BACKWARD STEP CONTROL FOR GLOBAL NEWTON-TYPE METHODS

ANDREAS POTSCHKA*

Abstract. We present and analyze a new damping approach called *backward step control* for the globalization of the convergence of Newton-type methods for the numerical solution of nonlinear root-finding problems. We provide and discuss reasonable assumptions that imply convergence of backward step control on the basis of generalized Newton paths in conjunction with a backward analysis argument. In particular, convergence to a specific solution and a-priori estimates on the residual reduction can be shown. Furthermore, we can guarantee a transition to full steps in the vicinity of a solution, which implies fast local convergence. We present two algorithmic realizations of backward step control and apply the method to more than one hundred examples, including comparisons with different globalization approaches for the minimization of the Rosenbrock function and to large-scale unconstrained optimization problems from the CUTEst benchmark library using backward step control for an inexact Newton method based on MINRES.

Key words. Newton-type methods, globalization, Newton path, backward analysis

AMS subject classifications. 58C15, 65H10, 65H20, 90C30

1. Introduction.

Let $D \subset \mathbb{R}^n$ be open and $F : D \to \mathbb{R}^n$ be a continuously differentiable function with Jacobian denoted by $J(x) = (\mathrm{d}F/\mathrm{d}x)(x)$. We consider the problem of finding an $x \in D$ such that

$$F(x) = 0. \tag{1.1}$$

Without doubt, (1.1) is one of the fundamental problems in numerical analysis. The most popular computational approach is to use Newton-type methods: Given an initial guess $x_0 \in D$, a continuous approximation $M(x)$ of the inverse $J^{-1}(x)$ of the Jacobian, and a sequence of step sizes (t_k) with $t_k \in (0, 1]$ for $k \in \mathbb{N}$, we iteratively compute the sequence (x_k) according to

$$x_{k+1} = x_k + t_k \Delta x_k, \quad \text{with } \Delta x_k = -M(x_k)F(x_k). \tag{1.2}$$

Usually, $M(x)$ is computed with partial or even complete knowledge of $J(x)$, but other iterative methods that evaluate $F(x_k)$ can also be cast in this form, e.g., a fixed-point iteration $x_{k+1} = x_k - t_k F(x_k)$ with $M(x) = \mathrm{I}_n$ and small $t_k > 0$. Often, $M(x)$ is implicitly given by a few iterations of an iterative method for the solution of $J(x)\Delta x_k = -F(x_k)$. These approaches are called inexact Newton methods (see, e.g., [13]) and we give an example of a general approach of investigating inexact Newton methods as Newton-type methods in Section 12.3.

It is well known that under certain conditions, iteration (1.2) with $t_k = 1$ and $M(x) = J^{-1}(x)$ converges quadratically provided that $x_0 \in D$ is chosen close enough to a solution $x^* \in D$ satisfying $F(x^*) = 0$. For a historical survey about local convergence theorems for this so-called full step Newton method, the reader is referred to [25, 16].

In practice, there are two main challenges: First, it might be more efficient to achieve a certain accuracy for (1.1) using an M that only approximately satisfies

*Interdisciplinary Center for Scientific Computing, Heidelberg University, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany (potschka@iwr.uni-heidelberg.de). We gratefully acknowledge support by the EU through S. Engell's and H.G. Bock's ERC Advanced Investigator Grant MOBOCON (291 458) and by the German Federal Ministry of Education and Research under grant 05M2013.

$J(x)M(x) \approx \mathrm{I}_n$, if the additional numerical effort for only linear or superlinear local convergence rate can be compensated by less numerical effort for the evaluation of $M(x)F(x)$ in comparison to $J^{-1}(x)F(x)$, which involves the (at least partial) computation of derivatives for $J(x)$ and the solution of a linear system with $J(x)$ and right-hand side $F(x)$. Second, no initial guess x_0 might be available that is sufficiently close to a solution x^*, in particular because the qualification "sufficiently close" depends on the nonlinearity of the problem at hand.

Throughout this article, $\langle .,. \rangle$ is an inner product of $\mathbb{R}^n$. For $v \in \mathbb{R}^n$, we denote the corresponding vector norm with $\|v\| = \sqrt{\langle v, v \rangle}$. Furthermore, we also use $\|A\|$ for the induced matrix sup-norm and the condition number $\operatorname{cond}(A) = \|A\| \, \|A^{-1}\|$ for a matrix $A \in \mathbb{R}^{n \times n}$ (or $\operatorname{cond}(A) = \infty$ if A is not invertible). The symbol I_n denotes the n-by-n identity matrix. We denote the gradient of a real-valued function $\phi : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ by $\nabla\phi(x) \in \mathbb{R}^n$, defined by

$$\langle \nabla\phi(x), v \rangle = \frac{\mathrm{d}\phi}{\mathrm{d}x}(x)v \quad \text{for all } v \in \mathbb{R}^n.$$

For any real n-by-n matrix A, we denote its determinant by $\det A$. Let $I \subset \mathbb{R}$ be a (possibly unbounded) interval and $y : I \to \mathbb{R}^n$ a continuously differentiable curve, then we denote its arc length by

$$\ell(y) := \int_I \|\dot{y}(t)\| \, \mathrm{d}t.$$

The (open) ball in $\mathbb{R}^n$ of radius $r \geq 0$ around $x \in \mathbb{R}^n$ will be denoted by

$$B(x, r) = \{y \in \mathbb{R}^n \mid \|x - y\| < r\}.$$

We also assume the convention $0 \in \mathbb{N}$. Euler's number is denoted by $e = \sum_{k=0}^{\infty} \frac{1}{k!}$ and the imaginary unit by $\mathrm{i} \in \mathbb{C}$.

2. Contributions. In this article, we present a novel step size strategy called *backward step control* for Newton-type methods. Given any $x_0 \in D$, we state reasonable assumptions that guarantee convergence to not just some, but a specific solution of (1.1) connected to x_0 by the so-called *generalized Newton path.* Furthermore, we provide an a-priori bound on $\|F(x_k)\|$ and show that full steps $t_k = 1$ are guaranteed in the vicinity of a solution. In addition, we present two algorithmic realization of backward step control, which are based on monotone iterations or a bisection procedure. For the bisection approach, we investigate the numerical efficiency and reliability of backward step control on various examples.

3. Summary. We provide an overview of the current state-of-the art in globalization methods in Section 4. As the basis for our convergence proof, we state and discuss the five main assumptions of this article in Section 5 and consider Newton-type methods with infinitesimally small step sizes in order to establish existence and uniqueness of the generalized Newton path. In Section 6, we present a local and global convergence theorem, which does not lend itself immediately to an efficient numerical method, but rather serves as a theoretical basis for the following sections. We then take a look at a geometric interpretation of iteration (1.2) in Section 7. We introduce backward step control in Section 8 and develop a convergence proof, followed by a discussion of affine invariance properties in Section 9. Two algorithmic realizations of backward step control are then presented in Section 10. A discussion of the algorithmic parameters follows in Section 11, before we finally demonstrate the numerical efficiency of backward step control in Section 12 on several examples, including more than one hundred benchmark examples from unconstrained optimization with up to $n = 100\,000$ variables.

4. Existing approaches. All popular globalization approaches rely directly or indirectly on monotonicity of a function of the iterates x_k. The minima of this function must correspond to solutions of (1.1) and Δx_k must point into a direction of descent. The function can belong to the class of generalized level set functions [16]

$$T(x|A) = \tfrac{1}{2} \left\| AF(x) \right\|^2, \quad \text{with } A \in \mathbb{R}^{n \times n} \text{ invertible,}$$

or, in the case of unconstrained optimization problems to be discussed in Section 12, it can coincide with the objective function.

With the canonical Euclidean inner product, the Levenberg-Marquardt method (see, e.g., [22, 29, 24, 16]) uses

$$M_\lambda(x) = (J(x)^T J(x) + \lambda \mathrm{I}_n)^{-1} J(x)^T, \quad \text{with } \lambda > 0.$$

If $J(x)$ is invertible, it holds that

$$\lim_{\lambda \to 0} M_\lambda(x) F(x) = J^{-1}(x) F(x) \quad \text{and} \quad \lim_{\lambda \to \infty} \lambda M_\lambda(x) F(x) = \nabla T(x|\mathrm{I}_n).$$

Thus, we obtain a Newton step for $\lambda \to 0$ and a damped gradient descent step on $T(x|\mathrm{I}_n)$ for $\lambda \to \infty$, which is the basis for global convergence, even if $J(x)$ is singular. It is well-known that each step of the Levenberg-Marquardt method is equivalent to solving for a trust region step with a trust region radius $\delta > 0$ (see, e.g., [11])

$$\min_{\Delta x_k \in \mathbb{R}^n} \tfrac{1}{2} \left\| J(x_k) \Delta x_k + F(x_k) \right\|^2 \quad \text{s.t.} \quad \left\| \Delta x_k \right\| \leq \delta.$$

The parameter λ takes the role of a Lagrange multiplier of the trust region constraint. The extra effort necessary to determine appropriate values for λ, however, constitutes an overhead that can render trust region approaches less efficient than other methods for many problems, as we shall see from the numerical results in Section 12.3.

Alternatively, line search methods enforce descent explicitly by successively reducing trial values for t_k (see, e.g., [24]). For many functions, including $T(x|\mathrm{I}_n)$, this requirement can lead to unnecessarily short steps $t_k \ll 1$ even close to a solution as we discuss in Section 12.2. To circumvent this inefficiency, Deuflhard [14, 15, 16] proposed a convergence proof that enforces descent in $T(x|J^{-1}(x^*))$, which, for twice continuously differentiable F, is an asymptotic distance function

$$T(x|J^{-1}(x^*)) = \tfrac{1}{2} \left\| x - x^* \right\|_2^2 + \mathcal{O}(\left\| x - x^* \right\|^3).$$

As a consequence, full steps in the vicinity of a solution x^* can be guaranteed. However, $T(x|J^{-1}(x^*))$ is not realizable in practice, because x^* is unknown. The substitute $T(x|J^{-1}(x_k))$ has been used with remarkable success in the form of a monotonicity requirement for a simplified Newton step. However, no convergence proof for these so-called natural level functions exists, which change from iteration to iteration. The Ascher-Osborne example [2] shows that the enforcement of natural monotonicity alone can lead to cycling of (1.2). The Restrictive Monotonicity Test [5] does not exhibit cycles on the Ascher-Osborne example, but still lacks a convergence proof.

5. The infinitesimal Newton-type method. Given an initial value $x_0 \in D$, we consider the initial value problem

$$\dot{x}(t) = -f(x(t)), \quad \text{with } f(x) := M(x)F(x) \text{ and } x(0) = x_0. \tag{5.1}$$

The use of differential equations for solving nonlinear systems dates back to [12]. We see that the Newton-type method (1.2) is really an Explicit Euler method for the solution of the initial value problem (5.1) for $t \to \infty$ with the step size sequence (t_k). It is thus not surprising that rather restrictive requirements on the step sizes t_k are necessary for the global convergence of iteration (1.2) due to the well-known stability limitations of the Explicit Euler method. The use of higher-order and implicit methods has been considered by many authors (compare, e.g., [6] and the references in [25, Section 7.5]), but these approaches usually require differentiability of $M(x)$ and lead to rather involved implementations. Our approach relies on an estimate of the distance between the actually computed explicit step and an implicit step that does not need to be computed. We obtain this estimate from a backward step, which comes at no extra cost provided that the first order Euler method and no higher order method of integrating (5.1) is used.

For the Newton method $M(x) = J^{-1}(x)$, the curve $x(t)$ that solves (5.1) can actually be derived as the solution $x(t)$ of the homotopy

$$\Phi(x,t) = F(x) - e^{-t}F(x_0) = 0 \quad \text{for all } t \in [0, \infty). \tag{5.2}$$

For general Newton-type methods, however, this remarkable property only holds in a weaker form that we show in Lemma 5.5 below. We now provide sufficient conditions for the existence of solutions of (5.1).

Definition 5.1 (Level sets). *Let $T : D \to \mathbb{R}$ denote the classical level function $T(x) := \frac{1}{2} \|F(x)\|^2 = \frac{1}{2} \langle F(x), F(x) \rangle$. We call*

$$\widetilde{\mathcal{T}}(x) := \{y \in D \,|\, T(y) \le T(x)\}$$

the level set of $x \in D$ and define the connected level set of $x \in D$ as

$$\mathcal{T}(x) := \{y \in \widetilde{\mathcal{T}}(x) \,|\, \textit{there is a path in } \widetilde{\mathcal{T}}(x) \textit{ connecting } x \textit{ and } y\}.$$

Definition 5.2 ($\bar{\varepsilon}$-regular and $\bar{\varepsilon}$-singular points). *For $\bar{\varepsilon} \ge 0$ we define*

$$\mathcal{R}_{\bar{\varepsilon}} = \{x \in D \,|\, \|x - y\| > \bar{\varepsilon} \textit{ for all } y \in D \textit{ with } \det J(y) = 0\}, \quad \mathcal{S}_{\bar{\varepsilon}} = D \setminus \mathcal{R}_{\bar{\varepsilon}}.$$

We can thus divide the domain D into the set of regular points $\mathcal{R}_0$ and the set of singularities $\mathcal{S}_0$. From now on we assume $\bar{\varepsilon} > 0$ to be fixed and use the abbreviations $\mathcal{S} := \mathcal{S}_{\bar{\varepsilon}}$ and $\mathcal{R} := \mathcal{R}_{\bar{\varepsilon}}$.

Let $x_0 \in D \cap \mathcal{R}$ with $\|F(x_0)\| > 0$ be given. For the remainder, we need the following assumptions:

A1. The connected level set $\mathcal{T}(x_0)$ is compact.

A2. There is a constant $\kappa < 1$ such that

$$\|\mathrm{I}_n - J(x)M(x)\| \le \kappa \quad \text{for all } x \in \mathcal{R} \cap \mathcal{T}(x_0).$$

A3. There is a constant $\omega < \infty$ such that

$$\|[J(x) - J(x - tf(x))]\, M(x)\| \le \omega t \,\|f(x)\| \quad \text{for all } x \in \mathcal{T}(x_0), t \in [0, 1].$$

A4. There is a Lipschitz constant $L < \infty$ such that

$$\|f(x) - f(y)\| \leq L \|x - y\| \quad \text{for all } x, y \in \mathcal{T}(x_0).$$

A5. For all $\Delta > 0$ there exist constants $\gamma, t_\gamma > 0$ such that

$$\|f(x - tf(x)) - f(x)\| \geq \gamma t$$

for all $t \in [0, t_\gamma]$ and all $x \in \mathcal{R} \cap \mathcal{T}(x_0)$ with $\|f(x)\| \geq \Delta$.

We feel that a few remarks are appropriate regarding assumptions A2–A5. In particular, we would like to point out that although the general blueprint of these conditions has been used before by other authors, we have devised subtle modifications that were necessary to complete the convergence proof for backward step control in Section 8.

To begin, the κ-condition A2 is a more restrictive version of a fairly general classical contraction assumption [25, Section 10.2.1], which requires for some $x^* \in D$ with $F(x^*) = 0$ that the spectral radius σ of the matrix $G(x^*) := \mathrm{I}_n - M(x^*)J(x^*)$ satisfy $\sigma < 1$, which then implies locally linear convergence with root-convergence factor equal to σ. If $J(x^*)$ is invertible, then σ is also the spectral radius of the similarity transform $G'(x^*) := \mathrm{I}_n - J(x^*)M(x^*) = J^{-1}(x^*)G(x^*)J(x^*)$ that appears in A2. The first appearance of a κ-condition with norms instead of the spectral radius for the characterization of the locally linear convergence factor can be found in [4] for the case of Gauss-Newton methods.

The ω-condition A3 is a specially weighted, weaker form of a Lipschitz condition on J: In order to see this, let us define the set of reachable points

$$U = \{y \in \mathbb{R}^n \mid y = x - tf(x) \text{ for some } x \in \mathcal{T}(x_0), t \in [0, 1]\}.$$

If A1 and $U \subseteq D$ hold and if in addition $J(x)$ is Lipschitz continuous in U with Lipschitz constant $L_J < \infty$, then U, as a continuous image of a compact set, is compact and there is an $\overline{M}_U < \infty$ such that $\|M(x)\| \leq \overline{M}_U$ for all $x \in U$. Thus, A3 is satisfied with $\omega = L_J \overline{M}_U$ due to

$$\|[J(x) - J(x - tf(x))]\, M(x)\| \leq L_J t \|f(x)\| \, \overline{M}_U = \omega t \|f(x)\|.$$

The same reasoning can be applied if we start with the assumptions of the classical Newton-Kantorovich Theorem (see, e.g., [25, Section 12.6.2]), which imply Lipschitz continuity of $J(x)$ and boundedness of $M(x) = J^{-1}(x)$, and thus also A3.

The ω-condition A3 has a close resemblence with affine covariant Lipschitz conditions, e.g., in the refined Newton-Mysovskii theorem [16, Theorem 2.3] where it is required for some open and convex $D^0 \subseteq D$ that

$$\|M(x)[J(y) - J(x)](y - x)\| \leq \omega \|y - x\|^2 \quad \text{for all } x, y \in D^0. \tag{5.3}$$

Condition (5.3) is implied by the affine covariant condition

$$\|M(x)[J(y) - J(x)]\| \leq \omega \|y - x\| \quad \text{for all } x, y \in D^0.$$

If we switch here to an affine contravariant condition by commuting the M and J terms according to

$$\|[J(y) - J(x)]M(x)\| \leq \omega \|y - x\| \quad \text{for all } x, y \in D^0,$$

we can deduce A3 for the choice $y = x - tf(x)$ provided that the set of reachable points U is contained in D^0. Thus, the Newton-Mysovskii condition (5.3) and A3 are implied by a stronger assumption in a covariant and a contravariant version, respectively.

If F is a linear function and M is constant, then we have $\omega = 0$. In this sense, ω is one way of quantifying the nonlinearity of F in $\mathcal{T}(x_0)$. In the case of the Newton method $M(x) = J^{-1}(x)$, we have $\kappa = 0$ provided that $\mathcal{S}_0 \cap \mathcal{T}(x_0) = \{\}$.

For the case of M being continuously differentiable, the γ-condition A5 implies a uniform lower bound on the derivative

$$\left\| \frac{\mathrm{d}f}{\mathrm{d}x}(x) f(x) \right\| \geq \gamma \quad \text{for all } x \in \mathcal{R} \cap \mathcal{T}(x_0) \text{ with } \|f(x)\| > \Delta,$$

which is, as we see later, a uniform lower bound on the norm of the second derivative of the generalized Newton path. This technical assumption is necessary to exclude pathological examples from our analysis. We also remark that we do not prescribe a bound on the curvature of the generalized Newton path, which is the second derivative in arc length parametrization.

Without loss of generality, we assume that A3 and A4 also hold for singular points. If this is not the case, we truncate $M(x)$ on $\mathcal{S}$. Even a continuous transition to 0 would be possible on $\mathcal{S}$.

We must assume $\bar{\varepsilon} > 0$ for the following reason: If there exists a $y \in \mathcal{S}_0 \cap \mathcal{T}(x_0)$, then A2 cannot hold by continuity of J and M and by virtue of

$$1 > \kappa \geq \|\mathrm{I}_n - J(y)M(y)\| = \sup_{\|v\|=1} \|v - J(y)M(y)v\| \geq 1,$$

which we obtain by choosing v in the nullspace of $M(y)$ or, if $M(y)$ is invertible, by choosing $M(y)v$ in the nullspace of $J(y)$. Conversely, A2 with $\bar{\varepsilon} > 0$ is sufficient for invertibility of $J(x)$ and $M(x)$ for all $x \in \mathcal{R} \cap \mathcal{T}(x_0)$.

The following lemma establishes ratios between $\|f(x)\|$ and $\|F(x)\|$.

Lemma 5.3. *If A1 holds, then there is a constant $\overline{M} < \infty$ such that*

$$\|f(x)\| \leq \overline{M} \, \|F(x)\| \quad \text{for all } x \in \mathcal{T}(x_0).$$

If, furthermore, A2 holds, then there is a constant $\overline{m} > 0$ such that

$$\overline{m} \, \|F(x)\| \leq \|f(x)\| \quad \text{for all } x \in \mathcal{R} \cap \mathcal{T}(x_0).$$

Proof. Because of A1 and continuity of $M(x)$ and $J(x)$ there exist constants $\overline{M}, \overline{J} \in (0, \infty)$ such that

$$\|M(x)\| \leq \overline{M} \quad \text{and} \quad \|J(x)\| \leq \overline{J} \quad \text{for all } x \in \mathcal{T}(x_0).$$

The first assertion follows then from

$$\|f(x)\| \leq \|M(x)\| \, \|F(x)\| \leq \overline{M} \, \|F(x)\| \, .$$

If A2 holds, we can estimate for $x \in \mathcal{R} \cap \mathcal{T}(x_0)$

$$\begin{aligned} \|F(x)\| &\leq \|(\mathrm{I}_n - J(x)M(x)) \, F(x)\| + \|J(x)M(x)F(x)\| \\ &\leq \kappa \, \|F(x)\| + \overline{J} \, \|f(x)\| \, , \end{aligned}$$

which is equivalent to

$$\overline{m}\,\|F(x)\| \leq \|f(x)\| \quad \text{with } \overline{m} := (1-\kappa)/\overline{J} > 0. \;\square$$

We now assume that $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$ for some $k \in \mathbb{N}$ and investigate the initial value problem

$$\dot{x}^k(t) = -f(x^k(t)), \quad x^k(0) = x_k. \tag{5.4}$$

DEFINITION 5.4. *We call a curve $x^k(t)$ that satisfies* (5.4) *a* generalized Newton path.

LEMMA 5.5. *If A2, A4, and $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$ hold, then* (5.4) *has a unique local solution $x^k(t) \in \mathcal{T}(x_k), t \in [0, \bar{t})$, and the level function satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t} T(x^k(t)) \leq -2(1-\kappa) T(x^k(t)) < 0 \quad \textit{for all } t \in [0, \bar{t}) \textit{ with } x^k(t) \in \mathcal{R}.$$

Moreover, the norm of F decreases exponentially along $x^k(t)$

$$\left\|F(x^k(t))\right\| \leq \|F(x_k)\|\, e^{-(1-\kappa)t}.$$

Proof. By A4 and the Picard-Lindelöf Theorem (see, e.g., [8]), there exists a unique local solution $x^k(t)$ in a neighborhood of $t = 0$. Using the abbreviation $x = x^k(t)$, we obtain the first assertion by application of the Cauchy-Schwarz inequality and A2 according to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} T(x) &= \langle F(x), J(x)\dot{x} \rangle = -\langle F(x), J(x)M(x)F(x) \rangle \\ &= -\langle F(x), F(x) \rangle + \langle F(x), (\mathrm{I}_n - J(x)M(x))\, F(x) \rangle \\ &\leq -2T(x) + \|F(x)\|\, \|\mathrm{I}_n - J(x)M(x)\|\, \|F(x)\| \leq -2(1-\kappa)T(x) < 0.\end{aligned}$$

The Gronwall inequality (see, e.g., [1]) then yields

$$T(x^k(t)) \leq T(x_k) e^{-2(1-\kappa)t},$$

which proves $x^k(t) \in \mathcal{T}(x_k)$ and the last assertion. $\square$

THEOREM 5.6. *If A1, A2, and A4 hold, then exactly one of the following holds:*

i) Equation (5.4) *has a unique global solution $x^k(t) \in \mathcal{R} \cap \mathcal{T}(x_k) \subseteq \mathcal{T}(x_0)$ of finite arc length*

$$\ell(x^k) \leq \frac{\overline{M}\,\|F(x_k)\|}{1-\kappa} \leq \frac{\overline{M}\,\|f(x_k)\|}{\overline{m}(1-\kappa)}$$

with the constants $\overline{m}$ and $\overline{M}$ from Lemma 5.3.

ii) There is a $t_S \in [0, \infty)$ such that (5.4) *has a solution for $t \in [0, t_S]$ and*

$$x^k(t) \in \mathcal{R} \cap \mathcal{T}(x_k) \quad \textit{for } t \in [0, t_S), \qquad\qquad x^k(t_S) \in \mathcal{S}.$$

Proof. Lemma 5.5 reveals that $x^k(t)$ stays in $\mathcal{R} \cap \mathcal{T}(x_k)$ for small t. Thus, by repeated application of the Picard-Lindelöf Theorem the local solution can be uniquely extended either to the whole interval $t \in [0, \infty)$ or until $x^k(t) \notin \mathcal{R}$. In the latter case, we immediately obtain assertion *ii)*. In the former case, we obtain *i)* from Lemmas 5.3 and 5.5 by virtue of

$$\begin{aligned}\ell(x^k) &= \int_0^\infty \left\|\dot{x}^k(t)\right\| \mathrm{d}t = \int_0^\infty \left\|M(x^k(t))F(x^k(t))\right\| \mathrm{d}t \\ &\leq \overline{M}\,\|F(x_k)\| \int_0^\infty e^{-(1-\kappa)t}\mathrm{d}t \leq \frac{\overline{M}\,\|F(x_k)\|}{1-\kappa} \leq \frac{\overline{M}\,\|f(x_k)\|}{\overline{m}(1-\kappa)} < \infty. \;\square\end{aligned}$$

Definition 5.7. *If it exists, we denote the end point of the curve x^k by*

$$x_k^* := \lim_{t\to\infty} x^k(t).$$

Lemma 5.5 immediately delivers $F(x_k^*) = 0$.

We now show with the help of the following lemma that regular solutions x^* of (1.1) are locally unique.

Lemma 5.8. *Let A1 hold and let (y_i) be a sequence in $\mathcal{T}(x_0)$ satisfying $F(y_i) = 0$ and $y_i \neq y_j$ for all $i, j \in \mathbb{N}$. Then, each accumulation point y^* of (y_i) satisfies $y^* \in \mathcal{S}_0 \cap \mathcal{T}(x_0)$.*

Proof. Due to A1, it holds that $y^* \in \mathcal{T}(x_0)$. Without loss of generality, we assume $y^* \neq y_i$ for all $i \in \mathbb{N}$ and define

$$d_i = \frac{y_i - y^*}{\|y_i - y^*\|} \quad \text{for all } i \in \mathbb{N}.$$

It holds that $\|d_i\| = 1$ for all $i \in \mathbb{N}$. By compactness of the unit sphere in $\mathbb{R}^n$, we can thus choose a convergent subsequence of (d_i). We denote its limit by d^*. By virtue of $F(y_i) = F(y^*) = 0$ for all $i \in \mathbb{N}$ we observe that

$$0 = \lim_{i\to\infty} \frac{F(y_i) - F(y^*)}{\|y_i - y^*\|} = \lim_{i\to\infty} \int_0^1 J(y^* + t(y_i - y^*))\frac{y_i - y^*}{\|y_i - y^*\|}\mathrm{d}t = J(y^*)d^*.$$

Thus, $J(y^*)$ is singular, which implies $y^* \in \mathcal{S}_0$. $\square$

Theorem 5.9. *Let A1 hold. If $x^* \in \mathcal{T}(x_0) \setminus \mathcal{S}_0$ satisfies $F(x^*) = 0$, then there exists an $\varepsilon > 0$ such that*

$$\|F(x)\| > 0 \quad \textit{for all } x \in B(x^*, \varepsilon) \setminus \{x^*\} \subset \mathcal{T}(x_0) \setminus \mathcal{S}_0.$$

Proof by contradiction. If we assume the contrary, we can construct a sequence (y_i) in $\mathcal{T}(x_0)$ converging to x^* which satisfies $F(y_i) = 0$. Then, Lemma 5.8 yields the contradiction $x^* \in \mathcal{S}_0$. $\square$

6. Global and local convergence. We now investigate the convergence of iteration (1.2). It turns out that under a reasonable condition on the step size sequence (t_k), x_k either becomes $\bar{\varepsilon}$-singular or converges to a solution of (1.1). In addition, full steps are allowed in the vicinity of the solution. First, we state an equivalent formulation of the κ-condition A2.

Lemma 6.1. *A2 holds if and only if for all $x \in \mathcal{R} \cap \mathcal{T}(x_0)$ and $t \in [0, 1]$*

$$\|\mathrm{I}_n - tJ(x)M(x)\| \leq 1 - (1-\kappa)t.$$

Proof. For $t = 1$, A2 follows. As a concatenation of a linear function and a norm, the function $\varphi(t) := \|\mathrm{I}_n - tJ(x)M(x)\|$ is convex. Thus, we obtain the assertion via

$$\varphi(t) = \varphi((1-t)\cdot 0 + t \cdot 1) \leq (1-t)\varphi(0) + t\varphi(1) \leq 1 - (1-\kappa)t. \quad \square$$

Lemma 6.2. *Let A1, A2, and A3 hold. If $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$, then*

$$\|F(x_{k+1})\| \leq \left[1 - (1-\kappa)t_k + (\omega/2)\|f(x_k)\|\, t_k^2\right] \|F(x_k)\|.$$

Furthermore, if there exists a $\theta < 1$ such that the step size sequence satisfies

$$\omega t_k \|f(x_k)\| \leq 2\theta(1-\kappa), \tag{6.1}$$

then

$$\|F(x_{k+1})\| \leq [1 - (1-\theta)(1-\kappa)t_k] \|F(x_k)\|.$$

Proof. Using the fundamental theorem of calculus and Lemma 6.1, we obtain

$$\begin{aligned}
\|F(x_{k+1})\| &= \|F(x_k) + F(x_k - t_k f(x_k)) - F(x_k)\| \\
&= \left\|F(x_k) - \textstyle\int_0^{t_k} J(x_k - \tau f(x_k)) f(x_k) \mathrm{d}\tau\right\| \\
&\leq \left\|\mathrm{I}_n - t_k J(x_k) M(x_k) + \textstyle\int_0^{t_k} [J(x_k) - J(x_k - \tau f(x_k))] M(x_k) \mathrm{d}\tau\right\| \|F(x_k)\| \\
&\leq \left[1 - (1-\kappa)t_k + \textstyle\int_0^{t_k} \|[J(x_k) - J(x_k - \tau f(x_k))]M(x_k)\| \mathrm{d}\tau\right] \|F(x_k)\| \\
&\leq \left[1 - (1-\kappa)t_k + (\omega/2)\|f(x_k)\|\, t_k^2\right] \|F(x_k)\|,
\end{aligned}$$

which shows the first assertion. The second assertion follows immediately from

$$\begin{aligned}
\|F(x_{k+1})\| &= \left[1 - (1 - \kappa - \tfrac{\omega}{2}\|f(x_k)\|\, t_k)t_k\right] \|F(x_k)\| \\
&\leq [1 - (1-\theta)(1-\kappa)t_k] \|F(x_k)\|. \quad \square
\end{aligned}$$

Lemma 6.3. *Let A1, A2, A3, and A4 hold and let $x^* \in \mathcal{T}(x_0) \setminus \mathcal{S}_0$ satisfy $F(x^*) = 0$. Then there exists a constant $\varepsilon^* > 0$ such that if $x_k \in B(x^*, \varepsilon^*)$ for some $k \in \mathbb{N}$, then exactly one of the following holds:*

i) For all $j \geq k$ it holds that $x_j \in \mathcal{R} \cap \mathcal{T}(x_k)$.

ii) There exists an $N \geq k$ such that for all $j \in \mathbb{N}$ with $k \leq j < N$, it holds that $x_j \in \mathcal{R} \cap \mathcal{T}(x_k)$ and $x_N \in \mathcal{S} \cap \mathcal{T}(x_k)$.

Furthermore, if $F(x) = 0$ for some $x \in \mathcal{T}(x_k)$, then $x = x^*$.

Proof. With $\varepsilon > 0$ from Theorem 5.9, there exists a $y \in B(x^*, \varepsilon) \subset \mathcal{T}(x_0) \setminus \mathcal{S}_0$ with $\|F(y)\| > 0$ and such that

$$\mathcal{U} := \mathcal{T}(y) \subset B(x^*, \varepsilon).$$

Without loss of generality, we can assume ε to be small enough to satisfy for some $\theta \in (0,1)$

$$\omega L \varepsilon \leq 2(1-\theta)\kappa.$$

Because the interior of $\mathcal{U}$ is not empty, there exists a constant $\varepsilon^* > 0$ such that

$$B(x^*, \varepsilon^*) \subset \mathcal{U} \subset B(x^*, \varepsilon).$$

Then, with $x_k \in B(x^*, \varepsilon^*) \subset \mathcal{U}$, we see that

$$\omega t_k \|f(x_k)\| = \omega t_k \|f(x_k) - f(x^*)\| \leq \omega L \|x_k - x^*\| \leq \omega L \varepsilon \leq 2(1-\theta)\kappa.$$

If $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$, then Lemma 6.2 then delivers

$$\|F(x_{k+1})\| \leq \|F(x_k)\| \leq \|F(y)\|,$$

which implies $x_{k+1} \in \mathcal{T}(x_k) \subset \mathcal{U} \subset B(x^*, \varepsilon)$. By induction, $x_j \in \mathcal{T}(x_k) \subset \mathcal{U} \subset B(x^*, \varepsilon)$ for all $j \geq k$ with $x_j \in \mathcal{R}$. Thus, either there exists an $N \in \mathbb{N}$ such that $x_j \in \mathcal{S}$, which implies assertion *ii)*, or not, which implies assertion *i)*. The last assertion follows from Theorem 5.9. □

Theorem 6.4. *Let A1, A2, and A3 hold. If there exists a $\theta < 1$ such that the step size sequence satisfies* (6.1) *for all $k \in \mathbb{N}$, then exactly one of the following holds:*

i) The sequence (x_k) stays in $\mathcal{R} \cap \mathcal{T}(x_0)$. If, furthermore, A4 holds and

$$t^* := \limsup_{k \to \infty} t_k > 0,$$

then (x_k) converges to some $x^* \in \mathcal{T}(x_0)$ with $F(x^*) = 0$.

ii) There is an $N \in \mathbb{N}$ such that

$$x_k \in \mathcal{R} \cap \mathcal{T}(x_0) \quad \textit{for all } k < N \quad \textit{and} \quad x_N \in \mathcal{S} \cap \mathcal{T}(x_0).$$

Proof. Let $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$, which holds for x_0 by our general assumption. Then, Lemma 6.2 delivers $x_{k+1} \in \mathcal{T}(x_k) \subset \mathcal{T}(x_0)$. If there is an $N \in \mathbb{N}$ such that $x_N \in \mathcal{S}$, then we can take the smallest such N and *ii)* holds. Otherwise, it holds that $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$ for all $k \in \mathbb{N}$ by induction, which proves the first assertion of *i)*. For the second assertion of *i)*, we now choose a subsequence of (t_k) via an index sequence (k_i) such that

$$k_i < k_{i+1} \quad \text{and} \quad t_{k_i} \geq t^*/2 \quad \text{for all } i \in \mathbb{N}.$$

This yields for $j \in \mathbb{N}$ that

$$\left\|F(x_{k_j})\right\| \leq [1 - (1-\theta)(1-\kappa)t^*/2]^j \|F(x_0)\|.$$

Thus, $F(x_k)$ converges to zero. Moreover, by A1, there is a convergent subsequence of (x_{k_i}) with limit point $x^* \in \mathcal{T}(x_0)$ satisfying $F(x^*) = 0$ by continuity of F. Then Lemma 6.3 guarantees that (x_k), and not just a subsequence, converges to x^*. □

Theorem 6.4 seems to suggest a rather simple preliminary step size selection: For a given constant $H' > 0$, we select

$$t_k = \min(1, H'/\|f(x_k)\|).$$

If H' is chosen small enough to satisfy $\omega H' \leq 2\theta(1-\kappa)$, then we obtain global convergence with guaranteed full steps in the vicinity of the solution by Theorem 6.4. However, we strongly discourage from the use of this simple strategy, because it takes only tiny steps $\|x_{k+1} - x_k\| \leq H'$ (with equality for $t_k < 1$) and is thus inefficient in most cases.

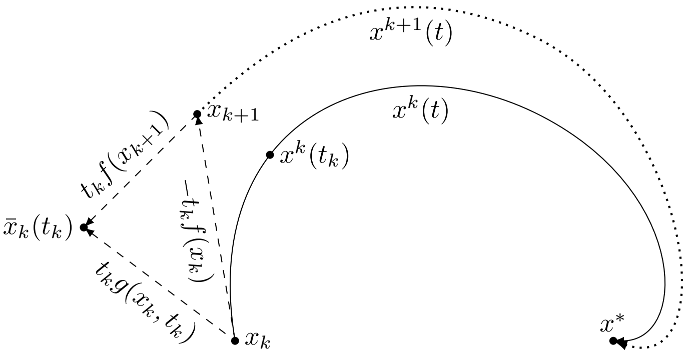

FIGURE 7.1. *The generalized Newton paths $x^k(t)$ and $x^{k+1}(t)$ are related by a tangential step $-t_k f(x_k)$ from x_k to x_{k+1}. The backward step $t_k f(x_{k+1})$ from x_{k+1} leads to $\bar{x}_k(t_k) = x_k + t_k g(x_k, t_k)$.*

7. Geometric interpretation. Let $k \in \mathbb{N}$. Starting from $x_k \in \mathcal{T}(x_0)$, we now take one step of iteration (1.2). If t_k is sufficiently small, $x_{k+1} \in \mathcal{T}(x_k)$ due to Lemma 6.2. The backward step control to be addressed in Section 8 is based on the backward iterate (compare Figure 7.1)

$$\bar{x}_k(t_k) = x_{k+1} + t_k f(x_{k+1}) = x_k + t_k g(x_k, t_k) \quad \text{with } g(x,t) = f(x - tf(x)) - f(x).$$

We first show that the deviation from the current generalized Newton path after one step decreases quadratically with the step size $t_k \to 0$.

LEMMA 7.1. *If A4 holds and $x^k(t), x_k - tf(x_k) \in \mathcal{T}(x_0)$ for all $t \in [0, t_k]$, then*

$$\left\|x^k(t_k) - x_{k+1}\right\| \leq \tfrac{1}{2} \left\|f(x_k)\right\| L e^{t_k L} t_k^2.$$

Proof. Under A4, we can estimate the distance of $x^k(t)$ from the line between x_{k+1} and x_k on the basis of the integral form of (5.4) according to

$$\begin{aligned}\left\|x^k(t) - [x_k - tf(x_k)]\right\| &= \left\|\int_0^t [f(x^k(\tau)) - f(x_k)] \mathrm{d}\tau\right\| \leq L \int_0^t \left\|x^k(\tau) - x_k\right\| \mathrm{d}\tau \\ &= L \int_0^t \left\|x^k(\tau) - [x_k - \tau f(x_k)] - \tau f(x_k)\right\| \mathrm{d}\tau \\ &\leq L \left\|f(x_k)\right\| \frac{t^2}{2} + L \int_0^t \left\|x^k(\tau) - [x_k - \tau f(x_k)]\right\| \mathrm{d}\tau.\end{aligned}$$

The integral form of the Gronwall inequality then yields

$$\left\|x^k(t) - [x_k - tf(x_k)]\right\| \leq \tfrac{1}{2} L \left\|f(x_k)\right\| e^{tL} t^2.$$

The assertion follows for $t = t_k$. □

As a consequence of the Ribbon Lemma (see supplementary material), if $x^k(t)$ and $x^{k+1}(t)$ converge to different zeros $x_k^* \neq x_{k+1}^*$, then there must be a $\tau \in [0, 1]$ such that the generalized Newton path $\tilde{x}(t)$ emanating from $\tilde{x}(0) = (1-\tau)x_k + \tau x_{k+1}$ hits a singularity.

8. Backward step control. From a computational vantage point, it is difficult to compute estimates of the constants that occur in (6.1) (compare, e.g., [4, 27]). Our aim here is to only use quantities for the step size strategy that need to be computed anyway once t_k is known. Our approach relies on the philosophy that it is important to solve (5.1) in a stable way, but not necessarily with high accuracy, because it is sufficient to steer the iterates into the domain of local full step convergence.

Our step size strategy relies on a backward analysis argument: The iterate x_{k+1} can be interpreted as one step of the Implicit Euler method on (5.1) with initial value $\bar{x}^k(t_k)$, which also converges to x_k quadratically for small t_k. It is well-known [30] that the region of absolute stability of the Implicit Euler method comprises almost the whole complex plane only without the disc of radius 1 centered at 1. Thus, we only take steps for which the explicit increment from x_k does not differ too much from the implicit increment from $\bar{x}_k(t_k)$.

Let us now consider the distance

$$\|\bar{x}_k(t_k) - x_k\| = t_k \|f(x_{k+1}) - f(x_k)\| = t_k \|g(x_k, t_k)\| .$$

This formula can be used for an asymptotically correct estimate of the second time derivative $\ddot{x}^k$ and thus the error committed in step k if we assume f to be continuously differentiable for a moment, because

$$\lim_{t_k \to 0} \frac{f(x_k - t_k f(x_k)) - f(x_k)}{t_k} = -\frac{\mathrm{d}f}{\mathrm{d}x}(x_k) f(x_k) = -\ddot{x}^k(0). \tag{8.1}$$

Lemma 8.1. *If f is continuously differentiable, then the local error*

$$e_k(t_k) := \left\|x_{k+1} - x^k(t_k)\right\|$$

satisfies

$$e_k(t_k) = \tfrac{1}{2} \|\bar{x}_k(t_k) - x_k\| + o(t_k^2) = \tfrac{1}{2} t_k \|f(x_{k+1}) - f(x_k)\| + o(t_k^2).$$

Proof. Under the stated assumption, we have that x^k is twice continuously differentiable. Hence, Taylor's theorem yields the existence of a function $\varphi(t)$ with $\lim_{t\to 0} \varphi(t) = 0$ such that

$$x^k(t_k) = x^k(0) + t_k \dot{x}^k(0) + \tfrac{1}{2} t_k^2 \ddot{x}^k(0) + t_k^2 \varphi(t_k) = x_{k+1} + \tfrac{1}{2} t_k^2 \ddot{x}^k(0) + t_k^2 \varphi(t_k).$$

We then obtain the assertion from (8.1) by

$$\left\|x_{k+1} - x^k(t_k)\right\| = \tfrac{1}{2} t_k^2 \left\|\ddot{x}^k(0) + 2\varphi(t_k)\right\| = \tfrac{1}{2} t_k \|f(x_{k+1}) - f(x_k)\| + o(t_k^2). \quad \square$$

Lemma 8.1 suggests a step size strategy that controls the local error in each step. For a given desired error tolerance $H/2 > 0$, we require

$$H/2 = e_k(t_k) = (t_k/2) \|f(x_{k+1}) - f(x_k)\| + o(t_k^2).$$

Neglecting higher-order terms, we thus require

$$t_k := \min \mathcal{B}_H(x_k) \quad \text{where } \mathcal{B}_H(x) = \{t \in [0,1] \mid H = t \|g(x,t)\|\} \cup \{1\}. \tag{BSC}$$

This implies the backward step control

$$H \geq \|\bar{x}_k(t_k) - x_k\| \quad \text{(with equality for } t_k < 1). \tag{8.2}$$

Equations (BSC) and (8.2) are also meaningful without the motivational assumption of f being continuously differentiable: In the sense of a backward analysis, x_{k+1}, the result of the possibly unstable Explicit Euler step, coincides with the result of a stable implicit Euler step for a perturbation of size H of the initial value x_k.

If A1 and A4 hold, full steps in the vicinity of a solution are guaranteed by (BSC), because if $\|f(x_k)\| < H/L$ (or $\|F(x_k)\| < H/(L\overline{M})$ with Lemma 5.3), then

$$t\,\|g(x_k, t)\| \leq Lt^2\,\|f(x_k)\| < H \quad \text{for all } t \leq 1.$$

Furthermore, a lower step size bound holds for $x_k \in \mathcal{T}(x_0)$, because if $t_k < 1$, then

$$t_k^2 = \frac{t_k H}{\|g(x_k, t_k)\|} \geq \frac{H}{L\,\|f(x_k)\|} \geq \frac{H}{L\overline{M}\,\|F(x_k)\|} \geq \frac{H}{L\overline{M}\,\|F(x_0)\|}. \tag{8.3}$$

Lemma 8.2. *Let A1, A2, A3, and A5 hold and let $\theta, \bar{t} \in (0,1)$. Then there exists an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ and $x \in \mathcal{R} \cap \mathcal{T}(x_0)$ with $\omega\,\|f(x)\| \geq 2\theta(1-\kappa)$ it holds that $\min \mathcal{B}_H(x) \leq \bar{t}$.*

Proof by contradiction. Nothing needs to be shown for $\omega = 0$. For $\omega > 0$, we assume that for all $\overline{H} > 0$ there exists an $H \in (0, \overline{H}]$ and a $x \in \mathcal{R} \cap \mathcal{T}(x_0)$ satisfying

$$\|f(x)\| \geq 2\theta(1-\kappa)/\omega =: \Delta \quad \text{and} \quad \min \mathcal{B}_H(x) > \bar{t}.$$

Then, A5 yields the existence of $\gamma, t_\gamma > 0$ such that for $t = \min(t_\gamma, \bar{t}) < \min \mathcal{B}_H(y)$ we obtain by (BSC) that

$$\overline{H} \geq H > t\,\|g(x,t)\| \geq \gamma t^2 > 0.$$

Because Δ and thus γ and t are independent of $\overline{H}$, we obtain a contradiction for $\overline{H} \to 0$. □

Lemma 8.3. *Let A1, A2, A3, A4, and A5 hold and let $\theta \in (0,1)$. Then there exists an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ and $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$ it holds that*

$$\omega\,\|f(x_k)\| \min \mathcal{B}_H(x_k) \leq 2\theta(1-\kappa).$$

Proof. Nothing needs to be shown for $\omega = 0$, thus we assume $\omega > 0$. With the constant $\overline{M}$ from Lemma 5.3, we choose

$$\bar{t} = \frac{2\theta(1-\kappa)}{\omega\overline{M}\,\|F(x_0)\|}.$$

Then, Lemma 8.2 delivers the existence of an $\overline{H} > 0$ and we denote $t_k = \min \mathcal{B}_H(x_k)$ for $H \in (0, \overline{H})$. In the case of $t_k > \bar{t}$, Lemma 8.2 delivers

$$\omega t_k\,\|f(x_k)\| \leq \omega\,\|f(x_k)\| < 2\theta(1-\kappa).$$

In the case of $t_k \leq \bar{t}$, we obtain

$$\omega t_k\,\|f(x_k)\| \leq \omega\bar{t}\overline{M}\,\|F(x_0)\| \leq 2\theta(1-\kappa). \ \square$$

Theorem 8.4 (BSC convergence). *Let A1, A2, A3, A4, and A5 hold. Then there exists an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ iteration* (1.2) *with* (BSC) *satisfies exactly one of the following:*

i) For all $k \in \mathbb{N}$ the iterates satisfy $x_k \in \mathcal{R} \cap \mathcal{T}(x_0)$ and there exists an H-independent constant $c > 0$ for the a-priori estimate

$$\sqrt{\|F(x_k)\|} \leq \sqrt{\|F(x_0)\|} - kc\sqrt{H} \quad \text{for all } k \leq \sqrt{\|F(x_0)\| / (c^2 H)}.$$

ii) There exists a $k \in \mathbb{N}$ with $x_j \in \mathcal{R} \cap \mathcal{T}(x_0)$ for all $j < k$ and $x_k \in \mathcal{S}$.

Proof. Let $\theta \in (0, 1)$. Except for the a-priori estimate, the result follows from Lemma 8.3 and Theorem 6.4. In case *i)*, Theorem 6.4 and the lower step size bound (8.3) deliver

$$\begin{aligned} \|F(x_{k+1})\| &\leq [1 - (1-\theta)(1-\kappa)t_k] \|F(x_k)\| \\ &\leq \|F(x_k)\| - \frac{(1-\theta)(1-\kappa)}{\sqrt{L\overline{M}}} \sqrt{H \|F(x_k)\|}. \end{aligned}$$

With $c = \frac{1}{2}(1-\theta)(1-\kappa)/\sqrt{L\overline{M}}$ and the abbreviation $a_k = \|F(x_k)\|$ we obtain

$$a_{k+1} \leq a_k - 2c\sqrt{Ha_k}. \tag{8.4}$$

The right-hand side of (8.4) is an Explicit Euler step with step size $c\sqrt{H}$ for the initial value problem

$$\dot{\alpha}(t) = -2\sqrt{\alpha(t)}, \quad \alpha(0) = a_0,$$

which has for $\alpha(0) > 0$ and $t \in [0, \sqrt{\alpha(0)}]$ the unique solution

$$\alpha(t) = \left(\sqrt{\alpha(0)} - t\right)^2.$$

The convexity of $\alpha(t)$ implies

$$\alpha(t + \tau) \geq \alpha(t) + \tau\dot{\alpha}(t) = \alpha(t) - 2\tau\sqrt{\alpha(t)}. \tag{8.5}$$

By induction, we can show that $a_k \leq \alpha(kc\sqrt{H})$: It holds for $k = 0$ by construction and for $k > 0$ we exploit (8.4) and (8.5) in order to show

$$a_{k+1} \leq a_k - 2c\sqrt{Ha_k} \leq \alpha(kc\sqrt{H}) - 2c\sqrt{H\alpha_k} \leq \alpha((k+1)c\sqrt{H}).$$

Thus, the a-priori estimate follows from

$$\|F(x_k)\| = a_k \leq \alpha(kc\sqrt{H}) = \left(\sqrt{\|F(x_0)\|} - kc\sqrt{H}\right)^2. \quad \square$$

The theorem after the following lemma shows that for H sufficiently small, convergence to x_0^* is guaranteed if the generalized Newton path emanating from the initial guess $x^0(t)$ lies safely within the set of regular points.

LEMMA 8.5. *Let A4 hold. If $x^k(t_k + \tau), x^{k+1}(\tau) \in \mathcal{T}(x_0)$ exist for all $\tau \leq t$, then*

$$\left\|x^k(t_k + t) - x^{k+1}(t)\right\| \leq \tfrac{1}{2} \|f(x_k)\| L e^{L(t_k+t)} t_k^2.$$

Proof. We consider

$$\begin{aligned} &\left\|x^k(t_k + t) - x^{k+1}(t)\right\| \\ &= \left\|x^k(t_k) - \int_0^t f(x^k(t_k + \tau))\mathrm{d}\tau - x^{k+1}(0) + \int_0^t f(x^{k+1}(\tau))\mathrm{d}\tau\right\| \\ &\leq \left\|x^k(t_k) - x_{k+1}\right\| + L\int_0^t \left\|x^k(t_k + \tau) - x^{k+1}(\tau)\right\| \mathrm{d}\tau. \end{aligned}$$

The Gronwall inequality and Lemma 7.1 yield the assertion. □

THEOREM 8.6. *Let A1, A2, A3, A4, and A5 hold. If the generalized Newton path x^0 is fully contained in $\mathcal{R}$, then there exists an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ iteration* (1.2) *with* (BSC) *converges to x_0^*.*

Proof. Under the stated assumptions and because $\mathcal{R}$ is open, there exists an $\varepsilon > 0$ such that

$$\inf_{y \in \mathcal{S}, t \geq 0} \left\| x^0(t) - y \right\| \geq \varepsilon.$$

Without loss of generality, we can assume that $\varepsilon \leq \varepsilon^*$ from Lemma 6.3. We now choose $T_* < \infty$ such that

$$\left\| x^0(t) - x_0^* \right\| \leq \varepsilon/2 \quad \text{for all } t \geq T_* - 1. \tag{8.6}$$

We observe that $\left\| f(x^0(t)) \right\| > 0$ for all $t \in [0, T]$, because otherwise there is a zero of f from which we can integrate (5.4) backwards to show $F(x^0) = 0$. Thus, we can choose an $\tilde{\varepsilon} \in (0, \varepsilon)$ such that there is a $\Delta > 0$ satisfying

$$\Delta \leq \|f(x)\| \quad \text{for all } t \in [0, T], x \in \mathcal{T}(x_0) \text{ with } \left\| x - x^0(t) \right\| \leq \tilde{\varepsilon}. \tag{8.7}$$

With this Δ, A5 yields constants $\gamma, t_\gamma > 0$ and Lemma 8.2 with $\bar{t} = t_\gamma$ and Theorem 8.4 deliver constants $\overline{H}$ that we can, without loss of generality, decrease to satisfy

$$0 < \overline{H} \leq \min\left(L\overline{M} \|F(x_0)\|, \frac{(\gamma \tilde{\varepsilon})^2}{(T_* e^{LT_*})^2 (L\overline{M} \|F(x_0)\|)^3} \right) \tag{8.8}$$

with $\overline{M}$ from Lemma 5.3. We now choose an H-dependent $\bar{k} \in \mathbb{N}$ that satisfies

$$T_* - 1 \leq \sum_{i=0}^{\bar{k}-1} t_i \leq T_*.$$

The lower (BSC) step size bound (8.3) then delivers a bound for $\bar{k}$ according to

$$T_* \geq \sum_{i=0}^{\bar{k}-1} t_i \geq \bar{k} \sqrt{\frac{H}{L\overline{M} \|F(x_0)\|}}, \quad \text{which implies} \quad \bar{k} \leq \frac{T_* \sqrt{L\overline{M} \|F(x_0)\|}}{\sqrt{H}}. \tag{8.9}$$

It follows from (BSC), (8.7), Lemma 8.2, and A5 that

$$\text{if} \quad \left\| x^0 \left(\textstyle\sum_{i=0}^{k-1} t_i \right) - x_k \right\| \leq \tilde{\varepsilon} \quad \text{then} \quad H \geq t_k \|g(x_k, t_k)\| \geq \gamma t_k^2.$$

Together with Lemma 8.5, (8.9), and (8.8), we obtain by induction that for all $k \leq \bar{k}$

$$\begin{aligned} \left\| x^0 \left(\textstyle\sum_{i=0}^{k-1} t_i \right) - x_k \right\| &\leq \sum_{j=0}^{k-1} \left\| x^j \left(\textstyle\sum_{i=j}^{k-1} t_i \right) - x^{j+1} \left(\textstyle\sum_{i=j+1}^{k-1} t_i \right) \right\| \\ &\leq \sum_{j=0}^{k-1} \frac{1}{2} \|f(x_j)\| L e^{LT_*} t_j^2 \leq \frac{L e^{LT_*} \overline{M} \|F(x_0)\|}{2\gamma} H k \\ &\leq \frac{T_* e^{LT_*} (L\overline{M} \|F(x_0)\|)^{3/2}}{2\gamma} \sqrt{H} \leq \frac{1}{2} \tilde{\varepsilon}. \end{aligned}$$

Therefore, $x_k \in \mathcal{R}$ for all $k \leq \bar{k}$ and we can use (8.6) to obtain

$$\|x_k - x_0^*\| \leq \left\| x_k - x^0 \left(\sum\nolimits_{i=0}^{k-1} t_i \right) \right\| + \left\| x^0 \left(\sum\nolimits_{i=0}^{k-1} t_i \right) - x_0^* \right\| \leq \varepsilon \leq \varepsilon^*.$$

The result then follows from Lemma 6.3. □

9. Affine invariance. We feel it is important here to discuss affine covariance principles of backward step control, because this concept is so closely linked to generalized level functions $T(x|A)$ and the efficiency on badly conditioned nonlinear problems [16]. To this end, let the matrices $A, B \in \mathbb{R}^{n\times n}$ be invertible. We consider the family of problems

$$\widetilde{F}(x) = AF(Bx) \quad \text{with Jacobian } \widetilde{J}(x) = AJ(Bx)B.$$

Furthermore, we assume that $\widetilde{M}(x) = B^{-1}M(Bx)A^{-1}$. The corresponding step function $\widetilde{f}$ then satisfies

$$\widetilde{f}(x) = \widetilde{M}(x)\widetilde{F}(x) = B^{-1}M(Bx)F(Bx).$$

Hence, f is an affine covariant quantity, because it does not depend on the choice of A. Conditions A4 and A5 are thus affine covariant, A2 is affine contravariant, and A3 is neither of the two, because its left-hand side is contravariant, but the right-hand side is covariant.

However, (BSC) is crafted in such a way, that all quantities that are not affine covariant only enter in H. Thus, from a practical perspective, backward step control is an essentially affine covariant method, even though $\overline{H}$ depends on A through A2 and A3. This property explains the practical efficiency of backward step control. The convergence proof, which depends on classical monotonicity arguments of $T(x) = T(x|\mathrm{I}_n)$, is to be understood as an asymptotic safeguard, even if the method itself does not enforce classical monotonicity for reasonable values of H.

10. Algorithmic realization. In order to solve (BSC) in each step, we need to find for given $x \in D$ the first zero of the univariate Lipschitz continuous function $t \mapsto t\,\|g(x,t)\| - H$ on the interval $[0,1]$ or guarantee that there are no zeros for $t \in [0,1]$. We propose two algorithmic realizations of backward step control. The first one is rigorous but costly, the second one is more efficient but can fail to detect the smallest zero in rare cases.

10.1. Monotone Iterations. The rigorous approach is based on a monotone iteration [19], for which we require an overestimate $L' > L$ of the Lipschitz constant in A4. We decompose the univariate function into a difference of strictly monotonically increasing functions $U : [0,\infty) \to [-H,\infty)$ minus $V : [0,1] \to \mathbb{R}$ defined as

$$\begin{aligned} U(t) &= t^2 L' \,\|f(x)\| - H, \\ V(t) &= U(t) - (t\,\|g(x,t)\| - H) = t^2 L' \,\|f(x)\| - t\,\|g(x,t)\|, \end{aligned}$$

where we have dropped the index k for simplicity. The parabola $U(t)$ is obviously strictly monotonically increasing on $[0,\infty)$ provided that $\|f(x)\| > 0$. We have the following lemma regarding the monotonicity of $V(t)$.

LEMMA 10.1. *Let A4 hold. If $\|f(x)\| > 0$, then $V(t)$ is strictly increasing.*

Proof. For $0 \le t_1 < t_2 \le 1$ the reverse triangle inequality and A4 yield

$$\begin{aligned}\|g(x,t_2)\| - \|g(x,t_1)\| &\le \|g(x,t_2) - g(x,t_1)\| \\ &= \|f(x - t_2 f(x)) - f(x) - f(x - t_1 f(x)) + f(x)\| \\ &\le (t_2 - t_1) L \|f(x)\|,\end{aligned}$$

from which we can deduce

$$\begin{aligned}t_2 \|g(x,t_2)\| - t_1 \|g(x,t_1)\| &= (t_2 - t_1)\|g(x,t_2)\| + t_1(\|g(x,t_2)\| - \|g(x,t_1)\|) \\ &\le (t_2 - t_1) t_2 L \|f(x)\| + t_1 (t_2 - t_1) L \|f(x)\| \\ &= (t_2^2 - t_1^2) L \|f(x)\|.\end{aligned}$$

Thus, it follows that

$$\begin{aligned}V(t_2) - V(t_1) &= (t_2^2 - t_1^2) L' \|f(x)\| - (t_2 \|g(x,t_2)\| - t_1 \|g(x,t_1)\|) \\ &\ge (L' - L)(t_2^2 - t_1^2) \|f(x)\| > 0,\end{aligned}$$

which shows the assertion. □

It follows from Lemma 10.1 and $V(0) = 0$ that $V(t)$ is non-negative. Hence, the iteration

$$t^0 = 0, \quad t^{j+1} = U^{-1}(V(t^j)) = \sqrt{(t^j)^2 + \frac{H - t^j \|g(x,t^j)\|}{L' \|f(x)\|}} \tag{10.1}$$

is well-defined. Furthermore, we can deduce from [19, Theorem 3] that (t^j) is monotonically increasing and either converges to the smallest zero in $[0,1]$ or leaves the interval in a finite number of steps. We have thus established that backward step control is realizable. For practical purposes, however, iteration (10.1) is too slow and it is hard to find good overestimates L' of the Lipschitz constant L.

Instead, we suggest to use a second approach, which is much more efficient from a computational point of view. A detrimental effect of the theoretical possibility of convergence to zeros t_k different from the smallest one was not observed in the practical examples in Section 12.3.

10.2. An efficient bisection procedure. Following [5], we propose to solve (BSC) for t_k only approximately using a simple bisection procedure outlined in Algorithm 1. The exact (BSC) satisfies $H = t_k^* \|g(x_k, t_k^*)\|$ if $t_k^* < 1$. We can thus use a bisection root-finding procedure on the scalar equation

$$0 = H_k'(t) - H, \quad \text{with } H_k'(t) = t \|g(x_k, t)\|.$$

Because there is ever only one H_k' needed at a time, we drop the index k in Algorithm 1. We accept t_k if $H_k'(t_k) \in [H^{\mathrm{l}}, H^{\mathrm{u}}]$, as can be seen from the step size adjustment loop in Algorithm 1, lines 8–19.

For the step size prediction in line 6, we pretend that $\|g(x,t)\|$ stays constant

$$H_{k-1}'/t_{k-1} = \|g(x_{k-1}, t_{k-1})\| \approx \|g(x_k, t_k^*)\| = H/t_k^*,$$

which lends itself to the prediction rule

$$t_k^* \approx t_{k-1}' = t_{k-1} H / H_{k-1}'.$$

Algorithm 1 Bisection procedure for backward step control

```
 1: function $x^* = \mathrm{BSC}(x_0, \mathrm{TOL}, H, H^{\mathrm{l}}, H^{\mathrm{u}}, \alpha, t_{\min}, t_{\mathrm{full}}, t_{\mathrm{stall}})$
 2:     $\Delta x_0 \leftarrow -M(x_0)F(x_0)$                                  ▷ evaluate
 3:     $H' \leftarrow H, t_0 \leftarrow 1$                                  ▷ initialize
 4:     for $k = 0, 1, \dots$ do
 5:         if $\|\Delta x_k\| \leq \mathrm{TOL}$ then return solution $x^* \leftarrow x_k$
 6:         $t_k \leftarrow \min(1, t_k(\alpha + (1-\alpha)H/H'))$          ▷ smoothed step size prediction
 7:         $t^{\mathrm{l}} \leftarrow 0, t^{\mathrm{u}} \leftarrow 1$
 8:         loop                                                          ▷ step size adjustment
 9:             if $t_k < t_{\min}$ then error: Minimal step size reached
10:             $x_+ \leftarrow x_k + t_k \Delta x_k, \Delta x_+ \leftarrow -M(x_+)F(x_+)$          ▷ evaluate
11:             $H' \leftarrow t_k \|\Delta x_+ - \Delta x_k\|, t' \leftarrow t_k$
12:             if $H' < H^{\mathrm{l}}$ and $t_k \leq t_{\mathrm{full}}$ then
13:                 $t^{\mathrm{l}} \leftarrow t_k, t_k \leftarrow (t^{\mathrm{u}} + t_k)/2$          ▷ increase
14:             else if $H' > H^{\mathrm{u}}$ then
15:                 $t^{\mathrm{u}} \leftarrow t_k, t_k \leftarrow (t^{\mathrm{l}} + t_k)/2$          ▷ decrease
16:             else break loop
17:             end if
18:             if $|t_k - t'| < t_{\mathrm{stall}} t_k$ then error: Bisection stalled
19:         end loop
20:         $x_{k+1} \leftarrow x_+, \Delta x_{k+1} \leftarrow \Delta x_+$          ▷ perform step
21:     end for
22: end function
```

In order to avoid oscillations in the step size sequence, we add an exponential smoother with smoothing factor $\alpha \in [0, 1]$ according to

$$t_k = \alpha t_{k-1} + (1-\alpha)t'_{k-1} = t_{k-1}\left(\alpha + (1-\alpha)H/H'_{k-1}\right).$$

The smoothed step size prediction plays a crucial role for the efficiency of Algorithm 1. In most cases, $H'_k \in [H^{\mathrm{l}}, H^{\mathrm{u}}]$ already holds for the predicted step size and thus the minimal number of only one evaluation of F and M is necessary in most iterations.

In some cases it is beneficial not to prescribe an absolute value for H but rather a relative $H_{\mathrm{rel}} > 0$ and to set

$$H = H_{\mathrm{rel}} \max(1, \|\Delta x_0\|)$$

after line 2 in Algorithm 1. Of course, also H^{l} and H^{l} must be adapted appropriately. This heuristic can sometimes improve the performance on badly scaled problems.

For the computations in this article we only adapt H and set the remaining values accordingly to $H^{\mathrm{l}} = H\min(0.1, H), H^{\mathrm{u}} = 2H, \alpha = 0.8, t_{\min} = 10^{-14}, t_{\mathrm{full}} = 0.999, t_{\mathrm{stall}} = 10^{-10}$. In Section 12.3, we set $H^{\mathrm{l}} = 0$ for reasons explained below.

11. Choice of H. The choice of H (or rather H_{rel}) is the most influential parameter in backward step control. Modifications of all other parameters recommended at the end of Section 10 play a secondary role for the performance of Algorithm 1. In general, no one-size-fits-all settings are known for any general-purpose solver, so we must necessarily excercise some modesty in our expectations of what can realistically be achieved here, too. Nonetheless, we discuss two use cases.

If we set out to solve a sequence of problems without supervision, then we could apply automatic tuning techniques (see, e.g., [23]) on a representative set of problems

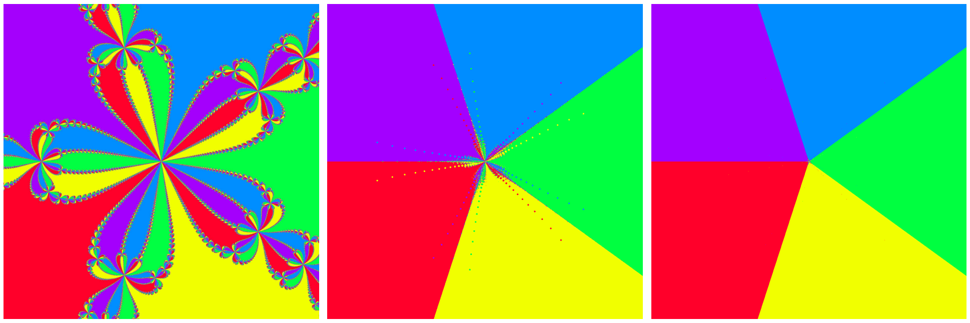

FIGURE 12.1. *Backward step control can eliminate the fractal behavior of the Newton method. The examples shown are for* $H = \infty, 0.1, 0.01$.

of similar size and nonlinearity beforehand. Backward step control is an excellent candidate for these techniques, because only one parameter H_{rel} needs to be adjusted.

If we want to solve a single problem instance under supervision, we recommend as a guiding principle based on the numerical results in Sections 12.2 and 12.3 that the distance $\|\bar{x}_0(t_0) - x_0\| = H$ relative to the size of $\|x_0\|$ is at most 50 %, i.e., to use $H_{\text{rel}} = 1/2$ as a default setting and to successively reduce H_{rel} if Algorithm 1 does not converge. Of course, if the generalized Newton path leads to a singularity, then we must either modify $M(x)$ or find another more suitable initial guess x_0.

12. Numerical results. In this section we illustrate the performance of backward step control for a number of root-finding problems, most of them from applications in the unconstrained optimization of a twice continuously differentiable objective function $\phi : \mathbb{R}^n \to \mathbb{R}$. To find extremal candidates x^*, we compute zeros of the gradient $F(x) := \nabla\phi(x)$. Then, J is the Hessian of ϕ and is thus symmetric.

All computations were performed using the canonical inner product of the Euclidean space $\langle u, v\rangle = u^T v$ for $u, v \in \mathbb{R}^n$. The source code can be found in the supplementary material of this article.

12.1. Newton fractals. It is well known that the convergence behavior of the Newton method exposes fractal structures (see, e.g., the survey article [3]). For instance, let us consider the polynomial function

$$F(x) = x^5 - 1$$

in the complex plane $\mathbb{C}$ identified with $\mathbb{R}^2$. The function F has the five zeros

$$x_j^* = e^{2\pi \mathrm{i} j/5}, \quad j = 0, \dots, 4.$$

We now perform the Newton method for each point x_0 on the unit square and color it with respect to $\arg(x_k)$, the angle of its respective iterate. In case of convergence, each point will be in the same color as the zero x_j^* it converges to. The question of the shape of these attraction basins dates back to the 19th century [7]. Depending on the step size strategy, we obtain different images depicted in Figure 12.1. For full step Newton ($H = \infty$) in the left picture, the fractal structure is clearly visible. Thus, small changes in the initial data can trigger convergence to a different solution. For backward step control (Algorithm 1) with $H = 0.1$ or $H = 0.01$ in the middle and right picture, this unwanted behavior can be virtually eliminated, except for starting points very close to the singularity at $x = 0$.

In order to explain this behavior, we need to compute the solutions of (5.1). An elegant solution can be obtained by considering the equivalent homotopy (5.2), from which it follows that

$$x(t)^5 = ((x_0)^5 - 1)e^{-t} + 1.$$

Thus, the Newton paths exhibit a rotational symmetry around 0 by $2\pi/5$ radians. If we assume the form $(x_0)^5 = a+1$ for some $a \in \mathbb{C}$, we obtain $x(t)^5 = ae^{-t} + 1$ and can determine the starting values that lead to the only singular point $x = 0$ according to

$$x(t) = 0 \quad \Leftrightarrow \quad ae^{-t} = -1 \quad \Leftrightarrow \quad a = -e^t.$$

Hence, the points on the star

$$\mathcal{S}^* = \{x \in \mathbb{C} \,|\, x = rx_j^*, r \leq 0, j = 0, \ldots, 4\}$$

are the only points with emanating Newton paths that lead to the singularity $x = 0$. Thus, the Ribbon Lemma (see supplementary material) says that they divide $\mathbb{C}$ into five sectors that contain each only one root of F. Obviously, all Newton paths emanating from $\mathbb{C} \setminus \mathcal{S}^*$ lead to the respective root and thus Theorem 8.6 explains the five clearly visible sectors for $H = 0.01$. This “defractalization” by backward step control reduces the effect of numerical rounding on the result of iteration (1.2) and reduces the possibility of varying results for the same problem on different hardware.

12.2. Globalization strategies for the Rosenbrock function. One of the most discussed unconstrained optimization problems is the minimization of the two-dimensional Rosenbrock function

$$\phi(x) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2.$$

We investigate globalization alternatives for the Newton method and an initial guess $x_0 = (-10, 10)$ far away from the unique minimizer $x^* = (1, 1)$ of $\phi(x)$. First, we observe that the full step Newton method needs only six iterations until the increment norm is less than 10^{-8}. Thus, a globalization safeguard is not needed and the Rosenbrock example here serves to asses the overhead introduced by the various globalization approaches.

The minimizer x^* lies in the infamous banana-shaped valley $x_2 = x_1^2$. The valley is rather flat at the bottom, but steep on the sides. This is reflected by the ratio of the eigenvalues $\mathrm{cond}(J(x^*)) \approx 2\,500$. This value is typically much higher in real-world applications. Furthermore, the Rosenbrock function is challenging for Newton-type methods because the Newton path lies very close to the set of singularities

$$\mathcal{S}_0 = \{x \in \mathbb{R}^2 \,|\, x_2 = x_1^2 + 1/200\}.$$

The curvature of the rather flat valley renders practically all damping approaches useless that enforce monotonicity of $\|F(x_k)\|$, because once the valley is reached, only small steps along the valley’s tangent are allowed before the objective gradient norm increases. The affine contravariant globalization strategy NLEQ-RES (see [16, Section 3.2]), for instance, takes more than 14,700 iterations with an average step size of 10^{-3}. This behavior is also typical for line-search methods on ϕ, although it is less pronounced in the Rosenbrock example. Qualitatively, this behavior can be observed also in this example with a typical line-search method (e.g., [24, Algorithm 3.1]) with

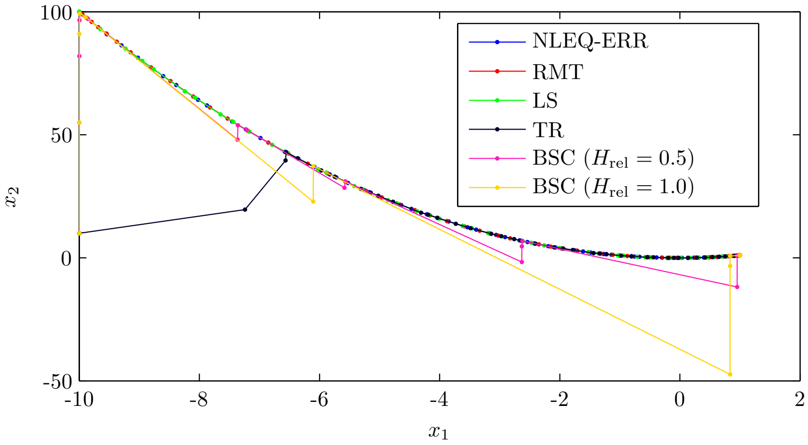

FIGURE 12.2. *Iterates of the different globalized methods of §12.2 for the minimization of the Rosenbrock function in $\mathbb{R}^2$ from the starting point $x_0 = (-10, 10)$ (on the left) to the solution $x^* = (1, 1)$ (on the right). All methods must eventually follow the Newton path, which leads from x_0 straight up close to $(-10, 100)$ and then turns right into the banana valley parabola $x_2 = x_1^2$ until it reaches the solution x^*. Methods that stay extremly close to the Newton path are required to do little steps along the parabola. BSC allows to relax this requirement by larger choices of H_{rel}, leading to long tangential steps (magenta and yellow), which make good progress towards x^*. The sawtooth-like vertical steps back up to the parabola are due to the new Newton path emanating from the current iterate x_k.*

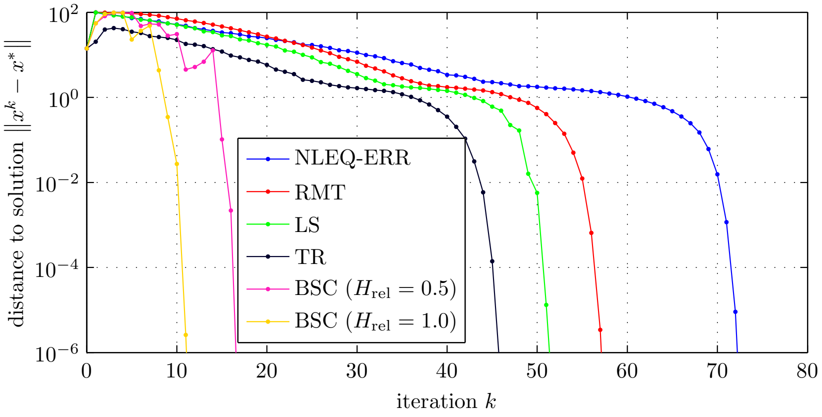

FIGURE 12.3. *Distance $\|x^k - x^*\|$ of the current iterate to the solution for the different globalized methods of §12.2.*

algorithmic parameters $\bar{\alpha} = 1, \rho = \sqrt{0.1}, c = 0.5$ (untypical), requiring a large amount of descent of ϕ in each step. The line-search method then needs 100 steps with an average step size of 0.39.

We now compare five different globalization strategies with commonly used algorithmic parameters:

BSC: Algorithm 1 with the parameters outlined in Section 8 and $H_{\mathrm{rel}} \in \{0.5, 1.0\}$ as motivated in Section 11.

LS: A line-search algorithm [24, Algorithm 3.1] with $\bar{\alpha} = 1, \rho = \sqrt{0.1}, c = 0.01$.

NLEQ-ERR: An error descent oriented affine covariant global Newton method [16,

	NLEQ-ERR	RMT	LS	TR	BSC ($H_{\mathrm{rel}} = 0.5$)	BSC ($H_{\mathrm{rel}} = 1.0$)
$\#\phi$	0	0	69	58	0	0
$\#F$	122	138	53	57	24	18
$\#M$	73	59	53	57	24	18

Table 12.1
Number of function evaluations needed for the minimization of the Rosenbrock function for different globalized methods.

Chapter 3.3] with $t_0 = 1$ and without switching to QNERR. In order to avoid stalling, we had to modify the algorithm in the branch $\lambda_k' \geq 4\lambda_k$ to set $\lambda_k \leftarrow \min(1, 4\lambda_k)$ if $\lambda_k' = 1$, instead of just $\lambda_k \leftarrow \lambda_k'$. The restricted and non-restricted versions yield the same results here.

RMT: The restrictive monotonicity test [5] with $\eta = 1, \eta_\star = 0.8, \eta^\star = 1.2$ without back-projection.

TR: A trust region algorithm [24, Algorithm 4.1] with $\hat{\Delta} = 10^4, \Delta_0 = 10, \eta = 0.2$. For simplicity, we solve the trust region subproblem with high accuracy.

We depict the iterates in Figure 12.2. We first observe that the Newton path heads up vertically from the initial guess and then continues along the banana valley. Only TR can take a shortcut, because it is not a damping strategy. Nonetheless, it must still follow the Newton path close to the solution, when the trust region constraint becomes inactive. In Figure 12.3, we plot the distance from the current iterate to the optimal solution x^*. The number of function evaluations is shown in Table 12.1. We can see that NLEQ-ERR requires the highest number of iterations and also evaluations of M. The RMT requires less evaluations of M but more evaluations of F. They are followed by LS, which relies on additional evaluations of ϕ. Its performance is similar to that that of TR. The two versions of BSC need the least amount of iterations and function evaluations.

12.3. MINRES-BSC for large-scale unconstrained optimization. Now we tackle the 170[1] unconstrained optimization problems from the CUTEst test set [20] with a variant of Algorithm 1 for different values of $H_{\mathrm{rel}} = \infty, 5, 1, 0.5, 0.1$ and compare the results with a trust region and a line search method. We choose the matrix M to be implicitly given by the iterative Krylov-space method MINRES [26] for solving

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

We remark here that in contrast to, e.g., CG [21], MINRES is not affine covariant. However, it is appropriate for semi- or even indefinite matrices and has advantages over CG for early termination and in view of a backward error analysis [18].

In the k-th Newton step and i-th MINRES iteration, the MINRES iterate satisfies

$$\Delta x_k^i = \underset{\Delta x_k^i \in \mathcal{K}_i}{\arg\min} \left\| J(x_k)\Delta x_k^i + F(x_k) \right\|,$$

where $\mathcal{K}_i$ denotes the i-th Krylov space, which depends on the initial residual and $J(x_k)$. In a backward analysis, we can now construct M_i^k according to

$$\Delta x_k^i = -M_i^k F(x_k) \quad \text{and} \quad v = J(x_k) M_i^k v \quad \text{for all } v \in \mathbb{R}^n, \langle v, F(x_k)\rangle = 0. \tag{12.1}$$

[1] We leave out problem FLETCHBV, because a corrected version is provided as FLETCBV2.

If MINRES converges with a given relative tolerance $\kappa' < 1$ in i iterations from the starting guess $\Delta x_k^0 = 0$, then we have with $M(x_k) = M_i^k$ that

$$\kappa' \|F(x_k)\| \geq \left\|J(x_k)\Delta x_k^i + F(x_k)\right\| = \|[\mathrm{I}_n - J(x_k)M(x_k)]\, F(x_k)\|\,, \tag{12.2}$$

which implies together with (12.1) the κ-condition A2 with $\kappa \leq \kappa'$. In this sense, MINRES minimizes κ over the Krylov-space. For the general case when $F(x)$ is not a gradient of a function and $J(x)$ is not symmetric, the same argument applies if we use GMRES [28] instead of MINRES. We prefer MINRES here, because it allows for the exploitation of a three-term recurrence so that we do not need to store an orthonormal basis of the Krylov subspace.

For the computations, we use a fixed relative tolerance of $\kappa' = 10^{-2}$ and at most 100 MINRES iterations, which in turn might lead to higher actual values of κ. We note that $M(x)$ is only piecewise continuous, because the number of MINRES iterations can change depending on x. Hence, we reduce the lower BSC bound H^1 to zero to avoid unnecessary stalling of the bisection procedure in Algorithm 1. Alternatively, $M(x)$ could be made continuous by linearly interpolating the last two iterates of MINRES such that (12.2) holds with equality. The interpolation, however, would lead to performance deterioration because the sometimes considerable gap between the left- and right-hand side of (12.2) cannot be exploited.

In the form described so far, the algorithm is suitable for strictly convex unconstrained nonlinear optimization problems, because then the matrices $J(x)$ and $M(x)$ are guaranteed to be positive definite. It then follows that besides the gradient norm $\|F(x(t))\|$ also the objective function $\phi(x(t))$ decreases along the generalized Newton path $x(t)$ due to

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(x(t)) = -\left\langle \nabla\phi(x(t)), f(x(t))\right\rangle = -\left\langle F(x(t)), M(x(t))F(x(t))\right\rangle\,, \tag{12.3}$$

which is less than zero as long as $\|F(x(t))\| > 0$.

In order to allow for more fairness in the comparison of methods, we use MINRES-BSC on the whole test set including the nonconvex problems. To this end, however, we need to modify $M(x)$, which we propose to do in the following way: On the basis of the MINRES increment $\Delta x_k^{\mathrm{MINRES}}$ we compute the quantity

$$s = \left\langle F(x_k), \Delta x_k^{\mathrm{MINRES}}\right\rangle / \|F(x_k)\|^2 \tag{12.4}$$

and use the following formula as a piecewise definition of a continuous $M(x_k)$

$$-M(x_k)F(x_k) = \begin{cases} \Delta x_k^{\mathrm{MINRES}} & \text{for } s < 0,\\ (1-s)\Delta x_k^{\mathrm{MINRES}} - sF(x_k) & \text{for } s \in [0,1],\\ -F(x_k) & \text{for } s > 1. \end{cases} \tag{12.5}$$

On the space of vectors orthogonal to $F(x_k)$, we can simply assume $M(x_k)$ to be the inverse of $J(x_k)$ like in (12.1). With this choice, we easily obtain ϕ-descent for $s < 0$ and $s > 1$ and nonascent of ϕ for $s \in [0,1]$ from (12.4) and (12.5) due to

$$-\left\langle F(x_k), M(x_k)F(x_k)\right\rangle = [(1-s)s - s]\,\|F(x_k)\|^2 = -s^2\,\|F(x_k)\|^2 \leq 0.$$

In fact, ϕ-stagnation is only possible for $s = 0$, which implies that the MINRES increment is tangent to an isoline. This case is so rare in practice that we condone its possibility without taking further precautions.

A word of caution is appropriate here with regard to nonconvex optimization problems: Although we can avoid attraction of the generalized Newton-path to local maxima and singular points of $J(x)$ satisfying $\|F(x)\| > 0$ with the modification of $M(x)$ by (12.5), the modification can lead to violation of the κ-condition A2, so that our convergence results do not apply. Nonetheless, we observe that Algorithm 1 still performs well in practice even on the nonconvex problems, because of its nature of following the generalized Newton-path, whose existence can also be based on the reduction of $\phi(x)$ instead of $T(x)$ in Lemma 5.5 due to (12.3) provided that there is a constant $\delta > 0$ satisfying $\langle F(x), f(x) \rangle > \delta$ for all $x \in \mathcal{T}(x_0)$.

We compare MINRES-BSC (with $H_{\text{rel}} = \infty, 5, 1, 0.5, 0.1$) with a line-search algorithm [24, Algorithms 3.4–3.5 with $c_1 = 10^{-4}, c_2 = 0.9$][2] using the search direction (12.5), and with a large-scale CG-based interior trust region algorithm [9, 10] implemented in in MathWorks Matlab as `fminunc` with the `Algorithm` option set to `trust-region`. Because large-scale versions of NLEQ-ERR and RMT from Section 12.2 with iterative linear algebra for approximation of the Newton direction are much more involved and not readily available in Matlab, we do not include results for these methods here. In order to keep the comparison fair, we adopt the termination criterion of `fminunc` for the other algorithms: For TOL $= 10^{-5}$, we terminate as soon as

a) $\|F(x_k)\|_\infty <$ TOL, or
b) $\|f(x_k)\|_2 <$ TOL, or
c) $k > 1, t = 1$, and $|\phi(x_k) - \phi(x_{k-1})| < \text{TOL}(1 + |\phi(x_{k-1})|)$.

We present performance profiles (compare [17]) for these methods on an Intel Core i7 2.67 GHz CPU with 24 GB of RAM running Ubuntu 12.04 Linux with CUTEst version 1.1 in Matlab R2013a restricted to one computational thread. We consider a problem unsolved by the respective solver if an error is reported or if the maximum time of 10 minutes per problem is exceeded. In performance profiles, fast methods have graphs that are close to the left axis, while reliable methods that solve many problems have graphs that are close to the upper axis.

In Figure 12.4 we depict the results for those 116 unconstrained problems of the CUTEst test set that satisfy the κ-condition A2 in every iteration of MINRES-BSC for $H_{\text{rel}} = \infty, 5, 1, 0.5, 0.1$, which we check numerically as in (12.2) via

$$\|F(x_k) + J(x_k)\Delta x_k\| < \|F(x_k)\|\,.$$

This condition is violated for some problems because we enforce ϕ-descent via (12.5).

We observe on the right axis that the line search method solves the least overall number of problems. The reasons are the reduction of the step size below 10^{-14} for 4 problems and a runtime above 10 minutes for 5 problems. However, we can see on the left axis that for 63 % of the problems, the line search method is actually the fastest.

The trust region method solves the most problems except for one due to excess of the runtime limit but is slower than the other methods on most problems. The modified MINRES search direction (12.5) with full steps performs surprisingly well on the test set, failing also only on one problem due to excess of the runtime limit. However, this approach lacks theoretical justification insofar as no convergence proof

[2] We use cubic interpolation for finding step sizes that satisfy the strong Wolfe conditions for $t < 1$. If $t > 1$ would be chosen in order to satisfy the curvature condition, we only enforce the sufficient decrease condition for $t = 1$. If an extra gradient evaluation is necessary for cubic interpolation, we use quadratic interpolation. If the interpolation nodes are closer than 10^{-4}, we use a fallback bisection procedure to avoid numerical instabilities.

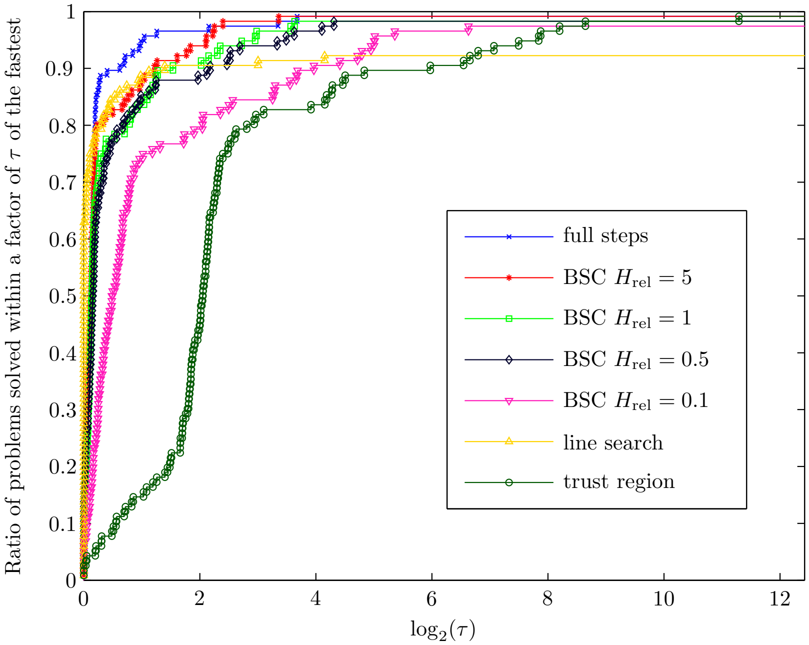

FIGURE 12.4. *Performance plot for those 116 unconstrained optimization problems from the CUTEst test set for which the κ-condition A2 is satisfied in every iteration of MINRES-BSC.*

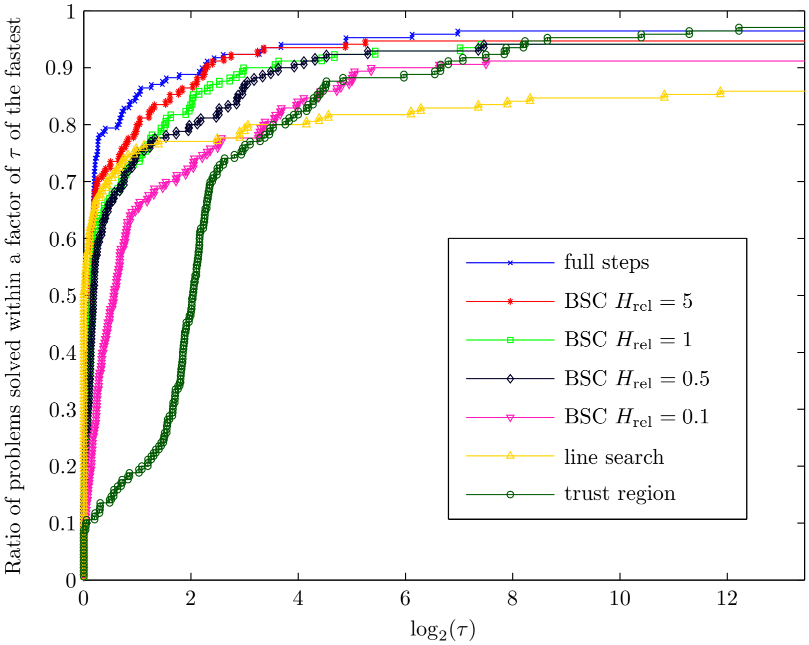

FIGURE 12.5. *Performance plot for all 170 unconstrained optimization problems from the CUTEst test set.*

is achievable for this method. The variants of backward step control combine the positive features of the line search, trust region, and full step methods, if reasonable values of H_{rel} are used (compare Section 11): Backward step control is faster than the considered trust region method on most problems, solves more problems than the considered line search method, and has the theoretical justification of a convergence proof. The number of problems not solved due to excess of the time limit are 1, 2, 2, 1 for $H_{\mathrm{rel}} = 5, 1, 0.5, 0.1$, respectively. For $H_{\mathrm{rel}} = 0.1$, 2 additional problems cannot be solved due to reduction of the step size below 10^{-14}. The one problem that cannot be solved with full steps and $H_{\mathrm{rel}} = 5$ is TOINTPSP [31], which can be solved in well under one second for the choices of $H_{\mathrm{rel}} = 1, 0.5, 0.1$.

We also present the performance profiles for the full set of unconstrained problems in Figure 12.5. The results are similar to the results on the reduced test set, even though the convergence proof of Section 8 does not apply.

13. Conclusions. We presented backward step control as a novel damping strategy for the globalization of convergence of Newton-type methods for the numerical solution of nonlinear equations. For the first time it is possible to give a convergence proof for an efficient globalization strategy based on the Newton path, or more precisely, even on generalized Newton paths. Although the assumptions and the proof are not oriented towards affine invariance principles, the method is essentially of affine covariant type, which makes it especially suited for highly nonlinear problems with high condition numbers of the Jacobian. The mathematical setting even allows for a-priori bounds on the residual reduction, for proving convergence to the specific solution x_0^*, and for guaranteed transition to full steps $t_k = 1$ in the vicinity of a solution. Moreover, we provided two algorithmic realization of backward step control, one based on monotone iterations and one on a bisection procedure. With the bisection approach, backward step control performs reliably and efficiently, also in conjunction with MINRES.

In order to address a wide audience, we have intentionally kept the setting here to the finite dimensional case while at the same time trying to keep the main lines of argument (e.g., the use of a general inner product, the Picard-Lindelöf Theorem) and Algorithm 1 in such a way that they can be used in an infinite dimensional Hilbert space setting as well. The subtle and intricate adjustment of assumptions, especially A1, or of the derivation of estimates like $\overline{m}$ in Lemma 5.3 shall be the subject of follow-up research.

Acknowledgments. We would like to thank E. Kostina, H.G. Bock, and J.P. Schlöder for advice, especially on the Restrictive Monotonicity Test, C. Kirches for comments on an earlier draft of this paper, and two unknown referees for their diligent and encouraging comments on this article.

REFERENCES

[1] H. Amann. *Ordinary differential equations*, volume 13 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1990.

[2] U. Ascher and M.R. Osborne. A note on solving nonlinear equations and the natural criterion function. *Journal of Optimization Theory and Applications*, 55(1):147–152, 1987.

[3] P. Blanchard. Complex analytic dynamics on the Riemann sphere. *Bull. Amer. Math. Soc. (N.S.)*, 11(1):85–141, 1984.

[4] H.G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, volume 183 of *Bonner Mathematische Schriften*. Universität Bonn, Bonn, 1987.

[5] H.G. Bock, E.A. Kostina, and J.P. Schlöder. On the role of natural level functions to achieve global convergence for damped Newton methods. In M.J.D. Powell and S. Scholtes, editors, *System Modelling and Optimization. Methods, Theory and Applications*, pages 51–74. Kluwer, 2000.

[6] P.T. Boggs. The solution of nonlinear systems of equations by A-stable integration techniques. *SIAM J. Numer. Anal.*, 8:767–785, 1971.

[7] A. Cayley. Desiderata and Suggestions: No. 3. The Newton-Fourier Imaginary Problem. *Amer. J. Math.*, 2(1):97, 1879.

[8] E.A. Coddington and N. Levinson. *Theory of ordinary differential equations.* Tata McGraw-Hill Publishing Company Limited, New Delhi, 1998.

[9] T.F. Coleman and Y. Li. On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Math. Program.*, 67(1-3):189–224, 1994.

[10] T.F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. on Optim.*, 6(2):418–445, 1996.

[11] A.R. Conn, N.I.M. Gould, and P.L. Toint. *Trust region methods*, volume 1. SIAM, 1987.

[12] D.F. Davidenko. On a new method of numerical solution of systems of nonlinear equations. *Doklady Akad. Nauk SSSR (N.S.)*, 88:601–602, 1953.

[13] R.S. Dembo, S.C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Numer. Anal.*, 19(2):400–408, 1982.

[14] P. Deuflhard. *Ein Newton-Verfahren bei fastsingulärer Funktionalmatrix zur Lösung von nichtlinearen Randwertaufgaben mit der Mehrzielmethode.* PhD thesis, Universität zu Köln, 1972.

[15] P. Deuflhard. A Modified Newton Method for the Solution of Ill-conditioned Systems of Nonlinear Equations with Applications to Multiple Shooting. *Numerische Mathematik*, 22:289–311, 1974.

[16] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, volume 35 of *Springer Series in Computational Mathematics.* Springer, 2006.

[17] E.D. Dolan and J.J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91(2):201–213, 2002.

[18] D.C.-L. Fong and M.A. Saunders. CG versus MINRES: An empirical comparison. *SQU Journal for Science*, 17(1):44–62, 2012.

[19] A. Galántai and J. Abaffy. Always convergent iteration methods for nonlinear equations of Lipschitz functions. *Numer. Algor.*, pages 1–11, 2014.

[20] N.I.M. Gould, D. Orban, and P.L. Toint. CUTEst: a constrained and unconstrained testing environment with safe threads. Technical report, Rutherford Appleton Laboratory, Chilton, Oxfordshire, England, EU, 2013.

[21] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952.

[22] J.J. Moré. The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis*, volume 630 of *Lecture Notes in Mathematics*, pages 105–116. Springer, 1978.

[23] K. Naono, K. Teranishi, J. Cavazos, and R. Suda, editors. *Software Automatic Tuning.* Springer, New York, 2010.

[24] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer Verlag, Berlin Heidelberg New York, second edition, 2006.

[25] J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Academic Press, New York, 1970.

[26] C.C. Paige and M.A. Saunders. Solutions of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975.

[27] A. Potschka. *A direct method for parabolic PDE constrained optimization problems.* Advances in Numerical Mathematics. Springer, 2013.

[28] Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 7:856–869, 1986.

[29] D.C. Sorensen. Newton's method with a model trust region modification. *SIAM J. on Numer. Anal.*, 19(2):409–426, 1982.

[30] J. Stoer and R. Bulirsch. *Introduction to numerical analysis.* Number 12 in Texts in applied mathematics. Springer, 2008.

[31] P.L. Toint. Some numerical results using a sparse matrix updating formula in unconstrained optimization. *Math. Comput.*, 32(1):839–852, 1978.