Numer. Math. 22, 289—315 (1974)

A Modified Newton Method for the Solution of Ill-Conditioned Systems of Nonlinear Equations with Application to Multiple Shooting

P. Deuflhard

Abstract. In this paper the well-known modified (underrelaxed, damped) Newton method is extended in such a way as to apply to the solution of ill-conditioned systems of nonlinear equations, i.e. systems having a "nearly singular" Jacobian at some iterate. A special technique also derived herein may be useful, if only bad initial guesses of the solution point are available. Difficulties that arose previously in the numerical solution of nonlinear two-point boundary value problems by multiple shooting techniques can be removed by means of the results presented below.

1. Introduction

In recent years considerable progress has been made in the numerical solution of system of nonlinear equations. The techniques that are in common use may, in principle, be derived from the steepest descent method (Cauchy [9], Rosenbloom [30]) and the Newton method (Kantorovič [18]). Among the most popular techniques are the Levenberg-Marquardt method (Levenberg [21], Marquardt [22], Bard [1]), Powell's new hybrid method (Powell [28], Broyden [5]), and the modified (underrelaxed, damped) Newton method (Goldstein [14], Stoer [31]). In [32] the latter method was selected to solve boundary value problems by multiple shooting techniques. (For reference see Keller [19], Osborne [25], Bulirsch [6], and Stoer/Bulirsch [32]. The notations used in this paper are close to [6] and [32].)

It may be recalled that in the *multiple shooting method* the two-point boundary value problem

$$\begin{aligned} &y' = f(x, y) \qquad x \in [a, b], \qquad y: [a, b] \Rightarrow \mathbb{R}^n \\ &r\big(y(a), y(b)\big) = 0 \end{aligned} \tag{1.1}$$

is replaced by $(m-1)$ initial value problems

$$\left.\begin{aligned} &y' = f(x, y) \qquad x \in [x_j, x_{j+1}] \\ &y(x_j; x_j, s_j) = s_j \end{aligned}\right\} \; j = 1, \ldots, m-1 \tag{1.2}$$

with a suitably chosen subdivision $a = x_1 < \cdots < x_m = b$.

The $n(m-1)$-vector $s^T = (s_1^T, \ldots, s_{m-1}^T)$ has to be selected so that the following conditions hold:

a) continuity conditions (for $m > 2$)

$$F_j(s_j, s_{j+1}) := y(x_{j+1}; x_j, s_j) - s_{j+1} = 0, \qquad j = 1, \ldots, m-2, \tag{1.3}$$

b) boundary conditions

$$F_{m-1}(s_1, s_{m-1}) := r\big(s_1, y(x_m; x_{m-1}, s_{m-1})\big) = 0.$$

Let $\Delta s^{(k)}$ denote the correction given by the *ordinary* Newton method at the iteration point $s^{(k)}$. Then by the *modified* Newton method

$$s^{(k+1)} = s^{(k)} + \lambda_k \Delta s^{(k)}, \quad 0 < \lambda_k \leqq 1. \tag{1.4}$$

The scalar sequence $\{\lambda_k\}$ is usually chosen in such a way as to ensure that

$$T(s^{(k+1)}) < T(s^{(k)}) \tag{1.5}$$

where

$$T(s) := \sum_{j=1}^{m-1} \|F_j(s)\|^2 = \sum_{j=1}^{m-1} F_j^T F_j.$$

($\|\cdot\|$ Euclidean vector or spectral matrix norm, respectively, throughout the paper.) As the iterates approach the solution point, the λ_k should approach 1, in order to take account of the quadratic convergence of the ordinary Newton method.

In some real life applications, however, the iteration terminates without convergence, if the Jacobian is singular at some iteration point. Even if the Jacobian is nonsingular in the strict mathematical sense, rounding and discretization errors can heavily disturb the actual computation: extremely small values of λ_k (say $\lambda_k \approx 1/1000$) are required (sometimes even negative). As a consequence, computing time may increase beyond a tolerable measure, or the algorithm may fail to converge. The purpose of this paper is to present a modification of Newton's method designed to overcome the difficulties mentioned.

For the case of singular Jacobian, some authors recommend the alternative use of steepest descent corrections as they do not involve the Jacobian inverse. That is why in § 2 Newton's method is compared with a class of steepest descent methods. This leads, at first, to a replacement of the usual monotonicity test (1.5) and, secondly, to a modified Newton method, which is in some respect similar to the least-squares method suggested by Ben-Israel [3] and extended by Fletcher [13]. In § 3, the method as proposed herein is backed by a constructive proof of global convergence which implies convergence results for the modified Newton method as known so far. Details of the numerical realization are worked out in § 4. *The results of* § 2–4 *apply to general systems of nonlinear equations.* In § 5 the method is specified taking account of the sparse structure of the Jacobian arising in multiple shooting. Numerical examples are presented in § 6.

2. A Steepest Descent Modification of Newton's Method

In this paragraph a modification of Newton's method for both nonsingular and singular Jacobian is derived from a comparison with a class of steepest descent methods.

Let

$$F(x) := \begin{bmatrix} f_1(\xi_1, \ldots, \xi_n) \\ \vdots \\ f_n(\xi_1, \ldots, \xi_n) \end{bmatrix}, \quad F(x^*) = 0$$

denote the system of nonlinear equations to be solved and let $J(x)$ denote the Jacobian (n, n)-matrix of the system.

Let a *level function* be defined by

$$T(x\,|\,A) := \tfrac{1}{2}\|AF(x)\|^2 = \tfrac{1}{2}\big(AF(x)\big)^T\big(AF(x)\big) \tag{2.1}$$

where A is an arbitrary constant nonsingular (n, n)-matrix. This type of function is useful since

$$\begin{aligned} &T(x\,|\,A)=0 \;\Leftrightarrow\; F(x)=0 \;\Leftrightarrow\; x=x^*\\ &T(x\,|\,A)>0 \;\Leftrightarrow\; F(x)\neq 0 \;\Leftrightarrow\; x\neq x^*. \end{aligned} \tag{2.2}$$

The following lemma is elementary.

Lemma 2.1. *Let $T(x|A)$ be defined for $x\in D\subset \mathbb{R}^n$. Let $p_k\neq 0$ denote a search direction in $x_k\in D$ and* $\operatorname{grad} T(x_k\,|\,A)$ *the gradient direction of $T(x\,|\,A)$ in x_k. Then if*

$$p_k^T \operatorname{grad} T(x_k\,|\,A) < 0$$

there exists a $\mu_k > 0$ such that

$$T(x_k+\lambda p_k\,|\,A) < T(x_k\,|\,A) \quad \text{for} \quad 0<\lambda\leqq\mu_k.$$

This inspires the following

Definition. *Let Δx_k denote the correction vector given in x_k by some iterative method for solving systems of nonlinear equations. Then a level function $T(x\,|\,A)$ is said to be "appropriate" for the iterative method in question, if and only if*

$$\Delta x_k^T \operatorname{grad} T(x_k\,|\,A) < 0 \quad \text{for all} \quad x_k\in D \quad \text{with} \quad \Delta x_k \neq 0. \tag{2.3}$$

a) Case of Nonsingular Jacobian. Assume $J(x_k)$ to be nonsingular for the time being. Then the *Newton correction vector* Δx_k is

$$\Delta x_k = -J(x_k)^{-1}F(x_k). \tag{2.4}$$

One obtains

$$\Delta x_k^T \operatorname{grad} T(x_k\,|\,A) = -2\,T(x_k\,|\,A) < 0 \quad \text{for all} \quad x_k\in D \quad \text{with} \quad \Delta x_k \neq 0. \tag{2.5}$$

Hence, according to Lemma 2.1 *each* level function $T(x\,|\,A)$ may be used to determine the sequence $\{\lambda_k\}$ by means of the *monotonicity test*

$$T(x_{k+1}\,|\,A) < T(x_k\,|\,A) \quad \text{with} \quad x_{k+1} := x_k + \lambda_k \Delta x_k. \tag{2.6}$$

Of course, the sequence $\{\lambda_k\}$—and the rate of convergence of the modified Newton method—depends on the choice of the matrix A. The question of how to choose A is resolved by means of the following

Theorem 2.2. *Let Δx_k denote the direction of steepest descent of $T(x|A)$ in x_k:*

$$\Delta x_k = -\operatorname{grad} T(x_k\,|\,A) = -\big(AJ(x_k)\big)^T AF(x_k). \tag{2.7}$$

Then there are level functions $T(x\,|\,B)$ satisfying $\Delta x_k^T \operatorname{grad} T(x_k\,|\,B) > 0$, unless

$$J(x_k)J(x_k)^T A^T AF(x_k) = \chi F(x_k) \quad \text{for some} \quad \chi\in\mathbb{R}^1.$$

Proof. Let $A = I$ be assumed (usual case). Write $\bar{A} := J(x_k) J(x_k)^T$, $\bar{B} := B^T B$, $F_k := F(x_k)$. Then $\Delta x_k^T \operatorname{grad} T(x_k | B) = -F_k^T \bar{A} \bar{B} F_k$. Now, choose $\bar{B} := \bar{A} + \mu y y^T$ with $\mu > 0$ and $F_k^T (\bar{A} + I) y = 0$, but $F_k^T y \neq 0$ (here the assumption $\bar{A} F_k \neq \chi F_k$ is required). For $\mu > \|\bar{A} F_k\|^2 / (F_k^T y)^2$ one obtains

$$\Delta x_k^T \operatorname{grad} T(x_k | B) = -\|\bar{A} F_k\|^2 + \mu (F_k^T y)^2 > 0.$$

The proof for the general case $A \neq I$ is omitted. ◊

The interesting part of this result is the special case

$$A = J(x_k)^{-1} \quad \text{in } x_k. \tag{2.8}$$

Upon substitution of (2.8) into (2.7) the *Newton correction* is obtained as *a particular steepest descent correction*:

$$\Delta x_k = -\operatorname{grad} T(x_k | J(x_k)^{-1}) = -J(x_k)^{-1} F(x_k). \tag{2.9}$$

Hence, (2.8) seems to point out a *natural scaling* of $T(x|A)$ in the monotonicity test (2.6). Moreover, it may be shown that in a neighborhood of x^* the *level surfaces*

$$T(x_k | A) = \text{const}$$

are *spheres* for this choice of A, while in general they are ellipsoidals. This follows from (compare e.g. J. Kowalik/M. R. Osborne [20])

$$T(x_k | A) = \tfrac{1}{2} \|A J(x^*)(x_k - x^*)\|^2 + O(\|x_k - x^*\|^3). \tag{2.10}$$

It may be recalled that for an exact sphere only one steepest descent correction is needed to find x^* —independent of the starting point x_0 (compare Greenstadt [16] for linear $F(x)$).

b) Case of Singular Jacobian. Without loss of generality the following special case is assumed (for a justification see § 4).

$$J(x_k) = \left[\begin{array}{c|c} J & j \\ \hline 0 & 0 \end{array}\right], \qquad s := \operatorname{rank}(J_k) = \operatorname{rank}(J), \tag{2.11}$$

where J is a nonsingular upper triangular (s, s)-matrix. Let P and $\bar{P}$ denote two projection matrices

$$P = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & 0 & \\ & & 1 & & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{bmatrix}, \qquad \bar{P} = I - P \text{ with } \operatorname{rank}(P) := s.$$

Remark. A sufficient condition for $\bar{P} J(x_k) = 0$ is

$$\bar{P} F(x) = \text{const}. \tag{2.12}$$

Upon partitioning the linear n-system $J(x_k)\Delta x_k = -F(x_k)$ one obtains

$$J\bar{P}\Delta x_k + j\bar{P}\Delta x_k = -PF(x_k) \tag{I}$$

$$0 = -\bar{P}F(x_k) \tag{II}$$

(II) is omitted (contradiction). (I) defines s necessary conditions for a modified Newton method

$$\Delta x_k \in N_k, \quad \textit{(Newton condition)} \tag{2.13}$$

where

$$N_k := \{\Delta x \in \mathbb{R}^n \mid JP\Delta x + j\bar{P}\Delta x + PF(x_k) = 0\}.$$

$(n-s)$ further conditions are obtained from the following

Lemma 2.3. *Let $\bar{P}J(x_k) = 0$ according to* (2.11). *Let Δx_k denote the direction of steepest descent of any level function $T(x \mid A)$* (*see* (2.7)). *Then, independent of A, the components of Δx_k satisfy*

$$\bar{P}\Delta x_k = \bar{j}^T P\Delta x_k \quad \textit{(steepest descent condition)} \tag{2.14}$$

where

$$\bar{j} := J^{-1}j.$$

Proof. Let $\bar{A} := A^T A$ be partitioned with respect to P and $\bar{P}$.

$$\bar{A} =: \left[\begin{array}{c|c} \bar{A}_0 & \bar{A}_1 \\ \hline \bar{A}_1^T & \bar{A}_2 \end{array}\right].$$

Upon substituting $\bar{A}$ into (2.7) it is obtained

$$P\Delta x_k = -J^T(\bar{A}_0 PF_k + \bar{A}_1 \bar{P}F_k), \quad \bar{P}\Delta x_k = -j^T(\bar{A}_0 PF_k + \bar{A}_1 \bar{P}F_k)$$

Since J is nonsingular, the result follows independent of A. ◊

Remark. (2.14) holds for the Levenberg-Marquardt method as well.

Theorem 2.4. *Let Δx_k denote a correction vector subject to both the Newton condition* (2.13) *and the steepest descent condition* (2.14). *Then Δx_k may be computed in the following way*

$$\begin{aligned} P\Delta y_k &:= -J^{-1}PF(x_k), \quad & \bar{P}\Delta y_k &:= \bar{j}^T P\Delta y_k \\ \bar{P}\Delta x_k &= (I_{n-s} + \bar{j}^T\bar{j})^{-1}\bar{P}\Delta y_k, \quad & P\Delta x_k &= P\Delta y_k - \bar{j}\,\bar{P}\Delta x_k \end{aligned} \tag{2.15}$$

Proof. Substitution of (2.14) into (2.13) yields

$$(J + j\bar{j}^T)P\Delta x_k = -PF_k,$$

where $(J + j\bar{j}^T) = J(I_s + \bar{j}\bar{j}^T)$ is nonsingular, as J is nonsingular from (2.11) and $(I_s + \bar{j}\bar{j}^T)$ is positive definite. Thus $P\Delta x_k$ can be determined. From (2.14), $\bar{P}\Delta x_k$ can be determined.

From a well-known theorem (see Householder [17]) one obtains

$$(I_s + \bar{j}\bar{j}^T)^{-1} = I_s - \bar{j}\,M^{-1}\bar{j}^T,$$

where $M := (I_{n-s} + \bar{J}^T \bar{J})$ is positive definite. With the auxiliary variable Δy_k as denoted above it is obtained

$$P\Delta x_k = (I + \bar{J}\bar{J}^T)^{-1} P\Delta y_k = P\Delta y_k - \bar{J} M^{-1} \bar{J}^T P\Delta y_k,$$
$$\bar{P}\Delta x_k = \bar{J}^T P\Delta x_k = M^{-1}\bar{J}^T P\Delta y_k \Rightarrow P\Delta x_k = P\Delta y_k - \bar{J}\bar{P}\Delta x_k. \quad \diamond$$

It is shown in [10] that (2.15) is a computationally convenient representation of

$$\Delta x_k = -J(x_k)^\dagger F(x_k), \tag{2.16}$$

where $A^\dagger$ denotes the *Penrose pseudoinverse* [26] of a real matrix A satisfying

$$(A^\dagger A)^T = A^\dagger A, \quad (AA^\dagger)^T = AA^\dagger, \quad A^\dagger AA^\dagger = A^\dagger, \quad AA^\dagger A = A \tag{2.17}$$

Thus, a method using the correction Δx_k from (2.15) with $\lambda_k = 1$, is equivalent to the Newton-Raphson method proposed by Ben-Israel [3]. In order to expand the domain of convergence of the method, Fletcher [13] proposed a modification for $0 < \lambda_k \leqq 1$ with $T(x|I)$ in the monotonicity test (2.6). He already gave a proof that $T(x|I)$ is an "*appropriate*" level function for (2.16). The whole class of "*appropriate*" level functions $T(x|A)$ is described by the following

Lemma 2.5. *Let* $\bar{P}J(x_k) = 0$, $\Delta x_k = -J(x_k)^\dagger F(x_k)$. *Then* $T(x|A)$ *is an "appropriate" level function, if and only if*

$$PA^T A\bar{P} = 0. \tag{2.18}$$

Proof. Let $\bar{A} := A^T A$. Then

$$\Delta x_k^T \operatorname{grad} T(x_k|A) = -(PF_k)^T \bar{A} PF_k - (PF_k)^T (P\bar{A}\bar{P})(\bar{P}F_k).$$

The first right-hand term is negative definite, if and only if $PF_k \neq 0$. $PF_k = 0$ is equivalent to $\Delta x_k = 0$. The second right-hand term may be positive. Therefore $P\bar{A}P = 0$ is necessary and sufficient that

$$\Delta x_k^T \operatorname{grad} T(x_k|A) < 0 \quad \text{for } \textit{all} \quad \Delta x_k \neq 0. \quad \diamond$$

For example, condition (2.18) holds for A orthogonal or diagonal (gauging of variables). The choice of A to be proposed herein may be motivated in the same way as for the case of nonsingular Jacobian. Upon employing (2.17) twice the analogue of (2.9) is obtained

$$\Delta x_k = -\operatorname{grad} T(x_k|J(x_k)^\dagger) = -J(x_k)^\dagger F(x_k). \tag{2.19}$$

Hence, for the case of singular Jacobian *natural scaling* of the level function $T(x|A)$ means

$$A = J(x_k)^\dagger \quad \text{in } x_k. \tag{2.20}$$

Obviously, condition (2.18) holds for (2.20). However, property (2.2) is modified as follows

$$\begin{aligned} &T(x_k|J(x_k)^\dagger) > 0 \Leftrightarrow PF(x_k) \neq 0 \Leftrightarrow \Delta x_k \neq 0 \\ &T(x_k|J(x_k)^\dagger) = 0 \Leftrightarrow PF(x_k) = 0 \Leftrightarrow \Delta x_k = 0. \\ &\text{If furthermore } \bar{P}F(x_k) = 0, \text{ then } x_k = x^*. \end{aligned} \tag{2.21}$$

A final comparison of the type of method presented here and the class of steepest descent methods may be made by the following

Lemma 2.6. *Let B, K be nonsingular (n, n)-matrices. Let two linear transformations be described by*

$$\text{(I)}\ F(x) \to G(x) = BF(x), \qquad \text{(II)}\ x \to y = Kx$$

a) *Let Δx_k denote the steepest descent correction (2.7) for $A = I$*

$$\Delta x_k = -J(x_k)^T F(x_k)$$

Then Δx_k is invariant to the transformations (I) *and* (II), *if and only if*

$$B^T B = K^T K = I. \tag{2.22}$$

b) *Let Δx_k denote the Newton-Raphson correction*

$$\Delta x_k = -J(x_k)^\dagger F(x_k)$$

where $\overline{P} J(x_k) = 0$. Then Δx_k is invariant to the transformations (I) *and* (II), *if and only if*

$$PB^T B\overline{P} = PK^T K\overline{P} = 0. \tag{2.23}$$

Remark. Condition (2.22) also holds for the Levenberg-Marquardt correction. Obviously (2.22) is contained in (2.23).

For the sake of completeness it may be noted that the results given here may easily be extended to apply to *least squares problems* as well (see Example 1 of § 6).

3. Convergence Results

It may be recalled from [3] that a method using the corrections (2.16) (with $\lambda_k = 1$) converges *locally* to a point x^* subject to

$$J(x^*)^T F(x^*) = 0. \tag{3.1}$$

That is, x^* is either a *solution point* (with $F(x^*) = 0$) or a *stationary point* (with $\overline{P} J(x^*) = 0$ and $PF(x^*) = 0$). In the latter case no information about $\overline{P} F(x^*)$ is given except the well-known least squares property

$$\|\overline{P} F(x^*)\|_2 = \min_{Z \in \mathbb{R}^n} \|J(x^*)Z + F(x^*)\|_2. \tag{3.2}$$

Hence, a reasonable assumption to obtain convergence to some *solution point* x^* is that the Jacobian be nonsingular in some domain D. Then, *local quadratic* convergence of the ordinary Newton method is guaranteed under the assumptions of the well-known Newton-Kantorovič theorem [18]. For the modified Newton method Stoer [31] gave a proof of *global* convergence showing the existence of some sequence $\{\lambda_k\}$ such that $\{T(x_k | I)\}$ is a monotonely decreasing sequence. The proof in [31] can easily be extended to apply to any $T(x | A)$, if A is contained in some matrix set

$$A_M := \{A \mid \operatorname{cond}_2(A) \leqq M < \infty\}. \tag{3.3}$$

Here A is assumed to be a *constant* matrix, independent of the iteration step k, in contrast with *natural scaling* as suggested in (2.8). The purpose of this paragraph is to give a theorem of global convergence of the following modified Newton method

$$\begin{aligned} &\Delta x_k := -J(x_k)^{-1}F(x_k) \\ &x_{k+1} := x_k + \lambda_k \Delta x_k \quad \text{with} \quad \lambda_k \in]0,1] \quad \text{such that} \\ &T\left(x_{k+1} \mid J(x_k)^{-1}\right) < T\left(x_k \mid J(x_k)^{-1}\right). \end{aligned} \tag{3.4}$$

Upon studying the monotonicity test (2.6) for arbitrary level function $T(x \mid A)$ a certain *extremal property* is shown to hold for *natural scaling* (2.8). Moreover, a *rank strategy* to be proposed in § 4 is motivated by the subsequent results.

Following Ortega/Rheinboldt [24] a *level set* of $T(x \mid A)$ is defined by

$$G_k(A) := \{x \in \mathbb{R}^n \mid T(x \mid A) \leqq T(x_k \mid A)\}. \tag{3.5}$$

In order to save space, standard assumptions are arranged once:

(3.6) **Standard Assumptions.**

I. *Let $F(x) \in C^1(D)$ with $G_0(A) \subset D \subset \mathbb{R}^n$ for some $A \in A_{3f}$. Let $J(x)$ denote the Jacobian matrix.*

II. *Let exist $J(x)^{-1}$ with $\|J(x)^{-1}\| \leqq \beta$ for all $x \in D$.*

III. $\|J(x) - J(y)\| \leqq \gamma \|x - y\|$ *for* $x, y \in D$.

IV. *Let the path-connected component of x_0 in $G_0(A)$ (in the sequel simply denoted as $G_0(A)$) be compact.*

It is well-known (see e.g. Rheinboldt [29]) that under the assumptions I, II, and IV there exists one and only one solution point x^* in (the path-connected component of x_0 in) $G_0(A)$.

The following notations will be frequently used

$$J_k := J(x_k), \quad F_k := F(x_k), \quad \alpha_k := \|\Delta x_k\|, \quad \beta_k := \|J_k^{-1}\|, \quad h_k := \beta^2 \gamma \|F_k\|. \tag{3.7}$$

A simple derivation of the modified Newton method is obtained from the following result.

Lemma 3.1. *Let (3.6) be assumed and let $x_k \in G_0(A)$. Let $GL(n)$ denote as usual the class of nonsingular (n, n)-matrices and define*

$$G_k := \bigcap_{A \in GL(n)} G_k(A). \tag{3.8}$$

Then

$$G_k = \{x \in \mathbb{R}^n \mid x = \bar{x}(\lambda),\ \lambda \in [0, 2]\}$$

where the path $\bar{x}(\lambda)$ (also called Newton path due to [23]) is defined by

$$\begin{aligned} &\text{a)} \quad F(\bar{x}) = (1 - \lambda) F_k, \quad \text{or by} \\ &\text{b)} \quad \frac{d\bar{x}}{d\lambda} = -J(\bar{x})^{-1} F_k, \qquad \bar{x}(0) = x_k \end{aligned} \tag{3.9}$$

Proof. Let $H_k(A)$ denote the set

$$H_k(A) = \{y \in \mathbb{R}^n \mid \|Ay\| \leqq \|AF_k\|\}$$

and let

$$H_k := \bigcap_{A \in GL(n)} H_k(A).$$

Select a subset $\hat{A}$ of $GL(n)$

$$\hat{A} := \{A \mid A^T A = \sum_i \varkappa_i q_i q_i^T \text{ with } q_1 := F_k / \|F_k\|\},$$

($\varkappa_i > 0$ eigenvalues, q_i eigenvectors of $A^T A$).

Let y have the components η_i, $i = 1, \ldots, n$, with respect to the basis $\{q_i\}$:

$$y = \sum_i \eta_i q_i.$$

Then

$$H_k(A) = \{y \in \mathbb{R}^n \mid \sum_i \eta_i^2 \varkappa_i \leqq \|F_k\|^2 \varkappa_1\} \quad \text{for} \quad A \hat{\in} A.$$

Now let $\hat{H}_k := \bigcap_{A \in \hat{A}} H_k(A)$. Then

$$\hat{H}_k = \{y = \eta_1 q_1 \mid \eta_1^2 \leqq \|F_k\|^2\} = \{y \in \mathbb{R}^n \mid y = (1-\lambda) F_k,\ \lambda \in [0, 2]\}.$$

It is easy to show that $H_k = \hat{H}_k$. Now set $y = F(x)$ and lift the path H_k to G_k (implicit function theorem). This proves statement *a*. The proof of statement *b* may be found in [23].

Remark. The interesting part of the path $\bar{x}(\lambda)$ is for $\lambda \in [0, 1]$, since $x^* = \bar{x}(1)$. It was shown in [29] that if $J(x)$ is *singular* for some $\bar{x}(\lambda^*)$ with $0 < \lambda^* < 1$, then G_k terminates at $\bar{x}(\lambda^*)$.

Along $\bar{x}(\lambda)$ the following property is common for all level functions:

$$T(\bar{x}(\lambda) \mid A) = (1-\lambda)^2 T(x_k \mid A). \tag{3.10}$$

Euler-discretization of the differential equation (3.9.b) yields (see e.g. Meyer [23]):

$$\frac{x_{k+1} - x_k}{\lambda_k} = -J_k^{-1} F_k, \qquad \lambda_k \in]0, 1]. \tag{3.11}$$

With these preparations the monotonicity test (2.6) can be studied for arbitrary level function $T(x \mid A)$.

Theorem 3.2. *Let* (3.6) *be assumed. Let* $x_k \in G_0(A)$ *with* $T(x_k \mid A) > 0$ *(without loss of generality).* Δx_k *denotes the Newton correction. Then*

$$T(x_k + \lambda \Delta x_k \mid A) \leqq t_k(\lambda \mid A)\, T(x_k \mid A) \quad \text{for} \quad \lambda \in [0, \bar{\mu}_k(A)]$$

where

$$t_k(\lambda \mid A) := [1 - \lambda + \lambda^2 h_k (1 + h_k) \text{ cond } (A J_k)]^2 \leqq 1$$

and

$$\bar{\mu}_k(A) := \min\{1,\ 1/[h_k(1+h_k) \text{ cond } (A J_k)]\}.$$

Proof. For convenience it is useful to introduce

$$D_k := \{x = x(\lambda, \delta) \mid x(\lambda, \delta) = \bar{x}(\lambda) + \delta(x_k + \lambda \Delta x_k - \bar{x}(\lambda)),\ (\lambda, \delta) \in [0, \bar{\mu}_k(A)] \times [0, \bar{\delta}(\lambda)]\}$$

(see Fig. 1).

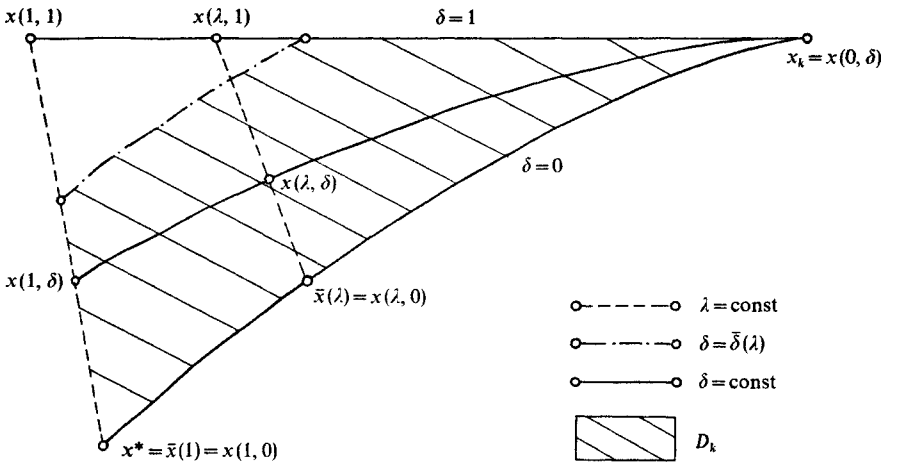


Fig. 1

Since $G_k \subset G_k(A)$, there exists a $\bar{\delta}(\lambda) \in]0, 1]$ for $\lambda \in [0, 1]$ such that $D_k \subset G_k(A)$. Moreover, $x_k \in G_0(A)$ implies $D_k \subset G_0(A)$.

1) The estimates to follow are valid for all $x(\lambda, \delta) \in D_k$. With the notations

$$\bar{x} := \bar{x}(\lambda) \quad \text{(for fixed } \lambda\text{) and} \quad \bar{J} := \frac{1}{\delta} \int_0^{\delta} J\big(x(\lambda, \delta')\big) d\delta'$$

one obtains

$$F\big(x(\lambda, \delta)\big) = F(\bar{x}) + \bar{J}\big(x(\lambda, \delta) - \bar{x}\big) \quad \text{(mean value theorem)}.$$

Employing the Cauchy-Schwarz inequality yields

$$T\big(x(\lambda, \delta) \,|\, A\big) = \tfrac{1}{2}\|AF\big(x(\lambda, \delta)\big)\|^2 \leqq \tfrac{1}{2}[\|AF(\bar{x})\| + \|A\bar{J}\big(x(\lambda, \delta) - \bar{x}\big)\|]^2.$$

A straightforward estimation using Lemma 3.1 gives

$$\|A\bar{J}\big(x(\lambda, \delta) - \bar{x}\big)\| \leqq \|AJ_k\| \, \|J_k^{-1}\bar{J}\| \, \|x(\lambda, \delta) - \bar{x}\|$$

$$\|x(\lambda, \delta) - \bar{x}\| \leqq \delta \lambda^2 \alpha_k h_k, \|J_k^{-1}\bar{J}\| \leqq 1 + h_k, \qquad \alpha_k \leqq \|J_k^{-1}A^{-1}\| \, \|AF_k\|.$$

Upon using (3.10) for $T(\bar{x} \,|\, A) \equiv \tfrac{1}{2}\|AF(\bar{x})\|^2$ it is obtained

$$\frac{\|AF(x(\lambda, \delta))\|}{\|AF_k\|} \leqq [1 - \lambda + \delta\lambda^2 h_k(1 + h_k) \operatorname{cond}(AJ_k)] =: t(\lambda, \delta). \tag{3.12}$$

2) Until now, no information about $\bar{\delta}(\lambda)$ was given. An elementary calculation shows that $\bar{\mu}_k(A)$ was selected in such a way as to ensure that

$$t(\lambda, \delta) < 1 \quad \text{for} \quad 0 < \lambda < \bar{\mu}_k(A), \quad 0 \leqq \delta \leqq 1, \quad t(0, \delta) = t\big(\bar{\mu}_k(A), 1\big) = 1. \tag{3.13}$$

By means of (3.13) and a continuity argument it can be shown, that $\bar{\delta}(\lambda) < 1$ leads to a contradiction for $0 < \lambda < \bar{\mu}_k(A)$. Thus

$$\bar{\delta}(\lambda) = 1 \quad \text{for} \quad 0 \leqq \lambda \leqq \bar{\mu}_k(A).$$

Hence, (3.12) holds for $0 \leqq \delta \leqq 1$, $0 \leqq \lambda \leqq \bar{\mu}_k(A)$. By definition $t_k(\lambda | A) \equiv t^2(\lambda, 1)$. This completes the proof. ◊

Corollary 1. The *optimal choice* $\lambda_k \in [0, 1]$ with respect to the function $t_k(\lambda | A)$ is

$$\lambda_k(A) = \min\{1, 1/[2h_k(1+h_k)\ \mathrm{cond}\ (AJ_k)]\}. \tag{3.14}$$

An estimate of the rate of convergence of the modified Newton method with λ_k from (3.14) gives

$$\begin{aligned} t_k(\lambda_k | A) &= \left(1 - \frac{\lambda_k}{2}\right)^2 \quad \text{for} \quad \lambda_k < 1 \\ t_k(\lambda_k | A) &\leqq \tfrac{1}{4} \qquad\quad \text{for} \quad \lambda_k = 1. \end{aligned} \tag{3.15}$$

Corollary 2. For *natural scaling* as proposed in (2.8) the following inequalities hold (*extremal property*)

$$\left.\begin{array}{lll} \text{a)} & t_k(\lambda | J_k^{-1}) \leqq t_k(\lambda | A) \\ \text{b)} & \bar{\mu}_k(J_k^{-1}) \geqq \bar{\mu}_k(A) \\ \text{c)} & \lambda_k(J_k^{-1}) \geqq \lambda_k(A) \end{array}\right\} \quad \text{for all} \quad A \in A_M. \tag{3.16}$$

For the proof of (3.16) just use $\mathrm{cond}(AJ_k) \geqq 1$, $\mathrm{cond}(I) = 1$.

Remark. As x_k approaches the solution point x^*, h_k approaches 0 (let $\gamma \neq 0$). For this case one obtains from (3.14) that $\lambda_k = 1$ is the optimal choice of λ-independent of the monotonicity test. This is the ordinary Newton method.

The following theorem of global convergence of the modified Newton method is closely related to the preceding theorem.

Theorem 3.3. *Let* (3.6) *be assumed. Let the sequence* $\{x_k\}$ *be obtained by the iteration*

$$x_{k+1} = x_k - \lambda_k J(x_k)^{-1} F(x_k).$$

If the sequence $\{\lambda_k\}$ *is selected so that*

$$\lambda_{\min} \leqq \lambda_k \leqq \mu_k(A),$$

where

$$\mu_k(A) := \min\{1, 1/[h_k(1+h_k)\ \mathrm{cond}\ (AJ(x_k))] - \lambda_{\min}\}$$

and $\lambda_{\min}$ *is any fixed value in the range*

$$0 < \lambda_{\min} \leqq \min\{1, 1/[2Mh_0(1+Mh_0)m]\}$$

with $m := \max_{x \in G_0(A)} \mathrm{cond}\ (AJ(x))$, *then the following results hold*:

a) $\{x_k\} \subset G_0(A)$.

b) *If* $T(x_k | A) > 0$, *then*

$$T(x_{k+1} | A) < T(x_k | A) \text{ and } T(x_{k+1} | J(x_k)^{-1}) < T(x_k | J(x_k)^{-1}).$$

c) *There exists an* $x^* \in G_0(A)$ *with* $x^* = \lim_k x_k$ *and* $F(x^*) = 0$.

Proof. 1) With the assumptions (3.6) it is readily obtained that $h_0 < \infty$ and $\mathrm{cond}(J(x)) \leqq \bar{M} < \infty$ for all $x \in G_0(A)$. Since $m \leqq M\bar{M} < \infty$, there always exists a $\lambda_{\min}$ subject to the conditions above. If $x_k \in G_0(A)$, then $h_k \leqq Mh_0$ (Kantorovič

inequality) and $\operatorname{cond}(AJ_k) \leqq m$. With the notation $\bar{\mu} := \min\{1, 1/[2Mh_0(1+Mh_0)]\}$ the following inequalities hold: $\mu_k(A) \geqq \bar{\mu} \geqq \lambda_{\min} > 0$. Hence, the interval $[\lambda_{\min}, \mu_k(A)]$ is non-empty.

2) Now, let $T(x_k|A) > 0$ and $x_k \in G_0(A)$ be assumed for the sake of induction. Then, from Theorem 3.2, $T(x_{k+1}|A) \leqq t_k(\lambda|A)\, T(x_k|A)$ for $0 \leqq \lambda \leqq \bar{\mu}_k(A)$. A straightforward calculation yields

$$t_k(\lambda|A) \leqq \left(1 - \frac{\lambda_{\min}}{2}\right)^2 \quad \text{for} \quad \lambda \in [\lambda_{\min}, \mu_k(A)].$$

This implies $x_{k+1} \in G_0(A)$. The assumption above is obviously true for $k=0$. The trivial case $T(x_k|A) = 0$ can be omitted. Thus, statement *a* and the first part of statement *b* are proven. The second part of statement *b* follows from (3.16).

3) The inequality $T(x_{k+1}|A) \leqq t_k(\lambda_k|A)\, T(x_k|A)$ may be applied consecutively giving

$$T(x_k|A) \leqq \prod_{i=0}^{k-1} t_i(\lambda_i|A)\, T(x_0|A) \leqq \left(1 - \frac{\lambda_{\min}}{2}\right)^{2k} T(x_0|A)$$

Since $\lambda_{\min} > 0$:

$$\lim_k T(x_k|A) = 0.$$

$G_0(A)$ is compact by assumption IV. Thus there exists an $x^* = \lim_k x_k$, $x^* \in G_0(A)$. With $T(x^*|A) = 0$ and $A \in A_M$ one obtains $\|F(x^*)\| = 0$. ◊

Finally, it may be summarized that for global convergence of the method (3.4) the *existence* of some *constant* scaling matrix $A \in A_M$ (so that the assumptions (3.6) are valid) is *sufficient*. In applications it is impossible to select a constant matrix A in such a way as to minimize the number of iterations required to find the solution. Because of the *extremal property* (3.16), however, *natural scaling* seems to be an approach into this direction. This is confirmed by numerical experience. (See e.g. Example 2 of § 6.)

4. Numerical Realization of the Method

The purpose of this paragraph is to discuss the question of *how to select the rank of the Jacobian* at each iteration step in actual computation. In general, since reducing the rank means just solving a subsystem (compare (2.21)), the use of maximum rank would be favorable. Exceptions are necessary for "nearly singular" Jacobian and may be useful for bad initial data.

At first, it may be summarized what *natural scaling* as proposed in (2.8) means in actual computation. Let Δx_k denote the correction of the *ordinary* Newton method and $\overline{\Delta x}_k$ the correction of the *simplified* Newton method

$$\begin{aligned} \Delta x_k &= -J(x_k)^{-1} F(x_k) \\ \overline{\Delta x}_{k+1} &= -J(x_k)^{-1} F(x_{k+1}) \end{aligned} \tag{4.1}$$

where

$$x_{k+1} = x_k + \lambda_k \Delta x_k.$$

Then λ_k is selected in such a way as to ensure that

$$\|\overline{\Delta x}_{k+1}\| < \|\Delta x_k\|. \tag{4.2}$$

The calculation of $\widetilde{\Delta x}_{k+1}$ requires no additional function evaluations, but only $O(n^2)$ operations more compared with $O(n^3)$ operations required by the usual Newton step (one decomposition of $J(x_k)$ at iteration step k). For the case of less than maximum rank of the Jacobian, $J(x_k)^{-1}$ is replaced by $J(x_k)^{\dagger}$ in (4.1).

a) A-priori Selection of the Pseudo-Rank of the Jacobian. The following method is derived from results due to Wilkinson [33]. Let Δx_k denote the exact solution of the *exact* linear system

$$J(x_k)\Delta x_k = -F(x_k).$$

In contrast with the notations (3.7), now let J_k denote the *approximate* Jacobian actually used in x_k (rounding errors included) and let $\widetilde{\Delta x}_k$ denote the exact solution of the *modofied* linear system

$$J_k\widetilde{\Delta x}_k = -F(x_k).$$

Then a well-known result [33] yields

$$\frac{\|\Delta x_k - \widetilde{\Delta x}_k\|}{\|\widetilde{\Delta x}_k\|} \leqq \frac{e_k \operatorname{cond}(J_k)}{1 - e_k \operatorname{cond}(J_k)}, \quad \text{if} \quad e_k \operatorname{cond}(J_k) < 1 \tag{4.3}$$

where

$$e_k := \frac{\|J(x_k) - J_k\|}{\|J_k\|}.$$

In what follows a *nonlinear system* $F(x)$ will be said to be *ill-conditioned in* x_k if

$$e_k \operatorname{cond}(J_k) \geqq 1, \tag{4.4}$$

i.e. if the linear Newton system in x_k is ill-conditioned (sometimes 1 is replaced by 1/2 in the right-hand side).

Monitoring the algorithm by means of (4.4), however, raises three numerical problems: First, the computation of $\operatorname{cond}(J_k)$, e.g. by singular value decomposition (for reference see Golub/Reinsch [15]) requires substantially more computing time than the solution of the linear Newton system. Secondly, in applications even the order of magnitude of e_k is usually unknown. Thirdly, the estimate (4.3) gives only an upper bound of the relative error of $\widetilde{\Delta x}_k$. The actual error may be much lower. Therefore, it seems to be sufficient in applications to use a *lower bound of* $\operatorname{cond}(J_k)$ instead of the correct value. This may be done by means of the following result.

Lemma 4.1. *Let the approximate Jacobian J_k be decomposed so that $J_k = QRC$ where Q is a Householder matrix, R is an upper triangular matrix and C is a column permutation matrix (updating the information of column pivoting due to* Businger/Golub [8]). *Let r_i denote the diagonal elements of the matrix R with $r_n \neq 0$. Then*

$$\operatorname{cond}(J_k) \geqq \frac{|r_1|}{|r_n|} \geqq \frac{|r_1|}{|r_{n-1}|} \geqq \cdots \geqq \frac{|r_1|}{|r_2|} \geqq 1. \tag{4.5}$$

Proof. The following inequality is known to hold for triangular matrices:

$$\operatorname{cond}(R) \geqq \max_{i,k} \frac{|r_i|}{|r_k|} \qquad \text{(see e.g. [31]).}$$

Column pivoting as proposed in [8] ensures that

$$|r_1| \geqq |r_2| \geqq \cdots \geqq |r_{n-1}| \geqq |r_n|.$$

Finally, since Q and C are orthogonal, it is obtained

$$\operatorname{cond}(J_k) = \operatorname{cond}(R). \quad \diamond$$

Moreover, column pivoting due to [8] ensures theoretically that the matrix R is of the form (2.11), if and only if J_k is singular. Since Q and C are orthogonal, the multiplication rule of the pseudoinverse of a product of matrices yields

$$J_k^\dagger = C^T R^\dagger Q.$$

$R^\dagger$ is calculated according to (2.15). This justifies the former restriction to the special case (2.11).

A-priori Test. *Same notations as in Lemma* 4.1. *Let* e *denote an input parameter subject to* $e \leqq e_k$ *for all* k. *If for some* $s < n$

$$|r_s| > |r_1|\, e \quad \textit{and} \quad |r_{s+1}| \leqq |r_1|\, e, \tag{4.6}$$

then let s *define the pseudo-rank of the Jacobian to be used in* (2.16).

Lemma 4.2. *If condition* (4.6) *holds, then* $F(x)$ *is ill-conditioned in* x_k.

Proof. $e_k \operatorname{cond}(J_k) \geqq e \dfrac{|r_1|}{|r_{s+1}|} \geqq 1. \quad \diamond$

The test (4.6) is performed during the decomposition of J_k, i.e. before the Newton correction is actually computed. That is why (4.6) is called *a-priori* test. It may be noticed that both the test (4.6) and the condition number depend on *scaling*. Therefore it is advisable to scale J_k before the actual decomposition (compare Bauer [2]). The input parameter e is usually selected to be of the order of magnitude of the relative machine precision.

b) A-posteriori Selection of the Pseudo-Rank of the Jacobian. Since the reverse statement of Lemma 4.2 is not true, there are cases of ill-condition that are not detected by the test (4.6). In the numerical solution of *linear* systems the question of how to classify a given matrix as to be singular or not, was discussed over years. The state-of-the-art decision is made by iterative refinement of the solution: no attempt at obtaining the solution is made unless the norm of the first correction is significantly smaller than the norm of the initially proposed solution (compare [8]). This method can be extended to the *non-linear* case recalling that iterative refinement just means Newton's method for linear $F(x)$ (with the exception of the fact that the residual vector is computed using double-length accumulation of inner products). The use of the monotonicity test (4.2) instead of the monotonicity test (1.5) is the nonlinear analogue of the fact that in the linear case the *corrections* are compared instead of the *residuals*. Finally $\lambda_k = 1$ (linear case) is extended to $\lambda_{\min} \leqq \lambda_k \leqq 1$, for some suitably chosen positive *input parameter* $\lambda_{\min}$.

(4.7) **Rank-Strategy.** *Let* $\lambda_{\min}$ *denote an input parameter in the range* $0 < \lambda_{\min} \leqq 1$. *At the iterate* x_k *let condition* (4.2) *be not satisfied for all* λ_k *tried in the range*

$$\lambda_{\min} \leqq \lambda_k \leqq 1$$

Then reduce the pseudo-rank of the Jacobian by one, compute the new correction Δx_k from (2.16), *and try once more, e.g. with* $\lambda_k = 1$. (*To be repeated, if necessary.*)

This type of *rank-strategy* is now backed theoretically by means of the result (3.14). To simplify notations, let the transition from the modified Newton method for pseudo-rank s (say $s = n$) of the Jacobian to the modified Newton method for pseudo-rank $s' < s$ be described by the following

(4.8) **Transition Tableau.**

$$
\begin{aligned}
s = n &\rightarrow s' < s \\
F(x_k) &\rightarrow PF(x_k), \quad \text{say} \quad \|\bar{P}F(x_k)\| > 0 \\
J(x_k) &\rightarrow J(x_k)' := \left[\begin{array}{c|c} J & j \\ \hline 0 & 0 \end{array}\right] \ (2.11)\ j := 0 \text{ without loss of generality} \\
\text{exist } \bar{J}(x_k)^{-1} &\rightarrow J(x_k)'^{\dagger} = \left[\begin{array}{c|c} J^{-1} & 0 \\ \hline 0 & 0 \end{array}\right] \\
\Delta x_k = -J(x_k)^{-1}F(x_k) &\rightarrow P\Delta x_k' = -J^{-1}PF(x_k), \qquad \bar{P}\Delta x_k = 0 \Rightarrow \bar{P}x \equiv \bar{P}x_k \\
T\left(x \mid J(x_k)^{-1}\right) &\rightarrow T\left(x \mid J(x_k)'^{\dagger}\right).
\end{aligned}
$$

It may be seen that the right-hand side refers to a modified Newton method for the reduced mapping $PF(Px)$. Hence, the results of theorem 3.2 for the mapping $F(x)$ may be compared with the results for the mapping $PF(Px)$.

Comparison Lemma 4.3. *Notations of the transition tableau* (4.8). *Let primed quantities correspond to the right-hand side of* (4.8). $\lambda_k(A)$ *and* $\lambda_k'(A)$ *are selected due to* (3.14). *Assume* $\lambda_k < 1$ *and* $\beta' < \beta = \beta_k$. *Then*

(4.9)
a) $\lambda_k'\left(J(x_k)^{\dagger}\right) > \lambda_k\left(J(x_k)^{-1}\right)$
b) $\lambda_k'(I) > \lambda_k(I)$.

Proof. a) Employ (3.14) for *natural scaling*. Then $\lambda_k < 1$ implies

$$\lambda_k = 1/[2h_k(1+h_k)] < 1 \quad \text{and} \quad \lambda_k' = \min\{1,\ 1/[2h_k'(1+h_k')]\}.$$

Moreover

$$\|PJ(x) - PJ(y)\| \leqq \|J(x) - J(y)\| \leqq \gamma \|x - y\|$$

implies

$$\exists \gamma' \leqq \gamma : \|PJ(x) - PJ(y)\| \leqq \gamma' \|x - y\|$$

and

$$h_k' = \beta'^2 \gamma' \|PF(x_k)\| < \beta^2 \gamma \|F(x_k)\| = h_k.$$

This proves statement *a*.

b) $\operatorname{cond}(J) = \|J\|\|J^{-1}\| = \|PJ(x_k)\|\beta_k' \leqq \|J(x_k)\|\beta' < \|J(x_k)\|\beta_k = \operatorname{cond}\left(J(x_k)\right)$. ◊

In applications, the selection of λ_k is not due to (3.14) but ad hoc. Nevertheless, numerical experience confirms the results (4.9). A reasonable choice of the range of the input parameter $\lambda_{\min}$ appeared to be

$$\tfrac{1}{16} \leqq \lambda_{\min} \leqq \tfrac{1}{8}.$$

c) A Rule of Thumb for Bad Initial Data. In some real life examples only bad initial guesses of the solution are available. A well-known rule of thumb for these cases is to start the modified Newton method with small values λ_0. This is backed by (3.14): x_k "far from x^*" means h_k "large", which implies $\lambda_k(A)$ "small". An efficient alternative rule of thumb can be derived from the result (4.9).

(4.10) **Rank-Strategy.** *Start with initial pseudo-rank of the Jacobian less than maximum rank. In the course of the iteration increase the rank gradually.*

Employing (4.10) gives usually smaller corrections $\Delta x_k'$, as for the Penrose pseudoinverse the following result is well-known:

$$\|\Delta x_k'\| \leqq \|\Delta x_k\| \Leftrightarrow T(x_k | J_k^\dagger) \leqq T(x_k | J_k^{-1}) \tag{4.11}$$

5. Application to the Solution of Multiple Shooting Equations

The purpose of this paragraph is to give a multiple shooting modification of Newton's method, similar to the method derived so far. This is necessary because of the sparse structure of the Jacobian $(n(m-1), n(m-1))$-matrix

$$J(s) = \begin{bmatrix} G_1, & & -I & & 0 \\ & \ddots & & \ddots & \\ & & G_{m-2}, & & -I \\ A & 0 & & & BG_{m-1} \end{bmatrix},$$

where

$$G_j := \frac{\partial y(x_{j+1}; x_j, s_j)}{\partial s_j}, \quad j = 1, \ldots, m-1$$

$$A := \frac{\partial r(y_a, y_b)}{\partial y_a}, \qquad B := \frac{\partial r(y_a, y_b)}{\partial y_b}$$

and the identity matrix I are (n, n)-matrices.

For *nonsingular* Jacobian the Newton correction Δs is computed following Stoer/Bulirsch [32]:

$$\begin{aligned} \Delta s_1 &= E^{-1} u \\ \Delta s_{j+1} &= G_j \Delta s_j + F_j, \quad j = 1, \ldots, m-2 \end{aligned} \tag{5.1}$$

where

$$\begin{aligned} E &:= A + BG_{m-1} \cdots G_1 \\ u &:= -F_{m-1} - BG_{m-1}F_{m-2} - \cdots - BG_{m-1} \cdots G_2 F_1. \end{aligned}$$

Remark. For a smooth trajectory (i.e. $F_1 = \cdots = F_{m-2} = 0$) E is the Fréchet-derivative of the operator that describes the two-point boundary value problem.

The (n, n)-matrix E is nonsingular, if and only if $J(s)$ is nonsingular, as may be seen from the following

Lemma 5.1. *Let C_i^l denote the (n,n)-matrices*

$$C_i^l = G_i G_{i-1} \ldots G_l \quad \text{for} \quad 1 \leqq l \leqq i \leqq m-1.$$

Let L, R, and S denote the $(n(m-1), n(m-1))$-matrices

$$L := \begin{bmatrix} BC_{m-1}^2, \ldots, BC_{m-1}^{m-1}, & I \\ -I \quad \ddots \quad 0 & 0 \\ \quad \ddots \quad \ddots & \vdots \\ 0 \quad -I, & 0 \end{bmatrix}$$

$$R := \begin{bmatrix} I & & 0 \\ C_1^1 & \ddots & \\ \vdots & \ddots & \ddots \\ C_{m-2}^1 & \cdots & C_{m-2}^{m-2} \; I \end{bmatrix}$$

$$S := \begin{bmatrix} E & & & \\ & I & & 0 \\ & & \ddots & \\ & 0 & & I \end{bmatrix}.$$

Then

$$\text{(5.2)} \qquad \begin{aligned} &\text{a)} \quad LJR = S \\ &\text{b)} \quad \det(J) = \det(E). \end{aligned}$$

The proof of a) is rather space consuming but straightforward.

The proof of b) is obtained from a) with $\det(L) = \det(R) = 1$.

If E is *singular*, then the system (5.1) cannot be "solved". The modification (2.16) of Newton's method using the Penrose pseudoinverse of the Jacobian as a whole would ignore the sparse structure of $J(s)$. Moreover, it is important for actual computation that independent of m only the (n, n)-matrix E has to be inverted. This inspires the following modification of (5.1):

$$\text{(5.3)} \qquad \begin{aligned} \Delta s_1 &= E^{\dagger} u \\ \Delta s_{j+1} &= G_j \Delta s_j + F_j, \quad j = 1, \ldots, m-2. \end{aligned}$$

The question of how to select the rank of E is decided by means of the results of § 4. Using the decomposition (5.2a) the system (5.3) may be written as

$$\Delta s = -J^- F = -(RR^T) \operatorname{grad} T(s \mid R^{-1} J^-)$$

where

$$J^- = RS^{\dagger} L.$$

Here J^- denotes a *generalized inverse* satisfying

$$\text{(5.4)} \qquad \begin{aligned} &(J^- J)^T = (RR^T)^{-1} J^- J (RR^T), \; (JJ^-)^T = (L^T L) JJ^- (L^T L)^{-1}, \\ &J^- JJ^- = J^-, \quad JJ^- J = J. \end{aligned}$$

The difference of the Penrose axioms (2.17) and the axioms (5.4) implies restrictions for the class of "*appropriate*" *level functions*.

Lemma 5.2. *Notations as before. $T(s|H)$ denotes a level function. Let $\bar{P}E = 0$, $\Delta s = -J^- F$. Then*

$$\text{(5.5)} \qquad \begin{aligned} &\text{a)} \; \Delta s^T \operatorname{grad} T(s \mid H) = -2T(s \mid H) + (H(I - JJ^-)F)^T HF \\ &\text{b)} \; (H(I - JJ^-)F)^T HF = -(\bar{P}u)^T (H^T HF)_{m-1}. \end{aligned}$$

Proof. a) $\Delta s^T \operatorname{grad} T(s|H) = -(J^-F)^T J^T H^T HF = -(H(JJ^-+I-I)F)^T HF$
$= -2T(s|H) + (H(I-JJ^-)F)^T HF$

b) Now $(I-JJ^-)F = F + J\Delta s = L^{-1}(LF + LJ\Delta s)$.

Using the notations of Lemma 5.1 it may be seen that

$$L^{-1}(LF+LJ\Delta s) =$$

$$\begin{bmatrix} 0, & -I & & \\ \vdots & & \ddots & \\ 0 & & & -I \\ I, & BC_{m-1}^2 & \cdots & BC_{m-1}^{m-1} \end{bmatrix} \begin{bmatrix} -u + E\Delta s_1 \\ -F_1 - G_1\Delta s_1 + \Delta s_2 \\ \vdots \\ -F_{m-2} - G_{m-2}\Delta s_{m-2} + \Delta s_{m-1} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ E\Delta s_1 - u \end{bmatrix}.$$

With $P(E\Delta s_1 - u) = 0$ and $\bar{P}E = 0$ the result above is obtained. ◊

An important consequence of this result is that the usual level function

$$T(s|I) = \tfrac{1}{2}\sum_{j=1}^{m-1} \|F_j\|^2$$

is *not "appropriate"* for the modification (5.3) of Newton's method, unless $m = 2$. Rather the *"appropriate"* level functions

$$T(s|J^-) = \tfrac{1}{2}\sum_{j=1}^{m-1} \|\Delta s_j\|^2$$

or

$$T(s|R^{-1}J^-) = \tfrac{1}{2}(\|\Delta s_1\|^2 + \sum_{j=1}^{m-2} \|F_j\|^2)$$

should be used to determine the sequence $\{\lambda_k\}$.

Proof. $\Delta s^T \operatorname{grad} T(s|I) = -\sum_{j=1}^{m-2} \|F_j\|^2 - \|PF_{m-1}\|^2 - (\bar{P}F_{m-1})^T(\bar{P}F_{m-1} + \bar{P}u)$. For arbitrary functions F_j the right-hand side is negative definite if and only if $m = 2$. The proof for $T(s|J^-)$ and $T(s|R^{-1}J^-)$ is easily obtained using (5.5a) and the axioms (5.4). ◊

As for convergence of this modification of Newton's method, the result (3.1) due to [3] may be applied. It can be shown that the method converges locally (for $\lambda_k = 1$) and yields a smooth trajectory $y(x)$ (satisfying $F_1 = \cdots = F_{m-2} = 0$ for $m > 2$) subject to

$$y' = f(x, y), \qquad Pr(y(a), y(b)) = 0. \tag{5.7}$$

No information is given about $\bar{P}r(y(a), y(b))$ except the analogue of the least squares property (3.2).

Remark. A sufficient condition for $\bar{P}E = 0$ is

$$\bar{P}r(y(a), y(b)) = \text{const} \tag{5.8}$$

(compare condition (2.12)).

6. Numerical Results

Example 1. Least squares problem due to Brown/Dennis [4]. Minimize

$$T(x|I) = \sum_{i=1}^{10} \left(f(\xi_1, \xi_2, p_i) - f(1, 10, p_i)\right)^2$$

where

$$p_i = 0.1 \cdot i, \quad i = 1(1)10, \quad f(\xi_1, \xi_2, p) := e^{-\xi_1 p} - e^{-\xi_2 p}.$$

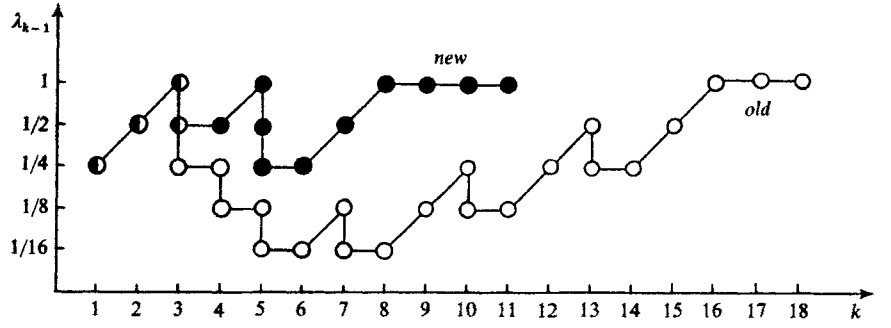
The experiment was run on the TR 440 of the *Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften.* The computations were performed in FORTRAN single precision with a 38 bit mantissa. The analytic expression of the Jacobian was used. Each evaluation of the Jacobian was counted for 2 function evaluations. Using the different starting points given in [4], the problem was solved by a (non-optimized) standard version of a least squares algorithm employing the results of this paper (see last row of Table A). In [4] all starting points (with the exception of the last one) were reported to lead to a failure (F) of the ordinary Newton method (there called Gauss-Newton method), see first row of Table A. For comparison the results presented in [4] for the Levenberg-Marquardt method (L.M.) and its derivative free version (D.F.L.M.) are arranged (second row of Table A).

Table A. Number of function evaluations required to reduce $T(x|I)$ to less than 10^{-5}

	Starting point x_0 and $T(x_0\|I)$				
	(0, 0) 3.06*	(0, 20) 2.09	(5, 0) 19.6	(5, 20) 1.81	(2.5, 10) 0.808
ordinary (Gauss-)Newton method	F	F	F	F	16
L.M. (D.F.L.M.)	22	25	25	31	16
modified (Gauss-)Newton method (this paper)	19	14	21	14	13

* It may be noted that for $x_0 = (0, 0)$ the Jacobian is *singular.*

The following three examples were run on an IBM 370/165 of the *Kernforschungsanlage Jülich* via an IBM 2780 terminal at the *Universität zu Köln.* The computations were performed in FORTRAN double precision with a 56 bit mantissa. The examples refer to multiple shooting. For an ALGOL program of the type of algorithm used herein see [7]. Unless noted, the Jacobian is approximated by numerical differentiation at each iterate. The damping factors λ_k are selected from the set $\{1, \frac{1}{2}, \ldots, \lambda_{\min}\}$, where $\lambda_{\min}$ denotes the input parameter as introduced for the *rank-strategy* (4.7). The following three figures show the comparative behavior of different algorithms by drawing the sequences $\{\lambda_k\}$ with unaccepted trials included. Each trial of λ means one function evaluation (in multiple shooting, that means the numerical solution of $(m-1)$ initial value problems for differential equations). The sign ▾ marks the reduction of the applied pseudorank of the Jacobian.



	monotonicity test	function eval.	level function
old	(6.1)	84	$T(s^{(18)}\mid I) = 1.3 \cdot 10^{-11}$
new	(6.2)	52	$T(s^{(11)}\mid I) = 0.4 \cdot 10^{-11}$

Fig. 2

Example 2. Re-entry auxiliary problem ($n = 6$, $m = 2$).

For reference to this boundary value problem see [32] (7.3.7.1), (7.3.7.3), (7.3.7.4), and (7.3.7.10). Since $m = 2$ (single shooting), *the method applied is the same as proposed for the solution of general systems of equations.* Fig. 2 shows a comparison of the sequence $\{\lambda_k\}$ using the monotonicity test

$$T(s^{(k+1)}\mid I) < T(s^{(k)}\mid I) \qquad (old) \tag{6.1}$$

and the sequence $\{\lambda_k\}$ using the monotonicity test

$$T\left(s^{(k+1)}\mid J(s^{(k)})^{-1}\right) < T\left(s^{(k)}\mid J(s^{(k)})^{-1}\right) \qquad (new) \tag{6.2}$$

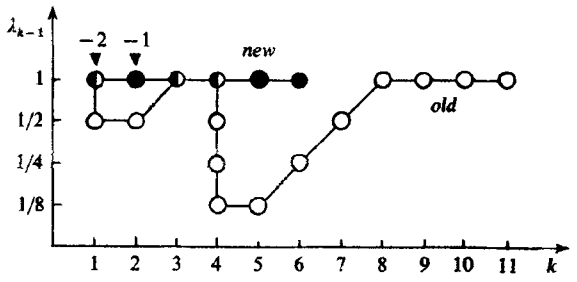
as proposed herein.

One evaluation of the Jacobian counts for 3 function evaluations. Throughout both of the iterations the Jacobian is of full rank. The experiment affirms the result (3.16c). For the common initial data $s^{(0)}$ one obtains $T(s^{(0)}\mid I) = 0.31$. The lower bounds of the *condition numbers* of the (3, 3)-matrix E (scaled) range from $0.4 \cdot 10^2$ to $0.5 \cdot 10^3$ (at the solution point $\tilde{s}$).

Example 3. Re-entry problem ($n = 7$, $m = 9$).

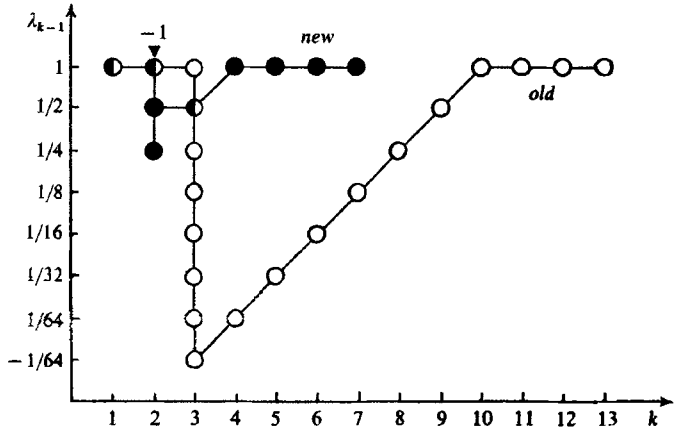
For reference see [32] (7.3.7.1), (7.3.7.6), and (7.3.7.7). For the initial data $s^{(0)}$ one obtains $T(s^{(0)}\mid I) = 1 \cdot 10^2$. Fig. 3 shows a comparison of the sequences $\{\lambda_k\}$ with and without *rank-strategy* (4.10). The actually applied pseudo-rank at $s^{(0)}$ is 2 less (and at $s^{(1)}$ 1 less) than the full rank of the matrix E (as is indicated by the negative numbers over ▼).

One evaluation of the Jacobian counts for 7 function evaluations. The lower bounds of the *condition numbers* of E (scaled) range from $3 \cdot 10^4$ (at $s^{(0)}$) to the maximum value $4 \cdot 10^6$; for the solution point one obtains $1 \cdot 10^6$. Reduction of the maximum rank by 2 (at $s^{(0)}$) yields $0.9 \cdot 10^1$, reduction by 1 (at $s^{(1)}$) yields $1 \cdot 10^4$.



		function eval.	level function
old	without (4.10)	100	$T(s^{(11)}\|I) = 0.9 \cdot 10^{-11}$
new	with (4.10)	56	$T(s^{(0)}\|I) = 0.3 \cdot 10^{-11}$

Fig. 3



	level function	function eval.	computing time
old	$T(s\|I)$	99	46.8 sec.
new	$T(s\|J(s^{(k)})^-)$	38	16.0 sec.

Fig. 4

Example 4. Optimal control for synenergetic orbit plane change of a spacecraft ($n = 13$, $m = 6$).

This example refers to a hypersonic maneuver of a spacecraft as described by Dickmanns [12]. Let a spacecraft fly on an orbit around a planet having a sufficiently dense atmosphere. The problem is to turn the orbit plane to a new one of the same altitude. The quantity to be minimized is the *total amount of fuel* required for the maneuver. It is important to limit the heating of the spacecraft.

The mathematical model as given in [6] leads to a two-point boundary value problem with 13 differential equations and several inequality constraints for the control variables. In actual computation the problem appeared to be *very sensitive*. Hence, the solution was attacked by computing a *homotopy sequence* of familiar problems using multiple shooting techniques (compare [6]). The example presented here is one of the problems from the homotopy chain. The number of nodes is $m = 6$.

Rank-1 corrections of the Jacobian due to [5] (see also [32] and [10]) are employed. Numerical differentiation is inserted only at suitably selected iterates. The example indicates the efficiency of the *rank-strategy* (4.7) for $m > 2$ in connection with the level function $T(s|J^-)$, even if only a poor approximation of the Jacobian is given. Fig. 4 shows a comparison of two algorithms using $T(s|I)$ and $T(s|J^-)$, respectively.

The algorithm using $T(s|I)$ would not converge unless two empirical rules of thumb were employed: (I) it is *not* required that $T(s|I)$ decreases at the first iteration step (actually one obtains $T(s^{(0)}|I) = 0.6 \cdot 10^{-2}$, $T(s^{(1)}|I) = 0.4 \cdot 10^{3}$, whereas $T(s^{(1)}|J(s^{(0)})^{-1}) < T(s^{(0)}|J(s^{(0)})^{-1})$!) (II) λ_2 is permitted to be *negative* (this means, that the Jacobian approximation is so poor that the result (2.5) cannot be claimed for the actual Newton correction $\Delta\tilde{s}^{(2)}$). Like in Fig. 3 the sign ▼ marks the reduction (by 1) of the pseudo-rank of the Jacobian. The lower bound of the *condition number* of the matrix E is $1.4 \cdot 10^5$ at the solution point.

Employing the results of this paper U. Zimmermann [34] computed the solutions of the complete homotopy sequence. An extensive comparison given in [34] clearly shows the efficiency of the method as proposed herein.

Example 5. Optimal descent of the second stage of a space shuttle subject to heating restrictions ($n = 13$, $m \geqq 9$).

This problem refers to a special type of spacecraft called *space shuttle*. The physical model is due to Dickmanns [11]. By means of the results of this paper Pesch [27] treated the optimal descent of the second stage. (The description given here follows the lines of [27] where the results of extensive computations were presented.) The quantity to be maximized is the *range*. As in example 4 *heating restrictions* are required.

The mathematical model leads to a two-point boundary value problem with 13 differential equations. For the *physical quantities* (*velocity* v, *height* h, *flight path angle* γ, *range* Λ etc.) the following 6 differential equations hold:

$$
\begin{aligned}
\dot{v} &= -\frac{W}{m} - g \cdot \sin\gamma \\
\dot{\chi} &= \frac{A}{m \cdot v} \cdot \frac{\sin\mu}{\cos\gamma} - \frac{v}{r} \cdot \cos\gamma \cdot \cos\chi \cdot \tan\Lambda \\
\dot{\gamma} &= \frac{A}{m \cdot v} \cdot \cos\mu - \left(\frac{g}{v} - \frac{v}{r}\right) \cdot \cos\gamma \qquad (6.3) \\
\dot{\Lambda} &= \frac{v}{r} \cdot \cos\gamma \cdot \sin\chi \\
\dot{h} &= v \cdot \sin\gamma \qquad (r = R + h) \\
\dot{\theta} &= \frac{v}{r} \cdot \frac{\cos\gamma}{\cos\Lambda} \cdot \cos\chi,
\end{aligned}
$$

where $A := c_A F \frac{\varrho}{2} v^2$, $W := c_W F \frac{\varrho}{2} v^2$, $c_W := c_{W_0} + k c_A^2$.

For the *lift coefficient* c_A (a *control* variable) three *inequality constraints* are required among which the heating restriction is the most severe:

$$
c_A \leqq c_{AH} = \sum_{j=1}^{5} B_j(h) H_j(h, v) + \Delta c_{AH}. \qquad (6.4)
$$

B_j, H_j, $i = 1, \ldots, 5$ are given functions, Δc_{AH} represents the heating of the space shuttle beyond the allowed limit temperature ($\Delta c_{AH} = 0$ means 2000 ° $F = 1093.3$ °C). The given *boundary conditions* are $v(0) = 7.85$ km/sec, $v(T) = 1.116$ km/sec, $\gamma(0) = -1.25$ °, $\gamma(T) = -2.7$ °, $h(0) = 95.0$ km, $h(T) = 30.0$ km, $\chi(0) = \Lambda(0) = \theta(0) = 0$ °.

The *total flight time* T is free ($\dot{T}=0$). From the calculus of variations one obtains 6 differential equations for the *adjoint variables* λ_v, λ_χ, λ_γ, λ_Λ, λ_h, λ_θ, e.g.

$$\begin{aligned}\dot{\lambda}_\gamma = -\Big\{&\lambda_v\Big[-g_0\Big(\frac{R}{R+h}\Big)^2\cos\gamma\Big]\\&+\lambda_\chi\Big[c_A\frac{F\varrho_0}{2m}e^{-\beta h}v\frac{\sin\mu}{\cos^2\gamma}\cdot\sin\gamma+\frac{v}{R+h}\sin\gamma\cos\chi\tan\Lambda\Big]\\&+\lambda_\gamma\Big[\Big(\frac{g_0}{v}\Big(\frac{R}{R+h}\Big)^2-\frac{v}{R+h}\Big)\sin\gamma\Big]\\&+\lambda_\Lambda\Big[-\frac{v}{R+h}\sin\gamma\sin\chi\Big]\\&+\lambda_h[v\cos\gamma]\\&+\lambda_\theta\Big[-\frac{v}{R+h}\frac{\sin\gamma\cos\chi}{\cos\Lambda}\Big]\Big\}.\end{aligned}$$

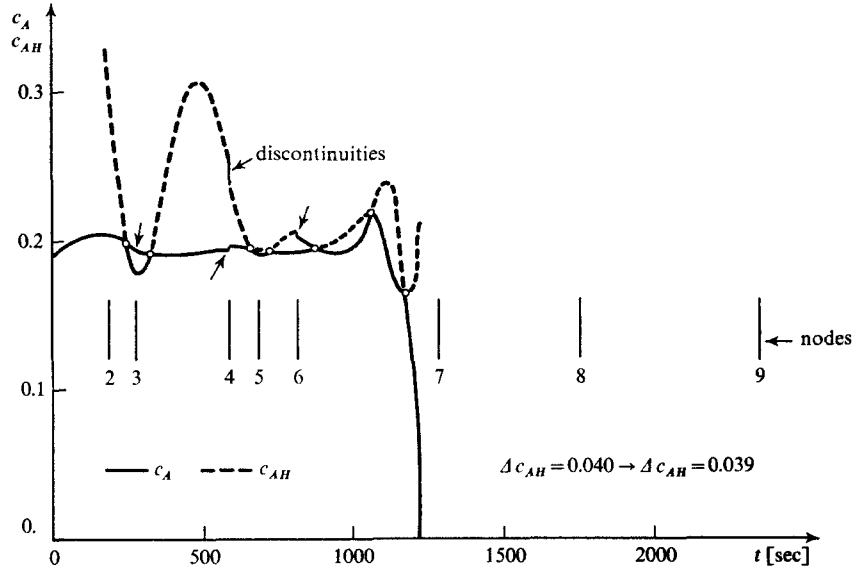
This equation *is nonlinear* since the controls c_A and μ are functions of the adjoint variables e.g.

$$\sin\mu=-\frac{\lambda_\chi}{w\cos\gamma},\qquad \cos\mu=-\frac{\lambda_\gamma}{w},\qquad w:=\sqrt{\Big(\frac{\lambda_\chi}{\cos\gamma}\Big)^2+\lambda_\gamma^2}.$$

Finally, the *transversality conditions* are 4 additional boundary conditions

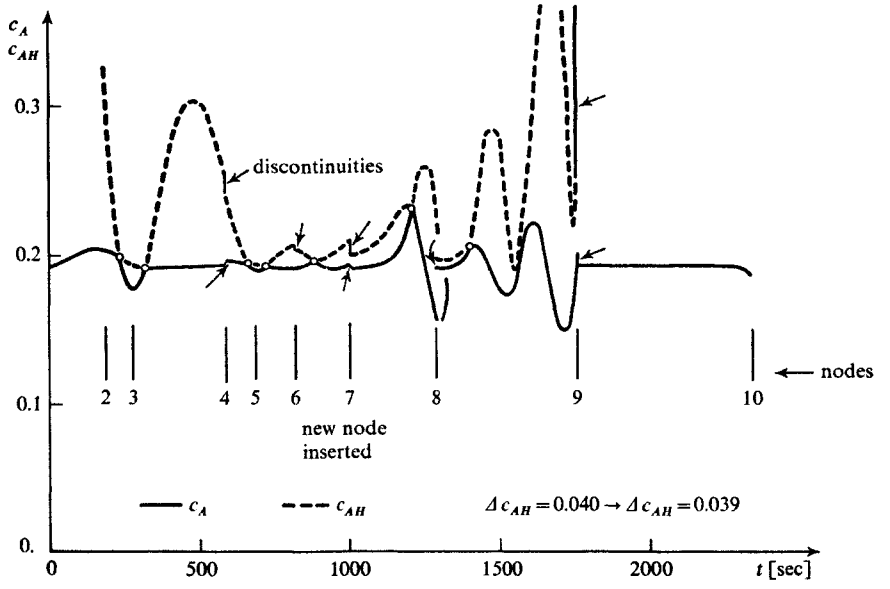
$$\lambda_\chi(T)=\lambda_\theta(T)=0.,\qquad \lambda_\Lambda(T)=-1.,\qquad [\lambda_v\dot{v}+\lambda_\chi\dot{\chi}+\lambda_\gamma\dot{\gamma}+\lambda_\Lambda\dot{\Lambda}+\lambda_h\dot{h}+\lambda_\theta\dot{\theta}]_{t=T}=0.$$

In actual computation the problem appeared to be *highly sensitive*. Hence, a sequence of familiar problems was constructed using Δc_{AH} from (6.4) as a homotopy parameter. To give an impression of the numerical difficulties three graphs for the control variable $c_A(t)$ are presented (due to [27]).



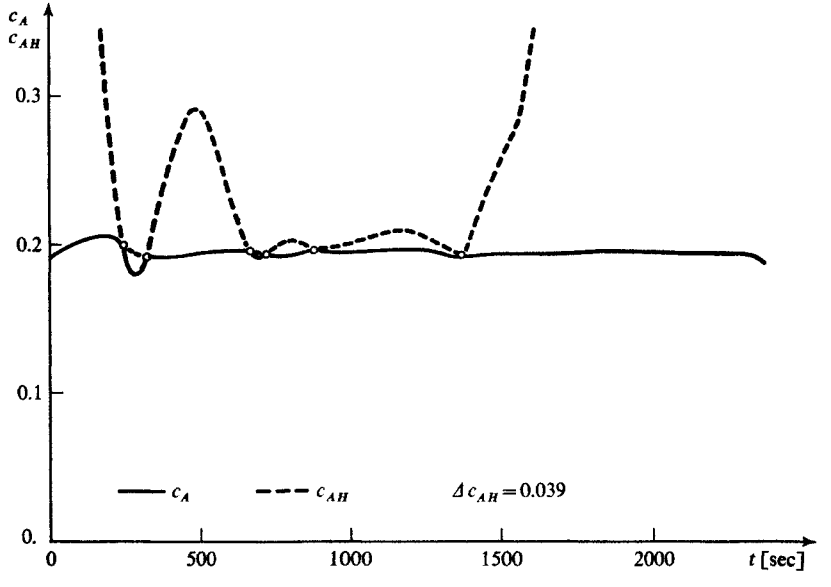
lift coefficient c_A *(control)*: Solution data for $\Delta c_{AH}=0.040$ do not yield a complete starting trajectory for $\Delta c_{AH}=0.039$ (exponential overflow)

Fig. 5



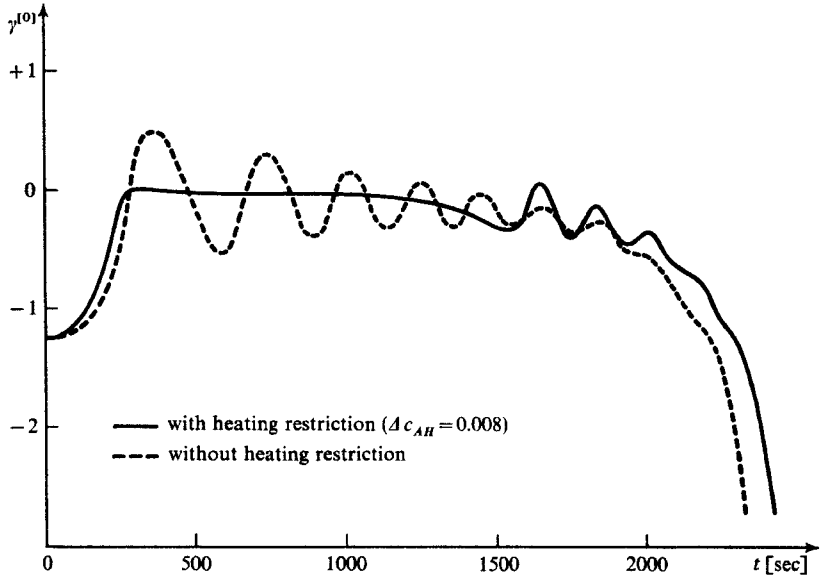
lift coefficient c_A *(control)*: insertion of a new node (7) generates a complete starting trajectory for $\Delta c_{AH} = 0.039$

Fig. 6



lift coefficient c_A *(control)*: solution for $\Delta c_{AH} = 0.039$

Fig. 7



flight path angle γ [27]

Fig. 8

In Fig. 5 the solution data for $\Delta c_{AH} = 0.040$ (at 9 nodes) were taken as the initial data for $\Delta c_{AH} = 0.039$: exponential overflow occurs in the subinterval 6. Upon inserting a new node the starting trajectory could be completed (see Fig. 6). The solution $c_A(t)$ for $\Delta c_{AH} = 0.039$ is shown in Fig. 7. The solution is correct to about 6–8 significants digits. Approaching $\Delta c_{AH} = 0.008$ the homotopy steps decreased and computing time increased: *for a change of* 0.0005 *in* Δc_{AH} *more than* 600 sec. *were required on the* IBM 370/165! Using a 56 bit mantissa the lower bound of the *condition number* of E (scaled) was about 10^{19} at the solution point. In Fig. 8 the graph $\gamma(t)$ for $\Delta c_{AH} = 0.008$ due to [27] is shown.

Conclusion

The modified Newton method as proposed herein, appears to work efficiently in solving ill-conditioned, highly nonlinear systems of equations (or non-linear least squares systems) as far as an approximation of the Jacobian is available and the dimension of the system is not too large. These properties hold for multiple shooting where the dimension of the system is equal to the number of differential equations to be solved simultaneously. By means of the results given here significant progress in solving sensitive optimal control problems could be made.

Acknowledgment. The author wishes to thank Prof. R. Bulirsch who stimulated and encouraged this work which is part of the author's dissertation. The author is indebted to Dr. E. D. Dickmanns who prepared initial data of Example 4 and 5 and to Prof. P. Frederickson for reading the manuscript.

References

1. Bard, Y.: Comparison of gradient methods for the solution of nonlinear parameter estimation problems. SIAM J. Numer. Anal. **7**, 157–186 (1970)
2. Bauer, F. L.: Optimally scaled matrices. Num. Math. **5**, 73–87 (1963)
3. Ben-Israel, Adi: A Newton-Raphson method for the solution of systems of equations. J. Math. Anal. Appl. **15**, 243–252 (1966)
4. Brown, K. M., Dennis, J. E., Jr.: Derivative-free analogues of the Levenberg-Marquardt and Gauss algorithms for nonlinear least squares approximation. Num. Math. **18**, 289–297 (1972)
5. Broyden, C. G.: A class of methods for solving nonlinear simultaneous equations. Math. Comp. **19**, 577–583 (1965)
6. Bulirsch, R.: Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung. Vortrag im Lehrgang „Flugbahnoptimierung" der Carl-Cranz-Gesellschaft e.V., Okt. 1971
7. Bulirsch, R., Stoer, J., Deuflhard, P.: Numerical solution of nonlinear two-point boundary value problems I. To be published in Num. Math., Handbook Series Approximation
8. Businger, P., Golub, G. H.: Linear least squares solutions by Householder transformations. Num. Math. **7**, 269–276 (1965)
9. Cauchy, A.: Méthode générale pour la résolution des systèmes d'équations simultanées. C.R. Acad. Sci. Paris **25**, 536–538 (1847)
10. Deuflhard, P.: Ein Newton-Verfahren bei fastsingulärer Funktionalmatrix zur Lösung von nichtlinearen Randwertaufgaben mit der Mehrzielmethode. Universität zu Köln, Mathematisches Institut: Dissertation, 1972
11. Dickmanns, E. D.: Optimale dreidimensionale Gleitflugbahnen beim Eintritt in Planetenatmosphäre. Raumfahrtforschung **14**, Heft 3 (1970)
12. Dickmanns, E. D.: Optimal control for synenergetic plane change. Proc. XXth Int. Astronautical Congress 597–631 (1969)
13. Fletcher, R.: Generalized inverse methods for the best least squares solution of systems of nonlinear equations. Comp. J. **10**, 392–399 (1968)
14. Goldstein, A. A.: Cauchy's Methode der Minimierung. Num. Math. **4**, 146–150 (1962)
15. Golub, G. H., Reinsch, C.: Singular value decomposition and least squares solutions. Num. Math. **14**, 403–420 (1970)
16. Greenstadt, J.: On the relative efficiencies of gradient methods. Math. Comp. **21**, 360–367 (1967)
17. Householder, A. S.: Principles of numerical analysis. New York: McGraw-Hill 1953
18. Kantorovič, L., Akilow, G.: Functional analysis in normed spaces. Moscow: Fizmatgiz 1959. German Translation: Berlin: Akademie-Verlag 1964
19. Keller, H. B.: Numerical methods for two-point boundary-value problems. London: Blaisdell 1968
20. Kowalik, J., Osborne, M. R.: Methods for unconstrained optimization problems. New York: American Elsevier Publ. Comp., Inc. 1968
21. Levenberg, K. A.: A method for the solution of certain nonlinear problems in least squares. Quart. Appl. Math. **2**, 164–168 (1944)
22. Marquardt, D. W.: An algorithm for least-squares-estimation of nonlinear parameters. SIAM J. Appl. Math. **11**, 431–441 (1963)
23. Meyer, G. H.: On solving nonlinear equations with a one-parameter operator imbedding. University of Maryland, Computer Science Center: Techn. Rep. 67–50 (Sept. 1967)
24. Ortega, J. M., Rheinboldt, W. C.: Iterative solution of nonlinear equations in several variables. New York-London: Academic Press 1970
25. Osborne, M. R.: On shooting methods for boundary value problems. J. Math. Anal. Appl. **27**, 417–433 (1969)
26. Penrose, R.: A generalized inverse for matrices. Proc. Cambridge Philos. Soc. **51**, 406–413 (1955)

27. Pesch, H.-J.: Numerische Berechnung optimaler Steuerungen mit Hilfe der Mehrzielmethode dokumentiert am Problem der Rückführung eines Raumgleiters unter Berücksichtigung von Aufheizungsbegrenzungen. Universität Köln, Mathematisches Institut: Diplomarbeit, 1973
28. Powell, M. J. D.: A FORTRAN subroutine for solving systems of non-linear algebraic equations. A.E.R.E. Rep. 5947, Harwell (1968)
29. Rheinboldt, W. C.: Local mapping relations and global implicit function theorems. University of Maryland, Computer Science Center: Techn. Rep. 68–52 (Febr. 1968)
30. Rosenbloom, P. C.: The method of steepest descent. Proc. Symp. Appl. Math. AMS 6, 127–176 (1956)
31. Stoer, J.: Einführung in die Numerische Mathematik I. Heidelberger Taschenbuch 105. Berlin-Heidelberg-New York: Springer 1972
32. Stoer, J., Bulirsch, R.: Einführung in die Numerische Mathematik II. Heidelberger Taschenbuch 114. Berlin-Heidelberg-New York: Springer 1973
33. Wilkinson, J. H.: Rounding errors in algebraic processes. London: Her Majesty's Stationary Office 1963 (German translation: Rundungsfehler. Heidelberger Taschenbuch 44. Berlin-Heidelberg-New York: Springer 1969)
34. Zimmermann, U.: Numerische Berechnung optimaler Steuerungen unter Verwendung der Mehrzielmethode bei mehrfacher Beschränkung im Phasenraum mit Beispielen aus der Flugbahnoptimierung. Universität Köln, Mathematisches Institut: Diplomarbeit, 1973

Peter Deuflhard
Mathematisches Institut der TU München
D-8000 München 2, Barerstr. 23
Bundesrepublik Deutschland