Pergamon

J. Appl. Maths Mechs, Vol. 58, No. 6, pp. 955–962, 1994
Copyright © 1995 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0021–8928/94 $24.00+0.00

0021–8928(94)00124-3

THE CAUCHY PROBLEM FOR MECHANICAL SYSTEMS WITH A FINITE NUMBER OF DEGREES OF FREEDOM AS A PROBLEM OF CONTINUATION ON THE BEST PARAMETER†

Ye. B. KUZNETSOV and V. I. SHALASHILIN

Moscow

(*Received* 2 *July* 1993)

The Cauchy problem for a system of ordinary differential equations is formulated as a problem of continuation on the best parameter. It is proved that the length of an integral curve of the problem is such a parameter. The merits of the proposed transformation are demonstrated by a test example in which a stiff system of equations describing the perturbed motion of an aircraft is solved numerically.

Many problems in the mechanics of rigid deformable bodies reduce to the solution of a system of non-linear algebraic, transcendental, differential or integral equations explicitly involving a parameter. The most widely used way of analysing the solutions of such systems is to trace how the solutions vary as a function of the parameter. Quite naturally, implementation of this approach involves continuation of solutions of the non-linear equations as functions of the parameter; that this can indeed be done is established by the implicit function theorem and its generalizations.

The effectiveness of any method for continuation of a solution depends on the successful choice of the continuation parameter. It was suggested in [1] that the length of the curve constituting the set of solutions of the system of equations is a satisfactory continuation parameter.

Riks [2] formulated the problem of choosing a continuation direction that will optimize the conditioning of the linearized system of equations. As a measure of conditioning Riks took the determinant of the system divided by the product of the squared norms of its rows. He was able to show that the system is best conditioned when the solutions are continued along a tangent to the curve of the set of solutions, thus supporting the suggestion made in [1].

A special case of this problem has been investigated [3], and the optimal continuation parameter thus obtained has been used to solve several specific problems of non-linear deformation.

In this paper we carry out a more detailed investigation, establishing necessary and sufficient conditions for the choice of the best continuation parameter. The results will be used to formulate a Cauchy problem which has several advantages.

1. Consider the solution of a system of n non-linear equations in n unknowns $x_1, \dots, x_n$ and a parameter p

$$F_j(x_1,\dots,x_n,p) = 0, \quad j = 1,2,\dots,n \tag{1.1}$$

We will write this system in a form which stresses the equivalence of the unknowns x_j and the parameter p, by working in Euclidean $(n + 1)$-space R^{n+1}: $\{x_1, x_2, \dots, x_n, x_{n+1} = p\}$. In this space the system of equations (1.1) may be written in the form

$$F_j(x_1, \dots, x_{n+1}) = 0 \tag{1.2}$$

The set of solutions of this system obtained by allowing the parameter p to take different values is a curve in the space R^{n+1}, which we will henceforth refer to as the solution set curve of system (1.2).

†*Prikl. Mat. Mekh.* Vol. 58, No. 6, pp. 14–21, 1994.

We shall assume that this curve is smooth. A solution of system (1.2) will be sought by the continuation method. It was pointed out [3] that continuation of a solution may be either discrete or continuous. In either case, however, the problem reduces at each stage of the continuation to solving a system of linear algebraic equations whose matrix is the Jacobian of system (1.2). Failure to converge in a numerical method may be related to the vanishing of the Jacobian, i.e. to the conditioning of the system of linear equations.

The problem amounts to choosing a parameter for the continuation of the solution in such a way as to ensure that the system is as well-conditioned as possible. Our candidate for the role of continuation parameter will be a quantity μ whose increment is assumed to be a linear combination such as

$$\Delta\mu = \alpha_i \Delta x_i, \quad i = 1, 2, \ldots, n+1 \tag{1.3}$$

Throughout this paper we shall use the summation convention for repeated indices.

Varying the sequences of numbers α_i, we can consider all the possible continuation parameters. For example, if $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$, $\alpha_{n+1} = 1$, the parameter will be the problem parameter $p = x_{n+1}$. If we define a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n+1})^T$ in R^{n+1}, it follows from formula (1.3) that the increment of μ is the scalar product of the vectors $\boldsymbol{\alpha}$ and $\Delta\mathbf{x} = (\Delta x_1, \ldots, \Delta x_{n+1})^T \in R^{n+1}$: $\Delta\mu = \boldsymbol{\alpha} \cdot \Delta\mathbf{x}$.

According to the representation (1.3), $\boldsymbol{\alpha}$ defines the direction in which the continuation parameter should be chosen. Thus, if α_i is taken to be the Kronecker delta $\alpha_i, = \delta_{ik}$, then $\boldsymbol{\alpha}$ is the unit vector along the x_k axis along which the continuation parameter has been chosen. Henceforth we shall indeed specify this direction by the unit vector $\boldsymbol{\alpha}$.

Equations continuing the solution as functions of a parameter μ will be constructed by differentiating equations (1.2) with respect to the parameter, on the assumption that $x_i = x_i(\mu)$, and then letting $\Delta\mu \to 0$ in (1.3) divided by $\Delta\mu$. This gives a system of linear equations for the components of the vector x_i, $\mu = dx_i/d\mu = (x_{1,\mu}; \ldots x_{n+1,\mu})^T$

$$\begin{Vmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{n+1} \\ F_{1,1} & F_{1,2} & \cdots & F_{1,n+1} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ F_{n,1} & F_{n,2} & \cdots & F_{n,n+1} \end{Vmatrix} \begin{Vmatrix} x_{1,\mu} \\ x_{2,\mu} \\ \cdots\cdots \\ x_{n+1,\mu} \end{Vmatrix} = \begin{Vmatrix} 1 \\ 0 \\ \cdots \\ 0 \end{Vmatrix} \tag{1.4}$$

The index following the dot in the element $F_{j,i}$ indicates the component with respect to which the differentiation is performed.

By an optimal continuation parameter we mean a parameter for which the linearized system (1.4) is best conditioned, in the sense that small variations of the matrix elements and the right-hand sides of the system will cause the smallest variations of the solution. We shall show that the path length λ of the solution set curve of system (1.2) is an optimal parameter. The measure of conditioning of system (1.4), which we denote by D, will be the determinant of the system divided by the product of the squared norms of its rows [2]. By Hadamard's inequality for determinants, $|D| \in [0, 1]$. It has been shown [4] that the maximum value of $|D|$ indicates the best conditioning of the system of equations. In the present case one can prove the following assertion.

Lemma 1. The absolute value of the determinant of the system of equations (1.4), divided by the product of the squared norms of its rows, achieves its maximum value when the vector $\boldsymbol{\alpha}$ is tangent to the solution set curve of system (1.2) at each point of the curve.

Proof. Letting Δ denote the determinant of system (1.4), let us investigate the function $D = \Delta/d$ for an extremum, where

$$\Delta = (-1)^{i+1} \alpha_i \Delta_i \tag{1.5}$$

$$d = \prod_{i=1}^{n+1} t_i, \quad t_{\beta+1} = (F_{\beta,i} F_{\beta,i})^{1/2}, \quad \beta = 1, 2, \ldots, n; \quad i = 1, 2, \ldots, n+1, \tag{1.6}$$

(no summation over β), with $t_1 = (\alpha_i\alpha_i)^{1/2} = 1$ since α is a unit vector. Consequently, d is independent of α_i.

To find an extremum of the function D provided that α is a unit vector, we construct the Lagrange function

$$L = (-1)^{i+1}\alpha_i\Delta_i / d + \gamma(1-\alpha_i\alpha_i), \quad i = 1,2,\ldots,n+1$$

where γ is an undetermined Lagrange multiplier. The extremum of this function is achieved at $\alpha_k = (-1)^{k+1}\Delta_k/(2\gamma d)$ ($k = 1, 2, \ldots, n+1$). Using the equality $\alpha_k\alpha_k = 1$, we determine the Lagrange multiplier. Thus, an extremum of the Lagrange function is reached at

$$\alpha_k = \pm(-1)^{k+1}\Delta_k / (\Delta_i\Delta_i)^{1/2}$$

Substituting this expression for α_k into (1.5), we see that the determinant of system (1.4) must satisfy the equality

$$\Delta = \pm(\Delta_i\Delta_i)^{1/2} \tag{1.7}$$

and the Lagrange function attains its extremum when

$$\alpha_k = (-1)^{k+1}\Delta_k / \Delta = dx_k / d\mu \tag{1.8}$$

The Lagrange multiplier is then equal to

$$\gamma = \Delta/(2d) = D/2$$

Analysis of the second differential of the Lagrange function as a quadratic form in the differentials $d\alpha_k$ shows that the absolute value of the function $D = \Delta/d$ will then take its maximum value $|D| = (\Delta_i\Delta_i)^{1/2}/d$. Indeed, the sign of the second differential of the Lagrange function

$$d^2L = -2\gamma(d\alpha_i d\alpha_i)$$

is determined by the Lagrange multiplier γ, which is positive if $D > 0$, and so D takes its maximum value. The sign of γ is negative if $D < 0$, in which case D takes its minimum value.

We have thus proved that the vector α, which determines the continuation parameter μ by (1.3), makes D take its largest possible value when it is a solution vector $(x_{1,\mu}; \ldots, x_{n+1,\mu})^T$ of the linearized system (1.4), i.e. when α points along the tangent to the solution set curve of system (1.2).

Let us examine the effect of perturbing the elements of the matrix of system (1.4) on its conditioning.

Lemma 2. The quadratic error in the solution of system (1.4) due to perturbation of the elements of its matrix is least when the vector α points along the tangent to the solution set curve of system (1.2) at each point of the curve.

Proof. Suppose that the first row in the matrix of system (1.4) is given with an error. Let α have the form $(\alpha_1, \ldots, \alpha_{j-1}, \alpha_j + \varepsilon, \alpha_{j+1}, \ldots, \alpha_{n+1})^T$. The determinant Δ_ε of the system can be expressed in terms of the determinant Δ of the original system

$$\Delta_\varepsilon = \Delta + (-1)^{j+1}\varepsilon\Delta_j = \Delta(1 + (-1)^{j+1}\Delta_j\varepsilon / \Delta)$$

Since we are considering small perturbations ε, the components of the perturbed solution $y_{i,\mu}$ may be written in the form

$$y_{i,\mu} = (-1)^{i+1}\Delta_i / \Delta_\varepsilon \approx (-1)^{i+1}\Delta_i / \Delta(1 - (-1)^{j+1}\varepsilon\Delta_j / \Delta)$$

Then the components δ_i of the error vector $\delta = (\delta_1, \ldots, \delta_{n+1})^T$ of the solution of the perturbed system may be calculated using the formulae

$$\delta_i = y_{i,\mu} - x_{i,\mu} = (-1)^{i+j+1} \varepsilon \Delta_j \Delta_i / \Delta^2$$

Let us investigate the squared error $\delta = \varepsilon^2 \Delta^2_j \Delta_i \Delta_i / \Delta^4$ for an extremum, on the assumption that α is a unit vector. The Lagrange function may be written in the form

$$L = \varepsilon^2 \Delta_j^2 \Delta_i \Delta_i / \Delta^4 + \gamma(\alpha_i \alpha_i - 1), \quad i = 1, 2, \ldots, n+1$$

This function is a minimum when

$$\alpha_k = 2\varepsilon^2 \Delta_j^2 \Delta_i \Delta_i (-1)^{k+1} \Delta_k / (\gamma \Delta^5), \quad k = 1, 2, \ldots, n+1 \tag{1.9}$$

Dividing the kth equation of these relations by the mth, we obtain an equality which enables us to express α_m in terms of α_k

$$\alpha_m = (-1)^{m-k} \alpha_k \Delta_m / \Delta_k \tag{1.10}$$

Then the determinant (1.5) of system (1.4) may be written

$$\begin{gathered} \Delta = (-1)^{m-1} \alpha_m \Delta_m = (-1)^{-k-1} \alpha_k \Delta_m \Delta_m / \Delta_k \\ m = 1, 2, \ldots, n+1. \end{gathered} \tag{1.11}$$

In that case the system of equations (1.9) is easily solved for α_k

$$\alpha_k = \left(\frac{2\varepsilon \Delta_j^2}{\gamma (\Delta_m \Delta_m)^4} \right)^{1/6} (-1)^{k+1} \Delta_k \tag{1.12}$$

Note that there is no summation over k in (1.10) and (1.11).

To find the Lagrange multiplier γ, we substitute (1.12) into the equality $\alpha_i \alpha_i = 1$. Then $\gamma = 2\varepsilon^2 \Delta_j^2 / (\Delta m \Delta m)$, and formula (1.12) becomes

$$\alpha_k = (-1)^{k+1} \Delta_k / (\Delta_m \Delta_m)^{1/2} \tag{1.13}$$

Substituting these values of α_k into (1.5), we see that $\Delta = (\Delta m \Delta m)^{1/2}$, and equalities (1.13) are identical with the equalities for $dx_k/d\mu$, i.e. equalities (1.8) are true, which it was required to prove.

Let us study the effect of perturbing the right-hand sides of system (1.4) on its conditioning.

Lemma 3. The squared error in the solution of system (1.4) due to perturbation of the right-hand sides of the system is least when the vector α points along the tangent to the solution set curve of system (1.2) at each point of the curve.

Proof. Suppose the vector of the perturbed right-hand side of system (1.4) has the form $(1 + \varepsilon, 0, \ldots, 0)^T$. Then the error vector will be $\delta = \varepsilon(x_{1,\mu}; \ldots, x_n+_{1,\mu})^T$ and the squared error becomes

$$\delta^2 = \varepsilon^2 \Delta_i \Delta_i / \Delta^2, \quad i = 1, 2, \ldots, n+1$$

Defining the Lagrange function as

$$L = \varepsilon^2 \Delta_i \Delta_i / \Delta^2 + \gamma(\alpha_i \alpha_i - 1)$$

and looking for its extremum as described in the previous lemmas, we see that the components of the vector must satisfy equalities (1.8) at an extremum point, which proves the lemma.

We can now finally prove the following theorem.

Theorem. The system of linearized equations (1.4) is best conditioned if and only if the continuation parameter for solutions of the system of non-linear equations (1.2) is the path length of the solution set curve of the latter system.

Proof. Necessity. According to our definition of conditioning, Lemmas 1–3, taken together, state the following: the system of linear equations (1.4) is best conditioned when the vector $\boldsymbol{\alpha}$ points along the tangent to the solution set curve of the non-linear system (1.2) at each point of the curve, i.e. when equalities (1.8) hold. In view of this fact, the equality $\alpha_i\alpha_i = 1$ may be written in the form

$$(d\mu)^2 = dx_i dx_i, \quad i = 1,2,\dots,n+1 \tag{1.14}$$

whence it follows that $d_\mu = (dx_i dx_i)^{1/2}$ is the differential of the path length of the solution set curve of system (1.2). If we assume that the initial point of the curve is that corresponding to $\mu = 0$, the continuation parameter will equal the length of the curve measured from that point. This proves necessity.

Sufficiency. We choose the path length of the solution set curve of system (1.2) to be the continuation parameter μ. The vector $\boldsymbol{\tau}$ tangent to the curve will be $\boldsymbol{\tau} = (x_{1,\mu}; \dots, x_{n+1,\mu})^T$. As pointed out previously, the unit vector $\boldsymbol{\alpha}$ determines the direction in which the solution of problem (1.2) is continued. Hence, by our special choice of continuation parameter, it must also point along the tangent to the solution set curve, i.e. the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ must be collinear. But they are also equal, since $\boldsymbol{\tau}$ is also a unit vector. Indeed, the differential of the continuation parameter, as an element of path length, must satisfy (1.14). If this equality is divided by $(d\mu)^2$, we obtain

$$x_{i,\mu}x_{i,\mu} = \boldsymbol{\tau}^2 = 1, \quad i = 1,2,\dots,n+1.$$

Since the vectors are equal, so are their components. The components $dx_k/d\mu = x_{k,\mu}$ for any continuation parameter μ must satisfy the system of linear equations (1.4).

Consequently, equalities (1.8) must hold. The left-hand sides of these equalities make the functions occurring in the lemmas reach their extremum values. This in turn ensures that system (1.4) is best conditioned, which it was required to prove.

2. We will use this theorem to formulate a Cauchy problem for a system of ordinary differential equations

$$dy_i / dt = f_i(t, y_1,\dots,y_n), \quad y_i(t_0) = y_{i0}, \quad i = 1,2,\dots,n \tag{2.1}$$

An integral of this problem

$$F_i(t, y_1, \dots, y_n) = 0 \tag{2.2}$$

defines a certain integral curve, whose construction may be interpreted as the process of continuing the solution y on the argument-parameter t. This interpretation enables us to consider the problem of how to choose the best parameter for the continuation of the solution, and hence of how to choose the best argument in problem (2.1).

To solve the problem, let us assume that y_i and t are functions of a certain parameter μ whose increment can be expressed at each point of the integral curve in the form

$$\Delta\mu = \alpha_i\Delta y_i + \alpha_{n+1}\Delta t, \quad i = 1,2,\dots,n \tag{2.3}$$

where Δy_i, Δt are the corresponding increments. We have already discussed the meaning of the coefficients α_j $(j = 1, 2, \dots, n + 1)$.

Dividing (2.3) by $\Delta\mu \to 0$ and noting that

$$dy_i / dt = (dy_i / d\mu)(dt / d\mu)^{-1}$$

we can express the solution of Eqs (2.3) and (2.1) in the form

$$\alpha_i y_{i,\mu} + \alpha_{n+1} t_{,\mu} = 1, \quad y_{i,\mu} - f_i t_{,\mu} = 0 \tag{2.4}$$

$$(y_{i,\mu} = dy_i / d\mu, \quad t_{,\mu} = dt / d\mu)$$

System (2.4) may be viewed as continuation equations when constructing the solution set curve for system (2.2), which is an integral curve of problem (2.1).

The theorem proved above states that system (2.4) will be best conditioned if the argument-parameter μ is taken to be the path length λ measured along the solution set curve of the system (2.2), i.e. along an integral curve of problem (2.1).

In that case, taking (1.8) into consideration, we can write system (2.4) in the form

$$y_{i,\lambda} y_{i,\lambda} + t_{,\lambda}^2 = 1, \quad y_{i,\lambda} - f_i t_{,\lambda} = 0$$

This system is easily solved for $y_{i,\lambda}, t_\lambda$. Assuming that the initial point of problem (2.1) corresponds to $\lambda = 0$, we obtain the following Cauchy problem

$$dy_i / d\lambda = f_i / (1 + f_j f_j)^{1/2}, \quad y_i(0) = y_{i0}$$

$$dt / d\lambda = 1 / (1 + f_j f_j)^{1/2}, \quad t(0) = t_0, \quad i, j = 1, 2, \ldots, n \tag{2.5}$$

In previous publications [5, 6] we discussed some advantages of solving problem (2.5) compared with problem (2.1). A further advantage may be illustrated through the following model problem

$$dy_1 / dt = a_1 y_1, \quad dy_2 / dt = a_2 y_2 \tag{2.6}$$

where a_1, a_2 are real numbers. A numerical solution of this problem will be sought using Euler's method.

We shall show that one can proceed from the initial point $A_0(t_0, y_{10}, y_{20})$ of an integral curve to the final point B (see Fig. 1) in the minimum number of steps by varying the parameter λ and not the parameter t. In other words, we shall show that at any point of the integral curve

$$H_\lambda \cos\theta \geqslant H_t \tag{2.7}$$

where θ is the angle between the tangent to the integral curve and the t axis, H_t and H_λ are the least integration steps with respect to t and λ at which the iterative process described by Euler's formula ceases to converge.

The explicit scheme of Euler's method for equation (2.6) is

$$y_{i,m+1} = y_{i,m} + h_t a_i y_{i,m} = (1 + h_t a_i) y_{i,m}, \quad m = 1, 2, \ldots, i = 1, 2.$$

(h_t is the integration step length with respect to t). This scheme will be stable if $|\ 1 + h_t a_i\ | < 1$, i.e. for $a_i < 0$

$$H_t = -2/a_1, \quad a_1 < a_2 \tag{2.8}$$

Transforming problem (2.6) to the form (2.5), we obtain a system of three differential equations, in which the solution of the equation for t will be determined by the solutions of the equations for the functions y_i. If these equations are linearized in the neighbourhood of some value $y_i = y_{i,m}$ [6], the condition or the transformed problem to be stable when $a_1 \ll a_2$ becomes

$$|1 + h_\lambda \Omega_{ij}| < 1; \quad i, j = 1, 2; \quad i \neq j$$

$$\Omega_{ij} = a_i (1 + a_j^2 y_{j,m}^2)(1 + a_1^2 y_{1,m}^2 + a_2^2 y_{2,m}^2)^{-3/2}$$

where h_λ is the integration step length for λ. This inequality will be satisfied if $a_i < 0$ and

$$H_\lambda = \min_{i,j}(-2 / \Omega_{ij}) \tag{2.9}$$

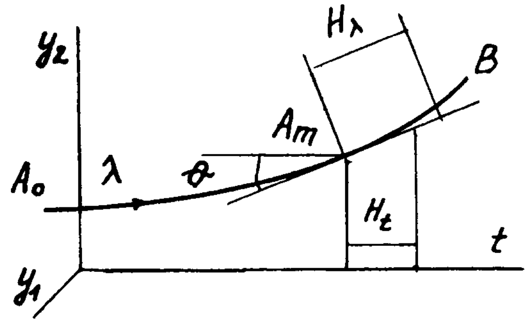

Fig. 1.

Since $\cos\theta = (1 + a_1^2 y_{1,m}^2 + a_2^2 y_{2,m}^2)^{-1/2}$, equalities (2.8) and (2.9) lead to a relationship that proves the truth of (2.7) and so demonstrates that the transformation proposed here is indeed effective.

3. As a test example, let us investigate the solution of a stiff system of equations. Consider the steady, straight-line flight of an aircraft in a plane without slipping, assuming that the parameters experience a small deviation from their initial values. The linearized equations of the perturbed motion of the aircraft may be written as follows [7]:

$$dy/dt = Ay, \quad y = (y_1, y_2, y_3, y_4)^T \tag{3.1}$$

$$A = \left\| \begin{matrix} -0.104 & 0.043 & -0.1 & 0 \\ -0.57 & -5.12 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ -12.574 & -43.68 & 0 & -9.672 \end{matrix} \right\|$$

The first two equations of system (3.1) describe the longitudinal and transverse perturbations of the aircraft's velocity, respectively. The last two equations describe the perturbation of the pitching angle.

It can be shown that the matrix A of system (3.1) has the following eigenvalues: $r_1 = 0.16$, $r_2 = -0.265$, $r_{3,4} = -7.4 \pm i6.2$. Clearly, $|r_1| < |r_2| \ll |r_3| = |r_4|$, implying that the system is stiff. Previously, using the PC1 program described in [8], we integrated this system, with initial data

$$y_1(0) = y_3(0) = y_4(0) = 0, \quad y_2(0) = 1 \tag{3.2}$$

Over the interval $t \in [0, 5]$. The computation required almost twice as much computer time compared with the solution of the same problem (3.1), (3.2) after a preliminary transformation to the form (2.5).

REFERENCES

1. VOROVICH I. I. and ZIPALOVA V. F., The solution of non-linear boundary-value problems of elasticity theory by the method of transformation to a Cauchy problem. *Prikl. Mat. Mekh.* **29**, 5, 894–901, 1965.
2. RIKS E., The application of Newton's method to the problem of elastic stability. *Trans. ASME, Ser. E, J. Appl. Mech.* **39**, 4, 1060–1065, 1972.
3. GRIGOLYUK E. I. and SHALASHILIN V. I., *Problems of Nonlinear Deformation*. Kluwer, Dordrecht, 1991.
4. ORTEGA J. M. and POOLE W. G., *An Introduction to Numerical Methods for Differential Equations*. Pitman, Marshfield, MA, 1981.
5. SHALASHILIN V. I. and KUZNETSOV Ye. B., The Cauchy problem for non-linearly deformable systems as a problem of continuing the solution on a parameter. *Dokl. Ross. Akad. Nauk* **329**, 4, 426–428, 1993.
6. KUZNETSOV Ye. B. and SHALASHILIN V. I., The Cauchy problem as a problem of continuing the solution as a function of a parameter. *Zh. Vychisl. Mat. Mat. Fiz.* **33**, 12, 1792–1805, 1993.
7. RAKITSKII Yu. V., USTINOV S. M. and CHERNORUTSKII I. G., *Numerical Methods for Solving Stiff Systems*. Nauka, Moscow, 1979.

8. KUZNETSOV Ye. B. and SHALASHILIN V. I., The Cauchy problem for deformable systems as a problem of continuing the solution as a function of a parameter. *Izv. Ross. Akad. Nauk. MTT* 6, 145–152, 1993.

Translated by D.L.