MATHEMATICS OF COMPUTATION
VOLUME 57, NUMBER 196
OCTOBER 1991, PAGES 673-701

THE MODIFIED NEWTON METHOD IN THE SOLUTION OF STIFF ORDINARY DIFFERENTIAL EQUATIONS

ROGER ALEXANDER

ABSTRACT. This paper presents an analysis of the modified Newton method as it is used in codes implementing implicit formulae for integrating stiff ordinary differential equations. We prove that near a smooth solution of the differential system, when the Jacobian is essentially negative dominant and slowly varying, the modified Newton iteration is contractive, converging to the locally unique solution—whose existence is hereby demonstrated—of the implicit equations. This analysis eliminates several common restrictive or unrealistic assumptions, and provides insight for the design of robust codes.

1. BACKGROUND, RESULTS, SIGNIFICANCE

1.1. **Prototype stiff problems. Their salient properties.** Note the structure o the solution of the model differential equation

$$y' = \lambda y + \cos t, \qquad \lambda \ll -1, \tag{1.1}$$

namely,

$$y(t) = e^{\lambda t}\left(y(0) + \frac{\lambda}{1+\lambda^2}\right) + \frac{1}{1+\lambda^2}(\sin t - \lambda \cos t).$$

There is an initial transient of duration $O(|\lambda^{-1}|\log|\lambda|)$, after which the tern $e^{\lambda t}$ is not active and the solution is as smooth as $\cos t$.

Under suitable conditions, see [29] and *infra*, solutions of the stiff time varying linear system

$$y' = B(t)y + g(t), \qquad y \in \mathbb{R}^N, \tag{1.2}$$

have the same structure: $y(t)$ is the sum of a smooth particular solution $y_s(t$ and a transient $v(t)$. The transient, a solution of the homogeneous equation

$$v'(t) = B(t)v(t), \qquad v(0) = y(0) - y_s(0),$$

expires after a short time. Meanwhile, $y_s(t)$ and its derivative have bounds expressible in terms of

$$B^{-1}(t)\frac{d^\nu B}{dt^\nu}, \qquad B^{-1}(t)\frac{d^\nu g}{dt^\nu}, \qquad \nu = 0, 1. \tag{1.3}$$

Received September 28, 1989; revised August 10, 1990.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 65L05, 65H10.
Key words and phrases. Stiff differential equations, modified Newton method.
Supported in part by NSF Grant MCS79-03408.

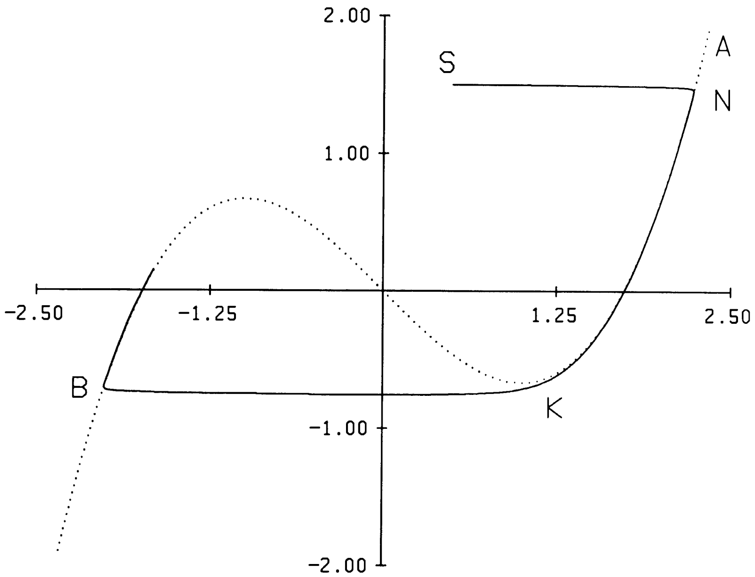

FIGURE 1
Van der Pol equation, $\varepsilon = 0.01$

No new transient appears so long as the quantities (1.3) are of moderate size.

In this work we show that the conditions responsible for this structure of solutions make possible the strategies embodied in codes for stiff problems. We consider one more example.

The solutions of a *nonlinear* stiff system may be smooth for only a limited time. Take Van der Pol's equation

$$\frac{d^2u}{d\tau^2} + \gamma(u^2 - 1)\frac{du}{d\tau} + u = 0, \qquad \gamma \gg 1.$$

Scale the independent variable by $\tau = t/\sqrt{\varepsilon}$, $\varepsilon = \gamma^{-2}$, and make the change of variables

$$x = u, \qquad y = \varepsilon\dot{u} + \tfrac{1}{3}u^3 - u, \qquad \cdot = d/dt,$$

to arrive at the singularly perturbed (Liénard) form

$$\varepsilon\dot{x} = y - (\tfrac{1}{3}x^3 - x), \qquad \dot{y} = -x. \tag{1.4}$$

Figure 1 displays the familiar approach to the limit cycle [31]. The trajectory originating at S undergoes an initial transient, moving rapidly to the right branch ANK of the cubic. The solution is smooth as it moves down the cubic, but this smooth phase endures only until the neighborhood of the knee of the cubic at K is reached. Then the solution "jumps" rapidly to the left branch of the cubic at B. These alternating smooth segments and jumps are characteristic of relaxation oscillators.

We now consider numerical methods. Stiff differential equations pose a fundamental problem: how to compute smooth solutions efficiently. But a smooth solution need not *remain* smooth, and in a robust code for stiff systems care will be taken that shortcuts adopted for efficiency's sake are justified.

1.2. **The modified Newton method.** Let the stiff problem be

$$y' = f(t, y), \qquad f: \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N, \qquad f \in C^{p+1}, \quad p \geq 1, \tag{1.5}$$

and suppose that a smooth phase of the solution has been reached. Integrating (1.5) with a stable implicit formula, one must solve at each time step one or more systems of equations of the form

$$y = \psi + \alpha h f(t, y) \tag{1.6}$$

for a vector $y \in \mathbb{R}^N$. Here, h is the step size, $\alpha > 0$ is a constant determined by the numerical method, t is a value of the independent variable, and ψ is a known vector. The equations arising from general implicit Runge-Kutta methods also have essentially this structure, as we discuss later on.

Functional iteration in (1.6) will not converge, so one resorts to a modification of Newton's method. The residual in (1.6) is

$$G(\psi, y) := \psi + \alpha h f(t, y) - y. \tag{1.7}$$

For an iteration matrix one takes

$$M = I - \alpha_0 h_0 J_0, \tag{1.8}$$

J_0 being the Jacobian $(\partial f_i / \partial y_j)$, or a divided difference approximation to it, at some point (t_0, y_0) in the past. Note that α_0, h_0 in (1.8) may differ from α, h in (1.6). From an initial guess y^0 for the solution of (1.6) a sequence of iterates is generated by

$$y^{m+1} = y^m + M^{-1} G(\psi, y^m), \qquad m = 0, 1, 2, \ldots . \tag{1.9}$$

Because it is expensive to evaluate Jacobian matrices and factor them, the LU decomposition of M is formed and then retained for several steps, possibly even through changes in step size, so long as convergence is satisfactory.

Is there convergence?

There is some relief in knowing that y^0 is ordinarily a good approximate solution of (1.6). Remember that we are computing a smooth trajectory, so that extrapolation from the solution history is justified. Hence we are not trying to achieve global convergence from a possibly poor initial guess. We expect (1.9) to terminate after just a few iterations with an acceptable solution. Indeed, it is good practice [37] to insist that (1.9) be contracting, not merely asymptotically convergent, even with a Jacobian that is considerably out of date.

In the present work we show that the algorithm in (1.9) is justified near a smooth trajectory of a stiff problem. We prove that in a suitable neighborhood of a smooth solution of (1.5)—its size not restricted by stiffness—the iteration (1.9) is contracting. Any y^0 in that neighborhood generates a sequence convergent to a locally unique fixed point, a solution of (1.6) whose existence is *demonstrated* by our analysis. This is the theoretical support for the strategy embodied in (1.9). From the analysis itself we shall see how better to make crucial decisions—e.g., whether in a case of too slow convergence to update the iteration matrix or to abandon the step. Moreover, we shall derive error estimates that provide a robust basis for stopping the iteration and accepting an approximate solution.

Our analysis also applies to Runge-Kutta methods, which we now describe. A q-stage Runge-Kutta formula is characterized by coefficient arrays (c, A, b), $A \in \mathbb{R}^{q\times q}$, $c, b \in \mathbb{R}^q$. To advance the approximate solution of (1.5) from (t, y_n) to $(t+h,\ y_{n+1})$, one must solve the system of equations

$$(1.10)\qquad y_{n,i} = y_n + \sum_{j=1}^{q} a_{ij} h f(t+hc_j,\ y_{n,j}), \qquad i = 1, 2, \dots, q,$$

for the $y_{n,i}$; then the solution is advanced by the quadrature formula

$$(1.11)\qquad y_{n+1} = y_n + \sum_{i=1}^{q} b_i h f(t+hc_i,\ y_{n,i}).$$

Now (1.6) to (1.9) are modified as follows. Consider first the case of invertible A. Refer to the equation (1.10) and put

$$(1.12)\qquad Y := \begin{pmatrix} y_{n,1} \\ \vdots \\ y_{n,q} \end{pmatrix}, \qquad F(Y) := \begin{pmatrix} f(t+hc_1,\ y_{n,1}) \\ \vdots \\ f(t+hc_q,\ y_{n,q}) \end{pmatrix}.$$

Finally, with $e := (1, \dots, 1)^T \in \mathbb{R}^q$, let

$$(1.13)\qquad \Psi := e \otimes y_n.$$

Then (1.10) takes the form

$$(1.6a)\qquad Y = \Psi + (A \otimes I_N) h F(Y).$$

(We designate the $\nu \times \nu$ identity matrix by I_ν, omitting the subscript when the context dictates its value.) The residual is

$$(1.7a)\qquad G(\Psi, Y) := \Psi + (A \otimes I) h F(Y) - Y.$$

Take for the iteration matrix

$$(1.8a)\qquad M := I_q \otimes I_N - A \otimes h_0 J_0;$$

Then the iteration (1.9) is replaced by

$$(1.9a)\qquad Y^{m+1} = Y^m + M^{-1} G(\Psi, Y^m), \qquad m = 0, 1, 2, \dots.$$

As written, (1.9a) entails the solution of an $Nq \times Nq$ linear system at every iteration, but substantial simplifications are possible. If A is lower triangular with equal diagonal entries $A_{ii} = \alpha$, then M in (1.8a) is block lower triangular, and each diagonal block is the M of (1.8). A general matrix A can be brought to triangular or other simple form by a similarity transformation, greatly reducing the operation count in solving linear systems with coefficient matrix M [45]. Formulas in which A has a single (positive) real eigenvalue have received major attention, see [2, 7, 3] and works cited therein. More recently, in [28] it is shown that on a parallel computer there is a significant advantage for formulas with distinct eigenvalues, i.e., diagonalizable A.

We have to modify the description (1.6a)–(1.9a) if A is not invertible. Instead of seeking the most generality, we consider here only the important class of *formulas of special type* ("Schémas particuliers de Runge-Kutta" [7]):

(1.14a) the first quadrature node is 0, and the first row of A is 0^T;

(1.14b) the last quadrature node is 1, and the last row of A is b^T;

(1.14c) the matrix derived from A by deleting the first column has linearly independent columns.

The Lobatto III A formulas (including the trapezoid rule) are among this class. A-stable diagonally implicit formulas of special type were first systematically studied by R. Alt in his 1973 Paris thesis, reported in [1, 7]. Recently, J. C. Butcher has renewed interest in formulas of special type [3].

If the formula is of special type, let us assign the index $i = 0$ to the first stage, and write the coefficient arrays as $(\hat{c}, \widehat{A}, \hat{b})$, with $\hat{c}, \hat{b} \in \mathbb{R}^{[0:q]}$, $\widehat{A} \in \mathbb{R}^{[0:q]\times[0;q]}$. We also write $a_0 = (\hat{a}_{10}, \hat{a}_{20}, \ldots \hat{a}_{q0})^T \in \mathbb{R}^q$ for the last q entries in the zeroth column of $\widehat{A}$. Since the first row of $\widehat{A}$ is $\widehat{0}^T \in \mathbb{R}^{[0:q]}$, equation (1.10) with $i = 0$ is just

$$y_{n,0} = y_n.$$

Thus, if we write A for the lower right $q \times q$ submatrix of $\widehat{A}$ and define Y and $F(Y)$ as in (1.12), then equation (1.10) becomes

$$Y = e \otimes y_n + a_0 \otimes hf(t, y_n) + (A \otimes I)hF(Y),$$

and this is in the form (1.6a) with an invertible A if we let

$$\Psi := e \otimes y_n + a_0 \otimes hf(t, y_n). \tag{1.13a}$$

We used only the properties (1.14a), (1.14c) of a formula of special type: (1.14b) was not needed. So already we can handle a wider class with singular A. It is easy to see in general that if A is singular then some $y_{n,i}$ in (1.10) can be expressed in terms of y_n and the other $y_{n,j}$, that is, to a zero eigenvalue of A there corresponds an explicit stage. Eliminating it leads to a smaller system resembling (1.10), with the multiplicity of the zero eigenvalue of A reduced by one. Repeating, if necessary, we eventually reach a system like (1.6a) with a nonsingular A. We omit the details, for we know no examples of a formula that would make it interesting or useful to carry them out.

1.3. **The neighborhood of a smooth solution.** The goal of this paper is to give conditions for convergence of iterations (1.9) and (1.9a), applied to compute smooth solutions in stiff differential equations. To describe the neighborhood of a smooth solution of a stiff problem, we need some concepts.

We use the maximum norm for vectors $x = (x_1, \dots, x_N)^T \in \mathbb{R}^N$: $|x| = \max_{1 \le i \le N} |x_i|$. The subordinate matrix norm will be written $|B|$, that is, $|B| = \max_{x \ne 0} |Bx|/|x|$.

Let $y = y(t)$ be a solution of (1.5) on some interval $T_0 \le t \le T_1$. With a constant τ exceeding the maximum tolerable deviation of a numerical approximation, let U be the tubular domain [10]

$$U = \{(t, y): |y - y(t)| < \tau,\ T_0 \le t \le T_1\}. \tag{1.15}$$

The next two definitions are slightly modified from [29].

Definition 1.1. A matrix function $\Lambda: U \to \mathbb{C}^{N \times N}$ is *negative dominant* in U if there is a constant ρ of moderate size such that for all $(t, y) \in U$

$$|\operatorname{Im} \Lambda_{ii}| \le \rho |\operatorname{Re} \Lambda_{ii}|, \qquad i = 1, 2, \dots, N, \tag{1.16}$$

and constants $\sigma > 0$ and δ with $0 < \delta < 1$ such that for all $(t, y) \in U$,

$$\operatorname{Re} \Lambda_{ii} < -\sigma, \qquad \sum_{\substack{j=1 \\ j \ne i}}^{N} |\Lambda_{ij}| < -(1 - \delta) \operatorname{Re} \Lambda_{ii}, \qquad i = 1, 2, \dots, N. \tag{1.17}$$

Λ is called *essentially negative dominant* if in place of (1.16) and (1.17) the inequalities

$$|\operatorname{Im} \Lambda_{ii}| \le \rho |\operatorname{Re} \Lambda_{ii}| + c, \quad i = 1, 2, \dots, N, \tag{1.16a}$$

$$\operatorname{Re} \Lambda_{ii} < c, \quad i = 1, 2, \dots, N, \tag{1.17a}$$

$$\sum_{\substack{j=1 \\ j \ne i}}^{N} |\Lambda_{ij}| \le \begin{cases} -(1 - \delta) \operatorname{Re} \Lambda_{ii} + c & \text{if } \operatorname{Re} \Lambda_{ii} < 0, \\ c - \operatorname{Re} \Lambda_{ii} & \text{if } \operatorname{Re} \Lambda_{ii} \ge 0, \end{cases} \tag{1.17b}$$

$i = 1, 2, \dots, N$, hold with a constant c of moderate size.

We designate quantities of moderate size to distinguish them from quantities like ε^{-1} in (1.4) characterizing the stiffness of the problem, which are permitted to be arbitrarily large.

Definition 1.2. A matrix function $\Lambda: U \to \mathbb{C}^{N \times N}$ is *slowly varying* to order p in U if there are constants $K_{1,\nu}$ of moderate size such that for $(t, y) \in U$ and multi-indices $\nu = (\nu_0, \nu_1, \dots, \nu_N)$ with $|\nu| = \nu_0 + \nu_1 + \dots + \nu_N \le p$, $D^\nu = \frac{\partial^{\nu_0}}{\partial t^{\nu_0}} \frac{\partial^{\nu_1}}{\partial y_1^{\nu_1}} \cdots \frac{\partial^{\nu_N}}{\partial y_N^{\nu_N}}$,

$$\min\left(|\Lambda_{ii}^{-1}|, 1\right) |D^\nu \Lambda_{ij}| \le K_{1,\nu}, \qquad i, j = 1, 2, \dots, N, \ \nu = 1, \dots, p. \tag{1.18}$$

Here is the basic condition imposed on (1.5).

Assumption 1.1. *The Jacobian* $J(t, y) = f_y(t, y)$ *is essentially negative dominant and slowly varying to order* 1 *in* U.

This is not too restrictive, even for linear differential equations; see the cogent examples presented in [29]. Curtis [8] warns that this analysis seems to rest on a segregation of the eigenvalues of A into "stiff" and "nonstiff" groups, but this separation is not rigid: no gap is assumed to exist, and Kreiss explicitly disavows an assumption that the number of "large" eigenvalues is constant. The experience with large, sparse stiff problems reported in [8] is that permitting only diagonal elements to be used as pivots in computing matrix factorizations gives acceptable stability, and this tends to confirm the supposition that off-diagonal elements are not arbitrarily large relative to the diagonal.

Singularly perturbed problems satisfy Assumption 1.1 in the neighborhood of a stable reduced solution; cf. [43]. In Van der Pol's equation (1.4) the Jacobian

$$J = \begin{bmatrix} -\varepsilon^{-1}(x^2 - 1) & \varepsilon^{-1} \\ -1 & 0 \end{bmatrix}$$

is essentially negative dominant if $|x| > \sqrt{(2 - \delta - c\varepsilon)/(1 - \delta)}$. Observe in Figure 1 that the segments of the cubic in $1 < |x| < \sqrt{2}$ are the approaches to the "jumps" at the knees of the cubic; in those intervals the smoothness of the solution deteriorates. The only nonzero term from (1.18) in J is

$$J_{11}^{-1} \frac{\partial J_{11}}{\partial x} = \frac{2x}{x^2 - 1};$$

this too is of moderate size for $|x| > 1 + \delta$, and grows without bound as either knee at $x = \pm 1$ is approached.

We also require smoothness of the solution of (1.5).

Assumption 1.2. *There exist an integer* $p \geq 1$ *and constants* $K_{2,j}$, $j = 1, 2, \ldots$, $p + 1$, *of moderate size such that* $|d^j y(t)/dt^j| \leq K_{2,j}$ *for* $T_0 \leq t \leq T_1$, $j = 1, 2, \ldots, p + 1$.

It follows from Assumption 1.2 that there is a constant K_0 depending only on τ, $K_{2,1}$ and $K_{2,2}$ such that any two points $(t, y) \in U$, $(u, z) \in U$, are connected by a broken line in U of length not greater than $K_0 \max\{|t-u|, |y-z|\}$.

A solution of (1.5) satisfying Assumption 1.2 will be called *smooth.* The examples of §1.1 illustrate the occurrence of smooth solutions. For many systems satisfying Assumption 1.1 it may be shown that *smooth solutions exist,* and the constants $K_{2,j}$ depend only on c, δ, σ, ρ and the $K_{1,\nu}$ for $|\nu| \leq j$: see [29] for linear time-varying systems, [43, Chapter 6] for singularly perturbed systems. In these cases, then, among others, Assumption 1.2 is not an additional restriction; it merely signifies an intention to study smooth solutions known to be present. If a different condition is imposed on f—we shall consider monotonicity, for example, later on—smooth solutions need not exist. The use of codes designed for stiff problems is questionable unless the solution is smooth sometime [42].

1.4. **Convergence of the iteration: the main theorems.** To the assumptions of §1.3 on the differential equation and its solution we add a condition on the numerical method. Consider first a consistent linear multistep formula leading to (1.6) [22]:

$$\sum_{i=0}^{k} [a_i y(t - s_i h) - h b_i y'(t - s_i h)] = O(h^2).$$

Take $s_0 = 0$; the common fixed-stepsize situation is $s_i = i$. We make

Assumption 1.3. $\alpha := b_0/a_0 > 0$.

For smooth functions y,

$$y(t) - \alpha h y'(t) = \sum_{i=1}^{k} \left[-\left(\frac{a_i}{a_0}\right) y(t - s_i h) + \left(\frac{b_i}{a_0}\right) y'(t - s_i h) \right] + O(h^2),$$

so the sum of bracketed terms is $y(t - \alpha h) + O(h^2)$. Now ψ in (1.6) is derived from this last expression by replacing $y(t - s_i h)$ and $y'(t - s_i h)$ by memorized approximations. This explains the condition in Theorem 1 below that ψ should lie in a neighborhood of $y(t - \alpha h)$. A similar observation holds for other representations of multistep methods, with step size fixed or not.

We write $U_t = \{y \in \mathbb{R}^N \mid (t, y) \in U\}$ for the t-sections of the tube U containing the smooth solution.

Finally, recall that a sequence y^m in $\mathbb{R}^N$ *converges q-linearly* to y^* if for a constant C with $0 < C < 1$

$$|y^{m+1} - y^*| < C|y^m - y^*|, \qquad m = 0, 1, 2, \ldots,$$

and C is called the *rate of convergence* [15, pp. 20–21].

Theorem 1. *Let k be given with $0 < k < 1$. There exist positive constants H_0, ξ_0, η_0, and neighborhoods V, B, W with*

$$y(t - \alpha h) \in V \subset U_{t-\alpha h}, \qquad y(t) \in B \subset U_t, \quad W \subset U,$$

all of sizes depending only on k and the constants in Assumptions 1.1 and 1.2, such that if $\psi \in V$, J_0 is $J(t_0, y_0)$ (or a difference approximation to it with step lengths less than ξ_0) with $(t_0, y_0) \in W$, $0 < h < H_0$, $|h\alpha/(h_0\alpha_0) - 1| < \eta_0$, and $y^0 \in B$ is arbitrary, then the iteration (1.9) commencing with y^0 generates a sequence in B converging q-linearly with rate not worse than k to a fixed point y^* of (1.9), the unique solution of (1.6) in B. Moreover $y^* = y^*(\psi)$ depends continuously on $\psi \in V$.

It is important that the constants and the neighborhoods are of moderate size commensurate with the smoothness of the solution, but independent of the stiffness of the problem. This is moreover a “semilocal” convergence theorem: the existence of the solution y^* is not assumed beforehand but instead is a consequence of the proof.

Consider next a Runge-Kutta formula (c, A, b). If the coefficient matrix A is lower triangular with all diagonal entries equal to $\alpha > 0$, Theorem 1 applies to the equations (1.10) taken in succession. For general Runge-Kutta formulae we need a hypothesis.

Assumption 1.3a. *The spectrum of A is contained in the open right half-plane together with $\{0\}$. If A is not invertible, the formula is of special type* (1.14a–c).

In the statement of Theorem 2 we use the local solution z of (1.5) with initial condition $z(t) = y_n$. We assume this solution exists and is smooth on an interval spanning the contemplated step, and write

$$Z(t) := (z(t+hc_1)^T, \dots, z(t+hc_q)^T)^T.$$

This is no restriction, for if transients are active in the local solution, the strategy described here should not be attempted.

Theorem 2. *Assume that the local solution z satisfies Assumption* 1.2. *Let k be given with $0 < k < 1$. There are neighborhoods V, B, W with*

$$e \otimes y_n \in V \subset (U_t)^q, \qquad Z(t) \in B \subset \prod_{i=1}^{q} U_{(t+c_ih)}, \qquad (t, y_n) \in W \subset U$$

(*if the formula is of special type, then $e \otimes y_n + a_0 \otimes hf(t, y_n) \in V \subset \prod_{i=1}^q U_{t+\hat{a}_{i0}h}$*) *and positive constants H_0, ξ_0, η, all of sizes depending only upon k and the constants in Assumptions* 1.1–1.3a, *such that if $\Psi \in V$, J_0 is $J(t_0, y_0)$ (or a divided difference approximation to it with step lengths less than ξ_0) with $(t_0, y_0) \in W$, $0 < h < H_0$, $|h/h_0 - 1| < \eta$, and $Y^0 \in B$ is arbitrary, then the iteration* (1.9a) *commencing with Y^0 generates a sequence in B converging q-linearly with rate at worst k to a fixed point Y^* of* (1.9a), *the unique solution in B of* (1.6a). *The fixed point $Y^* = Y^*(\Psi)$ depends continuously on Ψ.*

Again the step size h is restricted only by the smoothness of the local solution, and not by the stiffness.

1.5. **Discussion.** Previous studies of the solution of (1.6) or (1.6a) fall roughly into two groups. Analysts in the first group, e.g., [35, 37–38], assume the existence of a solution, and assume that the iteration is convergent, and study the effect of various economy-motivated compromises of Newton's method. This is reasonable because just this situation is encountered in practical codes. Our present work shows that their assumptions are actually *consequences* of the problem structure and that of the numerical method, and our analysis refines and extends their insights for algorithm design; see §4.

The second group of studies has been mainly concerned with establishing existence and uniqueness of the solution of the equations (1.6a) arising from Runge-Kutta methods, without regard to how that solution is to be computed.

Often, a fairly severe condition such as algebraic stability is imposed on the formula; contrast this with Assumption 1.3a.

At various phases in the analysis authors in both groups have assumed that J is similar to a diagonal matrix [35, 38, 14, §5.2]. We do not make this assumption.

In the analysis of Runge-Kutta methods it is often assumed that the differential equation is monotone [6, 27, 32, 33, 30, 44], i.e., there is an inner product $\langle \cdot , \cdot \rangle$ on $\mathbb{R}^N$ such that

$$\langle y - z , \ f(t, y) - f(t, z) \rangle \le 0.$$

Alternatively, [5, 10, 12, 14, Chapter 5; 25, 17], the differential equation is assumed to satisfy a one-sided Lipschitz condition,

$$\langle y - z , \ f(t, y) - f(t, z) \rangle \le \nu \langle y - z , \ y - z \rangle ,$$

with $\nu \le 0$ (f monotone) or $\nu > 0$ and of moderate size.

The drawbacks to the assumption that f is monotone have been explained with cogent examples [29]. Perfectly ordinary problems need not satisfy a one-sided Lipschitz condition, either: the Van der Pol equation (1.4) does not; or consider $f(t, y) = Jy$ with J the negative dominant matrix

$$\begin{bmatrix} -1 & 0 \\ (1-\delta)\varepsilon^{-1} & -\varepsilon^{-1} \end{bmatrix} , \qquad 0 < \delta < 1 , \quad 0 < \varepsilon \ll 1 .$$

(Take $y - z = (u_1 , u_2)^T$, with $u_1 = r\cos\theta$, $u_2 = r\sin\theta$, $\theta = \tan^{-1} \frac{1}{2}(1-\delta) < \frac{\pi}{4}$; then

$$(y-z)^T J (y-z) = r^2 \left(-\cos^2\theta + \varepsilon^{-1} \frac{\sin^2\theta}{\cos 2\theta} \right)$$

is not bounded above uniformly in r^2 as $\varepsilon \to 0^+$.) Thus, the one-sided Lipschitz condition is an unnecessary restriction.

Nor is it necessary to impose conditions on f to guarantee unique existence of a solution of (1.6) or (1.6a) globally in y, whatever be $h > 0$. Even the backward Cauchy-Euler method can have multiple solutions for problems that are not pathological at all—see §5 for a simple example. And step sizes used in practice are always constrained by the user's error tolerance [42]. Therefore, it suffices that numerical schemes be well defined locally, for a step size with respect to which the solution is smooth. We amplify this point by an example presently, and by another in §5.

Recall the care with which Dahlquist introduces the monotonicity condition [10]. Acknowledging that it is rarely valid in the whole space, he assumes only that it holds in a tube about a solution, of radius comparable to the maximum error tolerable in a numerical approximate solution. Global monotonicity of f is then imposed for mathematical convenience by altering f in the (uninteresting!) complement of the tube.

Subsequent authors have not always been so explicit about the realism of their assumptions. But still problems remain. When existence and uniqueness of solutions are established by appeal to topological principles, such as uniform monotonicity [10, 14, Chapter 5; 44, 27] or topological degree [17], it must still be shown that the solution lies in the "interesting" region and is not a mathematical phantom introduced by modifying f outside the tube. In [10] the results are justified a posteriori by error estimates. Later authors seldom address this difficulty at all. Finally, of course, these topological arguments say nothing about the convergence of practical algorithms such as (1.9), (1.9a).

A result somewhat related to ours is [23, Lemma 5], where the Kantorovich Theorem is invoked to obtain solutions of (1.6a) via convergence of the *unmodified* Newton method. The differential equation is taken to be in the singular perturbation form

$$dy/dt = f(y, z), \qquad \varepsilon dz/dt = g(y, z), \tag{1.19}$$

and it is assumed that

$$\mu(\partial g/\partial z) \le -1 \quad \text{in a neighborhood of the solution;} \tag{1.20}$$

μ denotes the logarithmic matrix norm with respect to some inner product [14, §1.5]. Assumption 1.1 for the problem (1.19) requires

$$[\partial g/\partial y, \ \partial g/\partial z] \quad \text{is negative dominant in } U. \tag{1.20a}$$

For comparison with (1.20), (1.20a) implies

$$\mu_\infty(\partial g/\partial z) \le -\delta\sigma \quad \text{in } U, \tag{1.20b}$$

μ_∞ being the logarithmic matrix norm subordinate to the norm $|\cdot|$ employed here [14, (1.5.9)]. If (1.20) is replaced by (1.20a), then [23, Lemma 5] is a special case of Theorem 2 here, and [23, Lemma 6] follows from Corollaries 4.2, 4.3 and the remarks following them.

We conclude this section with an example showing that solutions of dissipative differential equations need not be smooth, cf. [29]. Our example is inspired by the fact that examples showing nonexistence of a solution in (1.6a) all have J varying strongly over the length of a step [14, Examples 5.2.11, 5.3.1, 5.8.1, and Theorems 5.8.3, 5.8.4]. But one *should not attempt* to traverse such an interval in a single step, for the solution of the differential equation need not be smooth on such a scale [29].

Consider the differential equation

$$\begin{aligned} &\varepsilon\, dy/dt = -a(t)y + \varepsilon^\nu, \qquad && -1 \le t, \quad 0 \le \varepsilon \ll 1, \\ &a(t) = \varepsilon^\mu + \sin^2 \omega t, && 0 < \nu < \mu < 1. \end{aligned} \tag{1.21}$$

This differential equation is monotone. Consider first $\omega = \pi/2$. Then after a possible initial layer at $t = -1$ the solution is approximately given by

$$y(t) \sim \varepsilon^\nu / a(t),$$

exhibiting "spikes" of height $O(\varepsilon^{\nu-\mu}) \gg 1$ and duration $O(\varepsilon^{\nu/2})$ around $t = 0, 2, \ldots$. An algebraically stable and irreducible Runge-Kutta formula can be used with arbitrary $h > 0$ to compute a "solution" to this problem [14, Corollary 5.3.13], but the spikes will not be resolved unless h is small enough in their vicinity, so this fact confers no benefit. It could even be positively harmful if the error-estimating formula were likewise "stable". It has been suggested that the Gauss formulae be regarded as defective for just this reason [40].

Finally, if we take $\omega \gg 1$ in (1.21) we move the spikes close together and the solution is never smooth. This shows that a monotone differential equation need have no smooth solution.

2. The Jacobian matrix

In this section we determine bounds for matrix inverses and Lipschitz constants for matrix functions occurring in the analysis of the modified Newton method.

Lemma 2.1. *Suppose* Λ *is negative dominant with parameters* ρ, σ, δ. *Then* Λ *is invertible and* $|\Lambda^{-1}| < (\delta\sigma)^{-1}$.

Proof. Let $0 \neq x = (x_1, \dots, x_N)^T \in \mathbb{C}^N$ and let i be an index for which $|x_i| = |x|$. Then

$$\begin{aligned} |\Lambda x| &\geq \left| \Lambda_{ii}x_i + \sum_{j\neq i} \Lambda_{ij}x_j \right| \geq |\Lambda_{ii}x_i| - \sum_{j \neq i} |\Lambda_{ij}||x| \\ &\geq \left(|\operatorname{Re}\Lambda_{ii}| + (1-\delta)\operatorname{Re}\Lambda_{ii}\right)|x| = -\delta \operatorname{Re}\Lambda_{ii}|x|, \end{aligned} \tag{2.1}$$

by (1.17). By the first part of (1.17), $-\delta \operatorname{Re}\Lambda_{ii} > \delta\sigma$, and this completes the proof. □

We shall find (2.1) to be useful in the sequel.

Lemma 2.2. *Assume* J *is essentially negative dominant with parameters* ρ, δ, c, *and let* $\alpha = \beta + \gamma i$ *be a complex number with* $\beta > 0$. *Let*

$$\sigma := \tfrac{1}{4}\beta/|\alpha|^2; \tag{2.2}$$

then for any $h > 0$ *such that* $h < H_0$,

$$H_0 = H_0(\delta, c, \alpha) := \frac{\beta}{c|\alpha|^2}\min\{\tfrac{1}{4},\ 1-\delta\}, \tag{2.3}$$

the matrix $hJ - \alpha^{-1}I$ *is negative dominant with parameters* δ, σ *and*

$$\rho' = \rho'(\rho, \alpha) := \rho + 1 + 4|\gamma|/\beta. \tag{2.4}$$

Proof. To show the first part of (1.17) for $hJ - \alpha^{-1}I$ we have

$$\operatorname{Re}(hJ_{ii} - \alpha^{-1}) = h \operatorname{Re} J_{ii} - \beta/|\alpha|^2 < -\sigma,$$

by the definition of σ if $\operatorname{Re} J_{ii} \leq 0$, and by the condition on h if $\operatorname{Re} J_{ii} > 0$. For (1.16): if $\operatorname{Re} J_{ii} > 0$,

$$\frac{|\operatorname{Im}(hJ_{ii} - \alpha^{-1})|}{|\operatorname{Re}(hJ_{ii} - \alpha^{-1})|} = \frac{\left|h \operatorname{Im} J_{ii} + \frac{\gamma}{|\alpha|^2}\right|}{\left|h \operatorname{Re} J_{ii} - \frac{\beta}{|\alpha|^2}\right|} \leq \frac{hc(\rho+1) + \frac{|\gamma|}{|\alpha|^2}}{\sigma} < \rho',$$

by (1.16a), (1.17a), the assumptions on σ and h, and (2.2); if $\operatorname{Re} J_{ii} \leq 0$,

$$\frac{\left|h \ \operatorname{Im} J_{ii} + \frac{\gamma}{|\alpha|^2}\right|}{\left|h \ \operatorname{Re} J_{ii} - \frac{\beta}{|\alpha|^2}\right|} \leq \frac{\rho |\operatorname{Re} hJ_{ii}| + hc + \frac{|\gamma|}{|\alpha|^2}}{|\operatorname{Re} hJ_{ii}| + \frac{\beta}{|\alpha|^2}} \leq \max_{\lambda \geq 0} \frac{\rho\lambda + \sigma + \frac{|\gamma|}{|\alpha|^2}}{\lambda + \frac{\beta}{|\alpha|^2}} \leq \rho'$$

by (1.16a), the assumption on h, and (2.2). Finally, we verify the row-sum conditions in (1.17). If $\operatorname{Re} J_{ii} > 0$,

$$\begin{aligned}\sum_{j \neq i} |hJ_{ij}| &\leq hc - h \ \operatorname{Re} J_{ii} \\ &= -(1-\delta)\operatorname{Re}(hJ_{ii} - \alpha^{-1}) + (1-\delta)\operatorname{Re} hJ_{ii} \\ &\quad - (1-\delta)\frac{\beta}{|\alpha|^2} + hc - \operatorname{Re} hJ_{ii} \\ &< -(1-\delta)\operatorname{Re}(hJ_{ii} - \alpha^{-1})\end{aligned}$$

by (1.17b) and the assumption on h; and if $\operatorname{Re} J_{ii} < 0$,

$$\begin{aligned}\sum_{j \neq i} |hJ_{ij}| &\leq -(1-\delta)\operatorname{Re} hJ_{ii} + hc \\ &< -(1-\delta)\operatorname{Re}(hJ_{ii} - \alpha^{-1}) + hc - (1-\delta)\frac{\beta}{|\alpha|^2} \\ &< -(1-\delta)\operatorname{Re}(hJ_{ii} - \alpha^{-1})\end{aligned}$$

as before. □

Lemma 2.3. *Under the assumption of Lemma* 2.2 *there are constants* k_3', k_3 *depending only on* δ, α *such that*

$$|(I - h\alpha J)^{-1}| \leq k_3', \qquad |(I - h\alpha J)^{-1} h\alpha J| \leq k_3.$$

Proof. We have $I - h\alpha J = -\alpha(hJ - \alpha^{-1} I)$. The last factor is negative dominant, so by Lemma 2.1 $|(I - h\alpha J)^{-1}| < (|\alpha|\sigma\delta)^{-1}$. The second bound follows from $(I - h\alpha J)^{-1} h\alpha J = -I + (I - h\alpha J)^{-1}$. □

Lemma 2.4. *Let* U *be a tube in* $\mathbb{R} \times \mathbb{R}^N$ *surrounding a smooth curve, and let* $J \colon U \to \mathbb{R}^{N \times N}$ *be a matrix function that is essentially negative dominant with parameters* ρ, δ, c; *and slowly varying to order* 1 *with constants* $K_{1,\nu} \leq K_1$ *for all multi-indices* $|\nu| \leq 1$. *Let* α *be a complex number with positive real part. Let* $(t_0, y_0) \in U$ *and write* J_0 *for* $J(t_0, y_0)$. *Let* $0 < h < H_0$, *with* H_0 *given by* (2.3). *There is a constant* k_4 *depending only on* α, c, δ, ρ, *and* K_1 *such that for* $(t, y) \in U$,

$$|I - (I - \alpha h J_0)^{-1}(I - \alpha h J(t, y))| < k_4 \max\{|t - t_0|, \ |y - y_0|\}. \tag{2.5}$$

Proof. It suffices to prove the result when the line segment from (t_0, y_0) to (t, y) lies entirely in U. Let

$$\begin{aligned}M(\theta) &:= I - \alpha h J\left(t_0 + \theta(t - t_0), \ y_0 + \theta(y - y_0)\right), \qquad 0 \leq \theta \leq 1, \\ C(\theta) &:= M^{-1}(0) M(\theta);\end{aligned}$$

then

$$C(0) = I, \qquad C^H(\theta) = M^H(\theta)(M^H)^{-1}(0). \tag{2.6}$$

The Hermitian conjugate M^H satisfies the differential equation

$$\frac{dM^H}{d\theta} = \frac{dM^H}{d\theta}(M^H)^{-1}M^H = \left(M^{-1}\frac{dM}{d\theta}\right)^H M^H. \tag{2.7}$$

Since J is slowly varying, we have by (1.18) and the proof of Lemma 2.1,

$$\left|M^{-1}(\theta)\frac{dM}{d\theta}\right| \le \text{const} \cdot \max\{|t - t_0|,\ |y - y_0|\}, \tag{2.8}$$

with a constant depending only on α, c, δ, ρ, and K_1. Now according to (2.6), $C^H(\theta)$ is the solution operator for the differential equation (2.7). Applying the elementary estimate for this operator that follows from (2.8) yields

$$\begin{aligned}|C(0) - C(1)| &= |I - (I - \alpha hJ_0)^{-1}(I - \alpha hJ(t, y))| \\ &\le k_4 \max\{|t - t_0|,\ |y - y_0|\},\end{aligned}$$

with a constant $k_4 = k_4(\alpha, c, \delta, \rho, K_1)$, and the proof is complete. □

Note the identity

$$I - (I - \alpha hJ_0)^{-1}(I - \alpha hJ(t, y)) = (I - \alpha hJ_0)^{-1}\left(\alpha hJ(t, y) - \alpha hJ_0\right). \tag{2.9}$$

The estimate of Lemma 2.4, then, applies also to the matrix on the right side of (2.9), and exhibits k_4 as a relative Lipschitz constant. The author of [36] assumes a relative Lipschitz condition equivalent to Lemma 2.4 but does not exploit it as we do here. Deuflhard and Heindl [16] showed that the relative Lipschitz condition occurs naturally in affine-invariant convergence theorems for Newton-like methods. Their approach could be used to give an alternative proof of Theorems 1 and 2.

When an iteration matrix is retained through a change of step size or formula, we must account for the presence of distinct αh in the "numerator" and "denominator" matrices. In the notation of Lemma 2.4, with also $0 < h_0 < H_0(\alpha_0)$, we compute

$$\begin{aligned}I &- (I - \alpha_0 h_0 J_0)^{-1}(I - \alpha hJ(t, y)) \\ &= (I - \alpha_0 h_0 J_0)^{-1}\left(\alpha hJ(t, y) - \alpha_0 h_0 J_0\right) \\ &= (I - \alpha_0 h_0 J_0)^{-1}\left[\left(\alpha hJ(t, y) - \alpha hJ_0\right) + (\alpha h - \alpha_0 h_0)J_0\right].\end{aligned}$$

An application of Lemmas 2.4 and 2.1 now establishes the following estimate.

Lemma 2.5. *Under the conditions of Lemma* 2.4 *let also* $0 < h_0 < H_0(\alpha_0)$. *Then, with the constants of Lemmas* 2.1 *and* 2.2,

$$\begin{aligned}&|I - (I - \alpha_0 h_0 J_0)^{-1}(I - \alpha hJ(t, y))| \\ &\qquad \le |\alpha h/\alpha_0 h_0| k_4 \max\{|t - t_0|,\ |y - y_0|\} + |\alpha h/\alpha_0 h_0 - 1| k_3.\end{aligned} \tag{2.10}$$

The last compromise we need is the possible replacement of J_0 by a divided difference approximation. Of the many possibilities [15, §4.2; 47], we discuss only first-order differences; the results are readily extended to other schemes. Let $\xi = (\xi_1, \dots, \xi_N)^T \in \mathbb{R}^N$ be a vector of increments such that $(t_0, y_0 + \xi_j e_j) \in U$ for each j, and let the jth column of the Jacobian be approximated by

$$J_{0,j} := \xi_j^{-1}(f(t_0, y_0 + \xi_j e_j) - f(t_0, y_0)), \qquad j = 1, \dots, N. \tag{2.11}$$

We get the following replacement for (2.10).

Lemma 2.6. *Under the conditions of Lemma* 2.4, *with* J_0 *given by* (2.11), *we have that if* $|\xi|$ *is sufficiently small, then* $I - \alpha_0 h_0 J_0$ *is invertible, and*

$$\begin{aligned}
&|I - (I - \alpha_0 h_0 J_0)^{-1}(I - \alpha h J(t, y))| \\
&\quad = |(I - \alpha_0 h_0 J_0)^{-1}(\alpha h J(t, y) - \alpha_0 h_0 J_0)| \\
&\quad \le \left(1 - \frac{1}{2} N k_4 |\xi|\right)^{-1} \left(\left|\frac{\alpha h}{\alpha_0 h_0}\right| k_4 \max\{|t - t_0|,\ |y - y_0|\}\right. \\
&\qquad\qquad \left. + \left|\frac{\alpha h}{\alpha_0 h_0} - 1\right| k_3 + \frac{1}{2} N k_4 |\xi|\right).
\end{aligned} \tag{2.12}$$

Proof. By Lemma 2.4, the function $(I - \alpha_0 h_0 J(t_0, y_0))^{-1} J(t_0, y)$ satisfies a Lipschitz condition at y_0 with respect to y, with Lipschitz constant k_4. Therefore [15, Lemma 4.2.1],

$$\begin{aligned}
&|(I - \alpha_0 h_0 J(t_0, y_0))^{-1}[\alpha_0 h_0 f(t_0, y_0 + \xi_j e_j) - \alpha_0 h_0 f(t_0, y_0) \\
&\qquad\qquad - \alpha_0 h_0 J(t_0, y_0) \xi_j e_j]| \\
&\quad \le \tfrac{1}{2} k_4 \xi_j^2, \qquad j = 1, 2, \dots N.
\end{aligned} \tag{2.13}$$

Substituting (2.11) into (2.13) and dividing through by ξ_j gives

$$|(I - \alpha_0 h_0 J(t_0, y_0))^{-1} \alpha_0 h_0 (J_0 - J(t_0, y_0)) e_j| \le \tfrac{1}{2} k_4 |\xi_j|, \qquad j = 1, 2, \dots, N.$$

Hence, for the matrix norm we have

$$|I - (I - \alpha_0 h_0 J(t_0, y_0))^{-1}(I - \alpha_0 h_0 J_0)| \le \tfrac{N}{2} k_4 |\xi|. \tag{2.14}$$

It follows from the Neumann Lemma [15, Theorem 3.1.4; 34, 2.3.1] that $(I - \alpha_0 h_0 J(t_0, y_0))^{-1}(I - \alpha_0 h_0 J_0)$ is invertible (and thus so is $(I - \alpha_0 h_0 J_0)$) if $\frac{1}{2} N k_4 |\xi| < 1$, and then

$$|(I - \alpha_0 h_0 J_0)^{-1}(I - \alpha_0 h_0 J(t_0, y_0))| \le (1 - \tfrac{1}{2} N k_4 |\xi|)^{-1}. \tag{2.15}$$

The estimate (2.12) now follows from a computation:

$$\begin{aligned}
&I - (I - \alpha_0 h_0 J_0)^{-1}(I - \alpha h J(t, y)) \\
&\quad = (I - \alpha_0 h_0 J_0)^{-1}(\alpha h J(t, y) - \alpha_0 h_0 J_0) \\
&\quad = (I - \alpha_0 h_0 J_0)^{-1}(I - \alpha_0 h_0 J(t_0, y_0)) \\
&\qquad \cdot (I - \alpha_0 h_0 J(t_0, y_0))^{-1}[\alpha h (J(t, y) - J(t_0, y_0)) + (\alpha h - \alpha_0 h_0) J(t_0, y_0) \\
&\qquad\qquad + \alpha_0 h_0 (J(t_0, y_0) - J_0)]
\end{aligned}$$

Use Lemma 2.3 together with (2.14) and (2.15) here to complete the proof. □

For the remainder of the paper, J_0 denotes a matrix which can be either $J(t_0, y_0)$ or its divided difference approximation (2.11), as in the statements of Theorems 1 and 2. The estimate (2.12) always holds, for if J_0 is the analytical Jacobian, we take $\xi = 0$ in (2.12), reducing it to (2.10).

The last condition we need for the proof of Theorems 1 and 2 is Lipschitz continuity for $(t, y) \in U$ of the matrix function $(I - \alpha_0 h_0 J_0)^{-1} \alpha h J(t, y)$. It follows from Lemma 2.4 that this function satisfies a Lipschitz condition at (t_0, y_0). Here is the more general estimate that we require.

Lemma 2.7. *Under the conditions of Lemma 2.4 the matrix function*

$$(I - \alpha_0 h_0 J_0)^{-1} \alpha h J(t, y)$$

is Lipschitz continuous in $(t, y) \in U$ with a Lipschitz constant k_4' depending only on α_0, c, δ, ρ, K_1, $|\xi|$, diam U, *and* $|\alpha h/\alpha_0 h_0|$.

Proof. Let $(t, y), (u, z) \in U$. Then

$$\begin{aligned}(I - \alpha_0 h_0 J_0)^{-1} &(\alpha h J(t, y) - \alpha h J(u, z)) \\ &= (I - \alpha_0 h_0 J_0)^{-1} (I - \alpha_0 h_0 J(t_0, y_0)) \\ &\quad \cdot (I - \alpha_0 h_0 J(t_0, y_0))^{-1} (I - \alpha_0 h_0 J(t, y)) \\ &\quad \cdot (I - \alpha_0 h_0 J(t, y))^{-1} (\alpha h J(t, y) - \alpha h J(u, z)).\end{aligned}$$

Apply Lemmas 2.4 and 2.2 to estimate the matrix norm of the product by

$$\begin{aligned}(1 - \tfrac{1}{2} N k_4 |\xi|)^{-1} &(1 + k_4 \text{ diam } U) \cdot |\alpha h/\alpha_0 h_0| k_4 \max\{|t - u|, |y - z|\} \\ &\equiv k_4' (\alpha_0, c, \delta, \rho, K_1, |\xi|, \text{diam } U, |\alpha h/\alpha_0 h_0|) \max\{|t - u|, |y - z|\}. \quad \square\end{aligned}$$

For the proof of Theorem 2 we need estimates analogous to those of Lemmas 2.3–2.7 for the matrix (1.8a) and the Jacobian of (1.7a). The proofs are analogous, too, once the basic idea is understood.

Lemma 2.8. *Let J be an essentially negative dominant matrix, and A a Runge-Kutta matrix all of whose eigenvalues have positive real part. Denote these eigenvalues by $\{\alpha_j = \beta_j + i\gamma_j : j = 1, \dots, q\}$; let*

$$\sigma := \tfrac{1}{4} \min_j \beta_j / |\alpha_j|^2, \qquad H_0 := (\sigma/c) \min\{1, 4(1 - \delta)\}. \tag{2.16}$$

If $0 < h < H_0$, then there is a constant K_3 depending only on c, δ, and A such that $|(I_{qN} - A \otimes hJ)^{-1}| \le K_3$.

Proof. Let Q be a unitary matrix such that $Q^H A Q = R$ is upper triangular. Then

$$I_q \otimes I_N - A \otimes hJ = (Q \otimes I_N)(I_q \otimes I_N - R \otimes hJ)(Q^H \otimes I_N),$$

and the center factor on the right side is block upper triangular. The diagonal blocks are $I_N - \alpha_j hJ$, $j = 1, \dots, q$, and are invertible by Lemma 2.2 and (2.16). The result follows by back substitution and Lemma 2.3. □

Lemma 2.9. *Under the assumptions of Lemmas* 2.6 *and* 2.8, *let* $Y = (y_1^T, \dots, y_q^T)^T$ *with* $|y_j - y(t_n + hc_j)| < \tau$, $j = 1, 2, \dots, q$. *Recall the definition* (1.12) *of* $F(Y)$, *and write* $D_Y F$ *for the Jacobian of* F *with respect to* Y. *If* $0 < h < H_0$, H_0 *given by* (2.16), *then there is a constant* K_4 *depending only on* c, δ, A *and* K_1 *such that*

$$\begin{aligned}
&|I_{qN} - (I_{qN} - A \otimes h_0 J_0)^{-1}(I_{qN} - (A \otimes I)hD_Y F(Y))| \\
(2.17)\quad &\le (1 - K_4|\xi|)^{-1}\left\{(h/h_0)\left[K_4 \max_j\{\max|t_n + hc_j - t_0|, |y_j - y_0|\} + |\xi|\right]\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. + |h/h_0 - 1|(K_3 + 1)\right\}.
\end{aligned}$$

Proof. A brief computation gives

$$\begin{aligned}
I_{qN} &- (I_{qN} - A \otimes h_0 J_0)^{-1}(I_{qN} - (A \otimes I)hD_Y F(Y)) \\
&= (I_{qN} - A \otimes h_0 J_0)^{-1}(A \otimes I)(hD_Y F(Y) - I_q \otimes h_0 J_0) \\
&= (I_{qN} - A \otimes h_0 J_0)^{-1}(A \otimes I)\left[h(D_Y F(Y) - I \otimes J(t_0, y_0))\right. \\
&\qquad\qquad \left. + I \otimes h(J(t_0, y_0) - J_0) + (h - h_0)I \otimes J_0\right],
\end{aligned}$$

and the proof follows by triangularization of A and application of Lemmas 2.4–2.6 to estimate the terms in the square bracket. □

The same technique proves the following analogue of Lemma 2.7.

Lemma 2.10. *Under the conditions of Lemmas* 2.7, 2.8 *and* 2.9, *the matrix function* $(I_{qN} - A \otimes h_0 J_0)^{-1} hD_Y F(Y)$ *is Lipschitz continuous in* Y *with a Lipschitz constant* $K_4' = K_4'(c, \delta, \rho, K_1, A, |\xi|, \operatorname{diam} U, h/h_0)$.

3. Convergence of the iteration

We use the contraction mapping theorem in the following form.

Theorem 3.1 [18, 10.1.1]. *Let* $B = \{y : \|y - x\| < r\}$ *be an open ball in a Banach space* X. *Let* V *be an open set in* X *and let* $\varphi : V \times B \to X$ *be a continuous function satisfying*

$$(3.1)\qquad \|\varphi(v, y) - \varphi(v, z)\| \le k\|y - z\|$$

for all $v \in V$, $y, z \in B$, *with a constant* k, $0 \le k < 1$, *and*

$$(3.2)\qquad \|\varphi(v, x) - x\| \le (1 - k)r \quad \text{for any} \quad v \in V.$$

Then there exists a unique mapping g *of* V *into* B *such that*

$$g(v) = \varphi(v, g(v))$$

for any $v \in V$, and g is continuous in V. Indeed, for any $v \in V$ the prescription $y^0 \in B$ arbitrary,

$$y^{m+1} = \varphi(v, y^m), \qquad m = 0, 1, 2, \dots,$$

yields a sequence satisfying $y^m \in B$ for every m; $\lim_{m\to\infty} y^m = y^*$ exists and does not depend on y^0; g is defined by $g(v) = y^*$.

Theorem 1 is actually the special case $q = 1$, $e \otimes y_n = \psi$, $A = \alpha$, $c = \alpha$ of Theorem 2, so we do not present a separate proof.

Proof of Theorem 2. We show that any desired q-linear convergence rate $0 < k < 1$ can be achieved. We recall constants H_0, K_3 of Lemma 2.8, K_4 of Lemma 2.9, K_4' of Lemma 2.10.

Let $0 < \eta < 1$, $0 < C_1 < \tau$, $0 < \xi_0 < K_4$, $C_2 > 0$ be any constants satisfying

$$K_4' C_1 + (1 - K_4\xi_0)^{-1}\{(1+\eta)K_4(C_2 + \xi_0) + \eta(K_3 + 1)\} \le k. \tag{3.3}$$

Define the "internal order" $p_0 = \max\{p: Ac^{j-1} = c^j/j, \quad 1 \le j \le p\}$ if this set is nonempty, $p_0 = 0$ otherwise. Ordinarily, $p_0 \ge 1$, but this is not necessary for convergence [7]; [48] gives examples of formulae with $p_0 = 0$. Then choose $0 < h < H_0$, $0 < C_3 < \tau$ small enough that

$$K_3(C_3 + K_{2,p_0+1}h^{p_0+1}) < (1-k)C_1. \tag{3.4}$$

Assume that J_0 in (1.8a) is $J(t_0, y_0)$, with $(t_0, y_0) \in U$, or a divided difference approximation (2.11) to it with step lengths less than $|\xi|$. Let $z(t)$ be the local solution of (1.5) satisfying $z(t_n) = y_n$, and assume that

$$\max_{1\le j\le q} \max\{|t_n + hc_j - t_0|, |z(t_n + hc_j) - y_0|\} < C_2, \tag{3.5}$$

$$|\xi| < \xi_0, \tag{3.6}$$

$$\left|\frac{h}{h_0} - 1\right| < \eta. \tag{3.7}$$

Let

$$V := \{\Psi \in \mathbb{R}^{qN}: |\Psi - e \otimes y_n| < C_3\} \tag{3.8a}$$

if A is invertible,

$$V := \{\Psi \in \mathbb{R}^{qN}: |\Psi - (e \otimes y_n + a_0 \otimes hf(t, y_n))| < C_3\} \tag{3.8b}$$

if the Runge-Kutta formula is of special type. Using the local solution $z(t)$, define the vector $Z(t) = (z(t + hc_1)^T, \dots, z(t + hc_q)^T)^T$ and let

$$B := \{Y \in \mathbb{R}^{qN}: |Y - Z(t)| < C_1\}. \tag{3.9}$$

We show that the prescription (1.9a) defines a function $\varphi: V \times B \to \mathbb{R}^{qN}$ by

$$\varphi(\Psi, Y) = Y + M^{-1}G(\Psi, Y), \tag{3.10}$$

in which (1.7a) defines G, and that φ satisfies the conditions (3.1) and (3.2) of Theorem 3.1.

Now (3.10) is well defined, for Lemma 2.8 and the choice of h show that M is invertible. Moreover, if $\Psi = (\psi_1^T, \dots \psi_q^T)^T \in V$, then $(t, \psi_i) \in U$ (resp. $(t+ha_{i0}, \psi_i) \in U$) by the choice of C_3 and h; and if $Y = (y_1^T, \dots, y_q^T)^T \in B$, then $(t + hc_i, y_i) \in U$ by the choice of C_1. Thus the estimates of Lemmas 2.9–2.10 apply.

Let $Y, Z \in B$. Then

$$\begin{aligned} &\varphi(\Psi, Y) - \varphi(\Psi, Z) \\ &\quad = M^{-1}G(\Psi, Y) - M^{-1}G(\Psi, Z) - M^{-1}G_Y(\Psi, Z(t))(Y - Z) \\ &\qquad + (I + M^{-1}G_Y(\Psi, Z(t)))(Y - Z). \end{aligned} \tag{3.11}$$

By Lemma 2.10, $M^{-1}G(\Psi, Y)$ has a Lipschitz continuous partial derivative with respect to Y; therefore, by a mean value theorem [15, Lemma 4.1.15],

$$\begin{aligned} &|M^{-1}G(\Psi, Y) - M^{-1}G(\Psi, Z) - M^{-1}G_Y(\Psi, Z(t))(Y - Z)| \\ &\quad \le \tfrac{1}{2}K_4'(|Y - Z(t)| + |Z - Z(t)|)|Y - Z| \le K_4'C_1|Y - Z|. \end{aligned} \tag{3.12}$$

By Lemma 2.9,

$$\begin{aligned} &|I + M^{-1}G_Y(\Psi, Z(t))| \\ &\quad = |I_{qN} - (I_{qN} - A \otimes h_0 J_0)^{-1}(I_{qN} - (A \otimes I_N)hF_Y(Z(t)))| \\ &\quad \le (1 - K_4\xi_0)^{-1}\{(1 + \eta)K_4(C_2 + \xi_0) + \eta(1 + K_3)\} \end{aligned} \tag{3.13}$$

from (2.17), using (3.5), (3.6), (3.7). Inserting (3.12) and (3.13) into (3.11) and using (3.3) proves (3.1). To prove (3.2), consider for invertible A

$$\begin{aligned} |\varphi(\Psi, Z(t)) - Z(t)| &= |M^{-1}G(\Psi, Z(t))| \\ &\le |M^{-1}||\Psi - e \otimes y_n + e \otimes y_n + (A \otimes I)hF(Z(t)) - Z(t)| \\ &\le K_3\left(C_3 + \max_{1\le i\le q}\left|z(t) + h\sum_{j=1}^{q} a_{ij}z'(t + hc_j) - z(t + hc_i)\right|\right) \\ &\le K_3(C_3 + K_{2,p_0+1}h^{p_0+1}) < (1 - k)C_1 \end{aligned}$$

by (3.4), (3.8a), Lemma 2.8, and Assumption 1.2 for the local solution $z(t)$. The treatment of a formula of special type is similar, proceeding from (3.8b) instead of (3.8a). This shows (3.2) and completes the proof of Theorem 2. □

There is no difficulty in extending the theorem to cover general linear methods. One simply proves the analogues of Lemmas 2.8–2.10. No new ideas are required to establish (3.1) and (3.2).

4. Algorithmic consequences

Theorems 1 and 2 provide the theoretical support for the practice in codes for stiff problems of retaining the iteration matrix for many steps and even through changes of step size and formula. Our analysis shows that the achievement of a desired contraction rate k depends, according to (3.3), upon four quantities subject to control by the code: C_1, the accuracy of the predicted solution; C_2, the age of the Jacobian; η, the relative deviation of the current step size from that incorporated in the iteration matrix; and ξ_0, the size of increments if the Jacobian has been approximated by divided differences. For the iteration to be contracting to a *solution* we require the invariance of a certain ball, and this is achievable, according to (3.4), if the step size h is small enough, since for solving the original equation (1.6) or (1.6a) we may take $C_3 = 0$. The possibility of taking $C_3 > 0$ arises in error analysis, which we consider presently.

It is the task of the error-control mechanism in the code to propose a step size that is "on scale" for the local solution. This means that when the order of accuracy of the formula is p, the quantity $K_{2,p+1}h^{p+1}$ is roughly comparable to the user's error tolerance. Consequently, it is reasonable to expect (3.4) to hold with the step size being attempted. This means that algorithmic choices in the conduct of the iteration are made solely for the purpose of achieving the desired contraction rate. We should also mention that various relaxation [2] and acceleration [4] schemes have been tried; see the survey [46]. We do not attempt to analyze such schemes here.

Of the four quantities C_1, C_2, η, ξ_0 affecting the contraction rate, we need not consider ξ_0 and C_1 here. The optimal choice of ξ depends on f and properties of floating-point arithmetic independent of the application to solving differential equations [9; 15, §4.2]. C_1, the accuracy of the predicted solution, is also not a matter of concern for reasons already mentioned: if the local solution is smooth and the step size is on scale, then information from the solution history enables the code to make an accurate prediction.

We can get a rough idea of the convergence rate's sensitivity to the quantities C_2 and η by inspecting the coefficients multiplying them in (3.3). The coefficient of C_2 is a (relative) Lipschitz constant for J. The presumption that this constant is small is expressed by statements in the literature that the Jacobian is expected to be "nearly constant" in the neighborhood of a smooth solution [39, 42]; in the smooth part of the Van der Pol (1.4) limit cycle the Lipschitz constant is of order 1, however. Remember that C_2 itself, the "age" of the Jacobian, is ordinarily a not excessively large multiple of h. The coefficient of η, on the other hand, surely exceeds 1; this implies that η must be taken small relative to the desired convergence rate.

There are three remedies available in case of unsatisfactorily slow convergence. In order of increasing severity these are: having saved J_0, compute a new decomposition with the current h (making $\eta = 0$); form a new Jacobian at, say, (t_n, y_n), and compute its decomposition (this makes $\eta = 0$ and $C_2 = h \max_i |c_i|$); or, finally, abandon the step and retry with a new smaller h. We note that when Enright's device of reducing $(h\alpha)^{-1}I - J$ to Hessenberg form by a similarity transformation is in use, then $\eta = 0$ always, and the choice is between the latter alternatives [19, 40]. This remaining choice of remedies has been studied by Shampine. In [37] he says, "We suggest that a new J be formed at every convergence failure with J out of date. There is no reason to think the step size unsatisfactory, so we suggest trying it again." Two years later, discussing a somewhat different setup, he reaches a contrary conclusion [41]: "We argue that in conjunction with other aspects of our algorithm, reduction of the step size is always the appropriate response to failure of convergence." We shall see that the present analysis lends support to the second decision in some circumstances, but not always.

Let us consider first the situation when the rate of contraction is very slow, or the iteration is even diverging. Should we attempt to restore convergence by updating J, or do we infer that some fundamental assumption—smoothness of the solution, accuracy of the formula, slow variation of J, appropriateness of h—is invalid, so that a reduction in h is called for? Let us suppose for a moment that all the basic assumptions remain valid. On the last preceding step, (3.3) was satisfied. The only difference when we undertake the current step is that C_2 changes by $O(h)$—the Jacobian is one step older. It is not plausible that this single small change would turn a rapidly converging iteration into a slowly converging one; hence one or more of the basic assumptions is failing, and the appropriate decision *is* to reduce the step size.

The matter is less clear if the rate of convergence is only marginally unsatisfactory. We are not impelled, then, to give up any of the basic assumptions, for it is likely that the aging of J_0 is responsible for the deterioration of convergence, and that an updated J will make it rapid again. Shampine insists on reducing h here too: "There is no justification ... for going to the expense of forming a new Jacobian when a relatively small change of stepsize will suffice." The case is weaker here, however: in the situation we are describing, if we do not update J this time, we could be confronted with the same problem on the next step. This means that the decision comes down to proceeding with a smaller than optimal step size to avoid the expense of forming a new J, and this is a code- and perhaps even problem-dependent balance.

We turn now to error estimates. This has been studied heretofore using the theory of BSI stability [14, §5.4-5.10; 12, 13, 21, 26]. In our framework weaker hypotheses yield stronger results by elementary means. The conclusion of (3.2) is that there is a unique mapping $g: V \to B$ such that

$$G(\Psi, g(\Psi)) = 0$$

for all $\Psi \in V$, and that g is continuous. More is true. Writing Y^* for $g(\Psi)$, we have

$$D_Y G(\Psi, Y^*) = I_{qN} - (A \otimes I) h D_Y F(Y^*).$$

Under the hypotheses of Theorem 2 this matrix is nonsingular, so the implicit function theorem [18, 10.2.2] applies.

Corollary 4.1. *Under the conditions of Theorem* 2 *the mapping* $\Psi \to Y^* = g(\Psi)$ *is continuously differentiable, and*

$$Dg(\Psi) = [I_{qN} - (A \otimes I)hD_Y F(g(\Psi))]^{-1}.$$

Consider now an approximate solution $Y^m \in B$ of (1.6a) with residual r^m. Then by (1.7a),

$$r^m = \Psi + (A \otimes I)hF(Y^m) - Y^m$$

implies

$$G(\Psi - r^m, Y^m) = 0.$$

If $|r^m| < C_3$ in Theorem 2, then $Y^m = g(\Psi - r^m)$, and the Mean Value Theorem of differential calculus [18, 8.5.4], together with Lemma 2.8, yield an error bound.

Corollary 4.2. *If* $G(\Psi - r^m, Y^m) = 0$ *with* $Y^m \in B$, $|r^m| < C_3$, *then*

$$|Y^m - Y^*| < K_3 |r^m|.$$

K_3 is a constant of moderate size independent of stiffness. Shampine [38] obtains a similar bound, assuming that J is similar to a diagonal matrix, and that the matrix effecting the similarity is well conditioned. Curtis [8] reports, however, that an eigenbasis can be quite skew. Corollary 4.2 shows that a small residual *always* guarantees a small error; this is the theoretical support for the policy advocated in [38], that in the iterations (1.9), (1.9a) the smallness of the residual should be the stopping criterion. We can say more. By Lemma 2.10, the derivative Dg is Lipschitz continuous. By another mean value theorem [18, 8.6.2; 15, Lemma 4.1.12], the error is well approximated by the linearization.

Corollary 4.3. *There is a constant of moderate size such that*

$$\begin{aligned}&|g(\Psi - r^m) - g(\Psi) - Dg(\Psi)(-r^m)|\\ &\quad = |Y^m - Y^* - [I_{qN} - (A \otimes I)hD_Y F(Y^*)]^{-1} r^m| < \text{const}\,|r^m|^2.\end{aligned}$$

Now specialize to the case $q = 1$ which includes multistep methods. By Corollary 4.3,

$$\begin{aligned} y^m - y^* &= [I - \alpha hJ(y^*)]^{-1} r^m + O(|r^m|^2)\\ &\equiv -\alpha^{-1}(hJ^* - \alpha^{-1}I)^{-1} r^m + O(|r^m|^2).\end{aligned}$$

Now $hJ^* - \alpha^{-1}I$ is negative dominant. Consider the vector

$$\delta y^m := (I - \alpha hJ^*)^{-1} r^m = y^m - y^* + O(|r^m|^2).$$

Let i be the index of the maximum component of δy^m. Then by (2.1),

$$|(\delta y^m)_i| < \frac{|\alpha^{-1}(r^m)_i|}{-\delta(\operatorname{Re} hJ^*_{ii} - \alpha^{-1})},$$

and $|y^m - y^*|$ differs from this quantity by $O(|r^m|^2)$. It follows that if stiff components (corresponding to $\operatorname{Re} J^*_{ii} << -1$) predominate in the error, then the error is much smaller than the residual, while if smooth components (corresponding to $|h \text{ Re } J_{ii}^*| < 1$) predominate, the error and the residual are of comparable magnitude. Analogous results hold for Runge-Kutta methods.

A more precise estimate may be formulated in the case of singular perturbation problems, cf. [23, Lemma 6]. Then J has the block structure

$$J = \begin{bmatrix} J_{11} & J_{12} \\ \varepsilon^{-1} J_{21} & \varepsilon^{-1} J_{22}, \end{bmatrix},$$

and Assumption 1.1 means negative diagonal dominance in the rows multiplied by ε^{-1}. Let us write the solution of (1.6a) simply as Y in (1.12). Then by Corollary 4.3,

$$Y^m - Y = (I_q \otimes I_N - (A \otimes I)h \,\mathrm{diag}[J_1, \dots, J_q])^{-1} r^m + O(|r^m|^2),$$

with $J_i = J(t_n + hc_i, y_{n,i})$, $i = 1, 2, \dots, q$. Introduce δY^m for the first term on the right, and permute variables according to the block structure of J to get

$$\sum_{j=1}^{q} \left(\delta_{ij} \begin{bmatrix} I_{n_1} & 0 \\ 0 & I_{n_2} \end{bmatrix} - h a_{ij} \begin{bmatrix} J_{11,j} & J_{12,j} \\ \varepsilon^{-1} J_{21,j} & \varepsilon^{-1} J_{22,j} \end{bmatrix} \right) \delta Y_j^m = r_i^m$$

into the form

$$\begin{bmatrix} I + O(h) & O(h) \\ (h/\varepsilon) O(1) & I - (h/\varepsilon)(A \otimes I)\mathrm{diag}[J_{22,i}]_{i=1,\dots,q} \end{bmatrix} \begin{bmatrix} \delta Y^m(1) \\ \delta Y^m(2) \end{bmatrix} = \begin{bmatrix} r^m(1) \\ r^m(2) \end{bmatrix}.$$

Thus, by negative dominance,

$$\begin{aligned} |\delta Y^m(1)| &\le \mathrm{const}(|r^m(1)| + h|r^m(2)|), \\ |\delta Y^m(2)| &\le \mathrm{const}(|r^m(1)| + (\varepsilon/h)|r^m(2)|). \end{aligned}$$

Finally, for evaluating (1.11) we consider the accuracy of approximation to $K^* = hF(Y^*) = (A^{-1} \otimes I)(Y^* - \Psi)$. In order not to magnify errors in components of Y^m, as would occur if they were passed through f, we take

$$K^m := (A^{-1} \otimes I)(Y^m - \Psi). \tag{4.1}$$

From Corollary 4.2 we have an error estimate.

Corollary 4.4. *There holds* $|K^m - K^*| \le K_3 |A^{-1}|\, |r^m|$.

The solution is advanced by $y_{n+1}^* = y_n + (b^T \otimes I)K^*$. If we accept Y^m and define stage derivatives by (4.1), then we advance the solution by

$$\begin{aligned} y_{n+1} &= y_n + (b^T \otimes I)(A^{-1} \otimes I)(Y^m - \Psi) \\ &= y_n + (b^T A^{-1} \otimes I)(Y^m - \Psi). \end{aligned}$$

The error estimate follows from Corollary 4.4.

Corollary 4.5. *We have* $|y_{n+1} - y^*_{n+1}| \leq |b^T A^{-1}|K_3|r^m|$.

We note that the bound $|b^T A^{-1}|$ rather than $|b^T|\,|A^{-1}|$, which may be considerably larger, can be achieved in practice by ordering the computations astutely [7].

5. The Knee Problem revisited

We conclude with an example to highlight some requirements for a robust algorithm. Consider the "Knee Problem" [11]

$$\begin{aligned} &\varepsilon\frac{dy}{dt} = (1-t-y)y = f(t,y), \qquad 0 < \varepsilon << 1, \\ &y(0) = 1. \end{aligned} \tag{5.1}$$

The reduced problem ($\varepsilon = 0$) corresponding to this singular perturbation problem has two solution branches $y = 1 - t$ and $y = 0$. Since $\partial f/\partial y = 1 - t - 2y$, the first branch is stable for $0 < t < 1$, and the second branch is stable for $1 < t$. The solution of (5.1) thus follows the branch $y = 1 - t$ for $0 < t < 1 - O(\varepsilon^{1/2})$ and is well approximated on this interval by the asymptotic solution

$$y_a(t;\varepsilon) := 1 - t + \frac{\varepsilon}{1-t}. \tag{5.2}$$

(We ignore the boundary layer correction in y' near $t = 0$.) In the interval $|t-1| = O(\varepsilon^{1/2})$, a transition occurs, and then $y(t) = O(\varepsilon)$ for $t > 1 + O(\varepsilon^{1/2})$; see Figure 2.

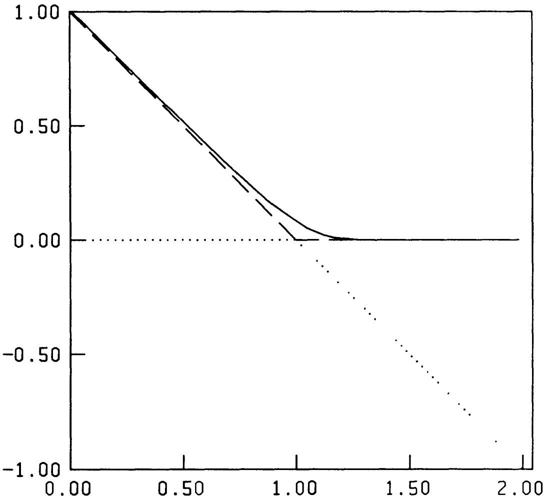

Figure 2
The Knee Problem

Consider now the approximate solution of (5.1) by the backward Cauchy-Euler method, with $0 \leq t_\nu < 1$:

$$\begin{aligned} y_{\nu+1} &= y_\nu + h_\nu \varepsilon^{-1} f(t_\nu + h_\nu,\ y_{\nu+1}) \\ &= y_\nu + h_\nu \varepsilon^{-1}(1 - t_\nu - h_\nu - y_{\nu+1}) y_{\nu+1}. \end{aligned} \tag{5.3}$$

This quadratic equation for $y_{\nu+1}$ has, naturally, two solutions—solutions of the implicit equations are seldom globally unique when f is nonlinear—provided $y_\nu > -(h_\nu(t_\nu + h_\nu - 1) + \varepsilon)^2/(4h_\nu\varepsilon)$. When $t_\nu + h_\nu < 1 - O(\varepsilon^{1/2})$ and $y_\nu \sim 1 - t_\nu$, these solutions have the form

$$y^*_{\nu+1} = 1 - t_\nu - h_\nu + O(\varepsilon), \tag{5.4}$$

$$y^\#_{\nu+1} = -\frac{y_\nu}{1 - t_\nu - h_\nu}\,\frac{\varepsilon}{h_\nu} + O(\varepsilon^2). \tag{5.5}$$

They are close to the (for $t < 1$!) stable and unstable solution branches, respectively, of the reduced problem. Dual solutions cause no difficulty away from $t = 1$. The iteration

$$y^{m+1}_{\nu+1} = y^m_{\nu+1} + (1 - h_0 J)^{-1} G(y_\nu, y^m_{\nu+1}) \tag{5.6}$$

with

$$G(y_\nu, y) = y_\nu + h_\nu \varepsilon^{-1}(1 - t_\nu - h_\nu - y)y - y \tag{5.7}$$

and

$$J = \varepsilon^{-1} \frac{\partial f}{\partial y}(0, 1) = -\varepsilon^{-1} \tag{5.8}$$

is rapidly convergent to the "correct" solution (5.4) of (5.3) from a reasonable initial guess, say

$$y^0_{\nu+1} = y_\nu + h_\nu \varepsilon^{-1} f(t_\nu, y_\nu), \tag{5.9}$$

whenever $y_\nu \sim 1 - t_\nu$.

We now consider the neighborhood of $t = 1$. Here the theory of §3 no longer justifies the procedure (5.6–5.9), and there is danger of finding the wrong solution. For example, if $t_\nu + \frac{1}{2}h_\nu = 1$—that is, $t = 1$ falls exactly in the middle of the step—and if y_ν lies on the asymptotic approximation (5.2), then $y^0_{\nu+1}$ in (5.9) is exactly the root (5.4) of (5.3). Now, however, this is the wrong root; (5.5) is the stable root for $t_{\nu+1} > 1$, just as in the differential equation.

Suppose now that we are not so unlucky as to predict the wrong root exactly. Take $(h_\nu = h)$

$$t_\nu = 1 - \tfrac{2}{3}h < 1, \qquad t_{\nu+1} = 1 + \tfrac{1}{3}h > 1,$$

$$y_\nu = y_a(t_\nu, \varepsilon) \ \text{ from (5.2)}, \qquad \text{and} \qquad y^0_{\nu+1} \ \text{ given by (5.9)},$$

so that

$$y^0_{\nu+1} = -\frac{h}{3} - \frac{3}{4}\frac{\varepsilon}{h}. \tag{5.10}$$

Then $y^{\#}_{\nu+1}$ (5.5) (the "correct" root) is an attracting fixed point for (5.6) while $y^{*}_{\nu+1}$ (5.4) (the "spurious" root) is repelling. Note that $y^{0}_{\nu+1}$ is closer to $y^{*}_{\nu+1}$ than to $y^{\#}_{\nu+1}$ for, say $h = 1/10$, $\varepsilon = 1/100$.

Upon performing the iteration (5.6) we find

$$G(y_\nu, y^0_{\nu+1}) = \frac{3}{4}h + \frac{27}{16}\varepsilon = O(h),$$

$$y^1_{\nu+1} - y^0_{\nu+1} = \frac{3}{16}\,\frac{4 + 9\varepsilon/h^2}{1 + \varepsilon/h}\varepsilon = O(\varepsilon).$$

If $0 < \varepsilon << h$ this is small enough to cause $y^1_{\nu+1}$ to be accepted by a test based on $|y^{m+1}_{\nu+1} - y^{m}_{\nu+1}|$, even though the iteration is diverging:

$$\begin{aligned} G(y_\nu, y^1_{\nu+1}) &= \frac{3}{4}h + \frac{1}{4}h^2 + \left(\frac{27}{16} + \frac{15}{16}h - \frac{13}{16}h^3\right)\varepsilon/h + O(\varepsilon^2/h^2) \\ &> G(y_\nu, y^0_{\nu+1}). \end{aligned}$$

Updating the Jacobian in (5.8) to $J = \frac{\partial f}{\partial y}(t_\nu, y_\nu)$ strengthens the respective attracting and repelling properties of $y^{\#}_{\nu+1}$ and $y^{*}_{\nu+1}$ without changing these observations in principle.

This example with a smooth differential equation reinforces the arguments of [37], which discusses an example with similar features: for it is common in problems of this type that a solution makes a fairly abrupt transition from one smooth manifold to another.

This has a clear implication for software. In a code for solving stiff problems by an implicit formula, one makes hypotheses about the smoothness of the solution and the vector field, in order to carry out expeditiously the task of solving the implicit equations to advance the step. Without a robust algorithm, however, the code can be deceived when these hypotheses turn out not to be correct. The experiment with the Knee Problem described in [11] apparently deceived the IMSL code in just this way.

Whether to use the residual or the difference of successive iterates in deciding to terminate the iteration has been discussed elsewhere [24]. These authors argue that the effect of a test based on residuals is to increase the number of iterations without substantially affecting either the step size history or the accuracy of the computed solution. The numerical tests supporting this conclusion, however, employ from the test set [20] only problems having a globally asymptotically stable rest point. Under these conditions an acceptance test based on closeness of successive iterates ordinarily yields an acceptable solution, for reasons already discussed in [38]. Thus these tests demonstrate nothing new, and they do not address the problem of robustness raised in [20, 37, 11].

Our analysis confirms the argument of [38] regarding the use of the residual as a stopping criterion for the modified Newton method. To overcome the relative lack of robustness of a test based on proximity of successive iterates, it appears that one must insist that at least two iterations be taken on every step and that their outcome should confirm—or rather, not refute—the hypothesis that the iteration is contracting at the required rate, cf. [38]. Even this is problematic if the first two iterates differ essentially by roundoff, as was already noticed in [38].

Acknowledgment

The author is grateful to have had the opportunity to reflect on this problem during a Faculty Improvement Leave in fall, 1986, and thanks H.F. Weinberger for hospitality at the Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota, where some of this work was done. Thanks are also due to Florian Potra for calling reference [16] to the author's attention.

Bibliography

1. R. Alt, *A-stable one-step methods with step-size control for stiff systems of ordinary differential equations*. J. Comput. Appl. Math. **4** (1978), 29–35.

→ K. Burrage, J. C. Butcher, and F. H. Chipman, *An implementation of singly-implicit Runge-Kutta methods*, BIT **20** (1980), 326–340.

3. J. C. Butcher, *Towards efficient implementation of singly-implicit methods*, ACM Trans. Math. Software **14** (1988), 68–75.

4. T. Chambers, *The use of numerical software in the digital simulation language PMSP*, Numerical Software—Needs and Availability (D. Jacobs, ed.), Academic Press, New York, 1978, pp. 257–278.

5. G. J. Cooper, *On the existence of solutions for algebraically stable Runge-Kutta methods*, IMA J. Numer. Anal. **6** (1986), 325–330.

→ M. Crouzeix, W. H. Hundsdorfer, and M. N. Spijker, *On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods*, BIT **23** (1983) , 84–91.

7. M. Crouzeix and P.-A. Raviart, *Approximation des problèmes d'évolution, Ch. 2: Méthodes de Runge-Kutta*, Lecture Notes, Université de Rennes, 1980.

8. A. R. Curtis, *Jacobian matrix properties and their impact on choice of software for stiff ODE systems*, IMA J. Numer. Anal. **3** (1983), 397–415.

9. A. R. Curtis and J. K. Reid, *The choice of step lengths when using differences to approximate Jacobian matrices*, J. Inst. Math. Appl. **13** (1974), 121–126.

10. G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, Numerical Analysis (Dundee 1975), Lecture Notes in Math., vol. 506, Springer, Berlin, 1976, pp. 60–72.

11. G. Dahlquist, L. Edsberg, G. Sköllermo, and G. Söderlind, *Are the numerical methods and software satisfactory for chemical kinetics*? Numerical Integration of Differential Equations and Large Linear Systems (J. Hinze, ed.), Lecture Notes in Math., vol. 968, Springer, Berlin, 1982, pp. 149–164.

12. K. Dekker, *Error bounds for the solution to the algebraic equations in Runge-Kutta methods*, BIT **24** (1984), 347–356.

13. K. Dekker and Ernst Hairer, *A necessary condition for BSI stability*, BIT **25** (1985), 285–288.

14. K. Dekker and J. G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, North-Holland, Amsterdam, 1984.

15. J. E. Dennis and R. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice-Hall, Englewood Cliffs, N.J., 1983.

16. P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal. **16** (1979), 1–10.

17. G. di Lena and R. I. Peluso, *On conditions for the existence and uniqueness of solutions to the algebraic equations in Runge-Kutta methods*, BIT **25** (1985), 223–232.

18. J. Dieudonné, *Foundations of modern analysis*, Academic Press, New York, 1960.

19. W. H. Enright, *Improving the efficiency of matrix operations in the numerical solution of stiff ordinary differential equations*, ACM Trans. Math. Software **4** (1978), 127–136.

20. W. H. Enright, T. E. Hull and B. Lindberg, *Comparing numerical methods for stiff systems of ODE:s*, BIT **15** (1975), 10–48.

21. R. Frank, J. Schneid, and C. W. Ueberhuber, *Stability properties of implicit Runge-Kutta methods*, SIAM J. Numer. Anal. **22** (1985), 497–514.

22. C. W. Gear, *Numerical initial value problems in ordinary differential equations*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

: → E. Hairer, Ch. Lubich, and M. Roche, *Error of Runge-Kutta methods for stiff problems studied via differential-algebraic equations*, BIT **28** (1988), 678–700.

24. N. Houbak, S. P. Nørsett, and P. G. Thomsen, *Displacement or residual test in the application of implicit methods for stiff problems*, IMA J. Numer. Anal. **5** (1985), 297–305.

25. W. H. Hundsdorfer, *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods*, CWI Tract 12, Centre for Mathematics and Computer Science, Amsterdam, 1985.

26. W. H. Hundsdorfer and J. Schneid, *On the equivalence of BS-stability and B-consistency*, BIT **29** (1989), 505–511.

27. W. H. Hundsdorfer and M. N. Spijker, *On the algebraic equations in implicit Runge-Kutta methods*, SIAM J. Numer. Anal. **24** (1987), 583–594.

28. O. A. Karakashian and W. Rust, *On the parallel implementation of implicit Runge-Kutta methods*, SIAM J. Sci. Statist. Comput. **9** (1988), 1085–1090.

29. H.-O. Kreiss, *Difference methods for stiff ordinary differential equations*, SIAM J. Numer. Anal. **15** (1978), 21–58.

30. J.-X. Kuang and J.-X. Xiang, *On the D-suitability of implicit Runge-Kutta methods*, BIT **29** (1989), 321–327.

31. S. Lefschetz, *Differential equations: Geometric theory*, Dover, New York, 1977.

32. M. Z. Liu, K. Dekker, and M. N. Spijker, *Suitability of Runge-Kutta methods*, J. Comput. Appl. Math. **20** (1987), 307–315.

33. M. Z. Liu and J. F. B. M. Kraaijevanger, *On the solvability of the systems of equations arising in implicit Runge-Kutta methods*, BIT **28** (1988), 825–838.

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

35. H. H. Robertson and J. Williams, *Some properties of algorithms for stiff differential equations*, J. Inst. Math. Appl. **16** (1975), 23–34.

36. B. A. Schmitt, *Stability of implicit Runge-Kutta methods for nonlinear stiff differential equations*, BIT **28** (1988), 884–897.

37. L. F. Shampine, *Implementation of implicit formulas for the solution of ODEs*, SIAM J. Sci. Statist. Comput. **1** (1980), 103–118.

38. ——, *Evaluation of implicit formulas for the solution of ODEs*, BIT **19** (1979), 495–502.

39. ——, *Solving ODEs in quasi steady state*, Numerical Integration of Differential Equations and Large Linear Systems (J. Hinze, ed.), Lecture Notes in Math., vol. 968, Springer, Berlin, 1982, pp. 234–245.

40. →—, *Implementation of Rosenbrock methods*, ACM Trans. Math. Software **8** (1982), 93–113.

41. ——, *Type-insensitive ODE codes based on implicit $A(\alpha)$-stable formulas*, Math. Comp. **39** (1982), 109–124.

42. L. F. Shampine and C. W. Gear, *A user's view of solving stiff ordinary differential equations*, SIAM Rev. **21** (1979), 1–17.

43. D. R. Smith, *Singular-perturbation theory*, Cambridge Univ. Press, Cambridge, 1985.

44. M. W. Spijker, *Feasibility and contractivity in implicit Runge-Kutta methods*, J. Comput. Appl. Math. **12** & **13** (1985), 563–578.

45. J. M. Varah, *On the efficient implementation of implicit Runge-Kutta methods*, Math. Comp. **33** (1979), 557–562.

46. T. J. Ypma, *Relaxed Newton-like methods for stiff differential systems*, J. Comput. Appl. Math. **16** (1986), 95–103.

47. ——, *Efficient estimation of sparse Jacobian matrices by differences*, J. Comput. Appl. Math. **18** (1987), 17–28.

48. Zahari Zlatev, *Modified diagonally implicit Runge-Kutta methods*, SIAM J. Sci. Statist. Comput. **2** (1981), 321–334.

DEPARTMENT OF MATHEMATICS, IOWA STATE UNIVERSITY, AMES, IOWA 50011
E-mail address: alex@pollux.math.iastate.edu