TRYK: DAIMI/RECAU

PB-164

J. Sand: On Contractive Linear Multistep and One-Leg Methods

ISSN 0105-8517

# On Contractive Linear Multistep and One-Leg Methods

Jørgen Sand

DAIMI PB-164
August 1983

Computer Science Department
**AARHUS UNIVERSITY**
Ny Munkegade – DK 8000 Aarhus C – DENMARK
*Telephone: 06 – 12 83 55*

On Contractive Linear Multistep and One-Leg Methods

by

Jørgen Sand

## Abstract

Contractive variable-formula methods for the integration of $y' = \lambda(t)y$ ($\lambda(t)$ restricted to some region of interest) are formed by using polyhedral norms for showing the contractivity of several one-leg or linear multistep formulas. When possible the results are given for variable step-size.

## 1. Introduction

This paper contains the main results of the author's thesis ([1], [2] and [3]) concerning stability and boundedness of one-leg and linear multistep methods applied to initial-value problems of the form

$$y' = f(t,y), \quad y(t_0) = y_0 \in \mathbb{R}^s, \quad t \geq t_0 \in \mathbb{R}. \qquad (1.1)$$

The numerical solution of (1.1) is a finite sequence $(y_0, y_1, \ldots, y_N)$ containing approximative values of the exact solution $\{y(t) \mid t \geq t_0\}$ (assumed to exist) on a grid $\{t_i \mid t_0 < t_1 < \ldots < t_N\}$. We will consider two classes of methods for the production of such a sequence, viz. the linear multistep methods (LMM) and the one-leg methods (OLM). Let k be a positive integer. Then a k-step LMM consists of finding $y_{n+k}$, $n = 1-k, 2-k, \ldots$ so that these satisfy the difference equations given by $k_n$-step linear multistep formulas (LMF):

$$\sum_{j=0}^{k} \alpha_{j,n} \cdot y_{n+j} = \bar{h}_{n+k} \cdot \sum_{j=0}^{k} \beta_{j,n} \cdot f(t_{n+j}, y_{n+j}), \qquad (1.2a)$$

where $k \geqq k_n \geqq 1$,

$$\alpha_{k,n} \neq 0, \quad |\alpha_{k-k_n,n}| + |\beta_{k-k_n,n}| \neq 0 \text{ and}$$
$$\alpha_{j,n} = \beta_{j,n} = 0 \text{ for } 0 \leqq j < k-k_n. \tag{1.2b}$$

Likewise, a k-step OLM consists of solving for $n=1-k,2-k,\ldots$ the equations given by $k_n$-step one-leg formulas (OLF):

$$\sum_{j=0}^{k} \alpha_{j,n} \cdot y_{n+j} = \bar{h}_{n+k} \cdot f\Big(\sum_{j=0}^{k} \beta_{j,n} \cdot t_{n+j}, \sum_{j=0}^{k} \beta_{j,n} \cdot y_{n+j}\Big), \tag{1.3a}$$

where

$$\alpha_{k,n} \neq 0, \quad |\alpha_{k-k_n,n}| + |\beta_{k-k_n,n}| \neq 0 \text{ and}$$
$$\alpha_{j,n} = \beta_{j,n} = 0 \text{ for } 0 \leqq j < k-k_n. \tag{1.3b}$$

The LMF's and OLF's are chosen so that

$$k_n \leqq n+k \tag{1.4a}$$

and the consistency (and normalization) requirements

$$\bar{h}_{n+k} = \sum_{j=0}^{k} \alpha_{j,n} \cdot t_{n+j}, \quad \sum_{j=0}^{k} \alpha_{j,n} = 0, \quad \sum_{j=0}^{k} \beta_{j,n} = 1 \tag{1.4b}$$

are met for $n = 1-k, 2-k, \ldots$ . Most often $\bar{h}_{n+k} \neq 0$ and may be replaced by the step-size $h_{n+k} = t_{n+k} - t_{n+k-1}$ used in the (n+k)'th step.

On certain mild assumptions the question of stability and boundedness of these methods can be dealt with by considering their application to pseudo-linear differential systems, i.e. problems (1.1), where $f(t,y) = \Lambda(t,y)y$ for some matrix-valued function $\Lambda$. If $\Lambda$ for all relevant values of t and y can be chosen sufficiently diagonally dominant, we may restrict ourselves to considering scalar equations only, i.e. problems of the form

$$y' = \lambda(t,y)y, \quad y(t_0) = y_0 \in \mathbb{C}, \quad t \geq t_0 \in \mathbb{R}. \tag{1.5}$$

Let $C_n$ denote the companion matrix of a k-step LMM or OLM applied to (1.5)

$$C_n = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ \gamma_{0,n} & \gamma_{1,n} & \cdots & \cdots & \gamma_{k-1,n} \end{bmatrix}, \quad n = 1-k, 2-k, \ldots,$$

where

$$\gamma_{j,n} = -(\alpha_{j,n} - q_{j,n} \cdot \beta_{j,n})/(\alpha_{k,n} - q_{k,n} \cdot \beta_{k,n}), \quad j = 0(1)k-1,$$

and

$$q_{j,n} = \begin{cases} \bar{h}_{n+k} \cdot \lambda(t_{n+j}, y_{n+j}) & \text{if an LMM is considered} \\ \bar{h}_{n+k} \cdot \lambda\left(\sum_{j=0}^{k} \beta_{j,n} t_{n+j}, \sum_{j=0}^{k} \beta_{j,n} y_{n+j}\right) & \text{if an OLM is considered} \end{cases} \tag{1.6}$$

We are then interested in showing the finiteness of

$$M = \sup_{1-k \leq N_1 \leq N_2 < \infty} \left\| \prod_{i=N_1}^{N_2} C_i \right\|, \tag{1.7}$$

where $\|\cdot\|$ denotes some operator norm. In the case of OLM's the contractivity condition

$$\| C_n \| \leq 1 \quad \text{for all } n \geq 1-k \tag{1.8}$$

becomes necessary (and sufficient) for the finiteness of M in (1.7), unless we place some restrictions on the order in which the different formulas and step-sizes are used.

In this paper we consider contractivity of OLM's and LMM's with respect to operator norms corresponding to vector norms with "corners" on their unit sphere, viz.

(i) The max-norm: $\| X \| = \max_{1 \leq i \leq k} | X^{(i)} |$.

(ii) Polyhedral norms: $\| X \| = \alpha$, where $X = \alpha X^*$,

$X^* \in \partial W$.

(W a convex balanced polyhedral neighbourhood of the origin)

(iii) Scaled max-norms: $\| X \| = \max_{1 \leq i \leq k} |(T^{-1}X)^{(i)}|$

The reason for not choosing e.g. an inner-product norm is the following fact that if all the formulas satisfy the algebraic condition for strong zero-stability

$$\rho_n(r) = 0 \Rightarrow (r = 1 \text{ or } |r| < 1), \quad \rho_n(r) \doteq \sum_{j=0}^{k} \alpha_{j,n} r^j ,$$

then ( cf. theorem 3.1 in [4]) condition (1.8) cannot be satisfied in a matrix norm corresponding to a differentiable vector norm if $\lambda \equiv 0$, unless $\rho_n$ is constant.

In section 2, we find that there exist LMM's with formulas of arbitrarily high order for which (1.8) is satisfied in the max-norm, as long as $q_{j,n} \in \{z \mid |z+a| \leq a\}$ for some constant $a > 0$ and the step-ratios $h_{n+k+1}/h_{n+k}$ do not exceed some upper limit greater than one. For fixed step-size a comparison of some of these formulas with the popular Adams-Moulton formulas shows that the "strongly 0-contractive" multistep formulas are reasonably stable and accurate, too.

For the integration of stiff differential equations, the backward differentiation formulas (the BDF's) are widely used and in section 3, we construct a polyhedral norm with respect to which (1.8) is satisfied for the 3-step LMM consisting of the BDF's of order one to three, as long as $q_{j,n} \leq 0$ and certain combinations of order and step-ratio are excluded.

Scaled norms are discussed in section 4, and by imbedding a one-parameter coordinate-transformation of $\check{C}^k$ in the max-norm, we develop a matrix-norm (called the (b,k)-norm), which is a generalization of a norm used by Brayton and Conley in [5]. By appropriate adjustment of k, which for a given set of formulas is only limited from below, and the parameter b, we are (in section 5) e.g. able to find for fixed step-size a 5-step OLM with formulas of order one to five for which (1.8) is satisfied in a (b,5)-norm, as long as $|q_{j,n}+a| \leq a$ for some $a > 0$. Furthermore, we find a 4-step OLM with fixed-step formulas of order one to four for which (1.8) is satisfied in a (b,4)-norm as long as

$$
\begin{array}{lll}
|\arg(-q_{j,n})| \leq 90^0 & \text{(if the first-order formula is used)} & \\
|\arg(-q_{j,n})| \lesssim 89^0 & \text{(- - second-order - - - )} & \\
|\arg(-q_{j,n})| \lesssim 45^0 & \text{(- - third-order - - - )} & (1.9) \\
|q_{j,n}+a| \leq a, a\approx 0.2 & \text{(- - fourth-order - - - )} &
\end{array}
$$

## 2. Strongly 0-contractive LMM's

In [1], we used the phrase "the LMM is $l_\infty$-decreasing in $Q \subseteq \tilde{C}$" if $\|C_n\|_\infty \leq 1$ was valid for all $q_{j,n} \in Q$. The new concept $l_\infty$-decrease was introduced to emphasize the variability of $\lambda$ in our test-equation $y' = \lambda(t)y$. Since we will not recommend the concept for general use, we shall only describe the LMM's by the adjective "contractive", although the ability of $\lambda$ to vary relatively freely prevents most LMM's from satisfying the contractivity condition (1.8).

Consider, for example, the principal root r of a zero-stable linear multistep formula. If $q_{j,n} = -a \cdot \delta_{1,j}$, then

$$r = 1 - a\beta_{1,n} / \rho_n'(1) + O(a^2), \quad \rho_n(\xi) \doteq \sum_{j=0}^{k} \alpha_{j,n} \xi^j$$

and thus

$$\alpha_{k,n} \cdot \beta_{1,n} \geq 0 \text{ for all } l,n \tag{2.1}$$

is a necessary requirement for the formula to be "strongly 0-contractive" in some norm (allowing $q_{j,n}$ to vary freely in some disk, tangential to the origin, lying in the left half-plane). We note that (2.1) rules out most of the well-known classes of linear multistep formulas. Among these the Adams formulas of order greater than two.

Condition (2.1) is clearly not sufficient for this kind of strong 0-contractivity, but if we choose to operate with the max-norm, necessary and sufficient conditions are easily derived.

Theorem 2.1

Let $C_n$ in (1.8) denote the companion matrix of a k-step LMM with $\alpha_{k,n}>0$. Then (1.8) is valid in the max-norm for all $(q_{0,n}, q_{1,n}, \dots, q_{k,n}) \in [-a_n, 0]^{k+1}$ ($a_n$ some positive real number), if and only if $\beta_{k,n} \geqq 0$ and

$$\left.\begin{array}{l} \alpha_{j,n} \leqq 0, \quad \beta_{j,n} \geqq 0 \\ \alpha_{j,n} = 0 \Rightarrow \beta_{j,n} = 0 \end{array}\right\} \quad \text{for } j = 0(1)k-1. \tag{2.2}$$

If $\beta_{j,n} = 0$ for all $j \in [0, k-1]$ then $\| C_n \|_\infty \leqq 1$ for all $(q_{0,n}, q_{1,n}, \dots, q_{k,n}) \in \{z \mid |z-(\alpha_{k,n}/\beta_{k,n})| \geqq |\alpha_{k,n}/\beta_{k,n}|\}^{k+1}$. Otherwise $\|C_n\|_\infty \leqq 1$ is valid for all $(q_{0,n}, q_{1,n}, \dots, q_{k,n}) \in \{z \mid |z+r_n| \leqq r_n\}^{k+1}$, where $r_n = \min\limits_{\substack{j=0(1)k-1 \\ \beta_{j,n} \neq 0}} |\alpha_{j,n}/\beta_{j,n}|$.

Proof:

Dropping the subscript n we now find necessary conditions for

$$\sum_{j=0}^{k-1} |\alpha_j - q_j \beta_j| \leqq |\alpha_k - q_k \beta_k|$$

to hold for all $q_j \in [-a, 0]$, a being some positive constant. First let $q_j = 0$, $j = 0(1)k$. We then find that

$$\sum_{j=0}^{k-1} |\alpha_j| = |\alpha_k|, \text{ i.e. } \underline{\alpha_j \leqq 0.}$$

Letting all the $q_j$'s but one be zero we find that

$$|\alpha_j - q_j \beta_j| \leqq |\alpha_j| \quad \text{for } j = 0(1)k-1. \tag{2.3}$$

It follows that <u>either $\beta_j$ is zero or</u>

$$|q_j - (\alpha_j/\beta_j)| \leq |\alpha_j/\beta_j|,$$

i.e. <u>$\alpha_j\beta_j < 0$.</u>

Letting all the $q_j$'s apart from $q_k$ be zero we find that

$$|\alpha_k - q_k\beta_k| \geq |\alpha_k| \tag{2.4}$$

and <u>$\beta_k \geq 0$</u> follows. We have now found all the conditions in the theorem to be necessary. That they are sufficient for $\|C_n\|_\infty \leq 1$ to hold, when $(q_0, q_1, \ldots, q_k)$ belongs to the regions described, is easily seen since (2.3) and (2.4) will be valid and thus

$$\sum_{j=0}^{k-1} |\alpha_j - q_j\beta_j| \leq \sum_{j=0}^{k-1} |\alpha_j| = |\alpha_k| \leq |\alpha_k - q_k\beta_k|. \qquad \square.$$

In order to check (2.2) for some linear multistep formulas we would like to have explicit expressions for the coefficients $\alpha_{j,n}$ and $\beta_{j,n}$, even in the case of variable step-size. By modifying the derivation of fixed-step formulas of order 2k (k=the step-number) made in [6], we obtain a lemma that can be applied for that purpose.

<u>Lemma 2.2</u>

For given sets $J_{\alpha,n}, J_{\beta,n}$ satisfying $J_{\beta,n} \subseteq J_{\alpha,n} \subseteq \{0,1,\ldots,k\}$ and $\max J_{\alpha,n} \equiv k$, there exists a linear multistep formula

$$\sum_{j=0}^{k} \alpha_{j,n} y_{n+j} = h_{n+k} \cdot \sum_{j=0}^{k} \beta_{j,n} f_{n+j} \quad \text{of order } r_n + s_n - 2,$$

where

$r_n \doteq$ number of elements in $J_{\alpha,n}$ (assumed $\geq 2$),

$s_n \doteq$ number of elements in $J_{\beta,n}$ (assumed $\geq 1$),

$\{j \mid \alpha_{j,n} \neq 0\} = J_{\alpha,n}$ and $\{j \mid \beta_{j,n} \neq 0\} = J_{\beta,n}$.

Apart from a normalization factor, this formula is uniquely determined by $J_{\alpha,n}$ and $J_{\beta,n}$ and given by

$$\beta_{j,n} = \Big(\prod_{i\varepsilon J_{\alpha,n}\setminus\{j\}} \tau_{j,i}\Big)\Big(\prod_{i\varepsilon J_{\beta,n}\setminus\{j\}} \tau_{j,i}\Big),\ \forall j\ \varepsilon J_{\beta,n},$$

$$\alpha_{j,n} = \begin{cases} \beta_{j,n}\cdot\Big[\sum_{i\varepsilon J_{\alpha,n}\setminus\{j\}} \tau_{j,i} + \sum_{i\varepsilon J_{\beta,n}\setminus\{j\}} \tau_{j,i}\Big] & \text{if } j\varepsilon J_{\beta,n}, \\ -\Big(\prod_{i\varepsilon J_{\alpha,n}\setminus\{j\}} \tau_{j,i}\Big)\cdot\Big(\prod_{i\varepsilon J_{\beta,n}} \tau_{j,i}\Big) & \text{if } j\varepsilon J_{\alpha,n}\setminus J_{\beta,n}, \end{cases}$$

where $\tau_{j,i} = h_{n+k}/(t_{n+j} - t_{n+i})$ for $i \neq j$.

Proof:

Given $r_n+s_n$ numbers $y_{n+i}$, $f_{n+j}$, $i\varepsilon J_{\alpha,n}$, $j\varepsilon J_{\beta,n}$, it is well-known from interpolation theory that there exists a unique polynomial $p$ of degree at most $r_n+s_n-1$ satisfying

$$p(t_{n+i}) = y_{n+i},\quad p'(t_{n+j}) = f_{n+j},\quad i\varepsilon J_{\alpha,n},\ j\varepsilon J_{\beta,n}. \tag{2.5}$$

If p is of degree $r_n+s_n-1$ exactly, (2.5) cannot be satisfied by a polynomial of degree less than $r_n+s_n-1$. In other words: all polynomials q of degree $r_n+s_n-2$ or less satisfying (2.5) will also satisfy

$$\sum_{j\varepsilon J_{\alpha,n}\setminus J_{\beta,n}} \{\phi_j(t_{n+j})\psi(t_{n+j})\}^{-1}\cdot q(t_{n+j}) + \sum_{j\varepsilon I_{\beta,n}} \{\phi_j(t_{n+j})\psi'(t_{n+j})\}^{-1}\cdot \tag{2.6}$$

$$[q'(t_{n+j}) - \{\phi_j'(t_{n+j})/\phi_j(t_{n+j}) + \psi_j'(t_{n+j})/\psi_j(t_{n+j})\}\,q(t_{n+j})] = 0$$

( (2.6) is obtained by setting the leading coefficient of p equal to zero and using $q(t_{n+i})=y_{n+i}$, $q'(t_{n+j})=f_{n+j}$). Here

$$\phi_j(x) \doteq \prod_{i\varepsilon J_{\alpha,n}\setminus\{j\}} (x - t_{n+i}), \quad j \varepsilon J_{\alpha,n},$$

$$\psi(x) \doteq \prod_{i\varepsilon J_{\beta,n}} (x - t_{n+i}), \quad \psi_j(x) \doteq \psi(x)/(x-t_{n+j}), \ j\varepsilon J_{\beta,n}.$$

The formula is now obtained by setting $q(t_{n+i})=y_{n+i}$, $q'(t_{n+j})=f_{n+j}$ in (2.6) and then multiplying on both sides by $h_{n+k}^{r_n+s_n-1}$. □.

Combining theorem 2.1 and lemma 2.2 we arrive at the following theoretical result.

Theorem 2.3

For any $p \geq 1$ there exists a linear multistep formula (1.2) of order p with the following properties:

1) The number of non-zero coefficients is p+2.
2) For all step-ratios $h_{n+k+1}/h_{n+k}$ belonging to a certain interval $(0,1+\varepsilon)$, $\varepsilon$ a positive constant, (1.8) is satisfied in the max-norm as long as $(q_{0,n},q_{1,n},\ldots,q_{k,n}) \ \varepsilon \ \{z \mid |z + r| \leq r\}^{k+1}$ for some $r>0$.

Proof:

For p even we may consider the formulas derived in lemma 2.2 with $J_{\alpha,n} = J_{\beta,n}$. It follows that

$$\beta_{j,n} = \Big(\prod_{i\varepsilon J_{\beta,n}\setminus\{j\}} \tau_{j,i}\Big)^2 > 0, \quad \forall j\varepsilon J_{\beta,n}$$

and

$$\alpha_{j,n} = 2\cdot\beta_{j,n}\cdot\Big(\sum_{i\varepsilon J_{\beta,n}\setminus\{j\}} \tau_{j,i}\Big), \quad \forall j\varepsilon J_{\alpha,n}. \tag{2.7}$$

Since sgn $(\tau_{j,i})$ = sgn $(j-i)$, $\alpha_{k,n} > 0$ and we have only to ensure that $\alpha_{j,n}<0$ for all $j\varepsilon J_{\alpha,n}\cap\{0,1,\ldots,k-1\}$. For fixed step-size this can be achieved by e.g. choosing $J_{\alpha,n}=J_{\beta,n}$ = $\{k,k-1,k-1-2^i \mid i=1(1)s\}$ with $s \geqq 0$. For s=0 we obtain the trapezoidal rule which satisfies the given conditions. For $s>0$ we get formulas with $k = 2^{s+1}$ and $p = 2(s+1)$, i.e. $p=k+1$ only if $p\leqq 6$. Let us now show these formulas to satisfy condition (2.2) for fixed step-size. From (2.7) we have

$$\alpha_{k-1,n}=2\cdot\beta_{k-1,n}\cdot\sum_{i\varepsilon J_{\beta,n}\setminus\{k-1\}}(k-1-i)^{-1}=2\cdot\beta_{k-1,n}\cdot\Big(-1+\sum_{i=1}^{s}2^{-i}\Big)<0$$

and for $l\,\varepsilon\,\{1,2,..,s\}$,

$$\begin{aligned}\alpha_{k-1-2^l,n}/(2\beta_{k-1-2^l,n}) &= \sum_{i\varepsilon J_{\beta,n}\setminus\{k-1-2^l\}}(k-1-2^l-i)^{-1}\\ &= -(1+2^l)^{-1} - 2^{-l} - \sum_{i=1}^{l-1}(2^l-2^i)^{-1} + \sum_{i=l+1}^{s}(2^i-2^l)^{-1}.\end{aligned} \tag{2.8}$$

Since

$$\sum_{i=l+1}^{s}(2^i-2^l)^{-1}\leqq\sum_{i=l+1}^{s}(2^{i-1})^{-1}=2^{-l+1}-2^{-s+1}<2^{-l+1}=(2^l-2^{l-1})^{-1},$$

this term is "outweighed" by the $(l-1)$'th term in $\sum_{i=1}^{l-1}(2^l-2^i)^{-1}$ if $l > 1$. For $l=s=1$ the term is zero whereas for $s > 1$ and $l = 1$

$$\sum_{i=l+1}^{s}(2^i-2^l)^{-1}=\frac{1}{2}+\sum_{i=3}^{s}(2^i-2)^{-1}\leqq\frac{1}{2}+\frac{1}{3}\cdot\sum_{i=3}^{s}(2^{i-2})^{-1}<\frac{1}{2}+\frac{1}{3}$$

and the term is "outweighed" by the first two terms in (2.8).

We have now shown that our formulas with even order and fixed step-size satisfy the conditions in theorem 2.1. In order to show 2) note first that, since the coefficients of the formula depend continuously on the step-ratios, the conditions $\alpha_{j,n} < 0$ for $j \varepsilon J_{\alpha,n}$ will remain satisfied for all step-ratios belonging to a certain interval around 1. If any of these step-ratios decreases it is easily seen that the quantities

$$h_{n+j}/|t_{n+j} - t_{n+i}|, \quad j = 1(1)k-1,$$

increase (or remain fixed) if $i > j$ and decrease (or remain fixed) if $i < j$. It follows from this observation that $h_{n+j} \cdot \alpha_{j,n}/h_{n+k}$ (and thus $\alpha_{j,n}$), $1 \leqq j \leqq k-1$, remain negative and the proof is immediate since $\alpha_{k,n} > 0, \alpha_{0,n} < 0$ and $\beta_{j,n} > 0$ (for $j \varepsilon J_{\beta,n}$) hold for all step-sequences.

To obtain formulas of odd order we may simply remove 0 from the set $J_{\beta,n}$ above. Let the new non-vanishing coefficients be denoted by $\tilde{\alpha}_{j,n}$ and $\tilde{\beta}_{j,n}$. Then

$$\tilde{\beta}_{j,n} = \beta_{j,n}/\tau_{j,0} > 0 \; , \tilde{\alpha}_{0,n} = -\tilde{\beta}_{0,n} < 0,$$

$$\tilde{\alpha}_{k,n} = \tilde{\beta}_{k,n} \cdot (2 \cdot \sum_{i \varepsilon J_{\tilde{\beta},n} \setminus \{k\}} \tau_{k,i} + \tau_{k,0}) > 0,$$

$$\tilde{\alpha}_{j,n} = \tilde{\beta}_{j,n} \cdot (2 \cdot \sum_{i \varepsilon J_{\beta,n} \setminus \{j\}} \tau_{j,i} - \tau_{j,0}) < 0 \text{ for } j \varepsilon J_{\alpha,n} \setminus \{0,k\}$$

and all the arguments above can be repeated. $\square$.

Although the theorem above is mainly of theoretical interest the proof gives us an idea of how formulas fulfilling condition 1) and 2) in theorem 2.3 may be derived. Using the computer we have examined all formulas with step-number not exceeding $2^{[(p-1)/2]}+1$ ($p$ = the order $\leqq 12$) and only $p+2$ non-zero coefficients. For even order we chose $J_\beta = J_\alpha$ and for odd order we selected $J_\beta = J_\alpha \setminus \{0\}$.

The formulas are listed for fixed step-size only, but corresponding variable step-size formulas are easily obtained from lemma 2.2 knowing $J_\alpha$ and $J_\beta$ . For $p \leqq 8$ the formulas with the largest radius r (among those examined!) are listed. ($C_{p+1}$ = the error-constant corresponding to $\sum_{j \varepsilon J_\beta} \beta_{j,n} = 1$)

p = 1:

$$y_{n+1}=y_n + hf_{n+1}, \qquad \underline{r=\infty}, \ \underline{C_2=-\tfrac{1}{2}}.$$

p = 2:

$$y_{n+1}=y_n + \frac{h}{2}(f_{n+1} + f_n), \qquad \underline{r=2}, \ \underline{C_3=-\tfrac{1}{12}}.$$

p = 3:

$$y_{n+2}=\frac{1}{5}(4y_{n+1}+y_n) + \frac{2h}{5}(f_{n+2}+2f_{n+1}), \qquad \underline{r=1}, \ \underline{C_4=-\tfrac{1}{36}}.$$

$$y_{n+3}=\frac{1}{28}(27y_{n+2}+y_n) + \frac{3h}{14}(2f_{n+3}+3f_{n+2}), \qquad \underline{r=\tfrac{3}{2}}, \ \underline{C_4=-\tfrac{1}{20}}.$$

p = 4:

$$y_{n+3}=\frac{1}{32}(27y_{n+2}+5y_n) +\frac{3h}{32}(4f_{n+3}+9f_{n+2}+f_n), \qquad \underline{r=1}, \ \underline{C_5=-\tfrac{3}{140}}.$$

p = 5:

$$y_{n+4}=\frac{1}{35}(24y_{n+3}+8y_{n+1}+3y_n) + \frac{12h}{35}(f_{n+4}+3f_{n+3}+f_{n+1}),$$
$$\underline{r=\tfrac{2}{3}}, \ \underline{C_6=-\tfrac{1}{100}}.$$

$$y_{n+5}=\frac{1}{2752}(2025y_{n+4}+700y_{n+2}+27y_n) + \frac{15h}{688}(16f_{n+5}+45f_{n+4}+10f_{n+2}), \qquad \underline{r=\tfrac{3}{4}}, \ \underline{C_6=-\tfrac{1}{71}}.$$

p = 6:

$$y_{n+5}=\frac{1}{5888}(3375y_{n+4}+2000y_{n+2}+513y_n) +\frac{15h}{2944}(64f_{n+5}+225f_{n+4}+100f_{n+2}+9f_n), \qquad \underline{r=\tfrac{1}{2}}, \ \underline{C_7=-\tfrac{10}{1393}}.$$

p = 7:

$$y_{n+6}=\frac{1}{3104}(1215y_{n+5}+1000y_{n+3}+729y_{n+1}+160y_n) +\frac{15h}{1552}(32f_{n+6}+135f_{n+5}+100f_{n+3}+27f_{n+1}), \qquad \underline{r=\tfrac{3}{10}}, \ \underline{C_8=-\tfrac{5}{1372}}.$$

$$y_{n+9}=\frac{1}{197632}(120393y_{n+8}+74088y_{n+6}+2808y_{n+2}+343y_n) +\frac{63h}{24704}(128f_{n+9}+441f_{n+8}+147f_{n+6}+9f_{n+2}), \qquad \underline{r=\tfrac{13}{24}}, \ \underline{C_8=-\tfrac{63}{3625}}.$$

p = 8:

$$y_{n+7}=\frac{1}{180224}(42875y_{n+6}+42875y_{n+4}+83349y_{n+2}+11125y_n) +\frac{105h}{90112}(256f_{n+7}+1225f_{n+6}+1225f_{n+4}+441f_{n+2}+25f_n),$$
$$\underline{r=\tfrac{1}{6}}, \ \underline{C_9=-\tfrac{35}{14274}}.$$

$$y_{n+9}=\frac{1}{928000}(413343y_{n+8}+432000y_{n+6}+76032y_{n+3}+6625y_n)$$
$$+\frac{9h}{46400}(1600f_{n+9}+6561f_{n+8}+3600f_{n+6}+576f_{n+3}+25f_n),$$
$$\underline{r=\frac{7}{20}},\quad \underline{C_9=-\frac{405}{43267}}.$$

p = 9:

$$y_{n+9}=\frac{1}{3386096}(455625y_{n+8}+592704y_{n+6}+2185596y_{n+4}+109296y_{n+1}+42875y_n)$$
$$+\frac{45h}{60466}(392f_{n+9}+2025f_{n+8}+2352f_{n+6}+882f_{n+4}+72f_{n+1}),$$
$$\underline{r=\frac{5}{56}},\quad \underline{C_{10}=-\frac{14}{5723}}.$$

p = 10:

$$y_{n+11}=\frac{1}{1078400000}(235782657y_{n+10}+561515625y_{n+8}+183997440y_{n+5}+70709375y_{n+2}+26394903y_n)$$
$$+\frac{99h}{53920000}(160000f_{n+11}+793881f_{n+10}+680625f_{n+8}+278784f_{n+5}+75625f_{n+2}+6561f_n),$$
$$\underline{r=\frac{3}{20}},\quad \underline{C_{11}=-\frac{7506675}{1145403224}}.$$

p = 11:

$$y_{n+12}=\frac{1}{276200000}(25332021y_{n+11}+137259375y_{n+9}+45999360y_{n+6}+20796875y_{n+3}+40212369y_{n+1}+6600000y_n)$$
$$+\frac{99h}{13810000}(40000f_{n+12}+216513f_{n+11}+226875f_{n+9}+139392f_{n+6}+75625f_{n+3}+19683f_{n+1}),$$
$$\underline{r=\frac{13}{220}},\quad \underline{C_{12}=-\frac{22275}{5026616}}.$$

p = 12:

$$y_{n+14}=\frac{1}{1301272115000}(29894619132y_{n+13}+626971072000y_{n+11}+243729729243y_{n+8}+292876876292y_{n+5}+85153523000y_{n+2}+22646295333y_n)$$
$$+\frac{63h}{454990250}(2044900f_{n+14}+11573604f_{n+13}+13249600f_{n+11}+9018009f_{n+8}+4008004f_{n+5}+828100f_{n+2}+59049f_n),$$
$$\underline{r=\frac{41}{2860}},\quad \underline{C_{13}=-\frac{243243}{40781266}}.$$

The main feature of the formulas above is that they all are "strongly 0-contractive" in the same norm and may thus be combined to form "contractive" LMM's (1.2). The Adams-Moulton formulas of order $p \geq 3$ do not satisfy condition (2.1), but they have reasonable absolute stability and accuracy properties for fixed step-size, and we therefore (in [1]) compared the "size" of the region of absolute stability and the error-constant of these formulas (denoted AM) to those of the formulas listed above with minimal step-number $k_n$ (denoted SC). In the table below

$p_n$, $k_n$ and $C_n$ denote the order, step-number and error-constant, respectively, and

$SI_n$ denotes the stability interval.

| | AM | | | SC | | |
|---|---|---|---|---|---|---|
| $p_n$ | $k_n$ | $C_n \cdot 10^3$ | $SI_n$ | $k_n$ | $C_n \cdot 10^3$ | $SI_n$ |
| 3 | 2 | -41.666 | [-6 ,0] | 2 | -27.777 | [-4 ,0] |
| 4 | 3 | -26.388 | [-3 ,0] | 3 | -21.429 | [-3.56 ,0] |
| 5 | 4 | -18.750 | [-1.84 ,0] | 4 | -10.000 | [-1.78 ,0] |
| 6 | 5 | -14.269 | [-1.18 ,0] | 5 | - 7.179 | [-1.45 ,0] |
| 7 | 6 | -11.367 | [-0.769,0] | 6 | - 3.644 | [-0.853,0] |
| 8 | 7 | - 9.357 | [-0.493,0] | 7 | - 2.452 | [-0.645,0] |
| 9 | 8 | - 7.893 | [-0.310,0] | 9 | - 2.446 | [-0.542,0] |
| 10 | 9 | - 6.786 | [-0.191,0] | 11 | - 6.554 | [-0.808,0] |
| 11 | 10 | - 5.924 | [-0.115,0] | 12 | - 4.431 | [-0.629,0] |
| 12 | 11 | - 5.237 | [-0.068,0] | 14 | - 5.965 | [-0.624,0] |

## 3. $A_0$-Contractivity of the BDF's of order $\leqq 3$.

In 1972 Brayton and Conley ([5]) showed that the usual variable-step version of the second-order BDF was contractive in a scaled max-norm when applied to any differential equation of the form

$$y' = \lambda(t) \cdot y, \quad |\arg(-\lambda(t))| \leqq \arctan(2\sqrt{2}) \approx 70^0 32',$$

with step-ratios $\gamma \leqq \frac{1}{2}(1+\sqrt{3}) \approx 1.366$. If we allow larger values of $\gamma$ the formula is not $A_0$-contractive in any norm. We shall sketch a proof of their result and show a similar result for another variable-step version of the second-order BDF since these proofs illustrate

a. A technique which in certain cases can be used for extending contractivity results in the fixed step-ratio case to results concerning variable step-ratios. This technique (a certain splitting of the companion matrix) will, for example, be used when proving the $A_0$-contractivity of the BDF's of order $p=1(1)3$ (theorem 3.4).

b. The strength of the scaled max-norm derived in section 4. Note that even for constant step-size the second-order BDF is not 0-contractive in the max-norm itself and when the step-size varies, the usual variable-step version does not remain 0-contractive in any (constant) inner-product norm (cf. [4, theorem 3.1]).

c. It may be worthwhile considering variable-step versions of the BDF's other than the usual one (cf. [1]).

To demonstrate the technique mentioned in a. consider the companion matrix of the second-order BDF:

$$\frac{1+2\gamma}{1+\gamma} y_{n+2} - (1+\gamma) y_{n+1} + \frac{\gamma^2}{1+\gamma} y_n = h_{n+2} \cdot f_{n+2}, \quad \gamma \doteq h_{n+2}/h_{n+1},$$

i.e.

$$C(q,\gamma) = \begin{bmatrix} 0 & 1 \\ -\dfrac{\gamma^2}{(1+2\gamma)-q(1+\gamma)} & \dfrac{(1+\gamma)^2}{(1+2\gamma)-q(1+\gamma)} \end{bmatrix}.$$

A simple calculation shows that

$$C(q,\gamma) = z(q,\gamma)\cdot C(0,\gamma) + [1-z(q,\gamma)]\cdot C(\infty,\gamma_{max}),$$

$$C(0,\gamma) = \frac{\gamma^2}{1+2\gamma}\cdot \varepsilon^{-1}\cdot C(0,\gamma_{max}) + [1-\frac{\gamma^2}{1+2\gamma}\cdot \varepsilon^{-1}]\cdot C(0,0),$$

where

$$z(q,\gamma) = \frac{1+2\gamma}{(1+2\gamma)-q\cdot(1+\gamma)} \quad \text{and} \quad \varepsilon = \frac{\gamma^2_{max}}{1+2\gamma_{max}}.$$

If $q \in [-\infty, 0]$, we see that $C(q,\gamma)$ for $\gamma \leqq \gamma_{max}$ is a convex combination of $C(0,\gamma_{max})$, $C(\infty,\gamma_{max})$ and $C(0,0)$ (the latter may be regarded as the companion matrix of the first-order BDF). For most families of one-leg formulas with the same (constant) leading coefficient or the same (constant) ratio between the two leading coefficients in the second characteristic polynomial a similar splitting is possible, but to ensure that the combination is convex certain pairs (formula, step-ratios) may have to be excluded (as we shall see in the proof of theorem 3.4).

<u>Lemma 3.1</u>

For $i = 1(1)k$ let $\tilde{C}_i(q,\gamma_{k+1-i}, \gamma_{k+2-i},\ldots,\gamma_{k-1})$ denote the companion matrix of an i-step consistent one-leg formula with step-ratios $\gamma_s = h_{n+s+1}/h_{n+s}$ and coefficients $((h_{n+k}/\bar{h}_n)\alpha^i_j, \beta^i_j)$, $j = k-i(1)k$. Assume that either $\beta^i_k \equiv b$ or $\beta^i_{k-1}/\beta^i_k \equiv B$ (b and B non-zero constants) for all i, and that $\bar{h}_n\cdot(\alpha^i_{k-i}/\alpha^i_k)\cdot(\beta^i_{k-i}/\beta^i_k) \neq 0$ when $\gamma_s$ are all equal to some number $\eta > 0$. Then, for any $m \in \{1,2,\ldots,k\}$ and any step-ratios $\gamma_s \geqq 0$,

$$C_m(q,\gamma_{k+1-m},\dots,\gamma_{k-1}) = z\cdot\sum_{i=1}^{k} a_i\cdot C_i(0,\eta) + (1-z)\cdot\sum_{i=1}^{k} b_i\cdot C_i(\infty,\eta)$$

for some scalars $a_i$, $b_i$, where $\Sigma a_i = \Sigma b_i = 1$, and $z \varepsilon (0,1]$ for $(\beta_k^m/\alpha_k^m)\cdot \bar{h}_n q \leqq 0$.

Here

$$C_m(q,\gamma_{k+1-m},\dots,\gamma_{k-1}) = \left[\begin{array}{ccccc} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & 0 & 1 & 0 \cdots & 0 \\ \vdots & \vdots & \multicolumn{3}{|c}{} \\ 0 \cdots & 0 & \multicolumn{3}{|c}{\tilde{\tilde{C}}_m(q,\gamma_{k+1-m},\dots,\gamma_{k-1})} \end{array}\right]$$

and $C_i(q,n) = C_i(q,\eta,\dots,\eta)$ denote $k\times k$ companion matrices.

Proof:

Let z be $\alpha_k^m/(\alpha_k^m - \bar{h}_n\cdot q\cdot\beta_k^m/h_{n+k})$. Then we only have to show the existence of scalars $\{a_i\}_{i=1}^{k}$, $\{b_i\}_{i=1}^{k}$ so that $\Sigma a_i = \Sigma b_i = 1$,

$$C_m(0,\gamma_{k+1-m},\dots,\gamma_{k-1}) = \sum_{i=1}^{k} a_i\cdot C_i(0,\eta)$$

and

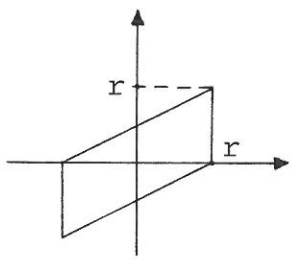
$$C_m(\infty,\gamma_{k+1-m},\dots,\gamma_{k-1}) = \sum_{i=1}^{k} b_i\cdot C_i(\infty,\eta).$$

By looking at the last row of these matrices we find that $a_k = a_{k-1} = \dots = a_{m+1} = b_k = b_{k-1} = \dots = b_{m+1} = 0$ and the rest of the scalars can be found by solving two non-singular triangular systems of equations. Premultiplying the coefficient matrix and the vector on the right-hand side in these systems by either the vector $(1,1,\dots,1)$ or the vector $(1,0,\dots,0)$ it follows that $\Sigma a_i = \Sigma b_i = 1$.

$\square$.

For the usual BDF's we need no splitting of $C_m(\infty,\gamma_{k+1-m},\ldots,\gamma_{k-1})$ and the lemma will be used (with q=0) in this section for splitting $C_m(0,\gamma_{k+1-m},\ldots,\gamma_{k-1})$ only. (Here $\beta^i_{k-i} \neq 0$ is not needed, but $\alpha^i_{k-i} \neq 0$ is).

Proving $A_0$-contractivity of the usual BDF's of order $p \leqq 2$ for all step-ratios $\gamma \leqq \gamma_{max}$ is thus equivalent to finding some norm in which $C(0,\gamma_{max})$, $C(\infty,\gamma_{max})$ and $C(0,0)$ all become contractive! Brayton and Conley succeeded in doing this by noting that for $\varepsilon = \gamma^2_{max}/(1+2\gamma_{max}) \leqq \frac{1}{2}$ all three matrices map a symmetric convex region of the form




into itself. For $\varepsilon > \frac{1}{2}$ (i.e. $\gamma_{max} > \frac{1}{2}(1+\sqrt{3}) \approx 1.366$), an unfortunate combination of the 3 matrices will produce an unstable matrix, since e.g.

$$C(0,\gamma_{max})^n \cdot C(\infty,\gamma_{max}) \cdot \begin{pmatrix}1\\1\end{pmatrix} = \frac{\varepsilon-\varepsilon^{n+1}}{\varepsilon-1} \cdot \begin{pmatrix}1\\1\end{pmatrix} + \varepsilon^n \cdot \begin{pmatrix}1\\0\end{pmatrix}$$

By noticing that the norm derived was of the form

$$\|C\| = \|T^{-1}CT\|_\infty, \quad T = \begin{bmatrix}1 & 0\\0.5 & 0.5\end{bmatrix},$$

they succeeded furthermore in establishing $A(\alpha)$-contractivity with $\alpha = \arctan(2\sqrt{2}) \approx 70^0 32'$. The (b,k)-norm developed in section 4 is a generalization of this norm to matrices of arbitrary dimension k.

Let us try the approach of Brayton and Conley on another variable-step version of the second-order BDF.

Proposition 3.2

The one-leg formula ($\gamma \doteq h_{n+2}/h_{n+1}$)

$$\frac{3}{2}y_{n+2}-2y_{n+1}+\frac{1}{2}y_n=\frac{3\gamma-1}{2\gamma}\cdot h_{n+2}f(t_{n+1}+\beta h_{n+2},\beta y_{n+2}+(1-\beta)y_{n+1}),$$

where $\beta=(3\gamma^2+1)/[2\gamma(3\gamma-1)]$, is contractive in a scaled max-norm (constructed below) for all step-ratios $\frac{1}{3}<\gamma\leq\frac{1}{6}(3+\sqrt{33})\approx 1.457$ when applied to any scalar equation of the form

$$y'=\lambda(t)\cdot y,\quad |\arg(-\lambda(t))|\leq\alpha(\gamma),\alpha(\gamma)>0,$$

$$[\alpha(\gamma)\geq\arccos(\tfrac{1}{5})\approx 78^0 27' \text{ for } \gamma\geq\tfrac{1}{12}(3+\sqrt{33})]$$

By allowing smaller and larger values of $\gamma$, the formula can be made unstable even for $\alpha(\gamma)\equiv 0$.

Proof:

The companion matrix is

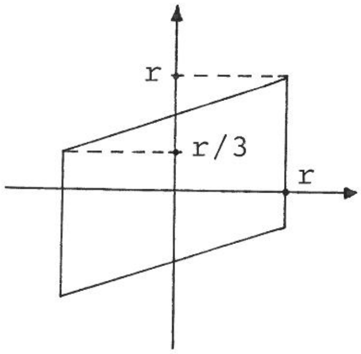
$$C(q,\gamma)=z\cdot C(0,\gamma)+(1-z)\cdot C(\infty,\gamma),$$

where

$$z=6\gamma^2/[6\gamma^2-q(3\gamma^2+1)],$$

$$C(0,\gamma)=\begin{bmatrix}0 & 1\\ -\frac{1}{3} & \frac{4}{3}\end{bmatrix}\text{ and } C(\infty,\gamma)=\begin{bmatrix}0 & 1\\ 0 & -\varepsilon\end{bmatrix},\quad \varepsilon=\frac{3\gamma^2-2\gamma-1}{3\gamma^2+1}.$$

In this case, further splitting is not needed since it is easily seen that for $-1\leq\varepsilon\leq\frac{1}{3}$ the matrices $C(0,\gamma)$ and $C(\infty,\gamma)$ map a symmetric region of the form




into itself.

For $\varepsilon < -1$, $C(\infty,\gamma)$ is unstable and for $\varepsilon > \frac{1}{3}$ an unfortunate combination of the two matrices may produce an unstable matrix:

$$C(0,\gamma)^n \cdot C(\infty,\gamma_{max}) \cdot \begin{pmatrix}1\\1\end{pmatrix} \xrightarrow[n\to\infty]{} -\frac{1}{2}(3\varepsilon+1)\cdot\begin{pmatrix}1\\1\end{pmatrix} .$$

We note that the formula has now been proved $A_0$-contractive for $\gamma \varepsilon (\frac{1}{3},\frac{1}{6}(3+\sqrt{33})]$ ($\gamma=\frac{1}{3}$ excluded per definition of $\beta$) in the matrix norm

$$||C|| = ||T^{-1}CT||_\infty \quad , \quad T = \begin{bmatrix} 1 & 0 \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix} .$$

The fact that the companion matrix is contractive in this norm may be expressed as

$$|q(6\gamma^2-3\gamma-1)| + |18\gamma^2+2q(6\gamma^2-3\gamma-1)| \leqq |18\gamma^2-3q(3\gamma^2+1)| .$$

For $\gamma \varepsilon [(3+\sqrt{33})/12,\ (3+\sqrt{33})/6]$ we find that $0 \leqq 6\gamma^2-3\gamma-1 \leqq 3\gamma^2+1$, and a sufficient condition for contractivity in this case is thus $[z \doteq q(6\gamma^2-3\gamma-1),\ a \doteq 18\gamma^2]$:

$$|z| + |a+2z| \leqq |a-3z|$$
$$\Updownarrow$$
$$|z| \cdot |a+2z| \leqq 2|z|^2 - 5a\cdot \operatorname{Re} z .$$

If $|\arg(-z)| \leqq \arccos(\frac{1}{5})$ then $-5\cdot\operatorname{Re} z \geqq |z|$ and the condition is satisfied. In general, A(0)-contractivity follows from a criterion in [7,p.468] (concerning the max-norm) since $|6\gamma^2-3\gamma-1| \leqq |3\gamma^2+1|$ for the step-ratios in question. $\square$.

Remark

Although the BDF is not $A_0$-contractive in the max-norm, we see from the polygons shown in this section that

$$\| \prod_{n=N_1}^{N_2} C(q_n,\gamma_n) \|_\infty \leqq 3\,(2,\text{ respectively}) \qquad \forall N_1, N_2$$

when $q_n \leqq 0$ and

$$\gamma_n \leqq \frac{1}{2}(1+\sqrt{3}) \quad (\frac{1}{3} \leqq \gamma_n \leqq \frac{1}{6}(3+\sqrt{33}),\text{ respectively})$$

for the usual and the "unusual" variable-step version, respectively, i.e. only a moderate expansion (measured in the max-norm) can occur in these cases. $\square$.

In the notation of section 4, the norm used by Brayton and Conley is the (0.5,2)-norm, whereas the one constructed in the proof of proposition 3.2 is the (1/3,2)-norm. The figures on the next page show for the fixed-step second-order BDF its region of absolute stability (the dashed line) and the regions of contractivity w.r.t. the two norms. The formula is not A-contractive in these norms.

We shall now apply the techniques demonstrated in this section to derive an $A_0$-contractivity result for a method consisting of the BDF's of order $p = 1(1)3$. We start by using lemma 3.1 for splitting $C_3(0,\gamma_1,\gamma_2)$ (although $\beta^i_{k-i} \neq 0$ is not satisfied).
The linear system for determination of $a_1$, $a_2$, $a_3$ (cf. lemma 3.1) is the following:

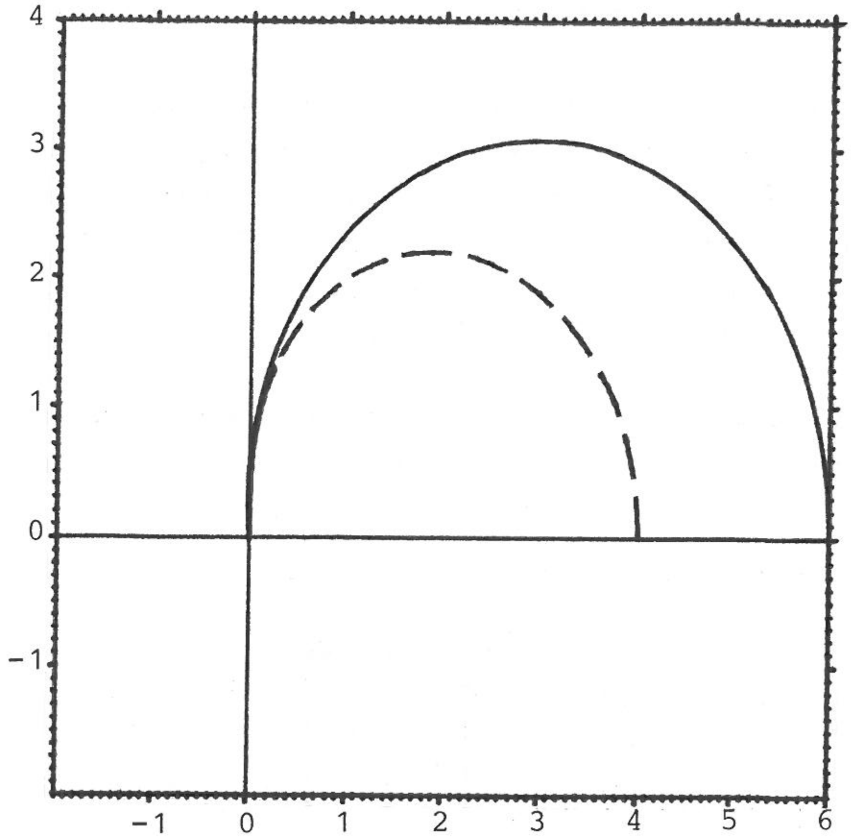



Fig. 3.1. The region of absolute stability and the region of contractivity in the (1/3,2)-norm of the fixed-step second-order BDF.

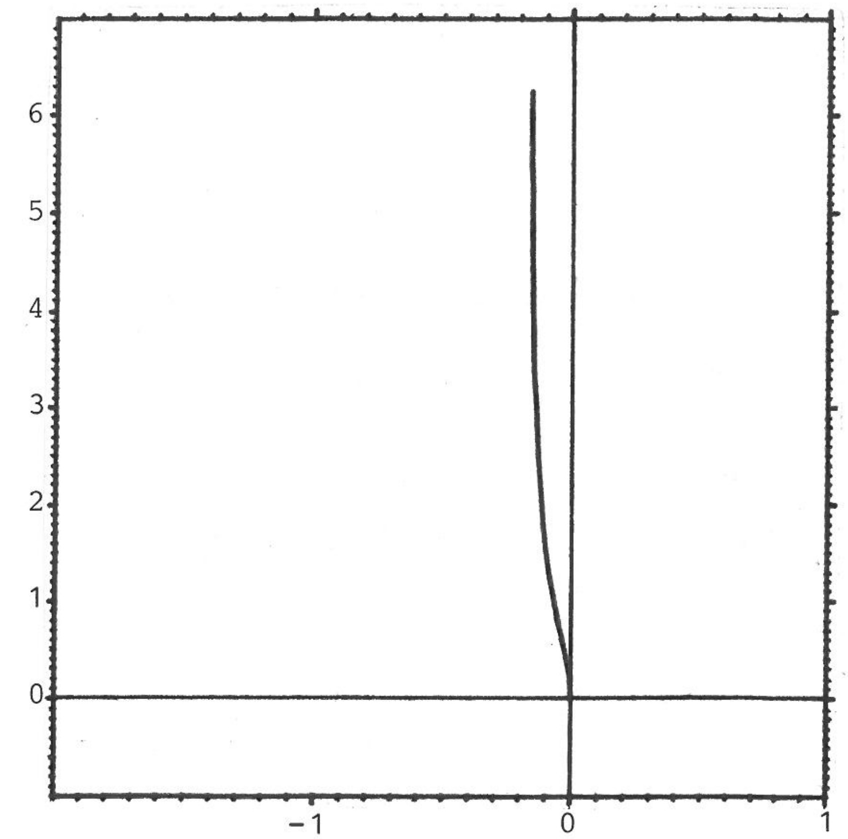



Fig. 3.2. The region of contractivity in the (0.5,2)-norm of the fixed-step second-order BDF.

$$\begin{bmatrix} 1 & \frac{(1+\eta)^2}{1+2\eta} & \frac{(1+\eta)(1+\eta+\eta^2)^2}{3\eta^3+4\eta^2+3\eta+1} \\ 0 & -\frac{\eta^2}{1+2\eta} & -\frac{\eta^2(1+\eta+\eta^2)^2}{3\eta^3+4\eta^2+3\eta+1} \\ 0 & 0 & \frac{\eta^5(1+\eta)}{3\eta^3+4\eta^2+3\eta+1} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \left\{ \begin{matrix} \text{last row in} \\ C_3(0,\gamma_1,\gamma_2) \\ \text{transposed} \end{matrix} \right\}$$

Setting $\gamma_1=\eta \cdot \frac{a}{1+a}$ and $\gamma_2= \eta \cdot \frac{b}{1+b}$ where a, $b \geqq 0$, the solution of this system will be

$\underline{a_3= p_3/q_3}$, where

$$p_3= a^3b^2(3\eta^3+4\eta^2+3\eta+1)(b^2\eta^2+2b(1+b)\eta+(1+b)^2),$$

$$q_3= (1+a)(1+b)^2(1+\eta)\left[3a^2b^2\eta^4+ab(4a+7ab+3b)\eta^3+a(1+b)(a+7ab+6b)\eta^2\right.$$
$$\left.+\ 2(1+a)(1+b)(a+2ab+b)\eta+(1+a)^2(1+b)^2\right].$$

$\underline{a_2= p_2/q_3}$, where

$$p_2= b^2(1+2\eta)\cdot\left[a^2b^2\eta^5+ a^2b(2a+5b+4)\eta^4 + a((1+b)(3a+2a^2+11ab+2b)\right.$$
$$+\ 2a^2 b)\eta^3 + a((1+b)(3+9a+3a^2+5b+13ab) +a^2b)\eta^2$$
$$\left.+\ (1+b)((1+b)(1+6a+9a^2)+2a^3)\eta+(1+b)^2(1+3a+3a^2)\right].$$

$\underline{a_1= p_1/q_3}$, where

$$p_1= (1+\eta)\cdot \sum_{i=0}^{3} \sum_{j=0}^{4} C_{ij}a^ib^j,$$

$C_{00}= 1$, $C_{10} = 2\eta+3$, $C_{01}=2\eta+4$, $C_{20}=\eta^2+4\eta+3$,

$C_{11}= 6\eta^2+14\eta+12$, $C_{02}=4\eta+5$, $C_{30}=(\eta+1)^2$,

$C_{21}= 4\eta^3+16\eta^2+22\eta+12$, $C_{12}=3\eta^3+12\eta^2+21\eta+15$,

$C_{03} = 2(\eta+1)$, $C_{31} = 2(\eta+1)(2\eta^2+3\eta+2)$,

$C_{22} = 3\eta^4+12\eta^3+27\eta^2+30\eta+15$, $C_{13} = 2(\eta+1)(\eta^2+\eta+3)$, $C_{04} = 0$,

$C_{32} = (\eta+1)(3\eta^3+8\eta^2+9\eta+5)$, $C_{23} = 2(\eta+1)(-\eta^3+\eta^2+2\eta+3)$,

$C_{14} = -\eta(\eta+1)^2$, $C_{33} = 2(\eta^2+\eta+1)(\eta+1)^2$,

$C_{24} = -2\eta(\eta^2+\eta+1)(\eta+1)^2$, $C_{34} = 0$.

Thus we find that $a_2$, $a_3$ are non-negative for all $(a,b)$, whereas some requirements on the relation between a and b (i.e. $\gamma_1$ and $\gamma_2$) are needed in order to ensure $a_1 \geqq 0$. Let us first find $\eta$, so as to make $C_3(0,\eta)$, $C_2(0,\eta)$, $C_1(0,\eta)$, $C_1(\infty,\eta)$ a stable family of matrices.

Lemma 3.3

$C_3(0,\eta)$, $C_2(0,\eta)$, $C_1(0,\eta)$ and $C_1(\infty,\eta)$ is a stable family of matrices if $\eta = 1.011588\ldots$ and not if $\eta$ is larger.

Proof:

Let $\varepsilon$ denote $\eta^2/(1+2\eta)$. Then for $|\varepsilon| < 1$

$$C_2(0,\eta)^n \xrightarrow[n\to\infty]{} \frac{1}{1-\varepsilon}(1,1,1)^T\cdot(0,-\varepsilon,1).$$

If we for notational convenience write $C_3(0,\eta)$ as

$$C_3(0,\eta) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \gamma_0 & \gamma_1 & \gamma_2 \end{bmatrix}$$

we find that for $|\varepsilon| < 1$

$$C_2(0,\eta)^n \cdot C_3(0,\eta)^2 \cdot C_1(\infty,\eta)\cdot(1,1,1)^T$$

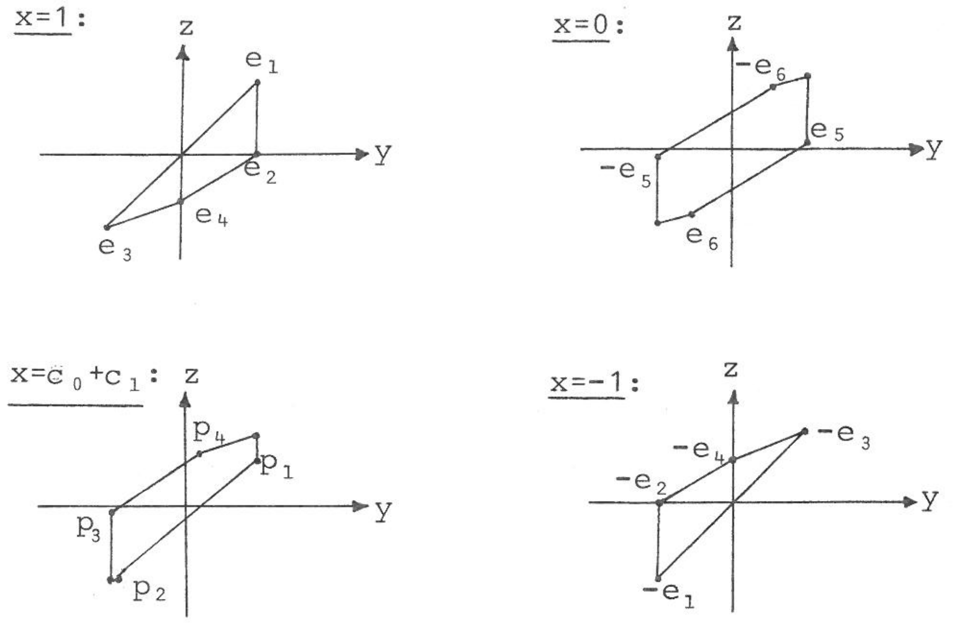
$$\xrightarrow[n\to\infty]{} [\gamma_0+\gamma_2(\gamma_0+\gamma_1)-\varepsilon(\gamma_0+\gamma_1)]\,/(1-\varepsilon)\cdot(1,1,1)^T.$$

If $\eta > 1.011588\ldots$ we will have $\gamma_0+\gamma_2(\gamma_0+\gamma_1)-\varepsilon(\gamma_0+\gamma_1)<\varepsilon-1$, and hence for n sufficiently large $C_2(0,\eta)^n \cdot C_3(0,\eta)^2 \cdot C_1(\infty,\eta)$ will possess an eigenvalue of modulus larger than one. On the other hand, if $\eta = 1.011588\ldots$, it is possible to find a symmetric convex neighbourhood W of the origin so that all the matrices in question map W into itself. The following region will do:

W = the convex hull of the set

$$\left\{\pm\begin{bmatrix}1\\1\\1\end{bmatrix},\ \pm\begin{bmatrix}1\\1\\0\end{bmatrix},\ \pm\begin{bmatrix}1\\-1\\-1\end{bmatrix},\ \pm\begin{bmatrix}1\\0\\c_0+c_1\end{bmatrix},\ \pm\begin{bmatrix}0\\1\\c_0\end{bmatrix},\ \pm\begin{bmatrix}0\\(c_0+c_1+1)-1\\ \mu(c_0+c_1+1)-1\end{bmatrix}\right\}$$

Let $\pm e_1, \pm e_2, \ldots, \pm e_6$ denote the extreme points of W. Then we can illustrate W in $\grave{R}^3 = \{(x,y,z)^T \mid x,y,z \in \grave{R}\}$ by considering intersections with different planes. (Note, however, that W may be larger than indicated by these intersections).




$p_1 = (1+x)e_5+xe_3$, $\qquad p_2 = (1+x)e_6+xe_1$,

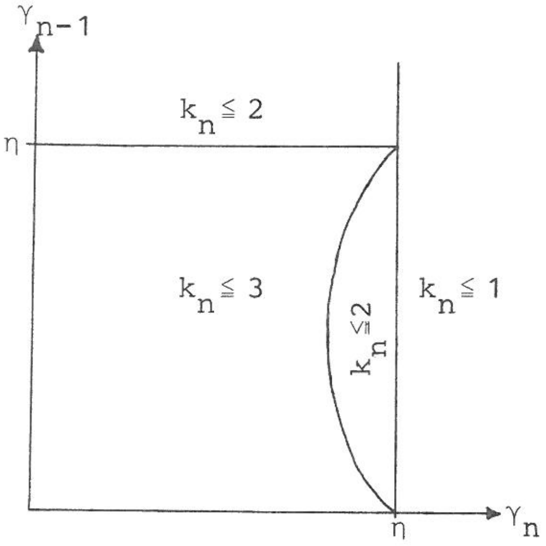
$p_3 = -(1+x)e_5+xe_2$, $\qquad p_4 = -(1+x)e_6+xe_4$.

Using these intersections, a straight-forward (but tedious) calculation will show that W is mapped into itself by any of the matrices in question if $\eta = 1.011588\ldots$ □.

By means of this lemma we can now show the main result of this section.

Theorem 3.4

Consider a variable step-size method based on the first, second and third order BDF. Let $k_n$ and $\gamma_n$ denote the order and the (last) step-ratio, respectively, used in the n'th integration step. If the variation of $(k_n, \gamma_n)$ for all n is confined as shown below, the method will produce bounded solutions to any (scalar) equation of the form $y' = \lambda(t,y)\cdot y$, $\lambda(t,u) \leqq 0 . \forall\, t,u$ .




$(\eta = 1.011588\ldots)$.

Proof:

The coefficients $C_{ij}$ on p. 23-24 are (rounded)

| | | | | |
|---|---|---|---|---|
| 1.000 | 6.023 | 9.046 | 4.023 | 0.000 |
| 5.023 | 32.302 | 51.629 | 20.256 | -4.093 |
| 8.070 | 54.769 | 88.541 | 20.161 | -24.846 |
| 4.046 | 28.490 | 51.087 | 24.561 | 0.000 |

We see that if b(cf. p.23) is "very large" $p_1 \geqq 0$ is possible only if a is "very large" or "very small". Using a root-seeking algorithm we find that $p_1 \geqq 0$ if

i) $(\gamma_n/\eta) \leqq .809..$

ii) $(\gamma_n/\eta) = .81$ and $|(\gamma_{n-1}/\eta)-.475| \geqq .036$

iii) $(\gamma_n/\eta) = .82$ and $|(\gamma_{n-1}/\eta)-.478| \geqq .137$

iv) $(\gamma_n/\eta) = .85$ and $|(\gamma_{n-1}/\eta)-.485| \geqq .259$

v) $(\gamma_n/\eta) = .90$ and $|(\gamma_{n-1}/\eta)-.493| \geqq .371$

vi) $(\gamma_n/\eta) = .99$ and $|(\gamma_{n-1}/\eta)-.4999| \geqq .4899$

vii) $(\gamma_n/\eta) = .999$ and $|(\gamma_{n-1}/\eta)-.5| \geqq .499...$

For $k_n = 2$, $a = 0$ and $p_1 \geqq 0$ is satisfied for all $b \geqq 0$ (all $\gamma_n \varepsilon [0,\eta]$). If $k_n = 1$ the companion matrix will map W (cf. the proof of lemma 3.3) into itself regardless of $(\gamma_{n-1},\gamma_n)$. $\square$.

Corollary 3.5

On the assumptions in theorem 3.4, the companion matrices satisfy

$$\|\prod_{n=N_1}^{N_2} C(q_n)\|_\infty \lesssim 3.83, \quad \forall\ N_1,\ N_2.$$

Proof:

We shall show that $A = \{(x,y,z)^T \mid \|(x,y,z)^T\|_\infty \leqq \delta\}$ is a subset of W (cf. the proof of lemma 3.3) if $\delta \lesssim 0.261414$, and the proof will follow, since $\delta^{-1} \lesssim 3.83$. A is convex and hence we only consider its extreme points. From the intersections shown p. 25 it is evident that $\pm(\delta,\delta,\delta)^T, \pm(\delta,-\delta,-\delta)^T \varepsilon W$.

Furthermore, for $\delta \approx 0.261414$, we obtain with $e_6 = (0,-a,-b)$:

$$\begin{bmatrix} \delta \\ -\delta \\ \delta \end{bmatrix} = \delta \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \frac{2\delta-(b-a)\cdot c}{1-\gamma_0} \cdot \begin{bmatrix} 0 \\ -1 \\ -\gamma_0 \end{bmatrix} + c \cdot \begin{bmatrix} 0 \\ a \\ b \end{bmatrix} ,$$

$$c = \frac{\delta(\gamma_0+3)+\gamma_0-1}{2b-a(\gamma_0+1)-(1-\gamma_0)}$$

and hence $\pm(\delta,-\delta,\delta)^T$ also lie in W. The last extreme points to be considered are $\pm(\delta,\delta,-\delta)^T$, but these points evidently lie in W, since for $\delta \approx 0.261414$:

$$\pm(0,\delta,-\delta)^T, \pm(1,\delta,-\delta)^T \varepsilon \, W .$$

□.

In the light of theorem 3.4, it is not surprising that unboundedness (or instability if $\lambda$ is independent of y) may occur even for fixed step-size if the 4'th order BDF is included in the method (and no restrictions are imposed on the variation of $\lambda$). This is a fact which Brayton and Conley pointed out in [5]. In our notation they showed that for a k-step consistent one-leg formula $(\rho_n, \sigma_n)$ the pair of matrices

$$H = \{C(0,\eta,\eta,\ldots,\eta),\ C(q,\eta,\eta,\ldots,\eta)\}$$

is unstable if q satisfies

$$\left| \frac{\alpha_{k,n} q}{\alpha_{k,n} - q\beta_{k,n}} + \rho_n'(1) \right| > \left| \rho_n'(1) \right| ,$$

where $(\rho_n, \sigma_n)$ is assumed constant for fixed step-ratio $\eta$. In the case of fixed step-size $(\eta=1)$, the k-step BDF has $\beta_k = 1$, $\alpha_k = \sum_{j=1}^{k} \frac{1}{j}$, $\rho'(1) = 1$ and hence we have an unstable pair if $k \geq 4$ and $q < -2 \cdot [1+24\cdot(1+\sum_{j=5}^{k} \frac{12}{j})^{-1}]$ .

## 4. Contractivity in Scaled Norms

In this section we shall define a generalization of the scaled max-norm used by Brayton and Conley when proving $A(\alpha)$-contractivity of the second-order BDF [5]. This generalized norm has the form $||C||_b = ||T_b^{-1} C T_b||_\infty$, where $T_b$ is a matrix dependent only on the parameter b. As an important by-product of our considerations, we will obtain a transformation which may be useful in showing contractivity even in the case of non-linear problems (cf. theorem 4.4). Before specifying $T_b$, let us indicate how powerful a tool similarity transformations is in connection with companion matrices.

Lemma 4.1

Let C be a companion matrix and A a matrix with the same characteristic polynomial. Then C is similar to A if and only if the minimal polynomial of A has the same degree as the characteristic polynomial.

Proof: Follows from Householder [8,p.150 and p.18].

We may express the condition in lemma 4.1. in another way if A is some "well behaved" transformation of a companion matrix. The transformations, g, which we will consider are analytic in some (open) region $\Omega$ containing the spectrum, $\Lambda$, of the companion matrix and univalent on $\Lambda$(i.e. $g(\lambda_i) \neq g(\lambda_j)$ when $\lambda_i \neq \lambda_j$ and $\lambda_i, \lambda_j \in \Lambda$).(*)

Lemma 4.2

Let C be a companion matrix with eigenvalues $\{\lambda_i\}_{i\in I}$ and $C_g$ the companion matrix with the eigenvalues $\{g(\lambda_i)\}_{i\in I}$,

(*) In the literature univalence is most often used in connection with regions.

where $g$ is analytic in a region $\Omega \supset \{\lambda_i\}_{i\in I}$ and univalent on $\{\lambda_i\}_{i\in I}$. Then

$$C_g = T^{-1}g(C)T \quad \text{for some matrix } T$$

if and only if

$$g'(\lambda)\neq 0 \text{ for all multiple eigenvalues of } C.$$

<u>Proof</u>:

We start by considering the Jordan Canonical Form (JCF) of the companion matrix $C$ (cf. e.g. [2,p.3]):

$$C = V\cdot J\cdot V^{-1}, \text{ i.e. } g(C) = V\cdot g(J)\cdot V^{-1}.$$

Since $g$ is univalent on the spectrum of $C$, $C_g$ can be represented by a JCF of the same block-structure as the JCF of $C$, say

$$C_g = V_g J_g V_g^{-1}.$$

It follows that $C_g$ is similar to $g(C)$ iff $J_g$ is similar to $g(J)$. We therefore consider the equation $J_g\cdot X = X\cdot g(J)$. Let $J_g = \underset{i=1(1)s}{\mathrm{blockdiag}}\{J_{g,i}\}$, then $X = \{X_{i,j}\}_{i,j=1}^{s}$, where $J_{g,i}\cdot X_{i,j} = X_{i,j}\cdot g(J_j)$. These equations have the form

$$\begin{bmatrix} g(\lambda_i) & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & g(\lambda_i)\end{bmatrix} \cdot \begin{bmatrix} x_1^T \\ \cdot \\ \cdot \\ \cdot \\ x_{m_i}^T\end{bmatrix} = \begin{bmatrix} x_1^T \\ \cdot \\ \cdot \\ \cdot \\ x_{m_i}^T\end{bmatrix} \cdot \begin{bmatrix} g(\lambda_j) & \cdots & \frac{1}{(m_j-1)!}g^{(m_j-1)}(\lambda_j) \\ 0 & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ 0 & \cdots\ 0 & g(\lambda_j)\end{bmatrix},$$

i.e.

$$\begin{bmatrix} x_2^T \\ \vdots \\ x_{m_i}^T \\ 0^T \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ \vdots \\ x_{m_i}^T \end{bmatrix} \cdot \begin{bmatrix} g(\lambda_j)-g(\lambda_i) & \cdots & \frac{1}{(m_j-1)!} g^{(m_j-1)}(\lambda_j) \\ 0 & \ddots & \vdots \\ 0 & 0 & g(\lambda_j)-g(\lambda_i) \end{bmatrix} \tag{4.1}$$

Comparing elements in the matrices on both sides (starting with the last row) we observe that $X_{ij} = 0$ if $i \neq j$. If $i = j$ we find by comparing the elements (now starting with the first row) that $X_{i,i}$ is upper triangular with the diagonal elements $g'(\lambda_i)^{\ell-1} \cdot (1,0,\dots,0) \cdot x_1$, $\ell = 1(1)m_i$. Since the multiplicity of $\lambda_r$ is $m_r$, we see that $J_g \cdot X = X \cdot g(J)$ has a non-singular solution <u>iff</u> $g'(\lambda_r) \neq 0$ for all multiple eigenvalues $\lambda_r$. On this assumption, $C_g = T^{-1} g(C) T$ follows for

$$T = VX^{-1}V_g^{-1} \tag{4.2}$$

□.

If the similarity transformation in the lemma above is to be incorporated in a norm, it will be most useful if T is independent of C.

<u>Theorem 4.3</u>

The matrix T in lemma 4.2 can be chosen independent of C if and only if g(C) has the form

$$g(C) = (-c \cdot C + a \cdot I)^{-1} \cdot (d \cdot C - b \cdot I), \quad ad \neq bc . \tag{4.3}$$

In this case $T = \{t_{ij}\}$ can be chosen so that for $k = \dim(C)$:

$$\sum_{j=1}^{k} t_{ij} \lambda^{j-1} = (c\lambda+d)^{k-1} \cdot \left(\frac{a\lambda+b}{c\lambda+d}\right)^{i-1}, \quad i = 1(1)k, \ \forall \lambda. \tag{4.4}$$

<u>Proof</u>:

In order to show that g must be of the stated form, split $V_g$, X and V into blocks:

$V_g = [V_{g,1}, V_{g,2}, \ldots V_{g,s}]$, $X = \underset{\ell=1(1)s}{\mathrm{blockdiag}} \{X_\ell\}$, $V = [V_1, V_2, \ldots, V_s]$.

According to (4.2) $T^{-1} =: U = \{u_{ij}\}$ must satisfy

$$V_{g,\ell} \cdot X_\ell = U \cdot V_\ell = \{u_i^{(j-1)}(\lambda_\ell)/(j-1)!\}_{i,j}, \quad \ell = 1(1)s,$$

where $u_i$ denotes the polynomial $u_i(\lambda) = \sum_{j=1}^{k} u_{ij} \lambda^{j-1}$.

Since $X_\ell$ is triangular (cf. the proof of lemma 4.2), we find by comparing the elements in the first column of $V_{g,\ell} \cdot X_\ell$ and $U \cdot V_\ell$ that, for $i = 1(1)k$:

$$g(\lambda_\ell)^{i-1} \cdot x_{1,1} = u_i(\lambda_\ell), \text{ i.e. } u_i(\lambda_\ell) \cdot u_1(\lambda_\ell)^{i-1} - u_1(\lambda_\ell) \cdot u_2(\lambda_\ell)^{i-1} = 0$$

Since $u_i$, $i = 1(1)k$, are independent of $\lambda_\ell$, $\ell = 1(1)s$, it follows that $u_i \cdot u_1^{i-1} - u_1 \cdot u_2^{i-1} \equiv 0$, $i = 1(1)k$, and, in particular, since $u_1 \neq 0$ is necessary for $U$ to be non-singular,

$$u_k \equiv \left(\frac{v_2}{v_1}\right)^{k-1} \cdot u_1, \text{ where } v_i = u_i / \gcd(u_1, u_2), \; i = 1,2.$$

We see that $v_1^{k-1}$ must divide $u_1$ and thus deg $(v_1) \leq 1$. Likewise, $v_2$ is at most of degree 1 since deg $(u_k) \leq k-1$. Hence $u_2/u_1$ must be of the form

$$\frac{u_2(\lambda)}{u_1(\lambda)} = \frac{d\lambda - b}{-c\lambda + a},$$

where $ad \neq bc$ because $U$ must be non-singular.

If we now compare the first two <u>rows</u> of $V_{g,\ell} \cdot X_\ell$ with those of $U \cdot V_\ell$, we notice that $(X_\ell =: \{x_{ij}^{(\ell)}\})$:

$$x_{1j}^{(\ell)} = u_1^{(j-1)}(\lambda_\ell)/(j-1)! \text{ and } x_{2j}^{(\ell)} = u_2^{(j-1)}(\lambda_\ell)/(j-1)! - x_{1j}^{(\ell)} g(\lambda_\ell).$$

On the other hand, (4.1) with $\lambda_i = \lambda_j = \lambda_\ell$ tells us that

$$x_{2j}^{(\ell)} = \sum_{i=1}^{j-1} x_{1i}^{(\ell)} \cdot g^{(j-i)}(\lambda_\ell)/(j-i)!, \quad j = 1(1)m_\ell.$$

Combining these relations, one finds that (using the chain rule)

$$u_2^{(j-1)}(\lambda_\ell) = (u_1 \cdot g)^{(j-1)}(\lambda_\ell), \quad j = 1(1)m_\ell,$$

and

$$g(C) = u_1(C)^{-1} \cdot u_2(C) = (-c \cdot C + aI)^{-1} \cdot (dC - bI)$$

follows easily.

In order to show that

$$(-cC+aI) \cdot T \cdot C_g = (dC-bI) \cdot T \tag{4.5}$$

holds for the matrix T given in (4.4), apply the matrix on the right-hand-side of (4.5) to the vector $(1,r,\dots,r^{k-1})^T$. It is easily seen that this gives the vector $v=(v_1,v_2,\dots v_k)^T$, where

$$v_i = (ad-bc) \cdot r \cdot (cr+d)^{k-2} \cdot \left(\frac{ar+b}{cr+d}\right)^{i-1}, \quad i = 1(1)k-1,$$

$$v_k = (cr+d)^{k-1}\left[(ad-bc) \cdot \frac{r}{cr+d} \cdot \frac{ar+b}{cr+d}^{k-1} - \frac{d}{p_k} \cdot p\left(\frac{ar+b}{cr+d}\right)\right].$$

Here $p(\lambda) = \sum_{j=0}^{k} p_j \lambda^j$ denotes the polynomial $p_k \cdot (-1)^k \cdot \det(C-\lambda I)$.

Likewise, we see that

$$C_g \cdot (1,r,\dots,r^{k-1})^T = (r,r^2,\dots,r^{k-1},r^k - \frac{1}{\tilde{p}_k} \cdot (cr+d)^k \cdot p\left(\frac{ar+b}{cr+d}\right)),$$

where $\tilde{p}_k = \sum_{j=0}^{k} p_j a^j c^{k-j}$. It follows that $T \cdot C_g \cdot (1,r,\dots r^{k-1})^T =$

$w = (w_1, w_2, \dots w_k)^T$, where

$$w_i = (cr+d)^{k-1} \cdot \left[r \cdot \left(\frac{ar+b}{cr+d}\right)^{i-1} - c^{k-i} a^{i-1} (cr+d) \cdot p\left(\frac{ar+b}{cr+d}\right)/\tilde{p}_k\right].$$

Some elementary calculations show that $(-cC+aI)w=v$, and (4.5) is proven. The theorem now follows from (4.5) since $-c\cdot C+aI$ and $T$ are non-singular. $T^{-1}=:\{u_{ij}\}$ is obtained by inserting $\lambda=(d\mu-b)/(-c\mu+a)$ in (4.4):

$$\sum_{j=1}^{k} u_{ij}\mu^{j-1}=\left(\frac{-c\mu+a}{ad-bc}\right)^{k-1}\cdot\left(\frac{d\mu-b}{-c\mu+a}\right)^{i-1},\quad i=1(1)k,\ \forall\mu. \tag{4.6}$$

$\square$.

It follows from (4.5) that

$$T^{-1}CT=(aC_g+bI)(cC_g+dI)^{-1}.$$

In a Hilbert space $\|C_g\|\leqq 1$ will thus imply $\|T^{-1}CT\|\leqq 1$ if

$$\left|\frac{az+b}{cz+d}\right|\leqq 1 \text{ for all } |z|\leqq 1. \tag{4.7}$$

If a, b, c and d are real, (4.7) is equivalent to

$$|a+b|\leqq|c+d|,\quad |a-b|\leqq|c-d| \text{ and } |c|<|d|.$$

For a general operator norm $\|C_g\|\leqq 1$ will imply

$$(|d|-|c|)\cdot\|T^{-1}CT\|\leqq|a|+|b|,$$

and thus $\|T^{-1}CT\|\leqq 1$ if $|a|+|b|\leqq|d|-|c|$. It is interesting to find a similar result even in the non-linear case:

Theorem 4.4

For any integer $s\geqq 1$ let $Z_n^s$ denote the super-vector $(z_n^T, z_{n+1}^T,\dots,z_{n+s-1}^T)^T$. Furthermore, let a, b, c and d denote scalars satisfying

$$ad\neq bc,\quad |a|+|b|\leqq|d|-|c| \text{ and } c^k\cdot\rho_n(a/c)\cdot c^k\cdot\sigma_n(a/c)\neq 0,$$

where $(\rho_n,\sigma_n)$ is a k-step one-leg formula.

Assume that, for some function f and some sequences $\{\bar{h}_n\}$, $\{\bar{t}_n\}$, any solution of the k-step one-leg equation

$$\tilde{\rho}_n(E)x_n=\bar{h}_n\cdot f(\bar{t}_n,\tilde{\sigma}_n(E)x_n), \tag{4.8}$$

$$\tilde{\rho}_n(r)=(cr+d)^k\cdot\rho_n\left(\frac{ar+b}{cr+d}\right),$$

$$\tilde{\sigma}_n(r)=(cr+d)^k\cdot\sigma_n\left(\frac{ar+b}{cr+d}\right),$$

satisfies

$$\|X^k_{n+1}\|\leq\|X^k_n\|,$$

where $\|\cdot\|$ is some vector-norm. Then any solution of the one-leg equation

$$\rho_n(E)y_n=\bar{h}_n\cdot f(\bar{t}_n,\sigma_n(E)y_n), \tag{4.9}$$

will satisfy

$$\|(T^{-1}\otimes I)\cdot Y^k_{n+1}\|\leq\|(T^{-1}\otimes I)\cdot Y^k_n\|,$$

where $T=\{t_{ij}\}$ is given by (4.4) and $\otimes$ denotes the Kronecker product.

Proof:

For any integer $s\geq 1$ let $T_s=\{t^s_{ij}\}$ denote the matrix of dimension s given by

$$\sum_{j=1}^{s} t^s_{ij}\cdot\lambda^{j-1}=(c\lambda+d)^{s-1}\cdot\left(\frac{a\lambda+b}{c\lambda+d}\right)^{i-1}, \quad i=1(1)s, \ \forall\lambda .$$

From (4.6) we have that $U_s=\{u^s_{ij}\}=T_s^{-1}$ is given by

$$\sum_{j=1}^{s} u^s_{ij}\xi^{j-1}=\left(\frac{-c\xi+a}{ad-bc}\right)^{s-1}\left(\frac{d\xi-b}{-c\xi+a}\right)^{i-1}, \quad i=1(1)s, \ \forall\xi.$$

Let $u^s_i(\xi)$ denote the polynomials $\sum_{j=1}^{s} u^s_{ij}\xi^{j-1}$. Then

$$u_i^{s+1}(\xi)=\left(\frac{-c\xi+a}{ad-bc}\right)\cdot u_i^s(\xi), \quad i=1(1)s,$$

$$u_i^{s+1}(\xi)=\left(\frac{d\xi-b}{ad-bc}\right)\cdot u_{i-1}^s(\xi), i=2(1)s+1.$$

In other words $T_{s+1}^{-1}$ can be decomposed in two ways

$$(ad-bc)\cdot T_{s+1}^{-1}=\left[\frac{-c\cdot[0|T_s^{-1}]+a\cdot[T_s^{-1}|0]}{a_1a_2\ldots\ldots\ldots\ldots a_{s+1}}\right]$$
$$=\left[\frac{b_1b_2\ldots\ldots\ldots b_{s+1}}{d\cdot[0|T_s^{-1}]-b\cdot[T_s^{-1}|0]}\right]. \tag{4.10}$$

Let now $Y_n^{k+1}$ satisfy (4.9). Then $X_n^{k+1}:=(T_{k+1}^{-1}\otimes I)\cdot Y_n^{k+1}$ satisfies (4.8) and thus $\|X_{n+1}^k\|\leq\|X_n^k\|$.

Making use of the decompositions in (4.10), we find that

$$|d|\cdot\|(T^{-1}\otimes I)Y_{n+1}^k\|-|b|\cdot\|(T^{-1}\otimes I)Y_n^k\|\leq$$

$$\|d\cdot(T_k^{-1}\otimes I)Y_{n+1}^k-b\cdot(T_k^{-1}\otimes I)Y_n^k\|=|ad-bc|\cdot\|X_{n+1}^k\|\leq$$

$$|ad-bc|\cdot\|X_n^k\|=\|-c\cdot(T_k^{-1}\otimes I)Y_{n+1}^k+a\cdot(T_k^{-1}\otimes I)Y_n^k\|\leq$$

$$|c|\cdot\|(T^{-1}\otimes I)Y_{n+1}^k\|+|a|\cdot\|(T^{-1}\otimes I)Y_n^k\|,$$

and the proof follows, since $0<|a|+|b|\leq|d|-|c|$. □.

Remark

It seems rather essential that our coordinate transformation represented by the matrix $T_{k+1}^{-1}$ is such that $T_{k+1}^{-1}$ can be decomposed in two ways

$$T_{k+1}^{-1}=\left[\frac{-\tilde{c}\cdot[0|R]+\tilde{a}[R|0]}{\tilde{a}_1\tilde{a}_2\ldots\ldots\tilde{a}_{k+1}}\right]=\left[\frac{\tilde{b}_1\tilde{b}_2\ldots\ldots\tilde{b}_{k+1}}{\tilde{d}[0|R]-\tilde{b}[R|0]}\right],$$

where $R = \{r_{ij}\}$ is an arbitrary non-singular matrix. It is therefore interesting to note that - apart from a normalization factor - the particular matrix $T_{k+1}^{-1}$ used in theorem 4.4 is the only matrix with this property. This fact follows from the necessary identities ($r_i(\xi) := \sum_{j=1}^{k} r_{ij} \cdot \xi^{j-1}$):

$$(\tilde{d}\xi-\tilde{b})\cdot r_{i-1}(\xi) = (-\tilde{c}\xi+\tilde{a})\cdot r_i(\xi), \quad i = 2(1)k.$$

From these we obtain that R must be such that

$$r_i(\xi) = \left(\frac{\tilde{d}\xi-\tilde{b}}{-\tilde{c}\xi+\tilde{a}}\right)^{i-1} \cdot r_1(\xi) = \text{const}\cdot(-\tilde{c}\xi+\tilde{a})^{k-1}\cdot\left(\frac{\tilde{d}\xi-\tilde{b}}{-\tilde{c}\xi+\tilde{a}}\right)^{i-1} .$$

$\square$.

As in the proof of theorem 4.3, we may calculate

$$T^{-1}CT\cdot(1,\lambda,\ldots,\lambda^{k-1})^T =: (\mu_1,\mu_2,\ldots,\mu_k)^T.$$

We find that

$$\mu_i = \frac{1}{c\lambda+d}\left[(a\lambda+b)\cdot\lambda^{i-1} - \frac{(-c)^{k-i}d^{i-1}}{p_k\cdot(ad-bc)^{k-1}}\cdot(c\lambda+d)^k\cdot p\left(\frac{a\lambda+b}{c\lambda+d}\right)\right]. \tag{4.11}$$

We notice that if and only if $c\cdot d = 0$, the matrix $T^{-1}CT$ will contain only one row dependent on p. Since we here wish to consider the criterion $\|T^{-1}CT\|_\infty \leqq 1$, we thus see that this criterion is greatly simplified by choosing $c\cdot d = 0$. Since $d = 0$ makes it impossible to satisfy the conditions of theorem 4.4 we shall choose $c = 0$ and (without loss of generality) $d = 1$. The conditions of theorem 4.4 then read $a \neq 0$ and $|a|+|b| \leqq 1$. If the theorem, however, is to be applied to a consistent formula $(\rho_n,\sigma_n)$, the polynomial $\tilde{\rho}_n$ will have a root in $(1-b)/a$ and we therefore require $|1-b| \leqq |a|$ to be satisfied, too. These conditions only leave the possibility $|a| = 1-b$, $0 \leqq b < 1$. Since $\|T^{-1}CT\|_\infty$ and $\|C_g\|_\infty$ are independent of the orientation of a we choose $a = 1-b$ and obtain the (b,k)-norm:

Definition 4.5

A norm is said to be a (b,k)-norm if it for some $b \in [0,1)$ has the form

$$\|A\|_b = \|T_b^{-1} A T_b\|_\infty ,$$

where $T_b = \{t_{ij}\}$ is the triangular matrix of dimension k given by

$$t_{ij} = \binom{i-1}{j-1} b^{i-j} (1-b)^{j-1}, \quad 1 \leq j \leq i \leq k \qquad (4.12)$$

□.

It is known that for any two ordered triples $\{\lambda_1, \lambda_2, \lambda_3\}$ and $\{\mu_1, \mu_2, \mu_3\}$ of distinct elements in $\check{C}$ there is one (and only one) linear fractional transformation which maps $\lambda_i$ to $\mu_i$, $i = 1(1)3$. Although it is intuitively clear from this that we lose some "adaptability" of our norm by withholding only one parameter in the fractional transformation, we shall see that this choice simplifies the analysis very much without eliminating all "adaptability" of the norm.

In certain cases we may have to shift the value of b in the (b,k)-norm during the integration, but the effect of this is easily calculated.

Lemma 4.6

Assume that

$$\|(T_n^{-1} \otimes I) \cdot Y_{n+1}\| \leq \|(T_n^{-1} \otimes I) \cdot Y_n\|, \quad n = 0,1,\ldots,$$

where $T_n = \{t_{ij}^{(n)}\}$ is the $k \times k$ matrix determined by

$$\sum_{j=1}^{k} t_{ij}^{(n)} \cdot \lambda^{j-1} = (c_n\lambda + d_n)^{k-1} \cdot \left(\frac{a_n\lambda + b_n}{c_n\lambda + d_n}\right)^{i-1}, \quad i = 1(1)k, \ \forall \lambda,$$

where $a_n d_n \neq b_n c_n$, $\forall n$. Then

$$\|Y_{n+1}\| \leq \left(\prod_{s=0}^{n-1} \|(T_{s+1}^{-1} \cdot T_s) \otimes I\|\right) \cdot \|T_n \otimes I\| \cdot \|T_0^{-1} \otimes I\| \cdot \|Y_0\|,$$

and if $\|\cdot\|=\|\cdot\|_\infty$, $c_n \equiv 0$, $d_n \equiv 1$, $a_n \equiv 1-b_n$, $b_n \varepsilon [0,1)$

$$\|Y_{n+1}\|_\infty \leq \left[\prod_{s=0}^{n-1} (1+\varepsilon_s)\cdot\left(\frac{1+b_0}{1-b_0}\right)\right]^{k-1} \cdot \|Y_0\|_\infty ,$$

$$\varepsilon_s = \begin{cases} 0 & \text{if } b_{s+1} \leq b_s \\ 2\cdot \dfrac{b_{s+1}-b_s}{1-b_{s+1}} & \text{otherwise.} \end{cases}$$

<u>Proof</u>:

For n = 1,2,...

$$\|(T_n^{-1}\otimes I)\cdot Y_{n+1}\| \leq \|(T_n^{-1}\otimes I)(T_{n-1}\otimes I)(T_{n-1}^{-1}\otimes I)\cdot Y_n\|$$

$$\leq \|(T_n^{-1}\cdot T_{n-1})\otimes I\|\cdot\|(T_{n-1}^{-1}\otimes I)\cdot Y_n\|.$$

Hence

$$\|Y_{n+1}\| = \|(T_n\otimes I)(T_n^{-1}\otimes I)\cdot Y_{n+1}\| \leq \|(T_n \otimes I\|\cdot$$

$$(\prod_{s=0}^{n-1}\|(T_{s+1}^{-1}\cdot T_s)\otimes I\|)\cdot\|T_0^{-1}\otimes I\|\cdot\|Y_0\|.$$

From (4.4) and (4.6) we find that $T_{s+1}^{-1}\cdot T_s =: R_s = \{r_{ij}^{(s)}\}$ is determined by the requirements

$$\sum_{j=1}^{k} r_{ij}^{(s)}\lambda^{j-1} = \left[\frac{-c_{s+1}(a_s\lambda+b_s)+a_{s+1}(c_s\lambda+d_s)}{a_{s+1}d_{s+1}-b_{s+1}c_{s+1}}\right]^{k-1}\cdot$$

$$\left[\frac{d_{s+1}(a_s\lambda+b_s)-b_{s+1}(c_s\lambda+d_s)}{-c_{s+1}(a_s\lambda+b_s)+a_{s+1}(c_s\lambda+d_s)}\right]^{i-1}, \quad i=1(1)k,\ \forall\lambda.$$

If $c_n \equiv 0$, $d_n \equiv 1$, $a_n \equiv 1-b_n$, $b_n \varepsilon [0,1)$ we thus have that

$$\|(T_{s+1}^{-1}T_s)\otimes I\|_\infty = \|T_{s+1}^{-1}T_s\|_\infty = \left[1+\frac{(b_{s+1}-b_s)+|b_{s+1}-b_s|}{1-b_{s+1}}\right]^{k-1},$$

$$||T_n \otimes I||_\infty = ||T_n||_\infty = [\, |1-b_n| + |b_n|\, ]^{k-1} = 1,$$

and

$$||T_0^{-1} \otimes I||_\infty = ||T_0^{-1}||_\infty = [(1+b_0)/(1-b_0)]^{k-1}, \text{ q.e.d.} \qquad \Box.$$

For the sake of reference we state some of the properties of the (b,k)-norm:

Theorem 4.7

Let C be the companion matrix of a polynomial p of degree $k \geq 2$. If there exists a real number $b \in [0,1)$ so that

$$\sum_{j=0}^{k-2} \frac{(1-b)^j}{j!} |p^{(j)}(b)| + \frac{(1-b)^{k-1}}{(k-1)!} \cdot \left|p^{(k-1)}(b) - \frac{b}{k} \cdot p^{(k)}(b)\right|$$
$$\leq \frac{(1-b)^{k-1}}{k!} \cdot |p^{(k)}(b)|, \tag{4.13}$$

then $||C||_b = ||(1-b)C_b + b \cdot I||_\infty \leq 1$, where $C_b$ is the companion matrix of the polynomial $q(r) = p((1-b)r+b)$, and

$$||C^n||_\infty \leq ||T_b||_\infty \cdot ||T_b^{-1}||_\infty = ||T_b^{-1}||_\infty = \left(\frac{1+b}{1-b}\right)^{k-1}, \quad \forall\, n \geq 0.$$

where $T_b$ is given by (4.12).

Proof:

(4.13) follows from (4.11) because

$$\mu_k = (1-b) \cdot \lambda^k + b\lambda^{k-1} - p((1-b)\lambda + b)/(p_k \cdot (1-b)^{k-1})$$
$$= -(p_k \cdot (1-b)^{k-1})^{-1} \cdot \left\{ (1-b)^{k-1} \cdot \left[\frac{p^{(k-1)}(b)}{(k-1)!} - bp_k\right] \cdot \lambda^{k-1} + \right.$$
$$\left. + \sum_{j=0}^{k-2} \frac{(1-b)^j}{j!} p^{(j)}(b) \lambda^j \right\} .$$

From the proof of theorem 4.4 we see that $T_b^{-1}$ is the triangular matrix with the elements

$$u_{ij}=\binom{i-1}{j-1}(-b)^{i-j}/(1-b)^{i-1}, \quad 1\leq j\leq i\leq k,$$

and the theorem follows easily from the previous discussion. □.

Because of the similarity between $||C||_b \leq 1$, $||C_b||_\infty \leq 1$ and $||C||_\infty \leq 1$ we may utilize some of the results in [7] and [9] to derive some relevant criteria.

Lemma 4.8

Let $C(q)$ denote the companion matrix of the polynomial $p(r,q)=\rho(r)-q\cdot\sigma(r)$, where $\rho(r)=\sum_{j=0}^{k}\alpha_j r^j$, $\sigma(r)=\sum_{j=0}^{k}\beta_j r^j$ and $p(1,q)=-q$. Furthermore, let $C_b(q)$ denote the companion matrix of the polynomials $p_b(r,q)=p((1-b)\cdot r+b,q)$, $0\leq b<1$. Then

$||C(0)||_b \leq 1$ if and only if

$$\alpha_k\cdot\rho^{(j)}(b)\leq 0,\ j=0(1)k-2, \text{ and } b\leq -\alpha_{k-1}/[(k-1)\alpha_k]; \tag{4.14}$$

$||C_b(0)||_\infty \leq 1$ if and only if

$$\alpha_k\cdot\rho^{(j)}(b)\leq 0,\ j=0(1)k-1 \tag{4.15}$$

If $\alpha_k>0$, and (4.14) or (4.15), respectively, holds with strict inequalities, then the following is valid

$\forall\ \alpha\ \varepsilon\,[0,\pi/2)\,\exists a>0 : ||C(q)||_b \leq 1$ or $||C_b(q)||_\infty \leq 1$, respectively,

holds for all $q\,\varepsilon\{z\,||\arg(-z)|\leq\alpha$ and $\mathrm{Re}(-z)\leq a\}$.

Proof:

$||C(q)||_b \leq 1$ and $||C_b(q)||_\infty \leq 1$ are both equivalent to criteria of the form

$$\sum_{j=0}^{k-1}|a_j-q\cdot b_j| \leq |a_k-q\cdot b_k|\ , \tag{4.16}$$

where

$$\tilde{p}(r,q) \doteq \sum_{j=0}^{k} (a_j - q\cdot b_j) r^j = p_b(r,q) + b\cdot(\alpha_k - q\beta_k)\cdot(r-1)\cdot[(1-b)r]^{k-1}$$

or $p_b(r,q)$, respectively (cf. theorem 4.7). We note that $\tilde{p}(1,0)=0$, and thus $\|C(0)\|_b \leq 1$ and $\|C_b(0)\|_\infty \leq 1$ are equivalent to $a_k \cdot a_j \leq 0$ for $j = 0(1)k-1$, which gives (4.14) and (4.15).

Assume now that $\alpha_k > 0$ (i.e. $a_k > 0$) and that $a_k \cdot a_j < 0$ for $j = 0(1)k-1$. Let $\alpha \in [0,\pi/2)$ be given. We must then find $a>0$ so that (4.16) holds for all $q = c\cdot(-1+ib)$, where $0 \leq c \leq a$ and $|b| \leq tg\ (\alpha)$. But if $q = c\cdot(-1+ib)$ we have as a tends to zero,

$$|a_j - qb_j| = |a_j|\cdot(1+c\cdot b_j/a_j + O(c^2)),$$

and thus (4.16) reads

$$0 \leq (|a_k| - \sum_{j=0}^{k-1} |a_j|) + \operatorname{sgn}(a_k)\cdot c\cdot(b_k + \sum_{j=0}^{k-1} b_j) + O(c^2) = c + O(c^2).$$

$\square$.

Remark

The conditions for strong 0-contractivity in the (b,k)-norm (i.e. $\exists r > 0$: $\|C(q)\|_b \leq 1$ for all $q \in \{z \,\|\, z+r| \leq r\}$) have later been found by Nevanlinna and the writer ([10]) to be:

$$\alpha_k > 0, \tag{4.17a}$$

$$\rho^{(j)}(b) \leq 0 \text{ and } \rho^{(j)}(b) = 0 \Rightarrow \sigma^{(j)}(b) = 0 \quad \text{for } j = 0(1)k-2, \tag{4.17b}$$

$$b \leq -\alpha_{k-1}/[(k-1)\alpha_k] \text{ and } b = -\alpha_{k-1}/[(k-1)\alpha_k] \Rightarrow b = -\beta_{k-1}/[(k-1)\beta_k] \tag{4.17c}$$

$\square$.

Necessary and sufficient criteria for $\|C(q)\|_b \leq 1$ or $\|C_b(q)\|_\infty \leq 1$ to hold for all $q \in \check{C}^-$ are found by using techniques very similar to those applied in [9] concerning the case $b = 0$:

Lemma 4.9

With the notation of lemma 4.8, assume that $\alpha_k/\beta_k>0$. Then $\|C(q)\|_b \leqq 1$ holds for all $q \in \check{C}^-$ if and only if

1) $\|C(0)\|_b \leqq 1$,

2) $\rho^{(j)}(b)=0 \Rightarrow \sigma^{(j)}(b) = 0$ for $j = 0(1)k-2$, and

$b = -\alpha_{k-1}/[(k-1)\alpha_k] \Rightarrow b = -\beta_{k-1}/[(k-1)\beta_k]$, and

3)
$$\alpha_k \cdot \left[\sum_{j=0}^{k-2} \frac{(1-b)^j}{j!} \frac{(\sigma^{(j)}(b))^2}{\rho^{(j)}(b)} + (1-b)^{k-1} \frac{(\beta_k(k-1)b+\beta_{k-1})^2}{\alpha_k(k-1)b+\alpha_{k-1}} + (1-b)^{k-1} \frac{\beta_k^2}{\alpha_k}\right] \geqq 0 .$$

$\|C_b(q)\|_\infty \leqq 1$ holds for all $q \in \check{C}^-$ if and only if

1') $\|C_b(0)\|_\infty \leqq 1$,

2') $\rho^{(j)}(b) = 0 \Rightarrow \sigma^{(j)}(b) = 0$ for $j = 0(1)k-1$, and

3')
$$\alpha_k \cdot \sum_{j=0}^{k} \frac{(1-b)^j}{j!} \cdot \frac{(\sigma^{(j)}(b))^2}{\rho^{(j)}(b)} \geqq 0$$

In 3) and 3') terms with zero denominator are to be removed.

Proof:

As mentioned in the proof of lemma 4.8, $\|C(q)\|_b \leqq 1$ and $\|C_b(q)\|_\infty$ both have the form

$$\sum_{j=0}^{k-1} |a_j-qb_j| \leqq |a_k-qb_k|, \text{ where } \sum_{j=0}^{k} a_j=1-\sum_{j=0}^{k} b_j=0. \qquad (4.18)$$

Let $J$ be the index set $\{j\,|\,a_j \neq 0\}$. Then for $q = iy$ purely imaginary, we have that

$$|a_j-qb_j| = \begin{cases} |a_j|+\dfrac{b_j^2}{2|a_j|} \cdot y^2 + O(y^4), & \text{if } j \in J \\ |b_j| \cdot y, & \text{otherwise.} \end{cases}$$

From the proof of lemma 4.8 we know that 1) and 1') imply that $\sum_{j=0}^{k-1} |a_j| = |a_k|$, and on this assumption we see that (4.18) for $q = iy$ has the form

$$y \cdot \sum_{\substack{j=0 \\ j\notin J}}^{k-1} |b_j| + \frac{y^2}{2} \cdot \sum_{\substack{j=0 \\ j\in J}}^{k-1} \frac{b_j^2}{|a_j|} \leq \frac{y^2}{2} \cdot \frac{b_k^2}{|a_k|} + O(y^4).$$

Hence, all conditions in the lemma are recognized as necessary. That they are sufficient follows from theorem 2.1 in [7], since they imply that (4.18) holds for all $q = iy$ (including $\infty$):

$$\sum_{j=0}^{k-1} |a_j - iyb_j| \overset{2),2')}{=} \sum_{\substack{j=0 \\ j\in J}}^{k-1} |a_j|^{\frac{1}{2}} \cdot \left(|a_j| + y^2 \cdot \frac{b_j^2}{|a_j|}\right)^{\frac{1}{2}}$$

$$\leq \left(\sum_{j=0}^{k-1} |a_j|\right)^{\frac{1}{2}} \cdot \left[\sum_{\substack{j=0 \\ j\in J}}^{k-1} \left(|a_j| + y^2 \cdot \frac{b_j^2}{|a_j|}\right)\right]^{\frac{1}{2}}$$

$$\overset{1),1')}{=} \left[a_k^2 + |a_k| \cdot y^2 \cdot \sum_{\substack{j=0 \\ j\in J}}^{k-1} \frac{b_j^2}{|a_j|}\right]^{\frac{1}{2}} \overset{3),3')}{\leq} |a_k - iy \cdot b_k| .$$

□.

## 5. Certain Contractive Variable-Formula Methods

It is well-known that the order of a one-leg formula

$$\rho(E)y_n = \bar{h} \cdot f(\Theta t_{n+k} + (1-\Theta) \cdot t_{n+k-1}, \sigma(E)y_n),$$

is $p = \min\{\ell, m\}$, where $\ell$ and $m$ are the largest integers for which the following hold:

| | | |
|---|---|---|
| | $\bar{h}^{-1}\rho(E)y(t_n) = y'(\Theta t_{n+k} + (1-\Theta)t_{n+k-1})$ | is exact for all polynomials $y$ of degree $\leq \ell$ |
| and | $\sigma(E)y(t_n) = y(\Theta t_{n+k} + (1-\Theta)t_{n+k-1})$ | is exact for all polynomials $y$ of degree $\leq m-1$. |

It follows that a $k_n$-step one-leg formula is of order $k_n$ iff

$$\bar{h}^{-1}\rho(E)y_n = \sum_{j=k-k_n}^{k} \frac{\Phi_j'(\Theta t_{n+k}+(1-\Theta)t_{n+k-1})}{\Phi_j(t_{n+j})} \cdot y_{n+j}$$

and $$\sigma(E)y_n = \sum_{j=k-k_n}^{k} \frac{\Phi_j(\Theta t_{n+k}+(1-\Theta)t_{n+k-1})+c}{\Phi_j(t_{n+j})} \cdot y_{n+j} \,,$$

where c is independent of j and $\Phi_j(t)$ denotes the polynomial $\prod_{i=k-k_n}^{k}(t-t_{n+i})/(t-t_{n+j})$. We replace the parameter c by B satisfying $B\cdot\beta_k+\beta_{k-1}=0$ and obtain the following fixed-step formulas:

$\underline{k_n=1:}$

$$\rho(r)/r^{k-1} = r-1,$$

$$(1-B)\sigma(r)/r^{k-1} = r-B.$$

$\underline{k_n=2:}$

$$2\cdot\rho(r)/r^{k-2} = (2\Theta+1)r^2-4\Theta r+(2\Theta-1),$$

$$(2-B)\sigma(r)/r^{k-2} = [(\Theta+1)r^2-(\Theta-1)]-B\cdot[(\Theta+1)r-\Theta].$$

$\underline{k_n=3:}$

$$6\cdot\rho(r)/r^{k-3} = (3\Theta^2+6\Theta+2)r^3-(9\Theta^2+12\Theta-3)r^2+(9\Theta^2+6\Theta-6)r-(3\Theta^2-1),$$

$$2(3-B)\sigma(r)/r^{k-3} = [(\Theta^2+3\Theta+2)r^3-3(\Theta^2+\Theta-2)r+2(\Theta^2-1)]$$

$$-B\cdot[(\Theta^2+3\Theta+2)r^2-2(\Theta^2+2\Theta)r+(\Theta^2+\Theta)].$$

$\underline{k_n=4:}$

$$12\cdot\rho(r)/r^{k-4} = (2\Theta^3+9\Theta^2+11\Theta+3)r^4-(8\Theta^3+30\Theta^2+20\Theta-10)r^3+$$

$$+(12\Theta^3+36\Theta^2+6\Theta-18)r^2-(8\Theta^3+18\Theta^2-4\Theta-6)r+(2\Theta^3+3\Theta^2-\Theta-1),$$

$$6(4-B)\sigma(r)/r^{k-4} = \begin{bmatrix}(\Theta^3+6\Theta^2+11\Theta+6)r^4-6(\Theta^3+4\Theta^2+\Theta-6)r^2+ \\ +\ 8(\Theta^3+3\Theta^2-\Theta-3)r-3(\Theta^3+2\Theta^2-\Theta-2)\end{bmatrix} -$$

$$-B\cdot\begin{bmatrix}(\Theta^3+6\Theta^2+11\Theta+6)r^3-3(\Theta^3+5\Theta^2+6\Theta)r^2+ \\ 3(\Theta^3+4\Theta^2+3\Theta)r-(\Theta^3+3\Theta^2+2\Theta)\end{bmatrix}.$$

$\underline{k_n=5:}$

$$\begin{aligned}120\rho(r)/r^{k-5} &= (5\Theta^4+40\Theta^3+105\Theta^2+100\Theta+24)r^5-(25\Theta^4+180\Theta^3+ \\ &\quad +375\Theta^2+150\Theta-130)r^4+(50\Theta^4+320\Theta^3+510\Theta^2-40\Theta- \\ &\quad -240)r^3-(50\Theta^4+280\Theta^3+330\Theta^2-140\Theta-120)r^2+ \\ &\quad (25\Theta^4+120\Theta^3+105\Theta^2-60\Theta-40)r-(5\Theta^4+20\Theta^3+15\Theta^2-10\Theta-6),\end{aligned}$$

$$24(5-B)\sigma(r)/r^{k-5} = \begin{bmatrix}(\Theta^4+10\Theta^3+35\Theta^2+50\Theta+24)r^5-10(\Theta^4+8\Theta^3+ \\ +17\Theta^2-2\Theta-24)r^3+20\cdot(\Theta^4+7\Theta^3+11\Theta^2-7\Theta-12)r^2- \\ -15(\Theta^4+6\Theta^3+7\Theta^2-6\Theta-8)r+4(\Theta^4+5\Theta^3+5\Theta^2-5\Theta-6)\end{bmatrix}$$

$$-B\cdot\begin{bmatrix}(\Theta^4+10\Theta^3+35\Theta^2+50\Theta+24)r^4-4(\Theta^4+9\Theta^3+26\Theta^2+ \\ +24\Theta)r^3+6(\Theta^4+8\Theta^3+19\Theta^2+12\Theta)r^2-4(\Theta^4+7\Theta^3+ \\ +14\Theta^2+8\Theta)r+(\Theta^4+6\Theta^3+11\Theta^2+6\Theta)\end{bmatrix}.$$

The reason for replacing the parameter c by B is two-fold. Firstly, setting B to the same constant in several formulas, we may apply lemma 3.1 to the resulting OLM (1.3). Secondly, our analysis in [2] (pp.34-44) shows that in most cases (even for variable step-sizes) the $\infty$-contractive two- or three-step formulas with minimal $\Theta$-value satisfy

$$(k-1)b\,\beta_k + \beta_{k-1} = 0.$$

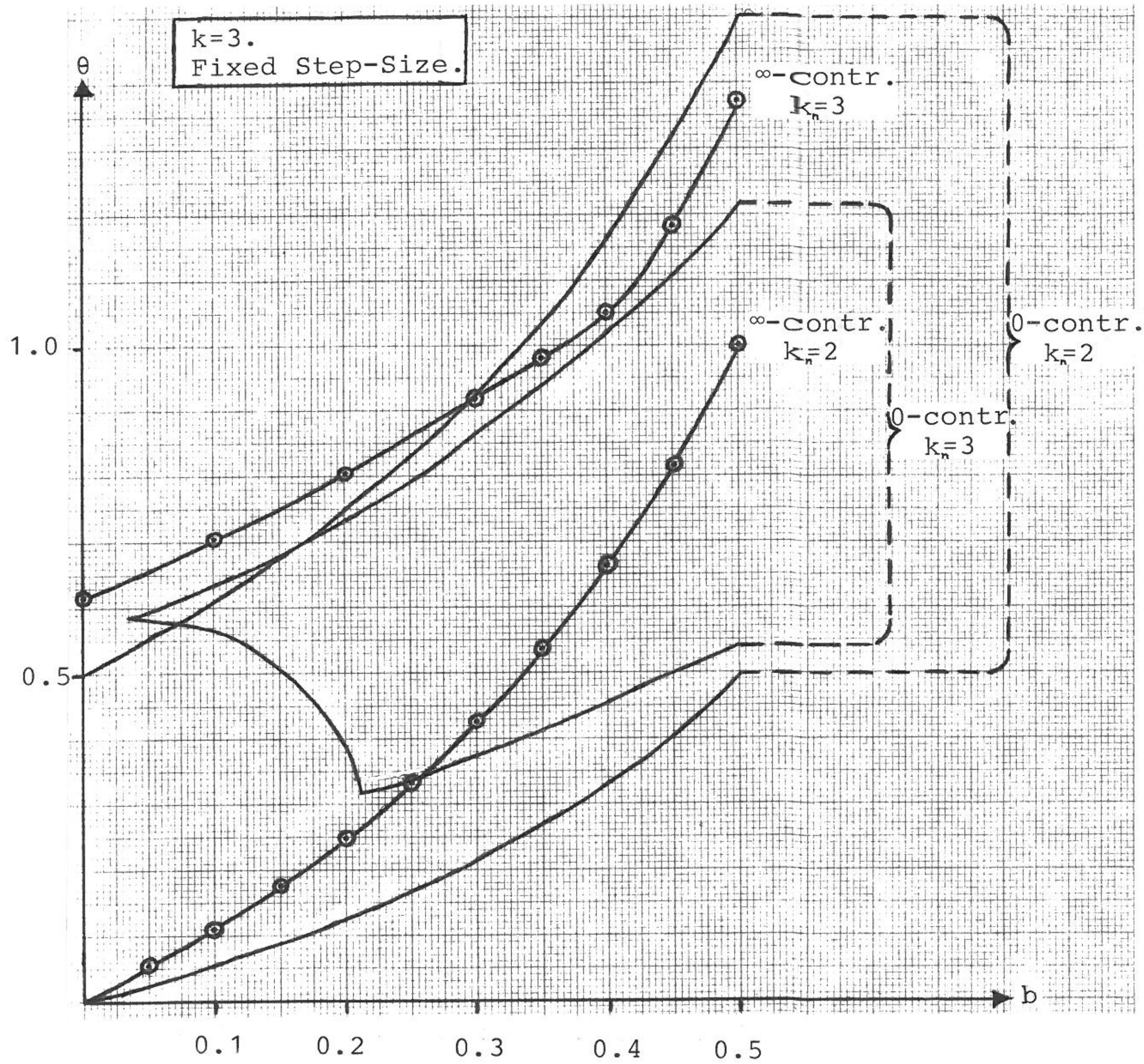
Since the $\Theta$-values for which a one-leg formula is 0-contractive w.r.t some (b,k)-norm usually are smaller than the

$\Theta$-values for which it becomes $\infty$-contractive, we shall here only consider OLM's with $B = (k-1)b$, $k \geq 3$ and refer the reader to [2] for a more thorough (but not that successful) treatment of one- and two-step formulas, $k = 2,3$.

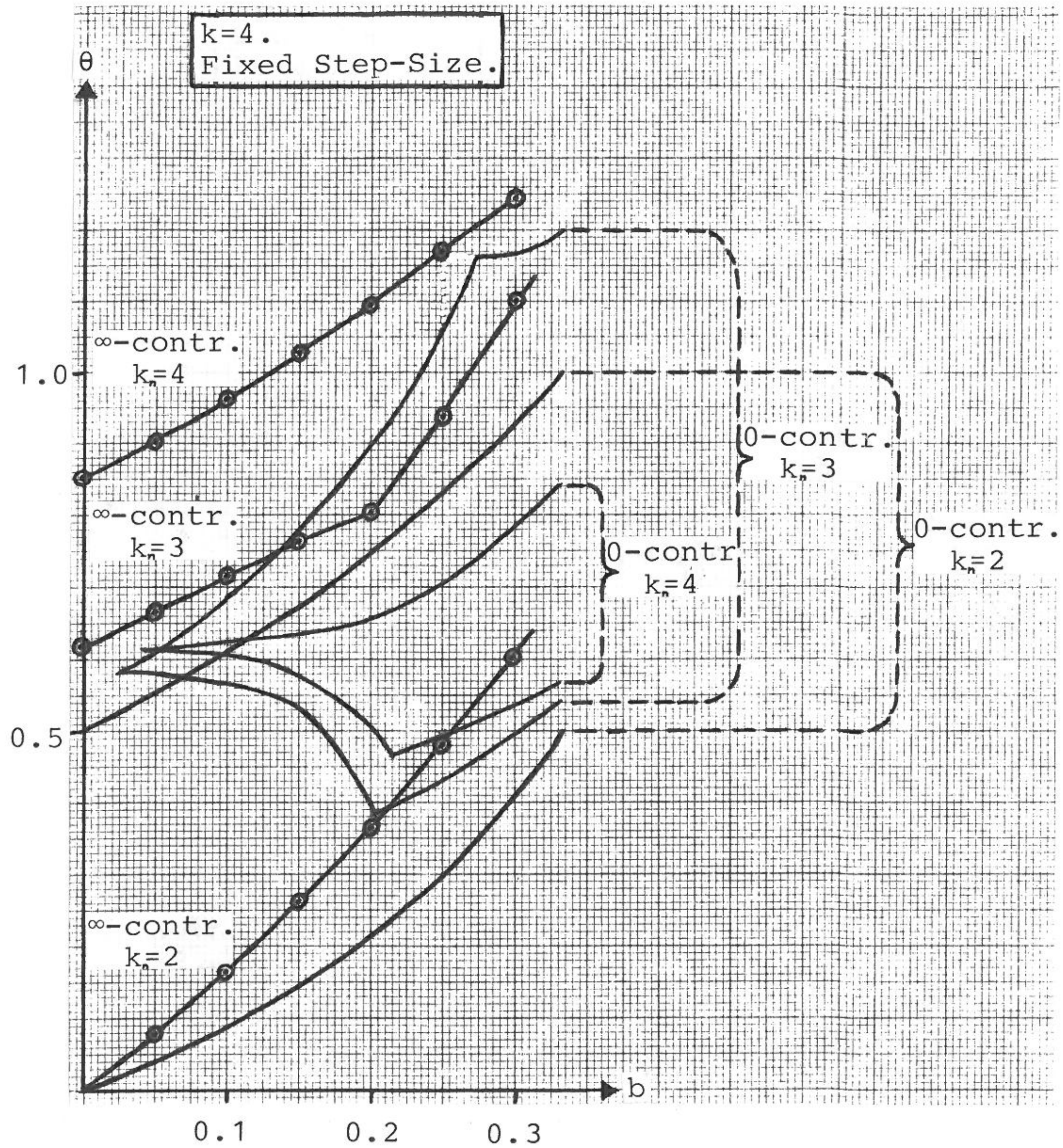
The following three figures illustrate the set of $(b,\Theta)$ for which the fixed-step k'th order formulas with $B = (k-1)b$ are 0-contractive in the $(b,k)$-norm. Moreover, the minimal $\Theta$-value for which the formulas are $\infty$-contractive in the $(b,k)$-norm is displayed. Below each figure the non-differentiable behaviour of some of the curves is explained.

We observe that for $k = 3$, no 3-step 3'rd order formula is $A_0$-contractive in a $(b,k)$-norm with $b \leq 1/(k-1)$. Fortunately, this is not the case for $k = 4$ or 5 and we find a second- and a third-order formula being $A(\alpha)$-contractive in the same $(b,k)$-norm and with a reasonably large $\alpha$ (approx. $89^0$ and $45^0$, respectively), cf. fig. 5.1 and fig. 5.2.

To construct a 4-step OLM we also need a one-step formula and from lemma 4.9 it is easily seen that choosing $\beta_k=1.1$ (say) will make the formula A-contractive in the $(0.2,4)$-norm. Finally, we may add a four-step formula although they seem to have a relatively small region of contractivity in the $(0.2,4)$-norm, cf. fig. 5.3.



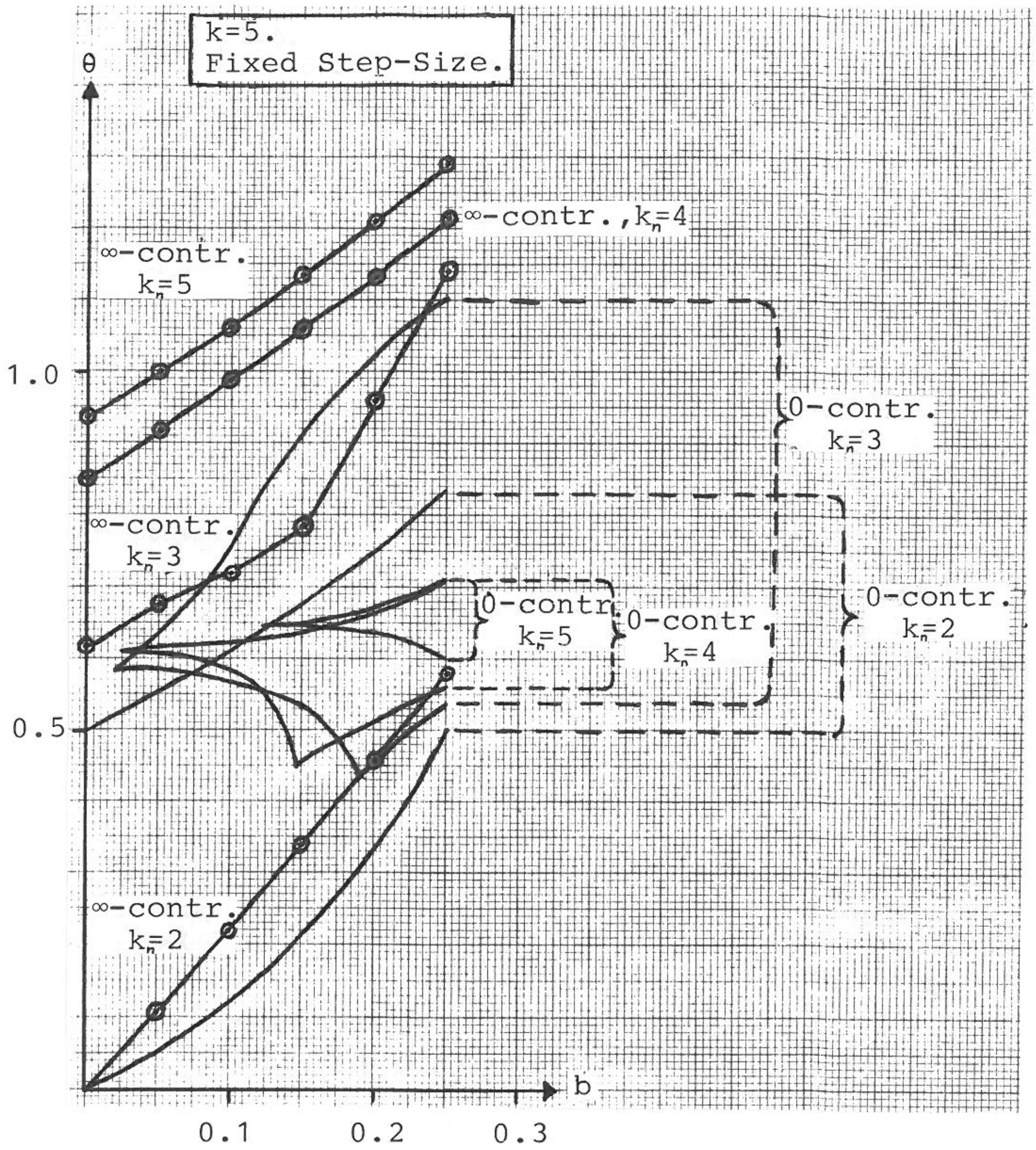

$k_n=3$: $\rho(b) = 2\alpha_3 b + \alpha_2 = 0$ for $\Theta \approx 0.3193995$ and $b \approx 0.2073370$.




$k_n$=3: $\rho(b) = 3\alpha_3 b + \alpha_2 = 0$ for $\theta \approx 0.3797959$ and $b \approx 0.2020410$.

$\rho'(b) = \rho''(b) = 0$ for $\theta \approx 1.1542916$ and $b \approx 0.2725283$.

$k_n$=4: $\rho'(b) = 3\alpha_4 b + \alpha_3 = 0$ for $\theta \approx 0.4648768$ and $b \approx 0.2139326$.




$k_n$=3: $\rho(b) = 4\alpha_3 b + \alpha_2 = 0$ for $\theta \approx 0.4363371$ and $b \approx 0.1902780$.

$k_n$=4: $\rho^{(2)}(b) = 4\alpha_4 b + \alpha_3 = 0$ for $\theta \approx 0.4489212$ and $b \approx 0.1446741$.

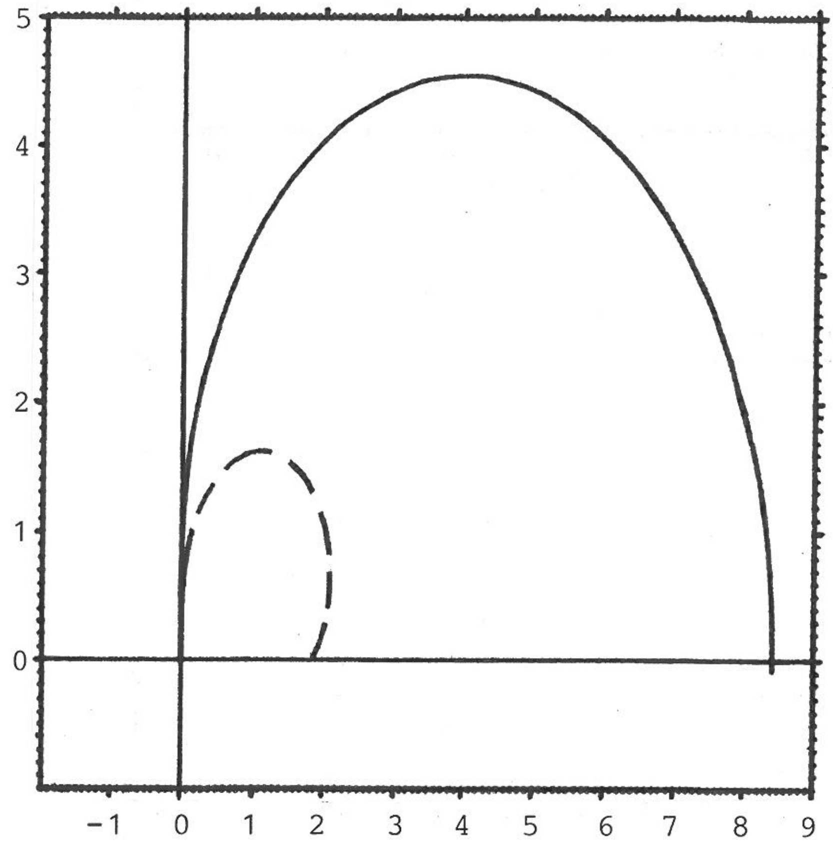



Fig. 5.1. The region of absolute stability (dashed line) and the region of contractivity in the (0.2,4)-norm of the second-order 2-step one-leg formula with $\Theta$=0.54, B=0.6.

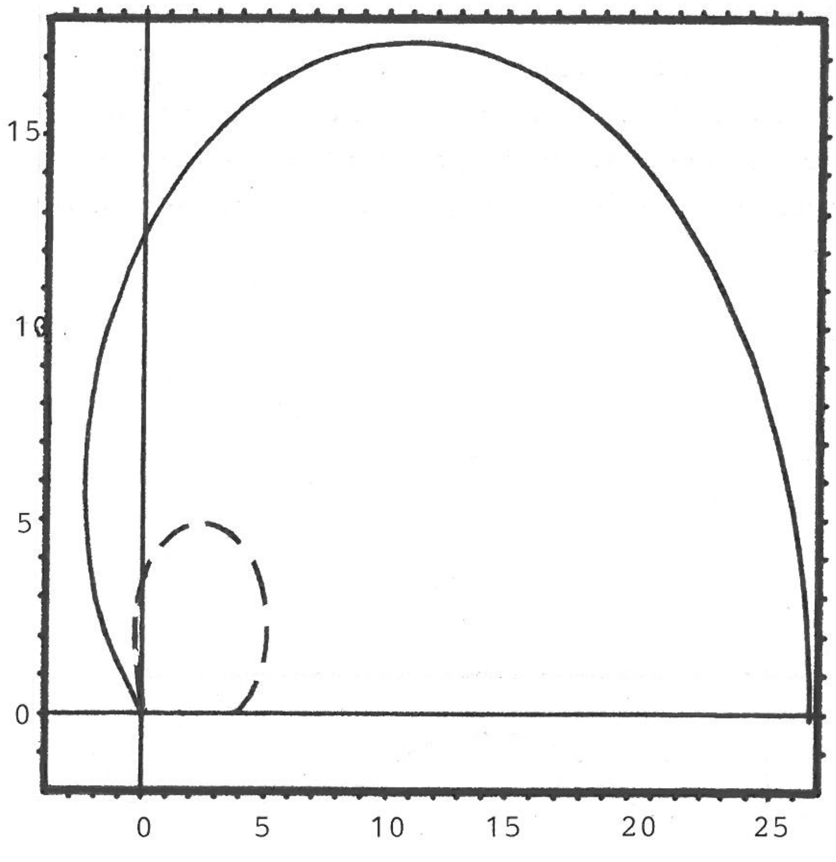



Fig. 5.2. The region of absolute stability (dashed line) and the region of contractivity in the (0.2,4)-norm of the third-order 3-step one-leg formula with $\Theta=\sqrt{5.53}-1.5$, B= 0.6.

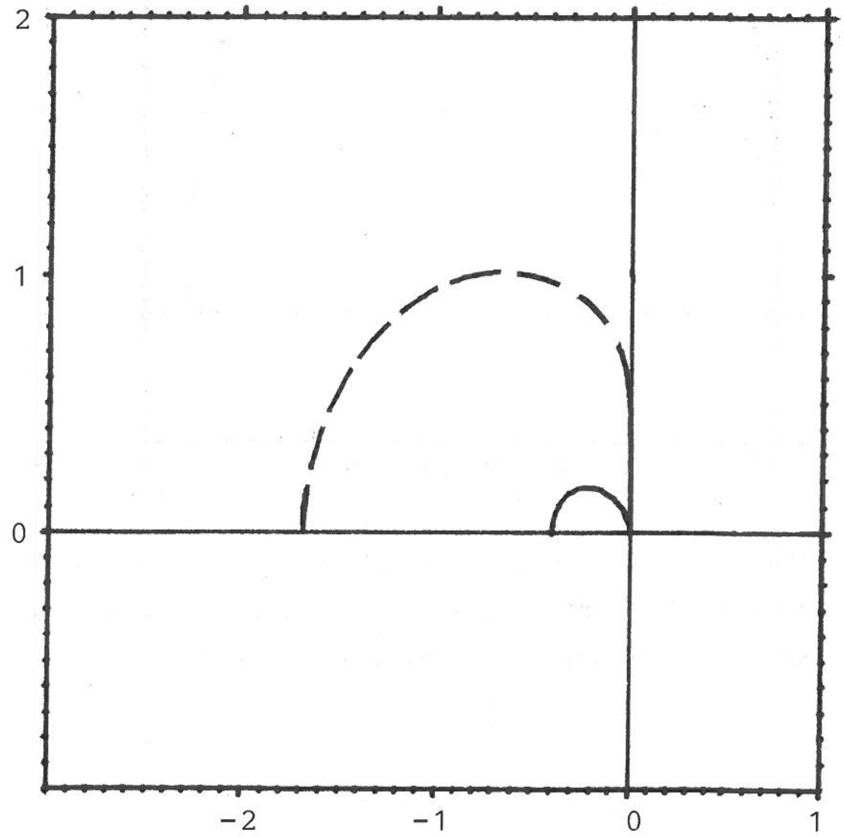



Fig. 5.3. The region of absolute stability (dashed line) and the region of contractivity in the (0.2,4)-norm of the fourth-order 4-step one-leg formula with $\Theta$=B=0.6.

Remark

The (b,2)-norm has turned out to be an appropriate measure of strong 0-contractivity of 2-step OLM's ([10]) and we hope to publish this result soon. □.

Acknowledgement

The thesis was written under the supervision of Professor G. Dahlquist at the Royal Institute of Technology, Stockholm, and I am very grateful to him as well as to the Danish Natural Science Research Council, which supported me financially during most of my stay at the institute.

References

1. J. Sand, On One-Leg and Linear Multistep Formulas with Variable Step-Sizes, Report TRITA-NA-8112, Dept. of Num. Anal. and Comp. Sci., The Royal Inst. of Techn., Stockholm (1981).

2. J. Sand, On Contractive One-Leg Methods, Report TRITA-NA-8218, Dept. of Num. Anal. and Comp. Sci., The Royal Inst. of Techn., Stockholm (1982).

3. J. Sand, Stability and Boundedness Results for Variable-Step Variable-Formula Methods, Report TRITA-NA-8219, Dept. of Num. Anal. and Comp. Sci., The Royal Inst. of Techn., Stockholm (1982).

4. R. K. Brayton and C. H. Tong, Stability of Dynamical Systems: a Constructive Approach, IBM Research Report RC-7027, IBM Watson Research Center, Yorktown Heights (1978).

5. R. K. Brayton and C. C. Conley, Some Results on the Stability and Instability of the Backward Differentiation Methods with Non-Uniform Time Steps, IBM Research Report RC-3964, IBM Watson Research Center, Yorktown Heights (1972).

6. G. Dahlquist, Convergence and Stability in the Numerical Integration of Ordinary Differential Equations, Math. Scand. 4 (1956), 33-53.

7. O. Nevanlinna and W. Liniger, Contractive Methods for Stiff Differential Equations, Part I, BIT 18 (1978), 457-474.

8. A. S. Householder, The Theory of Matrices in Numerical Analysis, Blaisdell Publishing Company, New York (1965).

9. W. Liniger, A-Contractivity of Second-Order Multistep Formulas with Variable Steps, To appear.

10. O. Nevanlinna and J. Sand, (not yet published).