ON THE ROLE OF NATURAL LEVEL FUNCTIONS TO ACHIEVE GLOBAL CONVERGENCE FOR DAMPED NEWTON METHODS

H. Georg Bock
Interdisciplinary Center for Scientific Computing (IWR),
University of Heidelberg,
Im Neuenheimer Feld 368, D-69120 Heidelberg, Germany.
bock@iwr.uni-heidelberg.de

Ekaterina Kostina
As above
ekaterina.kostina@iwr.uni-heidelberg.de

Johannes P. Schlöder
As above
j.schloeder@iwr.uni-heidelberg.de

Abstract The paper discusses a new view on globalization techniques for Newton's method. In particular, strategies based on "natural level functions" are considered and their properties are investigated. A "restrictive monotonicity test" is introduced and theoretically motivated. Numerical results for a highly nonlinear optimal control problem from aerospace engineering and a parameter estimation for a chemical process are presented.

1. INTRODUCTION

It is well-known that stepsize strategies based on suitable merit functions can globalize the convergence of the damped Newton method. Experience shows, however, that the standard choices for merit functions may enforce very small stepsizes when the problems are only mildly ill-conditioned, even in the domain of full-step local convergence, thus

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35514-6_15

M.J.D. Powell and S. Scholtes (Eds.), *System Modelling and Optimization: Methods, Theory and Applications.*

making the method very inefficient. So called "natural level functions", originally introduced by Deuflhard, can avoid this effect, but up to now lack a rigorous convergence theory. The present paper presents a new view on successful globalization strategies based on these merit functions. In particular, it is shown that a stepsize criterion given by the authors (the so called "restrictive monotonicity test") provides a theoretical justification. Extensions to the related problems of constrained least squares and constrained l_1 parameter estimation problems are suggested. Numerical results for real life applications from aerospace control problems and parameter estimation for chemical processes are given.

2. NEWTON'S METHOD FOR NONLINEAR EQUATIONS

We consider a finite dimensional but possibly large system of highly nonlinear equations

$$F(x) = 0.$$

Starting from an initial guess x^0, Newton's method improves a given estimate x^k iteratively by applying the formula

$$x^{k+1} = x^k + \Delta x^k. \tag{1}$$

The increment Δx^k solves the linear system of equations

$$F(x^k) + J(x^k)\Delta x^k = 0, \tag{2}$$
$$\text{where } J := \left(\frac{\partial F}{\partial x}\right).$$

The local convergence properties of this "full-step" version of Newton's method have been investigated thoroughly and may be formulated as follows.

Theorem 1 (local convergence properties) *Let $F : D \subset R^n \to R^n$ be twice continuously differentiable, $J(x)$ be nonsingular for all $x \in D$, and D be a domain. Assume further that*

$$||J(y)^{-1}(J(x + t\Delta x) - J(x))\Delta x|| \leq \omega t\, ||\Delta x||^2, \tag{3}$$
$$\omega \leq \infty,$$

for all $t \in]0, 1], x, y = x + \Delta x \in D$ with $\Delta x = -J(x)^{-1}F(x) \neq 0$, i.e. a global bound ω for the "curvature" exists, and that the initial guess x^0 is sufficiently near to a solution:

$$\delta^0 := \frac{\omega}{2}||\Delta x^0|| < 1. \tag{4}$$

Then the following holds:

- *if* $D^0 := B\left(x^0, ||\Delta x^0||/(1-\delta^0)\right) \subset D$, *then the sequence of iterates defined by (1) remains in* D^0,
- *there exists* $x^\star \in D^0$ *with* $F(x^\star) = 0$ *and* $x^k \to x^\star$ $(k \to \infty)$,
- *an a priori error estimate holds*

$$||x^k - x^\star|| \leq (\delta^0)^k \frac{||\Delta x^0||}{1-\delta^0},$$

- *and convergence is quadratic with*

$$||\Delta x^{k+1}|| \leq \frac{\omega}{2} ||\Delta x^k||^2.$$

The **proof** follows the lines of Banach's Fixed Point Theorem. Contractivity of the sequence of iterates is given by

$$\begin{aligned} ||\Delta x^{k+1}|| &= ||J(x^{k+1})^{-1}\left(F(x^{k+1}) - F(x^k) - J(x^k)(x^{k+1} - x^k)\right)|| \\ &= ||J(x^{k+1})^{-1} \int_0^1 \left(J(x^k + t\Delta x^k) - J(x^k)\right)\Delta x^k dt|| \\ &\leq \int_0^1 \omega t\, ||\Delta x^k||^2\, dt = \frac{\omega}{2}\, ||\Delta x^k||^2. \end{aligned}$$

Since from here

$$||x^{k+p} - x^k|| \leq \sum_{i=0}^{p-1} ||\Delta x^{k+i}|| \leq (\delta^0)^k \frac{||\Delta x^0||}{1-\delta^0},$$

we can conclude by induction that the sequence $\{x^k\}$ remains in D^0 and is a Cauchy sequence, so $x^\star$ exists. Finally, $F(x^\star) = 0$ follows from continuity and boundedness of $J(x)^{-1}$ on D^0. □

The local convergence theorem allows some interpretations.

1 The constant ω in (3) is a bound on the nonlinearity of the problem, and its inverse ω^{-1} characterizes the size of the region in which the linearization (2) is an acceptable approximation of F. Hence, condition (4) can also be read as

$$||\Delta x^0|| \leq \frac{\eta}{\omega}, \tag{5}$$

for some constant $\eta < 2$, i.e. the increment step should not exceed this region.

2 In the literature, condition (3) is typically replaced by the two conditions $||J(x)^{-1}|| \leq \beta < \infty$, $||J(y) - J(x)|| \leq \gamma ||y - x||$, $\gamma < \infty$. However, $\beta\gamma$ grossly over-estimates the weaker bound ω.

3 In highly nonlinear problems, though, even for the weaker bound ω one cannot expect the initial guess to be close enough to a solution for condition (4) or (5) to hold. One may rather expect

$$||\Delta x^0|| >> \frac{1}{\omega},$$

in which case convergence of the "full-step" Newton method from x^0 cannot be hoped for.

3. GLOBALIZATION BY UNDERRELAXATION

One way to globalize the convergence of Newton's method is by damping or underrelaxation. The iterates are then defined by

$$x^{k+1} = x^k + t^k \Delta x^k, \quad t^k \in]0, 1],$$

where t^k is a relaxation factor, also called the stepsize. The stepsize t^k is chosen such that the next iterate x^{k+1} is "better" than x^k. It is determined by a line search with respect to an appropriate "merit function" or "level function" $T(x)$.

Any piecewise continuously differentiable level function which satisfies the compatibility condition

$$\Delta x^k \neq 0 \;\Rightarrow\; \frac{d}{d\varepsilon} T(x^k + \varepsilon \Delta x^k)\Big|_{\varepsilon = 0+} < 0$$

is appropriate to ensure global convergence when the Jacobians are bounded away from singularity. (Note, that at a minimum $x^\star$ of a compatible level function $\Delta x^\star = 0$, hence $F(x^\star) = 0$.) The classical choice of a merit function for the underrelaxed Newton method is

$$T(x) := ||F(x)||_2^2$$

in any suitably scaled Euclidean norm of $F(x)$. For an exact or approximate line search for this level function one easily shows the property:

Lemma 1 (global convergence of damped Newton) *Assume that the level set*

$$N_\alpha := \{x \quad | \quad ||F(x)||_2 \leq \alpha\}$$

is compact and is contained in D, that F is twice continuously differentiable and that $J(x)$ is nonsingular on N_α. Then for all $x^0 \in N_\alpha$ there exist a stepsize sequence $\{t^k\}$ and $x^\star \in N_\alpha$ such that $x^k \to x^\star$ $(k \to \infty)$ with $F(x^\star) = 0$.

However, it is well known that already in mildly ill-conditioned problems such a stepsize strategy may be very inefficient since it may produce small stepsizes even in the domain where the full-step Newton's method converges according to Theorem 1. The reason is the following. In ill-conditioned cases, the Newton increment

$$\Delta x^k = -J(x^k)^{-1} F(x^k)$$

may be nearly orthogonal to the steepest descent direction

$$-\nabla T(x^k) = -2J(x^k)^T F(x^k),$$

so that enforcing descent of the level function leads to very small stepsizes. This is due to the fact that with high probability the cosine of the angle between two directions

$$\cos(\Delta x^k, -\nabla T) = \frac{F(x^k)^T F(x^k)}{||J(x^k)^{-1} F(x^k)||\ ||J(x^k)^T F(x^k)||} \geq \frac{1}{\text{cond } J(x^k)}$$

will actually be near its lower bound $(\text{cond } J(x^k))^{-1}$.

Example (Rosenbrock-type)
Let us consider the system of two nonlinear equations

$$F(x) = 0,\ F(x) = \begin{pmatrix} x_1/\sigma_1 \\ \left(x_2 + \frac{1}{200}(x_1 - 50)^2\right)/\sigma_2 \end{pmatrix},\quad \sigma_1 = 1,\ \sigma_2 = \frac{1}{50},$$

with the initial estimate $x^0 = (50, 1)^T$ and the solution $x^\star = (0, -12.5)^T$. In this example, the condition number of the Jacobian $J(x)$ is near 50 for all $x \in R^2$, which is very moderate compared to the practical nonlinear equations appearing typically in BVP. One can easily check that the conditions of Theorem 1 hold. For $D = R^2$ the curvature is bounded by $\omega \leq 0.01$. Convergence for the full step method is guaranteed in $[-100, 100] \times [-100, 100]$, where $\delta(x) := ||J(x)^{-1} F(x)||\omega/2 < 1$. For the initial point x^0 we have $\Delta x^0 = -(50, 1)^T$, and the estimate $\delta^0 \leq 50.01/200$ holds. The first iteration provides $x^1 = (0, 0)^T$ and the a priori estimate $||x^1 - x^\star|| \leq 17$ holds. The application of damped Newton with $||F(x)||_2^2$ as level function, however, gives the stepsize $t^0 \approx 0.077$ as the

optimal relaxation factor. Indeed, the direction of the steepest descent of the level function T at x^0, namely $-\nabla T(x^0)^T = -2F(x^0)^T J(x^0) = -100(1, 50)$, is almost orthogonal to the search direction Δx^0, the cosine of the angle between them being:

$$\frac{-\nabla T(x^0)^T \Delta x^0}{||\nabla T(x^0)||\ ||\Delta x^0||} \approx 0.040 \text{ (which corresponds to 87.71 degrees).}$$

Thus, although the local contraction conditions are fullfilled quite well, the slight nonlinearity together with the mild ill-conditioning of the problem leads to very small stepsizes.

3.1. NATURAL LEVEL FUNCTIONS

To avoid this effect, two different ways can be followed.

Modification of the search direction. The classical way is the Levenberg-Marquardt or trust region variant. Here, the search direction is replaced by

$$\Delta x^k(\gamma) = -(J(x^k)^T J(x^k) + \gamma I)^{-1} J(x^k)^T F(x^k),$$

i.e. it is turned towards $-\nabla T(x^k)$ for large γ.

Modification of the level function. Recall that any level function of the type

$$T_A(x) = ||AF(x)||_2^2,\ A \text{ nonsingular},$$

is compatible with Newton's method. The special choice $A := J(x^k)^{-1}$ yields a level function, called the "natural level function" by Deuflhard [8], [9], with some distinctive properties.

Lemma 2 (Deuflhard [8], [9])

1 *At the iterate x^k, the Newton direction Δx^k is the steepest descent direction of $T_A(x)$ with the choice $A := J(x^k)^{-1}$.*

2 *The level function is "affine invariant", i.e. invariant with respect to any affine transformation $F(x) \to BF(x)$, where B is a nonsingular matrix.*

Proof. *For $A = J(x^k)^{-1}$ the two vectors*

$$-\nabla T_A = -2J(x^k)^T A^T A F(x^k) \quad \textit{and} \quad \Delta x^k = -J(x^k)^{-1} F(x^k)$$

are obviously collinear, and for all B

$$||J(x^k)^{-1}F(x)||_2^2 = ||(BJ(x^k))^{-1}BF(x)||_2^2. \qquad \square$$

3 If the sequence $\{x^k\}$ *converges to a solution* $x^\star$*, then we have*

$$||J(x^k)^{-1}F(x)||_2 = ||x-x^\star||_2 \left(1 + O(||x-x^\star||_2) + O(||x^k-x^\star||_2)\right).$$

From the properties of Lemma 2 one may expect, that the "natural level function" approach should not suffer from the drawbacks demonstrated by the example. It may in fact be viewed as a (local) rescaling of F to AF, such that the condition number of $AJ(x)$ is optimal – namely 1. Figure 1 shows the steepest descent directions and contour lines for both the classical and natural level functions in the case of the Rosenbrock-type example.

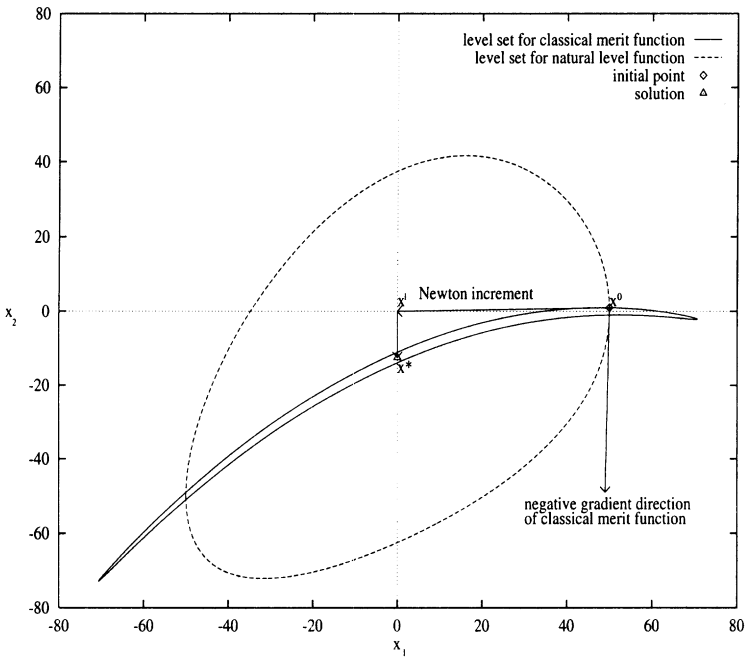

Figure 1 Rosenbrock-type example: natural level function vs classical merit function

Step size procedures based on natural level functions were first introduced by Deuflhard [8], based on a Goldstein type strategy. In Deuflhard [9], refined predictor-corrector strategies for approximate line searches were introduced, and combinations with rank-reduction strategies and

quasi-Newton modifications of the Jacobian were studied. The numerical results given by Deuflhard [8], [9], and also by other authors who adopted and modified the natural level function approach, showed very good practical results in difficult applications (e.g., Ascher et al [1], Bock [3], [4], [5], and Nowak and Weimann [10]).

A major deficiency of the approach, however, is the fact that the change of level functions in each step prevents the classical descent arguments of global convergence proofs to hold. Hence no global convergence proof has been given up to now. On the contrary, similar to the Chamberlain [6] result on cycling for SQP methods using the l_1-penalty level function, examples were constructed by Ascher and Osborne [2] and by Plitt [12] that even showed the existence of two-cycles.

4. THE RESTRICTIVE MONOTONICITY TEST (RMT)

In the following, we will derive a more restrictive stepsize strategy than exact or approximate line searches on the natural level function, which is a slight modification of techniques already successfully used in practice [5].

We will first show that these techniques may be interpreted as stepsize strategies, analogous to those used in numerical methods for the discretization of ODE with invariants, thus offering a global convergence argument which is not based on descent properties. In a second step, we will discuss modifications and extensions of damped Newton method that ensure global convergence as well as efficiency.

For the natural level function in step k, we establish the following quadratic bound that will provide a descent property.

Lemma 3 (Quadratic Upper Bound, where Δx^k is the Newton direction)

$$||J(x^k)^{-1}F(x^k + t\Delta x^k)|| \leq \left(1 - t + \frac{t^2}{2}\omega_1(t)||\Delta x^k||\right) ||J(x^k)^{-1}F(x^k)||, \tag{6}$$

where $\omega_1(t) = \sup_{0<s\leq t} \dfrac{||J(x^k)^{-1}\left(J(x^k + s\Delta x^k) - J(x^k)\right)||}{s||\Delta x^k||}$.

Proof.

$$\begin{aligned} & ||J(x^k)^{-1}F(x^k + t\Delta x^k)|| - (1-t)||J(x^k)^{-1}F(x^k)|| \\ \leq\ & ||J(x^k)^{-1}\Big(F(x^k + t\Delta x^k) - F(x^k) + tF(x^k) \\ & \qquad -tJ(x^k)\Delta x^k + tJ(x^k)\Delta x^k\Big)|| \\ =\ & ||J(x^k)^{-1}\Big(F(x^k + t\Delta x^k) - F(x^k) - tJ(x^k)\Delta x^k\Big)|| \end{aligned}$$

$$= ||J(x^k)^{-1} \int_0^t \Big(J(x^k + s\Delta x^k) - J(x^k) \Big) \Delta x^k \, ds||$$

$$\leq \omega_1(t) \, ||\Delta x^k||^2 \int_0^t s \, ds = \frac{t^2 \omega_1(t) \, ||\Delta x^k||^2}{2}.$$

From the previous relations it follows that

$$||J(x^k)^{-1} F(x^k + t\Delta x^k)|| \leq (1-t)||\Delta x^k|| + \frac{t^2}{2}\omega_1(t) \, ||\Delta x^k||^2$$

$$= \left(1 - t + \frac{t^2}{2}\omega_1(t) \, ||\Delta x^k||\right) ||J(x^k)^{-1} F(x^k)||. \qquad \square$$

Since $\omega_1(t)$, $t \geq 0$, is monotonically nondecreasing, one may choose the damping factor $t^k = t^k(\eta)$ in terms of this quadratic upper bound such that

$$t = \max! \quad \text{s.t.} \quad t \leq 1, \quad t\omega_1(t)||\Delta x^k|| \leq \eta,$$

for some prescribed $\eta < 2$. This means that we choose $t^k \leq 1$ maximal such that the "Restricted Monotonicity Test" (RMT)

$$t^k ||\Delta x^k|| \leq \min \left(\frac{\eta}{\omega_1(t^k)}, ||\Delta x^k|| \right), \tag{7}$$

is fulfilled. RMT (7) together with Lemma 3 ensures that the weaker traditional Armijo-type descent condition

$$||J(x^k)^{-1} F(x^k + t^k(\eta)\Delta x^k)|| \leq \left(1 - t^k(\eta)(1 - \frac{\eta}{2})\right) ||J(x^k)^{-1} F(x^k)|| \tag{8}$$

holds. Note, that $t^k(\eta)$ for $\eta = 1$ would minimize the right hand side of QUB (6) with respect to t if $\omega_1(t)$ were constant, or replaced by an upper bound.

Remark The RMT ensures that the actual length $t\Delta x$ does not exceed the $1/\omega$ region in which $J(x^k)$ is a valid approximation of $J(x)$ according to the definition of ω!

Similar to Lemma 3, one can show more generally, that

Lemma 4

$$||AF(x^k + t\Delta x^k)|| \leq \left(1 - t + \frac{t^2}{2}\omega_2^A(t) \, ||\Delta x^k||\right) ||AF(x^k)||,$$

where $\omega_2^A(t) = \sup_{0<s\le t} \dfrac{||A\left(J(x^k + s\Delta x^k) - J(x^k)\right)||}{s||AF(x^k)||}$.

Lemma 5 *Assume we choose* $\eta < 1$ *and* t^k *such that*

$$t^k \omega_1(t^k)\,||\Delta x^k|| \le \eta < 1. \tag{9}$$

Then

$$t^k \omega_2^A(t^k)\,||\Delta x^k|| \le \eta \frac{1+\eta}{1-\eta}$$

for all $A \in \{J(x^k + s\Delta x^k)^{-1} \mid s \le t^k\}$.

Proof. Let $J_0 = J(x^k)$ and $A = J(x^k + s\Delta x^k)^{-1}$ for some s, $0 < s \le t^k$. By definition

$$\omega_2^A(t^k) = \sup_{0<\tau\le t^k} \left(\frac{||A\left(J(x^k + \tau\Delta x^k) - J_0\right)||}{s||AF(x^k)||} \frac{||J_0^{-1}F(x^k)||}{||J_0^{-1}F(x^k)||} \right)$$

$$\le \sup_{0<\tau\le t^k} \left(\frac{||AJ_0||\;||J_0^{-1}\left(J(x^k+\tau\Delta x^k)-J_0\right)||}{s||AF(x^k)||} \frac{||J_0^{-1}A^{-1}||\;||AF(x^k)||}{||J_0^{-1}F(x^k)||} \right)$$

$$\le \omega_1(t^k)\,||AJ_0||\;||(AJ_0)^{-1}||.$$

Moreover, the choice of A, the definition of $\omega_1(t^k)$ and condition (9) imply

$$||J_0^{-1}(A^{-1} - J_0)|| \le s\omega_1(s)\,||\Delta x^k|| \le \eta < 1,$$

which gives the classical estimates

$$||AJ_0|| = ||\left(I - J_0^{-1}(J_0 - A^{-1})\right)^{-1}|| \le \frac{1}{1-\eta},$$

$$||J_0^{-1}A^{-1}|| = ||I + J_0^{-1}(A^{-1} - J_0)|| \le 1 + \eta.$$

It follows that

$$\omega_2^A(t^k) \le \omega_1(t^k)\frac{1+\eta}{1-\eta},$$

so another application of (9) provides the required result. □

Lemma 4 and Lemma 5 imply that, if η is chosen to satisfy

$$\eta\frac{1+\eta}{1-\eta} < 2, \quad \text{i.e.} \quad \eta < \frac{1}{2}\left(\sqrt{17} - 3\right),$$

then all level functions for intermediate choices of Jacobians also descend. In particular, two-cycles are impossible. For example, $\eta = 1/2$ yields the property

Lemma 6 (No Two-Cycles when $\eta = 1/2$) *For all* k

$$||J(x^k)^{-1}F(x^{k+1})|| \leq (1 - \frac{3t^k}{4})||J(x^k)^{-1}F(x^k)|| \qquad (10)$$

$$||J(x^{k+1})^{-1}F(x^{k+1})|| \leq (1 - \frac{t^k}{4})||J(x^{k+1})^{-1}F(x^k)|| \qquad (11)$$

hence

$$||J(x^{k+1})^{-1}F(x^{k+2})|| \leq (1 - \frac{3t^{k+1}}{4})(1 - \frac{t^k}{4})||J(x^{k+1})^{-1}F(x^k)||,$$

so that $x^k \neq x^{k+2}$.

Proof. Due to the choice $\eta = 1/2$, inequality (10) follows immediately from (8), and inequality (11) from Lemmas 4 and 5 in the case $A = J(x^{k+1})^{-1}$. □

Note, however, that this result still does not prove global convergence.

5. PRACTICAL REALIZATION OF THE RMT

The costly evaluation of $\omega_1(t)$ can be avoided. In the quadratic upper bound of Lemma 3, one can replace $\omega_1(t)$ by the weaker estimate for the curvature

$$\begin{aligned}\omega_3(t) &:= \frac{2||J(x^k)^{-1}\left(F(x^k + t\Delta x^k) - (1-t)F(x^k)\right)||}{t^2||\Delta x^k||^2} \\ &= \frac{2||J(x^k)^{-1}\int_0^t \left(J(x^k + s\Delta x^k) - J(x^k)\right)\Delta x^k\, ds||}{t^2||\Delta x^k||^2}.\end{aligned}$$

The estimate $\omega_3(t)$ is easy to evaluate. Indeed, it involves only the calculation of

$$\Delta x^k = -J(x^k)^{-1}F(x^k),$$

which is necessary anyway, and of

$$\bar{\Delta x}^k = -J(x^k)^{-1}F(x^k + t\Delta x^k),$$

just as in a line search procedure for the natural level function. However, instead of a one dimensional minimization, and analogously to the more restrictive test (7), we require the conditions

$$\eta_\star \leq t^k\omega_3(t^k)||\Delta x^k|| \leq \eta^\star \qquad (12)$$

with $\eta_\star < \eta < \eta^\star$. In the numerical tests of Section 9, $\eta = 1$, $\eta_\star = 0.8\eta$ and $\eta^\star = 1.2\eta$ were used.

As $\omega_3(t)$ is continuous, a simple rootfinding procedure for $t\omega_3(t)||\Delta x^k|| - \eta = 0$ is applied to satisfy (12). A good starting value for t^k is provided by the curvature estimate $\omega_3(t^{k-1})$ of the previous iteration, namely

$$t^k_{\text{start}} := \min\left(1, \frac{\eta}{\omega_3(t^{k-1})||\Delta x^k||}\right).$$

Thus, according to our experience, at most two F-evaluations per iteration are required.

This restrictive monotonicity test works very well in practical applications. Although the rigorous proof of Lemma 6 does not hold for the weaker curvature measure used here, cycling does not occur for the Ascher–Osborne example, and has not been observed in practical applications. It is hoped that a similar proof of non-cycling, possibly for sharper η, can be found.

Nevertheless, we keep in mind that even Lemma 6 does not provide a global convergence proof for either version (7) or (12) of the RMT based on descent arguments.

6. A DIFFERENT INTERPRETATION OF THE RMT

Let us consider, instead of a single mapping F, a family $H : D \times [0,1] \in R^{n+1} \to R^n$ depending on a parameter λ such that

$$H(x^0, 0) = 0, \; H(x, 1) = F(x),$$

where x^0 is some initial solution ($\lambda = 0$). For example, we can set

$$H(x, \lambda) = F(x) - (1 - \lambda)F(x^0). \tag{13}$$

Let us consider the equation

$$H(x(\lambda), \lambda) \equiv 0, \quad \lambda \in [0, 1]. \tag{14}$$

Under certain assumptions (see [11]), e.g. those of Lemma 1, (14) defines a unique continuously differentiable function $x(\lambda)$ (homotopy path), and $x(\lambda)$, $\lambda \in [0,1]$, satisfies the implicit, so-called Davidenko differential equations [7]

$$\left[\frac{\partial H(x(\lambda), \lambda)}{\partial x}\right] \dot{x} = -\frac{\partial H(x(\lambda), \lambda)}{\partial \lambda}, \quad \forall \lambda \in [0,1], \; x(0) = x^0.$$

For the choice (13), the last expression takes the form

$$\dot{x} = -J(x)^{-1}F(x^0) = -\frac{1}{1-\lambda}J(x)^{-1}F(x), \quad \forall\lambda \in [0,1],\ x(0) = x^0. \tag{15}$$

Let us introduce the change of variable $\lambda = 1 - e^{-\tau}$. Then τ varies from 0 to $+\infty$ as λ varies from 0 to 1, and (13) has the form

$$H(x,\tau) = F(x) - e^{-\tau}F(x^0), \quad \tau \in [0,+\infty). \tag{16}$$

The differential equation corresponding to (16) is given by

$$\frac{dx}{d\tau} = -J(x)^{-1}F(x), \quad \forall\tau \in [0,+\infty),\ x(0) = x^0. \tag{17}$$

We can consider the problem of constructing the trajectory $x(\lambda)$ of (15) (or $x(\tau)$ of (17)) as numerical integration of the ODE (15) (or (17)). If we integrate (17) by Euler's method with stepsizes t^k we obtain

$$x^{k+1} = x^k - t^k J(x^k)^{-1}F(x^k), \quad k = 0,1,...,$$

which is the damped Newton method. Thus, we can view the damped Newton method as an Euler approximation of the continuous Newton equation (17).

Next we will show that the RMT is nothing but a stepsize control for Euler's method applied to the Davidenko differential equations. First we estimate the local integration error. For simplicity we only consider the first iteration step.

Lemma 7 *The local integration error of Euler's method applied to (15) is given by*

$$\varepsilon(t) := -J(x^0)^{-1}\left(F(x^0 + t\Delta x^0) - (1-t)F(x^0)\right) + O(t^3).$$

Proof. The local error is defined as

$$\varepsilon(t) = x(t) - x^0 - t\Delta x^0,$$

where $x(t)$ satisfies the invariant

$$F(x(t)) = (1-t)F(x^0). \tag{18}$$

From a Taylor series expansion of (18) we have

$$\begin{aligned}
\dot{x}(0) &= -J(x^0)^{-1}F(x^0) = \Delta x^0, \\
\ddot{x}(0) &= -J(x^0)^{-1}\left(\frac{\partial}{\partial x}J(x^0)\dot{x}\right)\dot{x} \\
&= -\frac{2}{t^2}J(x^0)^{-1}\left(F(x^0 + t\Delta x^0) - (1-t)F(x^0)\right) + O(t),
\end{aligned}$$

and $x(t) = x^0 + t\Delta x^0 + \frac{t^2}{2}\ddot{x}(0) + O(t^3)$, which imply the required result.□

From the new point of view, the RMT

$$\begin{aligned} ||t\Delta x^0|| \frac{\eta_\star}{2} &\leq ||J(x^0)^{-1}\left(F(x^0 + t\Delta x^0) - (1-t)F(x^0)\right)|| \\ &\leq \frac{\eta^\star}{2}||t\Delta x^0|| \end{aligned} \tag{19}$$

is simply a stepsize control for Euler's method. By Lemma 7 the term controlled in formula (19) is an asymptotically correct estimate of the local integration error. It is kept small compared to the increment norm, which is controlled by the choice of η, in order to ensure that the Newton path is followed with a desired accuracy.

We can go one step further, if we take into account that (15) is an implicit ordinary differential equation with known invariant given by equation (18). Similar to techniques used in discretization methods for ODE or DAE with invariants, e.g. [13], we can therefore exploit the invariant for a stabilization step, which is a "back projection" of

$$x^{1,0} := x^0 + t\Delta x^0$$

to the invariant manifold, which is a curve in our case. This step can be performed by adding the correction term already computed for the RMT

$$\begin{aligned} x^{1,1} &:= x^{1,0} + \tilde{\Delta}x^0 = x^0 + t\Delta x^0 + \tilde{\Delta}x^0, \\ \tilde{\Delta}x^0 &:= -J(x^0)^{-1}\left(F(x^0 + t\Delta x^0) - (1-t)F(x^0)\right). \end{aligned}$$

Unlike Euler's method, which is of first order, the combined two step method is of second order. Note, that as soon as $t = 1$ is reached, the additional back projection step extends the quadratically convergent Newton's method to a well-known cubically convergent modification.

In terms of Newton's method, the combined scheme can be interpreted as the first two steps of a simplified full step Newton method to solve $F(x) - (1-t)F(x^0) = 0$, starting from x^0. The RMT then plays the role of a monitoring test to choose t sufficiently small, in order that contractivity (by $\eta^\star/2$) of this scheme is guaranteed. This projection could of course be repeated until the Newton path is met with a desired accuracy. A natural and necessary extension of the RMT is then to check sufficient contractivity of the iterations e.g.

$$\frac{||J(x^0)^{-1}\Big(F(x^{1,i+1}) - (1-t)F(x^0)\Big)||}{||J(x^0)^{-1}\Big(F(x^{1,i}) - (1-t)F(x^0)\Big)||} \leq \frac{\eta^{\star\star}}{2}, \quad \eta^\star \leq \eta^{\star\star} < 2.$$

If insufficient contractivity occurs, then starting over with reduced η (hence t) seems to be preferable to additional damping during "back projections" or to re-computing the Jacobian.

Remark. It can be shown that the stepsize strategy given here eventually leads to a full step method when the local convergence conditions of Theorem 1 are satisfied.

7. VARIATIONS OF THE DAMPED NEWTON METHOD

The interpretation as an error controlled integration method for an ODE with invariants allows various modifications and variations of the basic method.

Basic strategy

The basic strategy, which was used for the numerical computations presented below, is as follows:

1 compute the Newton direction Δx^k,

2 compute $x^{k+1} = x^k + t^k \Delta x^k$, where t^k satisfies the RMT as error control,

3 restart the homotopy path (equivalently, continue integration) from x^{k+1}.

Basic strategy with back projections

A more expensive strategy which needs one (or more) additional F-evaluation(s) is:

1 compute the Newton direction Δx^k,

2 compute $x^{k+1,0} = x^k + t^k \Delta x^k$, where t^k satisfies the RMT,

3 add one (or more) back projection step(s)

$$\begin{aligned} x^{k+1,i+1} &= x^{k+1,i} + \tilde{\Delta} x^{k,i}, \\ \tilde{\Delta} x^{k,i} &= -J(x^k)^{-1} \left(F(x^{k+1,i}) - (1 - t^k) F(x^k) \right), \end{aligned}$$

4 restart the homotopy path from the last $x^{k+1,i+1}$.

In this variant we first have to ensure local convergence to the Newton path using the convergence behaviour of the back projections for a reduction strategy for η, hence also for the stepsizes t. From this then

follows global convergence under certain assumptions like nonsingularity of Jacobians along the Newton path, since along this path all level functions $||AF(x)||_2^2$, with A nonsingular, descend by a factor of $1-t$. A termination criterion could be based on the latter property. In our numerical tests, however, repeated back projections were not found to be superior to the basic strategy. Apparently, the extra effort to iterate back to the continuous Newton path does not necessarily lead to a better iterate x^{k+1}, even though it guarantees global convergence.

Similar to techniques used in discretization methods for ODE with invariants, one can of course consider constructing higher order methods for integration of the Newton path. A few comments from the numerical ODE point of view may be made.

1 Since a highly accurate solution of the Newton path is unlikely to be necessary except maybe in extreme cases of ill-conditioning, low order methods should be most efficient.

2 The Davidenko equation is an implicit ODE and should be treated as such. In order to avoid frequent expensive unnecessary and possibly inaccurate evaluations of $J(x)^{-1}$, the use of higher order Runge-Kutta methods is not recommended.

3 Since back projection to the invariant curve effectively inhibits error propagation, consistency error considerations are sufficient for the construction of integration methods. Suitable candidates may be multistep methods based on solution values and occasional derivative evaluations.

A modification worth investigating may be to vary the steps t^k with the index i, e.g., as in the second order variant

$$\begin{aligned} x^{k+1,0} &= x^k + t^{k,0}\Delta x^k \\ \tilde{\Delta} x^{k+1,0} &= -J(x^k)^{-1}\left(F(x^{k+1,0}) - (1-t^{k+1,0})F(x^k)\right), \\ x^{k+1,1} &= x^k + t^{k,1}\Delta x^k + \frac{(t^{k,1})^2}{(t^{k,0})^2}\tilde{\Delta} x^{k+1,0}. \end{aligned}$$

8. EXTENSIONS TO L_2- AND L_1-PARAMETER ESTIMATION

Parameter estimation problems in dynamic processes can be expressed as nonlinearly constrained optimization problems in the general form

$$\begin{aligned} \min_x \quad & ||F_0(x)||_\nu, \quad \nu \in \{1,2\}, \\ & F_1(x) = 0, \quad F_2(x) \geq 0, \end{aligned} \tag{20}$$

where the cost functional is the l_2- or l_1-norm of the vector function $F_0(x)$. The vector of variables $x = (y, p)$ consists of "state" variables $y \in R^{n_y}$, typically discretization variables for underlying initial or boundary value problems in ODEs or DAEs, and unknown parameters p to be estimated.

Traditionally, the problem (20) is solved by the constrained Gauss–Newton method [3], according to which a new iterate is given by

$$x^{k+1} = x^k + t^k \Delta x^k,$$

where the increment Δx^k solves the linearly constrained problem

$$\begin{aligned} \min_x \quad & ||F_0(x^k) + J_0(x^k)\Delta x^k||_\nu, \quad \nu \in \{1, 2\}, \\ & F_1(x^k) + J_1(x^k)\Delta x^k = 0, \\ & F_2(x^k) + J_2(x^k)\Delta x^k \geq 0. \end{aligned} \tag{21}$$

In both cases the solution Δx^k of (21) can be represented in the form $\Delta x^k = -J^+(x^k)F(x^k)$, where $J(x^k)^+$ is a generalized inverse, i.e. it satisfies the condition

$$J^+JJ^+ = J^+, \quad J = \frac{\partial F}{\partial x}, \quad F = \begin{pmatrix} F_0 \\ F_1 \\ F_2 \end{pmatrix}.$$

For example, for the unconstrained l_2-case, $J^+(x^k)$ is the Moore–Penrose inverse of $J(x^k)$. The l_1-solution interpolates some of the measurements, i.e. some components of the linearized function

$$F_0(x) + J_0(x)\Delta x$$

are equal to zero ("active"). Under certain conditions, the generalized inverse of J is then the inverse of a projection of $J(x^k)$, which contains the active measurements, equality constraints and active inequality constraints.

Using the generalized inverse, the Gauss–Newton method becomes

$$x^{k+1} = x^k + t^k \Delta x^k, \quad \Delta x^k = -J^+(x^k)F(x^k). \tag{22}$$

With these preparations, we can formally extend our quadratic upper bound and restrictive monotonicity test to $||J^+(x^k)F(x^k + t\Delta x^k)||$. The Gauss–Newton method (22) can be then interpreted as a stepsize control for the Euler method applied to the continuous Gauss–Newton method. Note, however, that the active inequality constraints incorporated in $J^+(x^k)$ as well as the active measurements are changing along the solution of the continuous problem, so that the stepsize strategy must be combined with an additional monitoring of the changing active sets.

9. NUMERICAL RESULTS

With the new stepsize strategies two challenging test problems were treated. The optimal control problem of the re-entry of an Apollo spacecraft is known for its hard nonlinearities, due to the aerodynamic forces and has a very small region of feasible solutions [14]. In the estimation of the reaction constants in the nonlinear differential equation modelling the denitrogenization of pyridine, ill-conditioning and complicated stability problems for poor initial guesses of the parameters occur. The results show the potential of the new approach.

9.1. RE-ENTRY PROBLEM [14]

In this problem a control has to be chosen to minimize the heating of a space vehicle during the flight through the earth's atmosphere on the way back from the outer space. Numerical difficulties are caused by extreme instability properties due to the aerodynamic forces when entering the atmosphere. Convergence using Newton's method can be expected only if the initial guess is fairly close to the solution.

Applying the maximum principle to this optimal control problem results in a boundary value problem in the states v, γ, ξ, the adjoints λ_v, λ_γ, λ_ξ and the free final time T:

$$
\begin{aligned}
\dot{v} &= \left(-\frac{S\rho v^2}{2m}C_W(u) - \frac{g\sin\gamma}{(1+\xi)^2}\right)T,\\
\dot{\gamma} &= \left(\frac{S\rho v}{2m}C_A(u) + \frac{v\cos\gamma}{R(1+\xi)} - \frac{g\cos\gamma}{v(1+\xi)^2}\right)T,\\
\dot{\xi} &= \frac{v\sin\gamma}{R}T,\\
\dot{\lambda}_v &= \left(30v^2\sqrt{\rho} + \lambda_v\frac{S\rho v}{m}C_W(u)\right.\\
&\quad \left.-\lambda_\gamma\left\{\frac{S\rho}{2m}C_A(u) + \frac{\cos\gamma}{R(1+\xi)} + \frac{g\cos\gamma}{v^2(1+\xi)^2}\right\} - \lambda_\xi\frac{\sin\gamma}{R}\right)T,\\
\dot{\lambda}_\gamma &= \left(\lambda_v\frac{g\cos\gamma}{(1+\xi)^2} + \lambda_\gamma\left\{\frac{v\sin\gamma}{R(1+\xi)} - \frac{g\sin\gamma}{v(1+\xi)^2}\right\} - \lambda_\xi\frac{v\cos\gamma}{R}\right)T,\\
\dot{\lambda}_\xi &= \left(-5\beta Rv^3\sqrt{\rho} - \lambda_v\left\{\frac{\beta RS\rho v^2}{2m}C_W(u) + \frac{2g\sin\gamma}{(1+\xi)^3}\right\}\right.\\
&\quad \left.+\lambda_\xi\left\{\frac{\beta RS\rho v}{2m}C_A(u) + \frac{v\cos\gamma}{R(1+\xi)^2} - \frac{2g\cos\gamma}{v(1+\xi)^3}\right\}\right)T,
\end{aligned}
$$

with boundary conditions

$$v(0) = 0.36, \quad \gamma(0) = -8.1\pi/180, \quad \xi(0) = 4/R,$$
$$v(T) = 0.27, \quad \gamma(T) = 0, \quad \xi(0) = 2.5/R,$$
$$-10v(T)^3\sqrt{\rho} + \dot{v}(T)\lambda_v(T) + \dot{\gamma}(T)\lambda_\gamma(T) + \dot{\xi}(T)\lambda_\xi(T) = 0,$$

where $\rho_0 = 0.002704$, $R = 209.$, $\beta = 4.26$, $C_W(u) = 1.174 - 0.9\cos u$, $C_A(u) = 0.6\sin u$, $S/2m = 26.600$, $g = 3.2172 \times 10^{-4}$, and the control u is given by

$$\sin u = \frac{0.6\lambda_\gamma}{w}, \quad \cos u = \frac{0.9v\lambda_v}{w}, \quad w = \sqrt{(0.6\lambda_\gamma)^2 + (0.9v\lambda_v)^2}.$$

We parametrized this boundary value problem with a multiple shooting technique, using 6 equidistantly distributed nodes. The initial guesses were generated according to a technique described in [14]. For the solution of the resulting system of nonlinear equations with 37 variables Newton's method using the basic stepsize strategy was used. The solution (relative accuracy 10^{-4}) was achieved after 8 iterations, 4 with damped steps. Figure 2 shows the stepsizes in every iteration.

It is difficult to compare these results with ones documented in the literature [14], [8], [9], because there Broyden approximations and finite difference approximations were used. One may say, however, that our results are very competetive with the fastest results so far published.

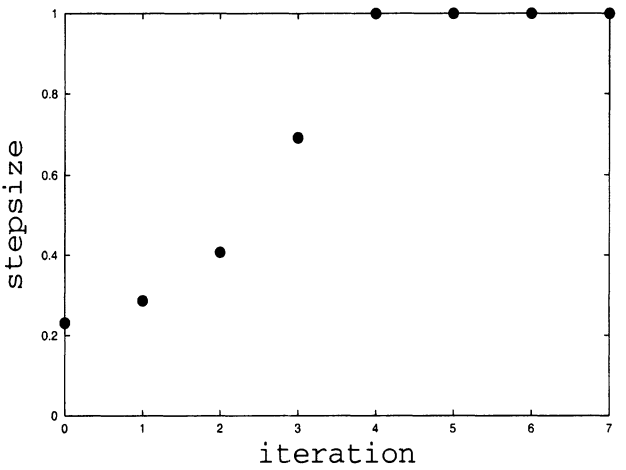

Figure 2 Stepsize history for the re-entry problem

9.2. PARAMETER ESTIMATION IN THE DENITROGENIZATION OF PYRIDINE

This problem (originally due to Zwaga [15]) was investigated in [3]. At first only pyridine is present which initiates a reaction process that can be described by ODEs with 7 state variables, the concentrations of the species:

Pyridine:	$\dot{A} = -p_1 A + p_9 B$
Piperidine:	$\dot{B} = p_1 A - p_2 B - p_3 CB + p_7 D - p_9 B + p_{10} DF$
Pentylamine:	$\dot{C} = p_2 B - p_3 BC - 2p_4 CC - p_6 C + p_8 E + p_{10} DF + 2p_{11} EF$
N-Pentylpiperidine:	$\dot{D} = p_3 BC - p_5 D - p_7 D - p_{10} DF$
Dipentylamine:	$\dot{E} = p_4 CC + p_5 D - p_8 E - p_{11} EF$
Ammonia:	$\dot{F} = p_3 BC + p_4 CC + p_6 C - p_{10} DF - p_{11} EF$
Pentane:	$\dot{G} = p_6 C + p_7 D + p_8 E.$

The rate constants $p_1, p_2, \ldots, p_{11}$ of these ODEs are unknown and have to be estimated from 77 measurements of the states at the times 0.5, 1, ..., 5.5.

We treated this parameter estimation problem with the multiple shooting code PARFIT [3], which has a generalized Gauss–Newton method as a core routine for the solution of the structured constrained least squares problems. For globalization we implemented the basic strategy ($\eta = 1$), replacing the inverse by the generalized inverse, that is the solution operator for the constrained linear least squares problems.

We performed 8 experiments with widely varying initial guesses for the 11 parameters. As initial guesses for the states we chose the measurements. The initial guesses for the parameters $p_i = \alpha, i = 1, ..., 7$, for different α were rather poor guesses, since the true values vary between 0.201 and 29.4. In all cases the algorithm safely converged. The history of stepsizes is shown in Figure 3. The number of iterations, the number of damped steps and the achieved accuracy are given in Table 1.

10. CONCLUSIONS

It is well known that Newton's method for nonlinear equations can be forced to converge globally in a domain where the Jacobian is nonsingular. However, the price one has to pay is unnecessarily small stepsizes in the local convergence domain of the full step method if the problem

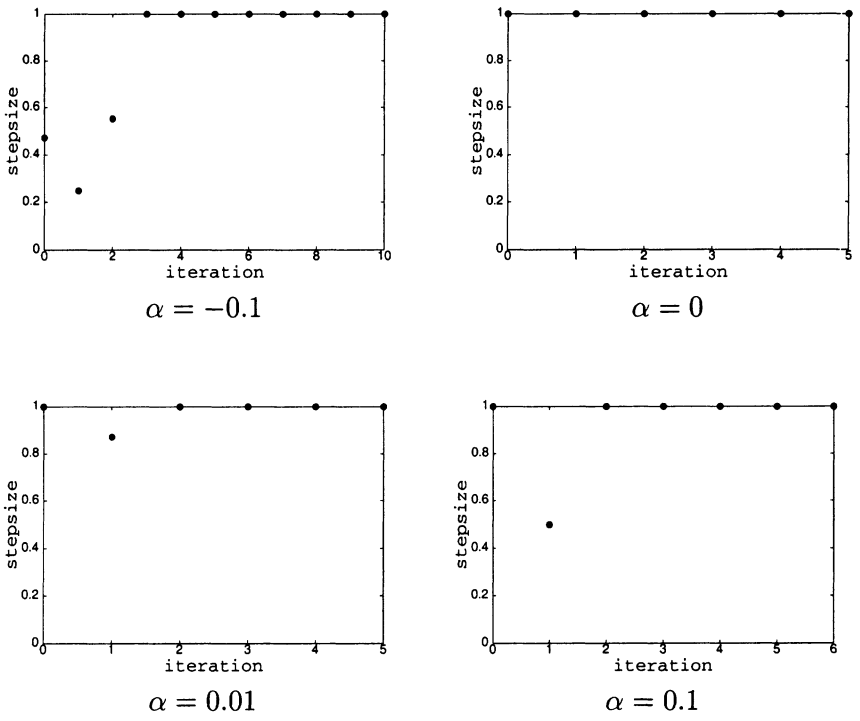

Figure 3 Stepsize histories for the pyridine problem

α	$\|\|\Delta x^0\|\|/\sqrt{n}$	$\|\|\Delta x^\star\|\|/\sqrt{n}$	iterations	damped steps
−0.1	236.04	4.91×10^{-4}	11	3
0.0	73.27	3.86×10^{-5}	6	0
0.01	661.18	1.28×10^{-4}	6	1
0.1	40.09	5.41×10^{-4}	7	1
0.2	16.70	7.82×10^{-5}	8	2
1.0	2.24	3.76×10^{-5}	9	2
5.0	2.65	3.68×10^{-5}	17	11
10.0	66.83	7.64×10^{-5}	27	22

Table 1 Convergence behaviour of the Gauss–Newton method with RMT for the pyridine problem for different initial guesses

is mildly ill-conditioned and nonlinear and one chooses classical merit functions.

The paper presents a new strategy which combines previously defined "natural level functions" with a restrictive monotonicity test. The new

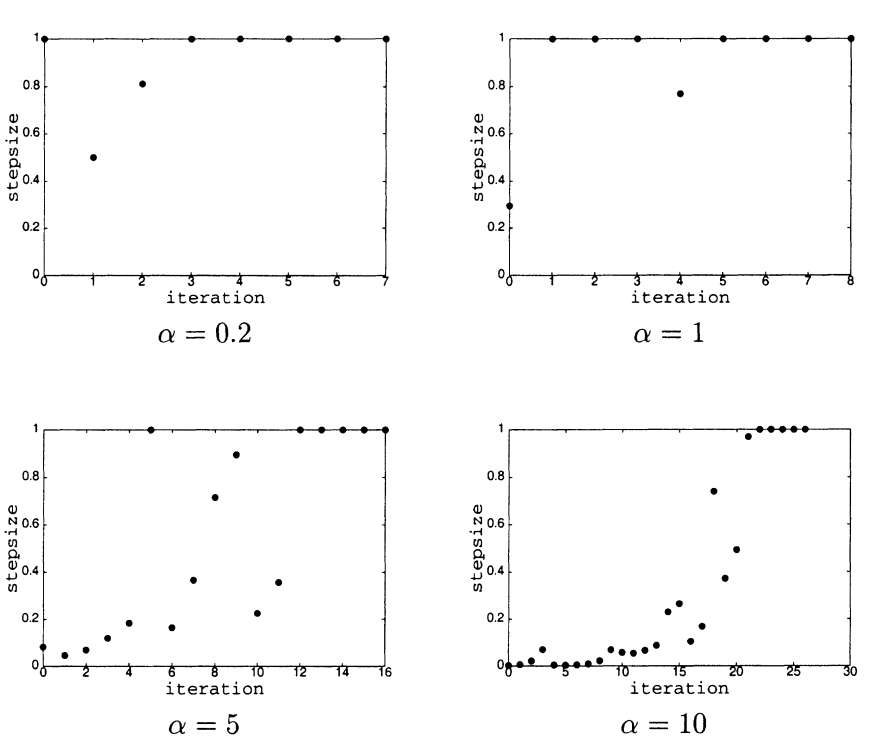

Figure 3 (continued) Stepsize histories for the pyridine problem

global convergence argument is quite different from the classical descent type proof. It is shown that this combination of natural level functions, for which descent proofs do not hold, and the restrictive monotonicity test overcome the problem of choosing too small steps. The new stepsize strategy is viewed as a stepsize control for the continuous Newton method which makes use of an invariant of the Newton path.

Three practical interpretations can be given. The first interpretation is a stepsize control of the continuous Newton method by means of an asymptotically correct estimate of the local error. Secondly, we interprete the damped Newton method as an attempt to solve a relaxed problem with a full step Newton method (with Jacobian kept constant), and the RMT is a test on sufficient contractivity. Thirdly, the stepsize is allowed to go as far as the approximation of the Jacobian is valid.

The second argument suggests generalizations to other stepsize and trust region strategies. Most of them can be interpreted as attempts to solve a relaxed version of the original nonlinear problem. In the spirit of this paper a stepsize (or trust region) strategy can be based on a control whether a sufficient (local) contraction of the method to the solution of

the relaxed problem is achieved. It is hoped that this approach proves to be equally effective in these other areas.

Numerical results to two demanding problems from optimal control and parameter estimation in ODE are given, which are notorious for their strong nonlinearities. They show a very nice and reliable convergence behaviour.

Acknowledgments

We would like to thank M.J.D. Powell and an anonymous referee for their valuable comments and suggestions that helped to improve the paper.

Also, we thank the Deutsche Forschungsgemeinschaft (DFG) for financial support through SFB 539.

References

[1] U. Ascher, R. M. M. Mathheij and R. D. Russell (1988), *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice Hall, Englewood Cliffs, New Jersey.

[2] U. Ascher and M. R. Osborne (1987), A note on solving nonlinear equations and the natural criterion function, *J. Optim. Theory Appl.*, 55, no. 1, pp. 147 – 152.

[3] H. G. Bock (1981), Numerical treatment of inverse problems in chemical reaction kinetics, in K. H. Ebert, P. Deuflhard and W. Jäger, eds, *Modelling of Chemical Reaction Systems*, Springer Series in Chemical Physics 18, Heidelberg.

[4] H. G. Bock (1983), Recent advances in parameter identification techniques for O. D. E., in P. Deuflhard and E. Hairer, eds, *Numerical Treatment of Inverse Problems, Progress in Scientific Computing*, 2, Birkhäuser, Boston, pp. 95 – 121.

[5] H. G. Bock (1987), *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, Bonner Mathematische Schriften, 187, Bonn.

[6] R. M. Chamberlain (1979), Some examples of cycling in variable metric methods for constrained minimization, *Math. Programming*, 16, pp. 378 – 383.

[7] D. Davidenko (1953), On the approximate solution of a system of nonlinear equations (Russian), *Ukrain. Mat. Zh.*, 5, pp. 196 – 206.

[8] P. Deuflhard (1974), A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with applications to multiple shooting, *Num. Math.*, 22, pp. 289 – 315.

[9] P. Deuflhard (1975), A relaxation strategy for the modified Newton method, in R. Bulirsch, W. Oettli and J. Stoer, eds, *Lecture Notes in Math.*, 447, Springer Verlag, pp. 59 – 73.

[10] U. Nowak and L. Weimann (1991), A family of Newton codes for systems of highly nonlinear equations, Technical Report TR-91-10, Konrad-Zuse-Zentrum für Informationstechnik Berlin.

[11] J. M. Ortega and W. C. Rheinboldt (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

[12] K. J. Plitt (1983), Private Communication.

[13] V. H. Schulz, H. G. Bock and M. C. Steinbach (1998), Exploiting invariants in the numerical solution of multipoint boundary value problems for DAE, *SIAM J. Sci. Comp.*, 19, no. 2, pp. 440 – 467.

[14] J. Stoer and R. Bulirsch (1973), *Einführung in die Numerische Mathematik II*, Heidelberger Taschenbuch 114, Springer, Berlin – Heidelberg – New York.

[15] P. Zwaga (1977), Private Communication.