Hence, every solution of (5.22) which starts in $\mathcal{N}$ will converge such that

$$\|y(x)\|^2 \leqslant 2\exp\left(-\tan(\delta k t + c)\right) \tag{5.27}$$

for

$$c = -\tan^{-1}\left(\ln\left(\tfrac{1}{2}\|y(0)\|^2\right)\right). \tag{5.28}$$

Acknowledgement

The author is indebted to Dr.William Schiesser for pointing out the utility of solving such problems by differential methods and for many useful discussions of the various points involved.

References

[1] W. Hahn, *Stability of Motion*, Springer, Berlin 1967.
[2] D. G. B. Edelen, Int. J. Engng. Sci. 12 (1974), p. 121.
[3] D. F. Davidenko, Ukr. Mat. Z. 5 (1953), p. 196; F. H. Branin, *Memoirs IEEE Conference on Systems, Networks and Computers*, Oaxtepec, Mexico 1971.

Presented to the Semester
Mathematical Models and Numerical Methods
(February 3–June 14, 1975)

BANACH CENTER PUBLICATIONS
VOLUME 3

DIFFERENTIAL PROCEDURES FOR SYSTEMS OF IMPLICIT RELATIONS AND IMPLICITLY COUPLED NONLINEAR BOUNDARY VALUE PROBLEMS

DOMINIC G. B. EDELEN

Center for the Application of Mathematics, Lehigh University, Bethlehem, Pennsylvania 18015, USA

1. Introduction and preliminary considerations

Theory, in the form of the implicit function theorem, is quite specific about solvability of implicit systems of relations. Numerical procedures for the realization of the solutions, when the conditions of the implicit function theorem are met, is quite another matter. The purpose of this note is to give families of differential procedures for solving systems of implicit relations.

Let x denote a vector (column matrix) in n-dimensional number space E_n and let α denote a vector in m-dimensional number space E_m. Let $f(x, \alpha)$ be a given vector valued function that is defined for all x in a given n-dimensional region R_n of E_n and all α in E_m and which takes its values in E_m. Thus, $f(x, \alpha)$ is a mapping of $R_n \times E_m$ into E_m. We further assume that f is continuous over its domain of definition and that its matrices of partial derivatives

$$A(x, \alpha) = V_x f, \quad B(x, \alpha) = V_\alpha f \tag{1.1}$$

are also continuous on the domain of definition of f. To be more specific, let i, j, k be indices which can take on values from 1 through n and let a, b, c be indices that can take on values from 1 through m. We then have

$$A = \left((\partial f_a/\partial x_i)\right), \quad B = \left((\partial f_a/\partial \alpha_b)\right),$$

so that A is an m-by-n matrix and B is an m-by-m matrix.

The problem we wish to solve is that of constructing differential procedures for obtaining $\alpha = \varphi(x)$ as a vector valued function of x such that

$$f(x, \varphi(x)) = 0 \tag{1.2}$$

is satisfied at all points x of E_n for which such solutions exist. The implicit function theorem tells us that if there are elements $\bar{x}$ and $\bar{\alpha}$ such that

$$f(\bar{x}, \bar{\alpha}) = 0 \tag{1.3}$$

and $\det\left(B(\bar{x}, \bar{\alpha})\right) \neq 0$, then there exists an n-dimensional neighborhood $N(\bar{x})$ of

$\bar{x}$ in R_n such that $\alpha = \varphi(x)$ exists on $N(\bar{x})$ and $f(x, \varphi(x)) = 0$ is identically satisfied for all x in $N(\bar{x})$. This neighborhood $N(\bar{x})$ has the additional property that $\det\left|B(x, \varphi(x))\right| \neq 0$. Further, for all continuously differentiable functions $x(t)$ with range in $N(\bar{x})$, we have

$$A\frac{dx}{dt} + B\frac{d\overset{*}{\varphi}}{dt} = 0 \tag{1.4}$$

(matrix multiplication); that is, $d\overset{*}{f}/dt = 0$. Here, we have set

$$\overset{*}{\varphi}(t) = \varphi(x(t)), \quad \overset{*}{f}(t) = f(x(t), \overset{*}{\varphi}(t)). \tag{1.5}$$

Clearly, there are three numerical problems. The first is that of finding two numerical vectors $\bar{x}$ and $\bar{\alpha}$ such that $f(\bar{x}, \bar{\alpha}) = 0$ and $\det(B(\bar{x}, \bar{\alpha})) \neq 0$. The second is that of finding values of $\varphi(x)$ for given values of x in $N(\bar{x})$ such that $\alpha = \varphi(x)$ satisfies $f(x, \alpha) = 0$. The third problem is that of finding other pairs of numerical vectors $\bar{y}$ and $\bar{\beta}$ such that $f(\bar{y}, \bar{\beta}) = 0$, $\det(B(\bar{y}, \bar{\beta})) \neq 0$, and the pair $(\bar{y}, \bar{\beta})$ does not belong to the solution set obtained by starting with the pair $(\bar{x}, \bar{\alpha})$. The possibility of the existence of such additional pairs $(\bar{y}, \bar{\beta})$ can not be excluded since the implicit function theorem only gives sufficient conditions for the existence of local single-valued solutions. For instance, consider the problem of solving $x^2 + \alpha^2 - r^2 = 0$. Clearly there are two solutions $\alpha = \pm\sqrt{r^2 - x^2}$ for each x in the open interval $(-r, r)$ and only one of these can be obtained by starting with the pair of points $x = 0$, $\alpha = r$.

The second problem is treated first in Section 2 since it is the easier of the three. The first problem is then solved in Section 3 and Section 4 takes up the third problem.

One of the most vexing classes of problems in which implicit relations occur consists of nonlinear boundary value problems in which there is coupling that is defined through a collection of implicit relations. Section 5 gives a method of solution for such problems by converting them to initial value problems in one additional variable t such that the large time limit of the solution of the initial value problem solves the given boundary value problem. Such a procedure is particularly useful in that the initial data can be chosen as any convenient vector of functions that satisfies the given boundary conditions.

2. Generation of a solution on a grid in E_n

We assume in this section that we know a pair of numerical vectors $\bar{x}$ and $\bar{\alpha}$ such that $f(\bar{x}, \bar{\alpha}) = 0$ and $\det(B(\bar{x}, \bar{\alpha})) \neq 0$. Let e_i, $i = 1, \ldots, n$, be an orthonormal system of basis vectors for E_n. The set of all points

$$x_1(t) = \bar{x} + te_1 \tag{2.1}$$

defines a line in E_n that passes through the point $\bar{x}$ and is parallel to the vector e_1. It is therefore meaningful to ask for a vector function $\alpha_1(t)$ such that

$$f(x_1(t), \alpha_1(t)) = 0, \tag{2.2}$$

$$\alpha_1(0) = \bar{\alpha}. \tag{2.3}$$

When (2.1) is used to calculate the time derivative of (2.2), we obtain the system of differential equations

$$B(\bar{x} + te_1, \alpha_1(t))\frac{d\alpha_1}{dt} = -A(\bar{x} + te_1, \alpha_1)e_1 \tag{2.4}$$

and the initial data (2.3) for the determination of the vector $\alpha_1(t)$. Since $\det(B(\bar{x}, \bar{\alpha})) \neq 0$ and B is a continuous matrix-valued functions of its arguments, it follows that there is a maximal open interval n_1 that contains the point $t = 0$ and is such that a solution exists to (2.4) subject to the initial data (2.2) for all t in n_1. Clearly, $\det\left|B(\bar{x} + te_1, \alpha_1(t))\right| \neq 0$ for t in n_1, and hence the system (2.4) can be solved for the vector $d\alpha_1(t)/dt$ and then integrated numerically. The reason that we can not guarantee existence of solutions to the system (2.4) for all t is that $\det\left|B(x + te_1, \alpha(t))\right|$ can go to zero as t approaches the endpoints of the closure of n_1. The solution $\alpha_1(t)$ allows us to generate a mesh of pairs of values by

$$x_k = \bar{x} + khe_1, \tag{2.5}$$

$$\alpha_k = \alpha_1(kh), \tag{2.6}$$

where h is a given constant (mesh constant) and k ranges over the integers such that kh belongs to n_1.

We now start with each of the points $\bar{x} + khe_1$ and generate the line

$$x_{2k}(t) = \bar{x} + khe_1 + te_2. \tag{2.7}$$

Since $\det\left|B(\bar{x} + khe_1, \alpha_1(kh))\right| \neq 0$, we can determine vector function $\alpha_{2k}(t)$ for each value of k by solving the system of differential equations

$$B(\bar{x} + khe_1 + te_2, \alpha_{2k}(t))\frac{d\alpha_{2k}}{dt} = -A(\bar{x} + khe_1 + te_2, \alpha_{2k}(t))e_2 \tag{2.8}$$

subject to the initial data

$$\alpha_{2k}(0) = \alpha_k = \alpha_1(kh), \tag{2.9}$$

for all t in a maximal open interval n_{2k} that contains $t = 0$. As before, we construct the two-dimensional mesh of pairs

$$x_{kj} = \bar{x} + khe_1 + jhe_2, \tag{2.10}$$

$$\alpha_{kj} = \alpha_{2k}(jh), \tag{2.11}$$

by allowing t to take on the values jh for j ranging over the integers such that jh belongs to n_{2k}.

This procedure can be continued by successive use of the basis vectors $e_3, e_4, \ldots$ until a mesh of pairs of numerical vectors x and α is built up such that the mesh of numerical vectors x constitutes a set of grid points on E_n with grid spacing h that spans an n-dimensional region R_n of E_n. Clearly, the mesh so generated constitutes a numerical solution of $f(x, \alpha) = 0$ on the region R_n. It is also clear that the region R_n is maximal since each of the open intervals n_1, n_{2k}, n_{3kj}, etc. is maximal.

3. Generation of the starting values for the grid

The problem we now have to solve is that of finding a pair of numerical vectors $\bar{x}$ and $\bar{\alpha}$ such that $f(\bar{x}, \bar{\alpha}) = 0$.

We start by picking a specific numerical vector $\bar{x}$. It is then useful to introduce the simplifying notation

$$\bar{f}(\alpha) = f(\bar{x}, \alpha), \tag{3.1}$$

$$\bar{B}(\alpha) = B(\bar{x}, \alpha). \tag{3.2}$$

In these terms, we need to find a vector α such that

$$\bar{f}(\alpha) = 0. \tag{3.3}$$

If there is *a priori* information to the effect that $\det(\bar{B}(\alpha)) \neq 0$ for all α in E_m, then an intrinsic economy can be achieved in the solution of (3.3). We therefore divide our considerations into two separate cases.

Case I: $\mathrm{Det}(\bar{B}(\alpha)) \neq 0$ *for all* α.

In this case, we can use the Davidenko method [1]. This method consists of picking an arbitrary initial vector α_0 for α and solving the system of autonomous differential equations

$$\frac{d\bar{f}(\alpha)}{dt} = \bar{B}(\alpha)\frac{d\alpha}{dt} = -K\bar{f}(\alpha), \tag{3.4}$$

where K is a positive constant. Since $\det(\bar{B}) \neq 0$ for all α, this system is equivalent to

$$\frac{d\alpha}{dt} = -K(\bar{B})^{-1}\bar{f}(\alpha) \tag{3.5}$$

where $(\bar{B})^{-1}$ is the inverse of the matrix $\bar{B}$. It then follows, on setting

$$V(\alpha) = \tfrac{1}{2}\bar{f}(\alpha)^{\mathrm{T}}\bar{f}(\alpha), \tag{3.6}$$

($\bar{f}^{\mathrm{T}}$ = transpose of $\bar{f}$) that

$$\frac{dV}{dt} = \bar{f}^{\mathrm{T}}B\frac{d\alpha}{dt} = -Kf^{\mathrm{T}}f = -2KV, \tag{3.7}$$

and hence

$$\overset{*}{V}(t) = \tfrac{1}{2}\bar{f}(\alpha(t))^{\mathrm{T}}\bar{f}(\alpha(t)) = V_0\exp(-2Kt). \tag{3.8}$$

Accordingly, $\lim_{t\to\infty}\{\alpha(t)\} = \bar{\alpha}$ exists and satisfies $\bar{f}(\bar{\alpha}) = 0$. This solves the problem for we then have the required pair of vectors $\bar{x}$ and $\bar{\alpha}$ such that $f(\bar{x}, \bar{\alpha}) = 0$.

It is to be noted in passing that $\bar{\alpha}$ is the only vector such that $f(\bar{x}, \bar{\alpha}) = 0$ for the given numerical vector $\bar{x}$. This follows from the fact that $\det(\bar{B}(\alpha)) \neq 0$ for all α is sufficient in order to insure that the mapping $\beta = \bar{f}(\alpha)$ is a one-to-one mapping of E_m to E_m.

Case II. *No a priori information about* $\det(\bar{B})$.

Clearly, the Davidenko method can not be used in this case since the coefficient matrix $\bar{B}(\alpha)$ of the differential system (3.4) may become singular for some values of the vector α. Instead of the system (3.4), we now use the autonomous differential system

$$\frac{d\alpha}{dt} = -K\bar{B}(\alpha)^{\mathrm{T}}\bar{f}(\alpha) \tag{3.9}$$

where K is a positive constant. If we assume that each of the entries of the matrix $\bar{B}(\alpha)$ is a C^1 function of the vector α, the right hand sides of (3.9) are C^1 functions of α so that there is no question about existence of solutions. Further, with $V(\alpha)$ defined by (3.6), the system (3.9) implies that

$$\frac{dV}{dt} = \bar{f}^{\mathrm{T}}\bar{B}\frac{d\alpha}{dt} = -K\bar{f}^{\mathrm{T}}\bar{B}\bar{B}^{\mathrm{T}}\bar{f} = \left(\frac{-1}{K}\right)\frac{d\alpha^{\mathrm{T}}}{dt}\frac{d\alpha}{dt} \leqslant 0 \tag{3.10}$$

with $dV/dt = 0$ if and only if $\bar{B}^{\mathrm{T}}f = \dfrac{d\alpha}{dt} = 0$. Accordingly, $V(\alpha(t))$ decreases along every solution of the system (3.9) whenever $d\alpha/dt \neq 0$. Since $V \geqslant 0$ with $V(\alpha) = 0$ if and only if $\bar{f}(\alpha) = 0$, every solution of the system (3.9), with an initial value of α such that $d\alpha(0)/dt \neq 0$, will converge for large values of t to values of α for which either $V(\alpha) = 0$ or $V(\alpha)$ = relative minimum. For those large time limit points such that $V(\alpha) = 0$, we have $\bar{f}(\alpha) = 0$ and the problem is solved. For those large time limit points such that $V(\alpha) \neq 0$, we have $\bar{B}^{\mathrm{T}}(\alpha)\bar{f}(\alpha) = 0$ but $\bar{f}(\alpha) \neq 0$, and hence such points satisfy $\det(\bar{B}(\alpha)) = 0$. We have thus established that the large time limit points of the system (3.9) exist for all initial data and contain the set of all points α for which $\bar{f}(\alpha) = 0$. We also obtain spurious convergence points where $\bar{B}^{\mathrm{T}}(\alpha)\bar{f}(\alpha) = 0$, but these are easily eliminated by testing to see whether the large time limit points satisfy $f(\alpha) = 0$. This testing procedure is a simple one since the value of the vector α to be tested is known.

We note in passing that $V(\alpha)$ is proportional to the square of the error in solving $\bar{f}(\alpha) = 0$, and that the system (3.9) can be written in the equivalent form

$$\frac{d\alpha}{dt} = -KV_{\alpha}V(\alpha). \tag{3.11}$$

Thus, the differential procedure (3.9) is the differential analog of steepest descent methods; that is, $d\alpha/dt$ is a vector that points in the direction of maximal decrease of the error.

4. Multiple branch solutions

If we know that $\det(B(\bar{x}, \alpha)) \neq 0$ for all α, then we have seen that there is only one vector $\bar{\alpha}$ such that $f(\bar{x}, \bar{\alpha}) = 0$. Thus, there is one and only one solution of $f(x, \alpha) = 0$ over the point $x = \bar{x}$. On the other hand, if $\det(B(\bar{x}, \alpha))$ can vanish

for some vector α, there can be more than one solution of $f(\bar{x}, \alpha) = 0$. This follows from the results established under Case II of the previous section. If there are several solutions $\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_p$ of $f(\bar{x}, \alpha) = 0$ that are obtained from the differential procedure (3.9) by allowing the initial α vector to change, we can use the method described in Section 2 to generate a solution grid for each starting pair $(\bar{x}, \bar{\alpha}_1), (\bar{x}, \bar{\alpha}_2), \ldots$ $\ldots, (\bar{x}, \bar{\alpha}_p)$. Clearly, in this instance, $f(x, \alpha) = 0$ has p-fold valued solutions. It is also clear that the size of the grid over which each of these solution meshes is defined can be different for each different starting point, $(\bar{x}, \bar{\alpha}_a)$. We do not consider in this note the very complex problems of whether the various value surfaces of the solution fit together where $\det(B) = 0$, the manner of this fitting together process, or the order in which the surfaces are fitted together.

5. Nonlinear boundary value problems with implicit coupling

The results established above provide a means whereby differential procedures can be constructed for the solution of nonlinear boundary value problems with implicit coupling.

We again suppose that we are given the vector valued function $f(x; \alpha)$ and that $\alpha = \varphi(x)$ is defined implicitly by

$$f(x, \alpha) = 0. \tag{5.1}$$

We further assume that the matrix $B(x, \alpha)$ defined by (1.1) is nonsingular so that

$$d\alpha = -B^{-1}A\,dx \tag{5.2}$$

holds for all vectors x. Now, consider the collection of vector valued functions $x = x(z, t)$ defined for all z in the closed interval $[a, b]$ and all t in $[0, \infty)$ and are such that

$$x(a, t) = \gamma_1, \quad x(b, t) = \gamma_2, \tag{5.3}$$

$x(z, t)$ has continuous second derivatives with respect to z, continuous first derivatives with respect to t and continuous mixed derivatives with respect to x and t. We wish to determine one such vector valued function which is such that it renders the functional

$$L(x; \alpha) = \int_a^b \mathscr{L}(z, x, \partial_z x, \alpha)\,dz \tag{5.4}$$

stationary in value. We assumed that $\mathscr{L}$ is an analytic function of its $2m+n+1$ arguments such that

$$L(x; \alpha) \geqslant J \tag{5.5}$$

for all $x(z, t)$ and all $\alpha(z, t)$ that are related by (5.1). Here $\partial_z x$ denotes the derivative with respect to z. We shall also use $\partial_t x$ to denote the derivative with respect to t. The standard methods from the calculus of variations [2] yield

$$\delta L(x; \alpha) = \int_a^b \left[\left\{ \frac{\partial \mathscr{L}}{\partial x} - \partial_z \left(\frac{\partial \mathscr{L}}{\partial(\partial_z x)} \right) \right\} \cdot \delta x + \frac{\partial \mathscr{L}}{\partial \alpha} \cdot \delta \alpha \right] dz. \tag{5.6}$$

If we define the row matrix of Euler–Lagrange derivatives by

$$\{E|\mathscr{L}\}_x = \frac{\partial \mathscr{L}}{\partial x} - \partial_z \left(\frac{\partial \mathscr{L}}{\partial(\partial_z x)} \right) \tag{5.7}$$

and note that (5.2) implies $\delta\alpha = B^{-1}A\delta x$, (5.6) becomes

$$\delta L(x; \alpha) = \int_a^b \left[\{E|\mathscr{L}\}_x - \frac{\partial \mathscr{L}}{\partial \alpha} B^{-1}A \right] \delta x\,dz, \tag{5.8}$$

so that we seek solutions to the implicit boundary value problem

$$\{E|\mathscr{L}\}_x - \frac{\partial \mathscr{L}}{\partial \alpha} B^{-1}A = 0, \qquad f(x, \alpha) = 0, \tag{5.9}$$

$$x(a, t) = \gamma_1, \quad x(b, t) = \gamma_2. \tag{5.10}$$

Now, as is well known, boundary value problems are difficult to solve. We thus proceed to construct an initial value problem whose solution has a large time limit that satisfies the given boundary value problem by making use of the occurrence of the variable t in the above formulation. Consider the system of equations

$$-\partial_t x = \{E|\mathscr{L}\}_x^{\mathrm{T}} - \left(\frac{\partial \mathscr{L}}{\partial \alpha} B^{-1}A \right)^{\mathrm{T}} \overset{\text{def}}{\equiv} w \quad (\mathrm{T} = \text{transpose}), \tag{5.11}$$

$$\partial_t \alpha = -B^{-1}A\partial_t x = -B^{-1}Aw, \tag{5.12}$$

$$x(a, t) = \gamma_1, \quad x(b, t) = \gamma_2, \tag{5.13}$$

$$x(z, 0) = x_0(z), \quad x_0(a) = \gamma_1, \quad x_0(b) = \gamma_2. \tag{5.14}$$

Use of the procedure given in Section 3 and the fact that $\det(B) \neq 0$ allows us to obtain the compatible initial data

$$\alpha(z, 0) = \alpha_0(z) \tag{5.15}$$

such that

$$f(x_0(z), \alpha_0(z)) = 0 \tag{5.16}$$

for all z in $[a, b]$. It now remains to show that this initial value problem will have solutions whose large time limits satisfy the given implicit boundary value problem.

It follows immediately from (5.8) with $\delta \equiv \partial_t$ that

$$\frac{d}{dt} L(x; \alpha) = \int_a^b \left\{ \{E|\mathscr{L}\}_x - \frac{\partial \mathscr{L}}{\partial \alpha} B^{-1}A \right\} \partial_t x\,dz,$$

and hence (5.11) yields

$$\frac{d}{dt} L(x; \alpha) = -\int_a^b ww^{\mathrm{T}} dz \leqslant 0 \tag{5.17}$$

with equality holding if and only if $w = 0$; that is, if and only if (5.8) holds, in which case (5.12) yields $\partial_t \alpha = 0$. Now, we have

$$L(x; \alpha) \geqslant J, \quad \frac{d}{dt} L(x; \alpha) \leqslant 0$$

and hence we obtain a contradiction unless $\lim_{t\to\infty} \frac{d}{dt} L(x; \alpha) = 0$. We have seen, however, that this can be the case if and only if $\lim_{t\to\infty} w = 0$, in which case we also obtain $\lim_{t\to\infty} \partial_t \alpha = 0$. The desired result is then established on noting that all solutions of (5.12) satisfy $f(x, \alpha) = 0$ since $\det(B) \neq 0$.

References

[1] D. F. Davidenko, Ukr. Mat. Z. 5 (1953), p. 196; F. H. Branin, *Memoirs IEEE Conference on Systems, Networks and Computers*, Oaxtepec, Mexico 1971.

[2] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs 1963.

Presented to the Semester
Mathematical Models and Numerical Methods
(February 3–June 14, 1975)

BANACH CENTER PUBLICATIONS
VOLUME 3

МЕТОДЫ РЕШЕНИЯ НЕКОРРЕКТНЫХ ЭКСТРЕМАЛЬНЫХ ЗАДАЧ

А. Н. ТИХОНОВ, Ф. П. ВАСИЛЬЕВ

Московский Государственный Университет, Факультет Вычислительной Математики и Кибернетики, Москва, СССР

При численном решении прикладных задач важное значение имеет тот факт, будет ли решение рассматриваемой задачи непрерывно зависеть от исходных данных или, иначе говоря, будет ли искомое решение устойчивым по отношению к возмущениям входных данных в той или иной топологии. Если решение устойчиво по входным данным, то можно быть уверенным в том, что достаточно малые погрешности в задании входных данных приведут к малым погрешностям в определении решения. Иное дело решать неустойчивую или, как говорят, некорректную задачу, решение которой не является непрерывно зависящим от входных данных: в этом случае приближённое решение задачи, отвечающее неточным входным данным, может как угодно сильно отличаться от искомого точного решения. Между тем некорректные задачи возникают в самых различных областях физики, техники, экономики и т.д. [1], и возникает важная проблема: как численно решать такие задачи?

Основы теории и методов решения некорректных задач заложены в работах А. Н. Тихонова, В. К. Иванова, М. М. Лаврентьева [1]–[15]. К настоящему времени создана достаточно полная общая теория некорректных задач, созданы приближённые методы решения таких задач, с помощью которых успешно решены и решаются многие прикладные задачи. По поводу общей теории некорректных задач и её приложений, а также библиографии по этим вопросам отсылаем читателя к [1], [15], [16].

В настоящей статье будут систематически и единообразны исследованы методы решения некорректных задач, непосредственно связанных с задачами математического программирования, оптимального управления и другими экстремальными задачами; из литературы здесь упомянем [1], [8]–[72].