U.S.S.R. Comput.Maths.Math.Phys.,Vol.27,No.3,pp.30-38,1987
Printed in Great Britain
0041-5553/87 $10.00+0.00

THE METHOD OF LOCAL LINEARIZATION IN THE NUMERICAL SOLUTION OF STIFF SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS*

B.V. PAVLOV and O.E. RODIONOVA

The method of local linearization in its simplest form has first order of accuracy and requires a considerable amount of computer time. A scheme of second order of accuracy is constructed which leads to a much higher algorithm efficiency.

Difference methods of numerical integration are not very suitable for solving stiff local-unstable systems of ordinary differential equations and sometimes lead to qualitatively incorrect results. The method of local linearization is free from this drawback.

When solving the Cauchy problem for the autonomous system

$$\dot{x}=f(x), \quad x(t_0)=x_0, \quad x=(x_1,\ldots,x_n), \quad f=(f_1,\ldots,f_n), \tag{1}$$

or its equivalent system in increments

$$\dot{u}(\tau)=f(x_0+u)=f(x_0)+Au+\varphi(u),\ u(0)=0, \tag{2}$$

$$\tau=t-t_0, \quad u(\tau)=x(t-t_0), \quad A=\frac{Df(x)}{Dx}\Big|_{x=x_0},$$

$$\varphi(u)=f(x_0+u)-f(x_0)-Au,$$

the method of local linearization reduces essentially to solving the following integral equation:

$$u(\tau)=C(\tau)f(x_0)+\int_0^\tau \exp[A(\tau-s)]\varphi(u(s))\,ds, \tag{3}$$

$$C(\tau)=[\exp(A\tau)-1]A^{-1},$$

in the interval $\tau\in[0, h]$, where $u(\tau)$ differs only slightly from $C(\tau)f(x_0)$, i.e. from the solution of the linearized system

$$u_l(\tau)=f(x_0)+Au_l, \quad u_l(0)=0.$$

The right-hand side of (3) contains no linear element Au, and hence the schemes obtained using (3) are only slightly sensitive to the stiffness of the initial system (1) or (2), connected with the ill-posed nature of the variational matrix $A=Df/Dx$. The main difficulties in realizing this approach are in evaluating the matrix $C(\tau)$ and the integral of the non-linear part $\varphi(u)$ in (3).

1. Evaluation of the matrix $C(\tau)$ and the implicit scheme.

The general method of calculating $C(\tau)$ is based on the use of the recurrence relation

$$C(2\tau)=C(\tau)+[E+C(\tau)A]C(\tau),$$

which follows from the functinal equation for $C(\tau)$

$$C(t_1+t_2)=C(t_1)+[E+C(t_1)A]C(t_2).$$

The initial value $C(\tau_0)$ when $\tau_0\|A\|<\varepsilon$ can be calculated using an expansion in series

$$C(\tau_0)=\tau_0\left[E+\frac{1}{2}A\tau_0+\ldots+\frac{1}{n!}(A\tau_0)^{n-1}+\ldots\right].$$

The formula has asymptotic stability, and experience has shown that this method of evaluating $C(\tau)$ has high accuracy and reliability (see /1/).

Henceforth, we will consider the case when the eigenvalues λ_i of the matrix A have small imaginary parts. Then, we can use for the inegral in (3) the quadrature formula

$$\int_0^\tau \exp[A(\tau-s)]\varphi(z(s))\,ds\approx\int_0^\tau \exp[A(\tau-s)]\varphi(z(\tau))\,ds=C(\tau)\varphi(z(\tau)).$$

Hence, we arrive at the implicit scheme of the method of local linearization /2/

*Zh.vychisl.Mat.mat.Fiz.,27,5,688-699,1987

$$\tilde{u}(\tau)=C(\tau)[f(x_0)+\varphi(\tilde{u}(\tau))]. \tag{1.1}$$

We note an important asymptotic property of the implicit scheme. For the slightly nonlinear system (2) with a stable matrix A a stable stationary point $u_c=u$ $(\tau\to\infty)$ exists which is given by the equation $f(x_0)+Au_c+\varphi(u_c)=0$. Then, if the solution of the implicit scheme $\tilde{u}(\tau)$ is bounded in $\tau\in[0,\infty)$, then $\tilde{u}(\tau\to\infty)=u_c$, since $\tilde{u}(\infty)=-A^{-1}[f(x_0)+\varphi(\tilde{u}(\infty))]$ as $\tau\to\infty$, and, consequently, $\tilde{u}(\infty)$ satisfies the equation

$$f(x_0)+A\tilde{u}(\infty)+\varphi(\tilde{u}(\infty))=0.$$

The solution $\tilde{u}(\tau)$ as a function of τ, satisfies the differential equation

$$\dot{\tilde{u}}=[E-C(\tau)D\varphi(\tilde{u})/D\tilde{u}]^{-1}f(x+\tilde{u}),\qquad \tilde{u}(0)=0,$$

which, when

$$\|C(\tau)D\varphi(\tilde{u})/D\tilde{u}\|\ll 1,\qquad \tau\in[0,h], \tag{1.2}$$

differs only slightly from the initial Eq.(2); condition (1.2) thereby limits the value of the interval $[0,h]$ in which $\tilde{u}\approx u$. A practical check of (1.2) can easily be made when solving (1.1) by direct iterations

$$\tilde{u}^{k+1}(\tau)=C(\tau)[f(x_0)+\varphi(u^k(\tau))].$$

The quantity

$$M=\max_k\frac{\|\tilde{u}^{k+1}-\tilde{u}^k\|}{|\tilde{u}^k-\tilde{u}^{k-1}|}=\max_k\frac{\|C[\varphi(\tilde{u}^k)-\varphi(\tilde{u}^{k-1})]\|}{|\tilde{u}^k-\tilde{u}^{k-1}|}$$

estimates $\|CD\varphi(\tilde{u})/D\tilde{u}\|$ in the sequence of vectors $(\tilde{u}^k-\tilde{u}^{k-1})$. Hence, direct iterations in this case are not only a method of solving (1.1) but also a method of choosing the integration step h: if h is such that $M\leqslant\varepsilon$, then $\|\tilde{u}(\tau)-C(\tau)f(x_0)\|\leqslant[\varepsilon/(1-\varepsilon)]\|C(\tau)f(x_0)\|$ and we can take as the error of the solution the quantity $C\varphi(\tilde{u})$. In the initial version of the method /1/, after calculating the solution using (1.1) with the necessary step h, linearization is carried out and the matrix A is calculated at the new point $x=x_1$. This organization required an unjustifiably large number of expensive evaluations of the matrix $C(\tau)$, which was the reason for the low efficiency of the algorithm. A modification of the algorithm, proposed in /3/, enabled the calculation of the matrix $C(\tau)$ to be used many times; the formula for the algorithm in this case has the form

$$z_i=C(h_i)[f(x_i)+\Delta_i z_i+\varphi(z_i)], \tag{1.3}$$

where $z_i=x_{i+1}-x_i$, $x_i=x(t_i)$, $h_i=t_{i+1}-t_i$, and $(\Delta_i)=A(x_i)-A(x_0)$ is the increment of the variational matrix to the i-th integration step. However, scheme (1.3), unlike scheme (1.1), has first order of accuracy and hence does not lead to any appreciable increase in the efficiency of the algorithm, since in view of the increase in the matrix (Δ_i), the integration step is regularly reduced. To increase the efficiency of the algorithm it is necessary to develop a scheme of higher order of accuracy. It is desirable in this case that it should have the asymptotic properties of schemes (1.1) and (1.3).

2. A method of increasing the order of accuracy.

Consider the system of differential equations

$$\dot{z}(\tau)=f+Az+\mu(z),\qquad z(0)=0 \tag{2.1}$$

and its equivalent integral equation

$$z(\tau)=C(\tau)f+\int_0^\tau\exp[A(\tau-s)]\mu(z(s))\,ds, \tag{2.2}$$

where $z=(z_1,\ldots,z_n)$, $f=(f_1,\ldots,f_n)$ and $A=(A_{ik})$ is an n-th order matrix, whose spectrum lies in the region of the real axis, $\Delta=(\Delta_{ik})$ is a certain n-th order matrix, and $\mu(z)=\Delta z+\varphi(z)$, $\varphi(z)$ is a polynomial function of z which does not contain constant and linear terms.

Our aim will be to construct a scheme of the second order of accuracy on the basis of the implicit scheme. Suppose

$$z_0(\tau)=C(\tau)[f+\mu(z_0)],\qquad \tau\in[0,T]. \tag{2.3}$$

Solution (2.2) will be sought on the basis of the iterational process

$$z_{n+1}(\tau)=z_0(\tau)+\int_0^\tau\exp[A(\tau-s)]\mu(z_n(s))\,ds-C(\tau)\mu(z_0(\tau)).$$

The interval $\tau\in[0,T]$ in which $z_0(\tau)$ will be taken as the initial approximation, and we will determine the condition

$$M=\max_k\frac{\|z_0^{k+1}(\tau)-z_0^k(\tau)\|}{\|z_0^k(\tau)-z_0^{k-1}(\tau)\|}<0.5,\qquad z_0^{k+1}=C(\tau)[f+\mu(z_0^k)],$$

for which $\|z_0(\tau)-C(\tau)f\|\leqslant\|C(\tau)f\|$. We will take as the approximate solution of (2.1) the quantity

$$z_1(\tau)=z_0(\tau)+y_1(\tau), \tag{2.4}$$

where

$$y_1(\tau)=\int_0^\tau \exp[A(\tau-s)]\mu(z_0(s))ds-C(\tau)\mu(z_0(\tau)).$$

The error of the approximate solution will be represented by the quantity

$$y_2(\tau)=\int_0^\tau \exp[A(\tau-s)][\mu(z_0(s)+y_1(s))-\mu(z_0(s))]ds.$$

The requirement that $\|y_2(\tau)\|$ should be small will also determine the value of the integration step $h=\tau_{max}\in[0, T]$.

3. Evaluation of the integrals.

We will evaluate the integrals using specially chosen approximating functions which can be integrated analytically. Consider integrals of the form

$$\Phi(\tau)=\int_0^\tau \exp[A(\tau-s)]\mu(z(s))ds,$$

where $z(s)=(z_1(s),\ldots,z_n(s))$. Henceforth, we will approximate $C(\tau-s)$ and not $\exp[A(\tau-s)]$ and hence we will rewrite the integral in the form

$$\Phi(\tau)=\frac{d}{d\tau}\int_0^\tau C(\tau-s)\mu(z(s))ds. \tag{3.1}$$

This form of writing the integral also determines the order of the operations after approximating the integrand, and it is also necessary to satify the following: if we approximate $C(\tau-s)$ by a certain continuous function $Q(\tau-s)$, then, according to (3.1),

$$\Phi(\tau)=Q(0)\mu(z(\tau))+\int_0^\tau Q'_{\tau-s}(\tau-s)\mu(z)ds,$$

which, when $Q(0)=0$, is equivalent to the singular approximation of the series

$$\exp[A(\tau-s)]=Q(0)\delta(\tau-s)+Q'_{\tau-s}(\tau-s).$$

Suppose we are given two matrices $Q_1(s, h)$ and $Q_2(s, h)$ which approximate $C(s)$, and two vectors $q_1(s, h)$ and $q_2(s, h)$ which approximates $z(s)$ in the section $[0, h]$, and Q_2 and q_2 approximate C and z better than Q_1 and q_1, i.e.

$$\|C(s)-Q_2(s)\|<\|C(s)-Q_1(s)\|,\quad \|z(s)-q_2(s)\|<\|z(s)-q_1(s)\|.$$

Then, assuming that $\delta C(\tau-s)\approx Q_2(\tau-s)-Q_1(\tau-s)$, $\delta z(s)\approx q_2(s)-q_1(s)$ we obtain expressions for the approximate quantity $\Phi(h)$ and its error

$$\Phi(h)=\left[\frac{d}{d\tau}\int_0^\tau Q_1(\tau-s)\mu(q_1(s))ds\right]_{\tau=h}, \tag{3.2a}$$

$$\delta\Phi(h)=\left\{\frac{d}{d\tau}\left[\int_0^\tau \delta C(\tau-s)\mu(q_1(s))ds+\int_0^\tau Q_1(\tau-s)\frac{D\mu(q_1)}{Dq_1}\delta z(s)ds\right]\right\}_{\tau=h}. \tag{3.2b}$$

4. Choice of the approximating functions.

Eq. (2.3) defines $z_0(\tau)$ as a certain function of the eigenvalues of the matrix $C(s)$, equal to $c_i(s)=\lambda_i^{-1}[\exp(\lambda_i s)-1]$, and hence when choosing the approximating functions we will be concerned with the properties of the function $c(\lambda s)=\lambda^{-1}[\exp(\lambda s)-1]$. We will consider two types of polynomials: with free terms which are equal to and differ from zero. For convenience we will denote the matrix polynomials differently

$$P^{(k)}(s)=\sum_{i=0}^{k} P_i^{(k)}s^i \quad \text{and} \quad R^{(k)}(s)=\sum_{i=1}^{k} R_i^{(k)}s^i$$

and the corresponding vector polynomials

$$p^{(k)}(s)=\sum_{i=0}^{k} p_i^{(k)}s^i \quad \text{and} \quad r^{(k)}(s)=\sum_{i=1}^{k} r_i^{(k)}s^i$$

of the first (P and p) and the second (R and r) types. The choice and verification of the suitability of the polynomials was made using the example of the evaluation of the matrix integral

$$\Phi^0(\tau)=\int_0^\tau \exp[A(\tau-s)]\Delta C(s)\,ds=\frac{d}{d\tau}\int_0^\tau C(\tau-s)\Delta C(s)\,ds,$$

which is the first iteration of the matrix integral equation

$$B(\tau)=C(\tau)+\int_0^\tau \exp[A(\tau-s)]\Delta B(s)\,ds$$

and is obtained by substituting into (3.1) the quantities $\mu(z(s))=\Delta z(s)$, $z(s)=C(s)$, where Δ is an arbitrary matrix. In the natural basis of the matrix A the matrix elements are as follows:

$$\hat{\Phi}_{ik}{}^0(\tau)=\int_0^\tau \exp[\lambda_i(\tau-s)][\exp(\lambda_k s)-1]\lambda_k{}^{-1}\hat{\Delta}_{ik}\,ds,$$

where $\hat{\Delta}_{ik}$ are the elements of the matrix Δ in the same basis. After integration we obtain the formula

$$\hat{\Phi}_{ik}{}^0(\tau)=(\lambda_i-\lambda_k)^{-1}[c_i(\tau)-c_k(\tau)]\hat{\Delta}_{ik}, \qquad (4.1)$$
$$c_i(\tau)=\lambda_i{}^{-1}[\exp(\lambda_i\tau)-1],$$

which is used for a direct check of the approximate values of $\Phi^0(h)$ obtained.

The case of a stable spectrum. If $\lambda<0$, then $c(\lambda s)$ is a monotonic function of s, bounded in $[0,\infty)$. Suppose $s\in[0,h]$ and $\lambda h\ll -1$, then, in almost the whole interval $[0,h]$, we have $c(\lambda s)\sim-\lambda^{-1}$. This important asymptotic property can be preserved if we take as the approximating function a polynomial of the first type: $c(\lambda s)\approx p^{(h)}(s)$. We will confine ourselves to considering polynomials of the first and second order. We will determine the coefficients of the polynomial $p^{(1)}(s,h)=p_0{}^{(1)}+p_1{}^{(1)}s$ from the condition that it should be identical with the approximated function $c(\lambda s)$ in the middle and at the end of the interval $[0,h]$

$$p^{(1)}(h,\lambda h)=c(\lambda h),\qquad p^{(1)}(h/2,\lambda h)=c(\lambda h/2).$$

Then

$$p_0{}^{(1)}=2c(\lambda h/2)-c(\lambda h),\qquad hp_1{}^{(1)}=2[c(\lambda h)-c(\lambda h/2)]. \qquad (4.2)$$

When determining the coefficients of the second-order polynomial $p^{(2)}(s,\lambda h)=p_0{}^{(2)}+p_1{}^{(2)}s+p_2{}^{(2)}s^2$ we will add the condition for the derivatives to be equal when $s=h$

$$\frac{d}{ds}p^{(2)}(s,h)=\frac{d}{ds}c(\lambda s);$$

then

$$p^{(2)}(h,\lambda h)=c(\lambda h),\qquad p^{(2)}(h/2,\lambda h)=c(\lambda h/2),\qquad [p^{(2)}(s,\lambda h)]'_{s=h}=\exp(\lambda h)$$

and for the coefficients $p_0{}^{(2)}, p_1{}^{(2)}, p_2{}^{(2)}$ we obtain

$$p_0{}^{(2)}=p_0{}^{(1)}+{}^1/_2 p_2{}^{(2)}h^2,\qquad hp_1{}^{(2)}=hp_1{}^{(1)}-{}^3/_2 p_2{}^{(2)}h^2,$$
$$2p_2{}^{(2)}h^2=[\exp(\lambda h)-p_1{}^{(1)}]h.$$

When $\lambda h\ll -1$ we have $c(\lambda h)\sim-\lambda^{-1}$, and then $p_1{}^{(1)}, p_1{}^{(2)}, p_2{}^{(2)}\sim 0$, while $p_0{}^{(1)}$ and $p_0{}^{(2)}$ approach $-\lambda^{-1}$. Hence, the approximating polynomials of the first type preserve the asymptotic properties of the function $c(\lambda s)$. Using the approximation considered and employing formulas (3.1) and (3.2), we can calculate the approximate values. Assuming $c(\tau-s)=c_i(\tau-s)$, $\mu(z)=\Delta_{ik}c_k(s)$ in (3.1) and $Q_i(\tau-s)=p^{(i)}(\tau-s,\lambda_i h)$, $q_i(s)=p^{(i)}(s,\lambda_k h)$ in (3.2) and using condition (4.2) to determine the coefficients, after appropriate calculations we obtain

$$\hat{\Phi}_{ik}{}^0(h)=\left\{c_i(h)c_k(h)-2\left[c_i(h)-c_i\left(\frac{h}{2}\right)\right]\times\left[c_k(h)-c_k\left(\frac{h}{2}\right)\right]\right\}\hat{\Delta}_{ik}. \qquad (4.3)$$

We now can make a direct comparison between the accurate value (4.1) and the approximate value (4.3) for $\Phi_{ik}{}^0(\lambda)$. A numerical check showed that in the quadrant $\lambda_i h,\ \lambda_k h<0.3$ expression (4.3) has an error in the second place. In the initial basis the formula for $\Phi^0(h)$ has the form

$$\Phi^0(h)=C(h)\Delta C(h)-2\left[C(h)-C\left(\frac{h}{2}\right)\right]\Delta\left[C(h)-C\left(\frac{h}{2}\right)\right].$$

The case of an unstable spectrum. When $\lambda>0$ the approximating function of the first type $p(s,h)$ satisfactorily approximates the function $\lambda^{-1}[\exp(As)-1]$ only in a small interval $h:\lambda h<0.3$ which makes it extremely difficult to monitor that this condition is satisfied.

Polynomials of the second type would be considerably better but they are not quite suitable for approximating $c(s\lambda)$ when $\lambda h \ll -1$. The complexity of the problem of approximating $C(As)$ for a stiff instability of the matrix A is exactly connected with the fact that the negative and positive eigenvalues require their own type of approximation. This can easily be done in the case of a diagonal (or triangular) form of the matrix A, but in general, we must use special methods. Experience has shown that quite satisfactory results can be obtained by using the combination of approximations of the first and second type given below.

We will divide the interval $[0, h]$ into two equal parts. Then, for $\Phi^0(h)$ we obtain

$$\hat{\Phi}_{ik}^{0}(h)=\left\{\exp\left(\lambda_i\frac{h}{2}\right)\left[\frac{d}{d\tau}\int_0^{\tau}c_i(\tau-s)c_k(s)\,ds\right]_{\tau=h/2}+\right.$$
$$\left.c_i\left(\frac{h}{2}\right)c_k\left(\frac{h}{2}\right)+\left[\frac{d}{d\tau}\int_0^{\tau}c_i(s)c_k(\tau-s)\,ds\right]_{\tau=h/2}\exp\left(\lambda_k\frac{h}{2}\right)\right\}\hat{\Delta}_{ik}.$$

This formula is constructed in such a way that $c_i(s)$ occurs with a factor $\exp(\lambda_i h/2)$ in the first interval, and occurs in the same way in the second interval for $c_k(s)$ so that when $\lambda_i h,\ \lambda_k h \ll -1$ these integrals are exponentially small irrespective of the type of approximation chosen for $c_i(s)$, $c_k(s)$. This fact enables us to approximate them by polynomials of the second type (without a free term), which can be easily done when $0<\lambda h<1$

$$c_i(s)\approx r_i^{(h)}(s)=\sum_{\nu=1}^{h}r_{i\nu}^{(h)}s^{\nu}.$$

Hence, as before by approximating $c_k(s)$ in the first integral and $c_i(s)$ in the second integral by polynomials of the first type, we can calculate $\Phi^0(h)$, confining ourselves to polynomials of the first order ($k=1$). Then, in the first integral $c_i(\tau-s)=r_1^{(1)}(\tau-s)$, $c_k(s)=p_0^{(1)}+p_1^{(1)}s$, and the coefficients $r_1^{(1)}$, $p_0^{(1)}$, $p_1^{(1)}$ can be found from the equations

$$r_1^{(1)}\frac{h}{2}=c_i\left(\frac{h}{2}\right),\qquad p_0^{(1)}+p_1^{(1)}\frac{h}{2}=c_k\left(\frac{h}{2}\right),$$
$$p_0^{(1)}+p_1^{(1)}\frac{h}{4}=c_k\left(\frac{h}{4}\right).$$

We have for the second integral $c_i(s)=p_0^{(1)}+p_1^{(1)}s$, $c_k(\tau-s)=r_1^{(1)}(\tau-s)$, where the coefficients can be found from the conditions

$$p_0^{(1)}+p_1^{(1)}\frac{h}{2}=c_i\left(\frac{h}{2}\right),\qquad p_0^{(1)}+p_1^{(1)}\frac{h}{4}=c_i\left(\frac{h}{4}\right),$$
$$r_1^{(1)}\frac{h}{2}=c_k\left(\frac{h}{2}\right).$$

Carrying out the necessary calculations we obtain

$$\hat{\Phi}_{ik}^{0}(h)=\left\{\left[c_i(h)-c_i\left(\frac{h}{2}\right)\right]c_k\left(\frac{h}{4}\right)+c_i\left(\frac{h}{2}\right)c_k\left(\frac{h}{2}\right)+\right.$$
$$\left.c_i\left(\frac{h}{4}\right)\left[c_k(h)-c_k\left(\frac{h}{2}\right)\right]\right\}\hat{\Delta}_{ik}. \tag{4.4}$$

A check showed that this "piecewise" method of approximation can be regarded as completely satisfactory: the approximate values for $\hat{\Phi}_{ik}^{0}(h)$ calculated from (4.4) with an error of less than 3%, are identical with the accurate values calculated from (4.1) for $\lambda_i h,\ \lambda_k h \leqslant 1$, and it is quite easy to monitor that this condition is satisfied (see Sect.5). In the initial basis the formula for $\Phi^0(h)$ will obviously have the form

$$\Phi^0(h)=\exp\left(A\frac{h}{2}\right)C\left(\frac{h}{2}\right)\Delta C\left(\frac{h}{4}\right)+C\left(\frac{h}{2}\right)\Delta C\left(\frac{h}{2}\right)+$$
$$C\left(\frac{h}{4}\right)\Delta C\left(\frac{h}{2}\right)\exp\left(A\frac{h}{2}\right)=\left[C(h)-C\left(\frac{h}{2}\right)\right]\Delta C\left(\frac{h}{4}\right)+$$
$$C\left(\frac{h}{2}\right)\Delta C\left(\frac{h}{2}\right)+C\left(\frac{h}{4}\right)\Delta\left[C(h)-C\left(\frac{h}{2}\right)\right].$$

By approximating $C(s)$ by second-order polynomials we can also evaluate $\delta\Phi_{ik}^{0}(h)$, but the formulas obtained are extremely complex. Without deriving them, we will merely note that they give the correct order of the error.

5. Solution of the integral equation.

Using the results obtained we will consider a procedure for solving an integral equation by the method of local linearization. According to Sect.2, we take as the solution of Eq. (2.2) the sum of the first two terms of the series, which is obtained as a result of the

iteration (2.2)

$$z(\tau)=z_0(\tau)+y_1(\tau).$$

The procedure consists of the following stages:

1) calculation of the initial approximation $z_0(\tau)$ by solving the equation

$$z_0(\tau)=C(\tau)[f+\mu(z_0(\tau))] \tag{5.1}$$

by the method of direct iteration; fairly rapid convergence is required which limits the value of the interval h in which the initial approximation $z_0(\tau)$ is acceptable;

2) calculation of the following term of the series:

$$y_1(h)=\int_0^h \exp[A(h-s)]\mu(z_0(s))ds-C(h)\mu(z_0(h)), \tag{5.2}$$

which reduces to evaluating the integral using some method of approximating $z_0(s)$ in $s\in[0, h]$;

3) calculation, using Eqs.(3.1), of the error δz_{ap} connected with the approximation;

4) evaluation of the integral

$$y_2(h)=\int_0^h \exp[A(h-s)][\mu(z_0(s)+y_1(s))-\mu(z_0(s))]ds,$$

which characterizes the value of the error connected with the "termination" of the iterational process of solving (2.2).

Note that calculation of the error δz_{ap} is only necessary when there are eigenvalues with large imaginary part in the spectrum of A. The approximations considered in Sect.4, in the case of a real spectrum, were checked directly and gave quite satisfactory results (for practical purposes). As regards $y_2(h)$ its role reduces to "refining" the value of the integration interval h from the smallness condition $\|y_2(h)\|$, which can be replaced with the same success by $\|y_1(h)\|$.

Hence, in the case of a real spectrum A the procedure for solving (2.2) essentially reduces to paragraphs 1) and 2) above. This fact simplifies the algorithm and increases its efficiency, since although the calculation of δz_{ap} and $y_2(h)$ does not present any difficulties in principle, it involves certain computer costs in view of the extremely complicated formulas.

We will now consider in more detail the calculation of $y_1(h)$ using the piecewise method of approximation, calculated assuming that positive eigenvalues are present in the spectrum A. According to (2.4)

$$y_1(h)=\Phi(h)-C(h)\mu(z_0(h)),$$

where

$$\Phi(h)=\left[\frac{d}{d\tau}\int_0^\tau C(\tau-s)\mu(z_0(s))ds\right]_{\tau=h}=$$
$$\exp\left(A\frac{h}{2}\right)\left[\frac{d}{d\tau}\int_0^\tau C(\tau-s)\mu(z_0(s))ds\right]_{\tau=h/2}+$$
$$C\left(\frac{h}{2}\right)\mu\left(z_0\left(\frac{h}{2}\right)\right)+$$
$$\left\{\frac{d}{d\tau}\int_0^\tau C(s)\left[\mu\left(z_0\left(\frac{h}{2}+\tau-s\right)\right)-\mu\left(z_0\left(\frac{h}{2}\right)\right)\right]ds\right\}_{\tau=h/2}.$$

Following the method of piecewise approximation, in the first integral we assume that $C(\tau-s)\approx R_1^{(1)}(\tau-s)$, where the coefficient $R_1^{(1)}$ is found from the condition $C(h/2)=R_1^{(1)}h/2$ and $z_0(s)\approx p_0^{(1)}+p_1^{(1)}s$, where the coefficients $p_0^{(1)}$ and $p_1^{(1)}$ are found from the conditions $z_0(h/2)=p_0^{(1)}+p_1^{(1)}h/2$, $z_0(h/4)=p_0^{(1)}+p_1^{(1)}h/4$, whence

$$p_0^{(1)}=2z_0\left(\frac{h}{4}\right)-z_0\left(\frac{h}{2}\right),\qquad p_1^{(1)}=\left(\frac{h}{4}\right)^{-1}\left[z_0\left(\frac{h}{2}\right)-z_0\left(\frac{h}{4}\right)\right].$$

In the second integral $C(s)=P_0^{(1)}+P_1^{(1)}s$, and the coefficients $P_0^{(1)}$, $P_1^{(1)}$ are found from the conditions $C(h/2)=P_0^{(1)}+P_1^{(1)}h/2$, $C(h/4)=P_0^{(1)}+P_1^{(1)}h/4$, whence

$$P_0^{(1)}=2C\left(\frac{h}{4}\right)-C\left(\frac{h}{2}\right),\qquad P_1^{(1)}=\left(\frac{h}{4}\right)^{-1}\left[C\left(\frac{h}{2}\right)-C\left(\frac{h}{4}\right)\right];$$

$z_0(h/2+\tau-s)-z_0(h/2)=r_1^{(1)}(\tau-s)$ and the coefficient $r_1^{(1)}$ is found from the condition $z_0(h)-z_0(h/2)=r_1^{(1)}h/2$.

Specific calculations can only be carried out when the structure of the function $\mu(z)$ is given. Suppose

$$\mu(z)=[\Delta+K(z)]z,$$

where Δ is a matrix with constant coefficients, and $K(z)$ is a matrix whose coefficients are linear uniform functions of z

$$K_{ik}=\Sigma l_{ik}^s z_s.$$

Then, after the necessary calculations we obtain

$$\Phi(h)=\exp\left(A\frac{h}{2}\right)C\left(\frac{h}{2}\right)\mu\left(z_0\left(\frac{h}{4}\right)\right)+C\left(\frac{h}{2}\right)\mu\left(z_0\left(\frac{h}{2}\right)\right)+$$
$$C\left(\frac{h}{4}\right)\left[\mu(z_0(h))-\mu\left(z_0\left(\frac{h}{2}\right)\right)\right]+$$
$$\frac{1}{96}h^3\left[\exp\left(A\frac{h}{2}\right)R_1^{(1)}\varphi(p_1^{(1)})-2P_1^{(1)}\varphi(r_1^{(1)})\right],$$

where $\varphi(z)=K(z)z$.

In this formula we can drop the term with the coefficient $h^3/96$, since the same term with a coefficient $\sim h^3/3$ occurs in the first part of the formula. As a result, we obtain for $\Phi(h)$ a convenient quadrature formula of the second order of accuracy in h, which holds for any structure of $\mu(z)$

$$\Phi(h)=\left[C(h)-C\left(\frac{h}{2}\right)\right]\mu\left(z_0\left(\frac{h}{4}\right)\right)+C\left(\frac{h}{2}\right)\mu\left(z_0\left(\frac{h}{2}\right)\right)+$$
$$C\left(\frac{h}{4}\right)\left[\mu(z_0(h))-\mu\left(z_0\left(\frac{h}{2}\right)\right)\right].$$

then, for $y_1(h)$ after identical transformations, we obtain

$$y_1(h)=-\left\{\left[C(h)-C\left(\frac{h}{2}\right)\right]\left[\mu\left(z_0\left(\frac{h}{2}\right)\right)-\mu\left(z_0\left(\frac{h}{4}\right)\right)\right]+\right.$$
$$\left.\left[C(h)-C\left(\frac{h}{4}\right)\right]\left[\mu(z_0(h))-\mu\left(z_0\left(\frac{h}{2}\right)\right)\right]\right\}. \tag{5.3}$$

It follows from (5.3) that if $z_0(h)$ is bounded in $[0, \infty)$ and the matrix A is stable, then $y_1(h\to\infty)\to 0$ since $C(h\to\infty)\to -A^{-1}$. Hence, the function $z_1(h)=z_0(h)+y_1(h)$ has the asymptotic property of $z_0(h)$.

Estimate of the right limit of the spectrum. Eqs. (5.2) and (5.3) give good results provided $\lambda h\leqslant 1$, and hence it is necessary to monitor that this is satisfied. When evaluating $\Phi(h)$, $h=2^m h_0$, when there is a set of matrices $\exp(Ah_k)$, $h_k=2^k h_0$, and, consequently, the values of the trace

$$\operatorname{Tr}\exp(Ah_k)=\sum_{i=1}^{n}\exp(\lambda_i h_k)$$

for all $k\leqslant m$; using the latter we can estimate extremely accurately the right limit of the spectrum of the matrix $\exp(Ah)$. However, for practical purposes when solving differential equations of "medium" dimensions of not more than 50, an estimate based on the use of a trinomial of the fourth order $p(x)=x^4-2x^2+x$, which is fairly small when $x\in[0, 1]$, is quite satisfactory:

$$\sup_{x\in[0,1]}|p(x)|\approx 1/8. \tag{5.4}$$

We will put

$$x_i=\exp(\lambda_i h),\qquad M_0=\operatorname{Tr}\exp(Ah)=\sum_{i=1}^{n}x_i,$$
$$M_1=\operatorname{Tr}\exp(2Ah)=\sum_{i=1}^{n}x_i^2,\qquad M_2=\operatorname{Tr}\exp(4Ah)=\sum_{i=1}^{n}x_i^4,$$
$$\sum_{i=1}^{n}p(x_i)=M_2-2M_1+M_0.$$

It is obvious that for any x_k the following holds:

$$p(x_k)=M_2-2M_1+M_0-\sum_{i\neq k}^{n}p(x_i),$$
$$p(x_k)\leqslant M_2-2M_1+M_0-\inf\left[\sum_{i\neq k}^{n}p(x_i)\right],\quad \inf p(x)\approx 0.075.$$

Hence it follows that $p(x_{\max})\leqslant M_2-2M_1+M_0+0.075(n-1)$, $x_{\max}=\exp(\lambda_{\max}h)$.

If

$$M_2-2M_1+M_0+0.075(n-1)\leqslant p(\exp(1))\approx 40, \tag{5.5}$$

then $\lambda_{\max}h<1$, so that (5.5) is the practical criterion for estimating $\lambda_{\max}h$. According to (5.4), for a stable matrix A we have

$$\sup(M_2-2M_1+M_0)\approx 0.125n,$$

so that (5.5) is satisfied for all $n<200$, and, consequently, (5.5) does not limit the step if A is stable.

6. Integration of the system of ordinary differential equations.

Assuming that

$$z=z^{n+1}=x^{n+1}-x^n, \qquad x^0=x(t_0), \qquad f=f(x^n),$$

$$A=\frac{Df(x)}{Dx}\Big|_{x=x_0}, \qquad \Delta=\frac{Df(x)}{Dx}\Big|_{x=x^n}-A,$$

$$\varphi(z)=\varphi(z^n)=f(x^n+z^n)-f(x^n)-Az^n, \qquad \mu(z)=\Delta z+\varphi(z)$$

and using formula (5.1) for $z_0(h)$ and formula (5.2) for $y_1(h)$, we obtain the scheme of the n-th step of the integration with fixed matrix $A=Df(x^0)/Dx$ for the system

$$\dot{x}=f(x), \quad x(t_0)=x^0, \qquad x^{n+1}=x^n+z_0^{n+1}(h)+y_1^{(n+1)}(h),$$

where $x^n=x(t_n)$.

In this scheme the step is monitored using two criteria:

1) by the limitation of the rate of convergence of the iterations of Eq.(5.1), i.e., by the condition $M\leqslant 0.5$ (this criterion determines the interval of integration h in which the initial approximation $z_0(h)$) is acceptable);

2) by the relative error $K=\|y_1(h)\|/\|x^{n+1}-x^0\|<\varepsilon$.

Condition 2) traces the local accuracy of the integration. In the limits defined by the criterion M, the step is chosen to correspond to the criterion K. If h is larger with respect to K than with respect to M, then at the point $x_1=x^{n+1}$ a new linearization is carried out and the process is repeated. This organization of the algorithm enables one to reduce the number of calculations of the matrix $C(h)$ considerably, and this is also related to its efficiency.

On the basis of the approach considered we developed an algorithm and compiled an experimental program for integrating a system of differential equations. Experiments on the integration of specific systems showed that there is a considerable gain in the speed of the calculation (by 1-2 orders of magnitude) compared with the method of local linearization of the first order (see /3/) for stiff systems with pronounced local instability /4/, when the trajectory of the solution passes through a region where $A=Df(x)/D(x)$ has large positive eigenvalues. Difference algorithms in this case may give qualitatively incorrect solutions if an insufficiently high local integration accuracy is specified. Tests were made on a specially constructed example of a locally unstable 16-dimensional system of differential equations of isothermal chemical kinetics, which models an explosive-type process with pronounced induction period. Characteristic graphs of the solutions for the initial and one of the intermediate materials are shown in Figs.1 and 2 respectively.

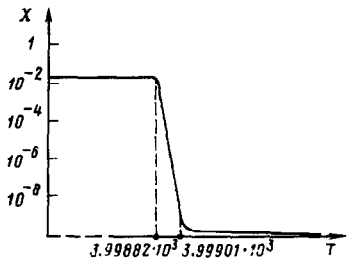

Fig.1

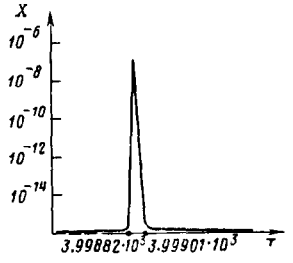

Fig.2

Hence, we can state that the method of integrating ordinary differential equations considered above is preferable to the widely used difference methods. The main interest is the polynomial approximation for $C(s)$ which preserves the asymptotic properties of this function as $h\to\infty$. When obtaining specific schemes, an analogous approximation was also used for the solution itself, but in other cases different methods of approximating the solution $z(s)$ may be preferable.

The authors thank K.I. Babenko and R.P. Fedorenko for discussing this paper.

REFERENCES

1. PAVLOV B.V., Numerical integration of "stiff" systems of ordinary differential equations, Proceedings of the National Conference, Computing Centre of the Academy of Sciences of the USSR, Novosibirsk, 1973.
2. PAVLOV B.V. and POVZNER A.YA., A method for the numerical integration of systems of ordinary differential equations, Zh. vychisl. Mat. mat. Fiz., 13, 4, 1056-1059, 1973.
3. GOL'DENBERG M.YA. and KRESTININ A.V., Numerical Integration of the Differential Equations of Chemical Kinetics, Preprint, Inst. Chem. Phys. Academy of Sciences of the USSR, Moscow, 1974.

4. RODIONOVA O.E. and PAVLOV B.V., Numerical integration of stiff locally unstable systems of ordinary differential equations, Proceedings of the National Conference, Groznyi, 1985.

Translated by R.C.G.

U.S.S.R. Comput.Maths.Math.Phys.,Vol.27,No.3,pp.38-43,1987 0041-5553/87 $10.00+0.00
Printed in Great Britain

A COMPOSITE METHOD OF SOLVING TWO-DIMENSIONAL STATIONARY SELFCONSISTENT PROBLEMS*

G.T. GOLOVIN

A new iterative method is described for solving two-dimensional stationary selfconsistent problems. The method is compared with well-known methods.

Introduction.

The main advantages and disadvantages of the two well-known iterative methods of solving stationary selfconsistent problems were analysed in detail in /1/; in these methods, the emission current density J on the cathode has to be found from the condition for the normal component of electric field-strength to vanish, namely,

$$E_n|_S=0, \tag{1}$$

where S is a given part (called the emission zone) of the cathode surface. In one method (call it method I) the density J on S is found at each iteration from the "$^3/_2$ power law", and in the other (call it method II, it is found from the integral equation of the 1st kind obtained in /2/, equivalent to condition (1). Approximation of this integral equation by means of quadrature formulae gives the algebraic equation

$$AJ+E_{n0}=0, \tag{2}$$

where E_{n0} is the normal component of the electrostatic field strength on the cathode, i.e., the field produced by the applied potential difference when there is no space charge, A is an $N\times N$ matrix, and N is the number of points in the emission zone S at which condition (1) has to be satisfied when it is discretized. The meaning of (2) is that the electric field induced by the space charge (the term AJ) must balance the electrostatic field E_{n0} with the opposite sign, which leads to satisfaction of condition (1) on the cathode.

Recall that the main merit of method I is the small volume of computations at each iteration, while the main drawback is the slow convergence of the iterations and the low accuracy of the numerical solution when the cathode surface has large curvature or the charged particle trajectory bends strongly close to the cathode surface, i.e., in other words, in cases when the "$^3/_2$ power law" used to find J ceases to hold close to the cathode.

The main merit of method II is the high accuracy of the numerical results for any types of cathode surface and trajectories, and also its rapid convergence /3/. The main drawback of the method is the need to solve at each iteration Eq.(2) to find J, which leads to a large volume of computations per iteration, since 80% of the time per iteration is spent on evaluating the elements of the matrix A. Note for comparison that, in method I not more than 10 arithmetic operations are needed to compute the function J at any point of the cathode. The volume of the other computations per iteration is virtually the same in both methods.

There may be a time difference due to the sequence of performing certain computations or due to the different methods used to compute certain functions. For instance, the electromagnetic fields may be found either by difference or by integral methods. There are many familiar methods suitable for numerical integration of the ordinary differential equations from which the charged particle trajectories are found. Thus there can be a great difference between the volumes of computations per iteration and hence a great difference in the times spent in methods I and II, due solely to the method of finding the function J.

1. Algorithm of the composite method.

A method combining the advantages and avoiding the disadvantages of methods I and II must, first, give satisfactory numerical accuracy where method I fails, and second, must have fewer computations per iteration than method II. In this "composite" method the current density J must be found on "poor" parts of the cathode surface (where the surface or the trajectories have large curvature) in the same way as in method II, and on "good" parts, in the same way as method I. In short, if J can be found satisfactorily on a part of the surface by a simple method, there is no need to use on this part a stronger method which requires a greater volume

Zh.vychisl.Mat.mat.Fiz.,27,5,700-710,1987