error estimation for such calculations by using thickening of nets and the Richardson method. The examples of calculations demonstrate that with the transition to the arc length, the greater the stiffness of a problem or its ill conditionality, the larger the gain in the accuracy. This gain can reach many orders of magnitude. It is shown that in order to get reliable results for hyperstiff problems (characterized by large difference in scales of speeds of various processes), it is necessary to make calculations with high digit capacity and/or by an analytical expression of a Jacobian matrix.

Keywords: stiff systems, arc length of integral curve, ODEs
DOI: 10.1134/S2070048215010044

1. PROBLEM

Consider the Cauchy problem for the system of ordinary differential equations (ODEs)

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}, t), \quad \mathbf{u}(0) = u_0, \quad 0 \le t \le T. \tag{1}$$

Here, t is a scalar argument and $\mathbf{u} = \{u_m, 1 \le m \le M\}$, $\mathbf{f} = \{f_m, 1 \le m \le M\}$ are M-dimensional vector functions. The numerical solution of this problem is still important and not simple. For example, assume that (1) describes the combustion reaction of methane in a coal mine. The fire originates from a random spark and at a low concentration of methane at first the fire is slow (it does not yet pose a serious danger). However, at some point in time the burning can dramatically accelerate and transit to detonation; this causes death and destruction. Identifying the instant of the transition is an ill-conditioned problem. Evidently, it is necessary to be able to solve such problems with the guaranteed mathematical error estimation.

The difficulties of problem (1) can be various: ill conditionality (integral curves quickly diverge), stiffness (integral curves quickly converge), and strong oscillations. In practice these difficulties often are not distinguished and are spoken of as "stiffness" [1]. All these difficulties are characterized by the large right-hand sides of (1) and the rapid changes of functions at some points in time. In the grid calculations one usually tries to find these points and significantly refine the grid steps in them. This is made by programs with automatic step selection. At the same time, the methodical calculations show that these programs do not guarantee the accuracy specified by the user (the difference often comes to three or four orders of magnitude).

A promising method for solving such problems consists of introducing a new argument, namely, the arc length of an integral curve in the $(M + 1)$-dimensional space $\{u_0 \equiv t, u_1, \ldots, u_M\}$. This method is proposed in [2] and since 1993 has been developed in detail by E.B. Kuznetsov in a cycle of works, including the monograph [3]. In particular, Kuznetsov proved the theorem stating that the introduction of the arc length provides the best conditionality of the Cauchy problem. Foreign publications on this topic are lacking (see [4]).

However, these works do not describe how to find the guaranteed error estimation of an obtained solution. Modern programs with automatic step selection do not make it possible to do this. The only way to get a posteriori asymptotically exact error estimation is by thickening the net by using the Richardson method [5, 6]. In this paper it is shown how to use the thickening of nets in the arc length method. Calculations with guaranteed error estimations demonstrate that with the transition to the arc length, the

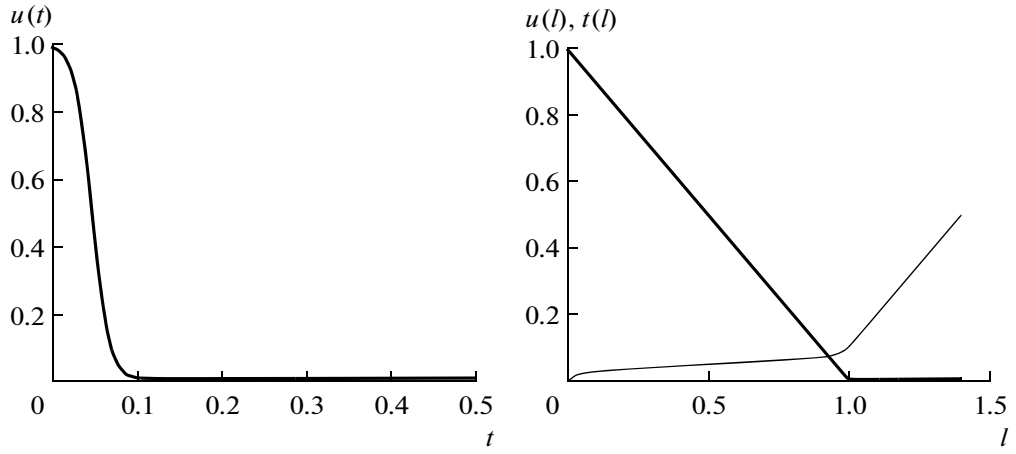

Fig. 1. (a) Graph of $u(t)$; (b) the thick line is the graph of $u(l)$ and the thin line is the graph of $t(l)$.

harder the problem (the greater the stiffness of initial problem (1) or its ill conditionality), the larger the gain in the accuracy. This gain can reach many orders of magnitude.

2. EQUATIONS

The arc length is determined by the relation

$$dl^2 = \sum_{m=0}^{M} du_m^2, \quad du_0 \equiv dt. \tag{2}$$

Assuming that l is a new argument, we obtain instead of (1) the following system:

$$\frac{du_m}{dl} = F_m(\mathbf{u}) \equiv \frac{f_m}{\sqrt{\sum_{m=0}^{M} f_m^2}}, \quad 0 \le m \le M, \quad f_0 \equiv 1. \tag{3}$$

The right-hand sides of (3) $F_m(\mathbf{u})$ do not depend on the new argument l; therefore, system (3) is autonomous. The relation $\sum_{m=0}^{M} F_m^2 = 1$ is true. Consequently, the right-hand sides of F_m are small and system (3) is smooth even if system (1) was ill-conditioned or stiff. System (1) should be integrated to the instant T. However, the value of $l = L$ to which system (3) needs to be integrated is not known. Since $F_0 > 0$, then $u_0 \equiv t$ is a monotonic increasing function of l. Hence, it is necessary to integrate (3) until condition $u_0(l) \ge T$ holds.

Let us illustrate the idea of the transition to the arc length. Consider an equation of the following form:

$$du/dt = -\lambda u(1-u), \quad \lambda \gg 1, \quad 0 < u_0 < 1. \tag{4}$$

For the sake of definiteness we take $\lambda = 100$ and $u_0 = 0.99$. The graph of the solution of Eq. (4) is presented in Fig. 1a: the function $u(t)$ decreases sharply until $u(t)$ is little different from a discontinuous function. After the transition to the arc length in accordance with formulas (3) (see Fig. 1b), the function $u(l)$ instead of the jump has the slope of 45°. The almost discontinuous curve $u(t)$ becomes the almost broken curve $u(l)$. We can see that it is much easier to integrate the latter curve numerically.

3. TEST PROBLEM

As a substantially difficult example from the set of tests in [1], we take the Van der Pol equation. It is the equation of an oscillator with nonlinear viscosity

$$\begin{cases} du/dt = v, \\ dv/dt = -\omega^2 u - \sigma(u^2 - 1)v. \end{cases} \tag{5}$$

If $\sigma = 0$, this is just a harmonic oscillator. If $\sigma > 0$, the solution in the phase plane is also a closed curve, but its form essentially differs from an ellipse. If $\sigma \gg 1$, problem (5) is difficult; here, the nature of the difficulty

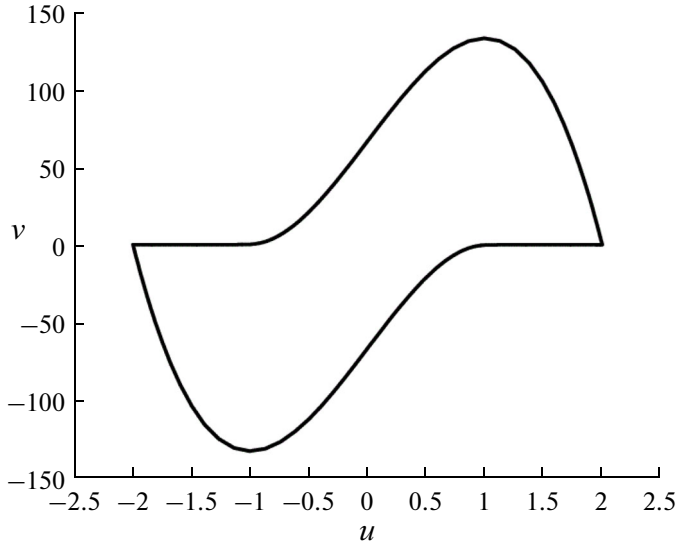

Fig. 2. Profile of the solution of Eq. (5) for $\sigma = 100$.

varies on different sections of the cycle. In some sections the problem is stiff, in others it is ill-conditioned, and in the third sections in the spectrum of the Jacobian large imaginary parts appear. The profile of the solution of Eq. (5) for $\sigma = 100$ in the phase variables $v(u)$ is presented in Fig. 2.

For the numerical integration the following schemes were used. From the family of explicit schemes we took the Runge–Kutta scheme of the second order of accuracy (*ERK*2)

$$\begin{cases} \mathbf{p}_1 = \mathbf{f}(\mathbf{u}), \\ \mathbf{p}_2 = \mathbf{f}(\mathbf{u} + 2\tau\mathbf{p}_1/3), \\ \hat{\mathbf{u}} = \mathbf{u} + \tau(\mathbf{p}_1 + 3\mathbf{p}_2)/4 \end{cases} \tag{6}$$

and the classical explicit Kutta scheme of the fourth order of accuracy (*ERK*4)

$$\begin{cases} \mathbf{p}_1 = \mathbf{f}(\mathbf{u}), \\ \mathbf{p}_2 = \mathbf{f}(\mathbf{u} + \tau\mathbf{p}_1/2), \\ \mathbf{p}_3 = \mathbf{f}(\mathbf{u} + \tau\mathbf{p}_2/2), \\ \mathbf{p}_4 = \mathbf{f}(\mathbf{u} + \tau\mathbf{p}_3), \\ \hat{\mathbf{u}} = \mathbf{u} + \tau(\mathbf{p}_1 + 2\mathbf{p}_2 + 2\mathbf{p}_3 + \mathbf{p}_4)/6. \end{cases} \tag{7}$$

Both of these schemes cannot possess the *A*-stability and are assumed to be unusable for stiff problems. From the family of explicit–implicit Rosenbrock schemes, we took a single-stage scheme with a complex coefficient of the second order of accuracy (*CROS*) [7]

$$\begin{cases} \left(E - \tau\dfrac{1+i}{2}\mathbf{f}_u\right)\mathbf{k} = \mathbf{f}(\mathbf{u}), \\ \hat{\mathbf{u}} = \mathbf{u} + \tau\,\mathrm{Re}\,\mathbf{k} \end{cases} \tag{8}$$

and the analogous two-stage scheme of the fourth order of accuracy (*CROS*4) [8]. This two-stage scheme is relatively bulky; therefore, we do not present it here. The *CROS* scheme is *L*2-stable and the *CROS*4 scheme is *L*4-stable. Such stability characteristics are unique: they must provide the rapid damping of stiff components of a solution. In addition, these schemes are noniterative; hence, they are economical and simple to implement.

From the family of fully implicit Runge–Kutta schemes we took the classic implicit Euler scheme of the first order of accuracy

$$\hat{\mathbf{u}} = \mathbf{u} + \tau\mathbf{f}(\mathbf{u}) \tag{9}$$

and the recursion inverse scheme of the second order of accuracy [9]

$$\hat{\mathbf{u}} = \mathbf{u} + \tau\mathbf{f}\left(\hat{\mathbf{u}} - \frac{\tau}{2}\mathbf{f}(\hat{\mathbf{u}})\right). \tag{10}$$

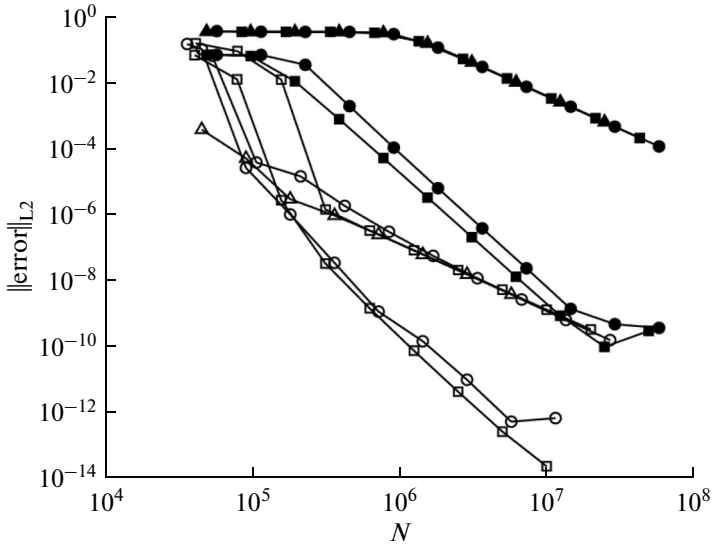

Fig. 3. Problem (5) for $\sigma = 100$. Black markers for argument t and light markers for the argument l; small squares for both *ERK* schemes, circles for both *CROS* schemes, and triangles for the recursion scheme.

These schemes are also $L1$-stable and $L2$-stable, respectively. Moreover, they are the most reliable schemes for solving stiff problems, due to the Newtonian iterations until convergence.

The calculations were made for argument t and for argument l on a sequence of uniform nets recurrently thickening twice. The exact solution of problem (5) is not expressed in elementary functions. Hence, the accuracy can only be checked by comparing the numerical solutions on pairs of neighboring nets. We mark the coinciding nodes of such nets with subscript n. The values of the net solutions on the first and second (more detailed) nets we denote by v'_n and v''_n (for definiteness we take velocity v, because its behavior is more important for the problem in question). It is convenient to use the following analog of the Hilbert norm:

$$\delta = \frac{1}{2^p - 1}\sum_n \frac{(v'_n - v''_n)^2}{v''^2_n}. \tag{11}$$

Outside the sum we have the conventional Richardson factor, where p is the order of accuracy of a scheme. The summation over n is produced by one cycle. We can consider δ as a mean-root square relative error.

3.1. Comparing the Schemes

The schemes were compared for $\sigma = 100$; this is considered to be quite a difficult problem (at all times it was taken $\omega = 1$). Calculations with a large step were impossible: the closure of a cycle did not occur in any of the schemes. Therefore, it was necessary to check the availability of the closure visually. The results of calculating the error are presented in Fig. 3. Here, on the abscissa the numbers of nodes for one cycle for argument t and for argument l are plotted. Let us discuss these results.

Each curve has a good straight (regular) segment, whose slope corresponds to the theoretical order of accuracy of a scheme. This demonstrates that the Richardson method is applicable, so that the obtained error estimates are asymptotically exact. Hence, various schemes can be reliably compared.

The upper line is the confluence of three curves for schemes of the second order in argument t. At first sight this seems strange: explicit and implicit schemes give the same accuracy, although explicit schemes are said to be bad for stiff problems. The reason is that (5) is not a pure stiff problem. In (5), sections of the ill conditionality and the complexity of the spectrum of the Jacobian matrix are longer than sections of the actual stiffness. In contrast, implicit schemes are efficient for pure stiff problems.

The third curve corresponds to the same schemes of the second order of accuracy, but in argument l. These schemes also give almost identical results. However, here the error is $\sim 10^6$ times smaller than in argument t. In this case, if we require the identical accuracy in the calculations, then it is necessary to take nodes along the arc length so that their number is smaller by a factor of $\sim 10^3$ (since these are schemes of

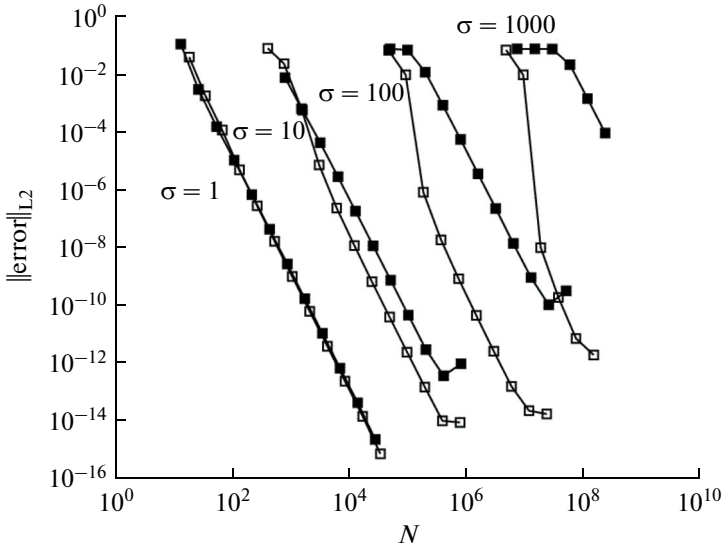

Fig. 4. Problem (5) and the *ERK*4 scheme; (■) calculations with respect to t and (□) calculations with respect to l. Near the lines values of σ are specified.

the second order of accuracy). This clearly shows that integrating over the arc length can give a sufficient quantitative gain.

For schemes of the fourth order the conclusions are qualitatively analogous. Here, curves for explicit and implicit schemes differ only slightly. The explicit Kutta scheme is just more accurate than the *CROS*4 scheme due to small coefficient in the remainder term. The difference in the accuracy for the identical numbers of nodes in the arguments t or l is $\sim 10^5$; this is somewhat less than for schemes of the second order. The labor input for identical accuracy differs by a factor of ~18. However, we cannot conclude that the arc length method is less useful for schemes of the fourth order. It is evident that schemes of the fourth order in l make it possible to attain the highest possible accuracy, namely, the accuracy of the rounding error (the break in the graph with the transition in a horizontal line).

3.2. Effect of the Stiffness

The effect of the stiffness is illustrated only by one explicit Kutta scheme of the fourth order (by the other schemes we have analogous results). In the calculations the values of the parameter from $\sigma = 1$ to $\sigma = 1000$ were taken. The first and the last values correspond to soft and very stiff problems, respectively. The results are presented in Fig. 4.

The difficulty of the problem is attested by the minimum number of nodes N whereby the closure of a cycle can be obtained. For $\sigma = 1$ we have $N \approx 100$, for $\sigma = 10$ we have $N \approx 1000$, for $\sigma = 100$ we have $N \approx 10^5$, and for $\sigma = 1000$ we have $N \approx 10^7$. This indicates the difficulty of the problem for $\sigma = 1000$.

All curves (even for the largest $\sigma = 1000$) have straight regular segments. Their slopes correspond to the theoretical order of accuracy $p = 4$. If $\sigma = 1$, the transition from t to l gives almost no gain. If $\sigma = 10$, the gain of using the arc length is ~10 times. If $\sigma = 100$, then the gain comprises $\sim 10^5$ times, whereas if $\sigma = 1000$, then the gain is $\sim 10^9$ times. The greater the stiffness, the larger the gain in the accuracy.

Hence, the arc length method makes it possible to confidently use the thickening of a net and estimating the accuracy according to Richardson even for very stiff problems. In this case, we have a great gain in the accuracy compared to integrating with respect to time.

4. HYPERSTIFFNESS

For the Van der Pol equation the characteristic stiffness constant is σ. According to the popular opinion, the problem with $\sigma = 100$ is already sufficiently stiff and the problem with $\sigma = 1000$ is very stiff. In most foreign tests, the characteristic stiffness parameter rarely exceeds 10^4. At the same time, there exist important classes of problems in which the stiffness parameter can exceed 10^{10}. For example, a problem

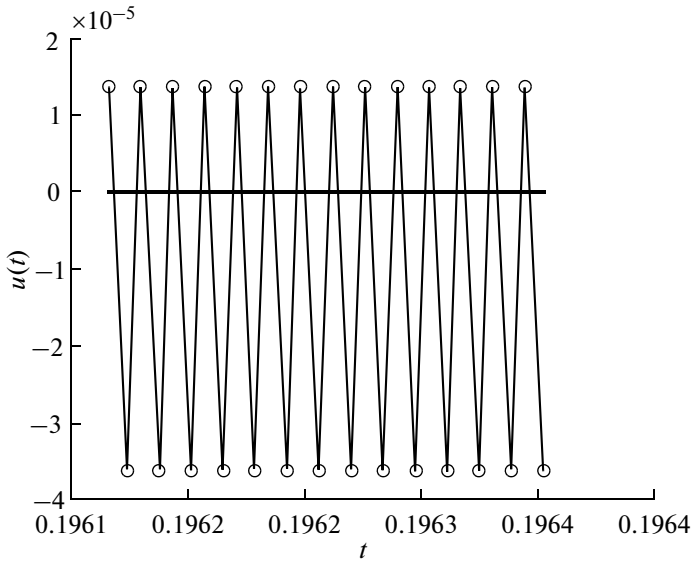

Fig. 5. Area of the solution of problem (13) by using a purely implicit Rosenbrock scheme for $l > 1$; the numerical solution (○) and the exact solution (thick line).

of this kind is the problem of chemical kinetics, since rates of chemical reactions can differ by many orders of magnitude. Such problems can be called hyperstiff (although it is more correct to call them hyperdifficult, since they include processes of ultrafast decay, as well as processes of very fast growth). They place much heavier demands on the reliability of the numerical methods.

We consider the arising difficulties by using the classic Dahlquist test with a very large constant

$$du/dt = -\lambda u, \quad 0 \le t \le 1, \quad u(0) = 1, \quad \lambda = 10^9 \gg 1. \tag{12}$$

The exact solution of this problem is $u(t) = \exp(-\lambda t)$; it is a positive function, which decreases very rapidly almost stepwise. If we go to the arc length, then (12) is replaced by the system

$$\begin{cases} du/dl = -\lambda u \Big/ \sqrt{1 + \lambda^2 u^2}, \\ dt/dl = 1 \Big/ \sqrt{1 + \lambda^2 u^2}, \end{cases} \quad u(0) = 1, \quad t(0) = 0. \tag{13}$$

The solution of this problem for $\lambda \gg 1$ is accurately described by the broken lines

$$u(l) \approx \begin{cases} 1 - l, \\ 0, \end{cases} \quad t(l) \approx \begin{cases} 0, \text{ for } 0 \le l \le 1, \\ l - 1, \text{ for } 1 < l \le 2. \end{cases} \tag{14}$$

This solution is much like the functions presented in Fig. 1. Let us require that the method gives a good solution of problem (12) with step $\tau \gg 1/\lambda$.

The Dahlquist problem for the independent variable t is successfully solved by using the well-known Rosenbrock, Rosenbrock–Wanner, and other methods [1]. However, in the transition to the arc length certain qualitative features of these methods (the monotonicity and the positiveness of the solution) are not retained. This is easily demonstrated by the numerical examples. Let us take the step in the arc length $h = 0.1$ such that the general number of steps $N = 20$. We describe the results of the calculations by using various implicit schemes.

(i) Implicit Euler scheme (9), in which the implicit algebraic system is solved by using Newtonian iterations, is said to be the most reliable one. These calculations were done with the difference computation of the Jacobian, as well as with analytical formulas. In both cases the calculations were well done to $l \approx 1$; then, the Newtonian iterations ceased to converge. Newtonian iterations in hyperstiff problems often do not converge on any implicit schemes. Hence, it is appropriate to use such schemes only for moderately stiff problems, in which the difficulty of a problem is associated not so much with the stiffness but with the nonlinearity of the problem.

(ii) It is assumed that among explicit–implicit (i.e., noniterative) schemes, the single-stage purely implicit Rosenbrock scheme is most reliable. It can be obtained from the *CROS* scheme (8), if we place 1 outside the Jacobian matrix instead of the parameter $(1 + i)/2$. One computation by such a scheme was done with the difference calculation of the Jacobian matrix and with 64-bit numbers. This calculation was

Table 1. Auxiliary variables of problem (16)

$r_1 = k_1u_1$	$r_{10} = k_{10}u_{11}u_1$	$r_{19} = k_{19}u_{16}$
$r_2 = k_2u_2u_4$	$r_{11} = k_{11}u_{13}$	$r_{20} = k_{20}u_{17}u_6$
$r_3 = k_3u_5u_2$	$r_{12} = k_{12}u_{10}u_2$	$r_{21} = k_{21}u_{19}$
$r_4 = k_4u_7$	$r_{13} = k_{13}u_{14}$	$r_{22} = k_{22}u_{19}$
$r_5 = k_5u_7$	$r_{14} = k_{14}u_1u_6$	$r_{23} = k_{23}u_1u_4$
$r_6 = k_6u_7u_6$	$r_{15} = k_{15}u_3$	$r_{24} = k_{24}u_{19}u_1$
$r_7 = k_7u_9$	$r_{16} = k_{16}u_4$	$r_{25} = k_{25}u_{20}$
$r_8 = k_8u_9u_6$	$r_{17} = k_{17}u_4$	
$r_9 = k_9u_{11}u_2$	$r_{18} = k_{18}u_{16}$	

excellent up to $l \approx 1$; here, u_n monotonically decreased and remained positive. However, upon the further increase of l, we saw that u_n began to alternately take values of different signs at the level of $\sim 10^{-5}$. In this case, t_n increased very slightly, while remaining close to 0 instead of approaching 1 (see Fig. 5). It was assumed that such behavior is related to rounding errors.

In order to check this assumption, the analogous calculation with a 128-bit numbers was made. The results for $l > 1$ dramatically improved. Here, u_n still remained alternating, but at a much smaller level (approximately $\pm 10^{-20}$ and less). Therewith, t_n increased in accordance with exact solution (14). All this confirms the hypothesis for the effect of rounding errors.

Still better results are obtained if the Jacobian matrix is found not by difference computations but by using exact formulas. The Jacobian matrix for problem (13) has the form

$$D = \begin{pmatrix} -\dfrac{\lambda}{(1+\lambda^2u^2)^{3/2}} & 0 \\ -\dfrac{\lambda^2u}{(1+\lambda^2u^2)^{3/2}} & 0 \end{pmatrix}. \tag{15}$$

In this case, even for 64-bit calculations excellent results are realized. Here, u_n monotonically decrease on the whole interval. For $l \le 1$, the decrease is linear in accordance with exact solution (14). For $l > 1$, u_n are very small and decrease in a geometric progression with ratio $(h\lambda)^{-1} = 10^{-8}$. Such behavior of a numerical solution can be considered as the standard behavior.

(iii) For the *CROS* scheme (8) with the difference calculation of the Jacobian matrix, the results are almost indistinguishable from the above-described results both of the 64-bit calculations and of the 128-bit calculations. However, with the analytical calculation of the Jacobian matrix and with the 64-bit calculations the results are somewhat worse than the standard results: for $l_n = 1.1$ the numerical solution becomes negative, $u_n \approx -10^{-30}$, and then remains negative, while rapidly decreasing in magnitude. This shows the slightly lower reliability of the *CROS* scheme even for the simplest Dahlquist test. For more difficult problems this effect can be stronger.

5. CHEMICAL KINETICS

Hyperstiff problems include tasks that describe changes in the concentrations of substances during chemical reactions: problems of chemical kinetics. In such problems both slow and very fast chemical reactions usually take place at the same time. In addition, practical problems bring about systems of equations of large dimensionality. Let us present a meaningful example.

5.1. Problem Statement

Consider a test from [11], which is taken from [12]. In the original work, burning a natural gas in air is described; here, the natural gas is methane with sulfur pollutions. In the test set [11] not a complete system, but only a part of it is presented; this part is responsible for the emission of harmful components in

Table 2. Constants of reactions of problem (16)

$k_1 = 0.350$	$k_{10} = 0.900 \times 10^4$	$k_{19} = 0.444 \times 10^{12}$
$k_2 = 0.266 \times 10^2$	$k_{11} = 0.220 \times 10^{-1}$	$k_{20} = 0.124 \times 10^4$
$k_3 = 0.123 \times 10^5$	$k_{12} = 0.120 \times 10^5$	$k_{21} = 0.210 \times 10$
$k_4 = 0.860 \times 10^{-3}$	$k_{13} = 0.188 \times 10$	$k_{22} = 0.578 \times 10$
$k_5 = 0.820 \times 10^{-3}$	$k_{14} = 0.163 \times 10^5$	$k_{23} = 0.474 \times 10^{-1}$
$k_6 = 0.150 \times 10^5$	$k_{15} = 0.480 \times 10^7$	$k_{24} = 0.178 \times 10^4$
$k_7 = 0.130 \times 10^{-3}$	$k_{16} = 0.350 \times 10^{-3}$	$k_{25} = 0.312 \times 10$
$k_8 = 0.240 \times 10^5$	$k_{17} = 0.175 \times 10^{-1}$	
$k_9 = 0.165 \times 10^5$	$k_{18} = 0.100 \times 10^9$	

the atmosphere and involves 25 chemical equations and 20 components. The system has canonical form (1). The right-hand sides of the system are presented as follows:

$$\mathbf{f} = \begin{pmatrix} -\sum\limits_{j\in\{1,10,14,23,24\}} r_j + \sum\limits_{j\in\{2,3,9,11,12,22,25\}} r_j \\ -r_2 - r_3 - r_9 - r_{12} + r_1 + r_{21} \\ -r_{15} + r_1 + r_{17} + r_{19} + r_{22} \\ -r_2 - r_{16} - r_{17} - r_{23} + r_{15} \\ -r_3 + 2r_4 + r_6 + r_7 + r_{13} + r_{20} \\ -r_6 - r_8 - r_{14} - r_{20} + r_3 + 2r_{18} \\ -r_4 - r_5 - r_6 + r_{13} \\ r_4 + r_5 + r_6 + r_7 \\ -r_7 - r_8 \\ -r_{12} + r_7 + r_9 \\ -r_9 - r_{10} + r_8 + r_{11} \\ r_9 \\ -r_{11} + r_{10} \\ -r_{13} + r_{12} \\ r_{14} \\ -r_{18} - r_{19} + r_{16} \\ -r_{20} \\ r_{20} \\ -r_{21} - r_{22} - r_{24} + r_{23} + r_{25} \\ -r_{25} + r_{24} \end{pmatrix}. \tag{16}$$

Expressions for the auxiliary variables r_j and the constants of reactions k_j are presented in Tables 1 and 2. The initial data are taken at the end of the main combustion and correspond to the afterburning:

$$\mathbf{u}_0 = (0;0.2;0;0.04;0;0;0.1;0.3;0.001;0;0;0;0;0;0;0;0.007;0;0;0). \tag{17}$$

For the interval of integrating with respect to time we take $0 \le t \le 1.2$.

Figure 6 shows the graphs of changes in some concentrations. We note that the scales of these concentrations are widely different.

5.2. Requirements for Difference Schemes

The explicit Runge–Kutta schemes proved to be unsuitable for solving such problems. When integrating with respect to time the calculations fall apart at the initial steps even with the very small steps $h_t \sim 10^{-5}$: the

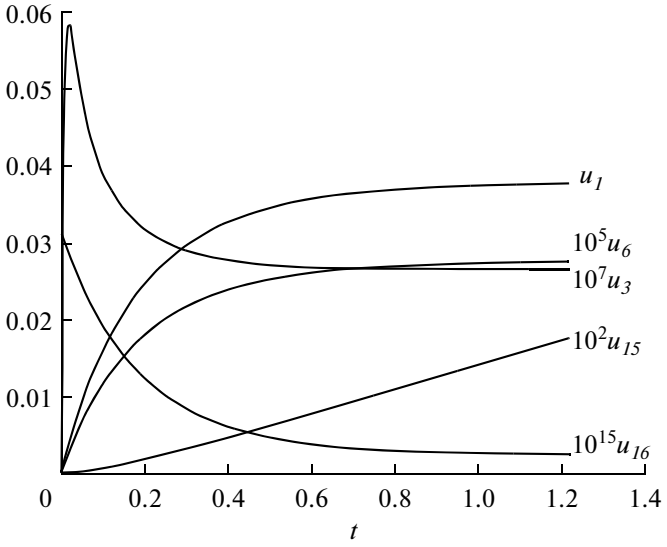

Fig. 6. Graphs of some concentrations of problem (16).

values of the concentrations increase in magnitude and exceed the numbers that can be represented on a computer.

On transition to the arc length the nature of the difficulties changes. The first step after the initial approximation gives a plausible solution. However, at subsequent steps, time t barely grows. Apparently, the difficulties are explained by the fact that the explicit Runge–Kutta schemes are nonmonotonic. The calculated graphs of the concentration become sawtoothed analogously to the graphs presented in Section 4 (see Fig. 5). Here, the amplitude of a saw can be some orders of magnitude greater than the exact values of the concentrations. As a consequence of this, the concentrations can even take negative values, which are chemically meaningless. Large concentrations lead to large values of the right-hand sides $\mathbf{f}(\mathbf{u})$. They are included in the denominator of the right-hand side $\mathbf{F}$ for function $t(l)$. As a result, this right-hand side is found to be vanishingly small and the calculated t barely grows from step to step.

This analysis demonstrates that in order to solve such problems it is necessary to use monotonic schemes. The monotonic schemes [10] include *CROS* schemes (8), the implicit Euler scheme (9), and the generally optimum inverse Runge–Kutta schemes [9]. When integrating both with respect to t and with respect to l, they give a good qualitative behavior of the numerical solution with no saws. These considerations are true in calculations with infinite digit capacity. If the digit capacity of calculations is finite, then one detail arises. It is impossible to carry out calculations with very different scales of constants of reactions with 32-digit numbers: rounding errors become comparable to the solution itself. With 64-digit numbers, good calculations can be done if the Jacobian matrix of the system is presented analytically. However, if the Jacobian matrix is found by difference calculations, 128-digit numbers are to be used for it (the obtained linearized system can also be solved with 64-digit numbers). This detail is especially sufficient if the argument is the arc length.

Balances. In the chemical reactions, molecules are formed and decomposed, but in this case the total number of atoms of each chemical element remains fixed. For a good qualitative behavior of a numerical solution, it is necessary that in the numerical calculation the number of atoms of each sort is also kept. Such balance relations are the first integrals of a system of chemical equations.

Each equation of a balance can be presented in the following way. Assume that each molecule u_j contains α_j atoms of a certain element. Then the total number of atoms in a system $\sum_j \alpha_j u_j$ is independent of time. One can readily see that here $\sum_j \alpha_j f_j = 0$: the number of atoms that come from some molecules is equal to the number of atoms that go to other molecules. It is convenient to present these relations in a matrix form. Assume that we have a column vector composed of components α_j; it can be considered as a rectangular matrix. Then the exact solutions fit the following balance relations:

$$\boldsymbol{\alpha}^T \mathbf{f} = 0, \quad \boldsymbol{\alpha}^T \mathbf{u} = 0, \tag{18}$$

where the multiplication of row $\boldsymbol{\alpha}^T$ by column $\mathbf{f}$ or $\mathbf{u}$ is performed according to the rules of matrix multiplication. The number of various columns $\boldsymbol{\alpha}$ is equal to the number of various chemical elements included in the chemical reactions; the initial system must fit all these balances. Let us demonstrate that the schemes used for the calculations in this work maintain the chemical balances of the system.

Theorem 1. *Runge–Kutta schemes maintain the chemical balance of the system.*

Proof. It is well known that formulas for the s-stage Runge–Kutta methods are presented in the general form as follows:

$$\hat{\mathbf{u}} - \mathbf{u} = \tau \sum_{k=1}^{s} b_k \mathbf{w}_k, \quad \mathbf{w}_k = \mathbf{f}\left(\mathbf{u} + \tau \sum_{l=1}^{L} a_{kl} \mathbf{w}_l, t + \tau a_k \right). \tag{19}$$

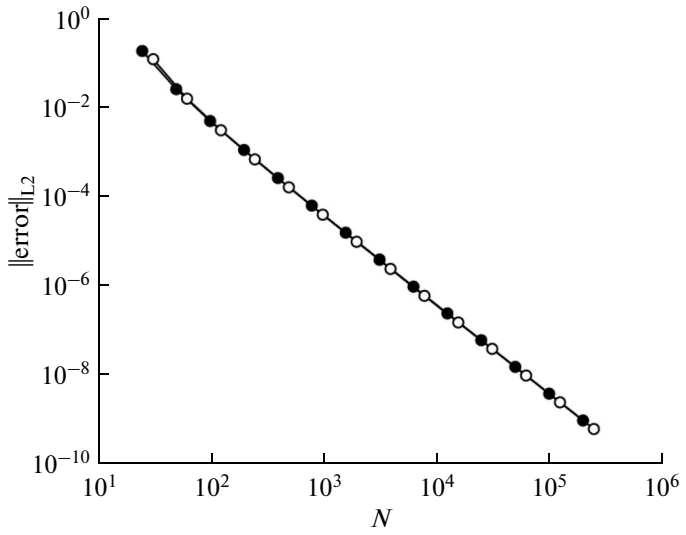

Fig. 7. Calculating for problem (16) by the *CROS* scheme on a series of thickening nets; (●) integrating with respect to time and (○) integrating with respect to the arc length.

We multiply from the left the first expression by $\boldsymbol{\alpha}^T$:

$$\boldsymbol{\alpha}^T\hat{\mathbf{u}} - \boldsymbol{\alpha}^T\mathbf{u} = \tau\sum_{k=1}^{s} b_k\boldsymbol{\alpha}^T\mathbf{w}_k. \tag{20}$$

Note that the expression for $\mathbf{w}_k$ is $\mathbf{f}$ of the shifted argument (see the second expression in (19)); here, this shift is fixed at each stage. Expression (18) must hold for any argument $\mathbf{f}$; therefore, $\boldsymbol{\alpha}^T\mathbf{w}_k = 0$. Hence, the whole sum in the right-hand side of expression (20) becomes zero. Consequently, $\boldsymbol{\alpha}^T\hat{\mathbf{u}} = \boldsymbol{\alpha}^T\mathbf{u}$; i.e., the chemical balance of the system is kept.

Note: The proof holds for the explicit, as well as for the diagonally implicit and fully implicit Runge–Kutta schemes (in particular, inverse ones [9]).

Theorem 2. *Single-stage Rosenbrock schemes maintain the chemical balance of the system.*

Proof. The proof is nontrivial in contrast to the proof of Theorem 1. In the general case, the formulas for the family of single-stage Rosenbrock schemes have the following form:

$$\begin{cases}\left(E - a\tau\mathbf{f}_u\right) = \mathbf{f}(\mathbf{u}),\\ \hat{\mathbf{u}} = \mathbf{u} + \tau b\mathbf{w}.\end{cases} \tag{21}$$

Here, E is a unit matrix, while a and b are the scalar parameters of a scheme. Let us multiply from the left the first equation from (21) by the row $\boldsymbol{\alpha}^T$:

$$\boldsymbol{\alpha}^T\left(E - a\tau\mathbf{f}_u\right)\mathbf{w} = \boldsymbol{\alpha}^T\mathbf{f} = 0. \tag{22}$$

Removing the brackets in the left-hand side, we get

$$\boldsymbol{\alpha}^T\mathbf{w} - a\tau\boldsymbol{\alpha}^T\mathbf{f}_u\mathbf{w} = 0. \tag{23}$$

From the condition $\boldsymbol{\alpha}^T\mathbf{f} = 0$ it follows that $\boldsymbol{\alpha}^T\mathbf{f}_u = 0$. Hence, we have

$$\boldsymbol{\alpha}^T\mathbf{w} = 0. \tag{24}$$

Let us multiply from the left the second equation from (21) by the row $\boldsymbol{\alpha}^T$:

$$\boldsymbol{\alpha}^T\hat{\mathbf{u}} = \boldsymbol{\alpha}^T\mathbf{u} + \tau b\boldsymbol{\alpha}^T\mathbf{w}. \tag{25}$$

In view of (24), we get $\boldsymbol{\alpha}^T\hat{\mathbf{u}} = \boldsymbol{\alpha}^T\mathbf{u}$; i.e., the chemical balance of the system is maintained.

Note: The proof holds for any value of parameter a (including a complex value) and the *CROS* scheme (8).

Generalization. Theorem 2 is generalized on the multistage Rosenbrock schemes, including schemes with complex coefficients.

In this way, the Runge–Kutta and Rosenbrock schemes fit the balance relation. Consequently, they are ***conservative*** in the sense given to this term by A.A. Samarskii.

In addition to difference schemes that maintain the chemical balance of a system, the same feature is possessed by the transform considered here of the transition to the arc length (3). We formulate this statement in the form of a theorem.

Theorem 3. *The transition to the arc length maintains the chemical balance of the system.*

Proof. In order to prove this theorem, it is necessary to demonstrate that if for the initial problem $\boldsymbol{\alpha}^T\mathbf{f} = 0$, then after the transition to the arc length, we shall have $\boldsymbol{\alpha}^T\mathbf{F} = 0$. Then schemes that keep the balance for the initial Cauchy problem will also maintain the chemical balance after the transition to the arc length. In fact,

$$\boldsymbol{\alpha}^T\mathbf{F} = \boldsymbol{\alpha}^T \frac{\mathbf{f}}{\sqrt{\sum_{m=0}^{M} f_m^2}}. \tag{26}$$

Since the root in the denominator of the fraction is the same for all components and $\boldsymbol{\alpha}^T\mathbf{f} = 0$, then the entire expression $\boldsymbol{\alpha}^T\mathbf{F} = 0$ as well.

5.3. Results of Calculations

Calculations of test problem (16) were made both for argument t and for argument l. In both cases good results are obtained by using the *CROS* scheme of the second order of accuracy; here, the Jacobian matrix was found by the difference calculations by using 128-digit numbers (the analytical calculations even for argument t are too cumbersome and require the use of a symbolic-computation program, while for argument l, the analytical calculations are practically nonimplementable). The other calculations were made with 64-digit numbers; in solving the auxiliary linear system this resulted in significant time savings.

The accuracy control was exerted by a series of calculations on the sequence of nets recurrently thickening twice. In this case, a posteriori asymptotically exact error estimation by the Richardson method was performed (see [5]). In Fig. 7 the dependence of the error on the number of nodes of the net is presented in the log-log scale. The good straight segment, whose slope 2 corresponds to the theoretical order of accuracy of a scheme, demonstrates that the Richardson method is applicable here.

It is seen that markers for arguments t and l lay down practically on a common straight line. Hence, in this problem the transition to the arc length gives no gain in accuracy as compared to argument t. This demonstrates that in problems of chemical kinetics, sections of the decay of components, as well as sections of the fast growth (of the ill conditionality), present equal difficulty. Apparently, in any problem of chemical kinetics with very different scales of constants of reactions, the transition to the arc length does not result in a gain.

Recommendations. It is appropriate to solve real problems of chemical kinetics by taking time t as an argument and using the single-stage *CROS* scheme (it is also possible to use the two-stage *CROS*4 scheme). In this case, it is necessary to make the calculations with numbers of not less than 64 digits and to find the Jacobian matrix by the difference method with 128-digit numbers.

6. APPLICATION DOMAINS OF THE METHOD

The use of the arc length method leads to a system of ODEs with significantly more cumbersome right-hand sides than in the initial system. The higher the order of a system the stronger this complication, which is required even in the case where explicit schemes are used. If implicit schemes are used, then it is necessary to calculate the Jacobian matrix of the right-hand sides; this further complicates the solution (makes it much more difficult). With these considerations let us discuss classes of problems for which the arc length method is effective or ineffective. It is clear that the method in question is highly effective for ill-conditioned systems of ODEs of not too high an order, since for such problems explicit schemes are usable. In particular, this includes problems with a singularity.

For systems of ODEs in which the stiffness strongly prevails over other types of the complexity, it is necessary to use implicit schemes. The same is related to problems in which the stiffness and the ill conditionality are almost equal (problems of chemical kinetics belong to this category). For such systems, making a step along the arc length is a considerably more cumbersome procedure than integrating with respect to time. In this case the arc length method can be ineffective and better results will be obtained by using Rosenbrock schemes with complex coefficients of argument t.

Attempts have been made to use the arc length for the one-dimensional partial differential equations solved by the method of straight lines [4]. The prospects for this avenue are poor. In fact, modern-day cal-

culations must include a posteriori error estimation, which is performed by thickening the nets in x and t. However, thickening a net in x increases the order of the system of ODEs and at the same time the volume of computations of each right-hand side. The full calculation can become too cumbersome even for explicit schemes and unacceptable for implicit ones. For two-dimensional partial differential equations this method is obviously ineffective.

However, the above-specified domains of the prospects of the method in question have many important applications; this makes the complications that arise in the case of the transition to the arc length, reasonable.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, grant nos. 11-01-00102 and 14-01-00161.

REFERENCES

1. E. Hairer and G. Vanner, *Solving Ordinary Differential Equations: 2. Stiff and Differential-Algebraic Problems*, 2nd ed. (Springer, Berlin, 1996; Mir, Moscow, 1999).
2. E. Riks, "The application of Newton's method to the problem of elastic stability," Journal of Applied Mechanics **39**, 1060–1065 (1972).
3. V. I. Shalashilin and E. B. Kuznetsov, *Method of the Continuation of Solving by a Parameter and the Best Parametrization* (Editorial URSS, Moscow, 1999) [in Russian].
4. J. K. Wu, W. H. Hui, and H. L. Ding, "Arc-length method for differential equations," Applied Mathematics and Mechanics **20** (8), 936–942 (1999).
5. N. N. Kalitkin, A. B. Al'shin, E. A. Al'shina, and B. V. Rogov, *Calculations on Quasi-Uniform Nets* (FIZMATLIT, Moscow, 2005) [in Russian].
6. G. I. Marchuk and V. V. Shaidurov, *Improving the Accuracy of Solutions of Difference Schemes* (Nauka, Moscow, 1979) [in Russian].
7. H. H. Rosenbrock, "Some general implicit processes for the numerical solution of differential equations," Comput. J., **5** (4), 329–330 (1964).
8. P. D. Shirkov, "Optimally damped schemes with complex coefficients for stiff systems of ODEs," Mat. Model. **4** (8), 47–57 (1992).
9. N. N. Kalitkin and I. P. Poshivailo, "Reverse Ls-stable Runge–Kutta schemes," Dokl. Akad. Nauk **442** (2), 1–5 (2012).
10. N. N. Kalitkin and L. V. Kuz'mina, Preprint No. 80, IPM RAN (Keldysh Inst. of Applied Mathematics, Russian Academy of Sciences, Moscow, 1981).
11. F. Mazzia and C. Magherini, *Test Set for Initial Value Problem Solvers: Release 2.4* (Univ. of Bari and INdAM, Research Unit of Bari, Italy, 2008).
12. J. G. Verwer, "Gauss–Seidel iteration for sti ODEs from chemical kinetics," SIAM J. Sci. Comput. **15** (5), 1243–1259 (1994).

Translated by L. Kartvelishvili