THE NUMERICAL SOLUTION OF NONLINEAR STIFF INITIAL VALUE PROBLEMS

AN ANALYSIS OF ONE-STEP METHODS

W.H. HUNDSDORFER

1. Zij k een geheel getal, $k \geq 0$. Beschouw de volgende drie aannamen.

(a1) X is een reële of complexe Hilbertruimte met inwendig produkt $(\cdot,\cdot)$ en norm $\|\cdot\|$;

(a2) $A: X \to X$ is lineair, $\mathrm{Re}(Ax,x) \leq 0$ en $\|Ax\| \leq \|x\|$ (voor alle $x \in X$);

(a3) De rationale functie ϕ , gedefinieerd door $\phi(\zeta) = (q_0+q_1\zeta+\ldots+q_k\zeta^k)^{-1}\,(p_0+p_1\zeta+\ldots+p_k\zeta^k)$ $(\zeta\in\mathbb{C})$, heeft reële coëfficiënten q_j,p_j $(0\leq j\leq k)$. Verder is $\phi(\zeta)$ regulier en $|\phi(\zeta)| \leq 1$ voor alle $\zeta \in \mathbb{C}$ met $\mathrm{Re}\,\zeta \leq 0$, $|\zeta| \leq 1$.

De volgende bewering geldt dan en slechts dan als $k \leq 1$.

(b) (a1), (a2) en (a3) tezamen impliceren dat $\phi(A) = (q_0I+q_1A+\ldots+q_kA^k)^{-1}\,(p_0I+p_1A+\ldots+p_kA^k)$ gedefinieerd is en $\|\phi(A)x\| \leq \|x\|$ (voor alle $x \in X$).

2. Voor gegeven getallen $\rho > 0$ en $\sigma \geq 0$ beschouwen we de klasse $F_{\rho,\sigma}$ van functies F die aan de volgende voorwaarden (1)-(4) voldoen. De Euclidische norm van $x \in \mathbb{R}^n$ duiden we aan met $\|x\| = (x^Tx)^{\frac{1}{2}}$.

(1) $F: \mathbb{R}^n \to \mathbb{R}^n$, $n \geq 1$, alle partiële afgeleiden van F bestaan en zijn continu op $\mathbb{R}^n$.

(2) Er is een $x^* \in \mathbb{R}^n$ met $F(x^*) = 0$.

(3) $(F(x)-F(y))^T(x-y) \geq 0$ (voor alle $x,y \in \mathbb{R}^n$).

(4) Voor alle $x,y \in \mathbb{R}^n$ met $\|x-x^*\| \leq \rho$, $\|y-x^*\| \leq \rho$ is er een lineaire afbeelding $E(x,y)$ van $\mathbb{R}^n$ naar $\mathbb{R}^n$ zodat $\|E(x,y)v\| \leq \sigma\|v\|$ (voor elke $v \in \mathbb{R}^n$) en

$$F'(x) = F'(y)(I+E(x,y)) .$$

Zij $\omega \geq 1$ en $\lambda > 0$. Voor gegeven $F \in F_{\rho,\sigma}$ beschouwen we het iteratieve proces

$$x_{k+1} = x_k - \omega[F'(x_k)+\lambda I]^{-1}F(x_k) \qquad (k=0,1,2,\ldots)$$

voor het benaderen van x^* .

Er geldt $\|x_{k+1}-x^*\| \leq \|x_k-x^*\|$ (voor alle $F \in F_{\rho,\sigma}$, $k \geq 0$, en startwaarden $x_0 \in \mathbb{R}^n$ met $\|x_0-x^*\| \leq \rho$) dan en slechts dan als $\omega \leq 2/(1+\sigma)$.

[1] ORTEGA, J.M., RHEINBOLDT, W.C., *Iterative solution of nonlinear equations in several variables*, Academic Press, New York 1970.

3. Voor vierkante complexe matrices $A = (a_{ij})$ definiëren we

$$\|A\| = \max_i \sum_j |a_{ij}| .$$

Zij $\theta > 0$. Beschouw de volgende twee uitspraken.

(i) A is een vierkante matrix met $\|A\| \leq \alpha$ en $\|\exp(tA)\| \leq 1$ (voor alle $t \geq 0$);

(ii) $I - \theta A$ is regulier, en $\|(I-\theta A)^{-1}(I+(1-\theta)A)\| \leq 1$.

We definiëren

$$\alpha^* = \sup\{\alpha : \alpha \geq 0 \text{ en (ii) geldt voor alle } A \text{ die aan (i) voldoen}\} .$$

Er geldt

$$\alpha^* > 0 \text{ d.e.s.d. als } \theta > \tfrac{1}{2} .$$

4. Zij G een Runge-Kutta methode met coëfficiënten a_{ij} en b_i $(1 \leq i,j \leq m)$. Neem aan dat er getallen d_j $(1 \leq j \leq m)$ bestaan zodat

$$b_i = \sum_{j=1}^{m} d_j a_{ji} \qquad (1 \leq i \leq m) .$$

Neem verder aan dat G BSI-stabiel is. Dan is G ook BS-stabiel.

[2] FRANK, R., SCHNEID, J., UBERHUBER, C.W., *Stability properties of implicit Runge-Kutta methods*, Report no. 52/82, Inst. Num. Math., Tech. Univ. Wien (1982).

5. Het bewijs van stelling 2 uit [3] is onjuist.

[3] BRENNER, P., THOMÉE, V., *On rational approximations of semi-groups*, SIAM J. Numer. Anal. 16, 683-694 (1979).

6. Zij A een reële $m \times m$-matrix, $m \geq 1$, en $\gamma > 0$. We beschouwen de collectie hypervlakken

$$H = \{H(I_1,I_2,\ldots,I_n): n \geq 1,\ I_1,I_2,\ldots,I_n \text{ zijn onderling disjuncte deelverzamelingen van } \{1,2,\ldots,m\} \text{ en elke } I_k\ (1 \leq k \leq n) \text{ bevat minstens twee elementen}\},$$

waarbij

$$H(I_1,I_2,\ldots,I_n) = \{v: v=(v_1,v_2,\ldots,v_m)^T \in \mathbb{R}^m,\ v_i = v_j \text{ als } i \text{ en } j \text{ tot een zelfde verzameling } I_k\ (1 \leq k \leq n) \text{ behoren}\}.$$

Neem aan dat er voor iedere $H \in H$ een $v \in H$ bestaat zodat $Av \notin H$. Dan zijn er vectoren $x = (x_1,x_2,\ldots,x_m)^T$ en $y = (y_1,y_2,\ldots,y_m)^T$ in $\mathbb{R}^m$ zodat $y = Ax$ en

$$x_i \neq x_j,\ (y_i-y_j)/(x_i-x_j) \leq -\gamma \quad (\text{voor } 1 \leq i < j \leq m).$$

[4] HUNDSDORFER, W.H., SPIJKER, M.N., *A note on* B-*stability of Runge-Kutta methods*, Numer. Math. 36, 319-331 (1981).

[5] BUTCHER, J.C., *A short proof concerning* B-*stability*, BIT 22, 528-529 (1982).

7. Middelbare scholieren die wiskunde willen gaan studeren moet worden ontraden om, ten koste van het aantal moderne talen, het vak "wiskunde II" in het eind-examen pakket op te nemen.

THE NUMERICAL SOLUTION OF NONLINEAR STIFF INITIAL VALUE PROBLEMS

AN ANALYSIS OF ONE-STEP METHODS

THE NUMERICAL SOLUTION OF NONLINEAR STIFF INITIAL VALUE PROBLEMS

AN ANALYSIS OF ONE-STEP METHODS

PROEFSCHRIFT

ter verkrijging van de graad van Doctor in de Wiskunde en Natuurwetenschappen aan de Rijksuniversiteit te Leiden, op gezag van de Rector Magnificus Dr. A.A.H. Kassenaar, Hoogleraar in de faculteit der Geneeskunde, volgens besluit van het College van Dekanen te verdedigen op woensdag 28 maart 1984 te klokke 15.15 uur

door

WILLEM HANS HUNDSDORFER

geboren te Graz in 1954.

1984
Centrum voor Wiskunde en Informatica, Amsterdam

PROMOTIECOMMISSIE

promotor	prof.dr. M.N. Spijker
referenten	prof.dr. P.J. van der Houwen
	prof.dr. M. van Veldhuizen
overige leden	prof.dr. G. van Dijk
	prof.dr.ir. L.A. Peletier
	dr. J.G. Verwer

CONTENTS

CHAPTER 1

INTRODUCTION

1.1. THE PROBLEM OF STIFFNESS

The subject of this monograph is the numerical solution of initial value problems for systems of ordinary differential equations. Numerical methods for such problems have already been known for a long time. The oldest and most simple one is Euler's method (cf. example 1.1.1). More sophisticated and accurate methods have been proposed afterwards, and nowadays very efficient procedures exist for a large class of problems arising in practice. Still there is a class of problems, the so-called stiff problems, for which the numerical procedures are less efficient and reliable. With such problems we shall be concerned in the following. In this section it will be explained what is meant by stiffness.

We consider an initial value problem for an autonomous system of s ordinary differential equations

$$U'(t) = f(U(t)) \qquad (0 \le t \le T) , \tag{1.1.1.a}$$

$$U(0) = u_0 . \tag{1.1.1.b}$$

This problem may be real or complex. The function $f: \mathbb{K}^s \to \mathbb{K}^s$, the vector $u_0 \in \mathbb{K}^s$, and the positive number T are given. Here $\mathbb{K}$ stands for either the set of real numbers $\mathbb{R}$ or the set of complex numbers $\mathbb{C}$, and s is a positive integer. It will always be assumed that the problem (1.1.1) is such that there exists a unique solution U.

For the numerical solution of (1.1.1) we shall deal with step-by-step methods, which produce approximations u_n to the solution U at *gridpoints* t_n. These gridpoints are defined by $t_n = t_{n-1} + h_n$ $(1 \le n \le N)$, $t_0 = 0$, where the numbers $h_n > 0$ are the *stepsizes* and $t_N = T$. If all stepsizes are equal, say $h_n = h$ $(1 \le n \le N)$, the grid $\{t_n\}$ is said to be *uniform*. For convenience this will be assumed in the subsequence.

Application of a step-by-step method to a concrete problem (1.1.1) results in a difference scheme from which the approximations u_n can be computed one after another. We shall deal only with *one-step-methods* where for the computation of u_n only u_{n-1} is needed and not the u_j with $j \le n-2$.

During the numerical integration errors will be introduced. Such errors are mainly caused by the fact that the differential equation is replaced by a difference scheme; these are the so-called local discretization errors (see section 3.1). Also by the computer arithmetic errors will be produced in the computation of u_n from u_{n-1} . The introduction of these local errors, which are small if the stepsize is chosen properly, does not lead to a bad overall result if the effect of such small errors remains small. Thus if we have two sequences $\{\tilde{u}_n\}$, $\{u_n\}$ satisfying the difference scheme, and $\tilde{u}_k = U(t_k)$, $u_k \neq U(t_k)$, we like to know whether $|\tilde{u}_{k+n}-u_{k+n}|$ remains small for all n ,if $|\tilde{u}_k-u_k|$ is so. We call a numerical scheme *stable* if there exists a constant $\sigma > 0$ such that

$$(1.1.2) \qquad |\tilde{u}_{k+n}-u_{k+n}| \le \sigma|\tilde{u}_k-u_k| \qquad (n=1,2,\ldots,N-k)$$

for all $\tilde{u}_k, u_k \in \mathbb{K}^s$ and $1 \le k < N$. Here $|\cdot|$ stands for some given norm on $\mathbb{K}^s$.

If a numerical scheme is known to be stable with a moderate stability constant σ , and the local errors are small, then also the global errors $|U(t_n)-u_n|$ $(1 \le n \le N)$ will be small.

For the stability analysis it is often assumed that the function f appearing in the right-hand side of the differential equation satisfies a Lipschitz condition

$$(1.1.3) \qquad |f(\tilde{x})-f(x)| \le L|\tilde{x}-x| \quad \text{(for all } \tilde{x},x \in \mathbb{K}^s\text{)},$$

which implies

$$(1.1.4) \qquad |\tilde{U}(t+\Delta t)-U(t+\Delta t)| \le e^{L\Delta t}\,|\tilde{U}(t)-U(t)| \qquad (0 \le t < t+\Delta t \le T)$$

for any two solutions $\tilde{U},U$ of the differential equation (1.1.1.a). There exists a rather satisfactory theory by means of which one can predict how well a numerical scheme will approximate the exact solution U of (1.1.1) ,

provided that the product TL is not too large and hL is sufficiently small. Important contributions to this theory can be found in the books by DAHLQUIST (1959), HENRICI (1962) and STETTER (1973). Actually the Lipschitz condition (1.1.3) only needs to hold on some tube around the solution U. Many practical problems meet this requirement with a Lipschitz constant L such that TL is not very large. In such a case the requirement on hL will not lead to an excessively large number of steps.

In (1.1.4) equality is possible for functions f satisfying (1.1.3) (e.g. with $\mathbb{K}^s = \mathbb{R}^1$, $f(x) = Lx$ $(x \in \mathbb{R})$). If this is the case and $\tilde{u}_N, u_N$ are good approximations to $\tilde{U}(t_N), U(t_N)$, respectively, we therefore have $|\tilde{u}_N - u_N| \approx e^{LT}|\tilde{u}_0 - u_0|$, and the stability inequality (1.1.2) will only hold with $\sigma \gtrapprox e^{LT}$. From this observation it can be seen that if TL is very large and (1.1.4) is not pessimistic any numerical scheme will encounter serious problems.

Suppose the norm $|\cdot|$ is generated by an inner product $\langle \cdot,\cdot \rangle$ on $\mathbb{K}^s$, and suppose, instead of or in addition to (1.1.3), that f satisfies a so-called *one-sided Lipschitz condition*

$$\text{(1.1.5)} \qquad \mathrm{Re}\,\langle f(\tilde{x}) - f(x), \tilde{x} - x \rangle \le \beta |\tilde{x} - x|^2 \qquad (\text{for all } \tilde{x}, x \in \mathbb{K}^s),$$

with $\beta \in \mathbb{R}$ given. This condition implies that we have for arbitrary solutions $\tilde{U}, U$ of (1.1.1.a)

$$\text{(1.1.6)} \qquad |\tilde{U}(t+\Delta t) - U(t+\Delta t)| \le e^{\beta \Delta t} |\tilde{U}(t) - U(t)| \qquad (0 \le t < t+\Delta t \le T)$$

(see e.g. DAHLQUIST (1959)). Even if the Lipschitz constant L is very large, the one-sided Lipschitz constant β may be close to zero or negative. The inequality (1.1.4) is then much too pessimistic. As an illustration we consider the following simple problem.

EXAMPLE 1.1.1. Let $\mathbb{K}^s = \mathbb{C}^1$, $T = 1$, and let f be defined by

$$f(x) = \lambda x \qquad (\text{for } x \in \mathbb{C})$$

where $\lambda \in \mathbb{C}$ satisfies $\mathrm{Re}\,\lambda \le 0$, $|\lambda| = 10^6$. With $\langle x, y \rangle = x\bar{y}$ (for $x, y \in \mathbb{C}$), (1.1.5) holds with $\beta = 0$, whereas we have to take $L = 10^6$ for (1.1.3) to hold.

The method of Euler is given by

$$u_n = u_{n-1} + hf(u_{n-1}) \qquad (1\le n\le N) \ ,$$

from which we obtain here the approximations $u_n = (1+h\lambda)^n u_0$ $(1\le n\le N)$ to the solution $U(t) = e^{\lambda t} u_0$ of (1.1.1). For two sequences of approximations computed with different starting values $\tilde{u}_0, u_0$, we have

$$|\tilde{u}_n - u_n| = |1+h\lambda|^n \, |\tilde{u}_0 - u_0| \qquad (1\le n\le N) \ .$$

If hL is large, say $h = 10^{-3}$, then $|1+h\lambda| \approx 10^3$ and $N = 10^3$. Thus we have the unfavourable result

$$|\tilde{u}_N - u_N| \approx (1000)^{1000} \, |\tilde{u}_0 - u_0| \ .$$

A completely different behaviour shows up with the Backward Euler method where the approximations are computed according to

$$u_n = u_{n-1} + hf(u_n) \qquad (1\le n\le N) \ .$$

For this case we obtain $\tilde{u}_n - u_n = (1-h\lambda)^{-n} (\tilde{u}_0 - u_0)$ $(1\le n\le N)$, and now we have

$$|\tilde{u}_n - u_n| \le |\tilde{u}_0 - u_0| \qquad (1\le n\le N)$$

for all stepsizes $h > 0$.

We thus see from this example that there are numerical methods, such as the Backward Euler method, which may produce stable approximations with a stability constant $\sigma \approx 1$ no matter how large hL is. If the solution U we want to approximate is smooth (slowly varying) then the local discretization errors will be small for stepsizes h which are of moderate size. For such a problem we then may obtain accurate approximations, even if hL is large.

Suppose the solution U of (1.1.1) is smooth, and (1.1.5) holds with $\beta \lessapprox 0$. We call this solution U *stiff* if for any open region $\mathcal{D} \subset \mathbb{K}^s$ containing $\{U(t): 0\le t\le T\}$ the Lipschitz constant L of the restriction of f to $\mathcal{D}$ is such that TL is very large.

Many numerical methods, for instance Euler's method, are unsuited to solve stiff problems (problems with a stiff solution), due to the fact that

a large value of TL will casue an extremely large stability constant unless hL is sufficiently small. Such a restriction on h leads to a very large number of steps. On the other hand there are methods, for example the Backward Euler method, for which the requirement on hL is unrealistic, and for such methods we may obtain accurate approximations to stiff solutions with a moderate number of steps.

Stiff initial value problems are encountered in many practical situations, for instance in control theory, chemical kinetics and diffusion processes. A detailed discussion and more examples can be found in WILLOUGHBY (1974) and BJUREL et al. (1970). Very often in these situations the solution U of (1.1.1) is such that U(t) varies wildly for a short period, the so-called transient phase. After this phase U(t) becomes smooth. We then only call the latter part of the solution stiff. In the transient phase the use of a small stepsize is inevitable for keeping the local errors small. Since this period is only very short this does not lead to an extremely large number of steps. After the transient phase a larger stepsize can be used, provided that the scheme remains stable, and there a numerical method which is suited for stiff problems is recommended.

REMARK 1.1.2. Instead of (1.1.1) we could also consider the more general (nonautonomous) problem

$$U'(t) = f(t,U(t)) \qquad (0\le t\le T)\,, \tag{1.1.7.a}$$

$$U(0) = u_0 \tag{1.1.7.b}$$

where $f: \mathbb{R} \times \mathbb{K}^s \to \mathbb{K}^s$ and $u_0 \in \mathbb{K}^s$. We will restrict ourselves to (1.1.1) for the sake of simplicity. The analysis presented in this monograph could also be given for nonautonomous problems (1.1.7), but this would only complicate the results in some chapters without leading to new insights. Moreover, any nonautonomous initial value problem (1.1.7) can be converted to an autonomous problem by adding the differential equation $U'_{s+1}(t) = 1$, with initial value $U_{s+1}(0) = 0$, to the system and replacing $f(t,U(t))$ by $f(U_{s+1}(t),U(t))$. After this conversion we obtain an initial value problem for an autonomous system of s+1 differential equations.

1.2. SOME STABILITY AND CONTRACTIVITY CONCEPTS

In this section some stability concepts will be introduced which indicate whether a given method is suited for stiff initial value problems. For stiff problems we would like to have schemes which remain stable with a moderate stability constant σ no matter how large L and T are. In the following we therefore take $T = \infty$ and consider classes of problems which are such that there is no (uniform) upper bound for the Lipschitz constants of the functions f in such a class.

Suppose the differential equation (1.1.1.a) is *dissipative*, i.e. for any two solutions $\tilde{U},U$ of (1.1.1.a) and arbitrary $t \geq 0$, $h > 0$ we have

$$(1.2.1) \qquad |\tilde{U}(t+h)-U(t+h)| \leq |\tilde{U}(t)-U(t)| \ .$$

A numerical scheme is said to be *contractive* if for arbitrary starting vectors $\tilde{u}_0,u_0$, the following discrete version of (1.2.1) holds,

$$(1.2.2) \qquad |\tilde{u}_{n+1}-u_{n+1}| \leq |\tilde{u}_n-u_n| \qquad (n=0,1,2,\ldots) \ .$$

For dissipative systems contractivity is a natural requirement. Moreover, contractivity implies stability (in the sense of (1.1.2)) with stability constant $\sigma = 1$.

As we shall see in chapter 5 the stability properties of a method for general stiff problems can be predicted to some extent by knowledge of the behaviour of the method on the class of simple one-dimensional linear problems where $\mathbb{K}^s = \mathbb{C}^1$ and

$$(1.2.3) \qquad f(x) = \lambda x \qquad (\text{for } x \in \mathbb{C}) \text{ with } \lambda \in \mathbb{C} \ .$$

The one-step methods we will consider are such that for these testproblems we get a scheme of the type

$$u_n = \phi(h\lambda)u_{n-1} \qquad (n=1,2,3,\ldots)$$

where $\phi: \mathbb{C} \to \mathbb{C}$ is a rational function which only depends on the method. For such schemes contractivity and stability (in the sense of (1.1.2) with $N = \infty$) are equivalent: if $|\phi(h\lambda)| \leq 1$ we have contractivity, and if $|\phi(h\lambda)| > 1$ then $|\tilde{u}_n-u_n| \to \infty$ $(n \to \infty)$ whenever $\tilde{u}_0 \neq u_0$.

ollowing definition is due to DAHLQUIST (1963). It is concerned
lass of problems where f satisfies (1.2.3) with Re $\lambda \leq 0$.

1.2.1. *A one-step method is said to be* A-stable *if the contract-*
tion (1.2.2) *holds for all dissipative problems satisfying* (1.2.3)
ary stepsizes $h > 0$.

bility has been established for many numerical methods. A funda-
tribution to this subject is given by WANNER, HAIRER and NØRSETT

eneral one-step methods A-stability is no guarantee that the method
ces stable results for arbitrary *nonlinear* stiff problems. It can
see e.g. SPIJKER (1982 B) and section 5.3 of this monograph, that
hod is A-stable, the differential equation (1.1.1.a) is linear
ative, and the norm $|\cdot|$ on $\mathbb{K}^s$ is generated by an inner product,
ontractivity relation (1.2.2) will hold for arbitrary stepsizes
e situation with arbitrary norms is totally different. For results
rection we refer to NEVANLINNA and LININGER (1978,1979), BRENNER
(1979) and SPIJKER (1982 A).
all confine ourselves to the case where the norm $|\cdot|$ is
by an inner product $\langle \cdot, \cdot \rangle$.

1.2.2. *A one-step method is said to be* B-contractive *if the*
ity relation (1.2.2) *holds whenever* $h > 0$, *and* f *satisfies*
ith $\beta = 0$.

concept, G-stability, was introduced by DAHLQUIST (1975) for
methods. Definition 1.2.2 is due to BUTCHER (1975), who used the
bility. For the schemes which arise if a one-step method is
a nonlinear problem, there is no equivalence between stability
ctivity. Therefore we have chosen the term B-contractivity in
1.2.2.
e definitions 1.2.1, 1.2.2 no restriction is imposed on the step-
a dissipative problem is solved numerically with a B-contractive
stepsize has only to be chosen in such a way that the local
ain small, and we need not worry about an unfavourable error
n.

1.3. THE SCOPE OF THE STUDY

In this monograph we shall mainly be concerned with implicit Runge-Kutta methods and semi-implicit one-step methods for nonlinear stiff initial value problems. These methods can be viewed as generalizations of explicit Runge-Kutta methods, and they constitute the class of one-step methods which are mostly used for stiff problems.

First, in chapter 2, useful technical results will be obtained.

In chapter 3 we take a closer look at the implicit Runge-Kutta methods and the semi-implicit methods, which have to be provided with a suitable approximation to the Jacobian f'. Some restrictions on the semi-implicit methods will be motivated in this chapter.

In chapter 4 we consider the following subject. If an implicit Runge-Kutta method is used for the numerical solution of a nonlinear initial value problem, at each step of the integration a nonlinear system of *algebraic* equations has to be solved. The approximations u_n are only well defined if the algebraic equations are uniquely solvable. It is known (see e.g. GRIGORIEFF (1972)) that there exists a unique solution if the product of the stepsize h with the Lipschitz constant L of f is sufficiently small. The implicit methods however are intended to deal with cases where hL is not small.

Using only the one-sided Lipschitz condition (1.1.5) and continuity as assumptions on f, and thus allowing L to be arbitrary large, sufficient conditions, on the coefficients of the Runge-Kutta method and the stepsize h, for having a unique solution to these algebraic equations will be presented in section 4.3. Similar results obtained in CROUZEIX, HUNDSDORFER and SPIJKER (1983) and DEKKER (1982) are slightly generalized. By a combination of these sufficient conditions with results of FRANK, SCHNEID and UEBERHUBER (1982 A), it can be shown that for many B-contractive Runge-Kutta methods the numerical approximations are always well defined for dissipative problems, without any restriction on the stepsize. This is a useful completion of the B-contractivity concept for these methods.

If we deal with linear differential equations or with semi-implicit methods only linear systems of algebraic equations have to be solved. These cases are considered in the sections 4.2 and 4.4.

In chapter 5 the error propagation in implicit and semi-implicit methods applied to nonlinear stiff initial value problems is studied. We shall assume that it is known in advance how the methods behave on the

scalar, linear testproblems (1.2.3), and consider questions of the following type. Suppose a method is A-stable. Can we then also apply this method to nonlinear, nonscalar, dissipative problems without having an unfavourable error propagation?

For a rather large class of implicit Runge-Kutta methods (the B-contractive ones) the answer is already known to be affirmative. The implicit Runge-Kutta methods are therefore only considered shortly in section 5.5. For the semi-implicit methods much less is known at present. It will be shown in section 5.4 that if we use a Rosenbrock method (a semi-implicit method using an exact Jacobian approximation) for approximating a slowly varying solution U of (1.1.1), and the variation of the Jacobian f' is restricted, the error propagation will not differ much from the case where we deal with a testproblem (1.2.3). Also for semi-implicit methods using a fixed Jacobian approximation we get this positive result, provided that the Jacobian approximation is good enough. These results indicate for what kind of nonlinear initial value problems the semi-implicit methods are suited. In this analysis essential results of HAIRER, BADER and LUBICH (1982) (on semi-implicit methods using a constant Jacobian approximation) and of HUNDSDORFER (1981) (on a small class of Rosenbrock methods) are generalized. Besides it will be proved that such positive results are not valid for all choices of the Jacobian approximation in semi-implicit methods.

For some simple implicit and semi-implicit methods the bounds on the error propagation obtained in chapter 5 are used in chapter 6 to derive convergence results where the initial value problem may be arbitrarily stiff.

CHAPTER 2

PRELIMINARIES

2.1. GENERAL NOTATIONS AND CONVENTIONS

In this section we introduce some notations that will be used throughout all chapters.

The set $\mathbb{K}$ stands consistently for either the set of real numbers $\mathbb{R}$ or the set of complex numbers $\mathbb{C}$. The set of natural numbers $\{1,2,3,\ldots\}$ is denoted by $\mathbb{N}$. Further we define $\mathbb{R}^+ = \{\xi:\xi\in\mathbb{R},\xi\geq 0\}$ and $\mathbb{C}^- = \{\zeta:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta\leq 0\}$.

Let $m \in \mathbb{N}$. The vectors $e_1^{(m)},e_2^{(m)},\ldots,e_m^{(m)}$ will stand for the unit vectors in $\mathbb{R}^m$; the j-th. component of $e_i^{(m)}$ equals 1 if $j = i$, and 0 otherwise. Generally we simply write e_i instead of $e_i^{(m)}$. The vector in $\mathbb{R}^m$ all of whose components equal 1 will be denoted by $e^{(m)}$ or e.

Let X and Y be finite dimensional normed vectorspaces over $\mathbb{K}$. Then $L(X,Y)$ denotes the space of linear operators from X to Y. On $L(X,Y)$ we will consider the operator norm which is induced by the norms on X and Y. We denote $L(X,X)$ shortly by $L(X)$. For given $A \in L(X)$, $B \in L(X,L(X))$, we define $AB \in L(X,L(X))$ by $(AB)x = A(Bx)$ (for $x \in X$).

If $A = (a_{ij})$ is an $n \times m$ matrix with entries a_{ij} in $\mathbb{K}$, we shall also write $A \in L(\mathbb{K}^m,\mathbb{K}^n)$. Thus no distinction will be made between a linear operator and its usual matrix representation.

The identity operator in $L(\mathbb{R}^m)$ will be denoted by $I^{(m)}$ or simply by I.

Let $s \in \mathbb{N}$, $\Lambda \in L(\mathbb{K}^s)$, and let $\langle\cdot,\cdot\rangle$ stand for an inner product on $\mathbb{K}^s$ with corresponding norm $|x| = \langle x,x\rangle^{\frac{1}{2}}$ (for $x \in \mathbb{K}^s$). We then also use $|\cdot|$ to denote the induced operator norm of Λ,

$$|\Lambda| = \sup\{|\Lambda x|:x\in\mathbb{K}^s,|x|=1\}\ . \tag{2.1.1}$$

The logarithmic norm $\mu[\Lambda]$ of Λ (w.r.t. the inner product $\langle\cdot,\cdot\rangle$) is given by

(2.1.2) $\mu[\Lambda] = \sup\{\mathrm{Re}\langle x,\Lambda x\rangle : x\in\mathbb{K}^s, |x|=1\}$

(cf. STRÖM (1975)). The set of eigenvalues (in $\mathbb{C}$) of the operator Λ will be denoted by $\sigma(\Lambda)$.

Let A be an $sn \times sm$ matrix and B an $sm \times sk$ matrix, which are both partitioned into blocks $A_{ij}, B_{ij} \in L(\mathbb{K}^s)$,

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nm} \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & \cdots & B_{1k} \\ \vdots & & \vdots \\ B_{m1} & \cdots & B_{mk} \end{pmatrix}.$$

The blocks A_{ij} are called the block-entries of A. We shall also write $[A]_{ij}$ to denote these blocks. The block-entries $C_{ij} \in L(\mathbb{K}^s)$ of $C = AB \in L(\mathbb{K}^{sk}, \mathbb{K}^{sn})$ are given by

$$C_{ij} = \sum_{\ell=1}^{m} A_{i\ell} B_{\ell j} \qquad (1\le i\le n,\ 1\le j\le k,)$$

(see e.g. GANTMACHER (1959)). If $D_1, D_2, \ldots, D_m \in L(\mathbb{K}^s)$, then $D = \mathrm{diag}(D_1, D_2, \ldots, D_m) \in L(\mathbb{K}^{sm})$ stands for the block-diagonal matrix with blocks $D_1, D_2, \ldots, D_m$ on the diagonal.

The Kronecker product (cf. MARCUS and MINC (1964)) of two matrices A and B will be denoted by $A \otimes B$.

Let $g: \mathbb{K}^s \to \mathbb{K}^s$ (or $L(\mathbb{K}^s)$) be a given function. The Gateaux-derivative of g will be denoted by g' (see section 2.3.1 for more details). We shall also write $D_x g(x)$ instead of $g'(x)$ (for $x \in \mathbb{K}^s$).

If $\psi: \mathbb{C} \to \mathbb{C}$ is such that $\lim_{|\zeta|\to\infty} \psi(\zeta)$ exists, we denote this limit by $\psi(\infty)$.

Finally, if $m,n \in \mathbb{N}$ and $m > n$, we use the conventions

$$\sum_{i=m}^{n} \ldots = 0 \quad \text{and} \quad \prod_{i=m}^{n} \ldots = 1 .$$

2.2. RATIONAL FUNCTIONS

2.2.1. Half-plane bounds for rational functions.

Consider a rational function ψ defined by

$$\psi(\zeta) = [q(\zeta)]^{-1}\, p(\zeta) \qquad (\zeta \in \mathbb{C}) , \tag{2.2.1}$$

where $q(\zeta) = q_0 + q_1\zeta + \ldots + q_k\zeta^k$, $p(\zeta) = p_0 + p_1\zeta + \ldots + p_k\zeta^k$, $k \in \mathbb{N}$, and all coefficients q_j, p_j are in $\mathbb{K}$. We assume that $q_k \neq 0$, and that $\sigma > 0$ is a given number such that ψ is analytic on $\{\zeta:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta<\sigma\}$. Further we put $\theta = \sigma^{-1}$.

We define the function $\Psi\colon \mathbb{R} \to \bar{\mathbb{R}}$ by

$$\Psi(t) = \sup\{|\psi(\zeta)|:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta\leq t,\psi \text{ is regular in } \zeta\} \quad (t \in \mathbb{R}) . \tag{2.2.2}$$

From the maximum modulus theorem it follows that $\Psi(t)$ equals $\sup\{|\psi(\zeta)|:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta=t\}$ whenever $t < \sigma$.

LEMMA 2.2.1. *Assume* $p_k = 0$. *Then there is a constant* $\omega > 0$ *such that*

$$\Psi(t) \leq \omega\ (1-\theta t)^{-1} \qquad (\textit{for}\ -\infty < t \leq \tfrac{1}{2}\sigma).$$

PROOF. Let ℓ be the largest index such that $p_\ell \neq 0$. Then

$$\psi(\zeta) = [q_k\zeta^k(1+O(\zeta^{-1}))]^{-1}\ [p_\ell\zeta^\ell(1+O(\zeta^{-1}))] \qquad (\zeta\to\infty)$$

Hence there exists a $\rho > 0$ such that

$$|\psi(\zeta)| \leq 2|p_\ell/q_k|\ |\zeta|^{\ell-k} \qquad (\text{for}\ |\zeta| > \rho).$$

Since $\ell - k \leq -1$, there is an $\omega_1 > 0$ such that

$$|\psi(\zeta)| \leq \omega_1(1+\theta|\zeta|)^{-1} \qquad (\text{for}\ |\zeta| > \rho). \tag{2.2.3}$$

If $\mathrm{Re}\ \zeta \leq t < -\rho$, then clearly $|\zeta| \geq -t > \rho$. Hence we obtain from (2.2.3)

$$|\psi(\zeta)| \le \omega_1 \, (1-\theta t)^{-1} \qquad (\text{for } \operatorname{Re}\zeta \le t < -\rho) .$$

Now assume that $\operatorname{Re}\zeta \le t$ and $-\rho \le t \le \frac{1}{2}\sigma$. Then

$$|\psi(\zeta)| \le \Psi(\tfrac{1}{2}\sigma) \le \Psi(\tfrac{1}{2}\sigma)\,(1+\theta\rho)\,(1-\theta t)^{-1} .$$

The lemma thus holds with $\omega = \max\{\omega_1, \Psi(\frac{1}{2}\sigma)(1+\theta\rho)\}$. □

LEMMA 2.2.2. *Assume that ψ is not constant and $\Psi(0) = 1$. Suppose further that $\zeta_0 \in \mathbb{C}$ is such that $\operatorname{Re}\zeta_0 = 0$ and $|\psi(\zeta_0)| = 1$. Then $\psi'(\zeta_0)/\psi(\zeta_0)$ is real and positive.*

PROOF. Consider $\zeta = \zeta_0 + \rho\, e^{i\tau}$ with $\rho > 0$ and $\frac{\pi}{2} \le \tau \le \frac{3\pi}{2}$.

Suppose $\psi'(\zeta_0) = 0$. Then there is a $j \ge 2$ such that

$$\psi(\zeta) = \psi(\zeta_0) + \frac{1}{j!}\,\rho^j e^{ij\tau}\,\psi^{(j)}(\zeta_0) + O(\rho^{j+1}) \quad (\rho \to 0)$$

with $\psi^{(j)}(\zeta_0) \ne 0$. We easily see that $|\psi(\zeta)| > 1$ for some $\tau \in [\frac{\pi}{2}, \frac{3\pi}{2}]$ and $\rho > 0$ sufficiently small. This contradicts the assumption of the lemma.

We thus have

$$\psi(\zeta) = \psi(\zeta_0) + \rho\, e^{i\tau}\,\psi'(\zeta_0) + O(\rho^2) \qquad (\rho \to 0)$$

with $\psi'(\zeta_0) \ne 0$. Hence

$$|\psi(\zeta)| = 1 + \rho\, \operatorname{Re}[\overline{\psi(\zeta_0)}\psi'(\zeta_0)e^{i\tau}] + O(\rho^2) \quad (\rho \to 0) .$$

Since $|\psi(\zeta)| \le 1$ for all $\tau \in [\frac{\pi}{2}, \frac{3\pi}{2}]$, we obtain

$$\operatorname{Re}[\overline{\psi(\zeta_0)}\psi'(\zeta_0)e^{i\tau}] \le 0 \qquad (\text{for all } \tau \in [\tfrac{\pi}{2}, \tfrac{3\pi}{2}]).$$

Thus

$\overline{\psi(\zeta_0)}\,\psi'(\zeta_0)$ is real and positive. □

LEMMA 2.2.3. *Assume that ψ is not constant and $\Psi(0) \le 1$. Then, for any $\zeta_0 \in \mathbb{C}$ with $\operatorname{Re}\zeta_0 = 0$, there are $\rho_0, \lambda_0 > 0$ such that*

$$|\psi(\zeta)| \le 1 + \lambda_0\xi \quad (\textit{for all } \zeta = \xi + i\eta \textit{ with } \xi \le 0\ ,\ |\zeta-\zeta_0| \le \rho_0\ ).$$

PROOF. If $|\psi(\zeta_0)| < 1$ the assertion is trivial.

Assume $|\psi(\zeta_0)| = 1$. We then have for $\zeta = \xi + i\eta$ with $\xi \le 0$,

$$|\psi(\zeta)| = |\psi(i\eta)+\xi \int_0^1 \psi'(i\eta+\tau\xi)d\tau\,| \le$$

$$\le |1+[\psi(i\eta)]^{-1}\xi \int_0^1 \psi'(i\eta+\tau\xi)d\tau|\ ,$$

provided that $|\zeta-\zeta_0| \le \rho_1$. Here ρ_1 is taken small enough to ensure that $\psi(i\eta) \neq 0$. In view of the previous lemma we thus have

$$(2.2.4) \qquad |\psi(\zeta)| \le |1+\lambda_1\xi+w(\zeta)\xi| \qquad (\text{for } \zeta = \xi + i\eta\ ,\ \xi \le 0\ ,\ |\zeta-\zeta_0| \le \rho_1)$$

where λ_1 is real and positive, and w is a function with $\lim_{\zeta\to\zeta_0} w(\zeta) = 0$.

From (2.2.4) it follows that, for any $\lambda_0 \in (0,\lambda_1)$, we can select a $\rho_0 > 0$ such that

$$|\psi(\zeta)| \le 1 + \lambda_0\xi \quad (\text{for } \zeta = \xi + i\eta\ ,\ \xi \le 0\ ,\ |\zeta-\zeta_0| \le \rho_0)\ . \quad \square$$

THEOREM 2.2.4. *Assume* $\Psi(t) < 1$ *for all* $t > 0$. *Then there are constants* $\lambda > 0$ *and* $t^* < 0$ *such that*

$$\Psi(t) \le \tilde{\Psi}(t) \quad (\textit{for} \quad t \le 0),$$

where the function $\tilde{\Psi}$ *is defined on* $(-\infty,0]$ *by*

$$\tilde{\Psi}(t) = 1 + \lambda t \quad \textit{if} \quad t^* < t \le 0\ ,$$

$$\tilde{\Psi}(t) = 1 + \lambda t^* \quad \textit{if} \quad t \le t^*\ .$$

PROOF. From the assumption it follows that there exist numbers $R > 0$ and $c \in (0,1)$ such that $|\psi(\zeta)| \le c$ whenever $|\zeta| \ge R$.

We know, from lemma 2.2.3, that for all $\eta_0 \in \mathbb{R}$ there are positive numbers $\rho(\eta_0)$ and $\lambda(\eta_0)$ such that $|\psi(\zeta)| \le 1 + \lambda(\eta_0)\xi$ (for $\zeta = \xi + i\eta$, $\xi \le 0$, $|\zeta-i\eta_0| \le \rho(\eta_0)$). For $\eta_0 \in \mathbb{R}$ we denote the sphere $\{\zeta:\zeta\in\mathbb{C},|\zeta-i\eta_0| \le \rho(\eta_0)\}$ by $B(\eta_0)$. Since the set $J = \{i\eta:\eta\in\mathbb{R},|\eta| \le R\}$ is compact, there is a finite covering $B(\eta_1),B(\eta_2),\ldots,B(\eta_n)$ of J . It

can be seen, using simple geometrical arguments, that there is an $r > 0$ such that $\zeta = \xi + i\eta$ belongs to some $B(\eta_j)$ if $|\xi| \le r$, $|\zeta| \le R$. We take $\lambda = \min_{1\le j\le n} \lambda(\eta_j)$. Then

$$|\psi(\zeta)| \le 1 + \lambda\xi \qquad (\text{for } \zeta = \xi + i\eta,\ -r \le \xi \le 0,\ |\zeta| \le R).$$

Put $\tilde{\Psi}(t) = \max\{1+\lambda t, \Psi(-r), c\}$ $(t \le 0)$. It follows that $\Psi(t) \le \tilde{\Psi}(t)$ for all $t \le 0$. Furthermore $\tilde{\Psi}$ has the desired form. □

REMARK 2.2.5. For rational functions ψ satisfying $\psi(\zeta) = e^{\zeta} + O(\zeta^2)$ $(\zeta \to 0)$, a result similar to theorem 2.2.4 can be found in CROUZEIX and RAVIART (1980).

If we assume that ψ is such that $\Psi(t) < 1$ (for all $t < 0$), $\psi(\zeta) = e^{\zeta} + O(\zeta^2)$ $(\zeta \to 0)$, and $|\psi(i\eta)| < 1$ (for all $\eta \in \mathbb{R}$ with $\eta \neq 0$), it has been shown by HAIRER, BADER and LUBICH (1982) that $\Psi(t) = e^t + O(t^2)$ $(t \to 0)$, and thus $\Psi(t) = 1 + t(1+O(t))$ $(t \to 0)$.

2.2.2. Rational functions with operator coefficients.

Let ψ, q and p be as in section 2.2.1 (see (2.2.1)), and let $s \in \mathbb{N}$. On the space $\mathbb{K}^s$ we consider a norm $|\cdot|$ generated by an inner product $\langle\cdot,\cdot\rangle$.

Let $\Lambda \in L(\mathbb{K}^s)$. We shall say that $\psi(\Lambda)$ *exists* whenever $q(\Lambda) = q_0 I + q_1\Lambda + \ldots + q_k\Lambda^k$ is regular. For such Λ we define $\psi(\Lambda)$ by

$$(2.2.5) \qquad \psi(\Lambda) = (q_0 I + q_1\Lambda + \ldots + q_k\Lambda^k)^{-1}\,(p_0 I + p_1\Lambda + \ldots + p_k\Lambda^k).$$

It is well-known (see e.g. GANTMACHER (1959)) that the spectrum $\sigma(q(\Lambda))$ of $q(\Lambda)$ is given by

$$\sigma(q(\Lambda)) = \{q(\lambda) : \lambda \in \sigma(\Lambda)\}.$$

We thus see that $q(\Lambda)$ is regular iff $\sigma(\Lambda)$ does not contain a zero of q. This leads to the following lemma.

LEMMA 2.2.6. *Let* $\tau \in \mathbb{R}$. *The matrix* $\psi(\Lambda)$ *exists for any* $\Lambda \in L(\mathbb{K}^s)$ *with* $s \ge 1$ *and* $\mu[\Lambda] \le \tau$ *iff* q *has no zeros in* $\{\zeta : \zeta \in \mathbb{C}, \mathrm{Re}\ \zeta \le \tau\}$.

PROOF. If $\mu[\Lambda] \le \tau$, it follows from (2.1.2) that $\mathrm{Re}\ \lambda \le \tau$ for all

$\lambda \in \sigma(\Lambda)$. Thus if $q(\zeta) \neq 0$ whenever $\mathrm{Re}\ \zeta \leq \tau$, and $\mu[\Lambda] \leq \tau$, then $q(\Lambda)$ is invertible.

On the other hand, suppose $q(\lambda_0) = 0$ with $\lambda_0 = \xi_0 + i\eta_0 \in \mathbb{C}$, $\xi_0 \leq \tau$. If $\mathbb{K} = \mathbb{C}$ we take $s = 1$ and $\Lambda = \lambda_0$. If $\mathbb{K} = \mathbb{R}$ we take $s = 2$. Let $d_1, d_2 \in \mathbb{R}^2$ be orthonormal w.r.t. the inner product $\langle \cdot, \cdot \rangle$. We define $\Lambda \in L(\mathbb{R}^2)$ by

$$\Lambda d_1 = \xi_0 d_1 - \eta_0 d_2 \ , \ \Lambda d_2 = \eta_0 d_1 + \xi_0 d_2 \ .$$

By some calculations it follows that $\mu[\Lambda] = \xi_0$, and

$$q(\Lambda)d_1 = \mathrm{Re}(q(\lambda_0))d_1 - \mathrm{Im}(q(\lambda_0))d_2 \ ,$$

$$q(\Lambda)d_2 = \mathrm{Im}(q(\lambda_0))d_1 + \mathrm{Re}(q(\lambda_0))d_2 \ .$$

Hence $q(\Lambda) = 0$. □

The following theorem is based on a result of J. VON NEUMANN (1951), who proved that if $|\Lambda| \leq 1$, then $|\psi(\Lambda)| \leq \sup\{|\psi(\zeta)| : \zeta \in \mathbb{C}, |\zeta| \leq 1\}$. In fact this result holds on arbitrary Hilbert spaces. Using the Cayley transformation it can be seen that this result is equivalent to the following theorem 2.2.7 (see e.g. CROUZEIX and RAVIART (1978)). A more direct proof of this theorem can be found in HAIRER, BADER and LUBICH (1982). In the subsequence we will use the function Ψ defined by (2.2.2).

THEOREM 2.2.7. *Let* $\tau \in \mathbb{R}$. *Suppose* $\Lambda \in L(\mathbb{K}^s)$ *is such that* $\mu[\Lambda] \leq \tau$ *and* $\psi(\Lambda)$ *exists. Then*

$$|\psi(\Lambda)| \leq \Psi(\tau) \ .$$

From theorem 2.2.7 we obtain the following corollaries 2.2.8-2.2.10. In these corollaries the existence of $\psi(\Lambda)$ is always guaranteed by lemma 2.2.6.

COROLLARY 2.2.8. *Let* $\sigma > 0$. *Suppose* $q_k \neq 0$ *and* ψ *is analytic on* $\{\zeta : \zeta \in \mathbb{C}, \mathrm{Re}\ \zeta < \sigma\}$. *Then there is a constant* $\omega > 0$ *such that*

$$|\psi(\Lambda)| \leq \omega$$

whenever $\Lambda \in L(\mathbb{K}^s)$ *with* $\mu[\Lambda] \le \tau \le \frac{1}{2}\sigma$.

PROOF. Since ψ is analytic on $\{\zeta:\zeta\in\mathbb{C},\text{Re }\zeta\le\frac{1}{2}\sigma\}$ and $q_k \neq 0$, $\Psi(\frac{1}{2}\sigma)$ is finite. We take $\omega = \Psi(\frac{1}{2}\sigma)$. The proof now follows from theorem 2.2.7. □

COROLLARY 2.2.9. *Let* $\sigma > 0$ *and* $\theta = \sigma^{-1}$. *Suppose* $p_k = 0$, $q_k \neq 0$, *and* ψ *is analytic on* $\{\zeta:\zeta\in\mathbb{C},\text{Re }\zeta<\sigma\}$. *Then there is a constant* $\omega > 0$ *such that*

$$|\psi(\zeta)| \le \omega\,(1-\theta\tau)^{-1}$$

whenever $\Lambda \in L(\mathbb{K}^s)$ *with* $\mu[\Lambda] \le \tau \le \frac{1}{2}\sigma$.

The proof of this corollary follows directly from lemma 2.2.1 and theorem 2.2.7.

By combining the theorems 2.2.7 and 2.2.4 we arrive at the following.

COROLLARY 2.2.10. *Assume* ψ *is analytic on* $\mathbb{C}^-$ *and* $\Psi(t) < 1$ *for all* $t > 0$. *Then there are constants* $\lambda > 0$ *and* $t^* < 0$ *such that*

$$|\psi(\Lambda)| \le \tilde{\Psi}(\tau)$$

whenever $\Lambda \in L(\mathbb{K}^s)$ *with* $\mu[\Lambda] \le \tau \le 0$. *Here* $\tilde{\Psi}$ *is defined as in theorem* 2.2.4.

As a particular application of theorem 2.2.7 and lemma 2.2.6 we consider $\psi(\zeta) = (1-\theta\zeta)^{-1}$ (for $\zeta \in \mathbb{C}$), where $\theta > 0$. It is easily seen that ψ is analytic on $\{\zeta:\zeta\in\mathbb{C},\text{Re }\zeta<\theta^{-1}\}$ and $\Psi(t) = (1-\theta t)^{-1}$ (for $t < \theta^{-1}$). Thus we obtain

COROLLARY 2.2.11. *Let* $\Lambda \in L(\mathbb{K}^s)$ *with* $\mu[\Lambda] \le \tau < \theta^{-1}$. *Then* $I - \theta\Lambda$ *is invertible, and*

$$|(I-\theta\Lambda)^{-1}| \le (1-\theta\tau)^{-1} .$$

2.3. DIFFERENTIATION

2.3.1. Basic properties.

In the following the differentiation concept of Gateaux will be used. For the definition of the Gateaux-derivative and its basic properties we refer to ORTEGA and RHEINBOLDT (1970) and MARTIN (1976).

Let X and Y be finite dimensional vector spaces. The results of this section will be used later on with $X = \mathbb{K}^s$, and $Y = \mathbb{K}^s$ or $L(\mathbb{K}^s)$. On X and Y we consider norms which will both be denoted by $|\cdot|$.

Consider a function $g: X \to Y$. The Gateaux-derivative of g will be denoted by g'. For any $x \in X$, $g'(x) \in L(X,Y)$ and $g''(x) = (g')'(x) \in \in L(X,L(X,Y))$. The norms $|g'(x)|$ and $|g''(x)|$ are given by the operator norms on $L(X,Y)$ and $L(X,L(X,Y))$. We shall also write $D_x g(x)$ for $g'(x)$.

Let $x \in X$. If g is Gateaux-differentiable on an open neighbourhood of x, and g' is continuous at x, we simply call g *continuously differentiable* at x. It is well known that g is continuously differentiable at x iff all partial derivatives of g exist on a neighbourhood of x and are continuous at x. If $\mathcal{D} \subset \mathbb{K}^s$ is open, and g is continuously differentiable at each point of $\mathcal{D}$, we call g continuously differentiable on $\mathcal{D}$.

The following version of the mean-value theorem will be used frequently in subsequent chapters.

THEOREM 2.3.1. *Let $\mathcal{D} \subset X$ be open and convex. Suppose g is continuously differentiable on $\mathcal{D}$, and $\tilde{x}, x \in \mathcal{D}$. Then*

$$g(\tilde{x}) - g(x) = \int_0^1 g'(x+t(\tilde{x}-x))\ (\tilde{x}-x)\ dt\ .$$

The proof of this theorem can be found in ORTEGA and RHEINBOLDT (1970).

From theorem 2.3.1 the following two lemmata 2.3.2 and 2.3.3 can be derived.

LEMMA 2.3.2. *Let $\mathcal{D} \subset X$ be open and convex, and let $\sigma > 0$ be given. Suppose g is continuously differentiable on $\mathcal{D}$. Then $|g(\tilde{x})-g(x)| \leq$ $\leq \sigma|\tilde{x}-x|$ (for all $\tilde{x}, x \in \mathcal{D}$) iff $|g'(x)| \leq \sigma$ (for all $x \in \mathcal{D}$).*

For the next lemma we consider a function $f: X \to X$, and we assume

that the norm $|\cdot|$ on X is generated by an inner product $\langle\cdot,\cdot\rangle$. We then have the following analogue of lemma 2.3.2.

LEMMA 2.3.3. *Let* $\mathcal{D} \subset X$ *be open and convex, and let* $\beta \in \mathbb{R}$. *Suppose* f *is continuously differentiable on* $\mathcal{D}$. *Then* $\operatorname{Re}\langle f(\tilde{x})-f(x), \tilde{x}-x\rangle \leq \beta|\tilde{x}-x|^2$ *(for all* $\tilde{x},x \in \mathcal{D}$*) iff* $\mu[f'(x)] \leq \beta$ *(for all* $x \in \mathcal{D}$*).*

We now derive some simple rules for the differentiation of products and inverses, which will be used in the next section. Relations similar to (2.3.1) and (2.3.3) can be found in DEN HEIJER (1979, lemma 4.2.2).

Let $\mathcal{D} \subset \mathbb{K}^s$ be an open set, and $k \in \mathbb{N}$. We consider given functions

$$\Lambda_1,\Lambda_2,\Lambda_3\colon X \to L(X) \quad \text{and} \quad f\colon X \to X .$$

It is assumed that Λ_3 is invertible for all $x \in \mathcal{D}$. The functions w,U,V,W are defined on $\mathcal{D}$ by

$$w(x) = \Lambda_1(x)f(x) \ , \ U(x) = \Lambda_1(x)\Lambda_2(x) \ ,$$

$$V(x) = \Lambda_3(x)^{-1} \ , \ W(x) = \Lambda_1(x)^k$$

(for $x \in \mathcal{D}$).

LEMMA 2.3.4. *Suppose* Λ_i $(i=1,2,3)$ *and* f *are Gateaux-differentiable on* $\mathcal{D}$. *Then also* w,U,V *and* W *are Gateaux-differentiable on* $\mathcal{D}$. *For arbitrary* $x \in \mathcal{D}$ *and* $v \in X$ *we have*

$$(2.3.1) \qquad w'(x)v = [\Lambda_1'(x)v]\, f(x) + \Lambda_1(x)\, [f'(x)v] \ ,$$

$$(2.3.2) \qquad U'(x)v = [\Lambda_1'(x)v]\, \Lambda_2(x) + \Lambda_1(x)\, [\Lambda_2'(x)v] \ ,$$

$$(2.3.3) \qquad V'(x)v = -\Lambda_3(x)^{-1}\, [\Lambda_3'(x)v]\, \Lambda_3(x)^{-1} \ ,$$

$$(2.3.4) \qquad W'(x)v = \sum_{\ell=1}^{k} \Lambda_1(x)^{\ell-1}\, [\Lambda_1'(x)v]\, \Lambda_1(x)^{k-\ell} \ .$$

PROOF. For $t > 0$ we have

$$w(x+tv) - w(x) = [\Lambda_1(x+tv)-\Lambda_1(x)]\, f(x) +$$

$$+ \Lambda_1(x)\, [f(x+tv)-f(x)] + [\Lambda_1(x+tv)-\Lambda_1(x)]\, [f(x+tv)-f(x)] \ .$$

For $t \neq 0$, $t^{-1}[\Lambda_1(x+tv)-\Lambda_1(x)]$ converges to $\Lambda_1'(x)v$, and similarly for f. We thus see that w is Gateaux-differentiable at x, and (2.3.1) holds.

In a similar way it can be shown that U and V are Gateaux-differentiable on $\mathcal{D}$, and that (2.3.2) and (2.3.3) hold. From (2.3.1) it can be shown, by induction with respect to k, that W is Gateaux-differentiable on $\mathcal{D}$, and that W' is given by (2.3.4). □

2.3.2. Differentiation of rational expressions of operator valued functions.

Let $p(\zeta) = p_0 + p_1\zeta + \ldots + p_k\zeta^k$, $q(\zeta) = q_0 + q_1(\zeta) + \ldots + q_k\zeta^k$ ($\zeta \in \mathbb{C}$) where $k \in \mathbb{N}$, $p_j, q_j \in \mathbb{K}$ ($1 \leq j \leq k$) and $q_k \neq 0$. Let $\mathcal{D}$ be an open set in $\mathbb{K}^s$. Assume that $\Lambda: \mathbb{K}^s \to L(\mathbb{K}^s)$ is continuously differentiable on $\mathcal{D}$, and $q(\Lambda(x))$ is invertible for all $x \in \mathcal{D}$. We consider the function $R: \mathbb{K}^s \to L(\mathbb{K}^s)$ defined by

$$(2.3.5) \qquad R(x) = [q(\Lambda(x))]^{-1}\, p(\Lambda(x)) \qquad (x \in \mathbb{K}^s).$$

LEMMA 2.3.5. *R is continuously differentiable on $\mathcal{D}$, and for all $x \in \mathcal{D}$, $v \in \mathbb{K}^s$ we have*

$$(2.3.6) \qquad R'(x)v = \sum_{j=1}^{k} \sum_{\ell=1}^{j} [q(\Lambda(x))^{-1}\Lambda(x)^{\ell-1}][\Lambda'(x)v]$$

$$[p_j\Lambda(x)^{j-\ell} - q_j\Lambda(x)^{j-\ell}R(x)] .$$

PROOF. We define $P,Q: \mathbb{K}^s \to L(\mathbb{K}^s)$ by $P(x) = p(\Lambda(x))$, $Q(x) = q(\Lambda(x))$ ($x \in \mathbb{K}^s$). From lemma 2.3.4 it follows that P,Q and R are Gateaux differentiable on $\mathcal{D}$, and that

$$R'(x)v = -Q(x)^{-1}\,[Q'(x)v]\,Q(x)^{-1}P(x) + Q(x)^{-1}\,[P'(x)v] .$$

In view of (2.3.4) we obtain

$$Q'(x)v = \sum_{j=1}^{k} q_j \sum_{\ell=1}^{j} \Lambda(x)^{\ell-1}\,[\Lambda'(x)v]\,\Lambda(x)^{j-\ell} ,$$

and a similar expression for $P'(x)v$. This leads to (2.3.6), from which we see that R' is continuous on $\mathcal{D}$. □

By using the fact that

$$\sum_{j=1}^{k}\sum_{\ell=1}^{j}\ \ldots\ =\ \sum_{\ell=1}^{k}\sum_{j=\ell}^{k}\ \ldots\ ,$$

we can rewrite (2.3.6) as

$$R'(x)v = \sum_{\ell=1}^{k} [q(\Lambda(x))^{-1}\Lambda(x)^{\ell-1}]\,[\Lambda'(x)v]\,[\psi_\ell(\Lambda(x))] \tag{2.3.7}$$

where the rational functions ψ_ℓ $(1\le\ell\le k)$ are defined by

$$\psi_\ell(\zeta) = q(\zeta)^{-1}\sum_{j=\ell}^{k}[p_j\zeta^{j-\ell}q(\zeta)-q_j\zeta^{j-\ell}p(\zeta)] \quad (\zeta\in\mathbb{C})\ .$$

We take a closer look at the degree of the numerator of the ψ_ℓ. At first sight this degree seems to be larger than k, which would imply $\lim_{\zeta\to\infty}\psi_\ell(\zeta) = \infty$. This is not so.

For arbitrary $\zeta\in\mathbb{C}$ the $\psi_\ell(\zeta)$ satisfy the recurrence relation

$$\psi_\ell(\zeta) = p_\ell - q_\ell q(\zeta)^{-1}p(\zeta) + \zeta\psi_{\ell+1}(\zeta) \qquad (1\le\ell\le k-1)\ .$$

Further we have

$$\psi_1(\zeta) = q(\zeta)^{-1}\sum_{j=1}^{k}[p_j\zeta^{j-1}q(\zeta)-q_j\zeta^{j-1}p(\zeta)] =$$

$$= q(\zeta)^{-1}\sum_{j=1}^{k}\sum_{i=0}^{k}(p_jq_i-q_jp_i)\,\zeta^{i+j-1} =$$

$$= q(\zeta)^{-1}\sum_{j=1}^{k}(p_jq_0-q_jp_0)\,\zeta^{j-1}\ .$$

Thus we see that

$$\psi_1(\zeta) = O(\zeta^{-1}) \quad (\zeta\to\infty)\ ,$$

and in view of the recurrence relation we get

$$\psi_\ell(\zeta) = O(\zeta^{-1}) \quad (\zeta \to \infty) \qquad (1 \le \ell \le k) \quad .$$

Since we also have

$$q(\zeta)^{-1}\, \zeta^{\ell-1} = O(\zeta^{-1}) \quad (\zeta \to \infty) \qquad (1 \le \ell \le k) \ ,$$

(2.3.7) leads to the following result.

THEOREM 2.3.6. *Let* R *be defined by* (2.3.5). R *is continuously differentiable on* $\mathcal{D}$ *, and for all* $x \in \mathcal{D}$, $v \in \mathbb{K}^s$ *we have*

$$R'(x)v = \sum_{\ell=1}^{k} \phi_\ell(\Lambda(x))\ [\Lambda'(x)v]\ \psi_\ell(\Lambda(x)) \ .$$

Here ϕ_ℓ, ψ_ℓ *are rational functions, which can be written with denominator* q , *and which satisfy* $\phi_\ell(\infty) = \psi_\ell(\infty) = 0$ $(1 \le \ell \le k)$.

2.4. SOME MATRIX RESULTS

2.4.1. Notation.

In this section we will use the following notation. Let $s \in \mathbb{N}$. On the space $\mathbb{K}^s$ we consider the inner product $\langle \cdot, \cdot \rangle$ and the related norm $|x| = \langle x,x \rangle^{\frac{1}{2}}$ (for $x \in \mathbb{K}^s$). For arbitrary $m \in \mathbb{N}$ we denote by $(\cdot,\cdot)$ the Euclidean inner product on $\mathbb{R}^m$. For a given positive definite diagonal matrix $D = \mathrm{diag}(d_1,\ldots,d_m)$ we define the inner product $[\cdot,\cdot]_D$ and norm $\|\cdot\|_D$ on $\mathbb{K}^{sm}$ by

$$[x,y]_D = \sum_{i=1}^{m} d_i \ \langle x_i, y_i \rangle \ , \quad \|x\|_D = [x,x]_D^{\frac{1}{2}}$$

for $x,y \in \mathbb{K}^{sm}$, where $x = (x_1^T, x_2^T, \ldots, x_m^T)^T$, $y = (y_1^T, y_2^T, \ldots, y_m^T)^T$, and all x_i, y_i are column-vectors in the $\mathbb{K}^s$.

Further we use in this section the following convention. Let $m \in \mathbb{N}$, $A = (a_{ij}) \in L(\mathbb{R}^m)$, $b = (b_i) \in \mathbb{R}^m$, and let $I^{(s)}$ be the $s \times s$ unit-matrix. Then the Kronecker products $A \otimes I^{(s)}$, $b \otimes I^{(s)}$ will also be denoted by A,b, respectively. Thus for $v = (v_1^T, v_2^T, \ldots, v_m^T)^T$ with all $v_i \in \mathbb{K}^s$ (or $L(\mathbb{K}^s)$), $w = Av$ is given by $w = (w_1^T, w_2^T, \ldots, w_m^T)^T$ with

$$w_i = \sum_{j=1}^{m} a_{ij} v_j \in \mathbb{K}^s \quad (\text{or} \quad L(\mathbb{K}^s)) \qquad (1 \le i \le m) .$$

2.4.2. Matrix results for implicit Runge-Kutta methods.

In this section we shall prove some results which will be of much use in the study of stability properties of implicit Runge-Kutta methods.

Let $m \in \mathbb{N}$ and $\sigma \in \mathbb{R}$. We consider the following condition on an arbitrary matrix $A \in L(\mathbb{R}^m)$.

(2.4.1) There is a matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_m)$ with $d_i > 0$ $(1 \le i \le m)$ such that

$$(v, DAv) \ge \sigma (Av, DAv) \qquad (\text{for all } v \in \mathbb{R}^m).$$

The class of regular matrices $A \in L(\mathbb{R}^m)$ satisfying (2.4.1) will be denoted by $A_m(\sigma)$.

Let $A \in L(\mathbb{R}^m)$. We note that $DA + A^T D$ is positive definite for some positive definite diagonal matrix $D \in L(\mathbb{R}^m)$ iff there is a $\sigma > 0$ such that $A \in A_m(\sigma)$.

The following lemma is a slight modification of a result of DAHLQUIST (1975, lemma 2.2). This lemma can be proved in the same way as Dahlquist's lemma.

LEMMA 2.4.1. *Let* $\sigma \in \mathbb{R}$. *Suppose* $A \in L(\mathbb{R}^m)$ *satisfies* (2.4.1). *Then*

$$\mathrm{Re}\ [w, Aw]_D \ge \sigma \| Aw \|_D^2 \qquad (\textit{for all } w \in \mathbb{K}^{sm}) .$$

The next lemma can be obtained as a corollary to theorem 4.3.1. Here we give a short direct proof.

LEMMA 2.4.2. *Let* $\sigma \in \mathbb{R}$, $A \in A_m(\sigma)$, *and let* $Z = \mathrm{diag}(z_1, z_2, \ldots, z_m)$ *with* $z_i \in L(\mathbb{K}^s)$, $\mu[z_i] \le \tau < \sigma$ $(1 \le i \le m)$. *Then* $I - AZ$ *is invertible.*

PROOF. Suppose that $u \in \mathbb{K}^{sm}$ is such that $(I-AZ)u = 0$. Then $A^{-1}u = Zu$, and therefore

$$[u, A^{-1}u]_D = [u, Zu]_D ,$$

where D is the matrix arising in (2.4.1).

Since all z_i satisfy $\mu[z_i] \le \tau$, it follows that $\mathrm{Re}\,[u,Zu]_D \le$ $\le \tau\|u\|_D^2$. Further we see from (2.4.1) and lemma 2.4.1 that $\mathrm{Re}\,[u,A^{-1}u]_D \ge$ $\ge \sigma\|u\|_D^2$. Thus we obtain

$$\sigma\|u\|_D^2 \le \tau\|u\|_D^2 ,$$

which can only hold if $u = 0$. □

LEMMA 2.4.3. *Let* $\sigma \in \mathbb{R}$, $A \in A_m(\sigma)$, *and let* D *be the matrix arising in* (2.4.1). *There is a constant* $\omega > 0$ *such that*

$$\|(I-AZ)^{-1}\|_D \le (\sigma-\tau)^{-1}\,\omega ,$$

$$\|Z(I-AZ)^{-1}\|_D \le \omega + (\sigma-\tau)^{-1}\,\omega^2 ,$$

for all matrices $Z = \mathrm{diag}(z_1,z_2,\ldots,z_m)$ *with* $z_i \in L(\mathbb{K}^s)$, $\mu[z_i] \le \tau < \sigma$ $(1\le i\le m)$.

PROOF. Let $u \in \mathbb{K}^{sm}$ with $u \neq 0$, and let $v = (I-AZ)^{-1}u$, $w = Z(I-AZ)^{-1}u$. Then

$$A^{-1}v - w = A^{-1}u ,$$

and hence

$$\mathrm{Re}\,[v,A^{-1}v]_D - \mathrm{Re}\,[v,w]_D = \mathrm{Re}\,[v,A^{-1}u]_D .$$

We have $\mathrm{Re}\,[v,A^{-1}u]_D \le \|v\|_D\,\|A^{-1}\|_D\,\|u\|_D$. Using the fact that $\mu[z_i] \le \tau$ $(1\le i\le m)$ and $w = Zv$, it easily follows that $\mathrm{Re}\,[v,w]_D \le$ $\le \tau\|v\|_D^2$. Further we see from lemma 2.4.1, that $\mathrm{Re}\,[v,A^{-1}v]_D \ge \sigma\|v\|_D^2$. Thus we obtain

$$\|v\|_D \le (\sigma-\tau)^{-1}\,\|A^{-1}\|_D\,\|u\|_D .$$

This proves the first inequality of the lemma with $\omega = \|A^{-1}\|_D$.

Writing

$$w = A^{-1}(v-u) ,$$

we see that

$$\|w\|_D \le \|A^{-1}\|_D (\|v\|_D+\|u\|_D) \le \|A^{-1}\|_D (\|A^{-1}\|_D(\sigma-\tau)^{-1}+1)\|u\|_D .$$

This yields the second inequality. □

Let $\sigma \in \mathbb{R}$ and $m \in \mathbb{N}$ with $m \ge 2$. In the following we deal with the class $\mathcal{B}_m(\sigma)$ consisting of the matrices $A \in L(\mathbb{R}^m)$ which are such that $e_1^T A = 0$ (the first row of A is zero), and

$$\bar{A} = \begin{pmatrix} a_{22} & \cdots & a_{2m} \\ \vdots & & \vdots \\ a_{m2} & \cdots & a_{mm} \end{pmatrix} \in L(\mathbb{R}^{m-1})$$

belongs to $A_{m-1}(\sigma)$.

LEMMA 2.4.4. *Let* $\sigma,\delta > 0$ *and* $A \in \mathcal{B}_m(\sigma)$. *Then* $I - AZ$ *is regular, and there are constants* $\omega_1,\omega_2 > 0$ *such that the* $s \times s$ *blocks of* $(I-AZ)^{-1}e \in$ $\in L(\mathbb{K}^s,\mathbb{K}^{sm})$ *satisfy*

$$|[(I-AZ)^{-1}e]_i| \le \omega_1 + \omega_2\delta(\sigma-\tau)^{-1} \qquad (1\le i\le m) ,$$

for all $Z = \mathrm{diag}(z_1,z_2,\ldots,z_m)$ *with* $z_i \in L(\mathbb{K}^s)$, $\mu[z_i] \le \tau \le \frac{1}{2}\sigma$, *and* $|z_i-z_1| \le \delta$ $(1\le i\le m)$.

PROOF. Let $\bar{a}_1 = (a_{21},a_{31},\ldots,a_{m1})^T$ and $\bar{e} = (1,1,\ldots,1)^T \in \mathbb{R}^{m-1}$. With the convention introduced in section 2.4.1 we can regard $\bar{a}_1$ and $\bar{e}$ also as operators in $L(\mathbb{K}^s,\mathbb{K}^{s(m-1)})$. Suppose $w = (w_1^T,w_2^T,\ldots w_m^T)^T$, with all $w_i \in L(\mathbb{K}^s)$, is such that $(I-AZ)w = e$. Then $w_1 = I$, and

$$\bar{w} - \bar{A}\bar{Z}\,\bar{w} - \bar{e} - \bar{a}_1 z_1 = 0 .$$

Here $\bar{w} = (w_2^T,w_3^T,\ldots,w_m^T)^T$ and $\bar{Z} = \mathrm{diag}(z_2,z_3,\ldots,z_m)$. Since $\bar{A} \in A_{m-1}(\sigma)$, we know from lemma 2.4.2 that $I - \bar{A}\bar{Z}$ is regular. It follows that

$$\bar{w} = (I-\bar{A}\bar{Z})^{-1}(\bar{e}+\bar{a}_1 z_1) ,$$

and $I - AZ$ is regular.

Let $\bar{Z}_1 \in L(\mathbb{K}^{s(m-1)})$ stand for the block-diagonal matrix with blocks

z_1 on the diagonal. We have

$$(I-\bar{\bar{A}}\bar{\bar{Z}})^{-1} - (I-\bar{\bar{A}}\bar{\bar{Z}}_1)^{-1} = (I-\bar{\bar{A}}\bar{\bar{Z}})^{-1}\ \bar{A}(\bar{Z}-\bar{Z}_1)\ (I-\bar{\bar{A}}\bar{\bar{Z}}_1)^{-1} \ .$$

Hence

$$\bar{w} = (I-\bar{\bar{A}}\bar{\bar{Z}}_1)^{-1}\ (\bar{e}+\bar{a}_1 z_1) + (I-\bar{\bar{A}}\bar{\bar{Z}})^{-1}\ \bar{A}(\bar{Z}-\bar{Z}_1)\ (I-\bar{\bar{A}}\bar{\bar{Z}}_1)^{-1}\ (\bar{e}+\bar{a}_1 z_1) \ . \tag{2.4.2}$$

Let $\bar{D} = \mathrm{diag}(d_1,d_2,\ldots,d_{m-1})$ be a positive definite matrix such that $(v,\bar{\bar{D}}\bar{A}v) \geq \sigma(\bar{A}v,\bar{\bar{D}}\bar{A}v)$ (for all $v \in \mathbb{R}^{m-1}$). From lemma 2.4.3 we know there exists a constant $\omega > 0$ such that $\|(I-\bar{\bar{A}}\bar{\bar{Z}})^{-1}\|_{\bar{D}} \leq (\sigma-\tau)^{-1}\omega$.

Let $\psi_1,\psi_2,\ldots,\psi_{m-1}$ be rational functions such that $(\psi_1(\zeta),\psi_2(\zeta),\ldots,\psi_{m-1}(\zeta))^T = (I-\bar{A}\zeta)^{-1}\ (\bar{e}+\bar{a}_1\zeta)$ $(\zeta\in\mathbb{C})$. It follows from lemma 2.4.2 that all ψ_i are analytic on $\{\zeta:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta<\sigma\}$, and, since $\bar{A}$ is regular, all $\psi_i(\zeta)$ remain bounded when $|\zeta| \to \infty$. By corollary 2.2.8 we therefore know that there exists a constant ω_1 such that $|\psi_i(z_1)| \leq \omega_1$ $(1\leq i\leq m-1)$. Moreover, by using lemma 2.4.6 it can easily be shown that $(I-\bar{\bar{A}}\bar{\bar{Z}}_1)^{-1}\ (\bar{e}+\bar{a}_1 z_1) = (\psi_1(z_1)^T,\psi_2(z_1)^T,\ldots,\psi_{m-1}(z_1)^T)^T$.

The proof of the lemma now follows from (2.4.2). □

The following lemma can be viewed as a corollary of lemma 2.4.6, corollary 2.2.9, and a result proved by CROUZEIX and RAVIART (1980). Since this reference is hardly available, the proof of Crouzeix and Raviart is incorporated in the following.

LEMMA 2.4.5. *Suppose* $A \in B_m(\sigma)$ *with* $\sigma > 0$, *and let* $b^T = e_m^T A$. *Then* $I - AZ_0$ *is regular, and there is a constant* $\omega > 0$ *such that all* $s \times s$ *block-elements of* $b^T(I-AZ_0)^{-1} \in L(\mathbb{K}^{sm},\mathbb{K}^s)$ *satisfy*

$$|[b^T(I-AZ_0)^{-1}]_i| \leq (\sigma-\tau)^{-1}\ \omega \qquad (1\leq i\leq m) \ ,$$

for all $Z_0 = \mathrm{diag}(z_0,z_0,\ldots,z_0) \in L(\mathbb{K}^{sm})$ *with* $z_0 \in L(\mathbb{K}^s)$, $\mu[z_0] \leq$ $\leq \tau \leq \frac{1}{2}\sigma$.

PROOF. The existence of $(I-AZ_0)^{-1}$ follows from lemma 2.4.4.

We introduce the rational functions ϕ_i $(1\leq i\leq m)$ satisfying $b^T(I-A\zeta)^{-1} = (\phi_1(\zeta),\phi_2(\zeta),\ldots,\phi_m(\zeta))$ $(\zeta \in \mathbb{C})$. We then have $b^T(I-AZ_0)^{-1} =$ $= (\phi_1(z_0),\phi_2(z_0),\ldots,\phi_m(z_0))$. This result can easily be proved by using

lemma 2.4.6.

In view of corollary 2.2.9 it is now sufficient to show that all ϕ_i are analytic on $\{\zeta:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta<\sigma\}$, and $\phi_i(\infty) = 0$ $(1\le i\le m)$.

The fact that all ϕ_i are analytic on the given set follows as in the proof of lemma 2.4.4, since $\det(I-A\zeta) = \det(I-\bar{A}\zeta)$ $(\zeta \in \mathbb{C})$.

Further we have

$$b^T(I-A\zeta)^{-1} = \zeta^{-1}\ e_m^T\ A\zeta(I-A\zeta)^{-1} = -\zeta^{-1}\ e_m^T(I-(I-A\zeta)^{-1}) \ .$$

The degree of the polynomial $\det(I-A\zeta)$ equals $m-1$ (not smaller). Thus all entries of $(I-A\zeta)^{-1}$ remain bounded when $|\zeta| \to \infty$, and we see that $\phi_i(\infty) = 0$ $(1\le i\le m)$. □

2.4.3. Miscellaneous results.

In this section some results on matrices are collected which are of a different type than the lemmata of section 2.4.2. We start with a result which has been used already in the proofs of the lemmata 2.4.4 and 2.4.5.

Let $m \in \mathbb{N}$, and let $V: \mathbb{C} \to L(\mathbb{C}^m)$ be a matrix-valued function such that the entries $v_{ij}(\zeta)$ of $V(\zeta)$ are rational functions, with real coefficients, in the variable ζ . For $\Lambda \in L(\mathbb{K}^s)$ $(s \in \mathbb{N})$ we denote by $V(\Lambda)$ the $sm \times sm$ matrix with block-entries $v_{ij}(\Lambda) \in L(\mathbb{K}^s)$. We shall say that $V(\Lambda)$ exists if all $v_{ij}(\Lambda)$ exist (cf. section 2.2.2).

LEMMA 2.4.6. *Let* V *be as above, and let* W *be a matrix-valued function such that* $W(\zeta) = V(\zeta)^{-1}$ *(whenever* $\zeta \in \mathbb{C}$ *and* $V(\zeta)^{-1}$ *is defined). Suppose* $\Lambda \in L(\mathbb{K}^s)$ *and* $V(\Lambda)$ *exists. Then*

$W(\Lambda)$ *exists iff* $V(\Lambda)$ *is regular.*

Further if $V(\Lambda)$ *is regular, we have* $W(\Lambda) = V(\Lambda)^{-1}$.

PROOF. We denote the entries of $W(\zeta)$ by $w_{ij}(\zeta)$ $(1\le i,j\le m\ ,\ \zeta \in \mathbb{C})$.

1. Let $\psi(\zeta) = \det(V(\zeta))$ $(\zeta \in \mathbb{C})$. Suppose $\psi(\lambda) = 0$ for some $\lambda \in \sigma(\Lambda)$. Let $x \in \mathbb{C}^s$ and $u \in \mathbb{C}^m$ be nonzero vectors such that $\Lambda x = \lambda x$ and $V(\lambda)u = 0$. Then we have

$$V(\Lambda)(u\otimes x) = (V(\lambda)u) \otimes x = 0 \ ,$$

and thus $V(\Lambda)$ is singular.

2. Assume $V(\Lambda)$ is regular. From the above we see that $\psi(\cdot)^{-1}$ is analytic on $\sigma(\Lambda)$. It follows that all w_{ij} are analytic on $\sigma(\Lambda)$, and thus $W(\Lambda)$ exists.

3. We now assume that $W(\Lambda)$ exists. It will be shown that $V(\Lambda)W(\Lambda) = I \in L(\mathbb{K}^{sm})$, which yields the proof of the lemma.

The block-entries of $V(\Lambda)W(\Lambda)$ are given by

$$\sum_{n=1}^{m} v_{jn}(\Lambda)\, w_{nk}(\Lambda) \qquad (1\leq j,k\leq m) .$$

Let Γ be a continuously differentiable, simple, closed curve in $\mathbb{C}$ with a positive orientation, such that $\sigma(\Lambda)$ is contained in the interior of Γ, and all v_{jk}, w_{jk} are analytic on and inside Γ. By applying Cauchy's integral formula for matrices (see e.g. DUNFORD and SCHWARTZ (1958)), we obtain

$$\sum_{n=1}^{m} v_{jn}(\Lambda)\, w_{nk}(\Lambda) = \frac{1}{2\pi i} \oint_{\Gamma} (\zeta I-\Lambda)^{-1} \Big[\sum_{n=1}^{m} v_{jn}(\zeta) w_{nk}(\zeta) \Big] d\zeta =$$

$$= \frac{1}{2\pi i} \oint_{\Gamma} (\zeta I-\Lambda)^{-1}\, \delta_{jk}\, d\zeta = I\, \delta_{jk} \qquad (1\leq j,k\leq m) .$$

Here δ_{jk} stands for the Kronecker delta. We thus see that $V(\Lambda)W(\Lambda)$ is the identity operator in $L(\mathbb{K}^{sm})$. □

Suppose $f: \mathbb{K}^s \to \mathbb{K}^s$ is a continuously differentiable function such that $\mathrm{Re}\ \langle f(\tilde{x})-f(x), \tilde{x}-x\rangle \leq 0$ (for all $\tilde{x}, x \in \mathbb{K}^s$). We know, by a combination of lemma 2.3.3 and theorem 2.3.1, that there exists a $\Lambda \in L(\mathbb{K}^s)$ such that $f(\tilde{x}) - f(x) = \Lambda(\tilde{x}-x)$ and $\mu[\Lambda] \leq 0$; namely $\Lambda = \int_0^1 f'(x+t(\tilde{x}-x))dt$. With the following lemma we see that such a Λ also exists if f is not continuously differentiable.

LEMMA 2.4.7. *Suppose* $u,v \in \mathbb{K}^s$ *with* $v \neq 0$ *and* $\mathrm{Re}\ \langle u,v\rangle \leq 0$. *Then there exists a* $\Lambda \in L(\mathbb{K}^s)$ *such that* $\mu[\Lambda] \leq 0$ *and* $u = \Lambda v$.

PROOF. Let $w_0 = v - u$, $w_1 = v + u$. Then it is easily seen that

(2.4.3) $\quad |w_1| \le |w_0|$, and $\langle w_1, w_0 \rangle + |w_0|^2 \neq 0$.

We define $M \in L(\mathbb{K}^s)$ by

$$Mx = |w_0|^{-2} \langle x, w_0 \rangle w_1 \qquad (\text{for } x \in \mathbb{K}^s).$$

Then $Mw_0 = w_1$, and using (2.4.3) it can easily be shown that $|M| \le 1$ and $M + I$ is regular.

Let $\Lambda = (M+I)^{-1}(M-I)$. Since $M(v-u) = v + u$, it follows that $(M-I)v = (M+I)u$, and therefore $\Lambda v = u$. Further we have for arbitrary $x \in \mathbb{K}^s$,

$$\text{Re } \langle x, \Lambda x \rangle = \text{Re } \langle (M+I)y, (M-I)y \rangle = |My|^2 - |y|^2 \le 0 .$$

Here y stands for $(M+I)^{-1}x$. Thus $\mu[\Lambda] \le 0$. □

COROLLARY 2.4.8. *Let* $\beta \in \mathbb{R}$. *Suppose* $u,v \in \mathbb{K}^s$ *with* $v \neq 0$, $\text{Re } \langle u,v \rangle \le \beta |v|^2$. *Then there exists a* $\Lambda \in L(\mathbb{K}^s)$ *such that* $\mu[\Lambda] \le \beta$ *and* $u = \Lambda v$.

PROOF. Let $\tilde{u} = u - \beta v$. Then $\text{Re } \langle \tilde{u}, v \rangle \le 0$. From the above lemma we know there exists a $\tilde{\Lambda} \in L(\mathbb{K}^s)$ with $\mu[\tilde{\Lambda}] \le 0$ and $\tilde{u} = \tilde{\Lambda} v$.

We take $\Lambda = \tilde{\Lambda} + \beta I$. Then $\mu[\Lambda] \le \beta$ and $u = \Lambda v$. □

In the following lemma we consider two matrices $V,W \in L(\mathbb{K}^{sm})$ with block-entries $V_{ij}, W_{ij} \in L(\mathbb{K}^s)$ $(1 \le i,j \le m)$.

LEMMA 2.4.9. *Assume* $V_{ij} = 0$ *(for* $1 \le i \le j \le m$*), and* $W = (I-V)^{-1}$. *Let* $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m > 0$ *be such that* $|V_{ij}| \le \varepsilon_j$ $(1 \le j < i \le m)$. *Then* $W_{ij} = 0$ *(for* $1 \le i < j \le m$*)* , $W_{ij} = I$ *for* $(1 \le i = j \le m)$, *and*

$$|W_{ij}| \le \varepsilon_j \prod_{k=j+1}^{i-1} (1+\varepsilon_k) \qquad (1 \le j < i \le m) .$$

PROOF. In order to find a relation between the block-entries V_{ij} and W_{ij} we consider the following equation with unknown u (and given v),

$$(I-V)\,u = v\ .$$

Here $u = (u_1^T, u_2^T, \ldots, u_m^T)^T$ and $v = (v_1^T, v_2^T, \ldots, v_m^T)^T \in L(\mathbb{K}^s, \mathbb{K}^{sm})$. This equation can be solved recursively. We obtain

$$(2.4.4)\qquad u_1 = v_1\ ,\ u_i = v_i + \sum_{k=1}^{i-1} V_{ik}\,u_k \qquad (2\le i\le m)\ .$$

Suppose $1 \le j \le m$, $v_j = I$, $v_k = 0$ (for $k \ne j$). Then $u_i = W_{ij}$ $(1\le i\le m)$. From (2.4.4) we therefore get

$$(2.4.5)\qquad W_{ij} = 0 \ \text{ if } \ i < j\ ,\ W_{ij} = I \ \text{ if } \ i = j\ ,$$

$$W_{ij} = \sum_{k=j}^{i-1} V_{ik}\,W_{kj} \ \text{ if } \ i > j\ .$$

Let $1 \le j \le m$, $\xi_n = |W_{j+n,j}|$ $(0\le n\le m-j)$. Then $\xi_0 = 1$ and it follows from (2.4.5) that

$$\xi_n \le \sum_{k=0}^{n-1} \varepsilon_{j+k}\,\xi_k \qquad (1\le n\le m-j)\ .$$

We define the numbers $\eta_0, \eta_1, \ldots, \eta_{m-j}$ by

$$\eta_0 = 1\ ,\ \eta_n = \sum_{k=0}^{n-1} \varepsilon_{j+k}\,\eta_k \qquad (1\le n\le m-j)\ .$$

Then $\eta_1 = \varepsilon_j$ and $\eta_n = (1+\varepsilon_{j+n-1})\,\eta_{n-1}$ $(2\le n\le m-j)$. By induction it is easily seen that $\xi_n \le \eta_n$ $(0\le n\le m-j)$, and

$$\eta_n = \varepsilon_j \prod_{k=j+1}^{j+n-1} (1+\varepsilon_k) \qquad (1\le n\le m-j)\ .\qquad \square$$

COROLLARY 2.4.10. *Let the assumptions of lemma* 2.4.9 *hold with* $\varepsilon_j = \varepsilon > 0$ $(1\le j\le m)$. *Then* $|W_{ij}| \le (1+\varepsilon)^{i-j}$ $(1\le j\le i\le m)$.

For the next lemma we consider rational functions a_{ij} and b_i $(1\le i,j\le m)$. We put $A(\zeta) = (a_{ij}(\zeta)) \in L(\mathbb{K}^m)$ and $b(\zeta) = (b_i(\zeta)) \in \mathbb{K}^m$

(for $\zeta \in \mathbb{C}$).

LEMMA 2.4.11. *Assume* $a_{ij} \equiv 0$ *(for* $1 \le i \le j \le m$ *), and all* a_{ij}, b_i *are analytic on* $\mathbb{C}^-$. *Then the statements* (i) *and* (ii) *are equivalent.*

(i) $a_{ij}(\infty) = b_i(\infty) = 0 \qquad (1 \le i,j \le m)$.

(ii) $\sup_{\zeta \in \mathbb{C}^-} |[(I-A(\zeta)\zeta)^{-1}]_{ij}| < \infty \qquad (1 \le i,j \le m)$, and

$\sup_{\zeta \in \mathbb{C}^-} |[b(\zeta)^T \zeta (I-A(\zeta)\zeta)^{-1}]_i| < \infty \qquad (1 \le i \le m)$.

PROOF. Let $\zeta \in \mathbb{C}$, $V(\zeta) = (v_{ij}(\zeta)) = A(\zeta)\zeta$ and $W(\zeta) = (w_{ij}(\zeta)) = (I-A(\zeta)\zeta)^{-1}$. Using the relation (2.4.5) it can be shown inductively that $w_{ij} = 0$ (for $1 \le i < j \le m$), $w_{jj} = 1$ (for $1 \le j \le m$), and

$$w_{ij}(\zeta) = v_{ij}(\zeta) + u_{ij}(\zeta) \qquad (1 \le j < i \le m) ,$$

where $u_{ij}(\zeta)$ is a sum of products of the entries $v_{k\ell}(\zeta)$ with $k \le i$, $\ell \ge j$ and $(k,\ell) \ne (i,j)$.

Therefore if (i) holds, we easily see that (ii) also holds.

Suppose (ii). Let i,j be such that $a_{ij}(\infty) \ne 0$, $a_{k\ell}(\infty) = 0$ for $k \le i$, $\ell \ge j$, $(k,\ell) \ne (i,j)$. This implies that u_{ij} is bounded on $\mathbb{C}^-$, whereas $|v_{ij}(\infty)| = \infty$. Hence w_{ij} is not bounded on $\mathbb{C}^-$, and (ii) is contradicted. We thus have $a_{ij}(\infty) = 0$ $(1 \le i,j \le m)$.

Now let $\psi_1, \psi_2, \ldots, \psi_m$ be rational functions such that $(\psi_1(\zeta), \psi_2(\zeta), \ldots, \psi_m(\zeta)) = b(\zeta)^T \zeta (I-A(\zeta)\zeta)^{-1}$ $(\zeta \in \mathbb{C})$, and let i be such that $b_i(\infty) \ne 0$, $b_k(\infty) = 0$ for $k > i$. Since $w_{k\ell}(\zeta) = 0$ for $k < \ell$, $w_{k\ell}(\zeta) = 1$ for $k = \ell$, we see that $\psi_m, \psi_{m-1}, \ldots, \psi_{i+1}$ are bounded on $\mathbb{C}^-$, but $|\psi_i(\infty)| = \infty$. □

CHAPTER 3

RUNGE-KUTTA METHODS AND GENERALIZATIONS

3.1. INTRODUCTION

Consider the initial value problem

(3.1.1) $\quad U'(t) = f(U(t)) \quad (t\geq 0)\ ,\ U(0) = u_0\ ,$

where $f\colon \mathbb{K}^s \to \mathbb{K}^s$ and $u_0 \in \mathbb{K}^s$ are known. Let $h > 0$ be a given stepsize. Application of a one-step method for the numerical solution of (3.1.1) results in a scheme which we write as

(3.1.2) $\quad u_{n+1} = G(u_n;h,f) \qquad (n=0,1,2,\ldots)\ .$

Here $G(\cdot;h,f)$ is defined on some suitable subset of $\mathbb{K}^s$, and depends on the stepsize h and the function f. If h and f are fixed we will also write $G(\cdot)$ instead of $G(\cdot;h,f)$.

The one-step method which determines the function $G(\cdot;\cdot,\cdot)$ (in the above sense) will be called the *method* G. The restriction to a constant stepsize h in the above is merely made for notational convenience. If we deal with a nonuniform grid $\{t_n : t_0=0, t_i=t_{i-1}+h_i\ (i\in\mathbb{N})\}$, application of method G to the problem (3.1.1) yields the scheme $u_{n+1} = G(u_n;h_{n+1},f)$ $(n=0,1,2,\ldots)$.

We recall that the *local discretization errors* $\ell_h(t_n)$ of method G w.r.t. the solution U of (3.1.1) are defined by

$$U(t_{n+1}) = G(U(t_n);h,f) + h\ell_h(t_n) \qquad (n=0,1,2,\ldots)\ .$$

The *order* p of the method G is the largest integer such that $\ell_h(t_n) = O(h^p)$ $(h \to 0$, uniformly in $n)$ whenever f is sufficiently often differentiable (cf. HENRICI (1962)).

The best known class of one-step methods is formed by the explicit

Runge-Kutta methods. With these methods $G(\cdot) = G(\cdot;h,f)$ is defined by

$$(3.1.3.a) \qquad G(x) = x + \sum_{i=1}^{m} b_i \, hf(y_i(x)) \ ,$$

where the internal vectors $y_i(x) \in \mathbb{K}^s$ satisfy

$$(3.1.3.b) \qquad y_i(x) = x + \sum_{j=1}^{i-1} a_{ij} \, hf(y_j(x)) \qquad (1 \le i \le m) \ .$$

Here m, the number of stages, is a positive integer, and a_{ij}, b_i are real parameters.

Explicit Runge-Kutta methods are frequently used for non-stiff initial value problems. However, because of their restricted stability properties (see remark 5.3.1), these methods are not suited for stiff problems. For this reason generalizations of the explicit Runge-Kutta methods have been introduced. Such generalizations will be considered in the following sections.

3.2. IMPLICIT RUNGE-KUTTA METHODS

Let $m \in \mathbb{N}$, $A = (a_{ij}) \in L(\mathbb{R}^m)$ and $b = (b_i) \in \mathbb{R}^m$. This set of parameters determines a Runge-Kutta method where $G(\cdot) = G(\cdot;h,f)$ is defined by

$$(3.2.1.a) \qquad G(x) = x + \sum_{i=1}^{m} b_i \, hf(y_i(x)) \ ,$$

$$(3.2.1.b) \qquad y_i(x) = x + \sum_{j=1}^{m} a_{ij} \, hf(y_j(x)) \qquad (1 \le i \le m) \ .$$

(for $x \in \mathbb{K}^s$, with $h > 0$ and $f: \mathbb{K}^s \to \mathbb{K}^s$ as in section 3.1). If A is strictly lower triangular, (3.2.1) reduces to (3.1.3), an explicit Runge-Kutta method. If $a_{ij} \neq 0$ for some i and j with $1 \le i \le j \le m$, we call the Runge-Kutta method *implicit*.

With implicit Runge-Kutta methods the problem of solving the internal vectors $y_i(x)$ from the system of algebraic equations (3.2.1.b) is no longer trivial, and in practical computations some numerical iterative

method is used for this purpose. This numerical method should also be convergent in cases where the product of h with the Lipschitz constant L of the function f is large, because the implicit Runge-Kutta methods are intended to treat such cases. For this reason one generally uses some modification of Newton's method, rather than direct (functional) iteration (see e.g. SHAMPINE (1980), FRANK, SCHNEID and UEBERHUBER (1982 B)).

By the work of Butcher (see e.g. BUTCHER (1965), GRIGORIEFF (1972)) it is known that by choosing the parameters a_{ij} and b_i in a suitable way, one can construct an implicit Runge-Kutta method with order $2m$, but not higher. These methods with maximal order (for a given m) are frequently called Gauss-methods. Other important classes of implicit Runge-Kutta methods can be found for instance in ALEXANDER (1977) and CHIPMAN (1971).

In recent years much research has been devoted to the stability and contractivity properties of implicit Runge-Kutta methods. Many implicit Runge-Kutta methods are known to be B-contractive or A-stable. A short review and references are given in chapter 5.

REMARK 3.2.1. For a nonautonomous initial value problem

$$U'(t) = f(t,U(t)) \quad (t\geq 0)\ ,\ U(0) = u_0\ ,$$

with $f\colon \mathbb{R} \times \mathbb{K}^s \to \mathbb{K}^s$, $u_0 \in \mathbb{K}^s$, a one-step method yields a scheme of the type

$$u_{n+1} = G(t_n,u_n) \qquad (n=0,1,2,\ldots)\ ,$$

where $G(\cdot,\cdot) = G(\cdot,\cdot;h,f)$ depends on f and the stepsize h. For Runge-Kutta methods $G(\cdot,\cdot)$ is then defined by

$$G(t,x) = x + \sum_{i=1}^{m} b_i\, hf(t+c_ih,y_i(t,x))\ ,$$

$$y_i(t,x) = x + \sum_{j=1}^{m} a_{ij}\, hf(t+c_jh,y_j(t,x)) \quad (1\leq i\leq m)$$

(for $t \geq 0$ and $x \in \mathbb{K}^s$). Generally it is assumed that $c_i = a_{i1} + a_{i2} + \ldots + a_{im}$ $(1\leq i\leq m)$. The Runge-Kutta method for nonautonomous problems is then again determined by the choice of the real $m \times m$ matrix $A = (a_{ij})$ and the vector $b = (b_i) \in \mathbb{R}^m$.

3.3. SEMI-IMPLICIT METHODS

3.3.1. Description of the methods.

The use of an implicit Runge-Kutta method for solving a nonlinear initial value problem (3.1.1) numerically, involves at each integration step the solution of a nonlinear system of algebraic equations. If we use a modified Newton method for this purpose the Jacobian $f'(x)$, or an approximation $J(x)$ to $f'(x)$, is needed. To avoid the nonlinear equations while retaining favourable stability properties, one can also incorporate the Jacobian directly in the method. This idea, which orginates with ROSENBROCK (1963), leads to the following generalization of the explicit Runge-Kutta methods.

Let $G(\cdot) = G(\cdot;h,f)$ be defined by

$$(3.3.1.a) \qquad G(x) = x + \sum_{i=1}^{m} b_i(hJ(x))\, hf(y_i(x)) ,$$

$$(3.3.1.b) \qquad y_i(x) = x + \sum_{j=1}^{i-1} a_{ij}(hJ(x))\, hf(y_j(x)) \qquad (1 \le i \le m) .$$

(for $x \in \mathbb{K}^s$, with $h > 0$ and $f: \mathbb{K}^s \to \mathbb{K}^s$ as in section 3.1). Here $J(\cdot) = J(\cdot;h,f): \mathbb{K}^s \to L(\mathbb{K}^s)$ is a given function depending on h and f, and the a_{ij}, b_i are rational functions with real coefficients (which we shall call the coefficients of G).

The class of methods (3.3.1) has been introduced by VAN DER HOUWEN (1977). In practical computations $J(x)$ will be an approximation to $f'(x)$. If some of the denominators of the rational functions a_{ij} and b_i are not identically equal to one, a method of the type (3.3.1) is called a *semi-implicit* method. To compute $G(x)$ for a given $x \in \mathbb{K}^s$, we then have to solve *linear* systems of algebraic equations of the form $q(hJ(x))v = w$ where $w \in \mathbb{K}^s$ is known and q is a denominator of one of the a_{ij} or b_i. Unless the matrix $J(x)$ has a special structure this is generally done by making an LU-decomposition of the matrix $q(hJ(x))$.

If $\rho \in \mathbb{R}$ is such that all rational functions a_{ij} and b_i only have a pole at ρ, it is sufficient to make a single LU-decomposition of $\rho I - hJ(x)$ to solve all the linear systems arising in (3.3.1). A class of semi-implicit methods which have this favourable property is formed by the *ROW-methods*. These methods are given by

$$(3.3.2.a)\qquad G(x) = x + \sum_{i=1}^{m} \beta_i \, hF_i(x) \ ,$$

$$(3.3.2.b)\qquad (I-\gamma hJ(x))\, F_i(x) = f\Big(x + \sum_{j=1}^{i-1} \alpha_{ij} hF_j(x)\Big) + \\ + \sum_{j=1}^{i-1} \gamma_{ij}\, hJ(x)\, F_j(x) + \sum_{j=1}^{i-1} \delta_{ij}\, F_j(x) \qquad (1\le i\le m) \ .$$

The real numbers $\alpha_{ij},\beta_i,\gamma_{ij},\delta_{ij}$ and γ are parameters defining the method. By taking

$$y_i(x) = x + \sum_{j=1}^{i-1} \alpha_{ij}\, hF_j(x) \qquad (1\le i\le m) \ ,$$

it can be seen by some calculations that (3.3.2) can also be written in the form (3.3.1) with coefficient functions a_{ij},b_i defined by

$$A(\zeta) = (a_{ij}(\zeta)) = \bar{A}(\bar{D}-\bar{C}\zeta)^{-1} \qquad (\zeta\in\mathbb{C}) \ ,$$

$$b(\zeta)^T = (b_1(\zeta),b_2(\zeta),\ldots,b_m(\zeta)) = \bar{b}^T(\bar{D}-\bar{C}\zeta)^{-1} \qquad (\zeta\in\mathbb{C}) \ ,$$

where $\bar{A} \in L(\mathbb{R}^m)$ is strictly lower triangular with entries α_{ij} , $\bar{b} \in \mathbb{R}^m$ has components β_i , $\bar{C} \in L(\mathbb{R}^m)$ is lower triangular with numbers γ on the diagonal and entries γ_{ij} below the diagonal, and $\bar{D} \in L(\mathbb{R}^m)$ is lower triangular with diagonal elements 1 and entries δ_{ij} below the diagonal. The ROW-method (3.3.2) can therefore be considered as a special semi-implicit method (3.3.1), with $1/\gamma$ being the only pole of all a_{ij} and b_i .

In (3.3.2) it is no restriction to take either all γ_{ij} or all δ_{ij} equal to zero (see KAPS and WANNER (1981), and VERWER, SCHOLZ, BLOM and LOUTER-NOOL (1982)). For actual computation it is convenient to take the γ_{ij} equal to zero, since then the matrix-vector products $J(x)F_j(x)$ do not have to be computed.

By ROSENBROCK (1963) semi-implicit one-step methods have been proposed which use more than one Jacobian evaluation per step. Since each new Jacobian (or Jacobian approximation) also involves a new LU-decomposition, these methods are unattractive from a computational point of view.

With the semi-implicit methods (3.3.1) we have the freedom of choosing an appropriate approximation $J(x) = J(x;h,f)$ to the Jacobian $f'(x)$. If the exact Jacobian $f'(x)$ is easily available we can take $J(x) = f'(x)$. We call the semi-implicit method (3.3.1) with $J \equiv f'$ a *Rosenbrock method*.

A second choice is to keep J fixed during the whole integration (or a large part of it). This case is also of much practical interest since solving the linear systems to compute the $y_i(u_n)$ and $u_{n+1} = G(u_n)$ from a given u_n will generally be rather expensive. If $J(u_n)$ remains constant for $n = 0,1,2,\ldots$, the amount of computational work is reduced considerably since the LU-decompositions of the denominators can be used throughout the integration proces, provided that the stepsize is not changed. If, for example, we know in advance that

$$(3.3.3) \qquad f(x) = \Lambda x + w(x) \qquad (\text{for } x \in \mathbb{K}^s),$$

where Λ is a linear operator from $\mathbb{K}^s$ to $\mathbb{K}^s$ and $w: \mathbb{K}^s \to \mathbb{K}^s$ has a small Lipschitz constant near the solution U of (3.1.1), the choice $J(\cdot) \equiv \Lambda$ seems very attractive from a computational point of view.

The situation (3.3.3) occurs for instance if an initial-boundary value problem for the semi-linear parabolic equation

$$\frac{\partial}{\partial t} u(x,t) = \frac{\partial^2}{\partial x^2} u(x,t) + g(u(x,t)) \qquad (0 \le x \le 1 \ , \ t \ge 0)$$

is solved by the method of lines, and g is a smooth function near the solution.

A strategy which has been followed by VERWER and SCHOLZ (1982) is to use the semi-implicit scheme $u_{n+1} = G(u_n)$ $(n \ge 0)$ with an exact Jacobian which is re-evaluated after every ν steps. This corresponds with the choice $J(u_{k\nu+\ell}) = f'(u_{k\nu})$ (for $\ell = 0,1,\ldots,\nu-1$ and $k = 0,1,2,\ldots$). The scheme which arises then can be viewed as an application of a Rosenbrock method with νm stages and with stepsize νh, by looking at the vectors $u_{k\nu+1}, u_{k\nu+2}, \ldots, u_{k\nu+\nu-1}$ as internal vectors $y_i(u_{k\nu})$.

For the order conditions for the general semi-implicit methods (3.3.1) the reader is referred to VAN DER HOUWEN (1972), NØRSETT and WOLFBRANDT (1979). The maximal attainable order $p^*(m)$ for a given m is known for $m \le 3$. These orders are

$$p^*(1) = 2 \ , \ p^*(2) = 4 \ , \ p^*(3) = 5 \ .$$

Let $q^*(m)$ stand for the maximal order which can be reached with an m-stage ROW-method. We have $q^*(m) \le m+1$ (see STEIHAUG and WOLFBRANDT (1979)). It was shown by KAPS and WANNER (1981) that order 5 is not possible for $m = 4$. For the ROW-methods we have the following results:

$$q^*(1) = 2 \ , \ q^*(2) = 3 \ , \ q^*(3) = 4 \ , \ q^*(4) = 4 \ .$$

REMARK 3.3.1. There are two options to adapt the semi-implicit methods to nonautonomous problems $U'(t) = f(t,U(t))$ $(t\ge0)$, $U(0) = u_0$. The first one is to use a nonautonomous version $G(t,x)$ of the $G(x)$ given by (3.3.1), where we read $J(t,x) \approx D_x f(t,x)$ instead of $J(x)$, and replace $f(y_i(x))$ by $f(t+c_i h, y_i(x))$ in (3.3.1). The second possibility is to convert the nonautonomous problem to an autonomous one (see remark 1.1.2). After this conversion (3.3.1) can be applied directly.

These two options yield different results - see VERWER (1981 A) for a more detailed discussion. Note that with the ordinary Runge-Kutta methods these options would lead to the same result because of the choice $c_i = a_{i1} + a_{i2} + \ldots + a_{im}$ $(1\le i\le m)$ (cf. remark 3.2.1).

3.3.2. Perturbed semi-implicit methods.

In this section we will motivate a condition on the coefficient functions a_{ij}, b_i of the semi-implicit methods (3.3.1), that will be convenient (though not strictly necessary) for the analysis in chapter 5. This condition is related to the BS- and BSI-stability concepts of FRANK, SCHNEID and UEBERHUBER (1982 A), add to the $\bar{A}$-stability concept of CROUZEIX and RAVIART (1980), for implicit Runge-Kutta methods.

We will consider the following class of functions $f: \mathbb{C} \to \mathbb{C}$ that was involved in the definition of A-stability (see section 1.2),

(3.3.4) $\qquad f(x) = \lambda x \quad$ (for $x \in \mathbb{C}$) with $\lambda \in \bar{\mathbb{C}}$.

Further $h > 0$ will be an arbitrary stepsize, and we assume that $J(x;h,f) = \lambda$ (for all $x \in \mathbb{C}$) for f given by (3.3.4).

If a semi-implicit method G of the type (3.3.1) is used for the solution of an initial value problem, at each step linear systems of

algebraic equations have to be solved. In dealing with stiff problems it is natural to require that solving these linear systems will not give rise to numerical difficulties for the class of simple testproblems where f is given by (3.3.4). To meet this requirement we have to assume that

(3.3.5) the rational functions a_{ij} and b_i $(1 \le i,j \le m)$ are analytic on $\bar{\mathbb{C}}^-$.

We then know that $G(x) = G(x;h,f)$ is defined for the class (3.3.4) (see also section 4.4). Further we will impose conditions on the method G to ensure that small perturbations will not disturb the computation of $G(x)$ and the internal vectors $y_i(x)$ too much.

Let $f: \mathbb{C} \to \mathbb{C}$ be given by (3.3.4), and let $x \in \mathbb{C}$. For $\zeta \in \mathbb{C}$ we denote by $A(\zeta)$ the strictly lower triangular $m \times m$ matrix with entries $a_{ij}(\zeta)$, and by $b(\zeta)$ the column vector in $\mathbb{C}^m$ with components $b_i(\zeta)$. Further $y(x)$ will stand for $(y_1(x),y_2(x),\ldots,y_m(x))^T$. Then $G(x) = G(x;h,f)$ is given by

$$(3.3.6) \qquad G(x) = x + h\lambda\, b(h\lambda)^T\, y(x)\ ,$$

$$y(x) = ex + h\lambda\, A(h\lambda)\, y(x)\ .$$

Besides (3.3.6) we consider the slightly perturbed version

$$(3.3.7) \qquad \tilde{G}(x) = x + h\lambda\, b(h\lambda)^T\, \tilde{y}(x) + w_0\ ,$$

$$\tilde{y}(x) = ex + h\lambda\, A(h\lambda)\, \tilde{y}(x) + w\ .$$

Here $w_0 \in \mathbb{C}$ and $w \in \mathbb{C}^m$ are small perturbations, for instance due to roundoff errors. From (3.3.6), (3.3.7) we obtain the relations

$$\tilde{y}(x) - y(x) = (I - h\lambda A(h\lambda))^{-1}\, w\ ,$$

$$\tilde{G}(x) - G(x) = h\lambda\, b(h\lambda)^T\, [\tilde{y}(x) - y(x)] + w_0\ .$$

The requirement that the computation of $G(x)$ and $y(x)$ cannot be disturbed unduely by small perturbations w_0, w for arbitrary f satisfying (3.3.4), thus leads to the following condition.

$$\text{(3.3.8.a)}\qquad \sup_{\zeta\in\mathbb{C}^-} |[(I-\zeta A(\zeta))^{-1}]_{ij}| < \infty \qquad (1\le i,j\le m)\ ,$$

$$\text{(3.3.8.b)}\qquad \sup_{\zeta\in\mathbb{C}^-} |[\zeta b(\zeta)^T(I-\zeta A(\zeta))^{-1}]_i| < \infty \qquad (1\le i\le m)\ .$$

From lemma 2.4.11 we see that this condition is equivalent to

$$\text{(3.3.9)}\qquad a_{ij}(\infty) = 0 \quad (\text{for } 1\le j<i\le m) \quad \text{and} \quad b_i(\infty) = 0 \quad (1\le i\le m)\ .$$

Summarizing the above we obtain the following result.

THEOREM 3.3.2. *Suppose* (3.3.4)-(3.3.7) *and* $h > 0$. *Let* $\|\cdot\|$ *be an arbitrary norm on* $\mathbb{C}^m$. *There are positive constants* σ_0 *and* σ *such that* $|\widetilde{G}(x)-G(x)| \le \sigma_0(|w_0|+\|w\|)$, $\|\widetilde{y}(x)-y(x)\| \le \sigma\|w\|$ *for all* $\lambda \in \mathbb{C}^-$, $w_0 \in \mathbb{C}$ *and* $w \in \mathbb{C}^m$ *iff* (3.3.9) *holds.*

EXAMPLE 3.3.3. Consider the methods given by

$$G_1(x;h,f) = x + (I-hf'(x))^{-1}\, hf(x)$$

($m = 1$, $b_1(\zeta) = (1-\zeta)^{-1}$), and

$$G_2(x;h,f) = x + hf(x+(I-hf'(x))^{-1}hf(x))$$

($m = 2$, $a_{21}(\zeta) = (1-\zeta)^{-1}$, $b_1(\zeta) \equiv 0$, $b_2(\zeta) \equiv 1$). Both methods can be viewed as a linearization of the Backward Euler method (method (3.2.1) with $m = 1$, $A = 1$ and $b = 1$). If f is linear and there are no perturbations, G_1 and G_2 will yield the same numerical results.

However, if we consider perturbations w_0, w as in (3.3.7) for the simple class of testfunctions (3.3.4), we get with the first method

$$\widetilde{G}_1(x) - G_1(x) = w_0 + h\lambda(1-h\lambda)^{-1}\, w \qquad (w_0, w \in \mathbb{C})\ ,$$

whereas we obtain with the second method

$$\widetilde{G}_2(x) - G_2(x) = w_0 + ((1+h\lambda)^{-1}\, h^2\lambda^2, h\lambda)w \quad (w_0 \in \mathbb{C}\ ,\ w \in \mathbb{C}^2)\ .$$

Thus for $|\lambda| \to \infty$, the effect of the perturbations remains bounded with method G_1 , which satisfies (3.3.9). For method G_2 we have

$$|\tilde{G}_2(x)-G_2(x)| \to \infty \qquad (\text{for } |\lambda| \to \infty).$$

In view of the foregoing we will confine ourselves in subsequent chapters to semi-implicit methods satisfying (3.3.9). Most well-known semi-implicit methods, such as the ROW-methods, do satisfy (3.3.9).

We note that (3.3.9) can be viewed as a BS + BSI-stability requirement on the semi-implicit methods for the class of problems (3.3.4). Also for the S-stability analysis in VERWER (1977), (3.3.9) was assumed (but not motivated as in this section).

3.4. ADAPTIVE RUNGE-KUTTA METHODS AND TRANSLATION INVARIANCE

The methods we have encountered thus far constitute the class of one-step methods which are mostly used for solving stiff problems, and we shall restrict ourselves to this class in the subsequent chapters.

A further generalization of the explicit Runge-Kutta methods was proposed by LAWSON (1967) (see also GRIGORIEFF (1972)). Modifications of Lawson's methods were given by EHLE and LAWSON (1975), FRIEDLI (1978) and STREHMEL (1981). In their general form these so called *adaptive Runge-Kutta methods* are given by the following function G. For $f: \mathbb{K}^s \to \mathbb{K}^s$ and $h > 0$ given, we define $G(\cdot) = G(\cdot;h,f)$ by

$$(3.4.1.a) \qquad G(x) = \phi(hJ(x))x + \sum_{i=1}^{m} \beta_i(hJ(x))\, hg_i(x) ,$$

$$(3.4.1.b) \qquad g_i(x) = f(y_i(x))-J(x)\, y_i(x) \qquad (1\le i\le m) ,$$

$$(3.4.1.c) \qquad y_i(x) = \phi(\gamma_i hJ(x))x + \sum_{j=1}^{i-1} \alpha_{ij}(hJ(x))\, hg_j(x) \qquad (1\le i\le m)$$

(for $x \in \mathbb{K}^s$). Here ϕ is a rational approximation to the exponential function, the α_{ij} and β_i are rational functions, and the γ_i are real parameters. By some manipulation it can be seen that (3.4.1) can also be written in the form

$$(3.4.2.a) \qquad G(x) = c_0(hJ(x))x + \sum_{i=1}^{m} b_i(hJ(x))\, hf(y_i(x)) ,$$

$$(3.4.2.b) \qquad y_i(x) = c_i(hJ(x))x + \sum_{j=1}^{i-1} a_{ij}(hJ(x))\, hf(y_j(x)) \qquad (1\le i\le m) ,$$

where the a_{ij}, b_i and c_i are rational functions. This seems like a real generalization of (3.3.1) where all c_i are identically equal to one. In this section it will be shown however, that if not all c_i can be taken identically equal to one, method (3.4.2) suffers from a drawback which is not present in the methods (3.3.1).

The first adaptive Runge-Kutta methods given by LAWSON (1967) are of the type (3.4.2) and cannot be written as (3.3.1). With many of the subsequent modifications the α_{ij}, β_i and γ_i in (3.4.1) were chosen such that the methods can be written as (3.3.1).

EXAMPLE 3.4.1. Lawson's generalization of Euler's method is given by

$$G(x) = \phi(hJ(x))\ (I-hJ(x))x + \phi(hJ(x))\ hf(x)\ ,$$

where ϕ is a rational approximation to the exponential function. Unless $\phi(\zeta) = (1-\zeta)^{-1}$ $(\zeta \in \mathbb{C})$, this method of the type (3.4.1) cannot be written in the form (3.3.1).

EXAMPLE 3.4.2. By FRIEDLI (1978) and STREHMEL (1981) the following method was considered.

$$G(x) = \phi_0(hJ(x))x + [\phi_1(hJ(x)) - \frac{1}{c}\phi_2(hJ(x))]\ hg_1(x) +$$

$$+ \frac{1}{c}\phi_2(hJ(x))\ \ hg_2(x)\ ,$$

$$y_1(x) = x\ ,$$

$$y_2(x) = \phi_0(chJ(x))x + c\phi_1(chJ(x))\ hg_1(x)\ ,$$

where $g_i(x) = f(y_i(x)) - J(x)\ y_i(x)$ $(i=1,2)$, $c \in \mathbb{R}$, ϕ_0 is a rational approximation to the exponential function, $\phi_1(\zeta) = \frac{1}{\zeta}[\phi_0(\zeta)-1]$, and $\phi_2(\zeta) = \frac{1}{\zeta}[\phi_1(\zeta)-1]$ $(\zeta \in \mathbb{C})$.

By some manipulation it can be seen that this method can be written in the form (3.3.1) with $m = 2$ and

$$a_{21}(\zeta) = c\phi_1(c\zeta)\ ,\ b_1(\zeta) = \phi_1(\zeta) - \frac{1}{c}\phi_2(\zeta),$$

$$b_2(\zeta) = \frac{1}{c}\phi_2(\zeta)\ \ (\zeta \in \mathbb{C})\ .$$

We will show in the subsequence that those adaptive Runge-Kutta methods which do not fit into the form (3.3.1) are not *translation invariant*. With this we mean the following.

DEFINITION 3.4.3. *A one-step method* G *is said to be translation invariant if*

$$G(x;h,f) = G(x-w;h,g) + w \qquad (\text{for all } x \in \mathbb{K}^s,\ h > 0)$$

for arbitrary $w \in \mathbb{K}^s$ *and* $f,g: \mathbb{K}^s \to \mathbb{K}^s$ *satisfying* $f(x) = g(x-w)$ *(for all* $x \in \mathbb{K}^s$*)*.

Translation invariance is a very natural requirement. Consider the initial value problems

$$V'(t) = g(V(t)) \quad (t \geq 0),\ V(0) = v_0,$$

and

$$U'(t) = f(U(t)) \quad (t \geq 0),\ U(0) = u_0,$$

where $u_0 = v_0 + w$, and w,f,g are as in defintion 3.4.3. The solutions U,V satisfy $U(t) = V(t) + w$ $(t \geq 0)$. If the method G is translation invariant, the numerical approximations $\{u_n\}, \{v_n\}$ to U,V, respectively, will satisfy $u_n = v_n + w$ $(n=0,1,2,\ldots)$.

If a method G is not translation invariant, the translation vector w will enter into the local discretization error of the method w.r.t. the solution U of the above initial value problem. Hence this local discretization error will not only depend on derivatives of U and on the function values of f and its derivatives near this solution. In particular, a method which is not translation invariant will also not be B-convergent (see chapter 6). This will generally lead to a bad behaviour for stiff initial value problems.

EXAMPLE 3.4.4. Consider again Lawson's generalization of Euler's method (see example 3.4.1) with $\phi(0) = 1$, $\phi(\zeta) \not\equiv (1-\zeta)^{-1}$.

Let $f(x) = \lambda(x-w)$ (for $x \in \mathbb{C}$) with $\lambda,w \in \mathbb{C}$ given and $w \neq 0$. Assume $J(x) = \lambda$ (for all $x \in \mathbb{C}$). The local discretization error

$\ell_h(t) = h^{-1}[U(t+h)-G(U(t))]$ w.r.t. the stationary solution $U(t) \equiv w$, equals

$$h^{-1}[1-\phi(h\lambda)(1-h\lambda)]\, w\ .$$

Suppose $L \in \mathbb{R}$ and $L >> 1$. Then $\ell_h(t)$ is small for all $\lambda \in \mathbb{C}$ with $\mathrm{Re}\,\lambda \le 0$, $|\lambda| \le L$ iff hL is small. Thus if we deal with a stiff initial value problem, the stepsize h has to be exessively small in order to ensure a small local error.

By some simple calculations it can be shown that all methods of the type (3.3.1) are translation invariant, provided that

(3.4.3) $J(x;h,f) = J(x-w;h,g)$ (for all $x \in \mathbb{K}^s$, $h > 0$) for arbitrary $w \in \mathbb{K}^s$ and $f,g: \mathbb{K}^s \to \mathbb{K}^s$ satisfying $f(x) = g(x-w)$ (for all $x \in \mathbb{K}^s$).

Note that $f'(x) = g'(x-w)$ if f and g are as in (3.4.3). The following theorem shows that the above is not true for those adaptive Runge-Kutta methods which cannot be written as (3.3.1).

THEOREM 3.4.5. *Assume* (3.4.3). *Let* G *be a method of the type* (3.4.2). *Then* G *is translation invariant iff* G *is equivalent to a method of the type* (3.3.1).

PROOF. Suppose the method G, given by (3.4.2), is translation invariant. By $\tilde{G}$ we denote the method given by (3.3.1) with the same coefficient functions a_{ij} and b_i as G.

We have $G(0;h,f) = \tilde{G}(0;h,f)$ for all $h > 0$ and $f: \mathbb{K}^s \to \mathbb{K}^s$. Since both G and $\tilde{G}$ are translation invariant it follows that $G(x;h,f) = \tilde{G}(x;h,f)$ for all $x \in \mathbb{K}^s$, $h > 0$ and $f: \mathbb{K}^s \to \mathbb{K}^s$. Thus G and $\tilde{G}$ always yield the same numerical approximations. □

CHAPTER 4

THE EXISTENCE OF UNIQUE SOLUTIONS TO THE ALGEBRAIC EQUATIONS IN IMPLICIT AND SEMI-IMPLICIT METHODS

4.1. INTRODUCTION

If we approximate the solution of the initial value problem $U'(t) = f(U(t))$ $(t\geq 0)$, $U(0) = u_0$, using an implicit Runge-Kutta method, at each step of the integration a system of *algebraic* equations of the following type has to be solved (cf. (3.1.2), (3.2.1.b)).

$$(4.1.1) \qquad y_i(u_n) = u_n + \sum_{j=1}^{m} a_{ij}\, hf(y_j(u_n)) \qquad (1\leq i\leq m) \ .$$

It is known (see e.g. GRIGORIEFF (1972)) that if $f: \mathbb{K}^s \to \mathbb{K}^s$ $(s\geq 1)$ satisfies a Lipschitz condition

$$|f(\tilde{x})-f(x)| \leq L|\tilde{x}-x| \qquad (\text{for all } \tilde{x},x \in \mathbb{K}^s),$$

and the product hL is sufficiently small, then (4.1.1) has a unique solution. However, if the initial value problem is stiff such a restriction on h is embarrassing.

In this chapter we consider the question whether the system (4.1.1) has a unique solution for all continuous functions $f: \mathbb{K}^s \to \mathbb{K}^s$ satisfying a one-sided Lipschitz condition

$$(4.1.2) \qquad \mathrm{Re}\ \langle f(\tilde{x})-f(x),\tilde{x}-x\rangle \leq \beta|\tilde{x}-x|^2 \qquad (\text{for all } \tilde{x},x \in \mathbb{K}^s) \ .$$

Here $\beta \in \mathbb{R}$, $\langle\cdot,\cdot\rangle$ is an arbitrary inner product on $\mathbb{K}^s$ and $|\cdot|$ stands for the corresponding norm.

If we deal with semi-implicit methods the similar question arises whether all the linear systems that have to be solved to compute an approximation u_{n+1} from a given u_n have unique solutions. This question is much easier to answer and will only be considered in section 4.4. Until

then we turn our attention to the implicit Runge-Kutta methods.

Instead of (4.1.1) we shall consider in the subsequence the more general system

$$(4.1.3) \qquad y_i = \sum_{j=1}^{m} a_{ij}\, hf_j(y_j) \qquad (1\le i\le m)\ ,$$

where all functions $f_j\colon \mathbb{K}^s \to \mathbb{K}^s$ are continuous and satisfy

$$(4.1.4) \qquad \mathrm{Re}\ \langle f_j(\tilde{x})-f_j(x),\tilde{x}-x\rangle \le \beta|\tilde{x}-x|^2 \qquad (\text{for all } \tilde{x},x \in \mathbb{K}^s)\ .$$

By taking $f_j(x) = f(x+u_n)$ $(x\in\mathbb{K}^s)$ and replacing y_j by $y_j(u_n) - u_n$ $(i\le j\le m)$, the system (4.1.3) reduces to (4.1.1). An advantage of (4.1.3) over (4.1.1) is that also perturbed systems (4.1.1) or the systems which arise for nonautonomous problems (see remark 3.2.1) can be dealt with. These cases are covered by considering

$$\tilde{y}_j(u_n) = u_n + \sum_{j=1}^{m} a_{ij}\, f(t_n+c_jh,\tilde{y}_j(u_n)) + v_i \quad (1\le i\le m)$$

with perturbations $v_i \in \mathbb{K}^s$, and $f\colon \mathbb{R}\times\mathbb{K}^s \to \mathbb{K}^s$, instead of (4.1.1). Such perturbed equations were considered by FRANK, SCHNEID and UEBERHUBER (1982 A,B) for proving B-consistency of some implicit Runge-Kutta methods.

In this chapter we use the notations and conventions that were introduced in section 2.4.1. Let $A = (a_{ij}) \in L(\mathbb{R}^m)$. With the convention that the Kronecker product $A \otimes I^{(s)}$ will also be denoted by A, the system (4.1.3) can be written as

$$(4.1.5) \qquad y - h\,AF(y) = 0$$

where $y = (y_1^T,y_2^T,\ldots,y_m^T)^T$ and $F(y) = (f_1(y_1)^T,f_2(y_2)^T,\ldots,f_m(y_m)^T)^T \in \in \mathbb{K}^{sm}$.

If $m = 1$, the system (4.1.5) is essentially the same as the system of algebraic equations that arises if a linear multistep method is used to compute the approximations u_n. For this case several authors have discussed the existence and uniqueness of solutions; see e.g. DAHLQUIST

(1975), DESOUR and HANEDA (1972), and WILLIAMS (1979). Their results can also be used for *diagonally implicit* Runge-Kutta methods, i.e. $a_{ij} = 0$ whenever $i < j$.

The question whether the system (4.1.1) has a unique solution for stiff initial value problems was considered in CROUZEIX and RAVIART (1980), and HUNDSDORFER and SPIJKER (1981). The results of these papers were combined in CROUZEIX, HUNDSDORFER and SPIJKER (1983). Results similar to those in CROUZEIX and RAVIART (1980) were obtained by DEKKER (1982).

We now give a short outline of this chapter. In the sections 4.2 and 4.3 we deal with implicit Runge-Kutta methods. The case where the differential equation is linear with constant coefficients will be considered in section 4.2. In section 4.3 sufficient conditions (on A , h and β) will be given that ensure the existence of a unique solution to (4.1.5). These results slightly generalize some results given in CROUZEIX, HUNDSDORFER and SPIJKER (1983) and DEKKER (1982). Finally, in section 4.4, the semi-implicit methods will be considered.

4.2. IMPLICIT RUNGE-KUTTA METHODS FOR LINEAR DIFFERENTIAL EQUATIONS

In this section we assume that the functions f_j , arising in (4.1.5), are of the form

$$(4.2.1) \qquad f_j(x) = \Lambda x + w_j \qquad (x \in \mathbb{K}^s , 1 \le j \le m)$$

where $\Lambda \in L(\mathbb{K}^s)$ and $w_j \in \mathbb{K}^s$ $(1 \le j \le m)$. This case occurs if the differential equation is linear with constant coefficients, $U'(t) = \Lambda U(t)$ $(t \ge 0)$ - or more general $U'(t) = \Lambda U(t) + w(t)$ $(t \ge 0)$.

For these functions f_j necessary and sufficient conditions (on A , h and β) for the unique solvability of (4.1.5) can be given quite easily.

Writing $w = (w_1^T, w_2^T, \ldots w_m^T)^T \in \mathbb{K}^{sm}$, $z_0 = h\Lambda$ and $Z_0 = \mathrm{diag}(z_0, z_0, \ldots, z_0) \in L(\mathbb{K}^{sm})$, (4.1.5) reads for this case

$$(4.2.2) \qquad (I - AZ_0)y = hAw .$$

Obviously this system has a unique solution iff the $sm \times sm$ matrix $I - AZ_0$ is regular.

In the following lemma we use the sets $C_r \subset \mathbb{C}$ defined by

$$C_r = \{\zeta:\zeta\in\mathbb{C}, |\zeta - \frac{1}{2r}| > |\frac{1}{2r}| \text{ or } \zeta=0\} \quad \text{if} \quad r < 0 ,$$

$$C_r = \{\zeta:\zeta\in\mathbb{C}, \mathrm{Re}\,\zeta>0 \text{ or } \zeta=0\} \quad \text{if} \quad r = 0 ,$$

$$C_r = \{\zeta:\zeta\in\mathbb{C}, |\zeta - \frac{1}{2r}| < |\frac{1}{2r}| \text{ or } \zeta=0\} \quad \text{if} \quad r > 0 .$$

THEOREM 4.2.1. *The following three statements are equivalent.*

(i) $I - AZ_0$ *is regular whenever* $Z_0 = \mathrm{diag}(z_0,z_0,\ldots,z_0) \in L(\mathbb{K}^{sm})$ *with* $z_0 \in L(\mathbb{K}^s)$, $s \geq 1$, $\mu[z_0] \leq h\beta$.

(ii) $I - A\zeta_0$ *is regular for all* $\zeta_0 \in \mathbb{C}$ *with* $\mathrm{Re}\,\zeta_0 \leq h\beta$.

(iii) $\sigma(A) \subset C_{h\beta}$.

PROOF. We observe that AZ_0 can be written as $A \otimes z_0$ for $Z_0 = \mathrm{diag}(z_0,z_0,\ldots,z_0)$. It follows (see e.g. MARCUS and MINC (1964)) that the spectrum $\sigma(AZ_0)$ equals $\sigma(A)\sigma(z_0) = \{\lambda\nu:\lambda\in\sigma(A),\nu\in\sigma(z_0)\}$. This fact will be used in the subsequence.

Obviously (i) implies (ii) if $\mathbb{K} = \mathbb{C}$. Assume $\mathbb{K} = \mathbb{R}$ and $I - A\zeta_0$ is singular, $\zeta_0 = \xi_0 + i\eta_0$ with $\xi_0 \leq h\beta$. Let d_1,d_2 be vectors in the $\mathbb{R}^2$ which are orthonormal w.r.t. the inner product $\langle\cdot,\cdot\rangle$. As in the proof of lemma 2.2.6 we define $z_0 \in L(\mathbb{R}^2)$ by

$$z_0d_1 = \xi_0d_1 - \eta_0d_2 , \quad z_0d_2 = \eta_0d_1 + \xi_0d_2 .$$

By some calculations it can be seen that $\mu[z_0] = \xi_0$ and $\zeta_0 \in \sigma(z_0)$. Hence $1 \in \sigma(A)\sigma(z_0) = \sigma(AZ_0)$, $I - AZ_0$ is singular. Thus (i) also implies (ii) if $\mathbb{K} = \mathbb{R}$.

Now suppose (ii) holds, and let $z_0 \in L(\mathbb{K}^s)$ with $\mu[z_0] \leq h\beta$. Then $\mathrm{Re}\,\nu \leq h\beta$ for any $\nu \in \sigma(z_0)$. It follows that $1 \notin \sigma(A)\sigma(z_0)$. Hence (i) holds.

The equivalence of (ii) and (iii) follows by some elementary calculations. □

We note that for $\beta = 0$ the above equivalence of (i) and (ii) was given already by SPIJKER (1982 A).

4.3. GENERAL RESULTS FOR IMPLICIT RUNGE-KUTTA METHODS

4.3.1. A sufficient condition for the existence of a unique solution to (4.1.5).

In this section sufficient conditions on A,h and β will be given that ensure the existence of a unique solution to (4.1.5) for nonlinear functions f_j.

The basic assumption on the matrix $A = (a_{ij})$ is the following (cf. (2.4.1)).

(4.3.1) There is a positive definite diagonal matrix $D \in L(\mathbb{R}^m)$ such that $(v,DAv) \geq \sigma(Av,DAv)$ (for all $v \in \mathbb{R}^m$).

Here $\sigma \in \mathbb{R}$ is a given number.

THEOREM 4.3.1. *Suppose the functions* f_j *are continuous and satisfy* (4.1.4), A *satisfies* (4.3.1), *and* $h\beta < \sigma$. *Then the system* (4.1.5) *has a unique solution* $y^* \in \mathbb{K}^{sm}$, *and* $\|y^*\|_D \leq (\sigma - h\beta)^{-1} h\|F(0)\|_D$.

PROOF. Consider the functions $\Phi,\Psi: \mathbb{K}^{sm} \to \mathbb{K}^{sm}$ defined by

$$\Phi(y) = y - hAF(y) \ , \ \Psi(y) = y - hF(Ay) \qquad (y \in \mathbb{K}^{sm}) .$$

We will first show that Ψ has a unique zero.

1. Suppose $I - ZA$ is singular, $Z = \mathrm{diag}(z_1,z_2,\ldots,z_m)$, $z_i \in L(\mathbb{K}^s)$, $\mu[z_i] \leq h\beta$ $(1 \leq i \leq m)$. It will be shown that this leads to a contradiction.

Take $y \in \mathbb{K}^{sm}$ such that $y \neq 0$, $y = ZAy$. Then $Ay \neq 0$, and using lemma 2.4.1 we see that

$$\mathrm{Re}[ZAy,Ay]_D = \mathrm{Re}[y,Ay]_D \geq \sigma\|Ay\|_D^2 .$$

Since $\mu[z_i] \leq h\beta$ $(1 \leq i \leq m)$, it also follows that

$$\mathrm{Re}[ZAy,Ay]_D \leq h\beta\|Ay\|_D^2 .$$

which contradicts the assumption $h\beta < \sigma$.

Thus $I - ZA$ is regular whenever $Z = \mathrm{diag}(z_1,z_2,\ldots,z_m)$ with

$z_i \in L(\mathbb{K}^s)$, $\mu[z_i] \le h\beta$ $(1 \le i \le m)$.

2. Suppose $\Psi(\tilde{y}) = \Psi(y)$ $(\tilde{y}, y \in \mathbb{K}^{sm})$. Then $\tilde{y}-y = hF(A\tilde{y}) - hF(Ay)$. From corollary 2.4.8 we know there exists a matrix $Z = \mathrm{diag}(z_1, z_2, \ldots, z_m)$ with $z_i \in L(\mathbb{K}^s)$, $\mu[z_i] \le h\beta$ $(1 \le i \le m)$ such that $hF(A\tilde{y}) - hF(Ay) = Z(A\tilde{y}-Ay)$. From part 1 of this proof it follows that $\tilde{y} = y$.

Thus Ψ is one-to-one.

3. We will now show that $\|\Psi(y)\|_D \to \infty$ (for $\|y\|_D \to \infty$).

For arbitrary $\tilde{y}, y \in \mathbb{K}^{sm}$ we have

$$\mathrm{Re}[\Psi(\tilde{y})-\Psi(y), A\tilde{y}-Ay]_D = \mathrm{Re}[\tilde{y}-y, A\tilde{y}-Ay]_D -$$

$$- h\,\mathrm{Re}[F(A\tilde{y})-F(Ay), A\tilde{y}-Ay]_D \ge (\sigma-h\beta)\,\|A\tilde{y}-Ay\|_D^2 .$$

Using the Schwarz inequality it follows that

$$\|\Psi(\tilde{y})-\Psi(y)\|_D \ge (\sigma-h\beta)\,\|A\tilde{y}-Ay\|_D .$$

Hence we obtain for any $y \in \mathbb{K}^{sm}$,

$$\|\Psi(y)\|_D \ge (\sigma-h\beta)\,\|Ay\|_D - \|\Psi(0)\|_D =$$

$$= (\sigma-h\beta)\,\|Ay\|_D - h\|F(0)\|_D .$$

On the other hand we have

$$\|\Psi(y)\|_D \ge \|y\|_D - h\|F(Ay)\|_D .$$

Thus we get for any $y \in \mathbb{K}^{sm}$ the relation

(4.3.2) $$\|\Psi(y)\|_D \ge \max\{(\sigma-h\beta)\|Ay\|_D - h\|F(0)\|_D, \|y\|_D - h\|F(Ay)\|_D\} .$$

Suppose there is a constant $K_0 > 0$ and a sequence $\{y_k\}$ in $\mathbb{K}^{sm}$ such that $\|y_k\|_D \to \infty$ while $\|\Psi(y_k)\|_D \le K_0$ (for all $k \in \mathbb{N}$). From (4.3.2) it follows that $\|Ay_k\|_D \le (\sigma-h\beta)^{-1}$ $(K_0 + h\|F(0)\|_D) = K_1$. Let $K_2 = \sup\{h\|F(y)\|_D : y \in \mathbb{K}^{sm}, \|y\|_D \le K_1\}$. We then obtain from (4.3.2),

$\|\Psi(y_k)\|_D \geq \|y_k\|_D - K_2$, which contradicts the boundedness of $\{\|\Psi(y_k)\|_D\}$.

Thus $\lim_{\|y\|_D \to \infty} \|\Psi(y)\|_D = \infty$.

4. From the results of the parts 2 and 3 of this proof, it can be seen that Ψ is a local homeomorfism at each point $y \in \mathbb{K}^{sm}$. From the norm-coerciveness theorem (see e.g. ORTEGA and RHEINBOLDT (1970)) it now follows that Ψ has a unique zero $y^0 \in \mathbb{K}^{sm}$. Moreover, using relation (4.3.2), it follows that $\|Ay^0\|_D \leq (\sigma - h\beta)^{-1} h\|F(0)\|_D$.

5. Between the functions Φ and Ψ the following relations exist.

$$\Phi(Ay) = A\Psi(y) \text{ (for all } y \in \mathbb{K}^{sm}),$$
$$\Phi(y) = 0 \text{ implies } \Psi(hF(y)) = 0 .$$

It follows that Φ also has a zero, namely $y^* = Ay^0$. Further we see that if $\Phi(y) = \Phi(\tilde{y}) = 0$, then $F(y) = F(\tilde{y})$ (since Ψ is injective), and $y - hAF(y) = \tilde{y} - hAF(\tilde{y})$. Hence $y = \tilde{y}$. □

COROLLARY 4.3.2. *Suppose* (4.3.1) *holds with a positive* σ . *Then the system* (4.1.5) *has a unique solution whenever* $h > 0$ *and the functions* f_j *are continuous and dissipative.*

We note that if $DA + A^TD$ is positive definite for some positive definite diagonal matrix D , then A is regular and (4.3.1) holds with a positive σ . On the other hand (4.3.1) may hold with a constant $\sigma > 0$ for a singular matrix A . If for instance all entries a_{ij} of A equal $a > 0$, we may take $D = I$ and $\sigma = (ma)^{-1}$. Therefore corollary 4.3.2 is a proper extension of theorem 1 in CROUZEIX, HUNDSDORFER and SPIJKER (1983).

Theorem 4.3.1 also contains the following result that was stated in the same paper.

COROLLARY 4.3.3. *Suppose* (4.3.1) *holds with* $\sigma = 0$. *Then the system* (4.1.5) *has a unique solution whenever* $h > 0$, *and the functions* f_j *are continuous and satisfy* (4.1.4) *with* $\beta < 0$.

REMARK 4.3.4. The B-contractive Runge-Kutta methods which are interesting from a practical point of view are such that $B = \mathrm{diag}(b_1, b_2, \ldots, b_m)$ is positive definite and $BA + A^TB - bb^T$ is positive semi-definite (see

section 5.5.2). The matrix $BA + A^TB$ is then positive semi-definite and corollary 4.3.3 is thus applicable.

By FRANK, SCHNEID and UEBERHUBER (1982 A) it was shown that many well-known B-contractive methods satisfy the condition of corollary 4.3.2, and for these methods the system (4.1.5) has a unique solution as soon as the functions f_j are continuous and dissipative.

In CROUZEIX, HUNDSDORFER and SPIJKER (1983) an algebraically contractive method and dissipative functions $f_j: \mathbb{C} \to \mathbb{C}$ were given such that (4.1.5) has no solution. This shows that the requirement $\sigma - h\beta > 0$ in theorem 4.3.1 is essential and cannot be replaced by the weaker condition $\sigma - h\beta \geq 0$.

4.3.2. Some extensions of the sufficient conditions.

For some special matrices A the sufficient conditions given in theorem 4.3.1 admit an obvious extension.

Let A_k be the $(m-1)\times(m-1)$ matrix that results if we remove the k-th. row and column from A . Using theorem 4.3.1, the proof of the following result is straightforward.

COROLLARY 4.3.5. *Suppose that the k-th. row or the k-th. column of* A *only contains zeros. If the requirement of theorem* 4.3.1 *holds for* A_k *(instead of* A *), the system* (4.1.5) *has a unique solution.*

Matrices A with a zero column or row occur if the Runge-Kutta method is a collocation method where the collocation points are based on Lobatto or Radau points (see e.g. GRIGORIEFF (1972, pp. 37-38)).

If the Runge-Kutta method is diagonally implicit (i.e. $a_{ij} = 0$ for $i < j$), we are actually dealing with m equations of the type (4.1.5)-with a_{ii} instead of A , and $m = 1$. These equations can be solved one after another. In this case theorem 4.3.1 can be applied with $m = 1$.

As an illustration of the above we consider the following example (see also DAHLQUIST (1975)).

EXAMPLE 4.3.6. For the trapezoidal rule we have

$$A = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} .$$

This method is diagonally implicit, and the first row of A is zero. Theorem 4.3.1 is not directly applicable. However, the implicit equation that has to be solved reads

$$y_2(u_n) = u_n + \tfrac{1}{2} hf(u_n) + \tfrac{1}{2} hf(y_2(u_n)) .$$

Taking $y = y_2(u_n) - u_n - \frac{1}{2} hf(u_n)$, $F(x) = f(x+u_n+\frac{1}{2}hf(u_n))$ $(x \in \mathbb{K}^s)$, this equation is tranformed into

$$y = \tfrac{1}{2} hF(y) .$$

From theorem 4.3.1 it can now be seen that if $h\beta < 2$ and F satisfies (4.1.4) (i.e. f satisfies (4.1.2)) this equation has a unique solution.

REMARK 4.3.7. Suppose $\beta = 0$ and $h > 0$. The question what conditions have to be imposed on A to ensure that the system (4.1.5) has a unique solution whenever (4.1.4) holds can be answered completely in two restricted cases.

In the first case we assume that $\mathbb{K}^s = \mathbb{R}^1$. Then the following result can be proved.

(4.3.2) The system (4.1.5) has a unique solution for all continuous functions $f_j : \mathbb{R} \to \mathbb{R}$ satisfying (4.1.4) with $\beta = 0$ $(1 \le j \le m)$ iff all principal subdeterminants of A are nonnegative.

Also if $m \le 2$ (and $\mathbb{K}^s$ is arbitrary) the question can be answered completely. Using essentially the same technique as in the proof of theorem 4.3.1, the following equivalence can be proved.

(4.3.3) If $m \le 2$, the statements (i), (ii), (iii) and (iv) are equivalent.
(i) System (4.1.5) has a unique solution for all continuous functions f_j $(1 \le j \le m)$ satisfying (4.1.4) with $\beta = 0$.

(ii) $I - AZ$ is regular whenever $Z = \mathrm{diag}(z_1, z_2, \ldots, z_m)$ with $z_i \in L(\mathbb{K}^s)$, $s \geq 1$, $\mu[z_i] \leq 0$ $(1 \leq i \leq m)$.

(iii) $I - AZ$ is regular whenever $Z = \mathrm{diag}(\zeta_1, \zeta_2, \ldots, \zeta_m)$ with $\zeta_i \in \mathbb{C}$, $\mathrm{Re}\ \zeta_i \leq 0$ $(1 \leq i \leq m)$.

(iv) All principal subdeterminants of A are zero, or all principal subdeterminants of A are nonnegative and $\mathrm{trace}\ (A) > 0$.

It is a conjecture that the above statements (i), (ii) and (iii) are also equivalent if $m > 2$.

By (4.3.3) it is shown that the condition on A imposed in (4.3.2) is not sufficient to ensure the existence and uniqueness of solutions to (4.1.5) in the general case $\mathbb{K}^s \neq \mathbb{R}^1$.

We note that the above results (4.3.2) and (4.3.3) can be extended to include arbitrary $\beta \in \mathbb{R}$.

4.4. Semi-implicit methods.

We now consider the case that a semi-implicit method (3.3.1) is used to approximate the solution U of $U'(t) = f(U(t))$ $(t \geq 0)$, $U(0) = u_0$. We assume that the approximation $J(\cdot) = J(\cdot; h, f)$ to f' satisfies

(4.4.1) $\qquad \mu[J(x)] \leq \beta \qquad$ (for all $x \in \mathbb{K}^s$).

Note that if the function f is continuously differentiable then the one-sided Lipschitz condition (4.1.2) is equivalent to $\mu[f'(x)] \leq \beta$ (for all $x \in \mathbb{K}^s$) - see lemma 2.3.3. The assumption (4.4.1) is therefore a natural one.

Let ψ stand for one of the coefficient functions a_{ij}, b_i of the semi-implicit method G. If we compute from a known $u_n \approx U(t_n)$ a consecutive approximation $u_{n+1} \approx U(t_n + h)$, we have to solve linear systems of the type

$$q(hJ(u_n))v = p(hJ(u_n))hf(y_j(u_n))$$

where $v \in \mathbb{K}^s$ is the unknown, $y_j(u_n)$ has been computed already, and p, q are polynomials such that $\psi(\zeta) = p(\zeta)/q(\zeta)$ $(\zeta \in \mathbb{C})$. Clearly this linear system has a unique solution iff $q(hJ(u_n))$ is regular, i.e.

$\psi(hJ(u_n))$ exists (cf. section 2.2.2). Using lemma 2.2.6 the question whether u_{n+1} is uniquely determined can be answered very easily.

COROLLARY 4.4.1. *Let* $h > 0$ *and* $\beta \in \mathbb{R}$ *be given. The linear systems of algebraic equations arising in* (3.3.1) *all have unique solutions whenever* $s \in \mathbb{N}$ *and* $J(\cdot): \mathbb{K}^s \to L(\mathbb{K}^s)$ *satisfies* (4.4.1) *iff all rational functions* a_{ij}, b_i *are analytic on* $\{\zeta : \zeta \in \mathbb{C}, \operatorname{Re} \zeta \leq h\beta\}$.

CHAPTER 5

CONTRACTIVITY AND ERROR PROPAGATION PER STEP

5.1. INTRODUCTION

In this chapter we consider again the initial value problem

(5.1.1.a) $U'(t) = f(U(t)) \quad (t \geq 0)$,

(5.1.1.b) $U(0) = u_0$,

where $f: \mathbb{K}^s \to \mathbb{K}^s$ and $u_0 \in \mathbb{K}^s$ are known. Further $\langle \cdot, \cdot \rangle$ will stand for an arbitrary inner product on $\mathbb{K}^s$, and $|x| = \langle x, x \rangle^{\frac{1}{2}}$ (for all $x \in \mathbb{K}^s$).

The one-step schemes for the numerical solution of (5.1.1) are written as

$$u_{n+1} = G(u_n; h, f) \qquad (n=0,1,2,\ldots) .$$

Here $h > 0$ is the stepsize and $u_n \approx U(nh)$ $(n=0,1,2,\ldots)$. We consider methods where $G(\cdot) = G(\cdot; h, f)$ is defined on some suitable subset of the $\mathbb{K}^s$ by

(5.1.2.a) $$G(x) = x + \sum_{i=1}^{m} b_i(hJ(x))\, hf(y_i(x)) ,$$

where the internal vectors $y_i(x) \in \mathbb{K}^s$ satisfy

(5.1.2.b) $$y_i(x) = x + \sum_{j=1}^{m} a_{ij}(hJ(x))\, hf(y_j(x)) \qquad (1 \leq i \leq m) .$$

The functions a_{ij} and b_i are rational functions with real coefficients and $J(\cdot) = J(\cdot; h, f): \mathbb{K}^s \to L(\mathbb{K}^s)$ is a given function. Generally we will have $J(x) \approx f'(x)$ $(x \in \mathbb{K}^s)$.

The semi-implicit methods of section 3.3 fit into the form (5.1.2) with $a_{ij} \equiv 0$ for $1 \leq i \leq j \leq m$. If all a_{ij} and b_i are constant,

(5.1.2) defines a Runge-Kutta method.

Let $\beta \in \mathbb{R}$, and let $\mathcal{D} \subset \mathbb{K}^s$ be open and convex. Assume f is continuous and satisfies the one-sided Lipschitz condition

$$\text{(5.1.3)} \qquad \operatorname{Re} \langle f(\tilde{x})-f(x), \tilde{x}-x \rangle \leq \beta |\tilde{x}-x|^2 \qquad \text{(for all } \tilde{x}, x \in \mathcal{D}\text{)} .$$

As is well known (see e.g. DAHLQUIST (1959)) (5.1.3) implies that for any two solutions $\tilde{U}, U$ of the differential equation (5.1.1.a) with trajectories in $\mathcal{D}$, we have

$$\text{(5.1.4)} \qquad |\tilde{U}(t+h)-U(t+h)| \leq e^{\beta h} |\tilde{U}(t)-U(t)|$$

for arbitrary $h > 0$ and $t \geq 0$. The reverse also holds. In (5.1.4) the Lipschitz constant of f is not involved. This makes the estimate useful for stiff problems. In particular if $\beta = 0$, we have for all $h > 0$ and $t \geq 0$,

$$\text{(5.1.5)} \qquad |\tilde{U}(t+h)-U(t+h)| \leq |\tilde{U}(t)-U(t)| .$$

We see that an error in the initial value u_0 will then not be amplified.

In this chapter we want to investigate in which cases the properties (5.1.4) and (5.1.5) carry over to the numerical approximations computed from a method (5.1.2). For this purpose we first take a look at the class of simple test problems where $\mathbb{K}^s = \mathbb{C}^1$ and f is linear,

$$\text{(5.1.6)} \qquad f(x) = \lambda x \quad \text{(for } x \in \mathbb{C}\text{)} \qquad \text{with} \quad \lambda \in \mathbb{C}, \ \operatorname{Re} \lambda \leq \beta .$$

This is a special case of (5.1.3). About the behaviour of numerical solutions for such simple functions f much is known already. We shall therefore not study this case for specific methods, but simply make assumptions about the numerical approximations for the problems (5.1.6).

The main question we consider in this chapter is the following.
To what extent do the conclusions about the numerical approximations which can be drawn for the simple test problems (5.1.6) *carry over to nonscalar and nonlinear differential equations satisfying* (5.1.3)*?*

We introduce some terminology for methods G given by (5.1.2).

DEFINITION 5.1.1. *Let* $f: \mathbb{K}^s \to \mathbb{K}^s$, *and let* $\mathcal{D}$ *be an arbitrary subset of* $\mathbb{K}^s$. *Assume* $h > 0$ *and* $G(\cdot;h,f)$ *is defined on* $\mathcal{D}$. *Method* G *is said to be* <u>*contractive*</u> *on* $\mathcal{D}$ *(for* f *and* h*) if*

$$|G(\tilde{x};h,f)-G(x;h,f)| \le |\tilde{x}-x| \qquad \text{(for all } \tilde{x},x \in \mathcal{D}\text{)}.$$

If the above holds for all $h > 0$ *the method* G *is called* <u>*unconditionally contractive*</u> *on* $\mathcal{D}$ *(for* f*).*

In the subsequence we will generally impose conditions on f such that $G(\cdot;h,f)$ is continuously differentiable at a given point of interest $x_0 \in \mathbb{K}^s$. For these cases we also consider

<u>DEFINITION 5.1.2.</u> *Let* $f: \mathbb{K}^s \to \mathbb{K}^s$ *and* $x_0 \in \mathbb{K}^s$. *Assume* $h > 0$, *and* $G(\cdot;h,f)$ *is defined on an open neighbourhood of* x_0 *and continuously differentiable at* x_0. *Method* G *is said to be* <u>*locally contractive*</u> *at* x_0 *(for* f *and* h*) if*

$$|G'(x_0;h,f)| \le 1 .$$

If this holds for all $h > 0$ *we call method* G <u>*unconditionally locally contractive*</u> *at* x_0 *(for* f*).*

From lemma 2.3.2 we obtain a result that gives a justification for the term "locally contractive".

<u>COROLLARY 5.1.3.</u> *Let* $f: \mathbb{K}^s \to \mathbb{K}^s$, $h > 0$, *and let* $\mathcal{D} \subset \mathbb{K}^s$ *be open and convex. Assume* $G(\cdot;h,f)$ *is defined and continuously differentiable on* $\mathcal{D}$. *Then* G *is contractive on* $\mathcal{D}$ *(for* f *and* h*) iff* G *is locally contractive at each point* $x_0 \in \mathcal{D}$ *(for* f *and* h*).*

If a method G is unconditionally contractive on $\mathcal{D}$ for f, then the analogue of (5.1.5) holds for the numerical approximations; if $\{\tilde{u}_n\}$, $\{u_n\}$ are two sequences in $\mathcal{D}$ computed from method G with starting vectors $\tilde{u}_0, u_0 \in \mathcal{D}$ and arbitrary $h > 0$, then

$$|\tilde{u}_{n+1}-u_{n+1}| \leq |\tilde{u}_n-u_n| \qquad (n=0,1,2,\ldots) \ .$$

With the above terminology we can reformulate the important concepts of A-stability and B-contractivity. A method G is A-stable if G is unconditionally contractive on $\mathbb{C}$ for all functions f satisfying (5.1.6) with $\beta = 0$. If G is unconditionally contractive on $\mathbb{K}^s$ for all f satisfying (5.1.3) with $\beta = 0$, G is said to be B-contractive.

In spite of the simplicity of the class of functions f (scalar and linear) involved in the definition of A-stability, numerical experience has shown that A-stable methods generally work rather well on much more complicated stiff problems. In the present chapter we shall mainly be concerned with methods which are A-stable but not B-contractive. For such methods conditions on dissipative functions f will be imposed which guarantee unconditional contractivity. Also conditional contractivity results will be presented where the restriction on the stepsize h is such that the product of h and the Lipschitz constant L of f may be arbitrary large.

We now give an outline for the rest of this chapter.

The subsequent analysis will be based on a result of section 5.2 where a useful expression for $G'(x;h,f)$ will be derived.

In section 5.3 we shall consider linear systems of differential equations. For such systems far reaching conclusions can be drawn from the behaviour of the methods for the simple scalar testproblems (5.1.6).

In section 5.4 semi-implicit methods are considered. We shall mainly deal with Rosenbrock methods $(J(x)=f'(x))$, and semi-implicit methods using a fixed Jacobian approximation. Although these methods are not B-contractive, it will be shown that they can be unconditionally contractive for dissipative functions f which, in some sense, do not differ too much from a linear function. For arbitrary nonlinear stiff systems conditional contractivity results are presented. Further it will be shown that these contractivity results do not hold for the semi-implicit methods with $J(x) = f'(x+chf(x))$, $c \neq 0$.

Finally, in section 5.5, we review some results on B-contractive implicit Runge-Kutta methods, and give a short analysis for a class of implicit Runge-Kutta methods which are not B-contractive.

5.2. PRELIMINARY RESULTS

Consider a method G defined by (5.1.2). Let $s \in \mathbb{N}$, $f: \mathbb{K}^s \to \mathbb{K}^s$ and $h > 0$ be given, and let $\mathcal{D} \subset \mathbb{K}^s$ be an open, convex set. We assume that $G(x)$ and the vectors $y_i(x)$ $(1 \le i \le m)$ are well defined by (5.1.2) for any point $x \in \mathcal{D}$ (i.e. the $a_{ij}(hJ(x))$ and $b_i(hJ(x))$ exist, and the system (5.1.2.b) has a unique solution).

Throughout this chapter the following notations will be used frequently. For a given $z \in L(\mathbb{K}^s)$ and $x \in \mathbb{K}^s$ we write

$$b(z)^T = (b_1(z), b_2(z), \ldots, b_m(z)) \in L(\mathbb{K}^{sm}, \mathbb{K}^s) ,$$

$$A(z) = \begin{pmatrix} a_{11}(z) & a_{12}(z) & \ldots & a_{1m}(z) \\ a_{21}(z) & & & \cdot \\ \vdots & & & \vdots \\ a_{m1}(z) & \ldots & & a_{mm}(z) \end{pmatrix} \in L(\mathbb{K}^{sm}) ,$$

$$y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_m(x) \end{pmatrix} \in \mathbb{K}^{sm} \quad \text{and} \quad F(y(x)) = \begin{pmatrix} f(y_1(x)) \\ f(y_2(x)) \\ \vdots \\ f(y_m(x)) \end{pmatrix} \in \mathbb{K}^{sm} .$$

Further e will stand for $e^{(m)} \otimes I^{(s)} \in L(\mathbb{K}^s, \mathbb{K}^{sm})$, where $e^{(m)} = (1,1,\ldots,1)^T \in \mathbb{R}^m$, $I^{(s)}$ is the $s \times s$ unit matrix, and $\otimes$ denotes the Kronecker product. Note that for any vector $v \in \mathbb{K}^s$ we have $ev = e^{(m)} \otimes v$.

With the above notation we can rewrite (5.1.2) as

(5.2.1.a) $$G(x) = x + b(z)^T hF(y(x)) ,$$

(5.2.1.b) $$y(x) = ex + A(z)\, hF(y(x)) ,$$

where z stands for $hJ(x)$.

For fixed $x_0 \in \mathcal{D}$ we use the notations

$$z_0 = hJ(x_0) , \quad z_i = hf'(y_i(x_0)) \qquad (1 \le i \le m) ,$$

$$Z = \operatorname{diag}(z_1, z_2, \ldots, z_m) \quad \text{and} \quad Z_0 = \operatorname{diag}(z_0, z_0, \ldots, z_0) \in L(\mathbb{K}^{sm}) .$$

Further we write for arbitrary $v \in \mathbb{K}^s$,

$$D_x b(z_0)^T v = (D_x b_1(hJ(x_0))v, D_x b_2(hJ(x_0))v, \ldots, D_x b_m(hJ(x_0))v) \in$$

$$\in L(\mathbb{K}^{sm}, \mathbb{K}^s) ,$$

where $D_x b_i(hJ(x_0))$ stands for the Gateaux-derivative of the function $b_i(hJ(\cdot))$ at x_0. In a similar way we write

$$D_x A(z_0)v \in L(\mathbb{K}^{sm}) \text{ for the matrix with block-entries}$$

$$D_x a_{ij}(hJ(x_0))v \in L(\mathbb{K}^s) \quad (1 \le i,j \le m) .$$

In the following theorems useful expressions will be presented for $G'(x_0) = G'(x_0;h,f)$. These expressions will be starting point for the analysis of the error propagation in the subsequent sections.

THEOREM 5.2.1. *Let* $\mathcal{D} \subset \mathbb{K}^s$ *be open and convex,* $h > 0$, *and* $x_0 \in \mathcal{D}$. *Assume* f *and* $J(\cdot) = J(\cdot;h,f)$ *are continuously differentiable on* $\mathcal{D}$, *and there exists an open neighbourhood* E *of* x_0 *such that* $G(x)$, $y_i(x)$ $(1 \le i \le m)$ *are well defined by* (5.2.1) *and the vectors* $y_i(x)$ $(1 \le i \le m)$ *depend continuously on* x *(for* $x \in E$*). Assume further that* $y_i(x_0) \in \mathcal{D}$ $(1 \le i \le m)$ *and* $I - A(z_0)Z$ *is regular. Then* $G(\cdot) = G(\cdot;h,f)$ *is continuously differentiable at* x_0, *and for any* $v \in \mathbb{K}^s$ *we have*

$$(5.2.2) \qquad G'(x_0)v = v + b(z_0)^T Z(I-A(z_0)Z)^{-1} ev +$$

$$+ [D_x b(z_0)^T v] hF(y(x_0)) +$$

$$+ b(z_0)^T Z(I-A(z_0)Z)^{-1} [D_x A(z_0)v] hF(x_0)) .$$

PROOF. Let $v \in \mathbb{K}^s$ and let $t > 0$ be small enough to have $\tilde{x}_0 = x_0 + tv \in \mathcal{D} \cap E$ and $y_i(\tilde{x}_0) \in \mathcal{D}$ $(1 \le i \le m)$. We put $\tilde{z}_0 = hJ(\tilde{x}_0)$,

$$\tilde{z}_i = \int_0^1 hf'(y_i(x_0) + \tau(y_i(\tilde{x}_0) - y_i(x_0)))d\tau \quad (1 \le i \le m) ,$$

and $\tilde{Z} = \mathrm{diag}(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_m)$.

In view of the mean-value theorem 2.3.1 we have

$$hF(y(\tilde{x}_0)) - hF(y(x_0)) = \tilde{Z}[y(\tilde{x}_0)-y(x_0)] .$$

Using (5.2.1.b) we thus obtain

$$y(\tilde{x}_0) - y(x_0) = e[\tilde{x}_0-x_0] + A(z_0)\,\tilde{Z}[y(\tilde{x}_0)-y(x_0)] +$$

$$+ [A(\tilde{z}_0)-A(z_0)]\,hF(y(\tilde{x}_0)) .$$

Since $\tilde{Z} \to Z$ for $t \downarrow 0$, $I - A(z_0)\tilde{Z}$ is regular for $t > 0$ sufficiently small. We then have

$$(5.2.3) \qquad y(\tilde{x}_0) - y(x_0) = (I-A(z_0)\tilde{Z})^{-1}\,(e[\tilde{x}_0-x_0]+[A(\tilde{z}_0)-A(z_0)]hF(y(\tilde{x}_0))) .$$

Further we obtain from (5.2.1.a),

$$(5.2.4) \qquad G(\tilde{x}_0) - G(x_0) = [\tilde{x}_0-x_0] + b(z_0)^T\,\tilde{Z}[y(\tilde{x}_0)-y(x_0)] +$$

$$+ [b(\tilde{z}_0)^T-b(z_0)^T]\,hF(y(\tilde{x}_0)) .$$

By using (5.2.3) and the relation

$$A(\tilde{z}_0) - A(z_0) = \int_0^1 D_x A(hJ(x_0+\tau tv))tv\, d\tau ,$$

we see that for $t \downarrow 0$, $\frac{1}{t}[y(x_0+tv)-y(x_0)]$ converges to

$$(5.2.5) \qquad (I-A(z_0)Z)^{-1}\,(ev+[D_xA(z_0)v]\,hF(y(x_0))) .$$

Hence y is Gateaux-differentiable at x_0, and $y'(x_0)v$ is given by (5.2.5). It follows that the derivative is a continuous function at x_0.

In a similar way, by using (5.2.4), it can be shown that G is continuously differentiable at x_0, and (5.2.2) holds. □

THEOREM 5.2.2. *Assume that the conditions of theorem* 5.2.1 *hold, and* $I - A(z_0)Z_0$ *is regular. Then we have for any* $v \in \mathbb{K}^s$,

$$(5.2.6) \qquad G'(x_0)v = \phi(z_0)v + b(z_0)^T\,(I-A(z_0)Z_0)^{-1}\,(Z-Z_0)\,(I-A(z_0)Z)^{-1}ev +$$

$$+ [D_xb(z_0)^Tv]\,hF(y(x_0)) + b(z_0)^TZ(I-A(z_0)Z)^{-1}\,[D_xA(z_0)v]\,hF(y(x_0)) ,$$

where

$$\phi(z_0) = I + b(z_0)^T Z_0 (I-A(z_0)Z_0)^{-1} e \in L(\mathbb{K}^s) . \tag{5.2.7}$$

PROOF. If ψ_1 and ψ_2 are two rational functions, then $\psi_1(z_0)$ and $\psi_2(z_0)$ will commute. Using this fact it is easily seen that

$$A(z_0)Z_0 = Z_0\, A(z_0) .$$

Further we have

$$Z(I-A(z_0)Z)^{-1} - (I-Z_0A(z_0))^{-1}Z_0 =$$

$$= (I-Z_0A(z_0))^{-1}\,(Z-Z_0)\,(I-A(z_0)Z)^{-1} ,$$

$$Z_0(I-A(z_0)Z_0)^{-1} - (I-Z_0A(z_0))^{-1}Z_0 = 0 .$$

From these relations and (5.2.2) the proof easily follows. □

In formula (5.2.6) several contributions to $G'(x_0)v$ can be seen. First there is the part $\phi(z_0)v$ which is present even if f is linear. The second part is $b(z_0)^T (I-A(z_0)Z_0)^{-1} (Z-Z_0) (I-A(z_0)Z)^{-1}ev$. This part vanishes if $Z = Z_0$, i.e. $f'(y_i(x_0)) = J(x_0)$ $(1\le i\le m)$, which is the case if f is affine and $J \equiv f'$. The remaining contribution is zero if $D_x b(z_0)^T = 0$ and $D_x A(z_0) = 0$. This happens if we are dealing with an ordinary Runge-Kutta method, or if J is constant (e.g. if f is affine, $J \equiv f'$). For the case that f is affine we therefore obtain the following result.

THEOREM 5.2.3. *Let* $f(x) = \Lambda x + w$ *and* $J(x) = \Lambda$ *(for all* $x \in \mathbb{K}^s$*),* *with* $\Lambda \in L(\mathbb{K}^s)$ *and* $w \in \mathbb{K}^s$. *Put* $z_0 = h\Lambda$. *Assume all* $a_{ij}(z_0)$ *and* $b_i(z_0)$ $(1\le i,j\le m)$ *exist, and* $I - A(z_0)Z_0$ *is regular. Then* $G(x;h,f)$ *is defined for all* $x \in \mathbb{K}^s$, *and*

$$G(\tilde{x};h,f) - G(x;h,f) = \phi(z_0)\,(\tilde{x}-x) \quad \text{(for all } \tilde{x},x \in \mathbb{K}^s) ,$$

where $\phi(z_0)$ *is given by* (5.2.7).

PROOF. From the assumptions it can easily be seen that $G(x)$ and $y(x)$ are well defined by (5.1.2) and $y(\tilde{x}) - y(x) = (I-A(z_0)Z_0)^{-1} e[\tilde{x}-x]$ for all $\tilde{x},x \in \mathbb{K}^s$ (cf. (5.2.3)). Thus $y(x)$ depends continuously on x (for any $x \in \mathbb{K}^s$).

Application of theorem 5.2.2, with $\mathcal{D} = \mathbb{K}^s$, yields

$$G'(x) = \phi(z_0) \qquad \text{(for all } x \in \mathbb{K}^s\text{)} .$$

The proof now follows from the mean-value theorem 2.3.1. □

COROLLARY 5.2.4. *Suppose the conditions of theorem* 5.2.3 *hold with* $w = 0$. *Then* $G(x;h,f) = \phi(z_0)x$ *(for all* $x \in \mathbb{K}^s$*).*

PROOF. The proof follows from theorem 5.2.3 by noticing that $G(0;h,f) = 0$, or, more directly, from (5.2.1.a) and (5.2.1.b). □

In the following lemma it will be shown that $\phi(z_0)$ is a rational expression in z_0. This is a generalization of a result obtained by STETTER (1973; pp. 132,152) for Runge-Kutta methods.

LEMMA 5.2.5. *Assume* $z_0 \in L(\mathbb{K}^s)$ *is such that all* $a_{ij}(z_0)$ *and* $b_i(z_0)$ $(1 \le i,j \le m)$ *exist, and* $I - A(z_0)Z_0$ *is regular. Let* $\phi(z_0)$ *be defined by* (5.2.7), *and let the rational function* ψ *be defined by*

$$(5.2.8) \qquad \psi(\zeta) = 1 + b(\zeta)^T \zeta(I-A(\zeta)\zeta)^{-1} e^{(m)}$$

(for all $\zeta \in \mathbb{C}$ *such that* $a_{ij}(\zeta)$, $b_i(\zeta)$ $(1 \le i,j \le m)$ *are regular, and* $I - A(\zeta)\zeta$ *is invertible). Then* $\psi(z_0)$ *exists, and* $\psi(z_0) = \phi(z_0)$.

PROOF. For all appropriate ζ we have

$$\psi(\zeta) = 1 + \sum_{i,j=1}^{m} b_i(\zeta)\zeta\, w_{ij}(\zeta) ,$$

where the $w_{ij}(\zeta)$ are the entries of $W(\zeta) = (I-A(\zeta)\zeta)^{-1}$. By using lemma 2.4.6 we see that $w_{ij}(z_0)$ is the i,j-th. block of $(I-A(z_0)Z_0)^{-1}$. It follows that $\psi(z_0) \in L(\mathbb{K}^s)$ exists, and

$$\psi(z_0) = I + \sum_{i,j=1}^{m} b_i(z_0) z_0 w_{ij}(z_0) .$$

Hence $\psi(z_0) = I + b(z_0)^T Z_0 (I - A(z_0)Z_0)^{-1} e$. □

In view of the above, the function defined by (5.2.8) will also be denoted by ϕ . This rational function is called the *stability function* of the method G .

5.3 LINEAR DIFFERENTIAL EQUATIONS

In this section we will assume that the Jacobian approximation J , which is used in (5.1.2), is such that $J(x) = f'(x)$ (for all $x \in \mathbb{K}^S$) if f is an affine function (i.e. f' is constant).

Suppose

$$f(x) = \Lambda x \qquad (\text{for } x \in \mathbb{K}^S)$$

where $\Lambda \in L(\mathbb{K}^S)$. In section 5.2 we have seen that the methods G given by (5.1.2) are such that

$$(5.3.1) \qquad G(x;h,f) = \phi(h\Lambda)x \qquad (\text{for } x \in \mathbb{K}^S,\ h > 0)$$

where ϕ is a rational function with real coefficients, the stability function of the method. Methods with this property are said to have a *rational structure* (cf. SPIJKER (1982 B)).

For a given stability function ϕ we define $\Phi: \mathbb{R} \to \bar{\mathbb{R}}$ by

$$(5.3.2) \qquad \Phi(t) = \sup\{|\phi(\zeta)| : \zeta \in \mathbb{C}, \operatorname{Re}\zeta \le t, \phi \text{ is regular at } \zeta\} \qquad (t \in \bar{\mathbb{R}}) .$$

With this notation a method G, with stability function ϕ , is A-stable iff $\Phi(0) \le 1$. Note that ϕ is an approximation to the exponential function: $\phi(\zeta) = e^{\zeta} + O(\zeta^{p+1})$ $(\zeta \to 0)$, where p is the order of the method. If $p \ge 0$ we therefore have $\Phi(0) \ge 1$.

REMARK 5.3.1. If the method G is explicit (all a_{ij} , b_i are polynomials and $a_{ij} \equiv 0$ $(1 \le i \le j \le m)$) and has order ≥ 1 , ϕ is a nonconstant polynomial. Therefore such a method cannot be A-stable.

DEFINITION 5.3.2. *Let ϕ be the stability function of method G . If $\Phi(t) < 1$ (for all $t < 0$), G is called strongly A-stable. In case G is*

A-*stable and* $|\phi(\infty)| = 0$, G *is said to be* L-*stable*.

Note that if a method G is A-stable, the stability function ϕ has no essential singularties in $\mathbb{C}^-$. From the maximum modulus theorem it therefore follows that G is strongly A-stable iff G is A-stable and $|\phi(\infty)| < 1$.

REMARK 5.3.3. Suppose $\mathbb{K}^s = \mathbb{C}^1$, $h > 0$ and $f(x) = \lambda x$ (for $x \in \mathbb{C}$) with $\lambda \in \mathbb{C}$, $\mathrm{Re}\ \lambda \leq \beta$. Here $\beta \in \mathbb{R}$ is a given number.

Let $\tilde{u}_0, u_0 \in \mathbb{C}$ be arbitrary. Consider the solutions $\tilde{U}, U$ of the differential equation $U'(t) = \lambda U(t)$ $(t \geq 0)$ with initial values $\tilde{u}_0, u_0$, respectively. Then

$$|\tilde{U}(t_1) - U(t_1)| = |e^{h\lambda}|\ |\tilde{u}_0 - u_0| \leq e^{h\beta}\ |\tilde{u}_0 - u_0| .$$

For the numerical approximations $\tilde{u}_1 = G(\tilde{u}_0; h, f)$, $u_1 = G(u_0; h, f)$ we have in view of (5.3.1),

$$|\tilde{u}_1 - u_1| = |\phi(h\lambda)|\ |\tilde{u}_0 - u_0| \leq \Phi(h\beta)\ |\tilde{u}_0 - u_0|$$

(provided that $\phi(h\lambda)$ is defined).

Suppose $\beta < 0$. Then $e^{h\beta} < 1$, and therefore the "error" $\tilde{u}_0 - u_0$ is damped out by the differential equation. If the numerical method G is strongly A-stable, the numerical approximations will show a similar behaviour. For an L-stable method this damping out is strong if $h\beta << 0$, becauce we have for such a method $\Phi(h\beta) \to 0$ (if $h\beta \to -\infty$).

There is however a drawback with strongly A-stable (and L-stable) methods: they may be too stable. Since we have $|\phi(\infty)| < 1$, there is a $\lambda > 0$ such that $|\phi(h\lambda)| < 1$ whereas $|e^{h\lambda}| > 1$. For this λ the numerical scheme is thus contractive although the difference between two exact solutions of the differential equation increases. Consequences of this too stable behaviour have been investigated by LINDBERG (1974).

The following result shows that the conclusions about the error propagation in the numerical schemes that can be drawn from the one-dimensional linear testequation (5.1.6) carry over to arbitrary systems of linear differential equations with constant coefficients. For this it is essential that we deal with a norm which is generated by an inner product (see e.g. SPIJKER (1982 B)).

COROLLARY 5.3.4. *Let* G *be a method of the type* (5.1.2), *and let* $\beta \in \mathbb{R}$ *and* $h > 0$. *Assume* $f(x) = \Lambda x + w$ *and* $J(x) = \Lambda$ *(for all* $x \in \mathbb{K}^s$ *)* *where* $w \in \mathbb{K}^s$ *and* $\Lambda \in L(\mathbb{K}^s)$ *satisfies* $\mu[\Lambda] \leq \beta$. Then

$$|G(\tilde{x};h,f)-G(x;h,f)| \leq \Phi(h\beta)\ |\tilde{x}-x|$$

whenever $G(\tilde{x};h,f)$ *and* $G(x;h,f)$ *are defined.*

The proof of this corollary follows directly from the theorems 5.2.3, 2.2.7 and lemma 5.2.5. We note that if all a_{ij} , b_i $(1\leq i,j\leq m)$ are analytic on $\{\zeta:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta\leq h\beta\}$, and $I - A(\zeta)\zeta$ is regular for all $\zeta \in \mathbb{C}$ with $\mathrm{Re}\ \zeta \leq h\beta$, then it follows from the lemmata 2.2.6 and 2.4.6 that $G(x;h,f)$ is defined for all $x \in \mathbb{K}^s$ in the above corollary.

In particular we get for dissipative linear differential equations with constant coefficients the following result.

COROLLARY 5.3.5. *Let* G *be an A-stable method of the type* (5.1.2). *Let* $f(x) = \Lambda x + w$ *and* $J(x) = \Lambda$ *(for all* $x \in \mathbb{K}^s$ *) where* $w \in \mathbb{K}^s$ *and* $\Lambda \in L(\mathbb{K}^s)$ *with* $\mu[\Lambda] \leq 0$. *Then any two sequences of numerical approximations* $\{\tilde{u}_n\},\{u_n\}$ *computed from this method with stepsize* $h > 0$ *and starting vectors* $\tilde{u}_0, u_0$, *respectively, satisfy the contractivity relation*

$$|\tilde{u}_{n+1}-u_{n+1}| \leq |\tilde{u}_n-u_n| \qquad (n=0,1,2,\ldots)\ .$$

REMARK 5.3.6. The nonlinear methods of LAMBERT (1974) and WAMBECQ (1978), which do not fit into the form (5.1.2), have no rational structure (although such methods are often called rational methods). For these methods the nice conclusions of the corollaries 5.3.4 and 5.3.5 cannot be drawn.

5.4. SEMI-IMPLICIT METHODS

5.4.1. Negative B-contractivity results.

In this section we consider the class of semi-implicit methods (3.3.1). We shall mainly deal with the Rosenbrock methods and semi-implicit methods using a fixed Jacobian approximation. To begin with it will be shown that these methods cannot be B-contractive.

The following theorem was proved by VANSELOV (1979) for a bit more

restricted class of methods. We note that by SANDBERG and SHICHMAN (1968) it was shown before that the simple Rosenbrock method given by $G(x;h,f) = x + (I-hf'(x))^{-1} hf(x)$ is not B-contractive. The proof presented here is a straightforward generalization of Vanselov's proof.

In the following theorems we will assume that the order of G is at least one, to exclude the trivial method $G(x;h,f) = x$ (for all $x \in \mathbb{K}^s$ ($s \in \mathbb{N}$), $h > 0$, $f: \mathbb{K}^s \to \mathbb{K}^s$).

THEOREM 5.4.1. *Let* G *be a Rosenbrock method of the type* (3.3.1) *with order* ≥ 1. *Then* G *is not* B-*contractive.*

PROOF. Beside the Rosenbrock method G given by

$$G(x;h,f) = x + \sum_{i=1}^{m} b_i(hf'(x))\, hf(y_i(x)) ,$$

$$y_i(x) = x + \sum_{j=1}^{i-1} a_{ij}(hf'(x))\, hf(y_j(x)) \qquad (1 \leq i \leq m) ,$$

we also consider the explicit Runge-Kutta method $\tilde{G}$

$$\tilde{G}(x;h,f) = x + \sum_{i=1}^{m} \beta_i\, hf(\tilde{y}_i(x)) ,$$

$$\tilde{y}_i(x) = x + \sum_{j=1}^{i-1} \alpha_{ij}\, hf(\tilde{y}_j(x)) \qquad (1 \leq i \leq m) ,$$

with $\alpha_{ij} = a_{ij}(0)$ $(1 \leq j < i \leq m)$, $\beta_i = b_i(0)$ $(1 \leq i \leq m)$. Note that since the order of G is at least one the a_{ij} and b_i can be defined in 0 (no essential singularities), and we have $\beta_1 + \beta_2 + \ldots + \beta_m = 1$ (see e.g. VAN DER HOUWEN (1977)). Therefore the order of $\tilde{G}$ is also at least one, and the stability function $\tilde{\phi}$ of $\tilde{G}$ is a nonconstant polynomial (cf. remark 5.3.1). It follows that there are numbers $\beta < 0$, $h > 0$ and $x_0 \neq 0$ such that, for $f(x) = \beta x$ ($x \in \mathbb{R}$),

$$|\tilde{G}(x_0;h,f)| > |x_0| .$$

Let $\varepsilon > 0$ be such that $\varepsilon < |x_0|$ and $\varepsilon < |\tilde{y}_i(x_0) - x_0|$ whenever $1 \leq i \leq m$, $\tilde{y}_i(x_0) \neq x_0$. We take

$$\tilde{f}(x) = f(x) + g(x) \qquad (x \in \mathbb{R}),$$

where g is a differentiable real function such that $g(x_0) = 0$, $g'(x_0) = -\beta$, $g(x) = 0$ (for $|x-x_0| > \varepsilon$), $|g'(x)| \le \beta$ (for all $x \in \mathbb{R}$). Then

$$G(x_0;h,\tilde{f}) = \tilde{G}(x_0;h,f) \ , \ G(0;h,\tilde{f}) = 0 \ .$$

Therefore

$$|G(x_0;h,\tilde{f})-G(0;h,\tilde{f})| > |x_0|$$

although $\tilde{f}$ is dissipative. □

THEOREM 5.4.2. *Let* G *be a semi-implicit method of the type* (3.3.1) *with a constant Jacobian approximation* J. *Suppose the order of* G *is at least one. Then* G *is not* B*-contractive.*

PROOF. Consider the dissipative function $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = -x^3$ ($x \in \mathbb{R}$). Assume $J(x) = J_0 \in \mathbb{R}$ (for all $x \in \mathbb{R}$), and let $h > 0$ be arbitrary. Using (3.3.1) it is easily seen that $G(x) = G(x;h,f)$ is a polynomial in x, and $G(x) = x + O(x^2)$ ($x \to 0$). Since the order of G is at least one, $G(x)$ is not identically equal to x, and therefore the degree of $G(x)$ is larger than one. Hence

$$|G(x;h,f)-G(0;h,f)| > |x|$$

for $x \in \mathbb{R}$ sufficiently large. □

Within the B-contractivity framework, i.e. we consider nonlinear functions f satisfying the one-sided Lipschitz condition (5.1.3), contractivity results and results on the error propagation per step for semi-implicit methods will be derived in the subsequent sections. There we will impose some additional conditions on the functions f. Some results in this direction can be found in the papers of SANDBERG and SHICHMAN (1968), TRIGIANTE (1977) and HUNDSDORFER (1981), for some simple Rosenbrock methods (covered by the methods considered in example 5.4.20).

For rather general semi-implicit methods using a fixed Jacobian

approximation J and functions f which are known to be of the type

$$f(x) = Jx + g(x) \qquad (\text{for } x \in \mathbb{K}^s),$$

where g is a function with a small Lipschitz constant on the $\mathbb{K}^s$, nice results have been obtained by HAIRER, BADER and LUBICH (1982) and, following the same line, by STREHMEL and WEINER (1982 A). These results are not valid for semi-implicit methods where $J(x)$ may be varying on the $\mathbb{K}^s$, such as the Rosenbrock methods.

A different approach was followed by VERWER (1977), who considered the general semi-implicit methods (3.3.1) for the nonautonomous, 1-dimensional, linear S-stability model problem

$$f(t,x) = \lambda(x-g(t)) + g'(t) \qquad (t \in \mathbb{R},\ x \in \mathbb{C})$$

with $\lambda \in \mathbb{C}$ and $g: \mathbb{R} \to \mathbb{C}$ a smooth function. Similar investigations have been carried out afterwards by STREHMEL (1981), STREHMEL and WEINER (1982 B). Although this model problem is more general than the A-stability model problem (5.1.6), it is still rather restricted (see e.g. remark 5.4.28).

In VERWER (1982) a more general nonlinear model problem was considered, consisting of two coupled differential equations of the singularly perturbed type. This system contains a small parameter ε, and if ε tends to zero the stiffness is more and more increased. Verwer was concerned with certain accuracy and boundedness properties which hold uniformly in ε. A similar boundedness property was considered by VAN VELDHUIZEN (1973,1974, 1981,1983) for a 2-dimensional linear nonautonomous model problem (the D-stability model problem).

These results of van Veldhuizen and Verwer are not covered by the results in this chapter. For instance the so-called *internal* A-stability concept of VERWER (1977), which was shown to be useful through numerical experiments, hardly shows up in our results (cf. example 5.4.11). The more restriced model problems seem to be better suited to detect differences within the class of Rosenbrock methods. The investigation carried out in this chapter is more devoted to the question what extra conditions on the nonlinear functions f satisfying (5.1.3) have to be imposed to ensure a favourable error propagation, for instance contractivity.

5.4.2. An upper bound for $|G'(x;h,f)|$.

In view of the negative results of the previous section (the theorems 5.4.1, 5.4.2) there is a need for a theory that shows for what kind of nonlinear functions f the error propagation will be favourable. In this section an upper bound for $|G'(x;h,f)|$ will be given. Using this upper bound positive results will be obtained in the subsequent sections.

We consider a semi-implicit method G of the type (3.3.1) satisfying (3.3.5) and (3.3.9). Let θ be the largest number such that all a_{ij} and b_i are analytic on $\{\zeta:\zeta\in\mathbb{C},\mathrm{Re}\ \zeta\leq\theta^{-1}\}$. From (3.3.5) it follows that $\theta > 0$.

With the notation introduced in section 5.2 the method G is given by

$$G(x_0;h,f) = x_0 + b(z_0)^T\, hF(y(x_0)) ,$$

$$y(x_0) = ex_0 + A(z_0)\, hF(y(x_0))$$

(for $x_0 \in \mathbb{K}^s$ with $h > 0$, $f: \mathbb{K}^s \to \mathbb{K}^s$, $s \in \mathbb{N}$). Here $y(x_0) = (y_1(x_0)^T,\ldots,y_m(x_0)^T)^T$, $F(y(x_0)) = (f(y_1(x_0))^T, f(y_2(x_0))^T,\ldots,f(y_m(x_0))^T)^T$ and z_0 stands for $hJ(x_0)$. As in section 5.2, $hf'(y_i(x_0))$ will be denoted by z_i $(1\leq i\leq m)$, $Z = \mathrm{diag}(z_1,z_2,\ldots,z_m)$, and $Z_0 = \mathrm{diag}(z_0,z_0,\ldots,z_0) \in \in L(\mathbb{K}^{sm})$.

Let $\beta \in \mathbb{R}$ and let $\alpha,\gamma,\delta,\varepsilon : \mathbb{R}^+ \to \mathbb{R}$ be given functions. For notational convenience we put

$$\Delta(t,\xi) = \delta(t) + \varepsilon(t)\xi \qquad (\text{for } t,\xi \in \mathbb{R}^+).$$

Further we define

(5.4.1) $\quad h_0 = (2\theta\beta)^{-1}$ if $\beta > 0$, $h_0 = \infty$ if $\beta \leq 0$.

Let $h \in (0,h_0)$. On the functions f and $J(\cdot) = J(\cdot;h,f)$ the following assumptions (5.4.2)-(5.4.7) will be made. This set of assumptions will be denoted by (A).

(5.4.2) $\quad |\cdot|$ is an inner product-norm on $\mathbb{K}^s$, $s \in \mathbb{N}$, $x_0 \in \mathbb{K}^s$, $r_0 > 0$, and $\mathcal{D}_0 = \{x:x\in\mathbb{K}^s,|x-x_0|<r_0\}$.

(5.4.3) $f: \mathbb{K}^s \to \mathbb{K}^s$ and $J: \mathbb{K}^s \to L(\mathbb{K}^s)$ are continuously differentiable on $\mathcal{D}_0$.

(5.4.4) $|(I-h\theta J(x_0))^{-1} hf(x_0)| \leq \alpha(h)$.

(5.4.5) $\mu[J(x_0)] \leq \beta$.

(5.4.6) $|(I-h\theta J(x_0))^{-1} hJ'(x_0)| \leq \gamma(h)$.

(5.4.7) $|(I-h\theta J(x_0))^{-1} h(J(x_0)-f'(x))| \leq \delta(h) + \varepsilon(h)\ |x-x_0|$
(for all $x \in \mathcal{D}_0$) .

In the applications in the subsequent sections the assumptions (5.4.4), (5.4.6) and (5.4.7) will be replaced by more simple, concrete ones, which are however more restrictive.

From section 4.4 it can be seen that the assumptions $h \in (0,h_0)$, (5.4.2) and (5.4.5) together imply that $G(x_0;h,f)$ is defined. From the same assumptions and the corollaries 2.2.8 and 2.2.9 it also follows that there is a constant $\omega > 0$, only depending on the coefficients of the method G , such that

$$(5.4.8)\qquad \max_{1\leq i,j\leq m} \{|a_{ij}(z_0)|,|b_i(z_0)|\} \leq \omega(1-h\theta\beta)^{-1} \text{ , and}$$

$$\max_{1\leq i,j\leq m} \{|a_{ij}(z_0)\theta z_0|,|a_{ij}(z_0)(I-\theta z_0)|,$$

$$|b_i(z_0)\theta z_0|,|b_i(z_0)(I-\theta z_0)|\} \leq \omega .$$

For $h \in (0,h_0)$ we define recursively for $i = 1,2,\ldots,m$,

$$(5.4.9.a)\qquad \rho_i(h) = \omega \sum_{j=1}^{i-1} \sigma_j(h) ,$$

$$(5.4.9.b)\qquad \sigma_i(h) = \alpha(h) + \theta^{-1} \rho_i(h) + \Delta(h,\rho_i(h))\ \rho_i(h) ,$$

$$(5.4.9.c)\qquad \tau_i(h) = \theta^{-1}\omega + \omega\ \Delta(h,\rho_i(h)) .$$

LEMMA 5.4.3. *Suppose the semi-implicit method* G *satisfies* (3.3.5) *and* (3.3.9). *Suppose further that* $h \in (0,h_0)$, (A) , $\rho_m(h) < r_0$. *Then all* $y_i(x_0) \in \mathcal{D}_0$, *and*

$$|y_i(x_0)-x_0| \le \rho_i(h) \qquad (1\le i\le m) ,$$

$$|(I-\theta z_0)^{-1} hf(y_i(x_0))| \le \sigma_i(h) \qquad (1\le i\le m) .$$

PROOF. The vectors $y_i = y_i(x_0)$ $(1\le i\le m)$ satisfy

$$(5.4.10) \qquad y_i - x_0 = \sum_{j=1}^{i-1} a_{ij}(z_0)\ (I-\theta z_0)\ (I-\theta z_0)^{-1}\ hf(y_j) .$$

If $y_i \in \mathcal{D}_0$ we obtain by Taylor expansion

$$hf(y_i) = hf(x_0) + \int_0^1 hf'(x_0+t(y_i-x_0))\ dt(y_i-x_0) =$$
$$= hf(x_0) + z_0(y_i-x_0) + \int_0^1 [hf'(x_0+t(y_i-x_0))-z_0]\ dt(y_i-x_0) .$$

Hence

$$(5.4.11) \qquad (I-\theta z_0)^{-1} hf(y_i) = (I-\theta z_0)^{-1} hf(x_0) +$$
$$+ \sum_{j=1}^{i-1} a_{ij}(z_0) z_0\ (I-\theta z_0)^{-1}\ hf(y_j) +$$
$$+ \int_0^1 (I-\theta z_0)^{-1}\ [hf'(x_0+t(y_i-x_0))-z_0]dt\ (y_i-x_0) .$$

The assertion of the lemma clearly holds for $i = 1$, since $y_1 = x_0$. Using (5.4.8)-(5.4.11) and $\rho_j(h) \le \rho_m(h)$ $(1\le j\le m)$, the assertion is easily proved for $i = 2,3,\ldots,m$ by induction. □

COROLLARY 5.4.4. *Under the assumptions of lemma* 5.4.3 *it follows that* $G(\cdot;h,f)$ *is continuously differentiable at* x_0.

PROOF. From the assumptions $h \in (0,h_0)$, (5.4.2), (5.4.3) and (5.4.5), and lemma 5.4.3, it can be seen that there exists an open, convex region $E \subset \mathcal{D}_0$ containing x_0 such that the following holds: $y_1(x),y_2(x),\ldots,y_m(x)$, $G(x)$ are defined and depend continuously on x (for all $x \in E$).

The proof is now an immediate consequence of lemma 5.4.3 and theorem 5.2.1. □

From lemma 5.4.3 and (5.4.7) we obtain

COROLLARY 5.4.5. *Under the assumptions of lemma* 5.4.3, *we have*

$$|(I-\theta z_0)^{-1}(z_i-z_0)| \le \Delta(h,\rho_i(h)) \qquad (1\le i\le m) .$$

LEMMA 5.4.6. *Let the assumptions of lemma* 5.4.3 *hold. Then*

$$|a_{ij}(z_0)z_j| \le \tau_j(h) , \ |b_i(z_0)z_i| \le \tau_i(h) \quad (1\le i,j\le m) .$$

For the i,j*-th.* s × s *block* $[(I-A(z_0)Z)^{-1}]_{ij}$ *of the* sm × sm *matrix* $(I-A(z_0)Z)^{-1}$ *we have*

$$|[(I-A(z_0)Z)^{-1}]_{ij}| \le \prod_{k=j}^{m-1} (1+\tau_k(h)) \qquad (1\le i,j\le m) .$$

PROOF. We have

$$|a_{ij}(z_0)z_j| \le |a_{ij}(z_0)z_0| + |a_{ij}(z_0)(I-\theta z_0)|\,|(I-\theta z_0)^{-1}(z_j-z_0)| .$$

Using (5.4.8), (5.4.9) and corollary 5.4.5 we thus get $|a_{ij}(z_0)z_j| \le$ $\le \tau_j(h)$. In the same way it follows that $|b_i(z_0)z_i| \le \tau_i(h)$.

From lemma 2.4.9 we see that

$$|[(I-A(z_0)Z)^{-1}]_{ij}| \le \tau_j(h) \prod_{k=j+1}^{i-1} (1+\tau_k(h)) \quad \text{if} \quad i > j ,$$

whereas $[(I-A(z_0)Z)^{-1}]_{ij}$ equals I if i = j , and 0 if i < j . The upper bound of the lemma thus follows. □

Using the above lemmata and the expression for $G'(x_0;h,f)$ given in theorem 5.2.2, we are now able to give an upper bound for $|G'(x_0;h,f)|$.

THEOREM 5.4.7. *Suppose the semi-implicit method* G *satisfies* (3.3.5) *and* (3.3.9), *and* $h \in (0,h_0)$, (A) , $\rho_m(h) < r_0$. *Let* ϕ *be the stability function of* G , *and let* Φ *be defined by* (5.3.2). *There are polynomials (in two variables)* P,Q *and* R , *with nonnegative coefficients which only depend on the coefficients of the method* G , *such that*

(5.4.12) $$|G'(x_0;h,f)| \le \Phi(h\beta) + \delta(h)\ P(\alpha(h)\varepsilon(h),\delta(h)) +$$

$$+ \alpha(h)\ \varepsilon(h)\ Q(\alpha(h)\varepsilon(h),\delta(h)) + \alpha(h)\ \gamma(h)\ R(\alpha(h)\varepsilon(h),\delta(h)) .$$

REMARK 5.4.8. In the following proof it will also be shown that there is a polynomial S, only depending on the coefficients of the method G, such that the condition

$$\rho_m(h) < r_0$$

can be written as

(5.4.13) $$\alpha(h)\ \ S(\alpha(h)\varepsilon(h),\delta(h)) < r_0 .$$

PROOF (of theorem 5.4.7 and remark 5.4.8).

We start by estimating the terms appearing on the right-hand side of the expression (5.2.6) for $G'(x_0;h,f)$. We will use the lemmata 5.4.3, 5.4.6, and corollary 5.4.5.

We have

$$|b(z_0)^T(I-A(z_0)Z_0)^{-1}(Z-Z_0)(I-A(z_0)Z)^{-1}e| =$$

$$= \Big|\sum_{i,j,k=1}^{m} b_i(z_0)[(I-A(z_0)Z_0)^{-1}]_{ij}(z_j-z_0)[(I-A(z_0)Z)^{-1}]_{jk}\Big| =$$

$$= \Big|\sum_{i,j,k=1}^{m} b_i(z_0)(I-\theta z_0)[(I-A(z_0)Z_0)^{-1}]_{ij}$$

$$(I-\theta z_0)^{-1}(z_j-z_0)[(I-A(z_0)Z)^{-1}]_{jk}\Big| \le$$

$$\le m^3 \max_{1\le i\le m} |b_i(z_0)(I-\theta z_0)| \max_{1\le i,j\le m} |[(I-A(z_0)Z_0)^{-1}]_{ij}|$$

$$\max_{1\le j\le m} |(I-\theta z_0)^{-1}(z_j-z_0)| \max_{1\le i,j\le m} |[(I-A(z_0)Z)^{-1}]_{ij}| \le$$

$$\le \Omega_1\ \Delta(h,\rho_m(h)) \prod_{k=1}^{m-1} (1+\tau_k(h)) ,$$

where $\Omega_1 = m^3\ \omega(1+\theta^{-1}\omega)^{m-1}$. Here we have used the fact that the i,j-th. block of $(1-A(z_0)Z_0)^{-1}$ is a rational expression in z_0 (see e.g. lemma 2.4.6), and that two rational expressions in z_0 commute. Besides the

norms of $[(I-A(z_0)Z_0)^{-1}]_{ij}$ are estimated by $(1+\theta^{-1}\omega)^{m-1}$ (see corollary 2.4.10 and (5.4.8)).

Using theorem 2.3.6 and corollary 2.2.8 it can be seen that there is an $\omega' > 0$ such that all $|[D_x b_i(z_0)v](I-\theta z_0)|$ and $|[D_x a_{ij}(z_0)v](I-\theta z_0)|$ are bounded by $\omega'|[(I-\theta z_0)^{-1}hJ'(x_0)v|$ for arbitrary $v \in \mathbb{K}^s$. Therefore

$$|[D_x b(z_0)^T v]hF(y(x_0))| \leq$$

$$\leq m \max_{1\leq i\leq m} |[D_x b_i(z_0)v](I-\theta z_0)| \max_{1\leq i\leq m} |(I-\theta z_0)^{-1}hf(y_i(x_0))| \leq$$

$$\leq \Omega_2 \gamma(h)\sigma_m(h)|v|$$

with $\Omega_2 = m\omega'$.

The last term on the right-hand side of (5.2.6) is estimated in a similar way. Note that $A(z_0)$, and therefore also $D_x A(z_0)v$, are strictly lower block-triangular.

$$|b(z_0)^T Z(I-A(z_0)Z)^{-1}[D_x A(z_0)v]hF(y(x_0))| =$$

$$= \Big|\sum_{i=1}^{m}\sum_{j=2}^{m}\sum_{k=1}^{m-1} b_i(z_0)z_i[(I-A(z_0)Z)^{-1}]_{ij}[D_x a_{jk}(z_0)v]hf(y_k(x_0))\Big| \leq$$

$$\leq m^3 \max_{1\leq i\leq m} |b_i(z_0)z_i| \max_{\substack{1\leq i\leq m\\ 2\leq j\leq m}} |[(I-A(z_0)Z)^{-1}]_{ij}|$$

$$\max_{1\leq j,k\leq m} |[D_x a_{jk}(z_0)v](I-\theta z_0)| \max_{1\leq k\leq m-1} |(I-\theta z_0)^{-1}hf(y_k(x_0))| \leq$$

$$\leq \Omega_3 \gamma(h)\,\sigma_{m-1}(h)\,\tau_m(h)\prod_{j=2}^{m-1}(1+\tau_k(h))$$

with $\Omega_3 = m^3\omega'$. We define $\sigma_0(h) = 0$ for the case that $m = 1$.

Let $\hat{\sigma}_i(h)$ $(1\leq i\leq m)$ be defined by

$$(5.4.14.a)\quad \hat{\sigma}_i(h) = 1 + \omega\Big[\sum_{j=1}^{i-1}\hat{\sigma}_j(h)\Big]\Big[\theta^{-1}+\delta(h)+\alpha(h)\varepsilon(h)\omega\sum_{j=1}^{i-1}\hat{\sigma}_j(h)\Big].$$

In view of (5.4.9) we have $\sigma_i(h) = \alpha(h)\,\hat{\sigma}_i(h)$ $(1\leq i\leq m)$, and

$$(5.4.14.b)\quad \tau_i(h) = \theta^{-1}\omega + \omega[\delta(h)+\alpha(h)\varepsilon(h)\omega \sum_{j=1}^{i-1} \hat{\sigma}_j(h)] \ .$$

Thus we see that the $\hat{\sigma}_i(h)$ and $\tau_i(h)$ can be written as polynomials in $\alpha(h)\varepsilon(h)$ and $\delta(h)$. We further have

$$\Delta(h,\rho_i(h)) = \delta(h) + \alpha(h)\varepsilon(h)\,\omega \sum_{j=1}^{i-1} \hat{\sigma}_j(h) \ .$$

Inserting this in the above estimates for the terms of (5.2.6), we obtain (5.4.12) with

$$(5.4.15.a)\quad P(\alpha(h)\varepsilon(h),\delta(h)) = \Omega_1 \prod_{k=1}^{m-1} (1+\tau_k(h)) \ ,$$

$$(5.4.15.b)\quad Q(\alpha(h)\varepsilon(h),\delta(h)) = \Omega_1\omega \sum_{j=1}^{m-1} \hat{\sigma}_j(h) \prod_{k=1}^{m-1} (1+\tau_k(h)) \ ,$$

$$(5.4.15.c)\quad R(\alpha(h)\varepsilon(h),\delta(h)) = \Omega_2\ \hat{\sigma}_m(h) + \Omega_3\ \hat{\sigma}_{m-1}(h)\ \tau_m(h) \prod_{k=2}^{m-1} (1+\tau_k(h)) \ .$$

Since $\rho_m(h) = \omega\alpha(h)\ [\hat{\sigma}_1(h)+\hat{\sigma}_2(h)+\ldots+\hat{\sigma}_{m-1}(h)]$, we also see from the above that the condition $\rho_m(h) < r_0$ can be written as (5.4.13) with

$$(5.4.15.d)\quad S(\alpha(h)\varepsilon(h),\delta(h)) = \omega \sum_{j=1}^{m-1} \hat{\sigma}_j(h) \ .$$

□

REMARK 5.4.9. If $m = 1$, we have $P \equiv \Omega_1$, $R \equiv \Omega_2$ and $Q \equiv S \equiv 0$. Here P,Q,R and S are the polynomials arising in theorem 5.4.7 and remark 5.4.8, and Ω_1,Ω_2 are the constants defined in the proof of theorem 5.4.7. Thus condition (5.4.13) is always fulfilled, and (5.4.12) reads

$$|G'(x_0;h,f)| \le \Phi(h\beta) + \Omega_1\ \delta(h) + \Omega_2\ \alpha(h)\ \gamma(h) \ .$$

EXAMPLE 5.4.10. Consider the one-stage method G , with one parameter $\theta \in (0,1]$, given by

$$G(x;h,f) = x + (I-h\theta J(x))^{-1}\ hf(x) \ .$$

The stability function of this method equals

$$\phi(\zeta) = (1-\theta\zeta)^{-1}\ (1+(1-\theta)\zeta) \qquad (\zeta \in \mathbb{C}) .$$

By some calculations it follows that

$$\Phi(t) = \max\{\theta^{-1}(1-\theta), (1-\theta t)^{-1}\ (1+(1-\theta)t)\} \quad (\text{for } t < \theta^{-1}).$$

We can take $\Omega_1 = 1$ and $\Omega_2 = \theta$ (namely $m = 1$, $\omega = 1$ and $\omega' = \theta$). Hence

$$|G'(x_0;h,f)| \leq \Phi(h\beta) + \delta(h) + \theta\alpha(h)\gamma(h)$$

whenever (A) holds and $1 - 2h\theta\beta > 0$ (i.e. $h \in (0,h_0)$).

In this case the stepsize restriction $1 - 2h\theta\beta > 0$ can be weakened to $1 - h\theta\beta > 0$.

EXAMPLE 5.4.11. Consider the two-stage method with

$$a_{21}(\zeta) = (1-\theta\zeta)^{-1}\ \xi \qquad (\text{for } \zeta \in \mathbb{C}),$$

$$b_i(\zeta) = (1-\theta\zeta)^{-1}\ \eta_i \qquad (\text{for } i = 1,2,\ \zeta \in \mathbb{C}),$$

where $\theta > 0$ and $\eta_1, \eta_2, \xi \in \mathbb{R}$. With $\theta = 1 - \frac{1}{2}\sqrt{2}$, $\eta_1 = 0$, $\eta_2 = 1$, $\xi = \frac{1}{2}(\sqrt{2}-1)$ we obtain a well known method proposed by ROSENBROCK (1963). With $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$, $\eta_1 = 3/4$, $\eta_2 = 1/4$ and $\xi = -\frac{2}{3}\sqrt{3}$ the third order method of CALAHAN (1968) appears.

From (5.4.14), (5.4.15) with $m = 2$, $\omega = \max\{|\eta_1|, |\eta_2|, |\xi|\}$, $\omega' = \theta\omega$, we obtain for this example

$$P(\alpha(h)\varepsilon(h),\delta(h)) = [\Omega_1(1+\theta^{-1}\omega)] + [\Omega_1\omega]\ \delta(h),$$

$$Q(\alpha(h)\varepsilon(h),\delta(h)) = [\Omega_1\omega(1+\theta^{-1}\omega)] + [\Omega_1\omega^2]\ \delta(h),$$

$$R(\alpha(h)\varepsilon(h),\delta(h)) = [\Omega_2(1+\theta^{-1}\omega)+\Omega_3\theta^{-1}\omega] + [(\Omega_2+\Omega_3)\omega]\ \delta(h) + \\ + [(\Omega_2+\Omega_3)\omega^2]\ \alpha(h)\varepsilon(h),$$

$$S(\alpha(h)\varepsilon(h),\delta(h)) = \omega,$$

with $\Omega_1 = 8\omega(1+\theta^{-1}\omega)$, $\Omega_2 = 2\theta\omega$ and $\Omega_3 = 8\theta\omega$. With these P,Q and R we get an upper bound for $G'(x_0;h,f)$ from theorem 5.4.7.

This result can be improved a little by using the *internal* stability function $\phi_1(\zeta) = 1 + a_{21}(\zeta)\zeta$ $(\zeta \in \mathbb{C})$, and $\Phi_1(t) = \sup\{|\phi_1(\zeta)| : \zeta\in\mathbb{C}, \mathrm{Re}\ \zeta\le t\}$. We have used in this section the estimate $|\phi_1(\zeta)| \le 1 + |a_{21}(\zeta)\zeta|$, but this may be crude, for instance if it is known that ϕ_1 is L-acceptable (i.e. $\Phi_1(0) \le 1$, $\phi_1(\infty) = 0$). For this example we give a more precise analysis.

We suppose that (A) holds, $y_1(x_0) \in \mathcal{D}_0$ and $h \in (0,h_0)$. By writing out (5.2.6) we get

$$
\begin{aligned}
G'(x_0)v = {} & \phi(z_0)v + \{\eta_1(I-\theta z_0)^{-1}(z_1-z_0) + \\
& + \eta_2(I-\theta z_0)^{-1}(z_2-z_0)(I+\xi(I-\theta z_0)^{-1}z_1) + \\
& + \eta_2(I-\theta z_0)^{-1}z_0(\xi(I-\theta z_0)^{-1}(z_1-z_0))\}v + \\
& + \{\eta_1\theta(I-\theta z_0)^{-1}[hJ'(x_0)v](I-\theta z_0)^{-1} hf(x_0) + \\
& + \eta_2\theta(I-\theta z_0)^{-1}[hJ'(x_0)v](I-\theta z_0)^{-1} hf(y_2(x_0))\} + \\
& + \{\eta_2\xi\theta(I-\theta z_0)^{-1}z_2(I-\theta z_0)^{-1}[hJ'(x_0)v](I-\theta z_0)^{-1}hf(x_0)\} .
\end{aligned}
$$

We have $y_2(x_0) - x_0 = \xi(I-\theta z_0)^{-1} hf(x_0)$, $I + \xi(I-\theta z_0)^{-1} z_1 = \phi_1(z_0) + \xi(I-\theta z_0)^{-1}(z_1-z_0)$, which leads to

$$
\begin{aligned}
& |y_2(x_0)-x_0| \le |\xi|\ \alpha(h) =: \rho_2^*(h) , \\
& |I+\xi(I-\theta z_0)^{-1}z_1| \le \Phi_1(h\beta) + |\xi|\ \delta(h) =: 1 + \tau_1^*(h) .
\end{aligned}
$$

Since $\rho_2^*(h) \le \rho_2(h)$, $\tau_1^*(h) \le \tau_1(h)$ these estimates are sharper than the ones obtained from the lemmata 5.4.3 and 5.4.6. We also have

$$
\begin{aligned}
|(I-\theta z_0)^{-1}hf(y_2(x_0))| \le {} & \Phi_1(h\beta)\alpha(h) + \\
& + [\delta(h)+\varepsilon(h)\rho_2^*(h)]\ \rho_2^*(h) =: \sigma_2^*(h) ,
\end{aligned}
$$

which follows from

$$(I-\theta z_0)^{-1}\, hf(y_2(x_0)) = \phi_1(z_0)\,(I-\theta z_0)^{-1}\, hf(x_0) +$$

$$+ \int_0^1 (I-\theta z_0)^{-1}\,[hf'(x_0+t(y_2(x_0)-x_0))-z_0]\,dt\,(y_2(x_0)-x_0)$$

(see (5.4.11)). Inserting these estimates in the expression for $G'(x_0)$ yields the upper bound

$$|G'(x_0)| \le \Phi(h\beta) + \{[|\eta_1|+|\eta_2|\Phi_1(h\beta)+|\eta_2\xi\theta^{-1}|]+|\eta_2\xi|\delta(h)\}\,\delta(h) +$$

$$+ \{[|\eta_2\xi|\Phi_1(h\beta)]+|\eta_2\xi^2|\delta(h)\}\,\alpha(h)\varepsilon(h) +$$

$$+ \{[|\eta_1\theta|+|\eta_2\xi|+|\eta_2\theta|\Phi_1(h\beta)]+2|\eta_2\xi\theta|\delta(h)+2|\eta_2\xi^2\theta|\alpha(h)\varepsilon(h)\}$$

$$\alpha(h)\gamma(h)\ .$$

This result is only a slight quantitative improvement over the result we obtained directly from theorem 5.4.7. In particular it does not reveal the importance of the internal A-stability concept (see VERWER (1977)), which has been shown to be useful by means of numerical experiments in VERWER (1977). More restricted model problems (see e.g. VAN VELDHUIZEN (1981), VERWER (1982)) seem to be better suited to deal with refinements on theorem 5.4.7.

5.4.3. A general contractivity result.

In this section theorem 5.4.7 will be applied to obtain contractivity results for semi-implicit methods. In the subsequent sections these results will be used for concrete choices of the Jacobian approximation J. The contractivity results show what kind of extra conditions on the dissipative functions f are sufficient to have for the numerical approximations the analogue of property (5.1.5),

$$|\tilde{U}(t+h)-U(t+h)| \le |\tilde{U}(t)-U(t)| \qquad \text{(for all } t \ge 0\ ,\ h > 0\ ).$$

A requirement on the method will be that the method is A-stable. Otherwise we cannot even get a contractivity result for the class of simple dissipative testproblems (5.1.6). It will turn out that we have to require a bit more than A-stability, namely strong A-stability (cf. definition 5.3.2), if we deal with nonlinear functions f (see the theorems

5.4.18, 5.4.24).

We shall be concerned with a semi-implicit method G which satisfies (3.3.5) and (3.3.9). The positive constant θ is defined as in section 5.4.2.

Let $\alpha_0,\gamma_0,\delta_0,\varepsilon_0 \geq 0$ and $\beta_0 \in \mathbb{R}$ be given constants, and let h_0 be defined by (5.4.1) with $\beta = \beta_0$. We consider the following set of assumptions (A_0) on the functions f and $J(\cdot) = J(\cdot;h,f)$. (A_0) consists of (5.4.2) and the assumption that (5.4.3) and the following conditions (5.4.16)-(5.4.19) hold for all $h \in (0,h_0)$.

$$|f(x_0)| \leq \alpha_0 , \tag{5.4.16}$$

$$\mu[J(x_0)] \leq \beta_0 , \tag{5.4.17}$$

$$|J'(x_0)| \leq \gamma_0 , \tag{5.4.18}$$

$$|J(x_0)-f'(x)| \leq \delta_0 + \varepsilon_0|x-x_0| \qquad \text{(for all } x \in \mathcal{D}_0\text{)}. \tag{5.4.19}$$

From corollary 2.2.11 we see that (5.4.17) implies

$$|(I-h\theta J(x_0))^{-1}| \leq (1-h\theta\beta_0)^{-1} ,$$

for all $h \in (0,h_0)$. Therefore (5.4.4)-(5.4.7) hold with $\alpha(h) = (1-h\theta\beta_0)^{-1} h\alpha_0$, $\beta = \beta_0$, $\gamma(h) = (1-h\theta\beta_0)^{-1} h\gamma_0$, $\delta(h) = (1-h\theta\beta_0)^{-1} h\delta_0$ and $\varepsilon(h) = (1-h\theta\beta_0)^{-1} h\varepsilon_0$. The results of section 5.4.2 will be applied with these choices.

THEOREM 5.4.12. *Let G be a strongly A-stable semi-implicit method satisfying* (3.3.5) *and* (3.3.9). *Then the following holds.*

(B_0) *There are constants $c_0,c_1,c_2,c_3 > 0$ such that G is unconditionally locally contractive at x_0 for f, whenever (A_0) holds with $\beta_0 < 0$, $\alpha_0/(|\beta_0|r_0) \leq c_0$, $\alpha_0\gamma_0/\beta_0^2 \leq c_1$, $\alpha_0\varepsilon_0/\beta_0^2 \leq c_2$ and $\delta_0/|\beta_0| \leq c_3$.*

These constants c_0,c_1,c_2 and c_3 only depend on the coefficients of the method G. Moreover there exists a constant $c_3^* > 0$, only depending on the coefficients of G, such that

(C_0) *For any given* $c_0,c_1,c_2 > 0$ *and* $c_3 \in (0,c_3^*)$, *there is a* $K > 0$ *such that* G *is locally contractive at* x_0 *for* f *and all* $h > 0$ *with* $h|\beta_0| \le K$, *whenever* (A_0) *holds with* $\beta_0 < 0$, $\alpha_0/(|\beta_0|r_0) \le c_0$, $\alpha_0\gamma_0/\beta_0^2 \le c_1$, $\alpha_0\varepsilon_0/\beta_0^2 \le c_2$ *and* $\delta_0/|\beta_0| \le c_3$.

Here K *only depends on* c_0,c_1,c_2,c_3 *and on the coefficients of* G.

PROOF. In this proof we apply theorem 5.4.7 with $\alpha(h),\beta,\gamma(h),\delta(h),\varepsilon(h)$ as indicated above this theorem.

From theorem 2.2.4 it is known that there are constants $\lambda > 0$ and $t^* < 0$ such that the function $\tilde{\Phi}$, defined on $(-\infty,0]$ by $\tilde{\Phi}(t) = 1 + \lambda t$ (for $t^* \le t \le 0$), $\tilde{\Phi}(t) = 1 + \lambda t^*$ (for $t < t^*$), satisfies

$$\Phi(t) \le \tilde{\Phi}(t) \quad \text{(for all } t \le 0\text{)}.$$

Let $c_1^{(0)},c_2^{(0)},c_3^{(0)}$ be positive numbers. Suppose $\beta_0 < 0$, $\alpha_0\gamma_0/\beta_0^2 \le c_1^{(0)}$, $\alpha_0\varepsilon_0/\beta_0^2 \le c_2^{(0)}$ and $\delta_0/|\beta_0| \le c_3^{(0)}$. We denote by P_0,Q_0,R_0,S_0 the function values at the point $(\theta^{-2}c_2^{(0)},\theta^{-1}c_3^{(0)})$ of the polynomials P,Q,R,S, respectively. These are the polynomials arising in theorem 5.4.7 and remark 5.4.8. By taking $c_0 > 0$ such that $c_0S_0 \le r_0$, it is easily seen that (5.4.13) holds for all $h > 0$ whenever $\alpha_0/(|\beta_0|r_0) \le c_0$.

Further we see from theorem 5.4.7 that

$$|G'(x_0;h,f)| \le \Phi(-k) + (1+\theta k)^{-1}\, k(\delta_0/|\beta_0|)\, P_0 +$$
$$+ (1+\theta k)^{-2}\, k^2(\alpha_0\varepsilon_0/\beta_0^2)\, Q_0 + (1+\theta k)^{-2}\, k^2(\alpha_0\gamma_0/\beta_0^2)\, R_0,$$

where k stands for $h|\beta_0|$. It follows that $|G'(x_0;h,f)| \le 1$ for all $h > 0$ if $\delta_0/|\beta_0|$, $\alpha_0\varepsilon_0/\beta_0^2$ and $\alpha_0\gamma_0/\beta_0^2$ are sufficiently small. Hence we obtain (B_0).

For proving (C_0) we also suppose that $c_0^{(0)} > 0$ and $\alpha_0/|\beta_0|r_0 \le c_0^{(0)}$. If $k = h|\beta_0|$ satisfies $(1+\theta k)^{-1}\, kc_0^{(0)}\, S_0 < 1$, (5.4.13) will hold. From (5.4.12) we see that

$$|G'(x_0;h,f)| \le 1 - \lambda k + k(\delta_0/|\beta_0|)\, P(0,0) + Mk^2 \quad (k \downarrow 0)$$

where M only depends on $\alpha_0\gamma_0/\beta_0^2$, $\alpha_0\varepsilon_0/\beta_0^2$ and $\delta_0/|\beta_0|$. Therefore (C_0) holds with $c_3^* = \lambda/P(0,0)$. □

REMARK 5.4.13. Statement (C_0) is equivalent to the following statement (D_0) , which seems at first sight to be a bit stronger.

(D_0) *For* $c_0,c_1,c_2 > 0$, $c_3 \in (0,c_3^*)$ *given, there is a* $K > 0$ *such that* G *is locally contractive at* x_0 *for* f *and all* $h > 0$ *with* $h\cdot\max\{\alpha_0/(r_0c_0),(\alpha_0\gamma_0/c_1)^{\frac{1}{2}},(\alpha_0\varepsilon_0/c_2)^{\frac{1}{2}},\delta_0/c_3\} \leq K$, *whenever* (A_0) *holds with* $\beta_0 > 0$, $\alpha_0/(|\beta_0|r_0) \leq c_0$, $\alpha_0\gamma_0/\beta_0^2 \leq c_1$, $\alpha_0\varepsilon_0/\beta_0^2 \leq c_2$ *and* $\delta_0/|\beta_0| \leq c_3$.

The equivalence of (C_0) and (D_0) can easily be proved by using the following fact: if (A_0) holds with constants $\alpha_0,\beta_0,\gamma_0,\delta_0,\varepsilon_0,r_0$, and if $\beta_0' > \beta_0$, then (A_0) also holds with constants $\alpha_0,\beta_0',\gamma_0,\delta_0,\varepsilon_0$ and r_0 .

REMARK 5.4.14. If $m = 1$ we may take $c_0 = c_2 = \infty$ in the statements (B_0) , (C_0) and (D_0) . This can be seen from remark 5.4.9 and the proof of theorem 5.4.12.

The reason for this is that the condition $\alpha_0/(|\beta_0|r_0) \leq c_0$ is only needed to ensure that all the intermediate vectors $y_1(x_0),y_2(x_0),\ldots,y_m(x_0)$ are in $\mathcal{D}_0$, and ε_0 is only used to give an upper bound for $|J(x_0)-f'(y_i(x_0))|$ $(i=2,3,\ldots,m)$. If $m = 1$, there is only one intermediate vector $y_1(x_0)$, and $y_1(x_0) = x_0$.

5.4.4. Rosenbrock methods.

The previous results will now be applied for concrete choices for the Jacobian approximations $J(\cdot) = J(\cdot;h,f)$. The first choice we consider is $J(x) \equiv f'(x)$, i.e. we are dealing with a Rosenbrock method.

We consider a Rosenbrock method G satisfying (3.3.5) and (3.3.9). Let $\theta > 0$ be defined as in section 5.4.2.

Suppose $\beta_0 \in \mathbb{R}$ and $\alpha_0,\gamma_0,\delta_0,\gamma_0 \in \mathbb{R}^+$. We will consider functions f that satisfy the set of assumptions (A_1) consisting of (5.4.2) and (5.4.20)-(5.4.23).

(5.4.20) f is twice continuously differentiable on $\mathcal{D}_0$.

(5.4.21) $|f(x_0)| \le \alpha_0$.

(5.4.22) $\mu[f'(x_0)] \le \beta_0$.

(5.4.23) $|f''(x)| \le \gamma_0$ (for all $x \in \mathcal{D}_0$).

Then (A_0) holds with $\delta_0 = 0$ and $\varepsilon_0 = \gamma_0$ (see (5.4.16)-(5.4.19)), and (A) holds with $\beta = \beta_0$, $\delta(h) = 0$, $\alpha(h) = (1-h\theta\beta)^{-1}h\alpha_0$ and $\gamma(h) = \varepsilon(h) = (1-h\theta\beta)^{-1}h\gamma_0$ (see (5.4.3)-(5.4.7)). Further we will take the same step-size restriction as in the previous sections; we take $h \in (0,h_0)$ where h_0 is defined by (5.4.1) with $\beta = \beta_0$.

In order to ensure that all the internal vectors $y_i(x_0)$ are in $\mathcal{D}_0$ and $G(\cdot;h,f)$ is continuously differentiable at x_0 we will require (instead of (5.4.13) for the general case)

(5.4.24) $(1-h\theta\beta_0)^{-1}\, h\alpha_0\, S_1((1-h\theta\beta_0)^{-2}h^2\alpha_0\gamma_0) < r_0$.

Here $S_1(\xi) = S(\xi,0)$ (for $\xi \in \mathbb{R}^+$) where S is the polynomial defined by (5.4.14.a) and (5.4.15.d). If $m = 1$ then S_1 is identically equal to zero, and in case $m = 2$ S_1 is constant.

By application of theorem 5.4.7 we obtain the following result.

THEOREM 5.4.15. *Let* G *be a Rosenbrock method satisfying* (3.3.5) *and* (3.3.9). *Assume* $h \in (0,h_0)$, (A_1) *and* (5.4.24). *Then we have*

(5.4.25) $|G'(x_0;h,f)| \le \Phi(h\beta_0) + P_1((1-h\theta\beta_0)^{-2}h^2\alpha_0\gamma_0)$

where P_1 *is a polynomial satisfying* $P_1(0) = 0$.

The above polynomial P_1 can be calculated from (5.4.14), (5.4.15) and the relation $P_1(\xi) = \xi Q(\xi,0) + \xi R(\xi,0)$ (for $\xi \in \mathbb{R}^+$). Note that P_1 only depends on the coefficients of the method G .

For fixed $h \in (0,h_0)$ and $\beta_0 \in \mathbb{R}$, the inequality (5.4.25) shows that the effect of the nonlinearity of the function f is small if $\alpha_0\gamma_0$ is small. It will be shown in example 5.4.29 that if the product $\alpha_0\gamma_0$ is allowed to be large the error propagation with the Rosenbrock method G may be unfavourable, no matter how good the method works on

linear stiff systems.

From theorem 5.4.12 we obtain

THEOREM 5.4.16. *Let* G *be a strongly* A*-stable Rosenbrock method satisfying* (3.3.5) *and* (3.3.9). *Then the following statements hold.*

(B_1) *There are* $c_0, c_1 > 0$ *such that* G *is unconditionally locally contractive at* x_0 *for* f *whenever* (A_1) *holds with* $\beta_0 < 0$, $\alpha_0/(|\beta_0| r_0) \le c_0$ *and* $\alpha_0\gamma_0/\beta_0^2 \le c_1$.

(C_1) *For any given* $c_0, c_1 \ge 0$ *there is a* $K > 0$ *such that* G *is locally contractive at* x_0 *for* f *and all* $h > 0$ *with* $h|\beta_0| \le K$, *whenever* (A_1) *holds with* $\beta_0 < 0$, $\alpha_0/(|\beta_0| r_0) \le c_0$ *and* $\alpha_0\gamma_0/\beta_0^2 \le c_1$.

In view of remark 5.4.14 we may take $c_0 = \infty$ in statement (B_1) if $m = 1$.

We consider the following application of theorem 5.4.16. Let $f: \mathbb{K}^s \to \mathbb{K}^s$ be twice continuously differentiable. Suppose U is a solution of the differential equation $U'(t) = f(U(t))$ $(t \ge 0)$, that converges to a stationary solution u^* satisfying

$$\mu[f'(u^*)] < 0 .$$

Let further G be a strongly A-stable Rosenbrock method satisfying (3.3.5) and (3.3.9). We define

$$\ell(h,x) = h^{-1}[V(h,x) - G(x;h,f)] \qquad (h > 0 , x \in \mathbb{K}^s)$$

where $V(h,x)$ stands for the solution at $t = h$ of the initial value problem $V'(t) = f(V(t))$ $(t \ge 0)$, $V(0) = x$. It can be proved that, for x sufficiently close to u^*, both $V(h,x)$ and $G(x;h,f)$ are defined for arbitrary $h > 0$, and depend continuously on x. Since $G(u^*;h,f) = u^*$ and $V(h,u^*) = u^*$ (for all $h > 0$), we have

$$\lim_{x \to u^*} \ell(h,x) = 0 \qquad \text{(for all } h > 0\text{)}.$$

The local discretization error of method G w.r.t. U at time t equals $\ell(h,U(t))$. Thus we see that as $U(t)$ approaches u^* , the local error will permit a (rather) large stepsize. The following corollary of theorem 5.4.16 shows that if the numerical approximation u_n to $U(t_n)$ is close to u^* , a large stepsize will not lead to an unfavourable error propagation.

COROLLARY 5.4.17. *Let* G *be a strongly* A*-stable Rosenbrock method satisfying* (3.3.5) *and* (3.3.9). *Suppose* $f: \mathbb{K}^s \to \mathbb{K}^s$ *is twice continuously differentiable on* $\mathbb{K}^s$, *and* $u^* \in \mathbb{K}^s$ *is such that* $f(u^*) = 0$, $\mu[f'(u^*)] < 0$. *Then there exists an open neighbourhood* $\mathcal{D}$ *of* u^* *such that* G *is unconditionally contractive on* $\mathcal{D}$ *for* f .

PROOF. Let $\beta_0 = \frac{1}{2}\mu[f'(u^*)]$, $\gamma_0 = 2|f''(u^*)|$, and let $r_0 > 0$ be such that $\mu[f'(x)] \le \beta_0$, $|f''(x)| \le \gamma_0$ for all $x \in \mathbb{K}^s$ with $|x-u^*| < 2r_0$. We take $\mathcal{D} \subset \mathbb{K}^s$ such that $\mathcal{D}$ is open and convex, $u^* \in \mathcal{D} \subset \{x: x \in \mathbb{K}^s, |x-u^*| < r_0\}$, $\alpha_0/(|\beta_0| r_0) \le c_0$ and $\alpha_0\gamma_0/\beta_0^2 \le c_1$, where c_0, c_1 are as in statement (B_1) of theorem 5.4.16, and $\alpha_0 = \sup\{|f(x)| : x \in \mathcal{D}\}$.

Then (5.4.20)-(5.4.23) are fulfilled for any $x_0 \in \mathcal{D}$. From theorem 5.4.16 it thus follows that G is unconditionally locally contractive at each point $x_0 \in \mathcal{D}$. In view of corollary 5.1.3 the proof is completed. □

In theorem 5.4.16 it has been required that G is strongly A-stable. It will now be shown that this requirement is essential.

THEOREM 5.4.18. *The conclusions* (B_1) *and* (C_1) *of theorem* 5.4.16 *do not hold on the class of* A*-stable Rosenbrock methods satisfying* (3.3.5) *and* (3.3.9).

PROOF. Consider an arbitrary Rosenbrock method G with $m = 1$, $b_1 \neq 0$, which is A-stable but not strongly A-stable. Let $\beta_0 < 0$ and $\alpha_0, \gamma_0, H > 0$ be arbitrary, and let $h \in (0, H)$ be such that $b_1(\zeta)$ is regular at $\zeta = h\beta_0$, and $b_1'(h\beta_0) \neq 0$. Note that $b_1(\infty)$ exists in $\mathbb{R}$ since G is A-stable and $\phi(\zeta) = 1 + \zeta b_1(\zeta)$ ($\zeta \in \mathbb{C}$) .

We take $\mathbb{K}^s = \mathbb{R}^2$. On the $\mathbb{R}^2$ we consider the Euclidean inner product. Let $x_0 = 0 \in \mathbb{R}^2$, $\sigma > 1$, and let $f: \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$f(x) = \begin{pmatrix} \alpha_0+\beta_0 x_1+\gamma_0 x_1 x_2 \\ \sigma\beta_0 x_2 \end{pmatrix} \qquad (\text{for } x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2).$$

This can be written as

$$f(x) = f_0 + f_0' x + \tfrac{1}{2}[f_0'' x]x \qquad (\text{for } x \in \mathbb{R}^2)$$

where $f_0 = \alpha_0 e_1$, $f_0' = \operatorname{diag}(\beta_0, \sigma\beta_0)$, and $f_0'' \in L(\mathbb{R}^2, L(\mathbb{R}^2))$ is defined by $[f_0''u]v = \gamma_0[\langle u,e_1\rangle\langle v,e_2\rangle + \langle u,e_2\rangle\langle v,e_1\rangle]e_1$ (for $u,v \in \mathbb{R}^2$). Here $e_1 = (1,0)^T$ and $e_2 = (0,1)^T$.

We have $f(x_0) = f_0$, $f'(x_0) = f_0'$, $f''(x) = f_0''$ (for all $x \in \mathbb{R}^2$). By some calculations it follows that $|f(x_0)| = \alpha_0$, $\mu[f'(x_0)] = \beta_0$ and $|f''(x)| = \gamma_0$ (for all $x \in \mathbb{R}^2$). Thus we see that the assumption (A_1) is fulfilled with $r_0 > 0$ arbitrary.

From formula (5.2.6) we obtain

$$G'(x_0;h,f)v = \phi(hf_0')v + [D_x b_1(hf_0')v]hf_0$$

for arbitrary $v \in \mathbb{R}^2$. Further we have in view of theorem 2.3.6

$$D_x b_1(hf_0')v = \sum_{i=1}^{k} c_i(hf_0')\,[hf_0''v]\,d_i(hf_0')$$

where $k \in \mathbb{N}$ and the c_i, d_i are rational functions with $b_1'(\zeta) = c_1(\zeta)d_1(\zeta) + c_2(\zeta)d_2(\zeta) + \ldots + c_k(\zeta)d_k(\zeta)$ ($\zeta \in \mathbb{C}$). Taking $v = e_2$ we thus obtain

$$[D_x b_1(hf_0')e_2]hf_0 = b_1'(h\beta_0)\, h^2\alpha_0\gamma_0\, e_1 .$$

Hence

$$G'(x_0;h,f)e_2 = \phi(h\sigma\beta_0)e_2 + b_1'(h\beta_0)\, h^2\alpha_0\gamma_0 e_1 .$$

Since G is merely A-stable, we have $\lim_{\sigma\to\infty} |\phi(h\sigma\beta_0)| = 1$. Therefore $|G'(x_0;h,f)| > 1$ if σ is chosen sufficiently large. □

The above theorem is a slight extension of a result given in HUNDSDORFER (1981). In the same paper a 2-dimensional example was presented showing that if $\mu[f'(x)] \le 0$ on some open region around x_0 and $\alpha_0\gamma_0 > 0$ is arbitrarily small we need not have contractivity for strongly A-stable methods. Therefore the requirement $\beta_0 < 0$ in theorem 5.4.16 is also essential.

For the cases that $\beta_0 = 0$ or G is A-stable but not strongly A-stable, we shall derive in the following remark a stability result on finite intervals for sufficiently small stepsizes. In this result the Lipschitz constant of the function f is not involved (in contrast to the classical stability results for nonstiff initial value problems such as in HENRICI (1962)).

REMARK 5.4.19. Let $\alpha_0,\gamma_0,r_0 > 0$ and $\beta_0 \le 0$. Let further $T > 0$ and $\mathcal{D} \subset \mathbb{K}^s$ be open and convex. Suppose $f: \mathbb{K}^s \to \mathbb{K}^s$ is such that (A_1) holds for any $x_0 \in \mathcal{D}$. We consider solutions U on the interval $[0,T]$ of the differential equation $U'(t) = f(U(t))$ $(0\le t\le T)$ with trajectories in $\mathcal{D}$.

Let $u_n,\tilde{u}_n \in \mathcal{D}$ $(1\le n\le T/h)$ be two sequences computed from an A-stable Rosenbrock method G satisfying (3.3.5) and (3.3.9), with stepsize $h \in (0,H]$ and starting vectors $u_0,\tilde{u}_0 \in \mathcal{D}$. Here $H \in (0,T]$ is such that (5.4.24) holds whenever $h \in (0,H]$.

If $\beta_0 = 0$ or G is not strongly A-stable, theorem 5.4.16 cannot be applied. However, we do have in view of theorem 5.4.15 and lemma 2.3.2,

$$|\tilde{u}_n-u_n| \le [1+P_1((1-h\theta\beta_0)^{-2}h^2\alpha_0\gamma_0)]\,|\tilde{u}_{n-1}-u_{n-1}|\ ,$$

and therefore

$$|\tilde{u}_n-u_n| \le [1+P_1((1-h\theta\beta_0)^{-2}h^2\alpha_0\gamma_0)]^n\,|\tilde{u}_0-u_0|\ ,$$

for all $n \in \mathbb{N}$ with $1 \le n \le T/h$. Since $P_1(0) = 0$, it follows that there exists a constant $c > 0$, which only depends on $\alpha_0\gamma_0,\beta_0$ and the coefficients of G, such that

$$|\tilde{u}_n-u_n| \le e^{chT}\,|\tilde{u}_0-u_0| \qquad \text{(for all } h \in (0, H],\ 1 \le n \le T/h\ ).$$

Thus we have *stability* on the finite interval $[0,T]$.

This result has only practical value if the interval $[0,T]$ is not too long, because the stepsize has to be restricted in order to get a moderate stability factor e^{chT}.

EXAMPLE 5.4.20. Consider the Rosenbrock method G given by

$$G(x;h,f) = x + (I-h\theta f'(x))^{-1} hf(x)$$

with $\theta \in (0,1]$. By example 5.4.10 (with $J(x) = f'(x)$, $\alpha(h) = (1-h\theta\beta_0)^{-1} h\alpha_0$, $\beta = \beta_0$, $\gamma(h) = (1-h\theta\beta_0)^{-1} h\gamma_0$ and $\delta(h) = 0$) we see that

$$|G'(x_0;h,f)| \le \Phi(h\beta_0) + \theta(1-h\theta\beta_0)^{-2} h^2\alpha_0\gamma_0$$

whenever (A_1) holds and $1 - h\theta\beta > 0$. Here $\Phi(t) = \max\{\theta^{-1}(1-\theta), (1-\theta t)^{-1} (1+(1-\theta)t)\}$ (for $t < \theta^{-1}$). If $\theta \ge \frac{1}{2}$ the method is A-stable. For $\theta > \frac{1}{2}$ we have strong A-stability.

Suppose $\theta > \frac{1}{2}$. By some calculations (see e.g. HUNDSDORFER (1981)) it can be shown that if

$$\beta_0 < 0 \quad \text{and} \quad \alpha_0\gamma_0/\beta_0^2 \le 2\theta - 1 ,$$

then $|G'(x_0;h,f)| \le 1$ for all $h > 0$ and f satisfying (A_1). Thus we may take $c_0 = \infty$, $c_1 = 2\theta - 1$ in statement (B_1) of theorem 5.4.10.

If $\theta = \frac{1}{2}$ we get no contractivity results. Using the stability property discussed in remark 5.4.19 it will be shown in section 6.3 that we do have convergence of the numerical approximations computed from this method to the exact solution U of (5.1.1) on a finite interval. This solution U may be arbitrarily stiff.

5.4.5. Semi-implicit methods with a constant Jacobian approximation.

In this section we regard semi-implicit methods G satisfying (3.3.5) and (3.3.9), which use a constant Jacobian approximation (i.e. $J(x;h,f)$ does not depend on x).

Let $\alpha_0,\delta_0 \ge 0$ and $\beta_0 \in \mathbb{R}$ be given constants. We will use again the stepsize restriction $h \in (0,h_0)$ where $h_0 > 0$ is defined by (5.4.1) with θ as in section 5.4.2 and $\beta = \beta_0$. Further we will assume that (A_0)

holds with $\gamma_0 = \varepsilon_0 = 0$ (see (5.4.18), (5.4.19)). This assumption will be denoted by (A_2).

We might also take the assumption (A_0) with $\gamma_0 = 0$ and $\varepsilon_0 \neq 0$, but this would only complicate the results without providing new insights. If it is known that $f(x) = \Lambda x + w(x)$ (for $x \in \mathbb{K}^s$) where $\Lambda \in L(\mathbb{K}^s)$ and $w: \mathbb{K}^s \to \mathbb{K}^s$ has a small Lipschitz constant near the solution U of (5.1.1), the assumption (A_2) is a natural one.

Results similar to the theorems 5.4.21 and 5.4.22 in this section have been obtained by HAIRER, BADER and LUBICH (1982) for the ROW-methods, and by STREHMEL and WEINER (1982 A) for a class of adaptive Runge-Kutta methods. In these papers it was assumed, in addition to (A_2), that $r_0 = \infty$. It will turn out in the subsequent that this extra condition makes the assumption $|f(x_0)| \leq \alpha_0$ unneccessary. However the class of functions f satisfying (A_2) is reduced considerably if we take $r_0 = \infty$.

Let $h \in (0,h_0)$, and let $S_2(\xi) = S(0,\xi)$ (for $\xi \in \mathbb{R}^+$) where S is the polynomial defined by (5.4.14.a), (5.4.15.d). In order to ensure that the internal vectors $y_i(x_0)$ are in $\mathcal{D}_0$, and $G(\cdot,h,f)$ is continuously differentiable at x_0 we will assume

$$(5.4.26) \qquad (1-h\theta\beta_0)^{-1} h\alpha_0 \, S_2((1-h\theta\beta_0)^{-1} h\delta_0) < r_0$$

(cf. lemma 5.4.3, corollary 5.4.4 and remark 5.4.8 with $\alpha(h) = (1-h\theta\beta_0)^{-1} h\alpha_0$, $\beta = \beta_0$, $\gamma(h) = 0$, $\delta(h) = (1-h\theta\beta_0)^{-1} h\delta_0$, $\varepsilon(h) = 0$). Again (5.4.26) will always hold if $m = 1$, since then $S_2 \equiv 0$. If $m = 2$ S_2 is constant.

From theorem 5.4.7 we obtain the following result that shows what kind of effect on the errorpropagation the use of a constant J will have in case the function f is nonlinear.

THEOREM 5.4.21. *Let G be a semi-implicit method using a constant Jacobian approximation, such that* (3.3.5) *and* (3.3.9) *hold. Assume* $h \in (0,h_0)$, (A_2) and (5.4.26). *Then we have*

$$(5.4.27) \qquad |G'(x_0;h,f)| \leq \Phi(h\beta_0) + P_2((1-h\theta\beta_0)^{-1} h\delta_0)$$

where P_2 is a polynomial with $P_2(0) = 0$, which only depends on the coefficients of G.

The above polynomial P_2 can be defined as $P_2(\xi) = \xi P(0,\xi)$ ($\xi \in \mathbb{R}^+$), where P is the polynomial defined by (5.4.14), (5.4.15.a).

As a consequence of theorem 5.4.12 we get the following contractivity result.

THEOREM 5.4.22. *Let* G *be a semi-implicit method using a constant Jacobian approximation, such that* (3.3.5) *and* (3.3.9) *hold. Suppose* G *is strongly* A*-stable. Then we have*

(B_2) *There are* $c_0, c_3 > 0$ *such that* G *is unconditionally locally contractive at* x_0 *for* f *whenever* (A_2) *holds with* $\beta_0 < 0$, $\alpha_0/(|\beta_0| r_0) \le c_0$ *and* $\delta_0/|\beta_0| \le c_3$.

Furthermore there exists a constant $c_3^* > 0$ *(only depending on the coefficients of* G *) such that*

(C_2) *For any given* $c_0 > 0$, $c_3 \in (0, c_3^*)$, *there is a* $K > 0$ *such that* G *is locally contractive at* x_0 *for* f *and all* $h > 0$ *with* $h|\beta_0| \le K$, *whenever* (A_2) *holds with* $\beta_0 < 0$, $\alpha_0/(|\beta_0| r_0) \le c_0$ *and* $\delta_0/|\beta_0| \le c_3$.

In the statement (C_2) the condition $h|\beta_0| \le K$ may be replaced by $h \max\{\alpha_0/(r_0 c_0), \delta_0/c_3\} \le K$ (see remark 5.4.13). For $r_0 = \infty$ we thus obtain a basic result of HAIRER, BADER and LUBICH (1982) (proved by them for the ROW-methods).

COROLLARY 5.4.23. *Let* G *be a semi-implicit method using a constant Jacobian approximation, such that* (3.3.5) *and* (3.3.9) *hold. Suppose* G *is strongly* A*-stable, and* $r_0 = \infty$. *Then there is a constant* $c_3^* > 0$ *such that*

(D_2) *For any* $c_3 \in (0, c_3^*)$ *there is a* $K > 0$ *such that* G *is locally contractive at* x_0 *for* f *and all* $h > 0$ *with* $h\delta_0 \le K$, *whenever* (A_2) *holds with* $\beta_0 < 0$ *and* $\delta_0/|\beta_0| \le c_3$.

From the proof of theorem 5.4.12 it follows that c_3^* can be taken as $\lambda/P_2'(0)$, where P_2 is the polynomial arising in (5.4.27) and $\lambda > 0$

is such that $\Phi(t) \le 1 + \lambda t + O(t^2)$ $(t \downarrow 0)$.

Note that if $m = 1$, we may take $c_0 = \infty$ in theorem 5.4.22. This follows from remark 5.4.14. Thus in that case there is no need for a bound on $|f(x_0)|$. Also if $r_0 = \infty$, the requirement $|f(x_0)| \le \alpha_0$ vanishes. The reason for this is that the condition on $\alpha_0/(|\beta_0| r_0)$ was only needed to make sure that the vectors $y_i(x_0)$ are in $\mathcal{D}_0$.

By a counter example the following result was proved by HAIRER, BADER and LUBICH (1982).

THEOREM 5.4.24. *The conclusions* (B_2) *and* (C_2) *do not hold on the class of A-stable semi-implicit methods using a constant Jacobian approximation* J , *and satisfying* (3.3.5) *and* (3.3.9).

REMARK 5.4.25. In case G is not strongly A-stable or $\beta_0 = 0$, we have no contractivity results of the type (B_2) , (C_2) or (D_2) . As with the Rosenbrock methods (cf. remark 5.4.19) we do have stability on finite intervals.

Let G be an A-stable semi-implicit method using a constant Jacobian approximation, such that (3.3.5) and (3.3.9) hold. Suppose $\alpha_0, \delta_0, r_0 > 0$ and $\beta_0 \le 0$. Suppose further that $\mathcal{D} \subset \mathbb{K}^s$ is open and convex, and $T > 0$. If (A_2) holds for all $x_0 \in \mathcal{D}$, and $h \in (0,T]$ is such that (5.4.26) is valid, we have

$$|\tilde{u}_n - u_n| \le [1 + P_2((1-h\theta\beta_0)^{-1} h\delta_0)]^n \, |\tilde{u}_0 - u_0|$$

whenever the $\tilde{u}_n, u_n \in \mathcal{D}$ $(1 \le n \le T/h)$ are computed from G with stepsize h . Here P_2 is the polynomial arising in theorem 5.4.21.

Since $P_2(0) = 0$ we have

$$[1 + P_2((1-h\theta\beta_0)^{-1} h\delta_0)]^n \to e^{P_2'(0)\delta_0 t} \qquad (h \downarrow 0 \ , \ nh = t \ )$$

It follows that we have the stability relation

$$|\tilde{u}_n - u_n| \le c|\tilde{u}_0 - u_0| \qquad (1 \le n \le T/h)$$

with a stability factor c which only depends on T, β_0, δ_0 and the coefficients of G .

Note that this stability relation is suited for stiff problems since the Lipschitz constant of f is not involved.

EXAMPLE 5.4.26. We consider the counterpart of the Rosenbrock method treated in example 5.4.20 (see also HAIRER, BADER and LUBICH (1982)). The method is given by

$$G(x;h,f) = x + (I-h\theta J_0)^{-1} hf(x) \qquad (x \in \mathbb{K}^s)$$

with J_0 depending on h and f, and $\theta \in (0,1]$. From remark 5.4.10 (with the appropriate $\alpha(h),\beta,\gamma(h)$ and $\delta(h)$) it is easily seen that

$$|G'(x_0;h,f)| \le \Phi(h\beta_0) + (1-h\theta\beta_0)^{-1} h\delta_0$$

whenever (A_2) holds and $1 - h\theta\beta_0 > 0$. Here $\Phi(t) = \max\{\theta^{-1}(1-\theta),(1-\theta t)^{-1}(1+(1-\theta)t)\}$ (for $t < \theta^{-1}$).

Suppose $\theta > \frac{1}{2}$. If we have

$$\beta_0 < 0 \quad \text{and} \quad \delta_0/|\beta_0| \le 2\theta - 1 ,$$

then $\Phi(h\beta_0) + (1-h\theta\beta_0)^{-1} h\delta_0 \le 1$ for all $h > 0$. Thus we may take $c_0 = \infty$ and $c_3 = 2\theta - 1$ in statement (B_2) of theorem 5.4.22.

For $\theta > \frac{1}{2}$ and $h > 0$ close to zero, we have

$$\Phi(h\beta_0) + (1-h\theta\beta_0)^{-1} h\delta_0 = 1 + (1-h\theta\beta_0)^{-1} h(\beta_0+\delta_0) .$$

Therefore we get no contractivity result at all if $\delta_0/|\beta_0| > 1$. By some calculations it can be shown that if $2\theta - 1 < \delta_0/|\beta_0| \le 1$, we have contractivity at x_0 for all stepsizes $h > 0$ with

$$h|\beta_0| \le \theta^{-1}(2\theta-1)\,[\delta_0/|\beta_0|-(2\theta-1)]^{-1} .$$

Thus we may take $c_3^* = 1$ and $K = \theta^{-1}(2\theta-1)\,[c_3-(2\theta-1)]^{-1}$ (for $c_3 \in (2\theta-1,1]$) in statement (C_2) of theorem 5.4.22.

5.4.6. A third choice for $J(\cdot;h,f)$.

In this section we will regard a choice for the Jacobian approximation where we do not have a bound for the error propagation per step without the Lipschitz constant of f being involved (as in the theorems 5.4.15, 5.4.21).

For $h > 0$, $f: \mathbb{K}^s \to \mathbb{K}^s$ we will consider $J(\cdot) = J(\cdot;h,f)$ given by

$$J(x) = f'(x+chf(x)) \qquad \text{(for } x \in \mathbb{K}^s\text{)}. \tag{5.4.28}$$

This choice was proposed by SCHOLZ (1978). The constant c can be used to increase the order of consistency of the method, without introducing any additional computational work (see e.g. VERWER (1980)). Using a result of VERWER (1977) it was shown that the nonautonomous version of these semi-implicit methods are S-stable as soon as they are strongly A-stable. In the nonautonomous version of semi-implicit methods we replace $J(x)$ by $J(t,x) \approx D_x f(t,x)$, and we read $f(t+c_i h, y_i(x))$ instead of $f(y_i(x))$ (cf. remark 3.3.1).

It will be shown that, in spite of the S-stability, one may expect numerical difficulties for these methods in dealing with stiff systems where f satisfies (A_1) , even with $\alpha_0\gamma_0$ small. At first sight this is somewhat surprising, since

$$|f'(x_0+chf(x_0))-f'(x_0)| \le |c|\, h\alpha_0\gamma_0$$

if $|c|\, h\alpha_0 < r_0$ and f satisfies (A_1) . Thus for moderate stepsizes h and small $\alpha_0\gamma_0$, the difference between $J(x_0)$ and $f'(x_0)$ is small. Nevertheless it will be shown in theorem 5.4.27, that the error propagation for the semi-implicit methods with $J(x) = f'(x+chf(x))$ can be considerably worse than with the methods where $J(x) = f'(x)$.

If we consider only scalar differential equations the situation is much better - see remark 5.4.28. For such equations this choice of J may well be favourable.

The occurrence of the bad error propagation is caused by the following. We have

$$J'(x_0)v = f''(x_0+chf(x_0))[(1+chf'(x_0))v] \qquad \text{(for } v \in \mathbb{K}^s\text{)}.$$

Thus $J'(x_0)v$ can be very large even if (A_1) holds with $\alpha_0\gamma_0$ small, due to the fact that $|f'(x_0)|$ is not restricted. The terms in the expression (5.2.6) for $G'(x_0;h,f)$ where $J'(x_0)v$ is involved will therefore generally blow up for stiff systems.

THEOREM 5.4.27. *Let* G *be an arbitrary semi-implicit method with* $m = 1$, $b_1 \neq 0$, *and* J *given by* (5.4.28) *with* $c \neq 0$. *Then we have for arbitrary* α_0,γ_0,r_0, $H > 0$ *and* $\beta_0 < 0$,

$$\sup\{|G'(x_0;h,f)|: f \text{ satisfies } (A_1), h\in(0,H]\} = \infty .$$

PROOF. From formula (5.2.6) we obtain, for suitable h and f,

$$G'(x_0;h,f)v = \phi(hJ(x_0))v + b_1(hJ(x_0))\,[hf'(x_0)-hJ(x_0)]v + \tag{5.4.29}$$
$$+ [D_x b_1(hJ(x_0))v]\, hf(x_0) .$$

We take the same testequation as in the proof of theorem 5.4.18. Thus $\mathbb{K}^s = \mathbb{R}^2$, $\langle .,. \rangle$ stands for the Euclidean inner product, $x_0 = 0$, and

$$f(x) = f_0 + f_0'x + \tfrac{1}{2}[f_0''x]x \qquad (\text{for } x \in \mathbb{R}^2),$$

where $f_0 = \alpha_0 e_1$, $f_0' = \mathrm{diag}(\beta_0,\sigma\beta_0)$, $[f_0''u]v = \gamma_0[\langle u,e_1\rangle\langle v,e_2\rangle + \langle u,e_2\rangle\langle v,e_1\rangle]e_1$ (for $u,v \in \mathbb{R}^2$). The constant $\sigma > 1$ will be specified later.

For this function f we have

$$J(x_0) = \begin{pmatrix} \beta_0 & hc\alpha_0\gamma_0 \\ 0 & \sigma\beta_0 \end{pmatrix}, \quad J'(x_0)e_2 = \begin{pmatrix} \gamma_0(1+\sigma ch\beta_0) & 0 \\ 0 & 0 \end{pmatrix} .$$

Note that $\mu[J(x_0)] < 0$ if $\beta_0 < 0$ and $h > 0$ is not too large. If $\sigma \to \infty$, $\mu[J(x_0)]$ converges to β_0.

In the same way as in the proof of theorem 5.4.18 we obtain

$$[D_x b_1(hJ(x_0))e_2]\, hf(x_0) = b_1'(h\beta_0)\; h^2\alpha_0\gamma_0(1+ch\sigma\beta_0)e_1 .$$

Further we have

$$b_1(hJ(x_0))\ [hf'(x_0)-hJ(x_0)]e_2 = -b_1(h\beta_0)\ h^2c\alpha_0\gamma_0 e_1 \ .$$

Thus we arrive at

$$(5.4.30)\qquad G'(x_0;h,f)\ e_2 = \phi(hJ(x_0))\ e_2 - b_1(h\beta_0)\ h^2c\alpha_0\gamma_0 e_1 +$$
$$+ b_1'(h\beta_0)\ h^2\alpha_0\gamma_0\ (1+ch\sigma\beta_0)\ e_1 \ .$$

Assume that ϕ is bounded on $\mathbb{C}^-$ (otherwise we even have a scalar, linear counter example). Then b_1 is not constant. Take $h \in (0,H]$ such that $b_1'(h\beta_0) \neq 0$. From (5.4.30) we now see that $|G'(x_0;h,f)| \to \infty$ (for $\sigma \to \infty$). □

REMARK 5.4.28. For $s = 1$ it can easily be proved, by using (5.4.29), that if $m = 1$, (3.3.5) and (3.3.9) hold, $\mu[J(x_0)] \leq \tilde{\beta}_0$, (A_1) , and $x_0 + chf(x_0) \in \mathcal{D}_0$, then

$$|G'(x_0;h,f)| \leq \Phi(h\tilde{\beta}_0) + \Omega(1-h\theta\tilde{\beta}_0)^{-1}\ h^2\alpha_0\gamma_0 \ .$$

Here Ω and θ are constants determined by the method. If f satisfies (A_1) and $|c|\ h\alpha_0 < r_0$ we have $\mu[J(x_0)] \leq \beta_0 + |c|\ h\alpha_0\gamma_0$. It follows that a (conditional) contractivity result of the type (C) as in the theorems 5.4.16, 5.4.22 holds for scalar initial value problems where f satisfies (A_1) . Thus here the situation is totally different from the one in the proof of theorem 5.4.27, where we took $s = 2$.

Thus we see that considering only one-dimensional test problems leads to much too optimistic results on the error propagation for such methods.

The S-stability model problem is one-dimensional (though nonautonomous). Moreover if the nonautonomous form of these semi-implicit methods with $J(t,x) = f_x(t,x+chf(t,x))$, is applied to the S-stability model problem $J(t,x)$ will be constant. Therefore the bad error propagation as in the proof of theorem 5.4.27 cannot occur.

5.4.7. Modifications of the results.

In this section we consider some modifications of the previous results.

We consider a Rosenbrock method satisfying (3.3.5) and (3.3.9). The estimate (5.4.25) for $|G'(x_0;h,f)|$ is only applicable if bounds for

$|f(x_0)|$ and $|f''(x)|$ $(x \approx x_0)$ are available.

In the general situation, where no special structure of f is assumed, these bounds are necessary as will be shown by the following example.

EXAMPLE 5.4.29. Consider an arbitrary Rosenbrock method with $m = 1$, $b_1 \neq 0$. Let $h > 0$, $x_0 \in \mathbb{R}$ and let $g: \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable. We assume that $b_1'(hg'(x_0)) \neq 0$.

Let σ_1 and σ_2 be positive numbers, and let the function $f: \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = g(x) + \sigma_1 + \tfrac{1}{2}\sigma_2(x-x_0)^2 \qquad (\text{for } x \in \mathbb{R}).$$

From formula (5.2.6) it follows that

$$G'(x_0;h,f) = \phi(hg'(x_0)) + b_1'(hg'(x_0))\, h^2(g''(x_0)+\sigma_2)(g(x_0)+\sigma_1) .$$

Thus we see that

$$|G'(x_0;h,f)| \to \infty \qquad (\text{for } \sigma_1 + \sigma_2 \to \infty) .$$

If we assume that the solution of the differential equation $U'(t) = f(U(t))$ passing through x_0 at time $t = t_0$ is slowly varying, then $|U'(t_0)| = |f(x_0)|$ is moderate. Thus if the initial value problem (5.1.1) with $u_0 = x_0$ is stiff, the assumption $|f(x_0)| \leq \alpha_0$, with α_0 moderate, is reasonable. It will be shown at the end of this section that contractivity results are sometimes possible for non-smooth solutions where $|f(x_0)|$ is very large.

The second assumption, $|f''(x)|$ is moderate for x near x_0, may be embarrassing for stiff problems.

EXAMPLE 5.4.30. Consider the following simple system of singularly perturbed differential equations

$$U_1'(t) = g_1(U_1(t),U_2(t)) ,$$

$$U_2'(t) = \sigma g_2(U_2(t)) .$$

Here $g_1: \mathbb{R}^2 \to \mathbb{R}$ and $g_2: \mathbb{R} \to \mathbb{R}$ are smooth functions with bounded second

Gateaux-derivatives, g_2 is dissipative, and σ is a large positive parameter.

Let $f_\sigma(x) = (g_1(x_1,x_2), \sigma g_2(x_2))^T$ (for $x = (x_1,x_2)^T \in \mathbb{R}^2$), $\alpha_0 > 0$, and $\mathcal{D}_\sigma = \{x_0 : x_0 \in \mathbb{R}^2, |f_\sigma(x_0)| \leq \alpha_0\}$. For any $x_0 = (x_1^{(0)}, x_2^{(0)})^T \in \mathbb{R}^2$ with $g_2''(x_2^{(0)}) \neq 0$, we have $|f_\sigma''(x_0)| \to \infty$ (for $\sigma \to \infty$), whereas $\mu[f_\sigma'(x_0)]$ is uniformly bounded for $\sigma > 0$. Thus the estimate (5.4.25) for $G'(x_0;h,f)$ (with $x_0 \in \mathcal{D}_\sigma$, $\beta_0 = \mu[f_\sigma'(x_0)]$, $\gamma_0 = \sup_{x \in \mathbb{R}^2} |f_\sigma''(x)|$) is only applicable if an upper bound for σ is known.

Modification 1. The results of section 5.4.4 for Rosenbrock methods will be modified in such a way that smooth solutions of the above singularly perturbed system can be treated no matter how large σ is. Such a modification was proposed to the author by M. VAN VELDHUIZEN.

The assumption $|f''(x)| \leq \gamma_0$ in (A_1) will be replaced by

$$(5.4.31) \qquad |(I - tf'(x_0))^{-1} tf''(x)| \leq \gamma_0^* \qquad \text{(for all } t > 0,\ x \in \mathcal{D}_0\text{)}.$$

By (A_1^*) we denote the set of assumptions (A_1) with the above modification. Thus (A_1^*) consists of (5.4.2) together with the assumptions that f is twice continuously differentiable on $\mathcal{D}_0$, $|f(x_0)| \leq \alpha_0$, $\mu[f'(x_0)] \leq \beta_0$ (cf. (5.4.20)-(5.4.22)), and (5.4.31). For convenience we will assume that $\beta_0 \leq 0$. We then known that $I - tf'(x_0)$ is regular for all $t > 0$ (see corollary 2.2.11).

Further we will consider again a Rosenbrock method G satisfying (3.3.5) and (3.3.9), and $\theta > 0$ is defined as in section 5.4.2. The general results of section 5.4.2 may now be applied with $\alpha(h) = (1 - h\theta\beta_0)^{-1} h\alpha_0$, $\beta = \beta_0$, $\gamma(h) = \varepsilon(h) = \theta^{-1}\gamma_0^*$, $\delta(h) = 0$ and $h_0 = \infty$.

In order to ensure that the intermediate vectors $y_i(x_0)$ all belong to $\mathcal{D}_0$, and that G is continuously differentiable at x_0, we now get the following requirement

$$(5.4.32) \qquad (1 - h\theta\beta_0)^{-1} h\alpha_0 \, S_1^*((1 - h\theta\beta_0)^{-1} h\alpha_0\gamma_0^*) < r_0,$$

where $S_1^*(t) = S_1(\theta^{-1}t)$ $(t \in \mathbb{R}^+)$, and S_1 is the polynomial we encountered in (5.4.24) (cf. also remark 5.4.8, (5.4.13)).

From theorem 5.4.7 we obtain a modified version of theorem 5.4.15 for the Rosenbrock methods. The polynomial P_1^* is given by $P_1^*(t) = P_1(\theta^{-1}t)$

: $\mathbb{R}^+$) , P_1 being the polynomial from theorem 5.4.15.

)REM 5.4.31. *Let* G *be a Rosenbrock method satisfying* (3.3.5), (3.3.9). *me* (A_1^*) *with* $\beta_0 \le 0$, *and* (5.4.32). *Then*

.33) $$|G'(x_0;h,f)| \le \Phi(h\beta_0) + P_1^*((1-h\theta\beta_0)^{-1}h\alpha_0\gamma_0^*)$$

e the polynomial P_1^* *is determined by the coefficients of the method nd satisfies* $P_1^*(0) = 0$.

The proof of the following theorem is essentially the same as the proof heorem 5.4.12.

)REM 5.4.32. *Suppose* G *is a strongly* A*-stable Rosenbrock method satis-ig* (3.3.5) *and* (3.3.9). *Then the following holds.*

) *There are* $c_0, c_1 > 0$ *such that* G *is unconditionally locally contractive at* x_0 *for* f , *whenever* (A_1^*) *holds with* $\beta_0 < 0$, $\alpha_0/(|\beta_0|r_0) \le c_0$ *and* $\alpha_0\gamma_0^*/|\beta_0| \le c_1$.

over there is a constant $c_1^* > 0$ *such that*

) *For any* $c_0 > 0$ *and* $c_1 \in (0,c_1^*)$ *there is a* $K > 0$ *such that* G *is locally contractive at* x_0 *for* f *and all* $h > 0$ *with* $h|\beta_0| \le K$, *whenever* (A_1^*) *holds with* $\beta_0 < 0$, $\alpha_0/(|\beta_0|r_0) \le c_0$ *and* $\alpha_0\gamma_0^*/|\beta_0| \le c_1$.

MPLE 5.4.33. Consider again the θ-Rosenbrock method of example 5.4.20,

$$G(x;h,f) = x + (I-h\theta f'(x))^{-1} hf(x)$$

h $\theta \in (0,1]$. Assuming that (A_1^*) holds with $\beta_0 \le 0$, we obtain m the upper bound given in example 5.4.10,

$$|G'(x_0;h,f)| \le \Phi(h\beta_0) + (1-h\theta\beta_0)^{-1} h\alpha_0\gamma_0^* .$$

s estimate has the same form as the one in example 5.4.26. If $\theta > \frac{1}{2}$ e. the method is strongly A-stable) and $\alpha_0\gamma_0^*/|\beta_0| \le 2\theta - 1$ we thus have

unconditional contractivity at x_0 . In case $2\theta - 1 < \alpha_0\gamma_0/|\beta_0| \leq 1$ we have contractivity for the stepsizes $h > 0$ such that $h|\beta_0| \leq$
$\leq \theta^{-1}(2\theta-1)\,[(\alpha_0\gamma_0^*/|\beta_0|) - (2\theta-1)]^{-1}$.

EXAMPLE 5.4.34. Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be given by

$$f(x) = \begin{pmatrix} \frac{1}{3} - x_1 + x_1x_2 \\ \\ \sigma(-x_2+x_2^2) \end{pmatrix} \qquad \left(\text{for } x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2\right)$$

with $\sigma > 1$, and let $x_0 = 0 \in \mathbb{R}^2$. On the $\mathbb{R}^2$ we consider the Euclidean norm.

In example 5.4.30 we have seen that the results of section 5.4.4 cannot be applied if σ is allowed to be arbitrary large. With the modified results we can prove for strongly A-stable Rosenbrock methods contractivity at x_0 uniformly for $\sigma > 0$.

By some calculations we obtain

$$(I-tf'(x_0))^{-1}\,[tf''(x)u]v = \begin{pmatrix} (1+t)^{-1}\; t(u_1v_2+u_2v_1) \\ \\ (1+t\sigma)^{-1}\; 2t\sigma u_2v_2 \end{pmatrix}$$

for $u = (u_1,u_2)^T$, $v = (v_1,v_2)^T \in \mathbb{R}^2$, and $x \in \mathbb{R}^2$ arbitrary. It follows that (A_1^*) holds with $r_0 = \infty$, $\alpha_0 = \frac{1}{3}$, $\beta_0 = -1$ and $\gamma_0^* = \sqrt{5}$ for any $\sigma > 0$.

Let G be the θ-Rosenbrock method of example 5.4.33. Suppose $\theta > \frac{1}{2} + \frac{1}{6}\sqrt{5}$. Then we have unconditional contractivity for this problem near $x_0 = 0$, no matter how large $\sigma > 0$ is.

Modification 2. We now consider an other modification of the results of section 5.4.4. For convenience we confine ourselves again to the Rosenbrock methods. This modification can however also be considered for the methods where J is constant.

It will be shown that the Rosenbrock methods can be locally contractive at a point $x_0 \in \mathbb{K}^s$ for $|f(x_0)|$ extremely large, if we have a situation that the (in modulus) large eigenvalues of $f'(x_0)$ are the cause for this large $|f(x_0)|$. Note that in example 5.4.29 this was not the case. The possibility of treating the case where $|f(x_0)|$ is large was pointed out by

E. HAIRER (private communications), who also suggested a modification of the following kind.

In the set of assumptions (A_1) we now replace the assumption $|f(x_0)| \le \alpha_0$ by

$$(5.4.34) \qquad |(I-tf'(x_0))^{-1} tf(x_0)| \le \alpha_0^* \qquad \text{(for all } t > 0\text{)}.$$

(A_1^{**}) stands for the assumptions (5.4.2), (5.4.34) together with the assumptions that f is twice continuously differentiable on $\mathcal{D}_0$, $\mu[f'(x_0)] \le \beta_0$ and $|f''(x)| \le \gamma_0$ (for all $x \in \mathcal{D}_0$) (cf. (5.4.20), (5.4.22), (5.4.23)). As in modification 1 we assume that $\beta_0 \le 0$.

The assumption (5.4.34) may hold for moderate α_0^* while $|f(x_0)|$ is very large.

EXAMPLE 5.4.35. Let $x_0 = (1,1)^T$ and $f: \mathbb{R}^2 \to \mathbb{R}^2$ be defined by

$$f(x) = (-x_1+x_2^2, -10^6 x_2) \qquad \text{(for } x = (x_1,x_2)^T \in \mathbb{R}^2\text{)}.$$

Then $\mu[f'(x_0)] \le -1$, $|f''(x)| \le 2$ (for all $x \in \mathbb{R}^2$). Further (5.4.34) holds with $\alpha_0^* = \sqrt{5}$, whereas $|f(x_0)| \approx 10^6$.

Starting from the assumptions (A_1^{**}) instead of (A_1), results similar to the theorems 5.4.15, 5.4.16 can be proved for the Rosenbrock methods G satisfying (3.3.5), (3.3.9). We apply the general results of section 5.4.2 with $\alpha(h) = \theta^{-1}\alpha_0^*$, $\beta = \beta_0$, $\gamma(h) = \varepsilon(h) = (1-h\theta\beta_0)^{-1}\gamma_0$, $\delta(h) = 0$ and $h_0 = \infty$.

Let P_1^*, S_1^* be the polynomials arising in (5.4.32), (5.4.33). If we have

$$(5.4.35) \qquad \alpha_0^* S_1^*((1-h\theta\beta_0)^{-1} h\alpha_0^*\gamma_0) < \theta r_0,$$

then all $y_i(x_0) \in \mathcal{D}_0$ and G is continuously differentiable at x_0 (see section 5.4.2 with $S_1^*(\xi) = S(\theta^{-1}\xi, 0)$ $(\xi \in \mathbb{R}^+)$). Note that (5.4.35) only holds if α_0^* is sufficiently small, unless $m = 1$ $(S_1^* \equiv 0)$.

Using theorem 5.4.7 it follows that

$$(5.4.36) \qquad |G'(x_0;h,f)| \le \Phi(h\beta_0) + P_1^*((1-h\theta\beta_0)^{-1} h\alpha_0^*\gamma_0),$$

whenever (A_1^{**}) holds with $\beta_0 \le 0$, and (5.4.35) is fulfilled.

With the above upper bound (5.4.36) we can prove, in the same way as theorem 5.4.12, the following contractivity properties of strongly A-stable Rosenbrock methods.

THEOREM 5.4.36. *Let* G *be a strongly* A-*stable Rosenbrock method satisfying* (3.3.5) *and* (3.3.9). *Then we have*

(B_1^{**}) *There are* $c_0, c_1 > 0$ *such that* G *is unconditionally locally contractive at* x_0 *for* f, *whenever* (A_1^{**}) *holds with* $\beta_0 < 0$, $\alpha_0^*/r_0 \le c_0$ *and* $\alpha_0^*\gamma_0/|\beta_0| \le c_1$.

Moreover there are constants $c_0^*, c_1^* > 0$, *which only depend on the coefficients of* G, *such that*

(C_1^{**}) *For* $c_0 \in (0, c_0^*)$, $c_1 \in (0, c_1^*)$ *given there is a* $K > 0$ *such that* G *is locally contractive at* x_0 *for* f *and all* $h > 0$ *with* $h|\beta_0| \le K$, *whenever* (A_1^{**}) *holds with* $\beta_0 < 0$, $\alpha_0^*/r_0 \le c_0$ *and* $\alpha_0^*\gamma_0/|\beta_0| \le c_1$.

The extra condition $c_0 < c_0^*$ in statement (C_1^{**}) which was not present in (C_1^*), is needed here because (5.4.35) does not automatically hold for $h > 0$ sufficiently small. For $m = 1$ we have $S_1^* \equiv 0$, and we may then take $c_0^* = \infty$.

The same modification - replacing $|f(x_0)| \le \alpha_0$ by (5.4.34) - can also be used for the semi-implicit methods using a constant Jacobian approximation J, thus yielding contractivity results where $|f(x_0)|$ may be large.

5.4.8. A numerical illustration.

The results on the various choices for the Jacobian approximation J will be illustrated by applying some simple semi-implicit methods to the differential equation

$$U_1'(t) = 1 - U_1(t) + \gamma U_1(t)U_2(t) ,$$

$$U_2'(t) = -10^6\, U_2(t) .$$

The parameter γ will be used to vary the nonlinearity of the problem. All solutions of this equation are stable and converge to the stationary solution $u^* = (1,0)^T$. The second component of a solution is always damped out very quickly.

As initial vectors we take $u_0 = (0,0)^T$ and $\tilde{u}_0 = (0,10^{-3})^T$. Let $\tilde{U},U$ stand for the solutions of the differential equation with $\tilde{U}(0) = \tilde{u}_0$, $U(0) = u_0$. For the first component U_1 of the solution U we have $U_1(t) = 1 - e^{-t}$, and for those values of γ that will be considered ($\gamma \in (0,10^6)$) we have $|\tilde{U}_1(t)-U_1(t)| \lesssim 10^{-3}$ (for all $t > 0$). The rounded values of $U_1(t)$ are listed in the following table for $t = 0.1,0.5,1,2$.

t	0.1	0.5	1	2
$U_1(t)$	0.095	0.393	0.632	0.865

The semi-implicit methods we will consider are given by

$$G(x;h,f) = x + (I-h\theta J(x))^{-1} hf(x) ,$$

where we take $\theta = \frac{1}{2}$ or 1, and $J(x) = f'(x)$, $f'(u_0)$ $(= \mathrm{diag}(-1,-10^6))$, $f'(x+hf(x))$. For the stepsize we take $h = 1/10$. The order of these methods is 2 if $\theta = \frac{1}{2}$ and $J(x) = f'(x)$ or $f'(x+hf(x))$, and 1 otherwise.

In the following tables the first components $u_{n,1} \approx U_1(nh)$ and $\tilde{u}_{n,1} \approx \tilde{U}_1(nh)$ are listed. If $J(x) = f'(x)$ or $J(x) = f'(u_0)$, and γ is not too large, these approximations also stay close to each other and converge to 1, as the exact solutions do. For $J(x) = f'(x+hf(x))$ the approximations are very inaccurate, even if γ is moderate.

CASE 1: $J(x) = f'(x)$.

$\theta = \frac{1}{2}$	all γ $u_{n,1}$	$\gamma = 1$ $\tilde{u}_{n,1}$	$\gamma = 10^2$ $\tilde{u}_{n,1}$	$\gamma = 10^4$ $\tilde{u}_{n,1}$	$\gamma = 10^6$ $\tilde{u}_{n,1}$
t = 0.1	0.095	0.095	0.096	0.182	$-2.0_{10}-3$
t = 0.5	0.394	0.394	0.394	0.521	$-2.4_{10}-3$
t = 1	0.632	0.632	0.632	0.737	$-4.0_{10}-4$
t = 2	0.865	0.865	0.865	0.931	$-8.0_{10}-4$

$\theta = 1$	all γ $u_{n,1}$	$\gamma = 1$ $\tilde{u}_{n,1}$	$\gamma = 10^2$ $\tilde{u}_{n,1}$	$\gamma = 10^4$ $\tilde{u}_{n,1}$	$\gamma = 10^6$ $\tilde{u}_{n,1}$
t = 0.1	0.091	0.091	0.092	1.000	-0.001
t = 0.5	0.379	0.379	0.380	1.000	0.316
t = 1	0.614	0.614	0.615	1.000	0.576
t = 2	0.851	0.851	0.851	1.000	0.836

CASE 2: $J(x) = f'(u_0)$.

$\theta = \frac{1}{2}$	all γ $u_{n,1}$	$\gamma = 1$ $\tilde{u}_{n,1}$	$\gamma = 10^2$ $\tilde{u}_{n,1}$	$\gamma = 10^4$ $\tilde{u}_{n,1}$	$\gamma = 10^6$ $\tilde{u}_{n,1}$
t = 0.1	0.095	0.095	0.095	0.095	0.095
t = 0.5	0.394	0.394	0.395	0.249	$7.7_{10}6$
t = 1	0.632	0.632	0.629	0.083	$-6.0_{10}16$
t = 2	0.865	0.869	0.860	0.084	$3.6_{10}36$

$\theta = 1$	all γ $u_{n,1}$	$\gamma = 1$ $\tilde{u}_{n,1}$	$\gamma = 10^2$ $\tilde{u}_{n,1}$	$\gamma = 10^4$ $\tilde{u}_{n,1}$	$\gamma = 10^6$ $\tilde{u}_{n,1}$
t = 0.1	0.091	0.091	0.091	0.091	0.091
t = 0.5	0.379	0.379	0.379	0.379	0.379
t = 1	0.614	0.614	0.614	0.614	0.614
t = 2	0.851	0.851	0.851	0.851	0.851

CASE 3: $J(x) = f'(x+hf(x))$.

$\theta = \frac{1}{2}$	all γ $u_{n,1}$	$\gamma = 0.1$ $\tilde{u}_{n,1}$	$\gamma = 1$ $\tilde{u}_{n,1}$
t = 0.1	0.095	0.065	0.017
t = 0.5	0.394	0.452	-0.025
t = 1	0.632	0.737	-0.043
t = 2	0.865	0.931	-0.087

$\theta = 1$	all γ $u_{n,1}$	$\gamma = 0.1$ $\tilde{u}_{n,1}$	$\gamma = 1$ $\tilde{u}_{n,1}$
t = 0.1	0.091	0.048	0.009
t = 0.5	0.379	0.349	0.323
t = 1	0.614	0.596	0.579
t = 2	0.815	0.844	0.837

Comparison of the methods with $\theta = 1$ and those with $\theta = \frac{1}{2}$ shows that if t and γ are large, the approximations with $\theta = 1$ are better. This is in agreement with the results on the error propagation of the previous sections (cf. the examples 5.4.20, 5.4.26). It should be noted that also the local errors are responsible for this behaviour. If $\theta = 1$ the second component of the numerical solution will always be damped out very quickly since the method is then L-stable. If $\theta = \frac{1}{2}$ this is not the case, and this will disturb the computation of the first component $\tilde{u}_{n,1}$.

Besides the bad behaviour of the methods with $J(x) = f'(x+hf(x))$, which was predicted in theorem 5.4.27, the most striking thing in these tables is the very robust character of the approximations $\tilde{u}_{n,1}$ computed with $\theta = 1$ and $J(x) = f'(u_0)$. Note that for large γ we have $|f''(\tilde{u}_0)||f(\tilde{u}_0)| \approx \gamma\, 10^3$ and $|f'(\tilde{u}_0)-f'(u_0)| = \gamma\, 10^{-3}$. Therefore the assumptions (A_1) on the Rosenbrock method only hold with $\alpha_0\gamma_0 \approx \gamma\, 10^3$ if $x_0 = \tilde{u}_0$, whereas the constant δ_0, arising in the assumptions (A_2) for the semi-implicit method which uses $f'(u_0)$ as the fixed Jacobian approximation, can be taken much smaller, $\delta_0 \approx \gamma\, 10^{-3}$.

The above tables do of course not tell us whether the semi-implicit methods are suited for more complicated stiff systems. Numerical tests for this purpose can be found for instance in GOTTWALD and WANNER (1981).

5.5. IMPLICIT RUNGE-KUTTA METHODS

5.5.1. An upper bound for the error propagation per step.

In this section we shall use the notation introduced in section 2.4. This notation is consistent with the one which has been used in the rest of this chapter.

Let G be an implicit m-stage Runge-Kutta method given by

$$(5.5.1.a) \qquad G(x;h,f) = x + b^T hF(y(x)) ,$$

where $y(x)$ satisfies

$$(5.5.1.b.) \qquad y(x) = ex + A\, hF(y(x))$$

for $x \in \mathbb{K}^s$ and $h > 0$, $f: \mathbb{K}^s \to \mathbb{K}^s$, $s \in \mathbb{N}$. Here $y(x) = (y_1(x)^T, y_2(x)^T, \ldots, y_m(x)^T)^T$ and $F(y(x)) = (f(y_1(x))^T, f(y_2(x))^T, \ldots$

$..,f(y_m(x))^T)^T$ are vectors in the $\mathbb{K}^{sm}$, and $A = (a_{ij}) \in L(\mathbb{R}^m)$, $b = (b_i) \in \mathbb{R}^m$ contain the coefficients of the method.

From the theorems 5.2.1 and 5.2.2 the following result is obtained.

THEOREM 5.5.1. *Let* $h > 0$, $x_0 \in \mathbb{K}^s$ *and* $z_0 \in L(\mathbb{K}^s)$. *Assume* $f: \mathbb{K}^s \to \mathbb{K}^s$ *is continuously differentiable on* $\mathbb{K}^s$, *and* (5.5.1.b) *has a unique solution* $y(x)$ *which depends continuously on* x *(for all* $x \in \mathbb{K}^s$*). Assume further that* $I - AZ_0$ *and* $I - AZ$ *are regular, where* $Z_0 = \mathrm{diag}(z_0,z_0,\ldots,z_0) \in \in L(\mathbb{K}^{sm})$, $Z = \mathrm{diag}(z_1,z_2,\ldots,z_m)$ *with* $z_i = hf'(y_i(x_0))$ $(1 \le i \le m)$. *Then* $G(\cdot) = G(\cdot;h,f)$ *is continuously differentiable at* x_0, *and we have*

$$G'(x_0;h,f) = I + b^T Z(I-AZ)^{-1} e, \tag{5.5.2}$$

$$G'(x_0;h,f) = \phi(z_0) + b^T(I-AZ_0)^{-1}(Z-Z_0)(I-AZ)^{-1} e \tag{5.5.3}$$

where ϕ *stands for the stability function of* G.

REMARK 5.5.2. Assume that f satisfies the one-sided Lipschitz condition (5.1.3) with $\mathcal{D} = \mathbb{K}^s$. Then the matrices $z_i \in L(\mathbb{K}^s)$ in theorem 5.5.1 satisfy $\mu[z_i] \le h\beta$ $(1 \le i \le m)$ (cf. lemma 2.3.3).

If f is not continuously differentiable we still know from corollary 2.4.8 that for arbitrary $\tilde{x},x \in \mathbb{K}^s$ there are $\tilde{z}_i \in L(\mathbb{K}^s)$ with $\mu[\tilde{z}_i] \le \le h\beta$ such that $hf(y_i(\tilde{x})) - hf(y_i(x)) = \tilde{z}_i(y_i(\tilde{x})-y_i(x))$ $(1 \le i \le m)$, provided that $y(\tilde{x}),y(x)$ are well defined by (5.5.1.b). Let $\tilde{Z} = \mathrm{diag}(\tilde{z}_1,\tilde{z}_2,\ldots,\tilde{z}_m)$. If $I - A\tilde{Z}$ is regular we obtain

$$G(\tilde{x};h,f) - G(x;h,f) = (I+b^T\tilde{Z}(I-A\tilde{Z})^{-1}e)(\tilde{x}-x).$$

From theorem 5.5.1 we obtain a similar result, but there we also know that $\tilde{z}_i \approx hf'(y_i(x))$ if $\tilde{x}$ is close to x. This property is useful for the nonlinear stability analysis of Runge-Kutta methods which are not B-contractive (see section 5.5.3).

Consider the function $\Psi: \mathbb{R} \to \mathbb{R}^+ \cup \{\infty\}$ defined by

$$\Psi(t) = \sup\{|I+b^TZ(I-AZ)^{-1}e| : Z=\mathrm{diag}(z_1,z_2,\ldots,z_m), z_i \in L(\mathbb{K}^s), \tag{5.5.4}$$

$$\mu[z_i] \le t\ (1 \le i \le m),\ s \in \mathbb{N},\ \text{and } (I-AZ)^{-1} \text{ exists}\}.$$

From (5.5.2) or remark 5.5.2 we see by comparison with corollary 5.3.4 that Ψ is an analogue of Φ (cf. (5.3.2)) for nonlinear differential equations; similar to corollary 5.3.4 we now have

COROLLARY 5.5.3. *Let* $h > 0$, $\beta \in \mathbb{R}$, *and suppose that* $f: \mathbb{K}^s \to \mathbb{K}^s$ *satisfies* $\mathrm{Re} \langle f(\tilde{x})-f(x), \tilde{x}-x \rangle \leq \beta |\tilde{x}-x|^2$ *(for all* $\tilde{x}, x \in \mathbb{K}^s$*). Then*

$$|G(\tilde{x};h,f)-G(x;h,f)| \leq \Psi(h\beta)\ |\tilde{x}-x|$$

whenever $G(\tilde{x};h,f)$ *and* $G(x;h,f)$ *are defined by* (5.5.1).

REMARK 5.5.4. If $m = 1$ we obtain from (5.5.2) the relation $G'(x_0;h,f) = \phi(z_1)$. It is easily seen that we now even have $\Psi \equiv \Phi$.

EXAMPLE 5.5.5. Let $m = 1$, $A = \theta \in (0,1]$ and $b = 1$. Let G be defined by (5.5.1). If $\theta = \frac{1}{2}$ this is the implicit midpoint rule, and if $\theta = 1$ the backward Euler method. The stability function of this method is given by

$$\phi(\zeta) = (1-\theta\zeta)^{-1}\ (1+(1-\theta)\zeta) \quad (\zeta \in \mathbb{C}).$$

We therefore have (see example 5.4.10 and remark 5.5.4)

$$\Psi(t) = \Phi(t) = \max\{\theta^{-1}(1-\theta), (1-\theta t)^{-1}(1+(1-\theta)t)\} \quad (\text{for } t < \theta^{-1}).$$

The upper bound for the error propagation per step we obtain for this method from corollary 5.5.3 can also be found in BURRAGE and BUTCHER (1979).

From corollary 5.5.3 we see in particular that if the method G is such that $\Psi(0) = 1$, then G is B-contractive. This result (even a stronger version) has been known already as we shall see in the next section.

5.5.2. Algebraically contractive Runge-Kutta methods.

In this section some nonlinear contractivity results on implicit Runge-Kutta methods will be reviewed.

DEFINITION 5.5.6. *The Runge-Kutta method* G *with coefficients* a_{ij}, b_i $(1 \leq i,j \leq m)$ *is called algebraically contractive if all* b_i *are nonnegative*

and the matrix $A^TB + BA - bb^T$ *is positive semi-definite. Here* $B = \mathrm{diag}(b_1, b_2, \ldots, b_m)$.

This definition is due to BURRAGE and BUTCHER (1979), who used the term algebraically stable, and CROUZEIX (1979). In the same papers the following theorem 5.5.7 was proved.

THEOREM 5.5.7. *Suppose the Runge-Kutta method* G *is algebraically contractive. Then* G *is* B*-contractive.*

For methods which are not equivalent to a method with fewer stages, algebraic contractivity is also necessary for B-contractivity (see HUNDSDORFER and SPIJKER (1981)), and we may strengthen the definition of algebraic contractivity by requiring that all the coefficients b_i are strictly positive (see DAHLQUIST and JELTSCH (1979)). Further important contributions to the theory of algebraically contractive Runge-Kutta methods can be found in HAIRER and WANNER (1981), HAIRER (1982), and HAIRER and TUERKE (1983).

The following theorem can easily be proved using the results of BURRAGE and BUTCHER (1979), or, more directly, by applying the lemma 3.2 of DAHLQUIST and JELTSCH (1979).

THEOREM 5.5.8. *The Runge-Kutta method* G *is algebraically contractive iff* $|1+b^TZ(I-AZ)^{-1}e| \le 1$ *for all* $Z = \mathrm{diag}(\zeta_1, \zeta_2, \ldots, \zeta_m)$ *such that* $\zeta_i \in \mathbb{C}$, $\mathrm{Re}\ \zeta_i \le 0$ $(1 \le i \le m)$, *and* $I - AZ$ *is regular.*

It thus follows from the theorems 5.5.7, 5.5.8 that for $t = 0$ we may restrict ourselves to $z_i \in \mathbb{C}$ instead of $z_i \in L(\mathbb{K}^s)$ $(s \in \mathbb{N})$ in the definition of $\Psi(t)$, without violating corollary 5.5.3.

5.5.3. A special class of non-B-contractive methods.

There are popular Runge-Kutta methods which are not B-contractive, but work rather well on stiff problems. The trapezoidal rule is the best known example of such a method. It is used for instance in the code TRAPEX (see e.g. ENRIGHT, HULL and LINDBERG (1975)).

In view of this observation we will present in this section a short analysis for a class of non-B-contractive Runge-Kutta methods including the trapezoidal rule. This analysis is similar to the one given in section 5.4 for the semi-implicit methods.

Let $\theta > 0$ be a given constant. In the following we consider the class of Runge-Kutta methods that satisfy the condition

(M) $$e_1^T A = 0 \ , \ e_m^T A = b^T \ , \text{ and}$$

$$\bar{A} = \begin{pmatrix} a_{22} & a_{23} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m2} & a_{m3} & \dots & a_{mm} \end{pmatrix} \in A_{m-1}(1/\theta) \ .$$

We recall (cf. section 2.4.2) that $\bar{A} \in A_{m-1}(1/\theta)$ iff $\bar{A} \in L(\mathbb{R}^{m-1})$ is regular and there exists a positive definite matrix $\bar{D} = \mathrm{diag}(d_1, d_2, \dots, d_{m-1})$ such that

$$(v, \bar{D}\bar{A}v) \geq (1/\theta)(\bar{A}v, \bar{D}\bar{A}v) \qquad (\text{for all } v \in \mathbb{R}^{m-1}) .$$

This class of Runge-Kutta methods has been considered in CROUZEIX and RAVIART (1980), where also some remarks on the computational efficiency and results on the order of such methods can be found.

Using $e_1^T A = 0$ (the first row of A is zero), it can easily be shown that a method which satisfies (M) cannot be algebraically contractive. Therefore, unless the method is reducible to a method with fewer stages, a method satisfying (M) is not B-contractive.

On the function f the following assumptions (5.5.5)-(5.5.9) will be made. We denote this set of assumptions by (A_3) .

(5.5.5) The norm $|\cdot|$ on the $\mathbb{K}^s$ is generated by an inner product $\langle \cdot, \cdot \rangle$, $s \in \mathbb{N}$, $x_0 \in \mathbb{K}^s$, $r_0 > 0$, and $\mathcal{D}_0 = \{x : x \in \mathbb{K}^s, |x - x_0| < r_0\}$,

(5.5.6) $f: \mathbb{K}^s \to \mathbb{K}^s$ is twice continuously differentiable on $\mathbb{K}^s$,

(5.5.7) $$|f(x_0)| \leq \alpha_0 \ ,$$

(5.5.8) $$\mu[f'(x)] \leq \beta_0 \qquad (\text{for all } x \in \mathbb{K}^s) \ ,$$

(5.5.9) $$|f''(x)| \leq \gamma_0 \qquad (\text{for all } x \in \mathcal{D}_0) .$$

Here $\alpha_0, \gamma_0 \geq 0$ and $\beta_0 \in \mathbb{R}$ are given constants.

Further we will use the stepsize restriction $h \in (0, h_0)$ where

(5.5.10) $$h_0 = 1/(2\theta\beta_0) \text{ if } \beta_0 > 0 \ , \ h_0 = \infty \text{ if } \beta_0 \leq 0 \ .$$

REMARK 5.5.9. By Theorem 4.3.1 and corollary 4.3.5 we know that the system of algebraic equations (5.5.1.b) has a unique solution whenever G satisfies (M), f satisfies (A_3), and $h \in (0,h_0)$.

If we assume in advance that (5.5.1.b) has a unique solution, the one-sided Lipschitz condition (5.5.8) would only need to hold for all $x \in \mathcal{D}_0$ in order that the following corollary 5.5.11 and theorem 5.5.12 remain valid. Lemma 5.5.10 would then have to be reformulated a little bit.

LEMMA 5.5.10. *Let* $\theta > 0$, *and let* G *be a Runge-Kutta method of type* (M). *We assume* (A_3) *and* $h \in (0,h_0)$. *There is a constant* $\omega > 0$, *which only depends on the coefficients of* G, *such that*

$$|y_i(x_0)-x_0| \leq (1-h\theta\beta_0)^{-1}\ \omega h\alpha_0 \qquad (1\leq i\leq m) .$$

PROOF. Let $\bar{a}_1 = (a_{21},a_{31},\ldots,a_{m1})^T$, $\bar{e} = (1,1,\ldots,1)^T \in \mathbb{R}^{m-1}$ and let $\bar{D} \in L(\mathbb{R}^{m-1})$ be a positive definite, diagonal matrix such that $(v,\bar{D}\bar{A}v) \geq$ $\geq (1/\theta)(\bar{A}v,\bar{D}\bar{A}v)$ (for all $v \in \mathbb{R}^{m-1}$). We will use the notations and conventions introduced in section 2.4.1.

From (5.5.1.b) we obtain $y_1(x_0) = x_0$, and

$$y_i(x_0) = x_0 + a_{i1}\ hf(x_0) + \sum_{j=2}^{m} a_{ij}\ hf(y_j(x_0)) \quad (2\leq i\leq m) .$$

We denote $\eta_i = y_i(x_0) - x_0$ $(2\leq i\leq m)$, $\bar{\eta} = (\eta_2^T,\eta_3^T,\ldots,\eta_m^T)$, $\bar{F}(\bar{\eta}) =$ $= (f(\eta_2+x_0)^T, f(\eta_3+x_0)^T,\ldots,f(\eta_m+x_0)^T)^T$, and $\chi(\bar{\eta}) =$ $= (\bar{A})^{-1}\ \bar{\eta} - h\bar{F}(\bar{\eta}) - (\bar{A})^{-1}\ \bar{a}_1\ hf(x_0)$. We then have

$$\mathrm{Re}[\bar{\eta},\chi(\bar{\eta})]_{\bar{D}} = \mathrm{Re}[\bar{\eta},(\bar{A})^{-1}\ \bar{\eta}]_{\bar{D}} - \mathrm{Re}[\bar{\eta},h\bar{F}(\bar{\eta})-h\bar{F}(0)]_{\bar{D}} -$$

$$- \mathrm{Re}[\bar{\eta},h\bar{F}(0)+(\bar{A})^{-1}\ \bar{a}_1 hf(x_0)]_{\bar{D}} \geq$$

$$\geq (\theta^{-1}-h\beta_0)(\|\bar{\eta}\|_{\bar{D}})^2 - \|\bar{\eta}\|_{\bar{D}}\ \|h\bar{F}(0)+(\bar{A})^{-1}\ \bar{a}_1 hf(x_0)\|_{\bar{D}} .$$

Hence the solution of $\chi(\bar{\eta}) = 0$ satisfies

$$\|\bar{\eta}\|_{\bar{D}} \leq (1-h\theta\beta_0)^{-1}\ \theta\|\bar{e}+(\bar{A})^{-1}\ \bar{a}_1\|_{\bar{D}}\ |hf(x_0)| .$$

From this inequality the proof follows with $\omega = \theta\|\bar{e}+(\bar{A})^{-1}\bar{a}_1\|_{\bar{D}}$ $(\min_{1\le i\le m-1} d_i)^{-1}$. □

COROLLARY 5.5.11. *If the assumptions of lemma* 5.5.10 *hold, and* $(1-h\theta\beta_0)^{-1}\ \omega h\alpha_0 < r_0$, *then*

$$|hf'(y_i(x_0))-hf'(x_0)| \le (1-h\theta\beta_0)^{-1}\ \omega h^2\alpha_0\gamma_0 \quad (1\le i\le m) .$$

THEOREM 5.5.12. *Let* $\theta > 0$, *and let* G *be a Runge-Kutta method satisfying* (M) . *Assume* f *satisfies* (A_3) , $h \in (0,h_0)$, *and* $(1-h\theta\beta_0)^{-1}\ \omega h\alpha_0 < r_0$, *where* ω *is the constant from lemma* 5.5.10 . *Then* $G(\cdot;h,f)$ *is continuously differentiable at* x_0 , *and there are constants* $\Omega_1,\Omega_2 \ge 0$ *such that*

$$(5.5.11)\qquad |G'(x_0;h,f)| \le \Phi(h\beta_0) + \Omega_1(1-h\theta\beta_0)^{-2}\ h^2\alpha_0\gamma_0 + \Omega_2(1-h\theta\beta_0)^{-4}\ h^4\alpha_0^2\gamma_0^2 .$$

PROOF. By lemma 2.4.4 we know that $I - A\tilde{Z}$ is regular for all matrices $\tilde{Z} = \mathrm{diag}(\tilde{z}_1,\tilde{z}_2,\ldots,\tilde{z}_m)$ with $\tilde{z}_i \in L(\mathbb{K}^s)$, $\mu[z_i] \le 1/(2\theta)$ $(1\le i\le m)$. In order to apply theorem 5.2.2 we will show that $y(\cdot)$ is continuous near x_0 .

For $\rho_0 > 0$ we put $E_0 = \{x : x\in\mathbb{K}^s, |x-x_0|<\rho_0\}$ and $\tilde{\alpha}_0 = \sup\{|f(x)| : x\in E_0\}$. We choose $\rho_0 > 0$ so small that $(1-h\theta\beta_0)^{-1}\ \omega h\tilde{\alpha}_0 < r_0-\rho_0$. Lemma 5.5.10 can then be applied for arbitrary $\tilde{x}_0 \in E_0$ with α_0 replaced by $\tilde{\alpha}_0$. Let $\tilde{x},x \in E_0$. Similar to corollary 5.5.11 we then have

$$|hf'(y_i(x)+t(y_i(\tilde{x})-y_i(x)))-hf'(x+t(\tilde{x}-x))| \le$$
$$\le (1-h\theta\beta_0)^{-1}\ \omega h^2\tilde{\alpha}_0\gamma_0$$

$(1\le i\le m\ ,\ t\in[0,1]\ )$. We define $\tilde{Z} = \mathrm{diag}(\tilde{z}_1,\tilde{z}_2,\ldots,\tilde{z}_m)$ with

$$\tilde{z}_i = \int_0^1 hf'(y_i(x)+t(y_i(\tilde{x})-y_i(x)))dt \qquad (1\le i\le m) .$$

We then know that $|\tilde{z}_i-\tilde{z}_1| \le (1-h\theta\beta_0)^{-1}\,\omega\, h^2\tilde{\alpha}_0\gamma_0$ and $\mu[\tilde{z}_i] \le h\beta_0$ $(1\le i\le m)$. Further it follows from the mean-value theorem 2.3.1 that $hF(y(\tilde{x})) - hF(y(x)) = \tilde{Z}(y(\tilde{x})-y(x))$, and therefore (see (5.5.1.b))

$$y(\tilde{x}) - y(x) = (I-A\tilde{Z})^{-1}e(\tilde{x}-x) .$$

The continuity of $y(\cdot)$ on E_0 now follows from lemma 2.4.4.

In view of the above we may apply theorem 5.2.2. It follows that (5.5.3) holds. With the same notation as in theorem 5.5.1 we thus get

$$|G'(x_0;h,f)| \leq |\phi(z_0)| +$$

$$+ m \max_{1\leq i\leq m} |[b^T(I-AZ_0)^{-1}]_i| \max_{1\leq i\leq m} |z_i-z_0| \max_{1\leq i\leq m} |[(I-AZ)^{-1}e]_i| ,$$

for all $z_0 \in L(\mathbb{K}^s)$ such that $I - AZ_0$ is regular.

By choosing $z_0 = z_j$ for some $j \in \{1,2,\ldots,m\}$ the proof now easily follows from the lemmata 2.4.4, 2.4.5 and corollary 5.5.11. □

The estimate (5.5.11) for the error propagation per step is of the same type as the one obtained in section 5.4.4 for the Rosenbrock methods (cf. (5.4.25)). Therefore contractivity results as in theorem 5.4.16 hold for the Runge-Kutta methods satisfying (M) .

EXAMPLE 5.5.13. Consider the 2-stage Runge-Kutta method with

$$A = \begin{pmatrix} 0 & 0 \\ 1-\theta & \theta \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1-\theta \\ \theta \end{pmatrix}$$

where $\theta \in (0,1]$. If $\theta = \frac{1}{2}$ we deal here with the trapezoidal rule, and if $\theta = 1$ with the backward Euler method.

By choosing $z_0 = z_2$ we obtain from (5.5.3)

$$G'(x_0;h,f) = \phi(z_2) + (1-\theta)\ (1-\theta z_2)^{-1}\ (z_1-z_2)$$

where $z_1 = hf'(x_0)$ and $z_2 = hf'(G(x_0;h,f))$. Here we have used the fact that $y_2(x_0) = G(x_0;h,f)$, which holds because $b^T = e_2^T A$.

In lemma 5.5.10 and corollary 5.5.11 we may take $\omega = 1$. It follows that

$$|G'(x_0;h,f)| \leq \Phi(h\beta_0) + (1-\theta)\ (1-h\theta\beta_0)^{-2}\ h^2\alpha_0\gamma_0 ,$$

whenever (A_3) holds and $h \in (0,h_0)$ (i.e. $1 - 2h\theta\beta_0 > 0$). In this case the stepsize restriction can be weakened to $1 - h\theta\beta_0 > 0$. The stability function for this method is given by $\phi(\zeta) = (1-\theta\zeta)^{-1}(1+(1-\theta)\zeta)$ $(\zeta\in\mathbb{C})$, and

therefore (see also the examples 5.4.10, 5.5.5)

$$\Phi(t) = \max\{\theta^{-1}(1-\theta),(1-\theta t)^{-1}(1+(1-\theta)t)\} \quad (\text{for } t < \theta^{-1}).$$

In case $\theta = \frac{1}{2}$ (the trapezoidal rule) we have $\Phi(t) \geq 1$ (for all t), and therefore no contractivity result is obtained. In the same way as in remark 5.4.19 we do get a stability result on finite intervals.

DAHLQUIST (1963) was the first who obtained nonlinear stability results for the trapezoidal rule. For his results however it was necessarily assumed that the stepsize was constant on the whole integration interval. With the above bound for the error propagation per step stability can also be proved on arbitrary nonuniform grids. On the other hand we need, apart from dissipativity, additional assumptions on f (see (A_3)). The necessity of such extra assumptions follows from an example of STETTER (1973, pp. 181, 182) (where a nonuniform grid was chosen).

CHAPTER 6

B-CONVERGENCE FOR SEVERAL θ-METHODS

6.1. INTRODUCTION

In this chapter the results on the error propagation of chapter 5 will be used to derive convergence results for some simple semi-implicit and implicit methods, where the initial value problem may be arbitrarily stiff.

We consider the solution U on the interval $[0,T]$ of the initial value problem

$$(6.1.1) \qquad U'(t) = f(U(t)) \quad (0\le t\le T) \ , \ U(0) = u_0 \ ,$$

with $T > 0$, $f: \mathbb{K}^s \to \mathbb{K}^s$ $(s\in\mathbb{N})$ and $u_0 \in \mathbb{K}^s$.

Application of a numerical method G yields approximations u_n to the solution U at the gridpoints $t_n = nh$ $(0\le n\le T/h)$. For convenience we consider here again uniform grids. Of course one wants to know whether the numerical approximations converge to U and how fast this convergence is. We are therefore looking for a natural number p (the order of convergence) such that $|U(t_n)-u_n| = O(h^p)$ for $h \downarrow 0$ and t_n fixed.

For the one-step methods G considered in chapter 3 convergence results are known, but in these results the Lipschitz constant of f may be involved (see e.g. HENRICI (1962), VAN DER HOUWEN (1977)). If f satisfies a Lipschitz condition with constant L on some tube around U , then there are $p \in \mathbb{N}$, $c > 0$ and $h^* > 0$ such that $|U(t_n)-u_n| \le c\, h^p$ for all $h \in (0,h^*]$ and all n with $0 \le n \le T/h$. However if L is very large, then h^* will be very small or c very large. This is due to the fact that the class of problems with large L contains ill conditioned problems (see also section 1.1). These classical convergence results are therefore unsuited for stiff problems.

In all of the following it will be assumed that f is twice continuously differentiable on $\mathbb{K}^s$ and satisfies the one-sided Lipschitz condition

(6.1.2) $\quad \mathrm{Re}\,\langle f(\tilde{x})-f(x),\tilde{x}-x\rangle \le \beta|\tilde{x}-x|^2 \qquad$ (for all $\tilde{x},x \in \mathbb{K}^s$) .

As in the preceding chapters $\langle\cdot,\cdot\rangle$ stands for an inner product on $\mathbb{K}^s$ and $|x| = \langle x,x\rangle^{\frac{1}{2}}$ (for all $x \in \mathbb{K}^s$) .

We will derive in this chapter some convergence results where no upper bound for the Lipschitz constant of f is assumed, but where the one-sided Lipschitz constant β plays a central role. Such convergence results, called B-convergence results (see definition 6.1.1), are useful if the solution U of (6.1.1) is stiff. The concept of B-convergence was introduced by FRANK, SCHNEID and UEBERHUBER (1981). The capital B indicates that we are working within the framework of B-contractivity, i.e. (6.1.2) holds.

Let G denote some one-step method, and let $H > 0$ be an upper bound for the stepsizes h used to solve (6.1.1) numerically,

$$0 < h \le H .$$

We assume that $\mathcal{D}$ is an open, bounded region in $\mathbb{K}^s$ such that we have

$$u_n + \sigma(U(t_n)-u_n) \in \mathcal{D} \qquad (0\le\sigma\le 1)$$

for all numerical approximations u_n to $U(t_n)$ $(0\le t_n\le T)$ computed from G with stepsizes $h \in (0,H]$. Note that such a bounded $\mathcal{D}$ exists. Namely, since f is continuously differentiable, f satisfies a Lipschitz condition near U . From the already mentioned classical convergence results it follows that $\{u_n : 0\le n\le T/h\}$ is uniformly bounded for $h \in (0,H]$.

For $j = 0,1,2,\ldots$ we define

$$M_j = \max\{|U^{(j)}(t)| : 0\le t\le T\} ,$$

$$K_j = \sup\{|f^{(j)}(x)| : x\in\mathcal{D}\} .$$

If $\mathcal{D}$ encloses $\{U(t):t\in[0,T]\}$ tightly, we have $K_0 \approx M_1$. Note that K_1 is the Lipschitz constant of f on $\mathcal{D}$.

Further we use the following notation. For a given function $g: \mathbb{R} \to \mathbb{R}$, $p \in \mathbb{N}$ and parameters $c_1,c_2,\ldots,c_k$ we write

$$g(h) = O(h^p | c_1, c_2, \ldots, c_k) \qquad (h \downarrow 0) \ ,$$

if there exist positive numbers c^* and h^* which depend only on $c_1, c_2, \ldots, c_k$ such that

$$|g(h)| \leq c^* h^p \qquad \text{for all} \quad h \in (0, h^*] \ .$$

Following essentially FRANK, SCHNEID and UEBERHUBER (1981), we give the following definition.

DEFINITION 6.1.1. *Method* G *is said to be* B-*convergent of order* p *w.r.t.* U *if we have for all* $t \in [0,T]$

$$|U(t) - u_n| = O(h^p | \Theta, \Pi) \qquad (h \downarrow 0 \ , \ nh = t) \ ,$$

where Θ *stands for the collection of coefficients of* G , *and* Π *is a set of problem dependent parameters containing* T , *some bounds on the derivatives of* U (M_j *with* $j \geq 1$), *and some bounds related to* f *and its derivates (* β *and* K_j *with* $j \neq 1$), *which permit an arbitrary stiffness with* U .

If Π *only contains* T, β *and some* M_j *with* $j \geq 1$, *we call* G *optimally* B-*convergent of order* p *w.r.t.* U .

We thus see that with B-convergence K_1 is not allowed to be involved. This is in contrast to the classical convergence results. Also any influence of M_0 is excluded. This is a very natural requirement - see section 3.4 where translation invariance is discussed. If G is optimally B-convergent w.r.t. U , then the convergence only depends on the smoothness of U , and, through β , on the stability the differential equation.

In the subsequence three different one-step methods, each containing one parameter $\theta \in [\frac{1}{2}, 1]$, will be considered. The methods are given by

$$(M_1) \qquad G(x;h,f) = x + (1-\theta)\, hf(x) + \theta hf(G(x;h,f)) \ ,$$

$$(M_2) \qquad G(x;h,f) = x + (I - h\theta f'(x))^{-1}\, hf(x) \ ,$$

$$(M_3) \qquad G(x;h,f) = x + (I - h\theta J)^{-1}\, hf(x) \ .$$

In (M_3) the matrix J stands for a fixed approximation to $f'(x)$. It will be assumed that

$$(6.1.3) \qquad \mu[J] \le \beta\,, \quad |f'(x)-J| \le \delta \qquad (\text{for all } x \in \mathcal{D}).$$

Further we assume that the maximal stepsize H is such that

$$1 - H\theta\beta > 0\,.$$

This condition ensures that the algebraic equations arising in (M_1)-(M_3) have a unique solution (see theorem 4.3.1 with corollary 4.3.5, and corollary 4.4.1).

In section 6.3 it will be shown how the results on the error propagation of chapter 5 can be used to obtain B-convergence results. For these convergence results we also need bounds for the local discretization errors where the Lipschitz constant of f is not involved. Such bounds will be given in section 6.2.

For the Runge-Kutta method (M_1) with $\theta = \frac{1}{2}, 1$ (the trapezoidal rule and the backward Euler method) B-convergence results have been given already in FRANK, SCHNEID and UEBERHUBER (1981). Their result on the trapezoidal rule will be improved by using the bound for the error propagation given in section 5.5.3. Further B-convergence results have thus far only been given for B-contractive Runge-Kutta methods (cf. FRANK, SCHNEID and UEBERHUBER (1981,1982 B)).

REMARK 6.1.2. For their investigations on B-contractive Runge-Kutta methods FRANK, SCHNEID and UEBERHUBER (1982 A,B) introduced two stability concepts, BS- and BSI-stability, which are related to the sensitivity of the methods to perturbations on the internal stages (where the vectors $y_i(x)$ are computed). By a complicated proof it was shown that if there is a positive definite diagonal matrix D such that $DA + A^TD$ is positive definite, then the Runge-Kutta method (with coefficient matrix A) is BS- and BSI-stable. A much shorter and transparant proof of this result can be obtained by using lemma 2.4.3 and corollary 2.4.8, or by slightly modifying the proof of lemma 2.4.3.

6.2. B-CONSISTENCY RESULTS

The local discretization error of a method G w.r.t. the solution U of (6.1.1) at time t equals

$$\ell_h(t) = h^{-1}\,[U(t+h)-G(U(t);h,f)]$$

(see also section 3.1).

Let $0 \le t < t+h \le T$, $h \in (0,H]$ and $x = U(t)$. We write $V(h) = U(t+h)$ and $v_h = G(U(t);h,f)$. With this notation $\ell_h(t)$ can be rewritten as $h^{-1}[V(h)-v_h]$.

In this section we want to find upper bounds for $|\ell_h(t)|$. If this upper bound only depends on h,Θ and Π, with Θ and Π as in the definition of B-convergence (def. 6.1.1), we will call this a B-*consistency* result. If Π only contains T,β and some M_j with $j \ge 1$, we use the term *optimally* B-*consistent*.

Method (M_1). For the Runge-Kutta method given by (M_1) we have

$$v_h = x + (1-\theta)\,hf(x) + \theta hf(v_h)\,.$$

Let $w_h \in \mathbb{K}^s$ satisfy the relation

$$V(h) = x + (1-\theta)\,hf(x) + \theta hf(V(h)) + w_h\,.$$

Then we have

$$V(h) - v_h = h\theta[f(V(h))-f(v_h)] + w_h\,.$$

By taking on both sides the inner product with $V(h) - v_h$, and using (6.1.2) and Schwarz's inequality, it follows that

$$(6.2.1) \qquad |V(h)-v_h| \le (1-h\theta\beta)^{-1}\,|w_h|\,.$$

Note that since $1 - H\theta\beta > 0$ and $h \in (0,H]$, we have $1 - h\theta\beta > 0$.

Taylor expansion of $V(h)$ and $V'(h)$ $(=f(V(h)))$ around $h = 0$ yields

$$V(h) - x - (1-\theta)\,hf(x) - \theta hf(V(h)) = (\tfrac{1}{2}-\theta)\,h^2V''(0) + R$$

where

$$R = \int_0^1 [\tfrac{1}{2}(1-\tau)^2 - \theta(1-\tau)]\ V'''(\tau h)\ d\tau\ .$$

By some calculations we thus obtain

$$(6.2.2) \qquad |w_h| \le \tfrac{1}{2}(2\theta-1)\ h^2 M_2 + \tfrac{1}{6}(3\theta-1)\ h^3 M_3\ .$$

A combination of (6.2.1) and (6.2.2) leads to the B-consistency result

$$(6.2.3) \qquad |\ell_h(t)| \le (1-h\theta\beta)^{-1}\ [\tfrac{1}{2}(2\theta-1)hM_2 + \tfrac{1}{6}(3\theta-1)h^2 M_3]\ .$$

We see that the θ-method (M_1) is optimally B-consistent with order 2 if $\theta = \frac{1}{2}$, and order 1 for $\theta \in (\frac{1}{2},1]$. This corresponds with the classical orders of consistency.

REMARK 6.2.1. By expanding $V(h)$ and $V'(h)$ only up to order 2 around $h = 0$ we obtain (instead of (6.2.2))

$$|w_h| \le \tfrac{1}{2}\ h^2 M_2\ .$$

For $\theta = 1$ this gives a slight improvement over (6.2.2).

Further we note that the bound for the local error given in FRANK, SCHNEID and UEBERHUBER (1981) for $\theta = \frac{1}{2}$ is somewhat larger than the bound we get from (6.2.3).

Method (M_2). For the semi-implicit method (M_2) we have

$$v_h = x + (I-h\theta f'(x))^{-1}\ hf(x)\ .$$

With $w_h \in \mathbb{K}^s$ such that

$$V(h) = x + (I-h\theta f'(x))^{-1}\ hf(x) + w_h\ ,$$

we now obtain

$$|V(h)-v_h| = w_h\ .$$

Taylor expansion of $V(h)$ around $h = 0$ leads directly to

$$w_h = hf(x) + \tfrac{1}{2}h^2 f'(x)f(x) - (I-h\theta f'(x))^{-1}\, hf(x) + h^3 R$$

with $|R| \le \frac{1}{6}M_3$. Hence

$$(6.2.4) \qquad w_h = (I-h\theta f'(x))^{-1}\,[\tfrac{1}{2}(1-2\theta)h^2 f'(x)f(x) - \tfrac{1}{2}\theta h^3 (f'(x))^2 f(x)] + h^3 R .$$

Starting from (6.2.4) we first derive an optimal B-consistency result. For this we rewrite (6.2.4) as

$$w_h = (1-h\theta f'(x))^{-1}\,(1-2\theta-\theta h f'(x))\,\tfrac{1}{2}h^2 f'(x)f(x) + h^3 R .$$

Since $\sup\{|(1-\theta\zeta)^{-1}(1-2\theta-\theta\zeta)| : \zeta\in\mathbb{C}, \mathrm{Re}\ \zeta\le t\} = \max\{1, |(1-\theta t)^{-1}\,(1-2\theta-\theta t)|\}$ it follows from theorem 2.2.7 that

$$(6.2.5) \qquad |\ell_h(t)| \le \max\{1, |(1-\theta h\beta)^{-1}(1-2\theta-\theta h\beta)|\}\,\frac{1}{2}hM_2 + \frac{1}{6}h^2 M_3 .$$

On the other hand, since $(f'(x))^2 f(x) = V'''(0) - f''(x)(f(x))^2$, we also obtain from (6.2.4)

$$(6.2.6) \qquad |\ell_h(t)| \le (1-h\theta\beta)^{-1}\,[\tfrac{1}{2}(2\theta-1)hM_2 + \tfrac{1}{2}\theta h^2 (M_3 + K_2 M_1^2)] + \frac{1}{6}h^2 M_3 .$$

This leads to a B-consistency result with order 2 if $\theta = \frac{1}{2}$, but with K_2 involved. From (6.2.5), which is an optimal B-consistency result, we only get order 1 for all $\theta \in [\frac{1}{2}, 1]$. Such a drop in the order has been observed for a large class of implicit Runge-Kutta methods by FRANK, SCHNEID and UEBERHUBER (1982 B).

It will turn out in the following section that the optimal B-consistency result (6.2.5) does not lead to an optimal B-convergence result, due to the fact that the method is not B-contractive.

Method (M_3). For this method we have

$$v_h = x + (I-h\theta J)^{-1}\, hf(x) ,$$

where $J \approx f'(x)$. We assume (6.1.3). Let $w_h \in \mathbb{K}^s$ be such that

$$V(h) = x + (I-h\theta J)^{-1} \quad hf(x) + w_h \ .$$

Obviously we then have

$$|V(h)-v_h| = |w_h| \ .$$

From the results obtained for the method (M_2) we easily get a result for this case by writing

$$w_h = V(h) - x - (I-h\theta f'(x))^{-1} \ hf(x) - \theta(I-h\theta J)^{-1} \ (hJ-hf'(x))$$
$$(I-h\theta f'(x))^{-1} \ hf(x) \ .$$

Using (6.2.5) we thus obtain

$$(6.2.7) \qquad |\ell_h(t)| \le \max\{1,|(1-h\theta\beta)^{-1}(1-2\theta-h\theta\beta)|\} \ \tfrac{1}{2}hM_2 +$$
$$+ \tfrac{1}{6}h^2M_3 + (1-h\theta\beta)^{-2} \ h^2\theta\delta M_1 \ .$$

This leads to a B-consistency result with order 1. Also for $\theta = \frac{1}{2}$ order 2 cannot be reached. If $\theta = \frac{1}{2}$ we can take $J = (1+\varepsilon) \ f'(x)$ with $\varepsilon > 0$ sufficiently small. We then have $w_h = V(h) - x - (I-h\tilde{\theta}f'(x))^{-1}hf(x)$ with $\tilde{\theta} = \theta(1+\varepsilon) \neq \frac{1}{2}$. Since the θ-method (M_2) only has (classical) order 1 if $\theta \neq \frac{1}{2}$, we get for (M_3) also order 1 for all $\theta \in [\frac{1}{2},1]$.

6.3. B-CONVERGENCE RESULTS

The bounds for the local discretization errors found in section 6.2, and the results of chapter 5 on the error propagation per step will now be used to derive B-convergence results for the θ-methods (M_1)-(M_3) .

This will be done along the following line. We have

$$|U(t_n)-u_n| \le |U(t_n)-G(U(t_{n-1});h,f)| +$$
$$+ \ |G(U(t_{n-1});h,f)-G(u_{n-1};h,f)| \ .$$

From section 6.2 we get an inequality

$$|U(t_n)-G(U(t_{n-1});h,f)| \le h \ \lambda_h \ ,$$

with λ_h such that $|\ell_h(t_{n-1})| \le \lambda_h$, and from chapter 5 we get an estimate of the form

$$|G(U(t_{n-1});h,f)-G(u_{n-1};h,f)| \le \Psi_h |U(t_{n-1})-u_{n-1}| .$$

It follows that

$$|U(t_n)-u_n| \le \Psi_h^n |U(t_0)-u_0| + (\Psi_h^{n-1}+\ldots+\Psi_h+1)\, h\lambda_h =$$

$$= (\Psi_h-1)^{-1}\,(\Psi_h^n-1)\, h\lambda_h .$$

In case $\Psi_h = 1$ we read n $(= \lim_{\zeta \to 1} (\zeta-1)^{-1}(\zeta^n-1))$ instead of $(\Psi_h-1)^{-1}(\Psi_h^n-1)$. The above inequality will yield the desired convergence results.

The methods (M_1)-(M_3) all have the same stability function $\phi(\zeta) = (1-\theta\zeta)^{-1}(1+(1-\theta)\zeta)$ $(\zeta \in \mathbb{C})$. The main contribution to Ψ_h is given by $\Phi(h\beta) = \sup\{|\phi(\zeta)| : \zeta \in \mathbb{C}, \mathrm{Re}\ \zeta \le h\beta\}$. We have (cf. example 5.4.10)

$$\Phi(h\beta) = \max\{\theta^{-1}(1-\theta), (1-h\theta\beta)^{-1}(1+h(1-\theta)\beta)\} . \tag{6.3.1}$$

For $\theta \in (\frac{1}{2},1]$, $\beta \neq 0$, and for $\theta = \frac{1}{2}$, $\beta > 0$ we see that

$$\Phi(h\beta) = (1-h\theta\beta)^{-1}(1+h(1-\theta)\beta)$$

for $|h\beta|$ sufficiently small (how small depends on θ). It easily follows that we have

$$\Phi(h\beta)^n = e^{\beta t} + O(h|\theta,\beta,T) \qquad (h\downarrow 0,\ nh=t) ,$$

$$(\Phi(h\beta)-1)^{-1}\,(\Phi(h\beta)^n-1)h = \beta^{-1}(e^{\beta t}-1) + O(h|\theta,\beta,T) \quad (h\downarrow 0,\ nh=t)$$

(uniformly for all $t \in [0,T]$).

If $\beta = 0$ or $\theta = \frac{1}{2}$, $\beta < 0$, we have $\Phi(h\beta) = 1$, and then $(\Phi(h\beta)-1)^{-1}(\Phi(h\beta)^n-1)h$ stands for nh.

Method (M_1). For this method we have (cf. example 5.5.13)

$$\Psi_h = \Phi(h\beta) + (1-h\theta\beta)^{-2}\,(1-\theta)\, h^2 K_0 K_2 .$$

with Φ as in (6.3.1). Method (M_1) is only B-contractive if $\theta = 1$. We

assume for the moment that $\theta \neq 1$. Then

$$\Psi_h = \Phi(h\beta) + O(h^2|\theta,\beta,K_0,K_2) \qquad (h\downarrow 0) \ .$$

It follows that for $\theta \in (\frac{1}{2},1)$, $\beta \neq 0$, and for $\theta = \frac{1}{2}$, $\beta > 0$,

$$(\Psi_h - 1)^{-1}(\Psi_h^n - 1)h = \beta^{-1}(e^{\beta t}-1) + O(h|\theta,T,\beta,K_0,K_2) \quad (h\downarrow 0 \ , \ nh=t\in[0,T]).$$

If $\beta = 0$ or $\theta = \frac{1}{2}$, $\beta < 0$, we get by some calculations

$$(\Psi_h - 1)^{-1}(\Psi_h^n - 1)h = t + O(h|\theta,T,\beta,K_0,K_2) \qquad (h\downarrow 0 \ , \ nh=t\in[0,T]) \ .$$

Together with the B-consistency result (6.2.3) this leads to the following.

THEOREM 6.3.1. *Consider the implicit Runge-Kutta method* (M_1) . *Let* Π *stand for* $\{T,\beta,M_2,M_3,K_0,K_2\}$.

Suppose $\theta \in (\frac{1}{2},1)$, $\beta \neq 0$, *or* $\theta = \frac{1}{2}$, $\beta > 0$. *Then we have*

$$(6.3.2) \quad |U(t)-u_n| \leq \tfrac{1}{2}(2\theta-1)\ \beta^{-1}(e^{\beta t}-1)hM_2 + O(h^2|\theta,\Pi) \quad (h\downarrow 0 \ , \ nh=t\in[0,T]).$$

In case $\beta = 0$, *or* $\theta = \frac{1}{2}$, $\beta < 0$, *we have*

$$(6.3.3) \quad |U(t)-u_n| \leq \tfrac{1}{2}(2\theta-1)\ thM_2 + O(h^2|\theta,\Pi) \qquad (h\downarrow 0 \ , \ nh=t\in[0,T]).$$

In spite of the optimal B-consistency result (6.2.3) we get no optimal B-convergence result for $\theta \in (\frac{1}{2},1)$ since (M_1) is then not B-contractive. For $\theta = 1$ we obtain, in view of remark 6.2.1, again (6.3.2) (if $\beta \neq 0$) and (6.3.3) (if $\beta = 0$), but now with $\Pi = \{T,\beta,M_2\}$. This optimal B-convergence result can also be found in FRANK, SCHNEID and UEBERHUBER (1981).

Method (M_2). For the Rosenbrock method (M_2) we have (see example 5.4.20)

$$\Psi_h = \Phi(h\beta) + (1-h\theta\beta)^{-2}\ \theta h^2 K_0 K_2 \ ,$$

with Φ as in (6.3.1).

We obtain a B-convergence result in the same way as with method (M_1). Since Ψ_h depends on K_0 and K_2 the optimal B-consistency result (6.2.5) does not lead to an optimal B-convergence result. We therefore use (6.2.6) which gives the order 2 if $\theta = \frac{1}{2}$.

THEOREM 6.3.2. *Consider the Rosenbrock method* (M_2) , *and let* Π *stand for* $\{T,\beta,M_1,M_2,M_3,K_0,K_2\}$.

Assume $\theta \in (\frac{1}{2},1]$, $\beta \neq 0$, *or* $\theta = \frac{1}{2}$, $\beta > 0$. *Then we have again the inequality* (6.3.2).

If $\beta = 0$, *or* $\theta = \frac{1}{2}$, $\beta < 0$, *then* (6.3.3) *holds.*

We thus see that there is hardly any difference between the results for the Rosenbrock method (M_2) and the implicit Runge-Kutta method (M_1), unless $\theta = 1$. This suggests, for $\theta \neq 1$, that the behaviour of both methods is comparable on the whole class of smooth problems (6.1.1) with $|f(x)| \leq K_0$ and $|f''(x)| \leq K_2$ for all x in a region $\mathcal{D}$ containing $\{U(t):t\in[0,T]\}$. Some care has to be taken with this conclusion since we only have upper bounds for the global error. Moreover, for each individual problem this upper bound (for a class of problems) may be pessimistic for one of the methods.

Method (M_3). For the semi-implicit method (M_3) we know from example 5.4.26 that

$$\Psi_h = \Phi(h\beta) + (1-h\theta\beta)^{-1} h\delta$$

with Φ as in (6.3.1). In a similar way as before we now get the following limits.

Suppose $\theta \in (\frac{1}{2},1]$, $\beta + \delta \neq 0$, or $\theta = \frac{1}{2}$, $\beta > 0$. Then

$$(\Psi_h-1)^{-1} (\Psi_h^n-1)h = (\beta+\delta)^{-1} (e^{(\beta+\delta)t}-1) +$$

$$+ \mathcal{O}(h|\theta,T,\beta,\delta) \qquad (h\downarrow 0 , nh=t\in[0,T]) .$$

If $\theta \in (\frac{1}{2},1]$ and $\beta + \delta = 0$, then

$$(\Psi_h-1)^{-1} (\Psi_h^n-1)h = t + \mathcal{O}(h|\theta,T,\beta,\delta) \qquad (h\downarrow 0 , nh=t\in[0,T]) .$$

In case $\theta = \frac{1}{2}$ and $\beta \leq 0$, we have

$$(\Psi_h-1)^{-1} (\Psi_h^n-1)h = \delta^{-1}(e^{\delta t}-1) +$$

$$+ \mathcal{O}(h|\theta,T,\beta,\delta) \qquad (h\downarrow 0 , nh=t\in[0,T]) .$$

By using the estimate (6.2.7) for the local discretization error we obtain the following result.

THEOREM 6.3.3. *Consider the semi-implicit method* (M_3) *with a fixed Jacobian* J *satisfying* (6.1.3). *Let* Π *stand for the set* $\{T,\beta,M_1,M_2,M_3,\delta\}$.

Assume $\theta \in (\frac{1}{2},1]$, $\beta + \delta \neq 0$, *or* $\theta = \frac{1}{2}$, $\beta > 0$. *Then we have*

$$(6.3.4) \qquad |U(t)-u_n| \leq \tfrac{1}{2}(\beta+\delta)^{-1}\,(e^{(\beta+\delta)t}-1)\;hM_2 + O(h^2|\theta,\Pi) \qquad (h\downarrow 0\,,\ nh=t\in[0,T])\ .$$

If $\theta \in (\frac{1}{2},1]$ and $\beta + \delta = 0$, then

$$(6.3.5) \qquad |U(t)-u_n| \leq \tfrac{1}{2}thM_2 + O(h^2|\theta,\Pi) \qquad (h\downarrow 0\,,\ nh=t\in[0,T])\ .$$

Finally if $\theta = \frac{1}{2}$ and $\beta \leq 0$, we get

$$(6.3.6) \qquad |U(t)-u_n| \leq \tfrac{1}{2}\delta^{-1}(e^{\delta t}-1)\;hM_2 + O(h^2|\theta,\Pi) \qquad (h\downarrow 0\,,\ nh=t\in[0,T])\ .$$

Comparision of the theorems 6.3.2 and 6.3.3 shows that the use of a fixed Jacobian has to be paid for by an order reduction in case $\theta = \frac{1}{2}$. This corresponds with the classical order result. Moreover, and this is probably more serious, the global error may grow exponentially if $\beta \lessapprox 0$ but $\delta > 0$ is not small, whereas the solution U remains bounded. Therefore the relative error $|U(t)-u_n|/|U(t)|$ will only be reasonable on a short interval.

REFERENCES

ALEXANDER, R. (1977), *Diagonally implicit Runge-Kutta methods for stiff ODE's*, SIAM J. Numer. Anal. 14, 1006-1021.

BJUREL, G., DAHLQUIST, G., LINDBERG, B., LINDE, S., ODEN, L. (1970), *Survey of stiff ordinary differential equations*, Report NA 70. 11, Dept. of Inf. Proc., Roy. Inst. of Techn., Stockholm.

BRENNER, P., THOMÉE, V. (1979), *On rational approximations of semi-groups*, SIAM J. Numer. Anal. 16, 683-694.

BUI, T.D. (1979 A), *A note on Rosenbrock procedures*, Math. Comp. 33, 971-975.

BUI, T.D. (1979 B), *Some A-stable and L-stable methods for the numerical solution of stiff ordinary differential equations*, J. ACM 26, 483-493.

BURRAGE, K., BUTCHER, J.C. (1979), *Stability criteria for implicit Runge-Kutta methods*, SIAM J. Numer. Anal. 16, 46-57.

BURRAGE, K., BUTCHER, J.C. (1980), *Nonlinear stability of a general class of differential equation methods*, BIT 20, 185-203.

BUTCHER, J.C. (1964), *Implicit Runge-Kutta processes*, Math. Comp. 18, 50-64.

BUTCHER, J.C. (1965), *On the attainable order of Runge-Kutta methods*, Math. Comp. 19, 408-417.

BUTCHER, J.C. (1975), *A stability property of implicit Runge-Kutta methods*, BIT 15, 358-361.

CALAHAN, D.A. (1968), *A stable, accurate method of numerical integration for nonlinear systems*, Proc. IEEE 56, 744-747.

CHIPMAN, F.H. (1971), *A-stable Runge-Kutta processes*, BIT 11, 384-388.

CROUZEIX, M. (1979), *Sur la B-stabilité des méthodes de Runge-Kutta*, Numer. Math. 32, 75-82.

CROUZEIX, M., HUNDSDORFER, W.H., SPIJKER, M.N. (1983), *On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods*, BIT 23, 84-91.

CROUZEIX, M., RAVIART, P.A. (1978), *Approximation d'équations d'évolution linéaires par des méthodes multipas*, In: Etude numérique des grands systèmes, Dunod, Paris.

CROUZEIX, M., RAVIART. P.A. (1980), Unpublished lecture notes, Université de Rennes.

CROUZEIX, M., RUAMPS, F. (1977), *On rational approximations to the exponential*, RAIRO Numer. Anal. 11, 241-243.

DAHLQUIST, G. (1959), *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. Roy. Inst. Techn., no. 130, Stockholm.

DAHLQUIST, G. (1963), *A special stability problem for linear multistep methods*, BIT 3, 27-43.

DAHLQUIST, G. (1975), *Error analysis for a class of methods for stiff nonlinear initial value problems*, Lecture Notes in Mathematics 506, Springer Verlag.

DAHLQUIST, G. (1978), *G-stability is equivalent to A-stability*, BIT 18, 384-401.

DAHLQUIST, G., JELTSCH, R. (1979), *Generalized disks of contractivity for explicit and implicit Runge-Kutta methods*, Report TRITA-NA-7906, Roy. Inst. Techn., Stockholm.

DAY, J.D., MURTHY, D.N.P. (1982), *An L(α)-stable fourth order Rosenbrock method with error estimator*, J. Comp. Appl. Math. 8, 21-27.

DEKKER, K. (1982), *On the iteration error in algebraically stable Runge-Kutta methods*, Report NW 138/82, Math. Centre, Amsterdam.

DESOUR, C.A., HANEDA, H. (1972), *The measure of a matrix as a tool to analyse computer algorithms for circuit analysis*, IEEE Trans. Circuit Theory 19, 480-486.

DUNFORD, N., SCHWARTZ, J.T. (1958), *Linear operators, part I*, Interscience Publ. Inc., New York.

EHLE, B.L., LAWSON, J.D. (1975), *Generalized Runge-Kutta processes for stiff initial value problems*, J. Inst. Math. Applics. 16, 11-21.

ENRIGHT, W.H., HULL, T.E., LINDBERG, B. (1975), *Comparing numerical methods for stiff systems of* ODE's, BIT 15, 10-48.

FRANK, R., SCHNEID, J., UEBERHUBER, C.W. (1981), *The concept of B-convergence*, SIAM J. Numer. Anal. 18, 753-780.

FRANK, R., SCHNEID, J., UEBERHUBER, C.W. (1982 A), *Stability properties of implicit Runge-Kutta methods*, Report no. 52/82, Inst. Angew. und Numer. Math., Techn. Univ. of Vienna.

FRANK, R., SCHNEID, J., UEBERHUBER, C.W. (1982 B), *Order results for implicit Runge-Kutta methods applied to stiff systems*, Report no. 53/82, Inst. Angew. und Numer. Math., Techn. Univ. of Vienna.

FRIEDLI, A. (1978), *Veralgemeinerte Runge-Kutta Verfahren zur Lösung steifer Differentialgleichungssysteme*, Lecture Notes in Mathematics 631, Springer Verlag.

GANTMACHER, F.R. (1959), *The theory of matrices*, Chelsea Publ. Co., New York.

GOTTWALD, B.A., WANNER, G. (1981), *A reliable Rosenbrock integrator for stiff differential equations*, Computing 26, 355-360.

GRIGORIEFF, R.D. (1972), *Numerik gewönlicher Differentialgleichungen*, Teubner, Stuttgart.

HAIRER, E. (1980), *Highest possible order of algebraically stable diagonally implicit Runge-Kutta methods*, BIT 20, 254-256.

HAIRER, E. (1982), *Constructive characterization of* A*-stable approximations to* exp(z) *and its connection with algebraically stable Runge-Kutta methods*, Numer. Math. 39, 247-258.

HAIRER, E., BADER, G., LUBICH, C. (1982), *On the stability of semi-implicit methods for ordinary differential equations*, BIT 22, 211-232.

HAIRER, E., TUERKE, H. (1983), *How* B*-stability is equivalent to* A*-stability*, Report 234, Inst. of Appl. Math. Univ. of Heidelberg.

HAIRER, E., WANNER, G. (1981), *Algebraically stable and implementable Runge-Kutta methods of high order*, SIAM J. Numer. Anal. 18, 1098-1108.

HENRICI, P. (1962), *Discrete variable methods in ordinary differential equations*, J. Wiley & Sons, New York.

HEYER, C. den (1979), *The numerical solution of nonlinear operator equations by imbedding methods*, Math. Centre Tracts 107, Amsterdam.

HOUWEN, P.J. van der (1977), *Construction of integration formulas for initial value problems*, North-Holland Publ. Co., Amsterdam.

HUNDSDORFER, W.H. (1981), *Nonlinear stability analysis for a simple Rosenbrock method*, Report 81-31, Univ. of Leiden.

HUNDSDORFER, W.H., SPIJKER, M.N. (1981), *A note on* B*-stability of Runge-Kutta methods*, Numer. Math. 36, 319-332.

HUNDSDORFER, W.H., SPIJKER, M.N. (1981), *On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods*, Report 81-49, Univ. of Leiden.

KAPS, P., RENTROP, P. (1979), *Generalized Runge-Kutta methods of order* 4 *with stepsize control for stiff ordinary differential equations*, Numer. Math. 33, 55-68.

KAPS, P., WANNER, G. (1981), *A study of Rosenbrock type methods of high order*, Numer. Math. 38, 279-298.

LAMBERT, J.D. (1973), *Computational methods in ordinary differential equations*, J. Wiley & Sons, London.

LAMBERT, J.D. (1974), *Two unconventional classes of methods for stiff systems*, In: Stiff differential systems, ed. R.A. Willoughby, Plenum press, New York.

LAMBERT, J.D., SIGURDSON, S.T. (1972), *Multistep methods with variable matrix coefficients*, SIAM J. Numer. Anal. 9, 715-733.

LAWSON, J. (1967), *Generalized Runge-Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal. 4, 372-380.

LINDBERG, B. (1974), *On a dangerous property of methods for stiff differential equations*, BIT 14, 430-436.

LININGER, W., WILLOUGHBY, R.A. (1970), *Efficient integration methods for stiff systems of ordinary differential equations*, SIAM J. Numer. Anal. 7, 47-66.

MARCUS, M., MINC, H. (1964), *A survey of matrix theory and matrix inequalities*, Allyn and Bacon Inc., Boston.

MARTIN, R.H. (1976), *Nonlinear operators and differential equations in Banach spaces*, J. Wiley & Sons, New York.

NEUMANN, J. von (1951), *Eine Spectraltheorie für allgemeine Operatoren eines unitären Raumes*, Math. Nachr. 4, 258-281.

NEVANLINNA, O., LININGER, W. (1978,1979), *Contractive methods for stiff differential equations*, BIT 18, 457-474 and BIT 19, 53-72.

NØRSETT, S.P., WOLFBRANDT, A. (1979), *Order conditions for Rosenbrock type methods*, Numer. Math. 32, 1-15.

ORTEGA, J.M., RHEINBOLDT, W.C. (1970), *Iterative solution of nonlinear equations in severable variables*, Academic Press, New York.

PROTHERO, A., ROBINSON, A. (1974), *On the stability and accuracy of one-step methods for solving stiff systems of* ODE's, Math. Comp. 28, 145-162.

ROSENBROCK, H.H. (1963), *Some general implicit processes for the numerical solution of differential equations*, Comp. J. 5, 329-330.

SANDBERG, I.W., SHICHMAN, H. (1968), *Numerical integration of systems of stiff nonlinear differential equations*, B.S.T.J. 47, 511-527.

SAWORIN, A.N. (1982), *Ueber die Effectivität einiger nichtlinearer Verfahren bei der Numerischen Behandlung steifer Differential-gleichungenssysteme*, Numer. Math. 40, 169-177.

SCHOLZ, S. (1978), S-*stabile modifizierte Rosenbrock-Verfahren 3. und 4. ordnung*, Sekt. Math., Techn. Univ. Dresden.

SHAMPINE, L.F. (1980), *Implementation of implicit formulas for the solution of* ODE's, SIAM J. Sci. Stat. Comput. 1, 103-118.

SHAMPINE, L.F. (1982), *Implementation of Rosenbrock methods*, ACM TOMS 8, 93-113.

SPIJKER, M.N. (1982 A), *Contractivity in the numerical solution of initial value problems*, Report 82-10, Univ. of Leiden.

SPIJKER, M.N. (1982 B), *Stability in the numerical solution of stiff initial value problems*, Nieuw Archief voor Wiskunde (3)XXX, 264-276.

SÖDERLIND, G. (1981), *On the efficient solution of nonlinear equations in numerical methods for stiff differential systems*, Report TRITA-NA-8114, Roy. Inst. Techn., Stockholm.

STEIHAUG, T., WOLFBRANDT, A. (1979), *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations*, Math. Comp. 33, 521-534.

STETTER, H.J. (1973), *Analysis of discretization methods for ordinary differential equations*, Springer Verlag, Berlin.

STETTER, H.J. (1975), *Towards a theory for discretizations of stiff differential systems*, Lecture Notes in Mathematics 506, Springer Verlag.

STREHMEL, K. (1981), *Stabilitätseigenschaften adaptiver Runge-Kutta-Verfahren*, ZAMM 61, 253-260.

STREHMEL, K., WEINER, R. (1982 A), *Nichtlineare Stabilität adaptiver Runge-Kutta Methoden*, Report No. 63, Martin-Luther-Univ., Halle a.d. Saale.

STREHMEL, K., WEINER, R. (1982 B), *Behandlung steifer Anfangswertprobleme gewönlicher Differentialgleichungen mit adaptiven Runge-Kutta-Methoden*, Computing 29, 153-165.

STRÖM, T. (1975), *On logarithmic norms*, SIAM J. Numer. Anal. 12, 741-753.

TRIGIANTE, D. (1977), *Asymptotic stability and discretization on an infinite interval*, Computing 18, 117-129.

VANSELOV, R. (1979), *Stabilität und Fehleruntersuchungen bei numerischen Verfahren zur Lösung nichtlinearer Anfangswertprobleme*, Diplomarbeit Sekt. Math. Techn. Univ. Dresden.

VELDHUIZEN, M. van (1973), *Convergence of one-step discretization methods for stiff differential equations (thesis)*, Math. Inst., Univ. of Utrecht.

VELDHUIZEN, M. van (1974), *Consistency and stability for one-step discretizations of stiff differential equations*, In: Stiff differential systems, ed. R.A. Willoughby, Plenum Press, New York.

VELDHUIZEN, M. van (1981), D-*Stabilty*, SIAM J. Numer. Anal. 18, 45-64.

VELDHUIZEN, M. van (1983), D-*Stability and Kaps-Rentrop methods*, Report 235, Vrije Universiteit Amsterdam.

VERWER, J.G. (1977), S-*stability properties of generalized Runge-Kutta methods*, Numer. Math. 27, 359-370.

VERWER, J.G. (1980), *On generalized Runge-Kutta methods using an exact Jacobian at non-step points*, ZAMM 60, 263-265.

VERWER, J.G. (1981 A), *Instructive experiments with some Runge-Kutta-Rosenbrock methods*, Report NW 100/81, Math. Centr., Amsterdam.

VERWER, J.G. (1981 B), *On the practical value of the notion of* BN-*stability*, BIT 21, 355-361.

VERWER, J.G. (1982), *An analysis of Rosenbrock methods for nonlinear stiff initial value problems*, SIAM J. Numer. Anal. 19, 155-170.

VERWER, J.G., SCHOLZ, S. (1982), *Rosenbrock methods and time-lagged Jacobian matrices*, Beiträge zur Numer. Math. 11.

VERWER, J.G., SCHOLZ, S., BLOM, J.G., LOUTER-NOOL, M. (1982), *A class of Runge Kutta-Rosenbrock methods for solving stiff differential equations*, Report NW 125/82, Math. Centr., Amsterdam.

WAMBECQ, A. (1978), *Rational Runge-Kutta methods for solving systems of ordinary differential equations*, Computing 20, 333-342.

WANNER, G. (1976), *A short proof of nonlinear* A-*stability*, BIT 16, 226-227.

WANNER, G. (1980), *Characterization of all* A-*stable methods of order* 2m - 4 , BIT 20, 367-374.

WANNER, G., HAIRER, E., NØRSETT, S.P. (1978), *Order stars and stability theorems*, BIT 18, 475-489.

WILLIAMS, J. (1979), *The problem of implicit formulas in numerical methods for stiff differential equations*, Numer. Anal. Report No. 40, Dept. of Math., Univ. of Manchester.

WILLOUGHBY, R.A. (ed.) (1974), *Stiff differential systems*, Plenum Press, New York.

SYMBOL INDEX

$\alpha(\cdot)$ function, 71
α_0 constant, 81,110
α_0^* constant, 101
β one-sided Lipschitz constant, 3, constant, 71
β_0 constant, 81,110
$\gamma(\cdot)$ function, 71
γ_0 constant, 81,110
γ_0^* constant, 98
Δ function, 71
$\delta(\cdot)$ function, 71
δ_0 constant, 81
$\varepsilon(\cdot)$ function, 71
ε_0 constant, 81
θ positive constant, 12,71,110
$\mu[\cdot]$ logarithmic norm, 10
ρ_i real numbers, 72
σ constant, 112
$\sigma(\cdot)$ spectrum, 11
σ_i real numbers, 72
τ_i real numbers, 72
ϕ stability function, 63,65
Φ half-plane bound for ϕ, 65, function, 49
ψ rational function, 12
Ψ half-plane bound for ψ, 12, function, 49, bound for error propagation, 107
Ψ_h bound for error propagation, 123
ω constant, 72

g' Gateaux-derivative of g, 11
$|\cdot|$ norm on $\mathbb{K}^s$, 10
$\|\cdot\|_D$ norm induced by D, 22
$\langle\cdot,\cdot\rangle$ inner product on $\mathbb{K}^s$, 10
$[\cdot,\cdot]_D$ inner product induced by D, 22

$(\cdot,\cdot)$	Euclidean inner product, 22
$[\cdot]_{ij}$	i,j-th. (block-) entry of a matrix, 11
$O(\cdot\|\cdots)$	order symbol, 117

SUBJECT INDEX

SAMENVATTING

In dit proefschrift wordt een analyse gegeven van een algemene klasse van eenstaps methoden voor het numeriek oplossen van stijve beginwaarde problemen voor stelsels gewone differentiaal vergelijkingen. Een stijf beginwaarde probleem wordt gekenmerkt door het feit dat de oplossing stabiel en glad is, terwijl het systeem eigenwaarden bezit die in modulus groot zijn. Deze grote eigenwaarden veroorzaken stabiliteits problemen bij vele bekende numerieke methoden, zoals de expliciete Runge-Kutta methoden.

De eenstaps methoden die in dit proefschrift beschouwd worden zijn generalisaties van expliciete Runge-Kutta methoden. Zulke generalisaties worden nader bekeken in hoofdstuk 3.

Numerieke methoden die geschikt zijn voor stijve beginwaarde problemen hebben noodzakelijkerwijs een impliciet karakter, d.w.z. tijdens het integratie proces dienen stelsels "algebraïsche" vergelijkingen opgelost te worden. Voor het goed gedefinieerd zijn van het numerieke proces is het nodig dat deze vergelijkingen eenduidig oplosbaar zijn. In hoofstuk 4 wordt nagegaan in hoeverre dit opgaat voor onze klasse van eenstaps methoden.

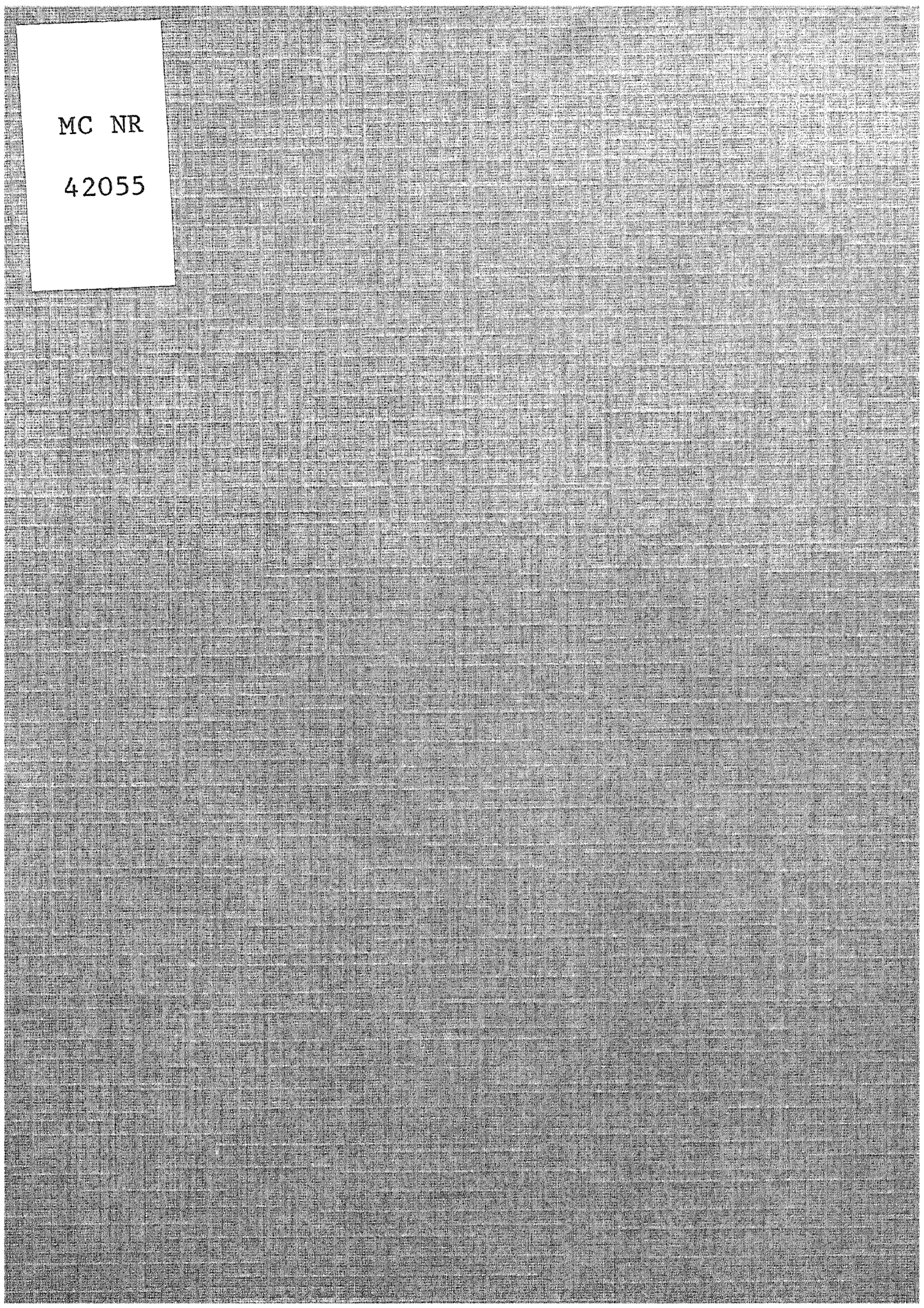
In hoofdstuk 5 wordt onderzocht hoe kleine storingen en fouten doorwerken wanneer de eenstaps methoden worden toegepast op stelsels niet-lineaire differentiaal vergelijkingen. Veel onderzoek op het gebied van de doorwerking van fouten is gedaan aan de hand van scalaire lineaire testvergelijkingen. Wij bekijken in welke mate de conclusies die getrokken kunnen worden uit zulke eenvoudige testvergelijkingen overdraagbaar zijn op gevallen waarbij we te maken hebben met stelsels niet-lineaire differentiaal vergelijkingen. Het onderzoek in dit hoofstuk spitst zich vooral toe op de zg. semi-impliciete methoden, waarvoor tot nu toe weinig bekend was op dit gebied.

Tenslotte wordt in hoofdstuk 6 een toepassing gegeven van de resultaten over de doorwerking van fouten uit hoofdstuk 5. Voor enkele eenvoudige methoden worden convergentie resultaten afgeleid voor willekeurig stijve, niet-lineaire, beginwaarde problemen.

CURRICULUM VITAE

De schrijver van dit proefschrift werd geboren op 31 augustus 1954 in Graz (Oostenrijk). Na in 1972 aan het Rotterdamsch Montessori Lyceum het eindexamen HBS-B te hebben afgelegd, begon hij met de studie Wis- en Natuurkunde aan de Rijksuniversiteit Leiden; de eerste twee jaar met hoofdvak natuurkunde, en na 1974 met hoofdvak wiskunde. In 1979 werd het doctoraal examen in de wiskunde met bijvak theoretische natuurkunde afgelegd.

Van 1979 tot 1983 was de auteur werkzaam als doctoraal assistent aan de Rijksuniversiteit Leiden. In deze periode werd het onderzoek voor dit proefschrift verricht.


MC NR
42055