KYBERNETIKA — VOLUME *15* (1979), NUMBER 3

On the Numerical Solution of Implicit Two-Point Boundary-Value Problems

JAROSLAV DOLEŽAL, JIŘÍ FIDLER

This paper presents an application of the so-called modified quasilinearization method to the solution nonlinear, implicite two-point boundary-value problems. Analogously as in the explicit case the resulting algorithm exhibits the descent property in the auxiliary performance index which measures the cumulative error in the differential equations and the boundary conditions. Practical importance of this algorithm is illustrated on several examples solved in detail. A comparison with the classical Newton-Raphson method is also included.

1. INTRODUCTION

A modified quasilinearization method for solving nonlinear two-point boundary-value problems was suggested and extensively studied by Miele et al. in [1–3]. An alternative derivation of this method including also the convergence proof is given by Roberts and Shipman in [4]. Thus the modified quasilinearization method is obtained when the appropriate realization of the abstract Newton method is applied to the studied problem and only a partial correction to the current profile is taken. For further details see [4].

The modified quasilinearization technique has two characteristic features: (i) the inhomogeneous terms in the linear differential equations and the corresponding boundary conditions, which define the sequence of linear two-point boundary-value problems, are multiplied by a scalar α, $0 < \alpha \leqq 1$, such that $\alpha = 1$ implies the Newton-Raphson method; and (ii) an auxiliary performance index $P(\alpha)$ is introduced which, after the appropriate choice of α, decreases in each iteration and exhibits the so-called descent property. In this way it is possible to overcome the difficulties of the original Newton-Raphson scheme caused by excessive magnitude of variations. This circumstance is further supported by a number of selected examples presented [1–3], so that the reader obtains an impression that the modified quasilinearization method can solve any given problem.

The authors of this paper studied and tested this method and also its discrete version in [5 – 6]. It has shown that there exist also problems which it is not possible to solve applying the modified quasilinearization method due to the descent property requirement. On the other hand, the classical Newton-Raphson method solved some (not all) of these problems quite easily not always decreasing the performance index P.

This behaviour of the modified quasilinearization method has grown more apparent when applying this method to the implicit two-point boundary-value problems, either continuous or discrete [6 – 9]. The word "implicit" denotes the problems having the equation not solved with respect to the highest derivative or difference. In fact, one can speak about several classes of problems according to their solvability by the both discussed methods.

Having this in mind the aim of this paper is twofold. First, a generalization of the modified quasilinearization method is presented to deal with implicit two-point boundary-value problems. These problems are not only interesting as such, but also arise, at least principally, in some other fields, e.g., in the calculus of variations (Euler-Lagrange equation, see [10]) and discrete optimal control (intrinsically discrete problems, see [6]). Up to now this type of boundary-value problems is not very much treated, especially from the computational point of view. To the authors' knowledge only the paper of Edelen [11] deals with these problems applying the implicit function theorem. The descibed procedure concerns primarily the nonlinear equations, but it can be applied also to boundary-value problems. To avoid the evident difficulties due to the eventual branching and non-uniqueness, only the "regular" case will be considered, because otherwise a general systematic approach would be hardly possible.

Second aim is to give an exhaustive classification of problems based upon their behaviour when treated by both, modified quasilinearization and Newton-Raphson methods. Several examples are solved in detail to illustrate this matter. Let us mention the fact that such classification of a particular problem can heavily depend also on the initial solution estimate.

2. MODIFIED QUASILINEARIZATION METHOD

For the sake of comparison and for the reader's convenience the notation of [1 – 3] is preserved whenever possible. Matrix notation will be used throughout the paper In this respect all appearing vectors are supposed to be column-vectors, while the gradients of various functions are always treated as row-vectors. All further defined functions are assumed to be continuously differentiable. Various partial derivatives will be indicated by the appropriate lower indices.

To begin, let us consider an implicit differential equation

$$\varphi(\dot{x}, x, t) = 0\,, \quad 0 \leqq t \leqq 1\,, \tag{1}$$

which is subject to the boundary conditions

$$f[x(0)] = 0\,, \quad g[x(1)] = 0\,, \quad h[x(0), x(1)] = 0\,. \tag{2}$$

Here $x \in E^n$, $\varphi : E^{2n+1} \to E^n$, $f : E^n \to E^p$, $g : E^n \to E^q$ and $h : E^{2n} \to E^r$. It is assumed that $p + q + r = n$. Without any loss of generality the unit interval for the time variable t is considered. Finally let the matrix $\varphi_{\dot{x}}$ be regular. The problem is to find the differentiable function $x(t)$, $0 \leqq t \leqq 1$, which satisfies (1)–(2), provided that such a function exists.

Following [1–3] let us define for any function $x(t)$, $0 \leqq t \leqq 1$, not necessarily satisfying (1)–(2), the performace index P as (arguments are omitted for the sake of simplicity):

$$P = \int_0^1 (\varphi^T \varphi)\, \mathrm{d}t + (f^T f + g^T g + h^T h)\,. \tag{3}$$

Here T indicates the transpose. This quadratic functional can be interpreted as the cumulative error in the implicit differential equation (1) and the boundary conditions (2). Clearly $P = 0$, iff $x(t)$, $0 \leqq t \leqq 1$, is the solution of (1)–(2); otherwise $P > 0$.

Taking now $x(t)$, $0 \leqq t \leqq 1$, as the nominal trajectory and perturbing it by $\Delta x(t)$, $0 \leqq t \leqq 1$, the new trajectory appears as

$$\tilde{x}(t) = x(t) + \Delta x(t)\,, \quad 0 \leqq t \leqq 1\,. \tag{4}$$

The result of this perturbation is that the performance index P changes to the first order

$$\delta P = 2 \int_0^1 \varphi^T \delta\varphi\, \mathrm{d}t + 2(f^T \delta f + g^T \delta g + h^T \delta h)\,. \tag{5}$$

As in [1–3] assume the linear variations

$$\delta\varphi = -\alpha\varphi\,, \tag{6}$$

$$\delta f = -\alpha f\,, \quad \delta g = -\alpha g\,, \quad \delta h = -\alpha h\,, \tag{7}$$

where α, $0 < \alpha \leqq 1$, is the stepsize to be determined later. Substitution of these terms in (6)–(7) yields the relation

$$\delta P = -2\alpha \int_0^1 \varphi^T \varphi\, \mathrm{d}t - 2\alpha(f^T f + g^T g + h^T h)\,, \tag{8}$$

which further implies that

$$\delta P = -2\alpha P < 0\,. \tag{9}$$

Hence, for α small enough, $0 \leqq P + \delta P < P$, and the descent property is established.

Expanding now (6)–(7) to the first-order terms one has

$$\delta\varphi = \varphi_{\dot{x}}\,\Delta\dot{x} + \varphi_x\,\Delta x\,, \quad 0 \leqq t \leqq 1\,, \tag{10}$$

and

$$\begin{gathered} \delta f = f_{x(0)}\,\Delta x(0)\,, \quad \delta g = g_{x(1)}\,\Delta x(1)\,, \\ \delta h = h_{x(0)}\,\Delta x(0) + h_{x(1)}\,\Delta x(1)\,. \end{gathered} \tag{11}$$

The dimensions of various matrices follow from the preceding considerations.

Substitution of (6)–(7) into (10)–(11), respectively, yields

$$\varphi_{\dot{x}}\,\Delta\dot{x} + \varphi_x\,\Delta x + \alpha\varphi = 0\,, \quad 0 \leqq t \leqq 1\,, \tag{12}$$

and

$$\begin{gathered} f_{x(0)}\,\Delta x(0) + \alpha f = 0\,, \quad g_{x(1)}\,\Delta x(1) + \alpha g = 0\,, \\ h_{x(0)}\,\Delta x(0) + h_{x(1)}\,\Delta x(1) + \alpha h = 0\,. \end{gathered} \tag{13}$$

Thus, for a given value of α, the linear implicit two-point boundary-value problem for the variable $\Delta x(t)$ is to be solved. Analogously as in [1–3], the resulting algorithm can be denoted as the modified quasilinearization algorithm. For $\alpha = 1$ we see that the classical Newton-Raphson method is obtained. Let us remark that in this case the descent property (9) need not necessarily hold.

As discussed in [4], the choice of the right-hand sides in (6)–(7) is crucial in the derivation of the modified quasilinearization method. This special choice causes, on the one hand, the quadratic properties of δP and, on the other hand, is responsible for the modified quasilinearization method being a Newton-like method. Any other choice would not lead to a Newton method.

The problem is simplified by introducing the auxiliary variable

$$y(t) = \Delta x(t)/\alpha\,, \quad 0 \leqq t \leqq 1\,. \tag{14}$$

Then (12)–(13) take the parameter-free form

$$\varphi_{\dot{x}}\dot{y} + \varphi_x y + \varphi = 0\,, \quad 0 \leqq t \leqq 1\,, \tag{15}$$

and

$$\begin{gathered} f_{x(0)}\,y(0) + f = 0\,, \quad g_{x(1)}\,y(1) + g = 0\,, \\ h_{x(0)}\,y(0) + h_{x(1)}\,y(1) + h = 0\,. \end{gathered} \tag{16}$$

This boundary problem can be solved without assigning a special value to α.

The assumption of the regularity of $\varphi_{\dot{x}}$ enables to resolve (15) as

$$\dot{y} + \varphi_{\dot{x}}^{-1}\varphi_x y + \varphi_{\dot{x}}^{-1}\varphi = 0\,, \quad 0 \leqq t \leqq 1\,, \tag{17}$$

and to use some of the known methods for its solution. For example, the method of particular solutions [1 – 3], the method of complementary functions [12], or the method of adjoints [12 – 13] can be applied. If the regularity of $\varphi_{\dot{x}}$ is not guaranteed, one can expect serious troubles connected with the nonuniqueness of the solution, branching of the solution, etc. Some hints in this respect are given for the static problem (nonlinear equations) in [11].

If the solution $y(t)$, $0 \leqq t \leqq 1$, of (16) – (17) is known, the varied solution is given as

$$\tilde{x}(t) = x(t) + \alpha\, y(t)\,, \quad 0 \leqq t \leqq 1\,. \tag{18}$$

For this one-parameter family of solutions the performance index P becomes a function of the stepsize α. Clearly

$$P_\alpha(0) = -2P(0)\,. \tag{19}$$

Hence, there exist a point α^* such that

$$P_\alpha(\alpha^*) = 0\,. \tag{20}$$

As the determination of the exact α^* might take excessive computer time, Miele et al. [1 – 3] suggest the noniterative procedure starting with $\alpha = 1$. The stepsize α is considered to be acceptable only if

$$P(\alpha) < P(0)\,. \tag{21}$$

Otherwise, the previous value of α must be reduced, e.g., invoking a bisection process, until (21) is met.

3. SUMMARY OF THE ALGORITHM

The described numerical procedure can be summarized as follows:

Step 1. Choose $\varepsilon > 0$ and the nominal function $x(t)$, $0 \leqq t \leqq 1$.

Step 2. Compute φ, $0 \leqq t \leqq 1$ and f, g, h. Evaluate P according to (3).

Step 3. If $P < \varepsilon$, then stop the computations; else go to Step 4.

Step 4. Compute $\varphi_{\dot{x}}$, φ_x, $0 \leqq t \leqq 1$, and $f_{x(0)}$, $g_{x(1)}$, $h_{x(0)}$, $h_{x(1)}$. Solve the linear two-point boundary-value problem (16) – (17).

Step 5. If Newton-Raphson method is requested, set $\alpha = 1$ and go to Step 7; else go to Step 6.

Step 6. Starting with $\alpha = 1$ perform a bisection process on α until (21) is met for the modified quasilinearization method.

Step 7. Set

$$x(t) \triangleq x(t) + \alpha\, y(t)\,, \quad 0 \leqq t \leqq 1\,, \tag{22}$$

and go to Step 2.

4. ILLUSTRATIVE EXAMPLES

The further presented examples were solved on an IBM 370/135 computer. The algorithm was programmed in PL/1 and similarly as in [1–3] double-precision arithmetic was used. As the stopping condition the value $\varepsilon = 10^{-20}$ was used. The same value was used also as a singularity level for the determinant of $\varphi_{\dot{x}}$. Finally, all variables through this section let be scalars.

As in [1–3] the bisection limit $N = 10$ is imposed to prevent extremely small changes.

The linear two-point boundary-value problem (16)–(17) was always solved using the method of adjoints, e.g., see [12–13]. Some examples are of rather simple structure and were constructed to admit analytical solutions. However, they are quite sufficient to illustrate the situations which can be encountered when solving general problems. Thus a principal comparison of the modified quasilinearization method (MQM) and the Newton-Raphson method (NMR) is established.

Example 1. Consider the differential equations

$$\exp(x_1 + \dot{x}_2) - \dot{x}_2^2 - x_2^2 = 0\,,$$
$$\dot{x}_1 - x_2 = 0\,,$$

subject to the boundary conditions

$$x_1(0) = 0\,, \quad x_1(1) = \sin 1\,.$$

The analytical solution

$$x_1(t) = \sin t\,, \quad x_2(t) = \cos t\,, \quad 0 \leqq t \leqq 1\,,$$

was obtained via MQM and NRM always with the same number of iterations (ranging from 5 to 35 iterations depending on the nominal estimate).

Example 2. Consider the differential equations

$$(2x_1\dot{x}_2 - x_2^2)^3 + 32\dot{x}_2(t\dot{x}_2 - x_2)^3 = 0\,,$$
$$\dot{x}_1 - x_2 = 0\,,$$

subject to the boundary conditions

$$x_1(0) = 3\,, \quad x_1(1) = 6\,.$$

The analytical solution

$$x_1(t) = (t + 1)^2 + 2\,, \quad x_2(t) = 2t + 2\,, \quad 0 \leqq t \leqq 1$$

was reached by the NRM in 11 iterations starting with

$$x_1(t) = 3t + 3\,, \quad x_2(t) = 0\,, \quad 0 \leqq t \leqq 1\,.$$

However, the MQM was not successful due to the excessive number of bisections.

Example 3. Consider the differential equations

$$\sin(x_1 + \dot{x}_2) - \dot{x}_2^2 - x_2^2 = 0\,,$$
$$\dot{x}_1 - x_2 = 0\,,$$

subject to the boundary conditions

$$x_1(0) = \pi/2\,, \quad x_1(1) = \pi/2 - \sin 1\,.$$

Also now the analytical solution

$$x_1(t) = \pi/2 - \sin t\,, \quad x_2(t) = -\cos t\,, \quad 0 \leqq t \leqq 1\,,$$

is known. This problem was solved for various initial solution estimates. For example, starting with

$$x_1(t) = \pi/2\,, \quad x_2(t) = t - 1\,, \quad 0 \leqq t \leqq 1\,,$$

the MQM was inefficient, while NRM has converged in 25 iterations. Analogously, starting with

$$x_1(t) = 1\,, \quad x_2(t) = t - 1\,, \quad 0 \leqq t \leqq 1\,,$$

the MQM was again not successful, while NRM reached the exact solution in 16 iterations.

Example 4. Consider the differential equations

$$\dot{x}_1 - 10x_2 = 0\,,$$
$$\dot{x}_2 - 10x_3 = 0\,,$$
$$(\sin^2 \dot{x}_3 + k\cos^2 \dot{x}_3)\,\dot{x}_3 + 5x_1x_2 = 0,$$

subject to the boundary conditions

$$x_1(0) = 0\,, \quad x_2(0) = 0\,, \quad x_2(1) = 1\,.$$

This problem was solved for various values of the parameter k causing the problem more or less "implicit". The initial solution estimate

$$x_1(t) = 0\,, \quad x_2(t) = t\,, \quad x_3(t) = 0\,, \quad 0 \leqq t \leqq 1\,,$$

was always applied.

a) If $k = 0$, the problem was solved by both methods in 2 iterations.

b) If $k = 0{\cdot}5$, the NMR diverged and in the MQM the bisection limit was reached, i.e., both methods were not successful. However, using "better" solution estimate

$$x_1(t) = 8t\,, \quad x_2(t) = t\,, \quad x_3(t) = 0\,, \quad 0 \leqq t \leqq 1\,, \tag{23}$$

the NRM again diverged, while MQM converged in 16 iterations with overall 31 bisections in various iterations.

c) If $k = 0{\cdot}8$, the NRM again diverged and MQM converged in 7 iterations with 4 bisections in the second iteration.

d) Also if $k = 0{\cdot}9$, the NMR has diverged, while using the MQM the solution is reached in 6 iterations with 3 bisections in the second iteration. Applying the initial solution estimate (23) the behaviour of both methods was identical and they converged in 6 iterations.

e) Finally if $k = 1$, the NRM diverged and MQM converged in 7 iterations with 3 bisections in the second iteration. See also [2, Example 3·5] in this respect.

5. CONCLUSIONS

The possibility of the numerical solution of implicit two-point boundary-value problems using the so-called modified quasilinearization method was explored in detail. Moreover a comparison of this method with the classical Newton-Raphson method was performed. This comparison has shown that generally the following four classes of problems are encountered:

I. Problems cannot be solved by any of these methods.

II. Problems can be solved by the NRM, but cannot be solved using MQM, resp. MQM requires higher number of iterations.

III. Problems solvable by the both methods, the behaviour of which is identical, i.e., $\alpha = 1$ during each iteration of MQM.

IV. Problems can be solved using MQM, but cannot be solved using NRM, resp. NRM requires higher number of iterations.

To be quite fair it would be also necessary to take into the account the total computer time required. It is namely clear that one iteration of MQM lasts longer due to the possible search for α. However, this circumstance was neglected when formulating this to the certain extent logical classification.

Especially Example 4 illustrates the various possibilities, which were not pointed out in the original papers [1 – 3]. Moreover, one can further see that such classification of a particular problem can depend also on the initial solution estimate. Further details concerning the computer implementation of both methods are to be found in [8].

(Received October 9, 1978.)

REFERENCES

[1] A. Miele, R. R. Iyer: General technique for solving nonlinear, two-point boundary-value problems via the method of particular solutions. J. Optimization Theory Appl. *5* (1970), 5, 382—399.

[2] A. Miele, R. R. Iyer: Modified quasilinearization method for solving nonlinear, two-point boundary-value problems. J. Math. Anal. Appl. *36* (1971), 3, 674—692.

[3] A. Miele, S. Naqui, A. V. Levy, R. R. Iyer: Numerical solution of nonlinear equations and nonlinear, two-point boundary-value problems. In "Advances in Control Systems: Theory and Applications", Vol. 8, C. T. Leondes (ed.), Academic Press, New York 1971, 189—215.

[4] S. M. Roberts, J. S. Shipman: On the Miele-Iyer modified quasilinearization method. J. Optimization Theory Appl. *14* (1974), 4, 381—391.

[5] J. Fidler: The application of the modified quasilinearization method for the solution of continuous time boundary-value problems. Research Report No. 819. Institute of Information Theory and Automation, Prague 1977. In Czech.

[6] J. Doležal: On the modified quasilinearization method for discrete two-point boundary-value problems. Research Report No. 788, Institute of Information Theory and Automation, Prague 1977.

[7] J. Doležal: On a certain type of discrete two-point boundary-value problems arising in discrete optimal control. EQUADIFF 4 Conference, Prague, August 22—26, 1977. See also: Kybernetika *15* (1979), 3, 215—221.

[8] J. Doležal, J. Fidler: To the problem of numerical solution of implicit two-point boundary-value problems. Research Report No. 857, Institute of Information Theory and Automation, Prague 1978. In Czech.

[9] J. Doležal: Modified quasilinearization method for the solution of implicit, nonlinear, two-point boundary-value problems for systems of difference equations. The 5th Symposium on Algorithms ALGORITMY' 79, High Tatras, April 23—27, 1979. In Czech.

[10] M. R. Hestenes: Calculus of Variations and Optimal Control Theory, Wiley, New York 1966.

[11] D. G. B. Edelen: Differential procedures for systems of implicit relations and implicitly coupled nonlinear boundary-value problems. In "Numerical Methods for Differential Systems: Recent Development in Algorithm, Software, and Applications", L. Lapidus, W. E. Schiesser (eds.), Academic Press, New York 1976, 85—95. See also: In "Mathematical Models and Numerical Methods", Banach Center Publications Vol. 3, A. N. Tichonov et al. (eds.), PWN-Polish Scientific Publishers, Warszawa 1978, 289—296.

[12] S. M. Roberts, J. S. Shipman: Two-Point Boundary Value Problems: Shooting Methods. American Elsevier, New York 1972.

[13] E. Polak: Computational Methods in Optimization: Unified Approach. Academic Press, New York 1971.

Ing. Jaroslav Doležal, CSc., Jiří Fidler, prom. mat., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.