Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

MATHEMATICAL CENTRE TRACTS 144

THE SOLUTION OF INITIAL VALUE PROBLEMS USING INTERVAL ARITHMETIC

FORMULATION AND ANALYSIS OF AN ALGORITHM

P. EIJGENRAAM

MATHEMATISCH CENTRUM AMSTERDAM 1981

1980 Mathematics subject classification: 65L05, 65G10, 65-04

ISBN 90 6196 230 7

ACKNOWLEDGEMENTS

First of all I want to express my gratitude to prof.dr. M.N. Spijker for his many critical and stimulating remarks.

Further I thank prof.dr. G. Alefeld and prof.dr. L.A. Peletier for their useful comments.

Finally, I am grateful to the Mathematical Centre for the opportunity to publish this monograph in their series Mathematical Centre Tracts, and all those at the Mathematical Centre who have contributed to its technical realization.

CONTENTS

CHAPTER 1

INTRODUCTION

In this monograph initial value problems are considered for systems of M ordinary differential equations, where $M \geq 1$. We deal with the problem of enclosing the solution numerically for arbitrary values of the independent variable t.

We allow that the initial value is not exactly known, but is only known to be contained in a given initial value set. In that case we want to enclose the corresponding set of solutions for arbitrary t.

In enclosing the solution or set of solutions we take rigorously into account all possible sources of errors, including rounding errors due to the finite precision of a computer. For this purpose use is made of rounded-interval arithmetic (see MOORE [1966]).

We will formulate and analyse a numerical method for solving the aforementioned problem. Other methods of solving the problem have been treated by MOORE [1966], KRÜCKEBERG [1969], HUNGER [1971], MARCOWITZ [1973, 1975], CONRADT [1980], STERN [1980] and others.

All these methods produce for certain grid-points $t_0 < t_1 < \ldots < t_N$ a set $\overline{y}_n$ enclosing the set of solutions at the grid-point t_n. For $n = 1,2,\ldots, N$ the set $\overline{y}_n$ is computed in such a way that it contains the value at $t = t_n$ of any solution of the differential system whose value at $t = t_{n-1}$ belongs to $\overline{y}_{n-1}$.

In the methods of Hunger and Marcowitz, and, at least for non-linear differential systems, in the methods of Conradt and Stern, these sets $\overline{y}_n$ are chosen so as to be vector intervals, i.e., M-dimensional blocks with their edges parallel to the coordinate axes. Moore has shown that in this case the diameter of the sets $\overline{y}_n$ may grow considerably faster, as n increases, than is inherent in the nature of the differential system. Therefore he and Krückeberg use sets $\overline{y}_n$ that are linear transformations of vector intervals. We will do the same in our method.

Hunger's method first computes an approximate solution and then solves a linear interval differential system for the error function (see also BAUCH [1977], which contains a survey of related methods).

The method of Marcowitz also starts with an approximate solution. However, in order to bound the error function a system of non-linear differential inequalities is solved. The number of inequalities of this system is twice the number of differential equations of the original system.

Conradt considers a variant of the method of Marcowitz. Furthermore, for linear differential systems only, Conradt discusses a method in which the sets $\overline{y}_n$ are polyhedra. This method is based on an idea, independently found by LOHNER & ADAMS [1978] and NICKEL [1979].

Stern works out theoretical details of the method of Marcowitz. Further he considers, for linear differential systems with constant coefficients only, a method in which the sets $\overline{y}_n$ are ellipsoids.

Krückeberg's method computes in each step an inclusion of the solution with an arbitrary fixed initial value $\hat{y}_{n-1} \in \overline{y}_{n-1}$ and then computes an inclusion of the perturbation of the solution due to a variation of the initial value $y \in \overline{y}_{n-1}$.

Moore's method is based on consecutive Taylor series expansions of the solutions at the points $t_0, t_1, \ldots, t_{N-1}$.

In this monograph we will deal with a method based on the same principle as Moore's method. However, we will work this principle out in a different way. Firstly, a step size $h_n = t_n - t_{n-1}$ is computed such that it approximates a prescibed value H_n. This improves the possibility of controlling the step size. Secondly, the set $\overline{y}_n$ enclosing the set of solutions, is not, as in Moore's method, obtained by transforming the differential system into a more complicated form, but it is computed in a more direct way.

In chapter 2 we will present definitions, notations and properties, mostly of interval analytic concepts. In section 2.1 concepts related to the ordering of $\mathbb{R}^M$ and $\mathbb{R}^{M,M}$ will be considered, especially the vector interval, the matrix interval and the rounding operator. Section 2.2 deals with operations on sets of vectors and matrices, in particular on vector intervals and matrix intervals. In section 2.3 the concepts of norm, diameter and distance are introduced and a number of properties are derived. Section 2.4 treats some notations and properties related to the initial value problem we are considering. Finally, in section 2.5 we consider the remaining definitions and properties we want to treat in this chapter.

Chapter 3 gives an outline of the numerical method we treat in this monograph. For certain grid-points $t_0 < t_1 < \ldots < t_N$ this method produces a set $\overline{y}_n$ enclosing the set of solutions at the grid-point t_n. This set $\overline{y}_n$ is a linear transformation of a vector interval. The n'th step of the method consists of the determination of a suitable new grid-point t_n and the set $\overline{y}_n$, for a given t_{n-1} and a given set $\overline{y}_{n-1}$. This n'th step consists of two parts, treated in chapters 4 and 5, respectively.

In chapter 4 the first part of the n'th step of the method is considered. After the introductory section 4.1 this part is described in section 4.2 by Algorithm I. The algorithm determines a suitable step size h_n and thus the new grid-point $t_n = t_{n-1} + h_n$. This is necessary since the step size cannot be prescribed arbitrarily. Further the algorithm computes a rough inclusion $\overline{b}_n$ of the set of solutions U(t) for $t_{n-1} \leq t \leq t_n$. We show that the algorithm, which contains an iteration process, is finite. In section 4.3 we prove that the vector interval $\overline{b}_n$ produced by Algorithm I is indeed an inclusion as required. Section 4.4 deals with the analysis of the step size h_n. In particular we consider how close this value is to the prescribed parameter H_n, which can be considered as the step size the algorithm aims at. In section 4.5 we analyse some variants of Algorithm I and compare them with the version of Algorithm I described in section 4.2. In particular we show that a variant suggested by MOORE [1966] is in general not a finite algorithm.

Chapter 5 treats the second part of the n'th step of the method. After an introductory section 5.1 this part is described in section 5.2 by Algorithm II. Using the inclusion $\overline{b}_n$ produced by Algorithm I, Algorithm II determines a set $\overline{y}_n$ enclosing the set of solutions $U(t_n)$. This set $\overline{y}_n$ is a linearly transformed vector interval, represented by a non-singular transformation matrix A_n and a vector interval $\overline{x}_n$. In section 5.3 we prove that the set $\overline{y}_n$ indeed encloses the set of solutions at the grid-point t_n. In section 5.4 the concept of local error is defined using the Hausdorff distance. Further we derive an estimate for the local error. Section 5.5 compares the method with other methods. We show that the transformation matrix used by MOORE [1966] can cause the local error to be essentially greater than in our method. Further we show that the same can hold for the local error of the method of KRÜCKEBERG [1969].

In chapter 6 the global behaviour of the method is studied. In section 6.1 the concept of global error is defined. Further a theorem is proved giving conditions under which the method is applicable on a prescribed interval [0,T], and giving a bound on the global error in terms of the

diameter of the initial value set, the maximal step size and the value of t. Section 6.2 deals with a further analysis of the global error and clarifies the bound obtained in section 6.1 by considering a limit case. In section 6.3 we consider the global error in the case where the set $\overline{y}_n$ is required to be a vector interval. HUNGER [1971] and MARCOWITZ [1973, 1975] give methods for which this limitation on $\overline{y}_n$ holds. An example of MOORE [1966], showing that in this case the global error can grow unfavourably, is analysed in more detail. In section 6.4 we show that the error term of order δ^2, where δ is the diameter of the initial value set, in the bound on the global error is unavoidable. Section 6.5 describes a method to vary the step size and the order of the method for each step, so as to satisfy a prescribed condition on the local error per unit step as efficiently as possible. Finally, section 6.6 discusses the effect of rounding errors on the global error.

Chapter 7 gives and explains a computer program for the method treated in this monograph. The program is written partly in Algol 60 and partly in Triplex-Algol 60, described in WIPPERMANN [1968].

Numerical results obtained with this program are given in chapter 8.

CHAPTER 2

PRELIMINARIES

2.1. ORDERING

2.1.1. Vector intervals and matrix intervals

In this monograph $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^M$ the set of real M-dimensional vectors and $\mathbb{R}^{M,M}$ the set of real M×M matrices. We identify $\mathbb{R}^1$ with $\mathbb{R}$.

For a vector $x \in \mathbb{R}^M$ the i'th component is denoted by $[x]_i$ $(1 \leq i \leq M)$. We write $[A]_{ij}$ for the element in the i'th row and j'th column of a matrix $A \in \mathbb{R}^{M,M}$ $(1 \leq i \leq M,\ 1 \leq j \leq M)$.

Symbols with a bar, like $\overline{x}$ and $\overline{A}$, always denote sets.

For a set $\overline{x} \subset \mathbb{R}^M$ we define

$$p_i(\overline{x}) = \{[x]_i \mid x \in \overline{x}\} \qquad (1 \leq i \leq M). \tag{2.1.1}$$

Similarly for a set $\overline{A} \subset \mathbb{R}^{M,M}$ we define

$$p_{ij}(\overline{A}) = \{[A]_{ij} \mid A \in \overline{A}\} \qquad (1 \leq i \leq M,\ 1 \leq j \leq M). \tag{2.1.2}$$

We define the relation "≤" on $\mathbb{R}^M$ by

$$a \leq b \iff [a]_i \leq [b]_i \quad (1 \leq i \leq M) \tag{2.1.3}$$

and on $\mathbb{R}^{M,M}$ by

$$A \leq B \iff [A]_{ij} \leq [B]_{ij} \quad (1 \leq i \leq M,\ 1 \leq j \leq M). \tag{2.1.4}$$

As can easily be verified this relation "≤" on V, with $V = \mathbb{R}^M$ or $V = \mathbb{R}^{M,M}$, is a partial ordering, i.e., it satisfies for all $a,b,c \in V$

$$(2.1.5)\qquad \begin{cases} a \le a & \text{(reflexivity)},\\ (a \le b,\ b \le c) \Rightarrow a \le c & \text{(transitivity)},\\ (a \le b,\ b \le a) \Rightarrow a = b & \text{(anti-symmetry)}. \end{cases}$$

Let $(V,\le)$ be a partially ordered set. For all $a,b \in V$ with $a \le b$ we define the *interval* $[a,b]$ by

$$(2.1.6)\qquad [a,b] = \{x \mid x \in V,\ a \le x \le b\}.$$

In accordance with the notation of KULISCH [1976] the set of all intervals in V is denoted by $\mathbb{I}V$. The elements of $\mathbb{I}\mathbb{R}$, $\mathbb{I}\mathbb{R}^M$ and $\mathbb{I}\mathbb{R}^{M,M}$ are called *real intervals*, *vector intervals* and *matrix intervals*, respectively. Elements of $\mathbb{I}\mathbb{R}^2$ can be interpreted geometrically as rectangles with their sides parallel to the coordinate axes. Similarly, elements of $\mathbb{I}\mathbb{R}^3$ can be interpreted as rectangular parallelepipeds, with their edges parallel to the coordinate axes.

We define for $\overline{\xi}_1,\ldots,\overline{\xi}_M \in \mathbb{I}\mathbb{R}$

$$(2.1.7)\qquad \begin{pmatrix} \overline{\xi}_1 \\ \vdots \\ \overline{\xi}_M \end{pmatrix} = \{x \mid x \in \mathbb{R}^M,\quad [x]_i \in \overline{\xi}_i\ (1 \le i \le M)\}$$

and for $\overline{\xi}_{ij} \in \mathbb{I}\mathbb{R}$ $\quad (1 \le i \le M,\ 1 \le j \le M)$

$$(2.1.8)\qquad \begin{pmatrix} \overline{\xi}_{11} & \cdots & \overline{\xi}_{1M} \\ \vdots & \cdots & \vdots \\ \overline{\xi}_{M1} & \cdots & \overline{\xi}_{MM} \end{pmatrix} = \{A \mid A \in \mathbb{R}^{M,M},\ [A]_{ij} \in \overline{\xi}_{ij}\ (1 \le i \le M,\ 1 \le j \le M)\}.$$

If we write a vector some of whose components are real intervals and others are real numbers then each real component ξ_i should be interpreted as the set $\{\xi_i\}$, in other words, as the interval $[\xi_i,\xi_i]$. The notation of a matrix whose elements are partly real intervals and partly real numbers should be interpreted correspondingly.

Using these notations we can formulate the following elementary properties.

$$(2.1.9)\qquad \begin{pmatrix}[\alpha_1,\beta_1]\\ \vdots \\ [\alpha_M,\beta_M]\end{pmatrix} = \left[\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_M\end{pmatrix}, \begin{pmatrix}\beta_1\\ \vdots\\ \beta_M\end{pmatrix}\right] \in \mathbb{II}\,\mathbb{R}^M$$

(for $\alpha_i,\beta_i \in \mathbb{R}$, $\alpha_i \le \beta_i$ $(1 \le i \le M)$),

$$(2.1.10)\qquad \begin{pmatrix}[\alpha_{11},\beta_{11}]\cdots[\alpha_{1M},\beta_{1M}]\\ \vdots \quad\cdots\quad \vdots \\ [\alpha_{M1},\beta_{M1}]\cdots[\alpha_{MM},\beta_{MM}]\end{pmatrix} = \left[\begin{pmatrix}\alpha_{11}\cdots\alpha_{1M}\\ \vdots\ \cdots\ \vdots\\ \alpha_{M1}\cdots\alpha_{MM}\end{pmatrix}, \begin{pmatrix}\beta_{11}\cdots\beta_{1M}\\ \vdots\ \cdots\ \vdots\\ \beta_{M1}\cdots\beta_{MM}\end{pmatrix}\right] \in \mathbb{II}\,\mathbb{R}^{M,M}$$

(for $\alpha_{ij},\beta_{ij} \in \mathbb{R}$, $\alpha_{ij} \le \beta_{ij}$ $(1 \le i \le M,\ 1 \le j \le M)$),

$$(2.1.11)\qquad \overline{x} = \begin{pmatrix}p_1(\overline{x})\\ \vdots\\ p_M(\overline{x})\end{pmatrix} \qquad (\text{for } \overline{x} \in \mathbb{II}\,\mathbb{R}^M),$$

$$(2.1.12)\qquad \overline{A} = \begin{pmatrix}p_{11}(\overline{A}) & \cdots & p_{1M}(\overline{A})\\ \vdots & \cdots & \vdots\\ p_{M1}(\overline{A}) & \cdots & p_{MM}(\overline{A})\end{pmatrix} \qquad (\text{for } \overline{A} \in \mathbb{II}\,\mathbb{R}^{M,M}),$$

$$(2.1.13)\qquad p_i([a,b]) = [[a]_i,[b]_i]$$

(for all $a,b \in \mathbb{R}^M$ with $a \le b$, and $1 \le i \le M$),

$$(2.1.14)\qquad p_{ij}([A,B]) = [[A]_{ij},[B]_{ij}]$$

(for all $A,B \in \mathbb{R}^{M,M}$ with $A \le B$, and $1 \le i \le M$, $1 \le j \le M$).

By virtue of (2.1.11) we find that a vector interval $\overline{x}$ is uniquely defined by the specification of the sets $p_i(\overline{x})$ $(1 \le i \le M)$. In view of (2.1.13) these sets are real intervals. A similar remark can be made for a matrix interval, due to (2.1.12) and (2.1.14).

REMARK 2.1.1. Let an *interval vector* be a vector whose components are real intervals. From (2.1.9) we see that a vector interval can be characterized by a corresponding interval vector. Several authors in the field of interval

arithmetic, for instance ALEFELD & HERZBERGER [1974], HANSEN [1965], HUNGER [1971], KRÜCKEBERG [1969] and MOORE [1966], start with the concept of interval vector. Some of them observe that an interval vector can be identified with a set of vectors.

However, although we will use notation (2.1.7) whenever this is convenient, we prefer to use the concept of vector interval, defined as a set of vectors of a special kind, rather than the concept of interval vector. It enables us to apply basic set-theoretic concepts like intersection and inclusion to vector intervals without specially defining them. Further motives of our choice will be given in section 2.2, where the distinction between vector intervals and interval vectors becomes more important.

Similar considerations apply to the concepts of matrix interval and *interval matrix*, respectively. □

With respect to the intersection and inclusion we easily obtain

$$(2.1.15)\quad \begin{cases} \overline{x} \cap \overline{y} \in \amalg \mathbb{R}^M, \\ p_i(\overline{x} \cap \overline{y}) = p_i(\overline{x}) \cap p_i(\overline{y}) \quad (1 \le i \le M) \end{cases}$$

(for all $\overline{x},\overline{y} \in \amalg \mathbb{R}^M$ with $\overline{x} \cap \overline{y} \neq \emptyset$),

$$(2.1.16)\quad \begin{cases} \overline{A} \cap \overline{B} \in \amalg \mathbb{R}^{M,M}, \\ p_{ij}(\overline{A} \cap \overline{B}) = p_{ij}(\overline{A}) \cap p_{ij}(\overline{B}) \quad (1 \le i \le M,\ 1 \le j \le M) \end{cases}$$

(for all $\overline{A},\overline{B} \in \amalg \mathbb{R}^{M,M}$ with $\overline{A} \cap \overline{B} \neq \emptyset$),

$$(2.1.17)\quad \overline{x} \subset \overline{y} \iff p_i(\overline{x}) \subset p_i(\overline{y}) \quad (1 \le i \le M)$$

(for $\overline{x},\overline{y} \in \amalg \mathbb{R}^M$),

$$(2.1.18)\quad \overline{A} \subset \overline{B} \iff p_{ij}(\overline{A}) \subset p_{ij}(\overline{B}) \quad (1 \le i \le M,\ 1 \le j \le M)$$

(for $\overline{A},\overline{B} \in \amalg \mathbb{R}^{M,M}$),

$$(2.1.19)\quad [a,b] \subset [c,d] \iff a \ge c,\ b \le d$$

(for all $a,b,c,d \in V$ with $a \le b$ and $c \le d$, where $V = \mathbb{R}^M$ or $V = \mathbb{R}^{M,M}$).

Rules (2.1.15) and (2.1.16), and, implicitly, rules (2.1.17) and (2.1.18) can be found in ALEFELD & HERZBERGER [1974].

2.1.2. The minimum, maximum, infimum and supremum

Let V be either $\mathbb{R}^M$ or $\mathbb{R}^{M,M}$, $\overline{x} \subset V$, $\overline{x} \neq \emptyset$.

$y \in V$ is said to be a *lower bound* of $\overline{x}$ if $y \leq x$ for all $x \in \overline{x}$, and an *upper bound* of $\overline{x}$ if $y \geq x$ for all $x \in \overline{x}$.

The *minimum* of $\overline{x}$, if it exists, is the $y \in \overline{x}$, which is a lower bound of $\overline{x}$, and it is denoted by "min $\overline{x}$". The *maximum* of $\overline{x}$, if it exists, is the $y \in \overline{x}$ which is an upper bound of $\overline{x}$, and it is denoted by "max $\overline{x}$". We observe that the concepts of minimum and maximum should not be confused with those of minimal and maximal element, respectively (defined for instance in KULISCH [1976]).

Obviously an interval in V has a minimum and a maximum and we have

$$(2.1.20) \quad a = \min [a,b], \quad b = \max [a,b]$$

(for all $a,b \in V$ with $a \leq b$),

$$(2.1.21) \quad \overline{x} = [\min \overline{x}, \max \overline{x}] \quad (\text{for } \overline{x} \in \mathbb{I}V).$$

Using (2.1.13), (2.1.14) and (2.1.20) we immediately obtain

$$(2.1.22) \quad \begin{cases} [\min \overline{x}]_i = \min p_i(\overline{x}), \\ [\max \overline{x}]_i = \max p_i(\overline{x}) \end{cases}$$

(for $\overline{x} \in \mathbb{I}\mathbb{R}^M$ and $1 \leq i \leq M$),

$$(2.1.23) \quad \begin{cases} [\min \overline{A}]_{ij} = \min p_{ij}(\overline{A}), \\ [\max \overline{A}]_{ij} = \max p_{ij}(\overline{A}) \end{cases}$$

(for $\overline{A} \in \mathbb{I}\mathbb{R}^{M,M}$ and $1 \leq i \leq M$, $1 \leq j \leq M$).

Let $\overline{x} \subset V$ be bounded and non-empty. We define the *infimum* of $\overline{x}$, denoted by "inf $\overline{x}$", if it exists, by

$$(2.1.24) \quad \inf \overline{x} = \max \{y \mid y \in V, y \text{ is a lower bound of } \overline{x}\},$$

and we define the *supremum* of $\overline{x}$, denoted by "sup $\overline{x}$", if it exists, by

(2.1.25) $\quad \sup \overline{x} = \min \{y \mid y \in V, y \text{ is an upper bound of } \overline{x}\}.$

If $\min \overline{x}$ exists then $\inf \overline{x}$ also exists and $\inf \overline{x} = \min \overline{x}$. Similarly, if $\max \overline{x}$ exists then $\sup \overline{x}$ exists and $\sup \overline{x} = \max \overline{x}$.

THEOREM 2.1.2. *For bounded and non-empty* $\overline{x} \subset \mathbb{R}^M$, $\inf \overline{x}$ *and* $\sup \overline{x}$ *exist and satisfy*

$$(2.1.26) \quad [\inf \overline{x}]_i = \inf p_i(\overline{x}) \qquad (1 \le i \le M),$$

$$(2.1.27) \quad [\sup \overline{x}]_i = \sup p_i(\overline{x}) \qquad (1 \le i \le M).$$

PROOF. For $1 \le i \le M$ the sets $p_i(\overline{x})$ are bounded and non-empty subsets of $\mathbb{R}$ and therefore have a finite infimum. Consider the vector

$$z = \begin{pmatrix} \inf p_1(\overline{x}) \\ \vdots \\ \inf p_M(\overline{x}) \end{pmatrix}.$$

First we show that z is a lower bound of $\overline{x}$. Let $x \in \overline{x}$. For $1 \le i \le M$ we have $[x]_i \in p_i(\overline{x})$ and hence $[x]_i \ge [z]_i$. Therefore $x \ge z$ for all $x \in \overline{x}$, which was to be shown.

Now assume that y is an arbitrary lower bound of $\overline{x}$. Let $1 \le i \le M$. For all $x \in \overline{x}$ we have $[y]_i \le [x]_i$, hence $[y]_i$ is a lower bound of $p_i(\overline{x})$ and therefore $[y]_i \le [z]_i$. Thus we have $y \le z$. Consequently z is the maximum of the set of lower bounds of $\overline{x}$, that is, $z = \inf x$. This proves (2.1.26).

The second part of the theorem can be proved analogously. □

THEOREM 2.1.3. *For bounded and non-empty* $\overline{A} \subset \mathbb{R}^{M,M}$, $\inf \overline{A}$ *and* $\sup \overline{A}$ *exist and satisfy*

$$(2.1.28) \quad [\inf \overline{A}]_{ij} = \inf p_{ij}(\overline{A}) \qquad (1 \le i \le M,\ 1 \le j \le M),$$

$$(2.1.29) \quad [\sup \overline{A}]_{ij} = \sup p_{ij}(\overline{A}) \qquad (1 \le i \le M,\ 1 \le j \le M).$$

PROOF. The proof is analogous to that of theorem 2.1.2. □

Finally we define for $\overline{x} \in \amalg \mathbb{R}^M$ and for $\overline{x} \in \amalg \mathbb{R}^{M,M}$

(2.1.30) $\quad \text{mean } \overline{x} = \frac{1}{2}(\min \overline{x} + \max \overline{x})$.

2.1.3. The rounding operator

Let $\overline{x} \subset V$ be bounded and non-empty, where $V = \mathbb{R}^M$ or $V = \mathbb{R}^{M,M}$. We define the *rounding operator* □ by

(2.1.31) $\quad \square \overline{x} = [\inf \overline{x}, \sup \overline{x}]$.

For $V = \mathbb{R}^2$ this is illustrated in figure 2.1.1.

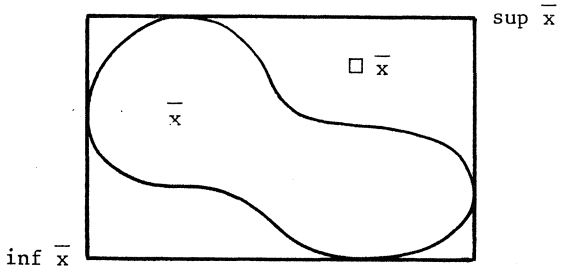

fig. 2.1.1

The following theorem gives some fundamental properties of the operator □ and can essentially be found in KULISCH [1976], p. 399.

THEOREM 2.1.4. *For* $V = \mathbb{R}^M$ *or* $V = \mathbb{R}^{M,M}$ *we have*

(2.1.32) $\quad \overline{x} \subset \square \overline{x}$ $\qquad$ *(for all bounded, non-empty* $\overline{x} \subset V$*)*,

(2.1.33) $\overline{x} \subset \overline{y} \Rightarrow \square\, \overline{x} \subset \overline{y}$

(*for all bounded, non-empty* $\overline{x} \subset V$ *and all* $\overline{y} \in \amalg V$),

(2.1.34) $\square\, \overline{x} = \bigcap_{\substack{\overline{y} \in \amalg V \\ \overline{y} \supset \overline{x}}} \overline{y}$ (*for all bounded, non-empty* $\overline{x} \subset V$)

(2.1.35) $\square\, \overline{x} = \overline{x}$ (*for* $\overline{x} \in \amalg V$),

(2.1.36) $\overline{x} \subset \overline{y} \Rightarrow \square\, \overline{x} \subset \square\, \overline{y}$ (*for all bounded, non-empty* $\overline{x}, \overline{y} \subset V$).

PROOF. Let $\overline{x} \subset V$ be bounded and non-empty.

(2.1.32) follows immediately from the fact that $\inf \overline{x}$ and $\sup \overline{x}$ are a lower bound and an upper bound, respectively, of the set $\overline{x}$.

Let $\overline{y} \supset \overline{x}$, $\overline{y} \in \amalg \mathbb{R}^M$. $\min \overline{y}$ is a lower bound of $\overline{y}$ and therefore of $\overline{x}$, hence $\min \overline{y} \leq \inf \overline{x}$. Similarly $\max \overline{y} \geq \sup \overline{x}$. Using (2.1.21), (2.1.19) and (2.1.31) we obtain

$$\overline{y} = [\min \overline{y},\ \max \overline{y}] \supset [\inf \overline{x},\ \sup \overline{x}] = \square\, \overline{x}$$

and (2.1.33) has been proved.

(2.1.34) immediately follows from (2.1.32) and (2.1.33).

For $\overline{x} \in \amalg V$ we have

$$\square\, \overline{x} = [\inf \overline{x},\ \sup \overline{x}] = [\min \overline{x},\ \max \overline{x}] = \overline{x},$$

and (2.1.35) has been proved.

Let $\overline{x} \subset \overline{y} \subset V$ with $\overline{x}$ and $\overline{y}$ bounded and non-empty. Using (2.1.32) we obtain $\overline{x} \subset \square\, \overline{y}$. Applying (2.1.33), with $\square\, \overline{y}$ substituted for $\overline{y}$, yields (2.1.36). □

By virtue of (2.1.32), (2.1.35) and (2.1.36) the operator $\square$ is, in the terminology of KULISCH [1976], a monotonous, upwards directed rounding with respect to the inclusion relation. Indeed, it is convenient to think of $\square\, \overline{x}$ as the result of *rounding* the set $\overline{x}$ to an interval, comparable with the rounding of a real number to a number representable in a computer. In fact $\square\, \overline{x}$ is the smallest interval containing the set $\overline{x}$, as is shown by (2.1.32) and (2.1.33).

THEOREM 2.1.5.

$$(2.1.37)\qquad \Box\,\overline{x} = \begin{pmatrix} \Box\, p_1(\overline{x}) \\ \vdots \\ \Box\, p_M(\overline{x}) \end{pmatrix}$$

(*for all bounded, non-empty* $\overline{x} \subset \mathbb{R}^M$),

$$(2.1.38)\qquad \Box\,\overline{A} = \begin{pmatrix} \Box\, p_{11}(\overline{A}) & \dots & \Box\, p_{1M}(\overline{A}) \\ \vdots & \dots & \vdots \\ \Box\, p_{M1}(\overline{A}) & \dots & \Box\, p_{MM}(\overline{A}) \end{pmatrix}$$

(*for all bounded, non-empty* $\overline{A} \subset \mathbb{R}^{M,M}$).

PROOF. Using (2.1.26) and (2.1.27) we find

$$\Box\,\overline{x} = \{\, y \mid y \in \mathbb{R}^M,\ \inf \overline{x} \le y \le \sup \overline{x}\}$$

$$= \{\, y \mid y \in \mathbb{R}^M,\ [\inf \overline{x}]_i \le [y]_i \le [\sup \overline{x}]_i\ (1 \le i \le M)\}$$

$$= \{\, y \mid y \in \mathbb{R}^M,\ \inf p_i(\overline{x}) \le [y]_i \le \sup p_i(\overline{x})\ (1 \le i \le M)\}$$

$$= \{\, y \mid y \in \mathbb{R}^M,\ [y]_i \in \Box\, p_i(\overline{x})\ (1 \le i \le M)\}$$

$$= \begin{pmatrix} \Box\, p_1(\overline{x}) \\ \vdots \\ \Box\, p_M(\overline{x}) \end{pmatrix}.$$

Using (2.1.28) and (2.1.29), (2.1.38) can be proved analogously. □

2.2. OPERATIONS ON SETS

2.2.1. General

For a function $g : X \to Y$ we define

$$g(\overline{x}) = \{g(x) \mid x \in \overline{x}\} \qquad (\text{for } \overline{x} \subset X). \tag{2.2.1}$$

This is a special case of the "united extension", used by MOORE [1966]. Since no confusion can arise we use the same symbol g in the expressions $g(\overline{x})$ and $g(x)$.

Note that, if we define for $x \in \mathbb{R}^M$, $1 \leq i \leq M$,

$$p_i(x) = [x]_i$$

and for $A \in \mathbb{R}^{M,M}$, $1 \leq i \leq M$, $1 \leq j \leq M$,

$$p_{ij}(A) = [A]_{ij},$$

then definitions (2.1.1) and (2.1.2) can be considered as special cases of definition (2.2.1).

For a function $h : Q \to Z$ with $Q \subset X * Y$ (where $*$ denotes the direct product) we define

$$h(\overline{x},\overline{y}) = \{h(x,y) \mid x \in \overline{x},\ y \in \overline{y}\} \quad (\text{for } \overline{x} * \overline{y} \subset Q), \tag{2.2.2}$$

$$h(\overline{x},y) = h(\overline{x},\{y\}) \quad (\text{for } \overline{x} * \{y\} \subset Q), \tag{2.2.3}$$

$$h(x,\overline{y}) = h(\{x\},\overline{y}) \quad (\text{for } \{x\} * \overline{y} \subset Q), \tag{2.2.4}$$

where, for any u, $\{u\}$ denotes of course the set with u as its only element.

Similarly, for a binary operation $\circledast : Q \to Z$ where $Q \subset X * Y$ we define

$$\overline{x} \circledast \overline{y} = \{x \circledast y \mid x \in \overline{x},\ y \in \overline{y}\} \quad (\text{for } \overline{x} * \overline{y} \subset Q), \tag{2.2.5}$$

$$\overline{x} \circledast y = \overline{x} \circledast \{y\} \quad (\text{for } \overline{x} * \{y\} \subset Q), \tag{2.2.6}$$

$$x \circledast \overline{y} = \{x\} \circledast \overline{y} \quad (\text{for } \{x\} * \overline{y} \subset Q). \tag{2.2.7}$$

In particular these definitions apply to the operations

$$(2.2.8)\qquad \begin{cases} +: V * V \to V & (\text{for } V = \mathbb{R}^M,\ \mathbb{R}^{M,M}), \\ -: V * V \to V & (\text{for } V = \mathbb{R}^M,\ \mathbb{R}^{M,M}), \\ \cdot: V * W \to W & (\text{for } V = \mathbb{R},\ \mathbb{R}^{M,M} \text{ and } W = \mathbb{R}^M,\ \mathbb{R}^{M,M}), \\ /: \mathbb{R} * (\mathbb{R} \setminus \{0\}) \to \mathbb{R}. \end{cases}$$

From definitions (2.2.1), (2.2.2) and (2.2.5) we immediately obtain the following fundamental properties:

$$(2.2.9)\qquad \overline{x}_1 \subset \overline{x}_2 \subset X \Rightarrow g(\overline{x}_1) \subset g(\overline{x}_2),$$

$$(2.2.10)\qquad \overline{x}_1 \subset \overline{x}_2,\ \overline{y}_1 \subset \overline{y}_2,\ \overline{x}_2 * \overline{y}_2 \subset Q \Rightarrow h(\overline{x}_1,\overline{y}_1) \subset h(\overline{x}_2,\overline{y}_2),$$

$$(2.2.11)\qquad \overline{x}_1 \subset \overline{x}_2,\ \overline{y}_1 \subset \overline{y}_2,\ \overline{x}_2 * \overline{y}_2 \subset Q \Rightarrow \overline{x}_1 \circledast \overline{y}_1 \subset \overline{x}_2 \circledast \overline{y}_2.$$

Similarly, for $V = \mathbb{R}^M,\ \mathbb{R}^{M,M}$ we define

$$(2.2.12)\qquad -\overline{x} = \{-x \mid x \in \overline{x}\} \qquad (\text{for } \overline{x} \subset V)$$

and obtain

$$(2.2.13)\qquad \overline{x}_1 \subset \overline{x}_2 \subset V \Rightarrow -\overline{x}_1 \subset -\overline{x}_2.$$

Finally, for an arbitrary integer $k \geq 1$ we define, similar to (2.2.1),

$$(2.2.14)\qquad \overline{\xi}^k = \{\xi^k \mid \xi \in \overline{\xi}\} \qquad (\text{for } \overline{\xi} \subset \mathbb{R})$$

and obtain

$$(2.2.15)\qquad \overline{\xi}_1 \subset \overline{\xi}_2 \subset \mathbb{R} \Rightarrow \overline{\xi}_1^k \subset \overline{\xi}_2^k.$$

Note that in general we do not have $\overline{\xi}^2 = \overline{\xi} \cdot \overline{\xi}$, etc. For instance, $[-1,1]^2 = [0,1]$, while $[-1,1] \cdot [-1,1] = [-1,1]$.

For $V = \mathbb{R}^M,\ \mathbb{R}^{M,M}$ and $\overline{x},\overline{y} \subset V$ we have

$$(2.2.16)\qquad \overline{x} - \overline{y} = \overline{x} + (-1).\overline{y},$$

(2.2.17) $-\overline{x} = (-1) \cdot \overline{x}$.

Using (2.2.16), from many properties of the addition, corresponding properties of the subtraction can easily be derived. We will not always state the latter.

In theorems 2.2.1, 2.2.2 and 2.2.3 we will treat some properties of operations on sets, as defined by (2.2.5).

THEOREM 2.2.1. *The following commutativity properties hold.*

(2.2.18) $\overline{x} + \overline{y} = \overline{y} + \overline{x}$ (*for* $\overline{x},\overline{y} \subset \mathbb{R}^M$),

(2.2.19) $\overline{A} + \overline{B} = \overline{B} + \overline{A}$ (*for* $\overline{A},\overline{B} \subset \mathbb{R}^{M,M}$),

(2.2.20) $\overline{\xi}\,\overline{\eta} = \overline{\eta}\,\overline{\xi}$ (*for* $\overline{\xi},\overline{\eta} \subset \mathbb{R}$).

PROOF. For $\overline{x},\overline{y} \subset \mathbb{R}^M$ we have

$$\begin{aligned}\overline{x} + \overline{y} &= \{x + y \mid x \in \overline{x},\ y \in \overline{y}\} \\ &= \{y + x \mid x \in \overline{x},\ y \in \overline{y}\} \\ &= \overline{y} + \overline{x}.\end{aligned}$$

The other properties are proved analogously. □

THEOREM 2.2.2. *For* $\overline{\lambda},\overline{\mu} \subset \mathbb{R}$, $\overline{x},\overline{y},\overline{z} \subset \mathbb{R}^M$, $\overline{A},\overline{B},\overline{C} \subset \mathbb{R}^{M,M}$ *we have*

(2.2.21) $(\overline{x} + \overline{y}) + \overline{z} = \overline{x} + (\overline{y} + \overline{z})$,

(2.2.22) $(\overline{A} + \overline{B}) + \overline{C} = \overline{A} + (\overline{B} + \overline{C})$,

(2.2.23) $(\overline{\lambda}\,\overline{\mu})\,\overline{x} = \overline{\lambda}\,(\overline{\mu}\,\overline{x}) = \overline{\mu}(\overline{\lambda}\,\overline{x})$,

(2.2.24) $(\overline{\lambda}\,\overline{\mu})\,\overline{A} = \overline{\lambda}\,(\overline{\mu}\,\overline{A}) = \overline{\mu}(\overline{\lambda}\,\overline{A})$,

(2.2.25) $(\overline{\lambda}\,\overline{A})\,\overline{x} = \overline{\lambda}(\overline{A}\,\overline{x}) = \overline{A}(\overline{\lambda}\,\overline{x})$,

(2.2.26) $(\overline{\lambda}\,\overline{A})\,\overline{B} = \overline{\lambda}(\overline{A}\,\overline{B}) = \overline{A}(\overline{\lambda}\,\overline{B})$,

(2.2.27) $(\overline{A}\,\overline{B})\,\overline{x} = \overline{A}(\overline{B}\,\overline{x})$,

(2.2.28) $(\overline{A}\,\overline{B})\,\overline{C} = \overline{A}(\overline{B}\,\overline{C})$.

PROOF. All properties are proved analogously. For instance

$$\begin{aligned}(\overline{\lambda}\,\overline{A})\,\overline{x} &= \{(\lambda A)\,x \mid \lambda \in \overline{\lambda},\ A \in \overline{A},\ x \in \overline{x}\}\\ &= \{\lambda(A\,x) \mid \lambda \in \overline{\lambda},\ A \in \overline{A},\ x \in \overline{x}\}\\ &= \overline{\lambda}(\overline{A}\,\overline{x})\,.\end{aligned}$$ □

By virtue of theorem 2.2.2 expressions like $\overline{x} + \overline{y} + \overline{z}$, $\overline{A} + \overline{B} + \overline{C}$, $\overline{\lambda}\,\overline{\mu}\,\overline{x}$, etc., without brackets, are not ambiguous. The same holds for more complicated expressions, such as $\overline{\lambda}\,\overline{\mu}\,\overline{A}\,\overline{B}\,\overline{x}$.

Let $V = \mathbb{R}^M$ or $V = \mathbb{R}^{M,M}$. For $\overline{x}_1,\overline{x}_2,\ldots,\overline{x}_k \subset V$ the expression $\pm\,\overline{x}_1 \pm \overline{x}_2 \pm \ldots \pm \overline{x}_k$ should be evaluated from left to right. Equivalently, it should be interpreted as the sum $(\pm\,\overline{x}_1) + (\pm\,\overline{x}_2) + \ldots + (\pm\,\overline{x}_k)$, where, of course, $+\,\overline{x}_\ell = \overline{x}_\ell$ $(1 \le \ell \le k)$.

Further we define for an arbitrary integer $k \ge 1$

(2.2.29) $$\sum_{\ell=1}^{k} \overline{x}_\ell = \overline{x}_1 + \overline{x}_2 + \ldots + \overline{x}_k \quad (\text{for } \overline{x}_1,\overline{x}_2,\ldots,\overline{x}_k \subset V).$$

THEOREM 2.2.3. *For* $V = \mathbb{R}$, $\mathbb{R}^{M,M}$ *and* $X = \mathbb{R}^M$, $\mathbb{R}^{M,M}$ *we have the distributivity properties*

(2.2.30) $v(\overline{x} + \overline{y}) = v\overline{x} + v\overline{y}$ (*for* $v \in V$ *and* $\overline{x},\overline{y} \subset X$),

(2.2.31) $(\overline{v} + \overline{w})x = \overline{v}x + \overline{w}x$ (*for* $\overline{v},\overline{w} \subset V$ *and* $x \in X$)

and the subdistributivity properties

(2.2.32) $\overline{v}(\overline{x} + \overline{y}) \subset \overline{v}\,\overline{x} + \overline{v}\,\overline{y}$ (*for* $\overline{v} \subset V$ *and* $\overline{x},\overline{y} \subset X$),

(2.2.33) $(\overline{v} + \overline{w})\overline{x} \subset \overline{v}\,\overline{x} + \overline{w}\,\overline{x}$ (*for* $\overline{v},\overline{w} \subset V$ *and* $\overline{x} \subset X$).

PROOF. For $v \in V$ and $\overline{x},\overline{y} \subset X$ we have

$$v(\overline{x} + \overline{y}) = \{v(x + y) \mid x \in \overline{x},\ y \in \overline{y}\}$$
$$= \{vx + vy \mid x \in \overline{x},\ y \in \overline{y}\}$$
$$= v\overline{x} + v\overline{y}.$$

(2.2.31) is proved analogously.

For $\overline{v} \subset V$ and $\overline{x},\overline{y} \subset X$ we have

$$\overline{v}(\overline{x} + \overline{y}) = \{v(x + y) \mid v \in \overline{v},\ x \in \overline{x},\ y \in \overline{y}\}$$
$$\subset \{v_1x + v_2y \mid v_1,v_2 \in \overline{v},\ x \in \overline{x},\ y \in \overline{y}\}$$
$$= \overline{v}\,\overline{x} + \overline{v}\,\overline{y}.$$

(2.2.33) is proved analogously. □

NOTE. YOUNG [1931] introduced definition (2.2.5) for sets of real numbers. The theorems 2.2.1, 2.2.2 and 2.2.3 are simple generalisations of properties observed in this work. □

For $\alpha,\beta,\gamma,\delta \in \mathbb{R}$ with $\alpha \le \beta$ and $\gamma \le \delta$ we have the following rules (for a proof, see for instance ALEFELD & HERZBERGER [1974]).

(2.2.34) $[\alpha,\beta] + [\gamma,\delta] = [\alpha + \gamma,\ \beta + \delta]$,

(2.2.35) $[\alpha,\beta] - [\gamma,\delta] = [\alpha - \delta,\ \beta - \gamma]$,

(2.2.36) $[\alpha,\beta] \cdot [\gamma,\delta] = [\min(\alpha\gamma,\alpha\delta,\beta\gamma,\beta\delta),\ \max(\alpha\gamma,\alpha\delta,\beta\gamma,\beta\delta)]$,

(2.2.37) $[\alpha,\beta] / [\gamma,\delta] = [\alpha,\beta] \cdot [1/\delta,1/\gamma]$ (if $0 \notin [\gamma,\delta]$).

In particular we observe that for $\circledast = +,-,\cdot,/$ and $\overline{\xi},\overline{\eta} \in \mathbb{II}\,\mathbb{R}$ (with $0 \notin \overline{\eta}$ if $\circledast = /$) we have

(2.2.38) $\overline{\xi} \circledast \overline{\eta} \in \mathbb{II}\,\mathbb{R}$.

For $\xi \in \mathbb{R}$ we define

(2.2.39) $\xi^+ = \max(\xi,0), \qquad \xi^- = \max(-\xi,0).$

Now we can easily derive from (2.2.36)

(2.2.40) $\quad [0,\lambda] \,.\, [\alpha,\beta] = [-\lambda\alpha^{-},\lambda\beta^{+}]$

(for all $\lambda,\alpha,\beta \in \mathbb{R}$ with $\lambda \geq 0$, $\alpha \leq \beta$).

2.2.2. Addition and subtraction

THEOREM 2.2.4. *Let the operator* $\circledast$ *be* + *or* −. *For* $\overline{\xi}_i$, $\overline{\eta}_i \in \mathbb{I}\mathbb{R}$ $(1 \leq i \leq M)$ *we have*

$$(2.2.41) \qquad \begin{pmatrix} \overline{\xi}_1 \\ \vdots \\ \overline{\xi}_M \end{pmatrix} \circledast \begin{pmatrix} \overline{\eta}_1 \\ \vdots \\ \overline{\eta}_M \end{pmatrix} = \begin{pmatrix} \overline{\xi}_1 \circledast \overline{\eta}_1 \\ \vdots \\ \overline{\xi}_M \circledast \overline{\eta}_M \end{pmatrix} \in \mathbb{I}\mathbb{R}^M,$$

and for $\overline{\xi}_{ij},\overline{\eta}_{ij} \in \mathbb{I}\mathbb{R}$ $\quad (1 \leq i \leq M,\ 1 \leq j \leq M)$ *we have*

$$(2.2.42) \begin{pmatrix} \overline{\xi}_{11} & \cdots & \overline{\xi}_{1M} \\ \vdots & \cdots & \vdots \\ \overline{\xi}_{M1} & \cdots & \overline{\xi}_{MM} \end{pmatrix} \circledast \begin{pmatrix} \overline{\eta}_{11} & \cdots & \overline{\eta}_{1M} \\ \vdots & \cdots & \vdots \\ \overline{\eta}_{M1} & \cdots & \overline{\eta}_{MM} \end{pmatrix} = \begin{pmatrix} \overline{\xi}_{11}\circledast\overline{\eta}_{11} & \cdots & \overline{\xi}_{1M}\circledast\overline{\eta}_{1M} \\ \vdots & \cdots & \vdots \\ \overline{\xi}_{M1}\circledast\overline{\eta}_{M1} & \cdots & \overline{\xi}_{MM}\circledast\overline{\eta}_{MM} \end{pmatrix} \in \mathbb{I}\mathbb{R}^{M,M}.$$

PROOF.

$$\begin{pmatrix} \overline{\xi}_1 \\ \vdots \\ \overline{\xi}_M \end{pmatrix} \circledast \begin{pmatrix} \overline{\eta}_1 \\ \vdots \\ \overline{\eta}_M \end{pmatrix} =$$

$$= \{x \circledast y \mid x,y \in \mathbb{R}^M,\ [x]_i \in \overline{\xi}_i,\ [y]_i \in \overline{\eta}_i \ (1 \leq i \leq M)\}$$

$$= \{z \mid z \in \mathbb{R}^M,\ [z]_i \in \overline{\xi}_i \circledast \overline{\eta}_i \ (1 \leq i \leq M)\}$$

$$= \begin{pmatrix} \overline{\xi}_1 \circledast \overline{\eta}_1 \\ \vdots \\ \overline{\xi}_M \circledast \overline{\eta}_M \end{pmatrix} \in \mathbb{I}\mathbb{R}^M.$$

(2.2.42) can be proved analogously. □

THEOREM 2.2.5. *For* $\overline{x},\overline{y} \subset \mathbb{R}^M$ *we have*

$$p_i(\overline{x} + \overline{y}) = p_i(\overline{x}) + p_i(\overline{y}), \tag{2.2.43}$$

and for $\overline{A},\overline{B} \subset \mathbb{R}^{M,M}$ *and* $1 \le i \le M$, $1 \le j \le M$ *we have*

$$p_{ij}(\overline{A} + \overline{B}) = p_{ij}(\overline{A}) + p_{ij}(\overline{B}). \tag{2.2.44}$$

PROOF.

$$\begin{aligned} p_i(\overline{x} + \overline{y}) &= \{[z]_i \mid z \in \overline{x} + \overline{y}\} \\ &= \{[x+y]_i \mid x \in \overline{x},\ y \in \overline{y}\} \\ &= \{[x]_i + [y]_i \mid x \in \overline{x},\ y \in \overline{y}\} \\ &= \{\xi + \eta \mid \xi \in p_i(\overline{x}),\ \eta \in p_i(\overline{y})\} \\ &= p_i(\overline{x}) + p_i(\overline{y}). \end{aligned}$$

(2.2.44) can be proved analogously. □

THEOREM 2.2.6. *For* $V = \mathbb{R}^M$, $\mathbb{R}^{M,M}$ *and for all* $a,b,c,d \in \amalg V$ *with* $a \le b$, $c \le d$ *we have*

$$[a,b] + [c,d] = [a+c,\ b+d]. \tag{2.2.45}$$

PROOF. Let $a,b,c,d \in \amalg \mathbb{R}^M$ with $a \le b$, $c \le d$. By virtue of (2.2.43), (2.1.13) and (2.2.34) we have for $1 \le i \le M$

$$\begin{aligned} &p_i([a,b] + [c,d]) = p_i([a,b]) + p_i([c,d]) = \\ &= [[a]_i,[b]_i] + [[c]_i,[d]_i] \\ &= [[a]_i + [c]_i,\ [b]_i + [d]_i] \\ &= [[a+c]_i,\ [b+d]_i] \\ &= p_i([a+c,\ b+d]). \end{aligned}$$

In view of (2.2.41) we have $[a,b]+[c,d] \in \amalg \mathbb{R}^M$. Using (2.1.11) we obtain (2.2.45).

For $V = \mathbb{R}^{M,M}$ the proof is obtained analogously. □

THEOREM 2.2.7. *For* $V = \mathbb{R}^M$, $\mathbb{R}^{M,M}$ *and for all bounded, non-empty* $\overline{x},\overline{y} \subset V$ *and* $y \in V$ *we have*

$$\inf(\overline{x} + \overline{y}) = \inf \overline{x} + \inf \overline{y}, \tag{2.2.46}$$

$$\sup(\overline{x} + \overline{y}) = \sup \overline{x} + \sup \overline{y}, \tag{2.2.47}$$

$$\square(\overline{x} + \overline{y}) = \square \overline{x} + \square \overline{y}, \tag{2.2.48}$$

$$\square(\overline{x} + y) = \square \overline{x} + y. \tag{2.2.49}$$

PROOF. Consider first the case $V = \mathbb{R}^M$. Let $\overline{x},\overline{y} \subset \mathbb{R}^M$ be bounded and non-empty.

By virtue of (2.1.26) and (2.2.43) we have for $1 \le i \le M$

$$\begin{aligned}
[\inf(\overline{x} + \overline{y})]_i &= \inf p_i(\overline{x} + \overline{y}) \\
&= \inf(p_i(\overline{x}) + p_i(\overline{y})) \\
&= \inf p_i(\overline{x}) + \inf p_i(\overline{y}) \\
&= [\inf \overline{x}]_i + [\inf \overline{y}]_i \\
&= [\inf \overline{x} + \inf \overline{y}]_i,
\end{aligned}$$

which proves (2.2.46).

(2.2.47) can be proved analogously.

Combining (2.2.46) and (2.2.47) and using (2.2.45) we obtain

$$\begin{aligned}
\square(\overline{x} + \overline{y}) &= [\inf(\overline{x} + \overline{y}), \sup(\overline{x} + \overline{y})] \\
&= [\inf \overline{x} + \inf \overline{y}, \sup \overline{x} + \sup \overline{y}] \\
&= [\inf \overline{x}, \sup \overline{x}] + [\inf \overline{y}, \sup \overline{y}] \\
&= \square \overline{x} + \square \overline{y} .
\end{aligned}$$

(2.2.49) is a direct consequence of (2.2.48)

For the case $V = \mathbb{R}^{M,M}$ the proof is obtained analogously. □

2.2.3. Multiplication

THEOREM 2.2.8. *For* $\lambda \in \mathbb{R}$ *and* $\overline{\xi}_1,\ldots,\overline{\xi}_M \in \amalg \mathbb{R}$ *we have*

$$\lambda\begin{pmatrix}\overline{\xi}_1\\ \vdots\\ \overline{\xi}_M\end{pmatrix} = \begin{pmatrix}\lambda\overline{\xi}_1\\ \vdots\\ \lambda\overline{\xi}_M\end{pmatrix} \in \amalg \mathbb{R}^M, \tag{2.2.50}$$

and for $\lambda \in \mathbb{R}$ *and* $\overline{\xi}_{ij} \in \amalg \mathbb{R}$ $(1 \le i \le M,\ 1 \le j \le M)$ *we have*

$$\lambda\begin{pmatrix}\overline{\xi}_{11} & \cdots & \overline{\xi}_{1M}\\ \vdots & \cdots & \vdots\\ \overline{\xi}_{M1} & \cdots & \overline{\xi}_{MM}\end{pmatrix} = \begin{pmatrix}\lambda\overline{\xi}_{11} & \cdots & \lambda\overline{\xi}_{1M}\\ \vdots & \cdots & \vdots\\ \lambda\overline{\xi}_{M1} & \cdots & \lambda\overline{\xi}_{MM}\end{pmatrix} \in \amalg \mathbb{R}^{M,M}. \tag{2.2.51}$$

PROOF.

$$\lambda\begin{pmatrix}\overline{\xi}_1\\ \vdots\\ \overline{\xi}_M\end{pmatrix} = \{\lambda x \mid x \in \mathbb{R}^M,\ [x]_i \in \overline{\xi}_i\ (1 \le i \le M)\}$$

$$= \{y \mid y \in \mathbb{R}^M,\ [y]_i \in \lambda\overline{\xi}_i\ (1 \le i \le M)\}$$

$$= \begin{pmatrix}\lambda\overline{\xi}_1\\ \vdots\\ \lambda\overline{\xi}_M\end{pmatrix} \in \amalg \mathbb{R}^M.$$

(2.2.51) can be proved analogously. □

THEOREM 2.2.9. *Let* $\overline{\lambda} \subset \mathbb{R}$. *For all* $\overline{x} \subset \mathbb{R}^M$ *and* $1 \le i \le M$ *we have*

$$p_i(\overline{\lambda}\,\overline{x}) = \overline{\lambda}\, p_i(\overline{x}) \tag{2.2.52}$$

and for all $\overline{A} \subset \mathbb{R}^{M,M}$ *and* $1 \le i \le M,\ 1 \le j \le M$ *we have*

$$p_{ij}(\overline{\lambda}\,\overline{A}) = \overline{\lambda}\, p_{ij}(\overline{A}). \tag{2.2.53}$$

PROOF.

$$\begin{aligned} p_i(\overline{\lambda}\,\overline{x}) &= \{[\lambda x]_i \mid \lambda \in \overline{\lambda},\ x \in \overline{x}\} \\ &= \{\lambda[x]_i \mid \lambda \in \overline{\lambda},\ x \in \overline{x}\} \\ &= \{\lambda\xi \mid \lambda \in \overline{\lambda},\ \xi \in p_i(\overline{x})\} \\ &= \overline{\lambda}\ p_i(\overline{x}). \end{aligned}$$

(2.2.53) can be proved analogously. □

NOTE. We use the notational convention that, if the rounding operator □ is followed by any product, then this product should be evaluated before applying the rounding operator. For instance, □ $\overline{A}\,\overline{x}$ should be interpreted as □$(\overline{A}\,\overline{x})$, for $\overline{A} \subset \mathbb{R}^{M,M}$, $\overline{x} \subset \mathbb{R}^M$, $\overline{A}$ and $\overline{x}$ bounded and non-empty. □

THEOREM 2.2.10. *Let* $\overline{\lambda} \in \amalg\mathbb{R}$. *For* $\overline{\xi}_1,\ldots,\overline{\xi}_M \in \amalg\mathbb{R}$ *we have*

$$(2.2.54)\qquad \square\,\overline{\lambda}\begin{pmatrix}\overline{\xi}_1\\ \vdots\\ \overline{\xi}_M\end{pmatrix} = \begin{pmatrix}\overline{\lambda}\,\overline{\xi}_1\\ \vdots\\ \overline{\lambda}\,\overline{\xi}_M\end{pmatrix}$$

and for $\overline{\xi}_{ij} \in \amalg\mathbb{R}$ $(1 \le i \le M,\ 1 \le j \le M)$ *we have*

$$(2.2.55)\qquad \square\,\overline{\lambda}\begin{pmatrix}\overline{\xi}_{11} & \cdots & \overline{\xi}_{1M}\\ \vdots & \cdots & \vdots\\ \overline{\xi}_{M1} & \cdots & \overline{\xi}_{MM}\end{pmatrix} = \begin{pmatrix}\overline{\lambda}\,\overline{\xi}_{11} & \cdots & \overline{\lambda}\,\overline{\xi}_{1M}\\ \vdots & \cdots & \vdots\\ \overline{\lambda}\,\overline{\xi}_{M1} & \cdots & \overline{\lambda}\,\overline{\xi}_{MM}\end{pmatrix}.$$

PROOF. For $1 \le i \le M$ we have, by virtue of (2.1.37), (2.2.52) and (2.2.38),

$$p_i\left(\square\,\overline{\lambda}\begin{pmatrix}\overline{\xi}_1\\ \vdots\\ \overline{\xi}_M\end{pmatrix}\right) = \square\ p_i\left(\overline{\lambda}\begin{pmatrix}\overline{\xi}_1\\ \vdots\\ \overline{\xi}_M\end{pmatrix}\right) = \square\ \overline{\lambda}\,\overline{\xi}_i = \overline{\lambda}\,\overline{\xi}_i.$$

In view of (2.1.11) this proves (2.2.54).

(2.2.55) can be proved analogously. □

REMARK 2.2.11. In general we do not have $\overline{\lambda}\,\overline{x} \in \amalg\, \mathbb{R}^M$ for $\overline{\lambda} \in \amalg\, \mathbb{R}$ and $\overline{x} \in \amalg\, \mathbb{R}^M$. This is illustrated in fig. 2.2.1 for M = 2, $\overline{\lambda} = [1,2]$ and $\overline{x} = \begin{pmatrix}[1,3]\\ [1,2]\end{pmatrix}$. In this figure the shaded hexagon represents $\overline{\lambda}\,\overline{x}$.

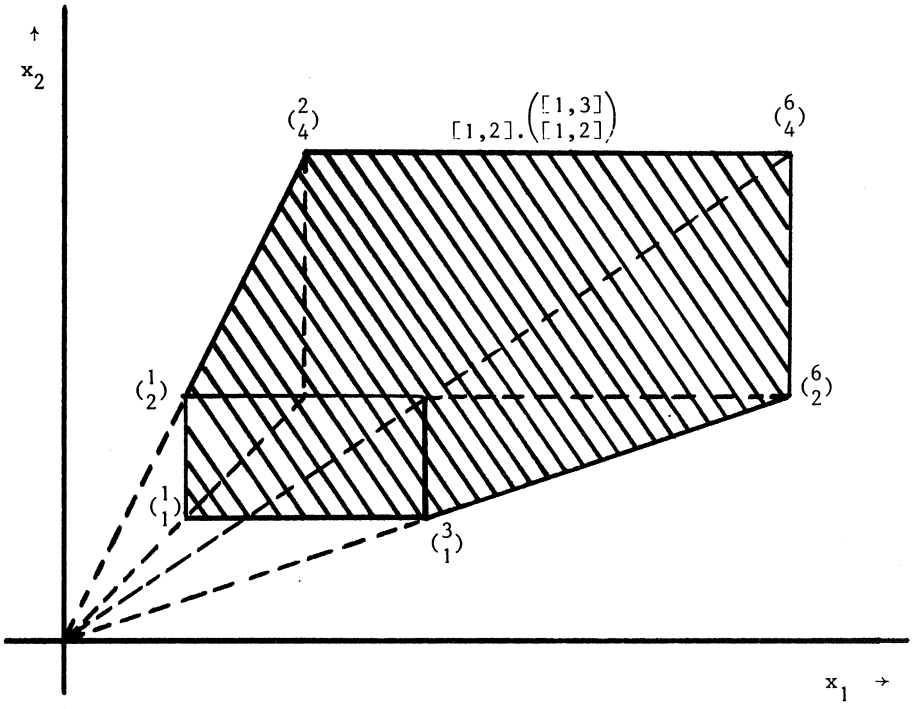

fig. 2.2.1

Similarly, in general we do not have $\overline{\lambda}\,\overline{A} \in \amalg\, \mathbb{R}^{M,M}$ for $\overline{\lambda} \in \amalg\, \mathbb{R}$ and $\overline{A} \in \amalg\, \mathbb{R}^{M,M}$. □

THEOREM 2.2.12. *For* $V = \mathbb{R}^M$, $\mathbb{R}^{M,M}$, $\lambda \in \mathbb{R}$ *and bounded and non-empty* $\overline{x} \subset V$ *we have*

(2.2.56) $\quad \square\, \lambda\, \overline{x} = \lambda\, \square\, \overline{x}.$

PROOF. Let $\overline{\xi} \subset \mathbb{R}$ be bounded and non-empty and let $\lambda \in \mathbb{R}$. For $\lambda \geq 0$ we have

$$[\inf \lambda\overline{\xi}\,,\ \sup \lambda\overline{\xi}\,] = [\lambda \inf \overline{\xi},\ \lambda \sup \overline{\xi}] = \lambda[\inf \overline{\xi},\ \sup \overline{\xi}]$$

and for $\lambda < 0$ we have

$$[\inf \lambda\overline{\xi}\,,\ \sup \lambda\overline{\xi}\,] = [\lambda \sup \overline{\xi},\ \lambda \inf \overline{\xi}] = \lambda[\inf \overline{\xi},\ \sup \overline{\xi}].$$

Therefore for all $\lambda \in \mathbb{R}$ we have

$$[\inf \lambda\overline{\xi}\,,\ \sup \lambda\overline{\xi}\,] = \lambda[\inf \overline{\xi},\ \sup \overline{\xi}]$$

and hence

$$\square\, \lambda\overline{\xi} = \lambda \square\, \overline{\xi}. \tag{2.2.57}$$

Now let $\overline{x} \subset \mathbb{R}^M$ be bounded and non-empty, and let again $\lambda \in \mathbb{R}$. By virtue of (2.1.37), (2.2.52) and (2.2.57) we have

$$p_i(\square\, \lambda\overline{x}) = \square\, p_i(\lambda\overline{x}) = \square\, \lambda\, p_i(\overline{x}) = \lambda \square\, p_i(\overline{x}) =$$
$$= \lambda\, p_i(\square\, \overline{x}) = p_i(\lambda \square\, \overline{x}).$$

We have $\square\, \lambda\overline{x} \in \amalg \mathbb{R}^M$ and, by virtue of (2.2.50), $\lambda \square\, \overline{x} \in \amalg \mathbb{R}^M$. Therefore, and in view of (2.1.11), we obtain (2.2.56).

For $V = \mathbb{R}^{M,M}$ the proof is obtained analogously. □

The following theorem can essentially be found in KULISCH [1976], pp. 412, 414.

THEOREM 2.2.13. *For* $\overline{A} \in \amalg \mathbb{R}^{M,M}$, $\overline{x} \in \amalg \mathbb{R}^M$ *and* $1 \le i \le M$ *we have*

$$p_i(\overline{A}\,\overline{x}) = \sum_{j=1}^{M} p_{ij}(\overline{A}).p_j(\overline{x}) \tag{2.2.58}$$

and for $\overline{A},\overline{B} \in \amalg \mathbb{R}^{M,M}$, $1 \le i \le M$ *and* $1 \le j \le M$ *we have*

$$p_{ij}(\overline{A}\,\overline{B}) = \sum_{k=1}^{M} p_{ik}(\overline{A}).p_{kj}(\overline{B}). \tag{2.2.59}$$

PROOF.

$$
\begin{aligned}
p_i(\overline{A}\,\overline{x}) &= \{[Ax]_i \mid A \in \overline{A},\ x \in \overline{x}\} \\
&= \{\sum_{j=1}^{M} [A]_{ij}[x]_j \mid A \in \overline{A},\ x \in \overline{x}\} \\
&= \{\sum_{j=1}^{M} \alpha_j\, \xi_j \mid \alpha_j \in p_{ij}(\overline{A}),\ \xi_j \in p_j(\overline{x})(1 \le j \le M)\} \\
&= \sum_{j=1}^{M} \{\alpha_j\, \xi_j \mid \alpha_j \in p_{ij}(\overline{A}),\ \xi_j \in p_j(\overline{x})\} \\
&= \sum_{j=1}^{M} p_{ij}(\overline{A})p_j(\overline{x}).
\end{aligned}
$$

$$
\begin{aligned}
p_{ij}(\overline{A}\,\overline{B}) &= \{[AB]_{ij} \mid A \in \overline{A},\ B \in \overline{B}\} \\
&= \{\sum_{k=1}^{M} [A]_{ik}[B]_{kj} \mid A \in \overline{A},\ B \in \overline{B}\} \\
&= \{\sum_{k=1}^{M} \alpha_k\, \beta_k \mid \alpha_k \in p_{ik}(\overline{A}),\ \beta_k \in p_{kj}(\overline{B})(1 \le k \le M)\} \\
&= \sum_{k=1}^{M} \{\alpha_k\, \beta_k \mid \alpha_k \in p_{ik}(\overline{A}),\ \beta_k \in p_{kj}(\overline{B})\} \\
&= \sum_{k=1}^{M} p_{ik}(\overline{A})p_{kj}(\overline{B}).
\end{aligned}
$$

□

REMARK 2.2.14. As we mentioned in remark 2.1.1, many authors use interval vectors and interval matrices rather than vector intervals and matrix intervals. Most of them (for instance ALEFELD & HERZBERGER [1974], HANSEN [1965], HUNGER [1971] and MOORE [1966]) define the product of two interval matrices A_I and B_I as the interval matrix A_IB_I satisfying

$$
(2.2.60) \qquad [A_IB_I]_{ij} = \sum_{k=1}^{M} [A_I]_{ik}[B_I]_{kj}.
$$

Here $[A_I]_{ij}$ is the element in the i'th row and j'th column of A_I. In view of theorem 2.2.13 the set of matrices corresponding to the interval matrix A_IB_I is $\square\, \overline{A}\,\overline{B}$, where $\overline{A}$ and $\overline{B}$ are the matrix intervals corresponding to A_I and B_I respectively.

Our main reason for using the concept of matrix interval as a special kind of set of matrices, and for using the general definition (2.2.5) for

operations on sets, is that we are thus able to distinguish between the sets $\overline{A}\,\overline{B}$ and $\square\,\overline{A}\,\overline{B}$.

Similar considerations apply to the product of a matrix interval and a vector interval, and to the product of a real interval and a vector interval or matrix interval. $\square$

2.2.4. Integration

Let $\alpha,\beta \in \mathbb{R}$ with $\alpha \le \beta$. For a continuous function $x: [\alpha,\beta] \to \mathbb{R}^M$ we define $\int_\alpha^\beta x(t)dt$, as usual, by

$$(2.2.61) \qquad [\int_\alpha^\beta x(t)dt]_i = \int_\alpha^\beta [x(t)]_i dt \qquad (1 \le i \le M).$$

Similarly, for a continuous function $A: [\alpha,\beta] \to \mathbb{R}^{M,M}$ we define $\int_\alpha^\beta A(t)dt$ by

$$(2.2.62) \qquad [\int_\alpha^\beta A(t)dt]_{ij} = \int_\alpha^\beta [A(t)]_{ij} dt \qquad (1 \le i \le M,\ 1 \le j \le M).$$

The following theorem gives an inclusion of these integrals.

THEOREM 2.2.15. *Let* $\alpha,\beta \in \mathbb{R}$ *with* $\alpha \le \beta$ *and let* V *be* $\mathbb{R}^M$ *or* $\mathbb{R}^{M,M}$. *For a continuous function* $x: [\alpha,\beta] \to V$ *we have*

$$(2.2.63) \qquad \int_\alpha^\beta x(t)dt \in (\beta-\alpha).\ \square\, x([\alpha,\beta]).$$

PROOF. Let $V = \mathbb{R}^M$, $\alpha < \beta$ and $1 \le i \le M$. For $\alpha \le t \le \beta$ we have

$$[x(t)]_i \in p_i(x([\alpha,\beta])),$$

hence

$$[x(t)]_i \ge \inf p_i(x([\alpha,\beta])).$$

Therefore, and by virtue of (2.1.26), we obtain

$$[\frac{1}{\beta-\alpha} \int_\alpha^\beta x(t)dt]_i = \frac{1}{\beta-\alpha} \int_\alpha^\beta [x(t)]_i dt$$

$$\ge \inf p_i(x([\alpha,\beta])) = [\inf x([\alpha,\beta])]_i.$$

This holds for $1 \le i \le M$, hence

$$\frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x(t)dt \geq \inf x([\alpha,\beta]).$$

Analogously we can derive

$$\frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x(t)dt \leq \sup x([\alpha,\beta]).$$

Consequently

$$\frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x(t)dt \in \square\, x([\alpha,\beta]),$$

which proves (2.2.29).

For $\alpha = \beta$ the proof is trivial.

For the case $V = \mathbb{R}^{M,M}$ the proof is obtained analogously. □

2.2.5. Taylor series

THEOREM 2.2.16. *Let* $\alpha,\beta \in \mathbb{R}$ *with* $\alpha \leq \beta$, *let* V *be* $\mathbb{R}^M$ *or* $\mathbb{R}^{M,M}$, *and let* k *be an integer with* $k \geq 1$. *Let* $x: [\alpha,\beta] \to V$ *be a* k *times continuously differentiable function. For* $t_0, t \in [\alpha,\beta]$ *we have*

$$(2.2.64)\quad x(t) \in x(t_0) + \sum_{j=1}^{k-1} \frac{(t-t_0)^j}{j!} x^{(j)}(t_0) + \frac{(t-t_0)^k}{k!}\, \square\, x^{(k)}([\alpha,\beta]).$$

PROOF. Let $V = \mathbb{R}^M$. We have

$$x(t) = x(t_0) + \sum_{j=1}^{k-1} \frac{(t-t_0)^j}{j!} x^{(j)}(t_0) + \frac{(t-t_0)^k}{k!} \cdot r,$$

where $r \in \mathbb{R}^M$ is such that for all i with $1 \leq i \leq M$ there is an $s_i \in [\alpha,\beta]$ such that

$$[r]_i = [x^{(k)}(s_i)]_i.$$

Therefore

$$[r]_i \in p_i(x^{(k)}([\alpha,\beta])) \subset \square\, p_i(x^{(k)}([\alpha,\beta])) \quad (1 \leq i \leq M).$$

Consequently, using (2.1.37) we obtain

$$r \in \square\, x^{(k)}([\alpha,\beta]),$$

which proves (2.2.64).

For the case $V = \mathbb{R}^{M,M}$ the proof is obtained analogously. □

NOTATION. Let $V \subset \mathbb{R}^M$ be open and let $g : V \to \mathbb{R}^M$ be given. If g is differentiable then we denote by $g'(x)$ the matrix of partial derivatives of g in x. □

LEMMA 2.2.17. *Let* $V \subset \mathbb{R}^M$ *be open and let* $g : V \to \mathbb{R}^M$ *be continuously differentiable. Let* $\overline{x} \subset V$ *be convex. For all* $x_1, x_2 \in \overline{x}$ *we have*

$$(2.2.65) \quad g(x_1) - g(x_2) = \left[\int_0^1 g'(x_2 + s(x_1 - x_2))ds\right](x_1 - x_2).$$

PROOF. We dispense with the simple proof. □

THEOREM 2.2.18. *Let* $V \subset \mathbb{R}^M$ *be open and let* $g : V \to \mathbb{R}^M$ *be continuously differentiable. Let* $\overline{x} \subset V$ *be bounded and convex. For all* $x_1, x_2 \in \overline{x}$ *we have*

$$(2.2.66) \quad g(x_1) - g(x_2) \in [\square\, g'(\overline{x})](x_1 - x_2).$$

PROOF. Combining lemma 2.2.17 and theorem 2.2.15 we obtain

$$g(x_1) - g(x_2) \in [\square\, g'(x_2 + [0,1](x_1 - x_2))](x_1 - x_2).$$

In view of the convexity of $\overline{x}$ this implies (2.2.66). □

2.2.6. Set valued functions

A set valued function is denoted by a symbol with a bar, for instance $\overline{f}$.

For a set V the set of all subsets of V (i.e., the power set of V) is denoted by $\mathbb{P}V$.

DEFINITION 2.2.19. Let V and W be sets and let $X \subset \mathbb{P}V$. A function $\overline{f} : X \to \mathbb{P}W$ is said to be *inclusion isotonic* if it satisfies

$$(2.2.67) \quad \overline{v}_1 \subset \overline{v}_2 \Rightarrow \overline{f}(\overline{v}_1) \subset \overline{f}(\overline{v}_2) \quad \text{(for all } \overline{v}_1, \overline{v}_2 \in X\text{)}. \quad \square$$

2.3. THE NORM, DIAMETER AND DISTANCE

2.3.1. The norm of a set

We define the *norm* of a vector $x \in \mathbb{R}^M$ by

$$\|x\| = \max_{1\le i\le M} |[x]_i| \tag{2.3.1}$$

and of a matrix $A \in \mathbb{R}^{M,M}$ by

$$\|A\| = \max_{1\le i\le M} \sum_{j=1}^{M} |[A]_{ij}|. \tag{2.3.2}$$

We thus use the maximum norm for vectors and the norm for matrices induced by this vector norm.

For a bounded, non-empty set $\overline{\xi} \subset \mathbb{R}$ we define

$$|\overline{\xi}| = \sup_{\xi\in\overline{\xi}} |\xi| \qquad \text{(the } \textit{absolute value of } \overline{\xi}\text{)} \tag{2.3.3}$$

and for bounded, non-empty sets $\overline{x} \subset \mathbb{R}^M$ and $\overline{A} \subset \mathbb{R}^{M,M}$ we define

$$\|\overline{x}\| = \sup_{x\in\overline{x}} \|x\| \tag{2.3.4}$$

and

$$\|\overline{A}\| = \sup_{A\in\overline{A}} \|A\|. \tag{2.3.5}$$

As in APOSTOLATOS & KULISCH [1968], where these definitions can be found, we will call $\|\overline{x}\|$ and $\|\overline{A}\|$ the *norm* of $\overline{x}$ and $\overline{A}$, respectively. Note however that these functions are not norms in the strict sense, since they are not defined on vector spaces.

NOTE. The real number $|\overline{\xi}|$ should not be confused with the set $g(\overline{\xi})$, where the function $g : \mathbb{R} \to \mathbb{R}$ is defined by $g(\xi) = |\xi|$. In fact $|\overline{\xi}| = \sup g(\overline{\xi})$. Similar remarks apply to the norm of a set of vectors and of a set of matrices. □

From the definitions (2.3.3), (2.3.4) and (2.3.5) we immediately obtain the following theorem.

THEOREM 2.3.1. *For bounded, non-empty* $\overline{\xi},\overline{\eta} \subset \mathbb{R}$ *we have*

$$(2.3.6) \qquad \overline{\xi} \subset \overline{\eta} \Rightarrow |\overline{\xi}| \leq |\overline{\eta}| .$$

For $V = \mathbb{R}^M$, $\mathbb{R}^{M,M}$ *and bounded, non-empty* $\overline{x},\overline{y} \subset V$ *we have*

$$(2.3.7) \qquad \overline{x} \subset \overline{y} \Rightarrow \|\overline{x}\| \leq \|\overline{y}\| . \qquad \square$$

THEOREM 2.3.2. *For* $\alpha,\beta \in \mathbb{R}$ *with* $\alpha \leq \beta$ *we have*

$$(2.3.8) \qquad |[\alpha,\beta]| = \max(|\alpha|,|\beta|) = \max(-\alpha,\beta).$$

For bounded, non-empty $\overline{x} \subset \mathbb{R}^M$ *we have*

$$(2.3.9) \qquad \|\overline{x}\| = \max_{1\leq i\leq M} |p_i(\overline{x})| .$$

For bounded, non-empty $\overline{A} \subset \mathbb{R}^{M,M}$ *we have*

$$(2.3.10) \qquad \|\overline{A}\| \leq \max_{1\leq i\leq M} \sum_{j=1}^{M} |p_{ij}(\overline{A})| .$$

If $\overline{A} \in \amalg \mathbb{R}^{M,M}$ *then* (2.3.10) *can be replaced by the stronger statement*

$$(2.3.11) \qquad \|\overline{A}\| = \max_{1\leq i\leq M} \sum_{j=1}^{M} |p_{ij}(\overline{A})| .$$

PROOF. (2.3.8) can easily be derived from definition (2.3.3).

$$\|\overline{x}\| = \sup_{x\in\overline{x}} \max_{1\leq i\leq M} |[x]_i|$$

$$= \max_{1\leq i\leq M} \sup_{\xi\in p_i(\overline{x})} |\xi|$$

$$= \max_{1\leq i\leq M} |p_i(\overline{x})| .$$

$$\|\overline{A}\| = \sup_{A\in\overline{A}} \max_{1\le i\le M} \sum_{j=1}^{M} |[A]_{ij}|$$

$$\le \max_{1\le i\le M} \sum_{j=1}^{M} \sup_{A\in\overline{A}} |[A]_{ij}|$$

$$= \max_{1\le i\le M} \sum_{j=1}^{M} |p_{ij}(\overline{A})|.$$

If $\overline{A} \in \amalg R^{M,M}$ then there is a matrix $A \in \overline{A}$ such that

$$|[A]_{ij}| = |p_{ij}(\overline{A})| \qquad (1 \le i \le M,\ 1 \le j \le M).$$

For this matrix A we have

$$\|A\| \le \|\overline{A}\| \le \max_{1\le i\le M} \sum_{j=1}^{M} |p_{ij}(\overline{A})| = \|A\|,$$

which implies (2.3.11). □

THEOREM 2.3.3. *Let* $V = \mathbb{R}^M$ *or* $V = \mathbb{R}^{M,M}$. *For all bounded, non-empty* $\overline{\lambda} \subset \mathbb{R}$, $\overline{x},\overline{y} \subset V$ *and* $\overline{A} \subset \mathbb{R}^{M,M}$ *we have*

(2.3.12) $\|\overline{x} \pm \overline{y}\| \le \|\overline{x}\| + \|\overline{y}\|$,

(2.3.13) $\|\overline{\lambda}\,\overline{x}\| = |\overline{\lambda}| \cdot \|\overline{x}\|$,

(2.3.14) $\|\overline{A}\,\overline{x}\| \le \|\overline{A}\| \cdot \|\overline{x}\|$.

PROOF.

$$\|\overline{x} \pm \overline{y}\| = \sup_{\substack{x\in\overline{x}\\ y\in\overline{y}}} \|x \pm y\| \le \sup_{\substack{x\in\overline{x}\\ y\in\overline{y}}} (\|x\| + \|y\|) = \|\overline{x}\| + \|\overline{y}\|.$$

$$\|\overline{\lambda}\,\overline{x}\| = \sup_{\substack{\lambda\in\overline{\lambda}\\ x\in\overline{x}}} \|\lambda x\| = \sup_{\substack{\lambda\in\overline{\lambda}\\ x\in\overline{x}}} (|\lambda| \cdot \|x\|) = |\overline{\lambda}| \cdot \|\overline{x}\|.$$

$$\|\overline{A}\,\overline{x}\| = \sup_{\substack{A\in\overline{A}\\ x\in\overline{x}}} \|Ax\| \le \sup_{\substack{A\in\overline{A}\\ x\in\overline{x}}} (\|A\|.\|x\|) = \|\overline{A}\|.\|\overline{x}\|. \qquad \square$$

THEOREM 2.3.4. *For all bounded, non-empty* $\overline{\xi} \subset \mathbb{R}$, $\overline{x} \subset \mathbb{R}^M$ *and* $\overline{A} \subset \mathbb{R}^{M,M}$ *we have*

$$|\,\square\,\overline{\xi}\,| = |\overline{\xi}|, \tag{2.3.15}$$

$$\|\,\square\,\overline{x}\| = \|\overline{x}\|, \tag{2.3.16}$$

$$\|\,\square\,\overline{A}\| \le M.\|\overline{A}\|. \tag{2.3.17}$$

PROOF. By virtue of (2.3.8) we have

$$|\,\square\,\overline{\xi}\,| = \max(-\inf\,\overline{\xi},\ \sup\,\overline{\xi}) = \sup_{\xi\in\overline{\xi}} \max(-\xi,\xi) = \sup_{\xi\in\overline{\xi}} |\xi| = |\overline{\xi}|.$$

Using (2.3.9), (2.1.37) and (2.3.15) we obtain

$$\|\,\square\,\overline{x}\| = \max_{1\le i\le M} |p_i(\,\square\,\overline{x})| = \max_{1\le i\le M} |\,\square\,p_i(\overline{x})|$$

$$= \max_{1\le i\le M} |p_i(\overline{x})| = \|\overline{x}\|.$$

In view of (2.3.11), (2.1.38) and (2.3.15) we find

$$\|\,\square\,\overline{A}\| = \max_{1\le i\le M} \sum_{j=1}^{M} |p_{ij}(\,\square\,\overline{A})| = \max_{1\le i\le M} \sum_{j=1}^{M} |\,\square\,p_{ij}(\overline{A})|$$

$$= \max_{1\le i\le M} \sum_{j=1}^{M} |p_{ij}(\overline{A})| = \max_{1\le i\le M} \sum_{j=1}^{M} \sup_{A\in\overline{A}} |[A]_{ij}|$$

$$\le \max_{1\le i\le M} \sum_{j=1}^{M} \sup_{A\in\overline{A}} \|A\| = M.\|\overline{A}\|. \qquad \square$$

REMARK 2.3.5. For arbitrary M there are non-trivial sets $\overline{A}$ for which we have equality in (2.3.17). For instance, let

$$\overline{A} = \{A \mid A \in \mathbb{R}^{M,M},\ \|A\| \le 1\},$$

then

$$\square\,\overline{A} = \{A \mid A \in \mathbb{R}^{M,M},\ |[A]_{ij}| \le 1 \quad (1 \le i \le M,\ 1 \le j \le M)\}.$$

Thus we see that $\|\overline{A}\| = 1$ and $\|\square\,\overline{A}\| = M$. □

We define $\overline{e} \in \mathbb{I}\mathbb{R}^M$ by

$$(2.3.18) \qquad \overline{e} = \begin{pmatrix} [-1,1] \\ \vdots \\ [-1,1] \end{pmatrix}.$$

The following theorem gives a number of properties of this vector interval.

THEOREM 2.3.6.

$$(2.3.19) \qquad \|\overline{e}\| = 1,$$

$$(2.3.20) \qquad \lambda\,\overline{e} = \{x \mid x \in \mathbb{R}^M,\ \|x\| \le |\lambda|\} \qquad (\textit{for } \lambda \in \mathbb{R}),$$

$$(2.3.21) \qquad \overline{x} \subset \|\overline{x}\|\,.\,\overline{e} \qquad (\textit{for bounded, non-empty } \overline{x} \subset \mathbb{R}^M),$$

$$(2.3.22) \qquad |\lambda| \le |\mu| \Rightarrow \lambda\,\overline{e} \subset \mu\,\overline{e} \qquad (\textit{for } \lambda,\mu \in \mathbb{R}),$$

$$(2.3.23) \qquad \overline{\lambda}\,\overline{e} = |\overline{\lambda}|\,.\,\overline{e} \qquad (\textit{for } \overline{\lambda} \in \mathbb{I}\mathbb{R}).$$

PROOF. (2.3.19) follows from (2.3.18) and (2.3.9).

For $\lambda \in \mathbb{R}$ we have by virtue of (2.2.50)

$$\lambda\,\overline{e} = \begin{pmatrix} \lambda\,.\,[-1,1] \\ \vdots \\ \lambda\,.\,[-1,1] \end{pmatrix} = \begin{pmatrix} [-|\lambda|,|\lambda|] \\ \vdots \\ [-|\lambda|,|\lambda|] \end{pmatrix}$$

$$= \{x \mid x \in \mathbb{R}^M,\ \|x\| \le |\lambda|\}.$$

(2.3.21), (2.3.22) and (2.3.23) are consequences of (2.3.20). □

THEOREM 2.3.7. *Let* $V \subset \mathbb{R}^M$ *be open and let* $g : V \to \mathbb{R}^M$ *be continuously differentiable. Let* $\overline{x} \subset V$ *be bounded and convex. For all* $x_1, x_2 \in \overline{x}$ *we have*

(2.3.24) $\|g(x_1) - g(x_2)\| \leq \|g'(\overline{x})\| \,.\, \|x_1 - x_2\|.$

PROOF. Using lemma 2.2.17 we obtain

$$\|g(x_1) - g(x_2)\| \leq \int_0^1 \|g'(x_2 + s(x_1 - x_2))\| ds \,.\, \|x_1 - x_2\|.$$

Due to the convexity of $\overline{x}$ this implies (2.3.24). □

2.3.2. The diameter of a set

For $V = \mathbb{R}^M$, $\mathbb{R}^{M,M}$ and bounded, non-empty $\overline{x} \in V$ we define the *diameter* of $\overline{x}$ by

(2.3.25) $$\operatorname{diam} \overline{x} = \sup_{x_1, x_2 \in \overline{x}} \|x_1 - x_2\|.$$

The following theorem gives some properties of the diameter, which are direct consequences of the definition.

THEOREM 2.3.8. *For* $V = \mathbb{R}^M$, $\mathbb{R}^{M,M}$ *and bounded, non-empty* $\overline{x}, \overline{y} \subset V$ *and* $\hat{x} \in \overline{x}$ *we have*

(2.3.26) $\overline{x} \subset \overline{y} \Rightarrow \operatorname{diam} \overline{x} \leq \operatorname{diam} \overline{y}$,

(2.3.27) $\operatorname{diam} \overline{x} = \|\overline{x} - \overline{x}\|$,

(2.3.28) $\operatorname{diam} \overline{x} \leq 2\|\overline{x}\|$,

(2.3.29) $\|\overline{x} - \hat{x}\| \leq \operatorname{diam} \overline{x}$. □

THEOREM 2.3.9. *For bounded, non-empty* $\overline{x} \subset \mathbb{R}^M$ *we have*

(2.3.30) $$\operatorname{diam} \overline{x} = \max_{1 \leq i \leq M} \operatorname{diam} p_i(\overline{x}),$$

(2.3.31) $\operatorname{diam} \square \overline{x} = \operatorname{diam} \overline{x}$.

PROOF.

$$\operatorname{diam} \bar{x} = \sup_{x_1, x_2 \in \bar{x}} \max_{1 \le i \le M} |[x_1]_i - [x_2]_i|$$

$$= \max_{1 \le i \le M} \sup_{\xi_1, \xi_2 \in p_i(\bar{x})} |\xi_1 - \xi_2|$$

$$= \max_{1 \le i \le M} \operatorname{diam} p_i(\bar{x}).$$

By virtue of (2.3.27) and (2.3.16) we have

$$\operatorname{diam} (\square \bar{x}) = \| \square \bar{x} - \square \bar{x} \| = \| \square (\bar{x} - \bar{x}) \|$$

$$= \|\bar{x} - \bar{x}\| = \operatorname{diam} \bar{x} . \qquad \square$$

THEOREM 2.3.10. *For all bounded, non-empty* $\bar{\lambda} \subset \mathbb{R}$ *and* $\bar{x}, \bar{y} \subset \mathbb{R}^M$ *and all* $\lambda \in \mathbb{R}$ *and* $y \in \mathbb{R}^M$ *we have*

(2.3.32) $\operatorname{diam} (\bar{x} + \bar{y}) \le \operatorname{diam} \bar{x} + \operatorname{diam} \bar{y}$,

(2.3.33) $\operatorname{diam} (\bar{x} + y) = \operatorname{diam} \bar{x}$,

(2.3.34) $\operatorname{diam} (\lambda \bar{x}) = |\lambda| \cdot \operatorname{diam} \bar{x}$,

(2.3.35) $\operatorname{diam} (\bar{\lambda} \bar{x}) \ge |\bar{\lambda}| \cdot \operatorname{diam} \bar{x}$.

PROOF.

$$\operatorname{diam} (\bar{x} + \bar{y}) = \|(\bar{x} + \bar{y}) - (\bar{x} + \bar{y})\|$$

$$= \|(\bar{x} - \bar{x}) + (\bar{y} - \bar{y})\| \le \|\bar{x} - \bar{x}\| + \|\bar{y} - \bar{y}\|$$

$$= \operatorname{diam} \bar{x} + \operatorname{diam} \bar{y}.$$

$$\operatorname{diam} (\bar{x} + y) = \|(\bar{x} + y) - (\bar{x} + y)\| = \|\bar{x} - \bar{x}\| = \operatorname{diam} \bar{x} .$$

Using (2.2.30) we obtain

$$\operatorname{diam}(\lambda \overline{x}) = \|\lambda \overline{x} - \lambda \overline{x}\| = \|\lambda(\overline{x} - \overline{x})\| = |\lambda| \cdot \|\overline{x} - \overline{x}\|$$

$$= |\lambda| \cdot \operatorname{diam} \overline{x} .$$

Similarly, in view of (2.2.32) we have

$$\operatorname{diam}(\overline{\lambda}\, \overline{x}) = \|\overline{\lambda}\, \overline{x} - \overline{\lambda}\, \overline{x}\| \geq \|\overline{\lambda}(\overline{x} - \overline{x})\| = |\overline{\lambda}| \cdot \|\overline{x} - \overline{x}\|$$

$$= |\overline{\lambda}| \cdot \operatorname{diam} \overline{x} .$$

THEOREM 2.3.11. *For* $\overline{x} \in \amalg \mathbb{R}^M$ *we have*

$$\|\overline{x} - \operatorname{mean} \overline{x}\| = \tfrac{1}{2} \operatorname{diam} \overline{x} . \tag{2.3.36}$$

PROOF. For $\alpha, \beta \in \mathbb{R}$ with $\alpha \leq \beta$ we have

$$|[\alpha,\beta] - \operatorname{mean} [\alpha,\beta]| = |[-\tfrac{1}{2}(\beta-\alpha), \tfrac{1}{2}(\beta-\alpha)]|$$

$$= \tfrac{1}{2}(\beta-\alpha) = \tfrac{1}{2} \operatorname{diam}([\alpha,\beta]).$$

Let $\overline{x} \in \amalg \mathbb{R}^M$, $\hat{x} = \operatorname{mean} \overline{x}$. We have

$$[\hat{x}]_i = \operatorname{mean} p_i(\overline{x}) \qquad (1 \leq i \leq M).$$

Hence, using (2.3.30) and (2.3.9) we obtain

$$\operatorname{diam} \overline{x} = \max_{1 \leq i \leq M} \operatorname{diam} p_i(\overline{x})$$

$$= 2 \max_{1 \leq i \leq M} |p_i(\overline{x}) - \operatorname{mean} p_i(\overline{x})|$$

$$= 2 \max_{1 \leq i \leq M} |p_i(\overline{x}) - [\hat{x}]_i| = 2 \max_{1 \leq i \leq M} |p_i(\overline{x} - \hat{x})|$$

$$= 2 \|\overline{x} - \hat{x}\| ,$$

which proves (2.3.36). □

THEOREM 2.3.12. *Let* $V \subset \mathbb{R}^M$ *be open and let* $g : V \to \mathbb{R}^M$ *be continuously differentiable. Let* $\overline{x} \subset V$ *be bounded and convex. For all non-empty* $\overline{y} \subset \overline{x}$ *we have*

$$\text{(2.3.37)} \qquad \operatorname{diam} g(\overline{y}) \leq \| g'(\overline{x}) \| \cdot \operatorname{diam} \overline{y}.$$

PROOF. The theorem is a direct consequence of theorem 2.3.7 and the definition of the diameter. □

2.3.3. The distance between sets

The set of compact, non-empty subsets of $\mathbb{R}^M$ is denoted by $\mathbb{K}$. We define the *distance* $q(\overline{x},\overline{y})$ between the sets $\overline{x},\overline{y} \in \mathbb{K}$ by

$$\text{(2.3.38)} \qquad q(\overline{x},\overline{y}) = \max\Big(\max_{x\in\overline{x}} \min_{y\in\overline{y}} \|x-y\| \,,\, \max_{y\in\overline{y}} \min_{x\in\overline{x}} \|x-y\|\Big).$$

For more general sets $\overline{x}$ and $\overline{y}$ the function $q(\overline{x},\overline{y})$ was introduced by HAUSDORFF [1914]. It was used in interval arithmetic by MOORE [1966], among others, but only for vector intervals (more accurately, for interval vectors).

THEOREM 2.3.13. *The maxima and minima in* (2.3.38) *exist. Furthermore, the function* q *is a metric, that is, it satisfies for all* $\overline{x},\overline{y},\overline{z} \in \mathbb{K}$

$$\text{(2.3.39)} \qquad q(\overline{x},\overline{y}) \geq 0, \text{ with equality if and only if } \overline{x} = \overline{y},$$

$$\text{(2.3.40)} \qquad q(\overline{x},\overline{y}) = q(\overline{y},\overline{x}),$$

$$\text{(2.3.41)} \qquad q(\overline{x},\overline{z}) \leq q(\overline{x},\overline{y}) + q(\overline{y},\overline{z})$$

(*the triangle inequality*).

PROOF. Let $\overline{x},\overline{y} \in \mathbb{K}$. For $x \in \overline{x}$, $\|x-y\|$ is a continuous function of $y \in \overline{y}$, and $\overline{y}$ is compact. Therefore

$$\min_{y\in\overline{y}} \|x-y\|$$

exists for all $x \in \overline{x}$.

Furthermore, this minimum is a continuous function of $x \in \overline{x}$, since

$$|\min_{y\in\underline{y}} \|x_1-y\| - \min_{y\in\underline{y}} \|x_2-y\| \,| \le \|x_1-x_2\|$$

for all $x_1,x_2 \in \overline{x}$. In view of the compactness of $\overline{x}$ this implies the existence of

$$\max_{x\in\underline{x}} \min_{y\in\underline{y}} \|x-y\| .$$

Similarly

$$\max_{y\in\underline{y}} \min_{x\in\underline{x}} \|x-y\|$$

exists.

For the proof that q is a metric, see HAUSDORFF [1914]. □

For $x,y \in \mathbb{R}^M$ and $\overline{x},\overline{y} \in \mathbb{K}$ we define

(2.3.42) $\quad q(\overline{x},y) = q(\overline{x},\{y\}),$

(2.3.43) $\quad q(x,\overline{y}) = q(\{x\},\overline{y}).$

THEOREM 2.3.14. *For* $\lambda \in \mathbb{R}$, $y \in \mathbb{R}^M$, $A \in \mathbb{R}^{M,M}$ *and* $\overline{x},\overline{y},\overline{z} \in \mathbb{K}$ *we have*

(2.3.44) $\quad \overline{y} \subset \overline{x} \Rightarrow q(\overline{x},\overline{y}) = \max_{x\in\underline{x}} \min_{y\in\underline{y}} \|x-y\| ,$

(2.3.45) $\quad \overline{z} \subset \overline{y} \subset \overline{x} \Rightarrow q(\overline{y},\overline{z}) \le q(\overline{x},\overline{z}),$

(2.3.46) $\quad q(\overline{x},y) = \|\overline{x}-y\| ,$

(2.3.47) $\quad q(\square\,\overline{x}, \overline{x}) \le \operatorname{diam} \overline{x} ,$

(2.3.48) $\quad q(\lambda\,\overline{x}, \lambda\,\overline{y}) = |\lambda|.q(\overline{x},\overline{y}),$

(2.3.49) $\quad q(A\,\overline{x}, A\,\overline{y}) \le \|A\| \,.\, q(\overline{x},\overline{y}),$

(2.3.50) $\quad \overline{x} + \overline{y} \in \mathbb{K}$ *and* $q(\overline{x}+\overline{y}, \overline{x}) \le \|\overline{y}\| ,$

(2.3.51) $\quad \overline{x} \subset \overline{y} + q(\overline{x},\overline{y}) \,.\, \overline{e},$

(2.3.52) $\quad \operatorname{diam} \overline{x} \le \operatorname{diam} \overline{y} + 2q(\overline{x},\overline{y}).$

PROOF. (2.3.44) follows immediately from the definition of q, and (2.3.45) is a consequence of (2.3.44).

$$q(\overline{x},y) = \max(\max_{x\in\overline{x}} \|x-y\| , \min_{x\in\overline{x}} \|x-y\|)$$

$$= \max_{x\in\overline{x}} \|x-y\| = \|\overline{x}-y\| .$$

$$q(\square\overline{x} , \overline{x}) = \max_{y\in\square\overline{x}} \min_{x\in\overline{x}} \|x-y\| \le \sup_{x,y\in\square\overline{x}} \|x-y\|$$

$$= \operatorname{diam} (\square\overline{x}) = \operatorname{diam} \overline{x}.$$

$$q(\lambda\overline{x} ,\lambda\overline{y}) = \max(\max_{x\in\overline{x}} \min_{y\in\overline{y}} \|\lambda x-\lambda y\| , \max_{y\in\overline{y}} \min_{x\in\overline{x}} \|\lambda x-\lambda y\|)$$

$$= |\lambda| \, . \, q(\overline{x},\overline{y}).$$

(2,3.49) can be proved analogously.

It is not difficult to prove $\overline{x} + \overline{y} \in \mathbb{K}$.

$$q(\overline{x} + \overline{y}, \overline{x}) = \max(\max_{\substack{x\in\overline{x}\\ y\in\overline{y}}} \min_{x_1\in\overline{x}} \|x + y - x_1\| , \max_{x\in\overline{x}} \min_{\substack{x_1\in\overline{x}\\ y\in\overline{y}}} \|x - x_1 - y\|) .$$

Choosing $x_1 = x$ we obtain

$$q(\overline{x} + \overline{y}, \overline{x}) \le \max(\max_{y\in\overline{y}} \|y\| , \min_{y\in\overline{y}} \|y\|) = \|\overline{y}\| .$$

Let $x \in \overline{x}$. We have

$$\min_{y\in\overline{y}} \|x-y\| \le q(\overline{x},\overline{y}).$$

Hence there is a $y \in \overline{y}$ such that

$$\|x-y\| \le q(\overline{x},\overline{y}).$$

Using (2.3.20) we obtain

$$x = y + (x-y) \in \overline{y} + q(\overline{x},\overline{y}) \, . \, \overline{e}.$$

Since this holds for all $x \in \overline{x}$, it proves (2.3.51).

By virtue of (2.3.51) we can derive

$$\operatorname{diam} \overline{x} \le \operatorname{diam}(\overline{y} + q(\overline{x},\overline{y})\overline{e}) \le \operatorname{diam} \overline{y} + q(\overline{x},\overline{y}) \operatorname{diam} \overline{e}$$

$$= \operatorname{diam} \overline{y} + 2q(\overline{x},\overline{y}). \qquad \square$$

LEMMA 2.3.15. *For* $\overline{x},\overline{y} \in \mathbb{IK}$ *we have*

$$(2.3.53) \quad q(\overline{x},\overline{y}) \ge \max_{1\le i\le M} q(p_i(\overline{x}), p_i(\overline{y})).$$

If $\overline{x},\overline{y} \in \amalg \mathbb{R}^M$ *then we have*

$$(2.3.54) \quad q(\overline{x},\overline{y}) = \max_{1\le i\le M} q(p_i(\overline{x}), p_i(\overline{y})).$$

PROOF. Let $\overline{x},\overline{y} \in \mathbb{IK}$.

$$\max_{x\in\overline{x}} \min_{y\in\overline{y}} \|x-y\| = \max_{x\in\overline{x}} \min_{y\in\overline{y}} \max_{1\le i\le M} |[x]_i - [y]_i|$$

$$\ge \max_{x\in\overline{x}} \max_{1\le i\le M} \min_{\eta\in p_i(\overline{y})} |[x]_i - \eta|$$

$$= \max_{1\le i\le M} \max_{\xi\in p_i(\overline{x})} \min_{\eta\in p_i(\overline{y})} |\xi-\eta|.$$

For $\overline{y} \in \amalg \mathbb{R}^M$ the inequality changes into an equality.

Similarly,

$$\max_{y\in\overline{y}} \min_{x\in\overline{x}} \|x-y\| \ge \max_{1\le i\le M} \max_{\eta\in p_i(\overline{y})} \min_{\xi\in p_i(\overline{x})} |\xi-\eta|,$$

with equality for $\overline{x} \in \amalg \mathbb{R}^M$.

Using the definition of q we thus arrive at (2.3.53) and (2.3.54). $\square$

LEMMA 2.3.16. *For* M = 1 *and* $\overline{\xi},\overline{\eta} \in \mathbb{IK}$ *we have*

$$(2.3.55) \quad q(\square\overline{\xi}, \square\overline{\eta}) \le q(\overline{\xi},\overline{\eta}).$$

PROOF. Since $\overline{\xi},\overline{\eta} \in \mathbb{IK}$ these sets have a minimum and a maximum. Let $\alpha = \min \overline{\xi}$, $\beta = \max \overline{\xi}$, $\gamma = \min \overline{\eta}$, $\delta = \max \overline{\eta}$.

$$\begin{aligned}
\max_{\xi\in\overline{\xi}} \min_{\eta\in\overline{\eta}} |\xi-\eta| &= \max_{\xi\in\overline{\xi}} \min_{\eta\in\overline{\eta}} \max(\xi-\eta,\eta-\xi) \\
&\geq \max_{\xi\in\overline{\xi}} \max(\min_{\eta\in\overline{\eta}}(\xi-\eta), \min_{\eta\in\overline{\eta}}(\eta-\xi)) \\
&= \max_{\xi\in\overline{\xi}} \max(\xi-\delta,\gamma-\xi) \\
&= \max(\beta-\delta,\gamma-\alpha).
\end{aligned}$$

Similarly we have

$$\max_{\eta\in\overline{\eta}} \min_{\xi\in\overline{\xi}} |\xi-\eta| \geq \max(\delta-\beta,\alpha-\gamma).$$

Consequently

$$\begin{aligned}
q(\overline{\xi},\overline{\eta}) &\geq \max(\alpha-\gamma,\gamma-\alpha,\beta-\delta,\delta-\beta) \\
&= \max(|\alpha-\gamma|, |\beta-\delta|).
\end{aligned}$$

It is easy to verify, and mentioned in ALEFELD & HERZBERGER [1974], that

(2.3.56) $\quad q([\alpha,\beta],[\gamma,\delta]) = \max(|\alpha-\gamma|, |\beta-\delta|)$.

Since $\square\overline{\xi} = [\alpha,\beta]$ and $\square\overline{\eta} = [\gamma,\delta]$ this proves the lemma. □

THEOREM 2.3.17. *For* $\overline{x},\overline{y} \in \mathbb{IK}$ *we have*

(2.3.57) $\quad q(\square\overline{x},\square\overline{y}) \leq q(\overline{x},\overline{y})$.

PROOF. Using (2.3.54), (2.3.55) and (2.3.53) we obtain

$$\begin{aligned}
q(\square\overline{x},\square\overline{y}) &= \max_{1\leq i\leq M} q(p_i(\square\overline{x}),p_i(\square\overline{y})) \\
&= \max_{1\leq i\leq M} q(\square p_i(\overline{x}),\square p_i(\overline{y})) \\
&\leq \max_{1\leq i\leq M} q(p_i(\overline{x}),p_i(\overline{y})) \leq q(\overline{x},\overline{y}).
\end{aligned}$$

□

THEOREM 2.3.18. *Let* $V \subset \mathbb{R}^M$ *be open and let* $g : V \to \mathbb{R}^M$ *be continuously differentiable. Let* $\overline{x} \subset \mathbb{R}^M$ *be bounded amd convex. For all* $\overline{x}_1, \overline{x}_2 \in \mathbb{K}$ *with* $\overline{x}_1 \subset \overline{x}$ *and* $\overline{x}_2 \subset \overline{x}$ *we have*

$$(2.3.58) \quad q(g(\overline{x}_1), g(\overline{x}_2)) \leq \|g'(\overline{x})\| \cdot q(\overline{x}_1, \overline{x}_2).$$

PROOF. The theorem is a consequence of theorem 2.3.7 and the definition of q. □

2.4. THE INITIAL VALUE PROBLEM

In this monograph we always consider the autonomous system of $M \geq 1$ differential equations

$$U'(t) = f(U(t)), \tag{2.4.1}$$

where $f : \mathbb{R}^M \to \mathbb{R}^M$ is a given continuously differentiable function (as is well-known, every system can be transformed into an autonomous system, see e.g. GEAR [1971]).

Let $t \geq 0$ and $x \in \mathbb{R}^M$. If the initial value problem

$$\begin{cases} U'(s) = f(U(s)) & (0 \leq s \leq t), \\ U(0) = x \end{cases} \tag{2.4.2}$$

has a solution, then it is well-known (see for instance COPPEL [1965]) that the solution is unique. In this case we denote the vector $U(t)$ by $U(t,x)$.

According to the general definitions (2.2.2) - (2.2.4) we have

$$U(t,\overline{x}) = \{U(t,x) \mid x \in \overline{x}\} \quad (\text{for } t \geq 0 \text{ and } \overline{x} \subset \mathbb{R}^M),$$

etc., provided that all the $U(t,x)$ concerned exist.

$D_t U$ and $D_x U$ denote the derivatives of the function U with respect to its first and second argument, respectively. $D_t^2 U$ means the second derivative of U with respect to its first argument, etc.

Let the set of pairs $(t,x) \in [0,\infty) * \mathbb{R}^M$ for which $U(t,x)$ exists be denoted by Q. We have the following theorems (see e.g. COPPEL [1965]).

THEOREM 2.4.1. *Q is open relative to $[0,\infty) * \mathbb{R}^M$.* □

THEOREM 2.4.2. *If f is k times continuously differentiable then $D_t^{k+1} U(t,x)$ and $D_t^k D_x U(t,x)$ exist and are continuous for $(t,x) \in Q$.* □

THEOREM 2.4.3. *For $(t_0,x) \in Q$ the matrix function $D_x U(t,x)$ satisfies the differential system*

$$\begin{cases} D_t D_x U(t,x) = f'(U(t,x)) D_x U(t,x) & (0 \leq t \leq t_0), \\ D_x U(0,x) = I. \end{cases} \tag{2.4.3}$$

□

For an arbitrary induced norm for matrices a corresponding *logarithmic norm* can be defined. This concept was introduced separately by DAHLQUIST [1959] and LOZINSKIĬ [1958]. For the norm defined in (2.3.2) the corresponding logarithmic norm of a matrix $A \in \mathbb{R}^{M,M}$ is

$$\mu[A] = \max_{1\le i\le M} \Big([A]_{ii} + \sum_{\substack{j=1\\ j\neq i}}^{M} |[A]_{ij}|\Big). \tag{2.4.4}$$

Note that $\mu[A]$ is not a norm because it can take negative values. Obviously for $A \in \mathbb{R}^{M,M}$ we have

$$\mu[A] + \mu[-A] \ge 0. \tag{2.4.5}$$

THEOREM 2.4.4. *For* $(t,x) \in Q$ *the matrix* $D_x U(t,x)$ *is regular (i.e., non-singular) and*

$$\|D_x U(t,x)\| \le \exp\Big(\int_0^t \mu[f'(U(s,x))]ds\Big), \tag{2.4.6}$$

$$\|[D_x U(t,x)]^{-1}\| \le \exp\Big(\int_0^t \mu[-f'(U(s,x))]ds\Big). \tag{2.4.7}$$

PROOF. Let $(t_0,x) \in Q$, and let $v \in \mathbb{R}^M$ be arbitrary. Define for $0 \le t \le t_0$

$$V(t) = D_x U(t,x)v. \tag{2.4.8}$$

In view of theorem 2.4.3 we have

$$V'(t) = f'(U(t,x))\, V(t) \quad (0 \le t \le t_0), \quad V(0) = v. \tag{2.4.9}$$

This implies (see e.g. COPPEL [1965])

$$\|v\| \exp\Big(-\int_0^{t_0} \mu[-f'(U(t,x))]dt\Big) \le \|V(t)\| \le \|v\| \exp\Big(\int_0^{t_0} \mu[f'(U(t,x))]dt\Big). \tag{2.4.10}$$

Since $v \in \mathbb{R}^M$ is arbitrary this proves the theorem. □

For $\overline{A} \subset \mathbb{R}^{M,M}$ the set $\mu[\overline{A}]$ is defined by

$$\mu[\overline{A}] = \{\mu[A] \mid A \in \overline{A}\}, \tag{2.4.11}$$

according to the general definition (2.2.1). Now from theorem 2.4.4 we immediately obtain the following corollary.

COROLLARY 2.4.5. *For* $(t,x) \in Q$ *we have*

$$\|D_x U(t,x)\| \leq \exp(t \max \mu[f'(U([0,t],x))]), \tag{2.4.12}$$

$$\|[D_x U(t,x)]^{-1}\| \leq \exp(t \max \mu[-f'(U([0,t],x))]). \tag{2.4.13}$$

□

2.5. MISCELLANEOUS DEFINITIONS AND PROPERTIES

In this section we will treat miscellaneous definitions, notational conventions and properties which we need in this monograph.

For all real numbers $\xi > 0$ we define $\frac{\xi}{0} = \infty$.

In addition to the definitions of minimum and maximum, given in section 2.1.2, we have of course $\min(\xi,\infty) = \xi$ for $\xi \in \mathbb{R}$, etc.

For $k > \ell$ we define

$$\sum_{i=k}^{\ell} \ldots = 0.$$

Similarly we define for instance $\sum_{j=1,j\neq i}^{M} \ldots = 0$ for $i = M = 1$.

We define the function $\omega : \mathbb{R} \to \mathbb{R}$ by

$$\omega(\xi) = \begin{cases} \dfrac{e^{\xi}-1}{\xi} & (\text{for } \xi \neq 0), \\ 1 & (\text{for } \xi = 0). \end{cases} \tag{2.5.1}$$

THEOREM 2.5.1. *For all $\xi \in \mathbb{R}$ we have*

$$\omega(\xi) > 0. \tag{2.5.2}$$

Furthermore, $\xi\omega(\alpha\xi)$ is a non-decreasing function of α and ξ.

PROOF. (2.5.2) is trivial. For $\xi \geq 0$ we have

$$\omega(\xi) = \sum_{i=0}^{\infty} \frac{\xi^i}{(i+1)!},$$

$$\omega'(\xi) = \sum_{i=1}^{\infty} \frac{i}{(i+1)!}\xi^{i-1} = \sum_{i=0}^{\infty} \frac{i+1}{(i+2)!}\xi^{i} \geq 0,$$

and for $\xi > 0$

$$\omega(-\xi) = \frac{e^{-\xi}-1}{-\xi} = e^{-\xi}\omega(\xi),$$

$$\omega'(-\xi) = -\frac{d}{d\xi}[\omega(-\xi)] = -\frac{d}{d\xi}(e^{-\xi}\omega(\xi))$$

$$= e^{-\xi}(\omega(\xi) - \omega'(\xi)) > 0$$

Thus we have

(2.5.3) $\omega'(\xi) \geq 0$ for all $\xi \in \mathbb{R}$.

Consequently

$$\frac{\partial}{\partial\alpha}\,[\xi\omega(\alpha\xi)] = \xi^2\omega'(\alpha\xi) \geq 0.$$

Finally

$$\frac{\partial}{\partial\xi}\,[\xi\omega(\alpha\xi)] = e^{\alpha\xi} > 0,$$

which proves the theorem. □

THEOREM 2.5.2. *For real* α,β,h,h_0 *with* $\beta \geq 0$, $0 \leq h \leq h_0$ *we have*

$$(2.5.4)\qquad \frac{1}{h}\log(e^{h\alpha} + h\beta) \leq \min(\alpha + \beta e^{h_0\alpha^-},\ \alpha\omega(-h_0\alpha^-) + \beta).$$

PROOF.

$$\frac{1}{h}\log(e^{h\alpha} + h\beta) = \alpha + \frac{1}{h}\log(1 + h\beta e^{-h\alpha}) \leq \alpha + \beta e^{-h\alpha}.$$

Since $h\alpha \geq -h_0\alpha^-$ we obtain

$$(2.5.5)\qquad \frac{1}{h}\log(e^{h\alpha} + h\beta) \leq \alpha + \beta e^{h_0\alpha^-}.$$

Furthermore

$$\frac{1}{h}\log(e^{h\alpha} + h\beta) = \frac{1}{h}\log(1 + h[\alpha\omega(h\alpha) + \beta]) \leq \alpha\omega(h\alpha) + \beta.$$

Since $\alpha\omega(h\alpha)$ is isotonic with respect to h we find

$$(2.5.6)\qquad \frac{1}{h}\log(e^{h\alpha} + h\beta) \leq \alpha\omega(h_0\alpha) + \beta.$$

For $\alpha \geq 0$ (2.5.4) follows from (2.5.5) and for $\alpha < 0$ it is the result of combining (2.5.5) and (2.5.6). □

THEOREM 2.5.3. *Let* $\alpha,\beta,H \in \mathbb{R}$, $\beta \geq 0$. *For* $1 \leq n \leq N$ *let* $0 < h_n \leq H$ *and define for* $0 \leq n \leq N$

$$t_n = \sum_{i=1}^{n} h_i.$$

If the real numbers $\xi_0,\xi_1,\ldots,\xi_N$ *satisfy*

$$\xi_n \leq e^{h_n\alpha}\,\xi_{n-1} + h_n\beta \qquad (1 \leq n \leq N), \quad \xi_0 = 0, \tag{2.5.7}$$

then we have

$$\xi_n \leq e^{H\alpha^-}\,t_n\omega(\alpha t_n)\beta \qquad (0 \leq n \leq N). \tag{2.5.8}$$

PROOF. From (2.5.7) we easily derive

$$\xi_n \leq \beta \sum_{i=1}^{n} h_i e^{\alpha(t_n-t_i)} \qquad (0 \leq n \leq N). \tag{2.5.9}$$

For $1 \leq i \leq n \leq N$ and $t_{i-1} \leq t \leq t_i$ we have

$$\begin{aligned}\alpha(t_n-t_i) &= \alpha(t_n-t) - \alpha(t_i-t)\\ &\leq \alpha(t_n-t) + h_i\alpha^-\\ &\leq \alpha(t_n-t) + H\alpha^-.\end{aligned}$$

Thus we obtain from (2.5.9)

$$\begin{aligned}\xi_n &\leq \beta \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} e^{\alpha(t_n-t_i)}\,dt\\ &\leq \beta \int_{0}^{t_n} e^{\alpha(t_n-t)+H\alpha^-}\,dt\\ &= \beta \cdot e^{H\alpha^-} \cdot t_n\omega(\alpha t_n).\end{aligned}$$

□

For a regular matrix $A \in \mathbb{R}^{M,M}$ the *condition*, denoted by cond A, is defined by

$$\text{cond } A = \|A\| \cdot \|A^{-1}\|. \tag{2.5.10}$$

THEOREM 2.5.4. *Let* $A,B \in \mathbb{R}^{M,M}$, *where* A *is a regular matrix, and assume*

$$\|A-B\| = \varepsilon, \tag{2.5.11}$$

$$\varepsilon\|A^{-1}\| = \theta < 1. \tag{2.5.12}$$

Then B *is regular and*

$$\text{cond } B \leq \text{cond } A + \frac{\theta}{1-\theta}(1 + \text{cond } A). \tag{2.5.13}$$

PROOF.

$$\|A^{-1}B-I\| \leq \|A^{-1}\| \cdot \|B-A\| = \theta < 1.$$

Consequently the matrix $A^{-1}B$ is regular and

$$\|(A^{-1}B)^{-1}\| \leq \frac{1}{1-\theta}$$

(see for instance STOER [1972]).
This implies that B is regular and we obtain

$$\begin{aligned}\text{cond } B = \|B\| \cdot \|B^{-1}\| &\leq \|B\| \cdot \|(A^{-1}B)^{-1}\| \cdot \|A^{-1}\| \\ &\leq \|B\| \cdot \frac{\|A^{-1}\|}{1-\theta} \leq (\|A\| + \varepsilon) \cdot \frac{\|A^{-1}\|}{1-\theta} \\ &= \frac{\text{cond } A}{1-\theta} + \frac{\theta}{1-\theta} = \text{cond } A + \frac{\theta}{1-\theta}(1 + \text{cond } A).\end{aligned}$$

□

As usual the expression

$$\phi_1(x) = \phi_2(x) + O(\psi(x)) \quad (x \to x_0) \tag{2.5.14}$$

with $x, x_0 \in \mathbb{R}^M$, $\phi_1(x), \phi_2(x), \psi(x) \in \mathbb{R}$, means that there are positive real numbers δ and K such that for all $x \in \mathbb{R}^M$ with $\|x-x_0\| \leq \delta$ we have

(2.5.15) $$|\phi_1(x) - \phi_2(x)| \leq K \cdot |\psi(x)|.$$

If we replace (2.5.14) by

(2.5.16) $$\phi_1(x) \geq \phi_2(x) + O(\psi(x)) \quad (x \to x_0)$$

then (2.5.15) has to be replaced by

(2.5.17) $$\phi_1(x) \geq \phi_2(x) - K \cdot |\psi(x)|.$$

If, for instance, $\phi_3(x) = \phi_4(x)\ \phi_1(x)$ and $\phi_1(x)$ satisfies (2.5.14), then we may write

$$\phi_3(x) = \phi_4(x) \cdot [\phi_2(x) + O(\psi(x))] \quad (x \to x_0).$$

CHAPTER 3

OUTLINE OF THE METHOD

Let the following be given:

1. an integer $M \geq 1$;
2. a continuously differentiable function $f : \mathbb{R}^M \to \mathbb{R}^M$;
3. a set $\overline{y}_0 \subset \mathbb{R}^M$ of initial values;
4. a real number $T > 0$.

Consider the differential system

$$U'(t) = f(U(t)) \qquad (0 \leq t \leq T). \tag{3.1}$$

Assume that for all $x \in \overline{y}_0$ there is a function U on $[0,T]$ satisfying (3.1) and $U(0) = x$. In other words, assume that $U(T,\overline{y}_0)$ exists.

We will deal with a numerical method of enclosing $U(T,\overline{y}_0)$. It proceeds in the following way.

For certain grid-points $0 = t_0 < t_1 < \ldots < t_N = T$ a set $\overline{y}_n$ is computed such that

$$U(t_n,\overline{y}_0) \subset \overline{y}_n, \tag{3.2}$$

consecutively for $n = 1,2,\ldots,N$. We use a one-step-method in the sense that $\overline{y}_n$ is computed from $\overline{y}_{n-1}$, without using $\overline{y}_i$ for $i < n-1$, namely such that

$$U(t_n - t_{n-1},\overline{y}_{n-1}) \subset \overline{y}_n \qquad \text{for } n = 1,2,\ldots,N.$$

From this the required inclusion (3.2) follows easily by induction.

The *step sizes*

$$h_n = t_n - t_{n-1} \qquad (1 \leq n \leq N) \tag{3.3}$$

and hence the grid-points t_n cannot be prescribed arbitrarily. They are found in the course of the process.

Thus, for a given set $\overline{y}_{n-1} \subset \mathbb{R}^M$, the n'th step of the method consists of the computation of a step size h_n and a set $\overline{y}_n \subset \mathbb{R}^M$ such that

$$U(h_n, \overline{y}_{n-1}) \subset \overline{y}_n \tag{3.4}$$

(see fig. 3.1).

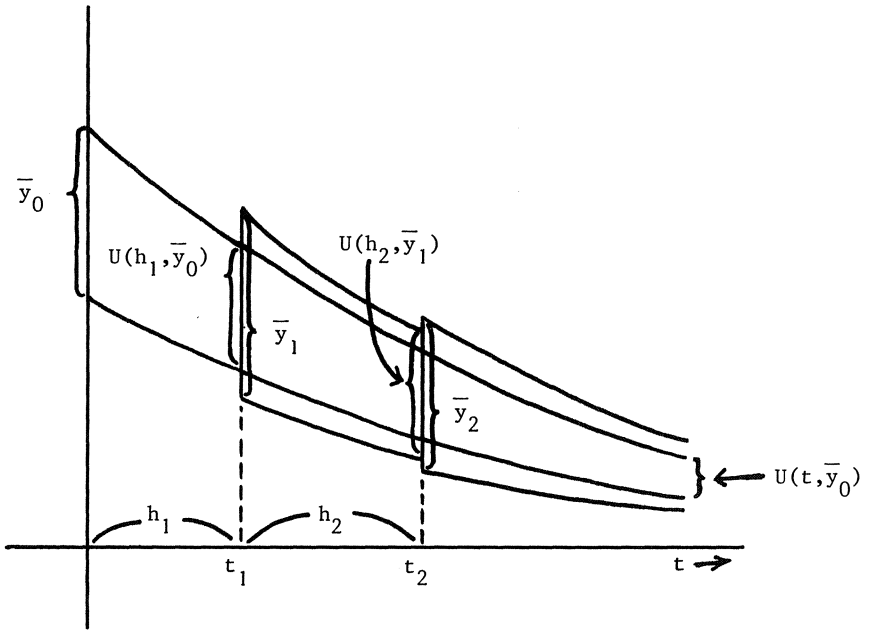

fig. 3.1

This process does not necessarily succeed in reaching the grid-point $t_N = T$ in a finite number of N steps. We may have an infinite process with $\lim_{n\to\infty} t_n \leq T$.

The existence of $U(T,\overline{y}_0)$ is a necessary condition for the finiteness of the process. However, it is not a sufficient condition. $U(T - t_n, \overline{y}_n)$ may not exist for some n, due to the fact that inclusion (3.2) is in general not an equality. Then the process is infinite all the same.

Therefore in chapter 6 we will give conditions under which we do reach the point $t_N = T$ in a finite number of N steps and hence succeed in enclosing $U(T,\overline{y}_0)$.

We note that, for our numerical method, (3.2) and (3.4) hold for all appropriate n, whether or not $U(T,\overline{y}_0)$ exists and whether or not the process is finite.

In practice the sets $\overline{y}_n$ have to be easily representable. Therefore we have to choose a suitable class $\mathbb{Y}$ of sets to which $\overline{y}_0$ is assumed to belong and from which we will choose the sets $\overline{y}_n$ $(n \geq 1)$.

An obvious choice is $\mathbb{Y} = \mathbb{I}\mathbb{R}^M$. HUNGER [1971] and MARCOWITZ [1973, 1975] use this class. However, the necessity to choose the sets $\overline{y}_n$ from this class $\mathbb{Y}$ can cause an unfavourable growth of diam $\overline{y}_n$, independently of the actual method and of the step sizes. We will show this in section 6.2 by an example.

The choice we make for the class $\mathbb{Y}$ is the following:

$$\mathbb{Y} = \{A\overline{x} \mid A \in \mathbb{R}^{M,M} \text{ is regular, } \overline{x} \in \mathbb{I}\mathbb{R}^M\}. \tag{3.5}$$

Now we can write

$$\overline{y}_n = A_n\overline{x}_n \quad (n \geq 0) \tag{3.6}$$

with a regular matrix $A_n \in \mathbb{R}^{M,M}$ and a vector interval $\overline{x}_n \in \mathbb{I}\mathbb{R}^M$.

In the n'th step we have at our disposal a regular matrix $A_{n-1} \in \mathbb{R}^{M,M}$, a vector interval $\overline{x}_{n-1} \in \mathbb{I}\mathbb{R}^M$ and a grid-point $t_{n-1} \in [0,T)$ such that

$$U(t_{n-1},\overline{y}_0) \subset \overline{y}_{n-1},$$

where $\overline{y}_0$ and $\overline{y}_{n-1}$ are defined by (3.6).

Performing the n'th step we first compute a suitable step size $h_n \in (0,T-t_{n-1}]$ and a rough inclusion $\overline{b}_n \in \mathbb{I}\mathbb{R}^M$ of the set $U([0,h_n],\overline{y}_{n-1})$. This part of the n'th step will be described and analysed in chapter 4. Using the inclusion $\overline{b}_n$ we compute a regular matrix $A_n \in \mathbb{R}^{M,M}$ and a vector interval $\overline{x}_n \in \mathbb{I}\mathbb{R}^M$ such that (3.4) holds, where $\overline{y}_n$ is defined by (3.6). This part of the n'th step will be described and analysed in chapter 5. Finally, the formal description of the global method will be given in chapter 6.

CHAPTER 4

COMPUTATION OF A SUITABLE STEP SIZE AND A ROUGH INCLUSION OF THE SOLUTION

4.1. INTRODUCTION

In this chapter we consider the first part of the n'th step of the method ($n \geq 1$). A regular matrix $A_{n-1} \in \mathbb{R}^{M,M}$, a vector interval $\overline{x}_{n-1} \in \mathbb{I}\mathbb{R}^M$ and a grid-point $t_{n-1} \in [0,T)$ are given such that for $\overline{y}_{n-1} = A_{n-1}\overline{x}_{n-1}$ we have

$$(4.1.1) \qquad U(t_{n-1},\overline{y}_0) \subset \overline{y}_{n-1}.$$

We will give an algorithm, which we will call Algorithm I, that computes a suitable step size $h_n \in (0,T-t_{n-1}]$ and a rough inclusion $\overline{b}_n \in \mathbb{I}\mathbb{R}^M$ of $U([0,h_n],\overline{y}_{n-1})$.

This algorithm contains a parameter $H_n > 0$, which may depend on n, such that the step size h_n is, if possible, approximately equal to this prescribed value H_n. This enables us to control the step size.

In general we cannot compute $f(\overline{x})$ or even $\square f(\overline{x})$ exactly for arbitrary $\overline{x} \in \mathbb{I}\mathbb{R}^M$. In practice we have available (due to the use of (rounded-) interval arithmetic, see chapter 7) a function $\overline{f}_0 : \mathbb{I}\mathbb{R}^M \to \mathbb{I}\mathbb{R}^M$ such that

$$(4.1.2) \qquad f(\overline{x}) \subset \overline{f}_0(\overline{x}) \qquad (\overline{x} \in \mathbb{I}\mathbb{R}^M).$$

Similarly we assume the availability of a function $\overline{g}_0 : \mathbb{I}\mathbb{R}^M \to \mathbb{I}\mathbb{R}^{M,M}$ satisfying

$$(4.1.3) \qquad f'(\overline{x}) \subset \overline{g}_0(\overline{x}) \qquad (\overline{x} \in \mathbb{I}\mathbb{R}^M).$$

In section 4.2 we describe Algorithm I and in section 4.3 we prove its correctness. Section 4.4 deals with bounds on the obtained step size h_n. Finally, in section 4.5 we compare our algorithm with some alternatives.

4.2. DESCRIPTION AND FINITENESS OF ALGORITHM I

In this section we describe Algorithm I and prove some of its properties, among which the finiteness of the algorithm.

Roughly speaking, Algorithm I consists of the following parts:

1. computation of a suitable vector interval to enclose U(t) (statements (4.2.1) - (4.2.6));
2. computation of a corresponding step size (statement (4.2.7));
3. if the step size is not satisfactory, reiteration of parts 1 and 2 with other values of parameter $\hat{H}$ (statements (4.2.8) - (4.2.10));
4. iterative improvement of the vector interval of part 1 (statements (4.2.11), (4.2.12)).

ALGORITHM I.

(4.2.1) $\overline{y}_{n-1} := A_{n-1}\,\overline{x}_{n-1}$.

(4.2.2) $\hat{H} := H_n$.

(4.2.3) $\hat{\alpha} := \hat{H} \,.\, \|\overline{g}_0(\square\,\overline{y}_{n-1})\|$.

(4.2.4) $$\alpha := \begin{cases} 1/10 & (\text{if } \hat{\alpha} < 1/10), \\ \hat{\alpha} & (\text{if } 1/10 \le \hat{\alpha} \le 1/2), \\ 1/2 & (\text{if } \hat{\alpha} > 1/2). \end{cases}$$

(4.2.5) $\beta := \frac{\alpha}{1-\alpha}\,\|\overline{f}_0(\square\,\overline{y}_{n-1})\|$.

(4.2.6) $\overline{b}^{(0)} := \square\,\overline{y}_{n-1} + \square\,[0,\hat{H}]\,\overline{f}_0(\square\,\overline{y}_{n-1}) + \hat{H}\beta\overline{e}$.

(4.2.7) $$\begin{cases} \hat{h} := \min\limits_{1\le i\le M} \min\left(\dfrac{\min p_i(\square\,\overline{y}_{n-1}) - \min p_i(\overline{b}^{(0)})}{[\min p_i(\overline{f}_0(\overline{b}^{(0)}))]^-}, \right. \\ \qquad\qquad \left. \dfrac{\max p_i(\overline{b}^{(0)}) - \max p_i(\square\,\overline{y}_{n-1})}{[\max p_i(\overline{f}_0(\overline{b}^{(0)}))]^+} \right) \qquad (\text{if } \overline{f}_0(\overline{b}^{(0)}) \neq \{0\}), \\ \hat{h} := H_n \qquad (\text{if } \overline{f}_0(\overline{b}^{(0)}) = \{0\}). \end{cases}$$

(4.2.8) $$\left\{\begin{array}{l}\underline{\text{if}}\\ (\hat{H} = H_n \ \underline{\text{and}}\ \hat{h} < \tfrac{1}{2}H_n) \ \underline{\text{or}}\ (\hat{H} < H_n \ \underline{\text{and}}\ h_{old} < \hat{h} < \hat{H})\\ \underline{\text{then}}\\ \underline{\text{begin}}\ \hat{H} := \hat{H}/2;\ h_{old} := \hat{h};\ \overline{b}_{old} := \overline{b}^{(0)};\ \underline{\text{goto}}\ (4.2.3);\ \underline{\text{end}}.\end{array}\right.$$

(4.2.9) $$\left\{\begin{array}{l}\underline{\text{if}}\ \hat{H} < H_n \ \underline{\text{and}}\ h_{old} \geq \hat{h}\ \underline{\text{then}}\\ \\ \underline{\text{begin}}\ \hat{H} := 2\hat{H};\ \hat{h} := h_{old};\ \overline{b}^{(0)} := \overline{b}_{old};\ \underline{\text{end}}.\end{array}\right.$$

(4.2.10) $h_n := \min(\hat{h}, H_n, T-t_{n-1})$.

(4.2.11) $\overline{b}^{(i)} := \overline{b}^{(i-1)} \cap [\ \square\ \overline{y}_{n-1} + \square\ [0,h_n]\ \overline{f}_0(\overline{b}^{(i-1)})]\ (i = 1,2,\ldots,i_n)$.

(4.2.12) $\overline{b}_n := \overline{b}^{(i_n)}$.

For i_n we choose the smallest value of $i \geq 1$ satisfying

(4.2.13) $\operatorname{diam} p_j(\overline{b}^{(i)}) \geq 0.9 \operatorname{diam} p_j(\overline{b}^{(i-1)})$ (for all j with $1 \leq j \leq M$),

but we maximally choose $i_n = 10$. In other words, we continue the iteration process (4.2.11) as long as the diameter of the real interval $p_j(\overline{b}^{(i)})$, for any j, is decreased by more than 10%, or until 10 iteration steps have been performed.

NOTE. In practice we calculate $\square\ \overline{y}_{n-1}$ directly from A_{n-1} and $\overline{x}_{n-1}$, instead of first calculating $\overline{y}_{n-1}$ and then $\square\ \overline{y}_{n-1}$ (which would be impossible). However, for convenience in the interpretation and analysis of our method we formulate the algorithm using the set $\overline{y}_{n-1}$, defined by (4.2.1). Because this set only occurs in the combination $\square\ \overline{y}_{n-1}$ in the algorithm, no practical difficulties can arise.

THEOREM 4.2.1. *The value of* $\hat{h}$ *is well-defined by* (4.2.7). *Moreover we have* $0 < \hat{h} < \infty$.

PROOF. Assume

(4.2.14) $\beta = 0$.

By virtue of (4.2.5) we have

$$(4.2.15)\qquad \overline{f}_0(\square\, \overline{y}_{n-1}) = \{0\}.$$

Consequently, using (4.2.6), we get $\overline{b}^{(0)} = \square\, \overline{y}_{n-1}$. Using (4.2.15) again we obtain

$$(4.2.16)\qquad \overline{f}_0(\overline{b}^{(0)}) = \{0\}.$$

Hence (4.2.14) implies (4.2.16).

Now assume

$$(4.2.17)\qquad \overline{f}_0(\overline{b}^{(0)}) \neq \{0\},$$

then $\beta > 0$. Hence, using (4.2.6), we obtain that $\square\, \overline{y}_{n-1}$ is in the interior of $\overline{b}^{(0)}$. Consequently the numerators occurring in (4.2.7) are all positive.

From (4.2.17) we conclude that for some i, $1 \le i \le M$, we have

$$p_i(\overline{f}_0(\overline{b}^{(0)})) \neq \{0\}$$

and therefore

$$\max p_i(\overline{f}_0(\overline{b}^{(0)})) > 0 \text{ or } \min p_i(\overline{f}_0(\overline{b}^{(0)})) < 0.$$

Hence the denominators occurring in (4.2.7) are not all zero. Moreover they are all non-negative.

Combining this with the result that the numerators are all positive yields that $\hat{h}$ is well-defined and that $0 < \hat{h} < \infty$. □

THEOREM 4.2.2. *Algorithm* I *is finite.*

PROOF. With each iteration of the statements (4.2.3) - (4.2.8) the value of $\hat{H}$ is halved, see (4.2.8). Therefore after a finite number of iterations we have $\hat{h} \le h_{old}$ (that is, $\hat{h}$ does not increase any more) or $\hat{h} \ge \hat{H}$. In both cases the iteration process is stopped. □

4.3. CORRECTNESS OF ALGORITHM I.

For methods that only *approximate* the solution of an initial value problem one is interested in the applicability (for instance for implicit methods) and the error. However, for methods that *enclose* solutions of initial value problems, it is essential to consider something else as well, namely the correctness. For the method of this chapter, this means that the following theorem holds.

THEOREM 4.3.1. *If the pair* $(h_n, \overline{b}_n)$ *is produced by Algorithm* I *then the set* $U([0,h_n], \overline{y}_{n-1})$ *exists and*

$$U([0,h_n], \overline{y}_{n-1}) \subset \overline{b}_n. \tag{4.3.1}$$

Before we give the proof of this theorem, we formulate a lemma.

LEMMA 4.3.2. *Each time that* (4.2.7) *is executed, the value* $\hat{h}$ *obtained by this statement is the greatest value of* h *satisfying*

$$\square \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\overline{b}^{(0)}) \subset \overline{b}^{(0)}, \tag{4.3.2}$$

provided that $\overline{f}_0(\overline{b}^{(0)}) \neq \{0\}$.

PROOF. (4.3.2) holds if and only if

$$\sup(\square \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\overline{b}^{(0)})) \leq \max \overline{b}^{(0)} \tag{4.3.2a}$$

and

$$\inf(\square \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\overline{b}^{(0)})) \geq \min \overline{b}^{(0)}. \tag{4.3.2b}$$

Let $1 \leq i \leq M$.

$$[\sup(\square \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\overline{b}^{(0)}))]_i =$$

$$[\max(\square \overline{y}_{n-1}) + \sup([0,h]\, \overline{f}_0(\overline{b}^{(0)}))]_i =$$

$$\max p_i(\square \overline{y}_{n-1}) + \sup p_i([0,h]\, \overline{f}_0(\overline{b}^{(0)})) =$$

$$\max p_i(\square \overline{y}_{n-1}) + \max([0,h]\, p_i(\overline{f}_0(\overline{b}^{(0)}))) =$$

$$\max p_i(\square \bar{y}_{n-1}) + h.[\max p_i(\bar{f}_0(\bar{b}^{(0)}))]^+.$$

Therefore (4.3.2a) holds if and only if

$$h \le \frac{\max p_i(\bar{b}^{(0)}) - \max p_i(\square \bar{y}_{n-1})}{[\max p_i(\bar{f}_0(\bar{b}^{(0)}))]^+} \qquad (1 \le i \le M).$$

Analogously (4.3.2b) holds if and only if

$$h \le \frac{\min p_i(\square \bar{y}_{n-1}) - \min p_i(\bar{b}^{(0)})}{[\min p_i(\bar{f}_0(\bar{b}^{(0)}))]^-} \qquad (1 \le i \le M).$$

Combining these results yields that (4.3.2) holds if and only if $h \le \hat{h}$, which proves the lemma. □

PROOF OF THEOREM 4.3.1. Let $\hat{h}$ and $\bar{b}^{(0)}$ have their final values.

Define

$$Z = \{U \mid U : [0,\hat{h}] \to \mathbb{R}^M, \ U \text{ continuous}\}, \tag{4.3.3}$$

$$X = \{U \mid U \in Z, \ U(t) \in \bar{b}^{(0)} \ (0 \le t \le \hat{h})\}. \tag{4.3.4}$$

Let $y \in \bar{y}_{n-1}$. Define $S : X \to Z$ by

$$(SU)(t) = y + \int_0^t f(U(s))ds \qquad (0 \le t \le \hat{h}). \tag{4.3.5}$$

By virtue of lemma 4.3.2, (4.3.2) holds for $h = \hat{h}$. Hence, using (2.2.63) and (4.1.2), we have

$$S(X) \subset X. \tag{4.3.6}$$

Define on Z the norm

$$\|U\| = \max_{0 \le t \le \hat{h}} (e^{-\hat{L}t}\|U(t)\|) \tag{4.3.7}$$

where

$$\hat{L} = \|f'(\bar{b}^{(0)})\|. \tag{4.3.8}$$

It is easy to verify that, with respect to this norm, S is a contraction mapping and X is closed (compare e.g. COPPEL [1965]). By virtue of the Contraction Mapping Theorem of Banach (see e.g. SMART [1974]) we may conclude that S has a fixed point. Consequently, $U([0,\hat{h}],y)$ exists and

$$U([0,\hat{h}],y) \subset \overline{b}^{(0)}. \tag{4.3.9}$$

This holds for all $y \in \overline{y}_{n-1}$, and $h_n \leq \hat{h}$. Hence

$$U([0,h_n], \overline{y}_{n-1}) \subset \overline{b}^{(0)}. \tag{4.3.10}$$

By induction (4.3.1) follows easily from (4.3.10), (4.2.11), (2.2.63) and (4.2.12). □

4.4. BOUNDS ON THE OBTAINED STEP SIZE

In this section we will give bounds on the obtained step size h_n, especially in relation to the prescribed value H_n, which is the step size we aim at.

Define for an arbitrary set $\overline{x} \subset \mathbb{R}^M$

$$(4.4.1) \qquad K(\overline{x}) = \sup \left\{ \frac{q(\overline{f}_0(\overline{y}),\overline{f}_0(\overline{z}))}{q(\overline{y},\overline{z})} \;\middle|\; \overline{y},\overline{z} \in \mathbb{I}\!\mathbb{I}\,\mathbb{R}^M;\ \overline{z} \subset \overline{y} \subset \overline{x};\ \overline{z} \neq \overline{y} \right\},$$

where $\sup \emptyset = 0$. Define

$$(4.4.2) \qquad K_0 = K(\overline{b}^{(0)}), \text{ with } \overline{b}^{(0)} \text{ defined by (4.2.2) - (4.2.6).}$$

Assume

$$(4.4.3) \qquad K_0 < \infty .$$

This will generally be true in practice for a reasonable choice of the function $\overline{f}_0$.

Let $\hat{H} = H_n$ and let $\hat{\alpha},\alpha,\beta,\overline{b}^{(0)}$ and $\hat{h}$ have the values corresponding to this value of $\hat{H}$. In view of (4.4.1), (4.4.2), (2.3.51) and (2.3.22) we have

$$(4.4.4) \qquad \overline{f}_0(\overline{b}^{(0)}) \subset \overline{f}_0(\square\, \overline{y}_{n-1}) + K_0 \cdot q(\overline{b}^{(0)}, \square\, \overline{y}_{n-1}) \cdot \overline{e}.$$

From (4.2.6), (2.3.50), (4.2.5) and (4.2.4) we have

$$\begin{aligned} q(\overline{b}^{(0)}, \square\, \overline{y}_{n-1}) &\leq \| \square\, [0,H_n]\, \overline{f}_0(\square\, \overline{y}_{n-1}) + H_n \beta\, \overline{e} \| \\ &\leq H_n \cdot \| \overline{f}_0(\square\, \overline{y}_{n-1}) \| + H_n \beta \\ &\leq H_n \cdot (\beta \cdot \frac{1-\alpha}{\alpha} + \beta) \\ &= H_n \cdot \frac{\beta}{\alpha} . \end{aligned}$$

Combining this with (4.4.4) yields

(4.4.5) $\quad \overline{f}_0(\overline{b}^{(0)}) \subset \overline{f}_0(\square\, \overline{y}_{n-1}) + K_0 \cdot H_n \cdot \frac{\beta}{\alpha} \cdot \overline{e}.$

Define

(4.4.6) $\quad h = \min\left(\frac{\alpha}{K_0}, H_n\right).$

By using (4.4.5), (4.2.6) and the subdistributivity property (2.2.32) we obtain the relations

$$\square\, \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\overline{b}^{(0)})$$

$$\subset \square\, \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\square\, \overline{y}_{n-1}) + [0,h] \cdot K_0 \cdot H_n \cdot \frac{\beta}{\alpha} \cdot \overline{e}$$

$$= \square\, \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\square\, \overline{y}_{n-1}) + H_n \beta \frac{hK_0}{\alpha}\, \overline{e}$$

$$\subset \square\, \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\square\, \overline{y}_{n-1}) + H_n \beta\, \overline{e}$$

$$\subset \overline{b}^{(0)}.$$

Consequently

$$\square\, \overline{y}_{n-1} + [0,h]\, \overline{f}_0(\overline{b}^{(0)}) \subset \overline{b}^{(0)}.$$

By virtue of lemma 4.3.2 and (4.2.7) we find $\hat{h} \geq h$ or $\hat{h} = H_n$, hence using definition (4.4.6) we have

(4.4.7) $\quad \hat{h} \geq \min\left(\frac{\alpha}{K_0}, H_n\right).$

By virtue of (4.2.8) and (4.2.9) the final value of $\hat{h}$ is not smaller than the first value. Therefore, using (4.2.10), we obtain

(4.4.8) $\quad h_n \geq \min(\hat{h}, H_n, T - t_{n-1})$ for the first value of $\hat{h}$.

Combining this with (4.4.7) we find

(4.4.9) $\quad h_n \geq \min\left(\frac{\alpha}{K_0}, H_n, T - t_{n-1}\right).$

In view of (4.2.4) we have $\alpha \geq \frac{1}{10}$, thus arriving at the following theorem.

THEOREM 4.4.1. *Assume that* $K_0 < \infty$, *where* K_0 *is defined by* (4.4.2). *Let* α *be defined by* (4.2.2), (4.2.3) *and* (4.2.4). *We have*

$$(4.4.10) \quad h_n \geq \min(\frac{\alpha}{K_0}, H_n, T-t_{n-1}) \geq \min(\frac{1}{10K_0}, H_n, T-t_{n-1}). \qquad \square$$

In chapter 6 we will use this result to prove the applicability of our method on the whole interval [0,T].

Define

$$(4.4.11) \quad K_1 = \|\overline{g}_0(\square\, \overline{y}_{n-1})\|,$$

then we have the following theorem.

THEOREM 4.4.2. *Let* K_0 *be defined by* (4.4.2) *and* K_1 *by* (4.4.11). *Assume* $K_0 < \infty$,

$$(4.4.12) \quad H_n \leq \frac{1}{2K_1}$$

and

$$(4.4.13) \quad t_{n-1} + h_n < T.$$

Then

$$(4.4.14) \quad \min\left(\frac{K_1}{K_0}, 1\right) \leq \frac{h_n}{H_n} \leq 1, \qquad \textit{where } \min(\tfrac{0}{0}, 1) = 1.$$

PROOF. Let $\hat{\alpha}$ and α be defined by (4.2.2), (4.2.3) and (4.2.4). By virtue of (4.2.3), (4.2.2) and (4.4.11) we have

$$(4.4.15) \quad \hat{\alpha} = H_n K_1.$$

Using (4.4.12) this implies $\hat{\alpha} \leq \frac{1}{2}$. Therefore, by virtue of (4.2.4) we have $\alpha \geq \hat{\alpha}$. Combine this with (4.4.15) and apply theorem 4.4.1, then we find

$$(4.4.16) \quad h_n \geq \min\left(\frac{H_n K_1}{K_0}, H_n, T-t_{n-1}\right).$$

Using (4.4.13) and (4.2.10) this proves the theorem. $\square$

If we define $\overline{f}_0$ by

$$(4.4.17) \quad \overline{f}_0(\overline{x}) = \square\, f(\overline{x}) \qquad (\overline{x} \in \amalg \mathbb{R}^M)$$

then in view of (4.1.2) and (2.1.33) the set $\overline{f}_0(\overline{x})$ is as small as possible for any $\overline{x} \in \amalg \mathbb{R}^M$. If f is a simple function, the function $\overline{f}_0$, which we must be able to evaluate practically, can often be chosen according to this definition.

THEOREM 4.4.3. *Let the function* $\overline{f}_0$ *be defined by* (4.4.17) *and* $\overline{b}^{(0)}$ *by* (4.2.2) - (4.2.6). *Assume that* (4.4.12) *and* (4.4.13) *hold, where* K_1 *is defined by* (4.4.11). *Then*

$$(4.4.18) \quad \min\left(\frac{\|\overline{g}_0(\square\, \overline{y}_{n-1})\|}{\|\overline{g}_0(\overline{b}^{(0)})\|}, 1 \right) \le \frac{h_n}{H_n} \le 1, \quad \textit{where } \min\left(\frac{0}{0}, 1\right) = 1.$$

PROOF. Let $\overline{y}, \overline{z} \in \amalg \mathbb{R}^M$, $\overline{y}, \overline{z} \subset \overline{b}^{(0)}$. Using (2.3.57) and (2.3.58) we find

$$\begin{aligned} q(\overline{f}_0(\overline{y}), \overline{f}_0(\overline{z})) &= q(\square\, f(\overline{y}), \square\, f(\overline{z})) \\ &\le q(f(\overline{y}), f(\overline{z})) \\ &\le \|f'(\overline{b}^{(0)})\| \cdot q(\overline{y}, \overline{z}). \end{aligned}$$

By virtue of (4.4.1) this implies $K_0 \le \|f'(\overline{b}^{(0)})\|$, where K_0 is defined by (4.4.2). Using (4.1.3) and (2.3.7) we find

$$(4.4.19) \quad K_0 \le \|\overline{g}_0(\overline{b}^{(0)})\|.$$

Applying theorem 4.4.2 we find (4.4.14). Combining this with (4.4.19) and (4.4.11) yields (4.4.18) and the theorem has been proved. □

INTERPRETATION OF THEOREM 4.4.3. *If* H_n *is small, then the sets* $\overline{b}^{(0)}$ *and* $\square\, \overline{y}_{n-1}$ *will not differ very much in general. Consequently we have*

$$(4.4.20) \quad \|\overline{g}_0(\overline{b}^{(0)})\| \approx \|\overline{g}_0(\square\, \overline{y}_{n-1})\|.$$

Therefore (4.4.18) implies

$$(4.4.21) \quad h_n \approx H_n,$$

which was our aim. □

COROLLARY 4.4.4. *Let the function* $\overline{g}_0$ *be defined by* $\overline{g}_0(\overline{x}) = \overline{C}$ $(\overline{x} \in \amalg \mathbb{R}^M)$, *where* $\overline{C} \in \amalg \mathbb{R}^{M,M}$, *and the function* $\overline{f}_0$ *by* (4.4.17). *Assume*

$$H_n \leq \frac{1}{2\|\overline{C}\|}$$

and $t_{n-1} + h_n < T$. *Then* $h_n = H_n$. □

If the function f is such that the set $f'(\mathbb{R}^M)$ is bounded and can practically be enclosed in a matrix interval, then the function $\overline{g}_0$ can be chosen to be constant, and corollary 4.4.4 can be applied.

EXAMPLE 4.4.5.

$B \in \mathbb{R}^{M,M}$, $c \in \mathbb{R}^M$.

$f(x) = Bx + c \qquad (x \in \mathbb{R}^M)$.

$\overline{f}_0(\overline{x}) = \square B\overline{x} + c \qquad (\overline{x} \in \amalg \mathbb{R}^M)$.

$\overline{g}_0(\overline{x}) = \{B\} \qquad (\overline{x} \in \amalg \mathbb{R}^M)$.

If

$$H_n \leq \frac{1}{2\|B\|}$$

and $t_{n-1} + h_n < T$ then $h_n = H_n$. □

4.5. COMPARISON WITH OTHER METHODS

4.5.1. Variants of Algorithm I

Let Algorithm I' be obtained from Algorithm I by leaving out statements (4.2.8) and (4.2.9). We will compare both algorithms and thus explain why these statements have been built in in Algorithm I. First we give a theorem showing that Algorithm I is, in a certain sense, at least as good as Algorithm I'. Then we will give example 4.5.2 showing that Algorithm I can be considerably better than Algorithm I'. Finally we motivate the presence of the term $\hat{H}\beta\overline{e}$ in (4.2.6).

THEOREM 4.5.1. *Let* h_n^* *and* $\overline{b}^{(0)*}$ *be the values of* h_n *and* $\overline{b}^{(0)}$ *respectively, produced by Algorithm* I'. *We have*

$$h_n^* \leq h_n \leq H_n \tag{4.5.1}$$

and

$$\overline{b}^{(0)} \subset \overline{b}^{(0)*}. \tag{4.5.2}$$

PROOF. According to (4.2.10) we have

$$h_n^* = \min(\hat{h}, H_n, T-t_{n-1}) \text{ for the first value of } \hat{h}. \tag{4.5.3}$$

Combining this with (4.4.8) yields $h_n \geq h_n^*$. Furthermore, by virtue of (4.2.10) we have $h_n \leq H_n$ and (4.5.1) has been proved.

In Algorithm I the final value of $\overline{b}^{(0)}$ is computed according to (4.2.3) - (4.2.6) with a value of $\hat{H}$ satisfying $\hat{H} \leq H_n$. In Algorithm I' the value $\hat{H} = H_n$ is used in (4.2.3) - (4.2.6). Therefore, using (4.2.2) - (4.2.6), we obtain (4.5.2) and the theorem has been proved. □

EXAMPLE 4.5.2.

$$M = 2$$

$$f\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}\right) = \begin{pmatrix} \xi_2^2 \\ 1 \end{pmatrix} \qquad (\xi_1, \xi_2 \in \mathbb{R}).$$

$$\overline{f}_0\left(\begin{pmatrix} \overline{\xi}_1 \\ \overline{\xi}_2 \end{pmatrix}\right) = \begin{pmatrix} (\overline{\xi}_2)^2 \\ 1 \end{pmatrix} \qquad (\overline{\xi}_1, \overline{\xi}_2 \in \mathbb{IR}).$$

$$\overline{g}_0\left(\left(\begin{array}{c}\overline{\xi}_1\\ \overline{\xi}_2\end{array}\right)\right) = \begin{pmatrix} 0 & 2\overline{\xi}_2 \\ 0 & 0 \end{pmatrix} \qquad (\overline{\xi}_1, \overline{\xi}_2 \in \mathbb{II}\,\mathbb{R}).$$

$$\overline{y}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The exact solution is

$$U(t) = \begin{pmatrix} \frac{1}{3}t^3 \\ t \end{pmatrix} \qquad (t \geq 0).$$

Let $n = 1$, hence $t_{n-1} = 0$, and let $T = 1$.

For any positive value of $\hat{H}$, the statements (4.2.3) - (4.2.7) give the following results.

$$\hat{\alpha} = 0,\ \alpha = \frac{1}{10},\ \beta = \frac{1}{9} \cdot \left\| \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\| = \frac{1}{9},$$

$$\overline{b}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \square\,[0,\hat{H}] \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \beta \hat{H} \overline{e} = \hat{H} \begin{pmatrix} [-\beta,\beta] \\ [-\beta, 1+\beta] \end{pmatrix},$$

$$\overline{f}_0(\overline{b}^{(0)}) = \begin{pmatrix} \hat{H}^2 \cdot [0,(1+\beta)^2] \\ 1 \end{pmatrix}.$$

$$\hat{h} = \min\left(\frac{\beta\hat{H}}{0}, \frac{\beta\hat{H}}{(1+\beta)^2\hat{H}^2}, \frac{\beta\hat{H}}{0}, \frac{(1+\beta)\hat{H}}{1} \right),$$

hence

$$\hat{h} = \min\left(\frac{\beta}{(1+\beta)^2\hat{H}}, (1+\beta)\hat{H} \right). \tag{4.5.4}$$

Substituting $\beta = \frac{1}{9}$ we obtain

$$\hat{h} = \min\left(\frac{9}{100\,\hat{H}}, \frac{10}{9}\hat{H} \right). \tag{4.5.5}$$

Thus we see that for large values of $\hat{H}$ the value of $\hat{h}$ is very small, and that halving such a large value of $\hat{H}$ causes a doubling of $\hat{h}$. Consequently, if H_n is large then h'_n is very small but h_n can be much larger than h'_n. Take for instance $H_n = 1$, then we have $h'_n = 0.09$, and executing Algorithm I we have successively $\hat{H} = 1$, $\hat{h} = 0.09$, $\hat{H} = 0.5$, $\hat{h} = 0.18$, $\hat{H} = 0.25$, $\hat{h} \approx 0.278$, and $h_n \approx 0.278$. Of course the quotient h_n/h'_n can even be larger if H_n is larger. □

Let Algorithm I" be obtained from Algorithm I by replacing (4.2.6) by

$$\overline{b}^{(0)} := \square\, \overline{y}_{n-1} + \square\, [0,\hat{H}]\, \overline{f}_0\, (\square\, \overline{y}_{n-1}),$$

or equivalently, by replacing (4.2.5) by $\beta := 0$. Let us apply this algorithm to the initial value problem of example 4.5.2. By substituting $\beta = 0$ in (4.5.4) we obtain $\hat{h} = 0$ for any value of $\hat{H}$. In particular, for $\hat{H} = H_n$ and for $\hat{H} = H_n/2$ we have $\hat{h} = 0$. Consequently $h_n = 0$, which would render the n'th step of the method (and all succeeding steps) useless. This shows the necessity of the term $\hat{H}\beta\,\overline{e}$ in (4.2.6).

4.5.2. Moore's method

Along the lines of MOORE [1966], p. 102, we can formulate the following alternative algorithm to compute the step size h_n.

(4.5.6) $\hat{H} := H_n$.

(4.5.7) $\overline{a} := \square\, \overline{y}_{n-1} + \square\, [0,\hat{H}]\, \overline{f}_0(\square\, \overline{y}_{n-1})$.

(4.5.8) if $\square\, \overline{y}_{n-1} + \square\, [0,\hat{H}]\, \overline{f}_0(\overline{a}) \subset \overline{a}$ then goto (4.5.13).

(4.5.9) $\overline{a} := \square\, \overline{y}_{n-1} + \square\, [0,\hat{H}]\, \overline{f}_0(\overline{a})$.

(4.5.10) if $\square\, \overline{y}_{n-1} + \square\, [0,\hat{H}]\, \overline{f}_0(\overline{a}) \subset \overline{a}$ then goto (4.5.13).

(4.5.11) $\hat{H} := \hat{H}/2$.

(4.5.12) goto (4.5.8).

(4.5.13) $h_n := \min(\hat{H}, T - t_{n-1})$.

Moore expects such an algorithm to be finite. However, we will give an example showing that this need not to be true.

EXAMPLE 4.5.3.

$M = 1$.

$f(x) = x^2 \quad (x \in \mathbb{R}), \quad \overline{f}_0(\overline{x}) = \overline{x}^2 \quad (\overline{x} \in \mathbb{I}\mathbb{R})$.

$$\overline{y}_0 \doteq \{1\}, \qquad n = 1, \quad H_1 = 1.$$

Write $a = \max \overline{a}$, then a necessary condition for

$$\square\, \overline{y}_{n-1} + \square\, [0,\hat{H}]\, \overline{f}_0(\overline{a}) \subset \overline{a}$$

is $1 + \hat{H}a^2 \leq a$, and therefore $\hat{H}a < 1$. For the values of $\hat{H}$ and $\overline{a}$ occurring in the course of the execution of the algorithm we have successively

$$\begin{aligned} \hat{H} = 1,\ a = 2,\ a &> 4,\ \hat{H} = 1/2, \\ a &> 8,\ \hat{H} = 1/4, \\ a &> 16,\ \hat{H} = 1/8, \\ a &> 32,\ \hat{H} = 1/16, \text{ etc.}, \end{aligned}$$

and the process never ends. □

CHAPTER 5

COMPUTATION OF THE FINAL INCLUSION OF THE SOLUTION

5.1. INTRODUCTION

In this chapter we consider the second part of the n'th step of the method ($n \geq 1$) which we will call Algorithm II.

A regular matrix $A_{n-1} \in \mathbb{R}^{M,M}$, a vector interval $\overline{x}_{n-1} \in \mathbb{I}\mathbb{R}^M$ and a grid-point $t_{n-1} \in [0,T)$ are given such that

$$U(t_{n-1},\overline{y}_0) \subset \overline{y}_{n-1}, \tag{5.1.1}$$

where

$$\overline{y}_{n-1} = A_{n-1}\overline{x}_{n-1}. \tag{5.1.2}$$

Furthermore, a step size $h_n \in (0,T-t_{n-1}]$ and a vector interval $\overline{b}_n \in \mathbb{I}\mathbb{R}^M$ are given such that $U(h_n,\overline{y}_{n-1})$ exists and we have the rough inclusion

$$U([0,h_n],\overline{y}_{n-1}) \subset \overline{b}_n. \tag{5.1.3}$$

We want to compute a regular matrix $A_n \in \mathbb{R}^{M,M}$ and a vector interval $\overline{x}_n \in \mathbb{I}\mathbb{R}^M$ such that

$$U(h_n,\overline{y}_{n-1}) \subset \overline{y}_n, \tag{5.1.4}$$

where

$$\overline{y}_n = A_n\overline{x}_n. \tag{5.1.5}$$

Our method contains an integer parameter $k \geq 2$. We assume that f is (k-1) times continuously differentiable.

Define $f_i : \mathbb{R}^M \to \mathbb{R}^M$ $(i = 0,1,\ldots,k-1)$ recursively by

$$(5.1.6) \qquad \begin{cases} f_0 = f, \\ f_i(x) = f'_{i-1}(x)f(x) \quad (x \in \mathbb{R}^M,\ i = 1,2,\ldots,k-1). \end{cases}$$

We assume that we can evaluate $f_i(x)$ and $f'_i(x)$ for $0 \leq i \leq k-2$ and every $x \in \mathbb{R}^M$.

Furthermore, we assume that we have available a function $\overline{f}_{k-1} : \mathbb{IR}^M \to \mathbb{IR}^M$ satisfying

$$(5.1.7) \qquad f_{k-1}(\overline{x}) \subset \overline{f}_{k-1}(\overline{x}) \quad (\overline{x} \in \mathbb{IR}^M)$$

and functions $\overline{g}_i : \mathbb{IR}^M \to \mathbb{IR}^{M,M}$ $(0 \leq i \leq k-2)$ satisfying

$$(5.1.8) \qquad f'_i(\overline{x}) \subset \overline{g}_i(\overline{x}) \quad (\overline{x} \in \mathbb{IR}^M,\ 0 \leq i \leq k-2).$$

Note that the function $\overline{g}_0$ has already been introduced (see section 4.1).

REMARK 5.1.1. For a large class of functions f the functions $\overline{f}_{k-1}$ and $\overline{g}_i$ $(0 \leq i \leq k-2)$ can be evaluated recursively. In that case explicit formulas for $\overline{f}_{k-1}$ and $\overline{g}_i$ (which may be very complicated) are not required (see MOORE [1966], chapter 11 and [1979], section 3.4). □

5.2. DESCRIPTION AND FINITENESS OF ALGORITHM II

In this section we describe Algorithm II and prove its finiteness.

ALGORITHM II.

(5.2.1) $\hat{y}_{n-1} := \text{mean}(\square\, \overline{y}_{n-1})$.

(5.2.2) $S_n := I + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} f'_{i-1}(\hat{y}_{n-1})$.

(5.2.3) $A_n := S_n A_{n-1}$.

(5.2.4) $$\overline{c} := \sum_{i=1}^{k-1} \frac{h_n^i}{i!} [f_{i-1}(\hat{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1})\hat{y}_{n-1}]$$
$$+ \square \left[\sum_{i=1}^{k-1} \frac{h_n^i}{i!} [\overline{g}_{i-1}(\square\, \overline{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1})] \right] (\square\, \overline{y}_{n-1} - \hat{y}_{n-1})$$
$$+ \frac{h_n^k}{k!} \overline{f}_{k-1}(\overline{b}_n).$$

(5.2.5) if A_n is singular then begin $h_n := h_n/2$; goto (5.2.2); end.

(5.2.6) $\overline{x}_n := \overline{x}_{n-1} + \square\, A_n^{-1}\, \overline{c}$.

REMARK 5.2.1. Statement (5.2.5), with A_n replaced by S_n, could have been inserted between (5.2.2) and (5.2.3). However, in practice the singularity of A_n is detected during the execution of statement (5.2.6). Therefore (5.2.5) has been placed immediately before (5.2.6). □

THEOREM 5.2.2. *Algorithm* II *is finite.*

PROOF. If $h_n \to 0$ then $\|S_n - I\| \to 0$ for the matrix S_n that corresponds to h_n according to (5.2.2). Therefore after a finite number of halvings of h_n we have $\|S_n - I\| < 1$. Then S_n is regular.

Furthermore the matrix A_{n-1} is regular. Therefore A_n is regular and the process of halving h_n is stopped. □

REMARK. 5.2.3. In the rest of this chapter the variables h_n, S_n, A_n and $\overline{c}$ refer to their final values in the course of the execution of Algorithm II. Note that (5.1.3) remains valid if h_n is interpreted according to this rule. □

5.3. CORRECTNESS OF ALGORITHM II

In this section we will prove the correctness of Algorithm II in the sense of the following theorem.

THEOREM 5.3.1. *For* h_n, A_n *and* $\overline{x}_n$ *produced by Algorithm* II *we have*

$$U(h_n,\overline{y}_{n-1}) \subset \overline{y}_n, \tag{5.3.1}$$

where

$$\overline{y}_n = A_n\overline{x}_n. \tag{5.3.2}$$

PROOF. Let $y \in \overline{y}_{n-1}$ be arbitrary. By virtue of (5.1.3), $U(h_n,y)$ exists and

$$U([0,h_n],y) \subset \overline{b}_n. \tag{5.3.3}$$

By induction we easily see that $U(t,y)$ is k times continuously differentiable with respect to t and that we have

$$\frac{\partial^i}{\partial t^i} U(t,y) = f_{i-1}(U(t,y)) \quad (0 \le t \le h_n,\ 1 \le i \le k). \tag{5.3.4}$$

Expanding $U(t,y)$ in a Taylor series with respect to t we find

$$\begin{cases} U(h_n,y) = y + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} D_t^i U(0,y) + R \\ \text{where } R \in \mathbb{R}^M \text{ is such that for } 1 \le j \le M \text{ we have} \\ [R]_j = \frac{h_n^k}{k!} [D_t^k U(s_j,y)]_j \text{ for some } s_j \in (0,h_n). \end{cases} \tag{5.3.5}$$

Using (5.3.4) we find that the vector R satisfies

$$[R]_j = \frac{h_n^k}{k!} [f_{k-1}(U(s_j,y))]_j \quad (1 \le j \le M). \tag{5.3.6}$$

By virtue of (5.1.3) we have

$$U(s_j,y) \in \overline{b}_n,$$

and therefore, using (5.1.7),

$$f_{k-1}(U(s_j,y)) \in f_{k-1}(\overline{b}_n) \subset \overline{f}_{k-1}(\overline{b}_n).$$

Consequently, using (5.3.6), we have

$$[R]_j \in \frac{h_n^k}{k!} p_j(\overline{f}_{k-1}(\overline{b}_n)) \quad (1 \le j \le M).$$

By virtue of (2.2.52), (2.1.11) and $\overline{f}_{k-1}(\overline{b}_n) \in \amalg \mathbb{R}^M$ we find

$$R \in \frac{h_n^k}{k!} \overline{f}_{k-1}(\overline{b}_n). \tag{5.3.7}$$

Using (5.3.5) and (5.3.4) yields

$$\begin{aligned} U(h_n,y) &= y + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} f_{i-1}(U(0,y)) + R \\ &= y + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} f_{i-1}(y) + R. \end{aligned}$$

Applying (2.2.66) and using $y,\hat{y}_{n-1} \in \square\, \overline{y}_{n-1}$ we find

$$\begin{aligned} U(h_n,y) \in\ & \hat{y}_{n-1} + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} f_{i-1}(\hat{y}_{n-1}) \\ & + [I + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} \square\, f'_{i-1}(\square\, \overline{y}_{n-1})](y-\hat{y}_{n-1}) + R. \end{aligned}$$

By virtue of (5.1.8) and (2.1.33) this yields

$$\begin{aligned} U(h_n,y) \in\ & \hat{y}_{n-1} + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} f_{i-1}(\hat{y}_{n-1}) \\ & + [I + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} \overline{g}_{i-1}(\square\, \overline{y}_{n-1})](y-\hat{y}_{n-1}) + R. \end{aligned}$$

Using (5.2.2), (2.2.30), (2.2.31) and $y \in \square\, \overline{y}_{n-1}$ we find

$$U(h_n,y) \in \hat{y}_{n-1} + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} f_{i-1}(\hat{y}_{n-1})$$

$$+ S_n(y - \hat{y}_{n-1})$$

$$+ \Big(\sum_{i=1}^{k-1} \frac{h_n^i}{i!} [\overline{g}_{i-1}(\square\, \overline{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1})]\Big)(y - \hat{y}_{n-1}) + R$$

$$\subset S_n y + \sum_{i=1}^{k-1} \frac{h_n^i}{i!} [f_{i-1}(\hat{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1})\hat{y}_{n-1}]$$

$$+ \Big(\sum_{i=1}^{k-1} \frac{h_n^i}{i!} [\overline{g}_{i-1}(\square\, \overline{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1})]\Big)(\square\, \overline{y}_{n-1} - \hat{y}_{n-1}) + R.$$

Combining this with (5.3.7) and (5.2.4) yields

(5.3.8) $\quad U(h_n,y) \in S_n y + \overline{c} \quad$ for all $y \in \overline{y}_{n-1}$.

Consequently, using (5.1.2), (5.2.3), (2.2.30), (5.2.6) and (5.3.2) we have

$$U(h_n,\overline{y}_{n-1}) \subset S_n\, \overline{y}_{n-1} + \overline{c}$$

$$= S_n A_{n-1}\, \overline{x}_{n-1} + \overline{c}$$

$$= A_n\, \overline{x}_{n-1} + \overline{c}$$

$$= A_n[\overline{x}_{n-1} + A_n^{-1}\, \overline{c}]$$

$$\subset A_n[\overline{x}_{n-1} + \square\, A_n^{-1}\, \overline{c}]$$

$$= A_n\, \overline{x}_n = \overline{y}_n,$$

and the theorem has been proved. □

5.4. THE LOCAL ERROR

In the n'th step of the method we compute a set $\overline{y}_n$ for which (5.1.4) holds. Of course we prefer $\overline{y}_n$ to be a narrow inclusion of $U(h_n,\overline{y}_{n-1})$. In order to measure how good (that is, how narrow) the inclusion is, we introduce the *local error* of the n'th step, defined by

$$\varepsilon_n = q(\overline{y}_n,\ U(h_n,\overline{y}_{n-1})). \tag{5.4.1}$$

Thus we use the Hausdorff distance between the enclosing and the enclosed set. Note that $U(h_n,\overline{y}_{n-1})$ is compact because of the compactness of $\overline{y}_{n-1}$ and the continuity of $U(h_n,x)$ with respect to x.

We will formulate a theorem that gives an upper bound of ε_n. In preparation for this theorem we give some definitions.

Define for $\overline{x} \subset \mathbb{R}^M$, $\overline{x} \neq \emptyset$,

$$\begin{cases} L(\overline{x}) = \sup\limits_{\substack{\underline{y}\subset\underline{x} \\ \overline{y}\in\mathbb{I}\mathbb{R}^M}} \dfrac{\operatorname{diam} \overline{f}_{k-1}(\overline{y})}{\operatorname{diam} \overline{y}} \\ \text{and} \\ L_i(\overline{x}) = \sup\limits_{\substack{\underline{y}\subset\underline{x} \\ \overline{y}\in\mathbb{I}\mathbb{R}^M}} \dfrac{\operatorname{diam} \overline{g}_i(\overline{y})}{\operatorname{diam} \overline{y}} \qquad (0 \le i \le k-2), \end{cases} \tag{5.4.2}$$

where $\frac{0}{0} = 0$. Furthermore, define

$$\delta_{n-1} = \operatorname{diam} \overline{y}_{n-1}. \tag{5.4.3}$$

THEOREM 5.4.1. *Assume*

$$L(\overline{b}_n) < \infty \tag{5.4.4}$$

and

$$L_i(\square\, \overline{y}_{n-1}) < \infty \quad (0 \le i \le k-2). \tag{5.4.5}$$

Then

$$\varepsilon_n \le \tfrac{1}{2}(1+\operatorname{cond} A_n)\Big[\sum_{i=1}^{k-1} \frac{1}{i!} L_{i-1}(\square\overline{y}_{n-1})\delta_{n-1}^2 h_n^i + \frac{1}{k!} L(\overline{b}_n) h_n^k \operatorname{diam} \overline{b}_n\Big]. \tag{5.4.6}$$

PROOF. Using (5.2.1), (2.3.36) and (2.3.31) we derive

$$\| \square \bar{y}_{n-1} - \hat{y}_{n-1} \| = \| \square \bar{y}_{n-1} - \text{mean}(\square \bar{y}_{n-1}) \|$$

$$= \tfrac{1}{2} \operatorname{diam}(\square \bar{y}_{n-1}) = \tfrac{1}{2} \operatorname{diam} \bar{y}_{n-1}.$$

Hence with definition (5.4.3) we find

$$\| \square \bar{y}_{n-1} - \hat{y}_{n-1} \| \le \tfrac{1}{2} \delta_{n-1}. \tag{5.4.7}$$

Using (5.2.4), (2.3.32), (2.3.31), (2.3.28), (5.4.2), (2.3.14), (2.3.12), (2.3.13), (2.3.29) and (5.4.7) we derive

$$\begin{aligned}
\operatorname{diam} \bar{c} &\le \operatorname{diam} \left(\left[\sum_{i=1}^{k-1} \frac{h_n^i}{i!} [\bar{g}_{i-1}(\square \bar{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1})] \right] (\square \bar{y}_{n-1} - \hat{y}_{n-1}) \right) \\
&\quad + \operatorname{diam} \left(\frac{h_n^k}{k!} \bar{f}_{k-1}(\bar{b}_n) \right) \\
&\le 2 \left\| \left[\sum_{i=1}^{k-1} \frac{h_n^i}{i!} [\bar{g}_{i-1}(\square \bar{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1})] \right] (\square \bar{y}_{n-1} - \hat{y}_{n-1}) \right\| \\
&\quad + \frac{h_n^k}{k!} L(\bar{b}_n) \operatorname{diam} \bar{b}_n \\
&\le 2 \cdot \left(\sum_{i=1}^{k-1} \frac{h_n^i}{i!} \| \bar{g}_{i-1}(\square \bar{y}_{n-1}) - f'_{i-1}(\hat{y}_{n-1}) \| \right) \cdot \| \square \bar{y}_{n-1} - \hat{y}_{n-1} \| \\
&\quad + \frac{h_n^k}{k!} L(\bar{b}_n) \operatorname{diam} \bar{b}_n \\
&\le 2 \cdot \left(\sum_{i=1}^{k-1} \frac{h_n^i}{i!} \operatorname{diam} \bar{g}_{i-1}(\square \bar{y}_{n-1}) \right) \cdot \tfrac{1}{2} \delta_{n-1} + \frac{h_n^k}{k!} L(\bar{b}_n) \operatorname{diam} \bar{b}_n.
\end{aligned}$$

Consequently, using (5.4.2) and (5.4.3) we have

$$\operatorname{diam} \bar{c} \le \sum_{i=1}^{k-1} \frac{h_n^i}{i!} L_{i-1}(\square \bar{y}_{n-1}) \cdot \delta_{n-1}^2 + \frac{h_n^k}{k!} L(\bar{b}_n) \operatorname{diam} \bar{b}_n. \tag{5.4.8}$$

By virtue of (5.4.1), (2.3.44) and (5.3.1) we have

$$\varepsilon_n = \max_{u\in\overline{y}_n} \min_{y\in\overline{y}_{n-1}} \|u-U(h_n,y)\|.$$

Combining this with (5.3.2), (5.2.6) and (5.3.8) we find

$$\varepsilon_n = \max_{x\in\overline{x}_n} \min_{y\in\overline{y}_{n-1}} \|A_n x - U(h_n,y)\|$$

$$\leq \max_{\substack{x_1\in\overline{x}_{n-1} \\ z\in\square A_n^{-1}\overline{c}}} \min_{y\in\overline{y}_{n-1}} \max_{c\in\overline{c}} \|A_n(x_1+z) - S_n y - c\|.$$

Choosing $y = A_{n-1}x_1$, defining $\hat{c}$ = mean $\overline{c}$ and using (5.2.3), (2.3.12), (2.2.30), (2.2.49), (2.3.14), (2.3.16) and (2.3.36) we obtain

$$\varepsilon_n \leq \max_{\substack{x_1\in\overline{x}_{n-1} \\ z\in\square A_n^{-1}\overline{c}}} \max_{c\in\overline{c}} \|A_n(x_1+z) - S_n A_{n-1}x_1 - c\|$$

$$= \max_{z\in\square A_n^{-1}\overline{c}} \max_{c\in\overline{c}} \|A_n z - c\|$$

$$= \|(A_n\square A_n^{-1}\overline{c} - \hat{c}) - (\overline{c}-\hat{c})\|$$

$$\leq \|A_n(\square A_n^{-1}\overline{c} - A_n^{-1}\hat{c})\| + \|\overline{c}-\hat{c}\|$$

$$= \|A_n\square(A_n^{-1}\overline{c} - A_n^{-1}\hat{c})\| + \|\overline{c}-\hat{c}\|$$

$$= \|A_n\square A_n^{-1}(\overline{c}-\hat{c})\| + \|\overline{c}-\hat{c}\|$$

$$\leq \|A_n\| \cdot \|\square A_n^{-1}(\overline{c}-\hat{c})\| + \|\overline{c}-\hat{c}\|$$

$$= \|A_n\| \cdot \|A_n^{-1}(\overline{c}-\hat{c})\| + \|\overline{c}-\hat{c}\|$$

$$\leq (\|A_n\| \cdot \|A_n^{-1}\| + 1) \cdot \|\overline{c}-\hat{c}\|$$

$$=(1 + \text{cond } A_n)\ \tfrac{1}{2}\ \text{diam } \overline{c}.$$

Combining this with (5.4.8) we obtain (5.4.6) and the theorem has been proved. □

5.5. COMPARISON WITH OTHER METHODS

5.5.1. Moore's method

In this section we will compare our method in some respects with the method of MOORE [1966], chapters 10 and 13.

Using our notation we can say that in MOORE [1966], like in our method, the set $U(t_n,\overline{y}_0)$ is enclosed in a set $\overline{y}_n = A_n\overline{x}_n$, with A_n a regular matrix and $\overline{x}_n \in \amalg \mathbb{R}^M$. The matrix A_n is still computed by $A_n = S_n A_{n-1}$ (i.e., according to (5.2.3)), but Moore chooses for S_n the matrix

$$S_n = I + h_n f'(\hat{y}_{n-1}) \tag{5.5.1}$$

(cf. (5.2.2)).

We will show by means of example 5.5.1 that this choice of S_n can cause the local error to be essentially greater than is ever possible in our method, independently of the actual method for computing $\overline{x}_n$.

EXAMPLE 5.5.1.

1. $$M = 2,\ B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},\ f(x) = Bx,\ A_0 = I,$$

$$\overline{y}_0 = \overline{x}_0 = \begin{pmatrix} [0,\delta] \\ [0,\delta] \end{pmatrix},\ \text{where } \delta > 0.$$

The general solution of the differential system is given by

$$U(t,x) = C(t)x,$$

where

$$C(t) = e^t \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

2. Let S_1 be defined by (5.5.1). Let $\overline{y}_1 = S_1\overline{x}_1$, with $\overline{x}_1 \in \amalg \mathbb{R}^M$, be an inclusion of $U(h_1,\overline{y}_0)$, produced by any method.

Consider the local error ε_1.

$$\varepsilon_1 = q(\overline{y}_1, U(h_1,\overline{y}_0))$$

$$= q(S_1\overline{x}_1, C(h_1)\overline{x}_0)$$

$$\geq \frac{1}{\|S_1^{-1}\|}\, q(\bar{x}_1, S_1^{-1}C(h_1)\bar{x}_0),$$

where the last inequality holds by virtue of (2.3.49). Since $\bar{x}_1 \supset S_1^{-1}C(h_1)\bar{x}_0$ and $\bar{x}_1 \in \amalg \mathbb{R}^M$, we have by virtue of (2.1.33) and (2.3.45)

$$\varepsilon_1 \geq \frac{1}{\|S_1^{-1}\|}\, q(\square\, S_1^{-1}C(h_1)\,\bar{x}_0,\ S_1^{-1}\,C(h_1)\bar{x}_0). \tag{5.5.2}$$

$$S_1 = I + h_1 f'(\hat{y}_0) = I + h_1 B = \begin{pmatrix} 1+h_1 & h_1 \\ 0 & 1+h_1 \end{pmatrix}.$$

$$S_1^{-1} = (1+h_1)^{-1} \begin{pmatrix} 1 & -h_1(1+h_1)^{-1} \\ 0 & 1 \end{pmatrix}.$$

$$S_1^{-1}C(h_1) = e^{h_1}(1+h_1)^{-1} \begin{pmatrix} 1 & h_1^2(1+h_1)^{-1} \\ 0 & 1 \end{pmatrix}.$$

By virtue of (5.5.2), (2.2.56) and (2.3.48) this implies

$$\varepsilon_1 \geq \frac{1}{\|S_1^{-1}\|}\, e^{h_1}(1+h_1)^{-1}\, q(\square\, Q\bar{x}_0,\ Q\bar{x}_0), \tag{5.5.3}$$

where

$$Q = \begin{pmatrix} 1 & q_0 \\ 0 & 1 \end{pmatrix} \text{ and } q_0 = h_1^2(1+h_1)^{-1}.$$

$$\square\, Q\bar{x}_0 = \square \begin{pmatrix} 1 & q_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} [0,\delta] \\ [0,\delta] \end{pmatrix} = \begin{pmatrix} (1+q_0)[0,\delta] \\ [0,\delta] \end{pmatrix},$$

hence

$$\begin{pmatrix} (1+q_0)\delta \\ 0 \end{pmatrix} \in \square\, Q\bar{x}_0.$$

Consequently,

$$q(\square\, Q\bar{x}_0, Q\bar{x}_0) \geq \min_{x \in \bar{x}_0} \left\| \begin{pmatrix} (1+q_0)\delta \\ 0 \end{pmatrix} - Qx \right\|$$

$$\doteq \min_{\substack{0\leq\xi_1\leq\delta \\ 0\leq\xi_2\leq\delta}} \max(|(1+q_0)\delta - (\xi_1 + q_0\xi_2)|, |\xi_2|)$$

$$= \min_{\substack{0\leq\xi_1\leq\delta \\ 0\leq\xi_2\leq\delta}} \max(q_0(\delta-\xi_2) + (\delta-\xi_1), \xi_2)$$

$$= \min_{0\le\xi_2\le\delta} \max(q_0(\delta-\xi_2),\xi_2)$$

$$= \frac{q_0}{1+q_0} \cdot \delta$$

$$= \delta h_1^2 \,(1 + O(h_1)) \qquad (h_1 \to 0).$$

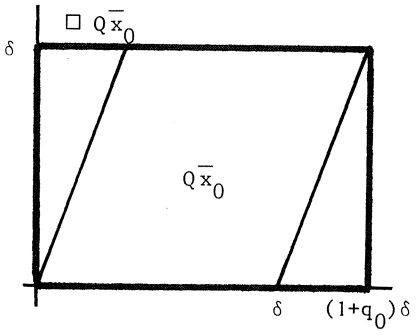

fig. 5.5.1

Furthermore,

$$\frac{1}{\|S_1^{-1}\|} e^{h_1}(1+h_1)^{-1} = 1 + O(h_1) \qquad (h_1 \to 0).$$

Therefore we obtain from (5.5.3)

(5.5.4) $\qquad \varepsilon_1 \ge \delta h_1^2(1 + O(h_1)) \qquad (h_1 \to 0).$

3. Now consider the local error ε_1 for our method, defined in sections 4.2 and 5.2. Under weak conditions it can be proved that

(5.5.5) $\qquad \operatorname{diam} \bar{b}_1 = \delta + O(h_1) \qquad (h_1,\delta \to 0)$

and we may derive from (5.4.6)

(5.5.6) $\qquad \varepsilon_1 = O(\delta^2 h_1 + \delta h_1^k + h_1^{k+1}) \qquad (h_1,\delta \to 0),$

or, equivalently,

(5.5.7) $\qquad \varepsilon_1 = O(\delta^2 h_1 + h_1^{k+1}) \qquad (h_1,\delta \to 0).$

The proof of (5.5.5), (5.5.6) and (5.5.7) and the precise conditions will not be given here, since these matters are treated in chapter 6 as part of the analysis of the global method (cf. (6.1.45), (6.1.54) and (6.1.55)).

If we put for instance $\delta = h_1^{3/2}$ and let $h_1 \to 0$ and hence also $\delta \to 0$, then (5.5.4) implies

$$\varepsilon_1 \ge h_1^{7/2} + O(h_1^{9/2}),$$

while we have for our method, in view of (5.5.6) or (5.5.7),

$$\varepsilon_1 = O(h_1^4)$$

if we take $k > 2$.

This shows that for $k > 2$ the local error ε_1 does not satisfy (5.5.6) or (5.5.7), whenever S_1 is defined by (5.5.1). For $k = 2$ the definitions (5.2.2) and (5.5.1) coincide. □

Concluding the comparison of the method of Moore with our method we observe that the former requires the inclusion of the set of inverses of all matrices in a given matrix interval, namely in

$$\square\,(I + \square\,[0,h_n]f'(\hat{y}_{n-1}))A_{n-1},$$

while the latter only requires the inversion of one matrix A_n. The first computation is more complicated, especially if some matrix in the interval is close to a singular matrix. Furthermore, if the matrix interval contains a singular matrix, then the computation is even impossible. This eventuality is more likely to occur than the singularity of our matrix A_n.

5.5.2. Krückeberg's method

In this section we will compare the method of KRÜCKEBERG [1969] with our method, and show by means of an example that the local error for the method of Krückeberg can be essentially greater than is ever possible for our method.

Let us consider the n'th step of his method, omitting details irrelevant for our purpose and using our own notation. As Krückeberg explains, there are a large number of numerical realizations for his method. However, we will try to make our considerations, as far as possible, independent of the realization actually chosen.

Like in our method, we have in the n'th step at our disposal a set $\overline{y}_{n-1} \in \mathbb{Y}$ (where $\mathbb{Y}$ is defined by (3.5)) and a grid-point $t_{n-1} \in [0,T)$ such that

$$U(t_{n-1},\overline{y}_0) \subset \overline{y}_{n-1}. \tag{5.5.8}$$

The n'th step is composed of three parts, process (I), process (II) and process (III).

Process (I) contains a parameter k. Let us choose k = 0, which, according to Krückeberg, is sufficient, and which is the only value for which process (III) is described. Then process (I) consists of the computation of a step size $h_n > 0$ and a vector interval $\overline{b}_n \in \mathbb{I}\mathbb{R}^M$ such that

$$U([0,h_n],\overline{y}_{n-1}) \subset \overline{b}_n. \tag{5.5.9}$$

For this process Krückeberg refers to MOORE [1966]. In section 4.5.2 we compared our Algorithm I with this process of Moore.

Process (II) consists of finding an inclusion of the vector $U(h_n,\hat{y}_{n-1})$, for some fixed vector $\hat{y}_{n-1} \in \overline{y}_{n-1}$.

In process (III) a matrix interval $\overline{S}_n$ is computed such that

$$U(h_n,\overline{y}_{n-1}) - U(h_n,\hat{y}_{n-1}) \in \overline{S}_n(\overline{y}_{n-1} - \hat{y}_{n-1}). \tag{5.5.10}$$

This $\overline{S}_n$ is defined by an infinite series. The precise computation of $\overline{S}_n$ is not specified, but for our comparison it suffices that $\overline{S}_n$ satisfies

$$\overline{S}_n = I + h_n\overline{g}_0(\overline{b}_n) + \frac{h_n^2}{2} \square[\overline{g}_0(\overline{b}_n)][\overline{g}_0(\overline{b}_n)] + \overline{R}, \tag{5.5.11}$$

where $\overline{R} \in \mathbb{I}\mathbb{R}^{M,M}$ depends on h_n and satisfies $\|\overline{R}\| = O(h_n^3)$, and the function $\overline{g}_0 : \mathbb{I}\mathbb{R}^M \to \mathbb{I}\mathbb{R}^{M,M}$ satisfies (4.1.3), just as for our method.

Combining the results of process (II) and process (III) Krückeberg obtains an inclusion $\overline{y}_n \in \mathbb{Y}$ of $U(h_n,\overline{y}_{n-1})$. This $\overline{y}_n$ satisfies

$$\overline{y}_n \supset U(h_n,\hat{y}_{n-1}) + \overline{S}_n(\overline{y}_{n-1} - \hat{y}_{n-1}). \tag{5.5.12}$$

Note that for process (III) the existence and availability of f' is necessary. Whether higher derivatives of f have to exist and be available depends on the actual realization of process (II).

EXAMPLE 5.5.2. M = 1, $f(x) = x^2$, $\overline{y}_0 = [1,1+\delta]$, where $\delta > 0$.

The general solution of the differential equation is given by

$$U(t,x) = \frac{1}{x^{-1}-t}.$$

Therefore

$$U(h_1,\overline{y}_0) = \frac{1}{[1,1+\delta]^{-1} - h_1} = \left[\frac{1}{1 - h_1} , \frac{1}{(1+\delta)^{-1} - h_1} \right]$$

$$= \left[\frac{1}{1 - h_1} , \frac{1 + \delta}{1 - h_1(1+\delta)} \right]$$

and

$$\operatorname{diam} U(h_1,\overline{y}_0) = \frac{1 + \delta}{1 - h_1(1+\delta)} - \frac{1}{1 - h_1}$$

$$= (1+\delta)[1 + h_1(1+\delta) + h_1^2(1+\delta)^2 + h_1^3]$$

$$- (1 + h_1 + h_1^2 + h_1^3) + O((\delta + h_1)^4)$$

$$= \delta + h_1(2\delta+\delta^2) + h_1^2 \cdot 3\delta + O(\delta+h_1)^4) \quad (\delta \to 0,\ h_1 \to 0).$$

Furthermore,

$$\overline{b}_1 \supset U([0,h_1],\overline{y}_0) = \frac{1}{[1,1+\delta]^{-1} - [0,h_1]}$$

$$= \left[1, \frac{1}{(1+\delta)^{-1} - h_1}\right] = \left[1,\ 1 + \delta + h_1 + O((\delta+h_1)^2)\right] \ (\delta \to 0,\ h_1 \to 0)$$

and

$$\overline{g}_0(\overline{b}_1) \supset f'(\overline{b}_1) = 2\overline{b}_1.$$

By virtue of (5.5.12), (2.3.35) and (5.5.11) we have in Krückeberg's method

$$\operatorname{diam} \overline{y}_1 \geq \operatorname{diam} [\overline{S}_1(\overline{y}_0 - \hat{y}_0)]$$

$$\geq |\overline{S}_1| \operatorname{diam} (\overline{y}_0 - \hat{y}_0) = |\overline{S}_1| \operatorname{diam} \overline{y}_0$$

$$= \delta \cdot |\overline{S}_1|$$

$$= \delta(1 + h_1 \max \overline{g}_0(\overline{b}_1) + \frac{h_1^2}{2} [\max \overline{g}_0(\overline{b}_1)]^2)$$

$$+ O((\delta+h_1)^4)$$

$$\geq \delta(1 + 2h_1(1 + \delta + h_1) + 2h_1^2) + O((\delta + h_1)^4)$$

$$= \delta + h_1(2\delta + 2\delta^2) + h_1^2 \cdot 4\delta + O((\delta + h_1)^4) \quad (\delta \to 0,\ h_1 \to 0).$$

In view of (2.3.52) the local error ε_1 in Krückeberg's method satisfies

$$2\varepsilon_1 \geq \operatorname{diam} \overline{y}_1 - \operatorname{diam} U(h_1, [1, 1+\delta])$$

$$\geq \delta^2 h_1 + \delta h_1^2 + O((\delta + h_1)^4) \quad (\delta \to 0,\ h_1 \to 0).$$

Hence

$$\varepsilon_1 \geq \tfrac{1}{2}\delta^2 h_1 + \tfrac{1}{2}\delta h_1^2 + O((\delta + h_1)^4) \quad (\delta \to 0,\ h_1 \to 0).$$

If we put for instance $\delta = h_1^{3/2}$ and let $h_1 \to 0$ and hence also $\delta \to 0$, then we conclude

$$\varepsilon_1 \geq \tfrac{1}{2}h_1^{7/2} + O(h_1^4).$$

We have for our method, in view of (5.5.7),

$$\varepsilon_1 = O(h_1^4)$$

if we take $k > 2$.

This shows that for Krückeberg's method the local error can be essentially larger than is ever possible for our method. □

CHAPTER 6

THE GLOBAL BEHAVIOUR OF THE METHOD

6.1. THE GLOBAL METHOD, ITS APPLICABILITY AND THE GLOBAL ERROR

In this chapter we will discuss the global method. The n'th step of this method is mainly composed of Algorithm I and Algorithm II, described in chapters 4 and 5, respectively.

Throughout this section we will assume the following to be given:

1. integers $M \geq 1$ and $k \geq 2$;
2. a (k-1) times continuously differentiable function $f : \mathbb{R}^M \to \mathbb{R}^M$;
3. functions $\overline{f}_0 : \mathbb{I}\mathbb{R}^M \to \mathbb{I}\mathbb{R}^M$, $\overline{f}_{k-1} : \mathbb{I}\mathbb{R}^M \to \mathbb{I}\mathbb{R}^M$ and $\overline{g}_i : \mathbb{I}\mathbb{R}^M \to \mathbb{I}\mathbb{R}^{M,M}$ $(0 \leq i \leq k-2)$, satisfying (4.1.2), (5.1.7) and (5.1.8) respectively;
4. a regular matrix $A_0 \in \mathbb{R}^{M,M}$;
5. a real number $T > 0$.

If we further assume that a vector interval $\overline{x}_0 \in \mathbb{I}\mathbb{R}^M$ is given, as well as values $H_n > 0$ for all appropriate $n \geq 1$, then the global method can be described as follows.

THE GLOBAL METHOD.

```
            t_0 := 0 ;  n := 0 ;

newstep :   n := n + 1;
            Algorithm I;
            Algorithm II;
            t_n := t_n-1 + h_n;

            if t_n < T then goto newstep.
```

Our method produces a grid-point t_n, a regular matrix A_n and a vector interval $\overline{x}_n$ for any $n \geq 1$, as long as we do not have $t_{n-1} = T$. For these t_n, A_n and $\overline{x}_n$ we have

$$(6.1.1) \qquad U(t_n,\overline{y}_0) \subset \overline{y}_n,$$

where, as previously defined,

$$(6.1.2) \qquad \overline{y}_i = A_i\overline{x}_i \text{ for all appropriate } i \geq 0.$$

Just as the narrowness of inclusion (5.1.4) is measured by the local error, defined by (5.4.1), we measure the narrowness of inclusion (6.1.1) by the *global error*, defined by

$$(6.1.3) \qquad \gamma_n = q(\overline{y}_n , U(t_n,\overline{y}_0)) \text{ for all appropriate } n \geq 0.$$

Thus in analogy to the local case we use the Hausdorff distance between the enclosing and the enclosed set.

We will formulate a theorem that gives an upper bound on γ_n.

As we pointed out in chapter 3, the assumption that $U(T,\overline{y}_0)$ exists does not simply imply that the global method is capable of reaching the grid-point $t_N = T$ in a finite number of N steps. Therefore we will give conditions under which the global method is applicable on the whole interval $[0,T]$.

The problems of applicability and bounding the global error are treated in one theorem because of their strong relationship.

THEOREM 6.1.1 (*about the applicability and the global error of the method*).

Let $\hat{x}_0 \in \mathbb{R}^M$ *and define* $\hat{y}_0 = A_0\hat{x}_0$. *Assume that* $U(T,\hat{y}_0)$ *exists. Furthermore, assume that* $K(\overline{x})$, $L(\overline{x})$ *and* $L_i(\overline{x})(0 \leq i \leq k-2)$, *defined by* (4.4.1) *and* (5.4.2), *as well as* $\sup\{\|\overline{f}_0(\overline{y})\| \,|\, \overline{y} \in \square\mathbb{R}^M, \overline{y} \subset \overline{x}\}$, *are finite for every nonempty bounded set* $\overline{x} \subset \mathbb{R}^M$.

Then there are values $\Delta > 0$, $H > 0$, $\beta_1 \geq 0$ *and* $\beta_2 \geq 0$ *such that conclusions I and II are valid whenever the following assumptions A and B hold.*

ASSUMPTIONS.

A. $\bar{x}_0 \in \amalg \mathbb{R}^M$, mean $\bar{x}_0 = \hat{x}_0$,
and with $\bar{y}_0 = A_0\bar{x}_0$ *and*

(6.1.4) $\quad \delta = \operatorname{diam} \bar{y}_0$

we have

(6.1.5) $\quad \delta \le \Delta.$

B.

(6.1.6) $\quad H_- \le H_n \le H$ *for all appropriate* n,

for some fixed value $H_- > 0$.

CONCLUSIONS.

I. The global method is applicable on the whole interval [0,T], *reaching the grid-point* $t_N = T$ *in a finite number of* N *steps.*

II. The global error γ_n, *defined by* (6.1.3), *satisfies*

(6.1.7) $\quad \gamma_n \le \beta_1\delta^2 + \beta_2(h_{max})^k \le \beta_1\delta^2 + \beta_2(H_{max})^k \quad (0 \le n \le N),$

where

(6.1.8) $\quad h_{max} = \max_{1\le n\le N} h_n, \qquad H_{max} = \max_{1\le n\le N} H_n.$

The values of Δ, H, β_1 *and* β_2 *can be defined according to* (6.1.9) - (6.1.28).

REMARK 6.1.2. Both as an extension of theorem 6.1.1 and in preparation for its proof we define a combination of possible values for Δ, H, β_1 and β_2 using a number of auxiliary sets and variables.

Let $\nu_1, \nu_2 > 0$ be arbitrary (we could have defined here for example $\nu_1 = 1$, $\nu_2 = 1$; however, our approach is somewhat more general and gives values of Δ, H, β_1 and β_2 depending on these parameters).

(6.1.9) $\quad \bar{Y} = U([0,T],\hat{y}_0) + \nu_1\bar{e},$

where $\bar{e}$ is defined by (2.3.18).

$$\bar{B} = \square\, \bar{Y} + \nu_2 \bar{e}. \tag{6.1.10}$$

$$\alpha_1 = \max \mu[f'(\bar{B})], \tag{6.1.11}$$

where μ denotes the logarithmic norm, defined by (2.4.4). Because of the compactness of the set $\bar{B}$ and the continuity of the functions f' and μ, the set $\mu[f'(\bar{B})]$ is compact and therefore has a maximum.

$$\alpha_2 = \max \mu[-f'(\bar{B})], \tag{6.1.12}$$

where the existence of the maximum is shown analogously to that of the maximum in (6.1.11).

$$\alpha_3 = \frac{1}{(k-1)!}\, \nu_3^{k-2}\, e^{\nu_3 \alpha_2^+} \left(M \cdot \| f'_{k-2}(\bar{B}) \| e^{\nu_3 \alpha_1^+} + \| f'_{k-2}(\bar{Y}) \|\right), \tag{6.1.13}$$

where $\nu_3 > 0$ is chosen such that we have

$$\alpha_3 \nu_3 < 1 \tag{6.1.14}$$

and

$$2\nu_3 \sup\{\| \bar{f}_0(\bar{x}) \| \mid \bar{x} \in \amalg\, \mathbb{R}^M,\ \bar{x} \subset \square\, \bar{Y}\} \le \nu_2. \tag{6.1.15}$$

$$\alpha_4 = \alpha_1 + \alpha_2 + \frac{\alpha_3}{1 - \alpha_3 \nu_3} \left(1 + e^{\nu_3(\alpha_1 + \alpha_2)}\right). \tag{6.1.16}$$

$$\alpha_5 = 1 + e^{\alpha_4 T} \cdot \operatorname{cond} A_0, \tag{6.1.17}$$

where $\operatorname{cond} A_0$ is the condition of the matrix A_0, defined by (2.5.10).

$$\alpha_6 = \frac{1}{2} \alpha_5 \sum_{i=1}^{k-1} \frac{1}{i!} L_{i-1}(\square\, \bar{Y}) \cdot \nu_3^{i-1}. \tag{6.1.18}$$

$$\alpha_7 = 6\, \alpha_6 \nu_1 + \frac{\alpha_5}{k!} L(\bar{B}) \nu_3^{k-1}. \tag{6.1.19}$$

$$\alpha_8 = \sup_{0<h\le\nu_3} \frac{1}{h} \log(e^{\alpha_1 h} + \alpha_7 h). \tag{6.1.20}$$

NOTE. By virtue of theorem 2.5.2, α_8 is finite and satisfies

$$\alpha_8 \leq \min(\alpha_1 + \alpha_7 e^{\nu_3 \bar{\alpha}_1}, \alpha_1 \omega(-\nu_3 \bar{\alpha}_1) + \alpha_7),$$

where the function ω is defined by (2.5.1). □

$$\alpha_9 = \alpha_6 e^{2\alpha_1^+ T}. \tag{6.1.21}$$

$$\alpha_{10} = \frac{\alpha_5}{2k!} e^{\alpha_1^+ T} L(\bar{B}). \tag{6.1.22}$$

$$\alpha_{11} = \frac{\alpha_5}{2k!} L(\bar{B}) \sup\{\operatorname{diam}([0,1]\bar{f}_0(\bar{x})) \mid \bar{x} \in \amalg \mathbb{R}^M, \bar{x} \subset \bar{B}\}. \tag{6.1.23}$$

NOTE. The finiteness of the supremum in (6.1.23) follows from the finiteness of

$$\sup\{\|\bar{f}_0(\bar{x})\| \mid \bar{x} \in \amalg \mathbb{R}^M, \bar{x} \subset \bar{B}\},$$

since

$$\operatorname{diam}([0,1]\bar{f}_0(\bar{x})) \leq 2\|[0,1]\bar{f}_0(\bar{x})\| = 2\|\bar{f}_0(\bar{x})\|. \qquad \square$$

Choose H and Δ such that

$$0 < H \leq \nu_3, \tag{6.1.24}$$

$$0 < \Delta \leq \nu_1 e^{-\alpha_1^+ T} \tag{6.1.25}$$

and

$$e^{\bar{\alpha}_8 H} T \omega(\alpha_8 T)[\alpha_9 \Delta^2 + \alpha_{10} \Delta H^{k-1} + \alpha_{11} H^k] \leq \tfrac{1}{2} \nu_1. \tag{6.1.26}$$

Then, finally, we can define β_1 and β_2 as follows.

$$\beta_1 = e^{\bar{\alpha}_8 H} T \omega(\alpha_8 T)(\alpha_9 + \tfrac{1}{2}\alpha_{10}). \tag{6.1.27}$$

$$\beta_2 = e^{\bar{\alpha}_8 H} T \omega(\alpha_8 T)(\alpha_{11} + \tfrac{1}{2}\alpha_{10} H^{k-2}). \tag{6.1.28}$$

□

PROOF OF THEOREM 6.1.1.

1. Let the values Δ, H, β_1 and β_2 and all auxiliary values be defined and chosen according to (6.1.9) - (6.1.28). Let $\bar{x}_0 \in \amalg \mathbb{R}^M$, such that mean $\bar{x}_0 = \hat{x}_0$ and (6.1.5) holds. Furthermore, let $H_- > 0$ and assume (6.1.6).

2. We will prove by induction

$$\gamma_i \leq e^{h_i\alpha_8}\gamma_{i-1} + \alpha_9\delta^2 h_i + \alpha_{10}\delta h_i^k + \alpha_{11}h_i^{k+1}, \tag{6.1.29}$$

$$U([0,t_i],\bar{y}_0) \subset \bar{B}, \tag{6.1.30}$$

$$\begin{cases} \text{cond } S_i \leq e^{h_i\alpha_4} \\ \text{and } h_i \text{ need not be halved according to (5.2.5),} \end{cases} \tag{6.1.31}$$

for all $i \geq 1$ for which the i'th step is performed.

Assume that (6.1.29) - (6.1.31) hold for $1 \leq i \leq n-1$, where $n \geq 1$ is such that $t_{n-1} < T$. We will prove (6.1.29) - (6.1.31) for $i = n$. This proof will be given in the parts 3 - 8. Note that for $n = 1$ we do not assume (6.1.29) - (6.1.31) for any i.

3. Assume $n > 1$. Define

$$h^* = \max_{1\leq i\leq n-1} h_i, \tag{6.1.32}$$

then from (6.1.29) ($1 \leq i \leq n-1$) and $\alpha_{10},\alpha_{11} \geq 0$ we have

$$\gamma_i \leq e^{h_i\alpha_8}\gamma_{i-1} + h_i[\alpha_9\delta^2 + \alpha_{10}.\delta.(h^*)^{k-1} + \alpha_{11}(h^*)^k] \quad (1 \leq i \leq n-1). \tag{6.1.33}$$

By virtue of theorem 2.5.3 and using $\gamma_0 = 0$ we find

$$\gamma_{n-1} \leq e^{\alpha_8\bar{h}^*}t_{n-1}\omega(\alpha_8 t_{n-1})[\alpha_9\delta^2 + \alpha_{10}\delta(h^*)^{k-1} + \alpha_{11}(h^*)^k]. \tag{6.1.34}$$

By virtue of (4.2.10), (6.1.6) and (6.1.32) we have

$$h^* \leq \max_{1\leq i\leq n-1} H_i \leq H.$$

In view of theorem 2.5.1, (6.1.5), (6.1.34) and $\alpha_9,\alpha_{10},\alpha_{11} \geq 0$ we thus have

$$\gamma_{n-1} \le e^{\alpha_8 \bar{H}} T\omega(\alpha_8 T)[\alpha_9 \Delta^2 + \alpha_{10}\Delta H^{k-1} + \alpha_{11} H^k]. \tag{6.1.35}$$

Using (6.1.26) we conclude

$$\gamma_{n-1} \le \tfrac{1}{2}\nu_1 , \tag{6.1.36}$$

which is also (trivially) true for n = 1.

Assume again n > 1. By virtue of (6.1.30) for i = n-1, (2.4.12) and (6.1.11) we have

$$\|D_x U(t_{n-1}, \bar{y}_0)\| \le \exp(t_{n-1} \max \mu[f'(U([0,t_{n-1}], \bar{y}_0))])$$

$$\le \exp(t_{n-1} \max \mu[f'(\bar{B})]) = \exp(t_{n-1}\alpha_1).$$

Therefore

$$\|D_x U(t_{n-1}, \bar{y}_0)\| \le e^{\alpha_1^+ T}, \tag{6.1.37}$$

which is also true for n = 1, because $D_x U(0, \bar{y}_0) = I$. Now drop the assumption n > 1.

It is easy to verify that we have

$$\hat{y}_0 = A_0 \hat{x}_0 = A_0 \text{ mean } \bar{x}_0 = \text{mean } \square\, A_0 \bar{x}_0 = \text{mean } \square\, \bar{y}_0.$$

Thus the definitions of $\hat{y}_0$ and $\bar{y}_0$ used in this chapter are in accordance with (5.2.1).

Using (2.3.24), (6.1.37), (5.4.7), (6.1.4) and (6.1.5) we obtain

$$\|U(t_{n-1}, \bar{y}_0) - U(t_{n-1}, \hat{y}_0)\|$$

$$\le \|D_x U(t_{n-1}, \bar{y}_0)\| . \|\bar{y}_0 - \hat{y}_0\|$$

$$\le e^{\alpha_1^+ T} . \tfrac{1}{2} \text{ diam } \bar{y}_0 \le e^{\alpha_1^+ T} . \tfrac{1}{2}\Delta .$$

Combining this with (6.1.25) we find

$$\|U(t_{n-1}, \bar{y}_0) - U(t_{n-1}, \hat{y}_0)\| \le \tfrac{1}{2}\nu_1 . \tag{6.1.38}$$

By virtue of the triangle inequality for the Hausdorff distance, and using (6.1.36) and (6.1.38) we obtain

$$
\begin{aligned}
& q(\overline{y}_{n-1}, U(t_{n-1}, \hat{y}_0)) \\
& \leq q(\overline{y}_{n-1}, U(t_{n-1}, \overline{y}_0)) + q(U(t_{n-1}, \overline{y}_0), U(t_{n-1}, \hat{y}_0)) \\
& = \gamma_{n-1} + \| U(t_{n-1}, \overline{y}_0) - U(t_{n-1}, \hat{y}_0) \| \\
& \leq \tfrac{1}{2}\nu_1 + \tfrac{1}{2}\nu_1 = \nu_1 .
\end{aligned}
$$

In view of (2.3.51) and (2.3.22) this implies

$$\overline{y}_{n-1} \subset U(t_{n-1}, \hat{y}_0) + \nu_1 \overline{e}.$$

Using (6.1.9) we find

$$\overline{y}_{n-1} \subset \overline{Y}. \tag{6.1.39}$$

4. Let the variable $\overline{b}^{(0)}$ have one of the values assigned to it in the course of the execution of Algorithm I. According to (4.2.6) we have

$$\overline{b}^{(0)} = \square\, \overline{y}_{n-1} + \square\, [0, \hat{H}]\, \overline{f}_0(\square\, \overline{y}_{n-1}) + \hat{H} \beta\, \overline{e}, \tag{6.1.40}$$

where β is defined by (4.2.3) - (4.2.5) and $\hat{H}$ has a value satisfying $\hat{H} \leq H_n$. Combining this with (6.1.6) and (6.1.24) we find

$$\hat{H} \leq H_n \leq H \leq \nu_3. \tag{6.1.41}$$

By virtue of (4.2.4) we have $\alpha \leq \frac{1}{2}$, which implies, using (4.2.5),

$$\beta \leq \| \overline{f}_0(\square\, \overline{y}_{n-1}) \|. \tag{6.1.42}$$

Furthermore

$$
\begin{aligned}
& \| \square\, [0, \hat{H}]\, \overline{f}_0(\square\, \overline{y}_{n-1}) + \hat{H} \beta\, \overline{e} \| \\
& \leq \| \square\, [0, \hat{H}]\, \overline{f}_0(\square\, \overline{y}_{n-1}) \| + \| \hat{H} \beta\, \overline{e} \|
\end{aligned}
$$

$$= \| [0,\hat{H}]\ \overline{f}_0(\square\ \overline{y}_{n-1})\| + \hat{H}\beta$$

$$\leq \hat{H}\| \overline{f}_0(\square\ \overline{y}_{n-1})\| + \hat{H}\beta.$$

Using (6.1.42), (6.1.41), (6.1.15) and (6.1.39) this implies

$$\| \square\ [0,\hat{H}]\ \overline{f}_0(\square\ \overline{y}_{n-1}) + \hat{H}\beta\,\overline{e}\| \leq \nu_2.$$

Consequently, by virtue of (6.1.40) and (6.1.39) we have

$$\overline{b}^{(0)} \subset \square\ \overline{Y} + \nu_2\overline{e}.$$

From this we conclude, using (6.1.10),

$$\overline{b}^{(0)} \subset \overline{B}. \tag{6.1.43}$$

Let $\overline{b}^{(0)}$ have its final value. As long as we have not proved that h_n is not halved according to (5.2.5), let h_n have the value produced by Algorithm I, and let the matrix S_n be defined by (5.2.2) for this value of h_n.

In view of (4.3.1), (4.2.12), (4.2.11) and (6.1.43) we have

$$U([0,h_n],\overline{y}_{n-1}) \subset \overline{b}_n = \overline{b}^{(i_n)} \subset \overline{b}^{(i_n-1)} \subset \ldots \subset \overline{b}^{(1)} \subset \overline{b}^{(0)} \subset \overline{B}. \tag{6.1.44}$$

From (4.2.11), (4.2.12) and (5.4.3) we have

$$\operatorname{diam} \overline{b}_n \leq \operatorname{diam}(\square\ \overline{y}_{n-1} + \square\ [0,h_n]\ \overline{f}_0(\overline{b}^{(i_n-1)}))$$

$$\leq \operatorname{diam}(\square\ \overline{y}_{n-1}) + \operatorname{diam}(\square\ [0,h_n]\ \overline{f}_0(\overline{b}^{(i_n-1)}))$$

$$= \operatorname{diam} \overline{y}_{n-1} + \operatorname{diam}([0,h_n]\ \overline{f}_0(\overline{b}^{(i_n-1)}))$$

$$= \delta_{n-1} + h_n \operatorname{diam}([0,1]\ \overline{f}_0(\overline{b}^{(i_n-1)})).$$

Using (6.1.44) and (6.1.23) this implies

$$L(\overline{B}) \cdot \operatorname{diam} \overline{b}_n \leq L(\overline{B}) \cdot \delta_{n-1} + h_n \cdot \frac{2k!\alpha_{11}}{\alpha_5}\ . \tag{6.1.45}$$

5. By virtue of (5.4.3), (2.3.52), (2.3.37), (6.1.3), (6.1.37) and (6.1.4) we have

$$\begin{aligned}\delta_{n-1} &= \operatorname{diam} \bar{y}_{n-1} \\ &\le \operatorname{diam} U(t_{n-1},\bar{y}_0) + 2q(\bar{y}_{n-1},U(t_{n-1},\bar{y}_0)) \\ &\le \|D_x U(t_{n-1},\bar{y}_0)\| \cdot \operatorname{diam} \bar{y}_0 + 2\gamma_{n-1} \\ &\le e^{\alpha_1^+ T}\delta + 2\gamma_{n-1}.\end{aligned}$$

Thus we have obtained

$$\delta_{n-1} \le e^{\alpha_1^+ T}\delta + 2\gamma_{n-1}. \tag{6.1.46}$$

6. In view of (2.4.12), (2.4.13), (6.1.44), (6.1.11) and (6.1.12) we find

$$\|D_x U(t,\bar{y}_{n-1})\| \le e^{\alpha_1 t} \qquad (0 \le t \le h_n) \tag{6.1.47}$$

and

$$\|[D_x U(t,\bar{y}_{n-1})]^{-1}\| \le e^{\alpha_2 t} \quad (0 \le t \le h_n). \tag{6.1.48}$$

The function f is (k-1) times continuously differentiable. Therefore the functions $D_t^i D_x U(t,x)$ exist for $0 \le i \le k-1$, for values of t and x for which U(t,x) exists.

Expanding the matrix function $D_x U(t,\hat{y}_{n-1})$ in a Taylor series around t = 0 and using (2.2.64), (5.3.4) and (5.2.2) we have

$$\begin{aligned}D_x U(h_n,\hat{y}_{n-1}) &\in I + \sum_{i=1}^{k-2} \frac{h_n^i}{i!} D_t^i D_x U(0,\hat{y}_{n-1}) \\ &\quad + \frac{h_n^{k-1}}{(k-1)!} \square D_t^{k-1} D_x U([0,h_n],\hat{y}_{n-1}) \\ &\subset I + \sum_{i=1}^{k-2} \frac{h_n^i}{i!} f_{i-1}^{!}(\hat{y}_{n-1})\end{aligned}$$

$$+ \frac{h_n^{k-1}}{(k-1)!} \square \{D_x[f_{k-2}(U(t,x))] \mid 0 \le t \le h_n,\ x = \hat{y}_{n-1}\}$$

$$\subset S_n + \frac{h_n^{k-1}}{(k-1)!} [\square f'_{k-2}(U([0,h_n],\hat{y}_{n-1}))D_x U([0,h_n],\hat{y}_{n-1})$$

$$- f'_{k-2}(\hat{y}_{n-1})].$$

Using (2.3.17), (5.2.1), (6.1.44), (6.1.47) and (6.1.39) this implies

$$\|D_x U(h_n,\hat{y}_{n-1}) - S_n\|$$

$$\le \frac{h_n^{k-1}}{(k-1)!} \cdot \| \square f'_{k-2}(U([0,h_n],\hat{y}_{n-1}))D_x U([0,h_n],\hat{y}_{n-1})$$

$$- f'_{k-2}(\hat{y}_{n-1})\|$$

$$\le \frac{h_n^{k-1}}{(k-1)!} \cdot [M \cdot \|f'_{k-2}(U([0,h_n],\hat{y}_{n-1}))\| \cdot \|D_x U([0,h_n],\hat{y}_{n-1})\|$$

$$+ \|f'_{k-2}(\hat{y}_{n-1})\|]$$

$$\le \frac{h_n^{k-1}}{(k-1)!} \cdot [M \cdot \|f'_{k-2}(\overline{B})\| \cdot e^{h_n\alpha_1^+} + \|f'_{k-2}(\overline{Y})\|].$$

By virtue of (4.2.10) and (6.1.41) we have

(6.1.49) $\quad h_n \le \nu_3.$

Therefore we may conclude, using (6.1.13),

(6.1.50) $\quad \|D_x U(h_n,\hat{y}_{n-1}) - S_n\| \le h_n\alpha_3 e^{-\nu_3\alpha_2^+}.$

For any matrix $A \in \mathbb{R}^{M,M}$ we have, according to (2.4.5),

$$\mu[A] + \mu[-A] \ge 0.$$

Therefore, using (6.1.11) and (6.1.12), we have

(6.1.51) $\alpha_1 + \alpha_2 \geq 0.$

Consider theorem 2.5.4 with $D_x U(h_n, \hat{y}_{n-1})$ and S_n substituted for A and B, respectively. By virtue of (6.1.50), (6.1.48), (6.1.49) and (6.1.14) we have

$$\varepsilon = \|A - B\| \leq h_n \alpha_3 e^{-\nu_3 \alpha_2^+},$$

$$\theta = \varepsilon \|A^{-1}\| \leq \varepsilon \cdot e^{\alpha_2 h_n} \leq \varepsilon \cdot e^{\nu_3 \alpha_2^+} \leq h_n \alpha_3 \leq \nu_3 \alpha_3 < 1$$

and

$$\text{cond } A_n = \|A\| . \|A^{-1}\| \leq e^{\alpha_1 h_n} . e^{\alpha_2 h_n} = e^{h_n(\alpha_1+\alpha_2)}.$$

Hence we can apply the theorem so as to obtain the following results.

In the first place the matrix S_n and hence also A_n is regular. Therefore no halvings of h_n according to (5.2.5) are needed, and we need not distinguish any more between the values of h_n produced by Algorithm I and Algorithm II, nor between the corresponding matrices S_n.

In the second place we obtain

$$\text{cond } S_n \leq \text{cond } A + \frac{\theta}{1-\theta} (1 + \text{cond } A)$$

$$\leq e^{h_n(\alpha_1+\alpha_2)} + h_n \cdot \frac{\alpha_3}{1-\alpha_3\nu_3} (1 + e^{h_n(\alpha_1+\alpha_2)}).$$

Using (6.1.51), (6.1.49), (2.5.4) and (6.1.16) this implies

(6.1.52) $\text{cond } S_n \leq e^{h_n \alpha_4},$

and (6.1.31) has been proved for i = n.

By virtue of (6.1.1) with n replaced by n-1, (6.1.44), and (6.1.30) for i = n-1, we obtain (6.1.30) for i = n. Thus only (6.1.29) for i = n remains to be proved.

7. By virtue of (6.1.31) for $1 \leq i \leq n$ we have

$$\text{cond } A_n \leq \left(\prod_{i=1}^{n} \text{cond } S_i \right) . \text{cond } A_0 \leq e^{t_n \alpha_4} . \text{cond } A_0.$$

Using (6.1.17) and $\alpha_4 \geq 0$ this implies

(6.1.53) $\quad \text{cond } A_n \leq \alpha_5 - 1.$

Define ε_n by (5.4.1) and combine (5.4.6), (6.1.53), (6.1.39), (6.1.44) and (5.4.2) so as to obtain

$$\varepsilon_n \leq \frac{1}{2}\alpha_5\Big[\sum_{i=1}^{k-1}\frac{1}{i!}L_{i-1}(\square\,\overline{Y})\delta_{n-1}^2 h_n^i + \frac{1}{k!}L(\overline{B})\,h_n^k \operatorname{diam}\overline{b}_n\Big].$$

By virtue of (6.1.18) and (6.1.49) this implies

(6.1.54) $\quad \varepsilon_n \leq \alpha_6\delta_{n-1}^2 h_n + \frac{\alpha_5}{2k!}L(\overline{B})\,h_n^k \operatorname{diam}\overline{b}_n.$

Using (6.1.54), (6.1.45), (6.1.46), (6.1.21) and (6.1.22) we find

$$\begin{aligned}\varepsilon_n &\leq \alpha_6\delta_{n-1}^2 h_n + \frac{\alpha_5}{2k!}L(\overline{B})\delta_{n-1}h_n^k + \alpha_{11}h_n^{k+1}\\ &\leq \alpha_6(e^{\alpha_1^+\cdot T}\delta + 2\gamma_{n-1})^2 h_n\\ &\quad + \frac{\alpha_5}{2k!}L(\overline{B})(e^{\alpha_1^+ T}\delta + 2\gamma_{n-1})h_n^k + \alpha_{11}h_n^{k+1}\\ &\leq h_n\gamma_{n-1}\Big[4\alpha_6 e^{\alpha_1^+ T}\delta + 4\alpha_6\gamma_{n-1} + \frac{\alpha_5}{k!}L(\overline{B})\,h_n^{k-1}\Big]\\ &\quad + \alpha_9\delta^2 h_n + \alpha_{10}\delta h_n^k + \alpha_{11}h_n^{k+1}.\end{aligned}$$

By virtue of (6.1.5), (6.1.19), (6.1.25), (6.1.36) and (6.1.49) this implies

(6.1.55) $\quad \varepsilon_n \leq \alpha_7 h_n\gamma_{n-1} + \alpha_9\delta^2 h_n + \alpha_{10}\delta\,h_n^k + \alpha_{11}h_n^{k+1}.$

8. Using (6.1.3), the triangle inequality for the Hausdorff distance, (5.4.1) and (2.3.58) we have

$$\gamma_n = q(\bar{y}_n, U(t_n, \bar{y}_0))$$

$$\leq q(U(h_n, \bar{y}_{n-1}), U(t_n, \bar{y}_0)) + q(\bar{y}_n, U(h_n, \bar{y}_{n-1}))$$

$$= q(U(h_n, \bar{y}_{n-1}), U(h_n, U(t_{n-1}, \bar{y}_0))) + \varepsilon_n$$

$$\leq \| D_x U(h_n, \bar{y}_{n-1}) \| \cdot q(\bar{y}_{n-1}, U(t_{n-1}, \bar{y}_0)) + \varepsilon_n$$

$$\leq \| D_x U(h_n, \bar{y}_{n-1}) \| \cdot \gamma_{n-1} + \varepsilon_n .$$

Combining this with (6.1.47) and (6.1.55) we obtain

$$\gamma_n \leq (e^{h_n \alpha_1 \alpha_7 h_n}) \gamma_{n-1} + \alpha_9 \delta^2 h_n + \alpha_{10} \delta h_n^k + \alpha_{11} h_n^{k+1} . \tag{6.1.56}$$

Using (6.1.20) and (6.1.49) this proves (6.1.29) for i = n.

Thus we have completed the induction step and proved (6.1.29), (6.1.30) and (6.1.31) for all i ≥ 1 for which the i'th step of the method is performed.

9. Now we will show that after a finite number of N steps the grid-point $t_N = T$ is reached.

Let n ≥ 1 be such that the n'th step of the method is performed. Using (6.1.43) and (4.4.1) we find $K_0 \leq K(\bar{B})$, with K_0 defined by (4.4.2). By virtue of theorem 4.4.1 and (6.1.6) we obtain

$$h_n \geq \min\left(\frac{1}{10K_0}, H_n, T - t_{n-1}\right) \geq \min\left(\frac{1}{10K(\bar{B})}, H_-, T - t_{n-1}\right).$$

Combined with (4.2.10) this shows that the grid-point $t_N = T$ is reached for some N with

$$N < \frac{T}{\min\left(\frac{1}{10K(\bar{B})}, H_-\right)} + 1,$$

hence

$$N < T \cdot \max\left(10K(\bar{B}), \frac{1}{H_-}\right) + 1. \tag{6.1.57}$$

10. Let $1 \leq n \leq N$. Analogously to the derivation of (6.1.34) we may deduce, using (6.1.8),

$$\gamma_n \leq e^{\bar{\alpha}_8 h_{max}} t_n \omega(\alpha_8 t_n)[\alpha_9\delta^2 + \alpha_{10}\delta(h_{max})^{k-1} + \alpha_{11}(h_{max})^k]. \tag{6.1.58}$$

By virtue of (6.1.8), (4.2.10) and (6.1.6) we have

$$h_{max} \leq H_{max} \leq H. \tag{6.1.59}$$

Using (6.1.58), (6.1.59), theorem 2.5.1 and $\gamma_0 = 0$ we find

$$\gamma_n \leq e^{\bar{\alpha}_8 H} T \omega(\alpha_8 T)[\alpha_9\delta^2 + \alpha_{10}\delta(h_{max})^{k-1} + \alpha_{11}(h_{max})^k] \quad (0 \leq n \leq N). \tag{6.1.60}$$

Furthermore, by virtue of (6.1.59) we have

$$2\delta(h_{max})^{k-1} \leq \delta^2 + (h_{max})^{2(k-1)} \leq \delta^2 + H^{k-2}(h_{max})^k.$$

Together with (6.1.60), (6.1.27) and (6.1.28) this implies

$$\gamma_n \leq \beta_1\delta^2 + \beta_2(h_{max})^k \quad (0 \leq n \leq N). \tag{6.1.61}$$

Combining this with (6.1.59) yields (6.1.7) and the theorem has been proved. □

REMARK 6.1.3. In practice one will often be interested in the vector interval $\square\, \bar{y}_n$, instead of A_n and $\bar{x}_n$, together representing the set $\bar{y}_n$. The smallest vector interval enclosing the set $U(t_n,\bar{y}_0)$ is $\square\, U(t_n,\bar{y}_0)$. Therefore, in addition to the quantity γ_n, defined by (6.1.3), an important quantity is γ_n', defined by

$$\gamma_n' = q(\square\, \bar{y}_n, \square\, U(t_n,\bar{y}_0)). \tag{6.1.62}$$

By virtue of (2.3.57) we have a simple relation between γ_n and γ_n', namely

$$\gamma_n' \leq \gamma_n. \tag{6.1.63}$$

Therefore theorem 6.1.1 is also valid if we replace γ_n in (6.1.7) by γ_n'. □

REMARK 6.1.4. Assume $H_n = H_-$ $(1 \le n \le N)$,

$$0 < H_- \le \frac{1}{10K(\bar{B})}, \quad H_- \le H.$$

Combining (6.1.7) and (6.1.57) we obtain

$$\gamma_n \le \beta_1 \delta^2 + \beta_2 \cdot \left(\frac{T}{N-1}\right)^k. \tag{6.1.64}$$

Thus we have a simple relation between the number of steps and the global error. □

6.2. A FURTHER ANALYSIS OF THE GLOBAL ERROR

6.2.1. The condition of A_n

Let the assumptions of theorem 6.1.1 and those formulated in part 1 of its proof, hold. Let $1 \leq n \leq N$.

From part 8 of the proof we obtain

$$\gamma_n \leq \|D_x U(h_n, \bar{y}_{n-1})\| \cdot \gamma_{n-1} + \varepsilon_n. \tag{6.2.1}$$

Using (6.1.47) this implies

$$\gamma_n \leq e^{\alpha_1 h_n} \gamma_{n-1} + \varepsilon_n. \tag{6.2.2}$$

Since $\gamma_0 = 0$ we therefore have

$$\gamma_n \leq \sum_{i=1}^{n} e^{\alpha_1 (t_n - t_i)} \varepsilon_i. \tag{6.2.3}$$

Thus we see that the global error can be bounded by the sum of the local errors, each multiplied by a factor. This factor depends on the distance $t_n - t_i$ between the two grid-points concerned and on the number α_1, which according to (6.1.11) is defined by

$$\alpha_1 = \max \mu[f'(\bar{B})], \tag{6.2.4}$$

where $\bar{B}$ is defined by (6.1.11). Thus α_1 depends on the nature of the differential system. Since its value may be negative, the factor $e^{\alpha_1 (t_n - t_i)}$ may be smaller than 1. In that case the local error of the i'th step propagates in a favourable way. If we would have used the norm of $f'(\bar{B})$ instead of the logarithmic norm μ, this property could not have been revealed.

Let us now consider the local error ε_n. From the proof of theorem 5.4.1 we obtain

$$\varepsilon_n \leq \frac{1}{2}(1 + \text{cond } A_n) \cdot \text{diam } \bar{c}, \tag{6.2.5}$$

where $\bar{c}$ is defined by (5.2.4) and depends on n. Let us consider the origin and necessity of the factor $\frac{1}{2}(1 + \text{cond } A_n)$.

According to (5.3.8) we have

$$U(h_n,y) \in S_n y + \bar{c} \tag{6.2.6}$$

for all $y \in \bar{y}_{n-1}$. Thus we have

$$U(h_n,\bar{y}_{n-1}) \subset S_n \bar{y}_{n-1} + \bar{c}. \tag{6.2.7}$$

Therefore if we would not demand $\bar{y}_n \in \mathcal{Y}$, then we could have chosen $\bar{y}_n = S_n \bar{y}_{n-1} + \bar{c}$. Then the local error would be

$$\begin{aligned} & q(S_n \bar{y}_{n-1} + \bar{c},\ U(h_n,\bar{y}_{n-1})) \\ &= \max_{\substack{y\in\bar{y}_{n-1} \\ c\in\bar{c}}} \min_{y_1\in\bar{y}_{n-1}} \|S_n y + c - U(h_n,y_1)\| \\ &\le \max_{\substack{y\in\bar{y}_{n-1} \\ c\in\bar{c}}} \|S_n y + c - U(h_n,y)\| \\ &\le \max_{\substack{y\in\bar{y}_{n-1} \\ c\in\bar{c} \\ c_1\in\bar{c}}} \|S_n y + c - (S_n y + c_1)\| \\ &= \max_{c_1 c_1\in\bar{c}} \|c-c_1\| = \operatorname{diam} \bar{c}. \end{aligned}$$

Thus we have arrived at the following estimate.

$$q(S_n\bar{y}_{n-1} + \bar{c},\ U(h_n,\bar{y}_{n-1})) \le \operatorname{diam} \bar{c}. \tag{6.2.8}$$

From the proof of theorem 5.3.1 we obtain

$$U(h_n,\bar{y}_{n-1}) \subset S_n\bar{y}_{n-1} + \bar{c} = A_n \bar{x}_{n-1} + \bar{c} = A_n[\bar{x}_{n-1} + A_n^{-1}\bar{c}]. \tag{6.2.9}$$

By virtue of (6.1.2) and (5.2.6) we have

$$(6.2.10)\qquad \overline{y}_n = A_n \square (\overline{x}_{n-1} + A_n^{-1}\,\overline{c}) = A_n\overline{x}_{n-1} + A_n \square A_n^{-1}\,\overline{c}.$$

Comparing (6.2.9) and (6.2.10) we see that $\overline{y}_n$ is (with respect to the inclusion) the smallest set of the form $A_n\overline{x}_n$ with $\overline{x}_n \in \mathbb{I}\mathbb{R}^M$ for which we have

$$(6.2.11)\qquad S_n\overline{y}_{n-1} + \overline{c} \subset \overline{y}_n.$$

Thus, if we compare (6.2.8) and (6.2.5) it becomes plausible that the factor $\frac{1}{2}(1 + \operatorname{cond} A_n)$ in (6.2.5) is a consequence of the necessity to choose for $\overline{y}_n$ a set of the form $A_n\overline{x}_n$ with $\overline{x}_n \in \mathbb{I}\mathbb{R}^M$.

A similar phenomenon is observed by JACKSON [1975]. He considers linear differential systems and he deals with the ideal case that

$$(6.2.12)\qquad A_n = D_x U(t_n,x)$$

(which does not depend on x).

Starting from the inclusion

$$(6.2.13)\qquad U(t_{n-1},\overline{y}_0) \subset A_{n-1}\overline{x}_{n-1},$$

where $\overline{x}_{n-1} \in \mathbb{I}\mathbb{R}^M$, he assumes that the n'th step of a numerical process produces an inclusion

$$(6.2.14)\qquad U(t_n,\overline{y}_0) \subset A_n\,\overline{x} + \overline{r},$$

where $\overline{x} \in \mathbb{I}\mathbb{R}^M$ is such that

$$U(h_n,A_{n-1}\overline{x}_{n-1}) = A_n\overline{x}$$

and $\overline{r} \in \mathbb{I}\mathbb{R}^M$ is a set of error vectors. In order to obtain from (6.2.14) an inclusion of the form

$$(6.2.15)\qquad U(t_n,\overline{y}_0) \subset \overline{y}_n = A_n\overline{x}_n$$

with $\overline{x}_n \in \mathbb{I}\mathbb{R}^M$, this $\overline{x}_n$ has to satisfy

$$(6.2.16)\qquad \overline{x}_n \supset \overline{x} + \square A_n^{-1}\,\overline{r}.$$

In the best case we have $\overline{x}_n = \overline{x} + \square A_n^{-1}\,\overline{r}$ and we obtain the inclusion

$$(6.2.17)\qquad U(t_n,\overline{y}_0) \subset \overline{y}_n = A_n\overline{x} + A_n \square A_n^{-1}\,\overline{r}.$$

Comparing this with (6.2.14) we see that the set of errors $\overline{r}$ has been "blown up" to the set $A_n \square A_n^{-1}\,\overline{r}$. This effect is similar to the replacement of $A_n[\overline{x}_{n-1} + A_n^{-1}\,\overline{c}]$ by $A_n \square [\overline{x}_{n-1} + A_n^{-1}\,\overline{c}]$ (see (6.2.9) and (6.2.10)).

Jackson considers the differential system

$$(6.2.18)\qquad U'(t) = B\,U(t), \text{ where } B = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

for which we have

$$(6.2.19)\qquad D_x U(t,x) = \frac{1}{2}\begin{pmatrix} 1+e^{-2t} & 1-e^{-2t} \\ 1-e^{-2t} & 1+e^{-2t} \end{pmatrix},$$

$$(6.2.20)\qquad [D_x U(t,x)]^{-1} = \frac{1}{2}\begin{pmatrix} 1+e^{2t} & 1-e^{2t} \\ 1-e^{2t} & 1+e^{2t} \end{pmatrix}.$$

Choosing $\overline{r}$ such that $\overline{r} = \|\overline{r}\| \cdot \overline{e}$ he obtains

$$(6.2.21)\qquad \square A_n \square A_n^{-1}\,\overline{r} = e^{2t_n}\,\overline{r}$$

and therefore

$$(6.2.22)\qquad \|A_n \square A_n^{-1}\,\overline{r}\| = e^{2t_n}\,\|\overline{r}\|.$$

Using the condition of A_n we have in general

$$(6.2.23)\qquad \|A_n \square A_n^{-1}\,\overline{r}\| \le (\operatorname{cond} A_n) \cdot \|\overline{r}\|.$$

In this example we have

$$\operatorname{cond} A_n = \|D_x U(t_n,x)\| \cdot \|[D_x U(t_n,x)]^{-1}\|$$
$$= 1 \cdot e^{2t_n} = e^{2t_n}.$$

Thus we have in (6.2.23) an equality.

Let us now return to our method, described in section 4.2, 5.2 and 6.1. If we apply this method to the differential system (6.2.18) we have

$$S_n \approx D_x U(h_n,x),$$

and therefore, if we choose $A_0 = I$,

$$A_n \approx D_x U(t_n,x). \tag{6.2.24}$$

Thus we obtain

$$\text{cond } A_n \approx \text{cond } D_x U(t_n,x) = e^{2t_n}. \tag{6.2.25}$$

Let

$$\hat{c} = \text{mean } \bar{c} \tag{6.2.26}$$

and assume

$$\bar{c} - \hat{c} = \|\bar{c} - \hat{c}\| \cdot \bar{e}, \tag{6.2.27}$$

then we can derive that we have, similar to (6.2.22),

$$\|A_n \square A_n^{-1} (\bar{c} - \hat{c})\| \approx e^{2t_n} \|\bar{c} - \hat{c}\| . \tag{6.2.28}$$

Thus we see that also in this case the set $\bar{c} - \hat{c}$, which can be considered as the "error set", has been "blown up" to the set $A_n \square A_n^{-1}(\bar{c} - \hat{c})$ with a norm which is e^{2t_n} times as large. Let us now consider the effect of this phenomenon on the local error ε_n. Let

$$\bar{u} = A_n \bar{x}_{n-1} + \hat{c}, \quad \bar{v} = \bar{c} - \hat{c}, \quad \bar{w} = A_n \square A_n^{-1}(\bar{c} - \hat{c}). \tag{6.2.29}$$

By virtue of (6.2.9), (6.2.10) and (2.3.53) we have

$$\varepsilon_n = q(\bar{y}_n, U(h_n, \bar{y}_{n-1}))$$

$$\geq q(\bar{u} + \bar{w}, \bar{u} + \bar{v})$$

$$\geq \max_{1\leq i\leq M} q(p_i(\overline{u}) + p_i(\overline{w}),\ p_i(\overline{u}) + p_i(\overline{v})).$$

From (2.3.56) we easily obtain

$$(6.2.30)\qquad q(\overline{\xi}+\overline{\eta},\ \overline{\xi}+\overline{\zeta}) = q(\overline{\eta},\overline{\zeta})$$

for $\overline{\xi},\overline{\eta},\overline{\zeta} \in \mathbb{I}\mathbb{R}$.

Further we have $p_i(\overline{u})$, $p_i(\overline{v})$, $p_i(\overline{w}) \in \mathbb{I}\mathbb{R}$ $(1 \leq i \leq M)$. Thus we obtain, using (2.3.41), (2.3.46) and (2.3.9),

$$\begin{aligned}
\varepsilon_n &\geq \max_{1\leq i\leq M} q(p_i(\overline{w}),\ p_i(\overline{v})) \\
&\geq \max_{1\leq i\leq M} |q(p_i(\overline{w}),0) - q(p_i(\overline{v}),0)| \\
&= \max_{1\leq i\leq M} \big|\, |p_i(\overline{w})| - |p_i(\overline{v})| \,\big| \\
&\geq \Big|\, \max_{1\leq i\leq M} |p_i(\overline{w})| - \max_{1\leq i\leq M} |p_i(\overline{v})| \,\Big| \\
&= \big|\, \|\overline{w}\| - \|\overline{v}\| \,\big|.
\end{aligned}$$

Combining this with (6.2.29) and (6.2.28) we obtain for the differential system (6.2.18)

$$(6.2.31)\qquad \varepsilon_n \gtrsim (e^{2t_n} - 1)\,.\,\|\overline{c} - \hat{c}\|.$$

By virtue of (6.2.26), (2.3.29) and (6.2.25) we find

$$(6.2.32)\qquad \varepsilon_n \gtrsim \tfrac{1}{2}(e^{2t_n} - 1)\,.\,\operatorname{diam}\overline{c} \approx \tfrac{1}{2}(\operatorname{cond} A_n - 1)\,.\,\operatorname{diam}\overline{c}.$$

Thus we see that cond A_n can grow exponentially in t and that in such cases (6.2.5) can give a realistic impression of the local error. This phenomenon may limit the length of the integration interval for which our method, described in sections 4.2, 5.2 and 6.1, can be applied successfully. Adopting a suggestion of JACKSON [1975], the following change to our method might be an improvement.

If for some n, (cond A_n) has become large, put

$$(6.2.33)\qquad \begin{cases} \overline{x}_n := \square\, \overline{y}_n, \\ A_n := I. \end{cases}$$

Although this application of the rounding operator to the set $\overline{y}_n$ causes an extra error, it might be worthwhile since it reduces the conditions of the matrices A_n and may thus reduce the local errors ε_n for the following steps. Whenever (cond A_n) has become large again, (6.2.33) may be repeated.

6.2.2. The global error for small H and Δ

In order to obtain a better insight into the global error γ_n and the estimate (6.1.7), let us consider the coefficients β_1 and β_2 in (6.1.7) if $\nu_1 \to 0$, $\nu_2 \to 0$, $\nu_3 \to 0$ and $H \to 0$. For simplicity assume that $A_0 = I$, that $\overline{f}_0$ is inclusion isotonic and that $\overline{f}_0$, L and L_0 are continuous with respect to the Hausdorff distance.

Let $\overline{Y}_1 = U([0,T],\hat{y}_0)$, then $\alpha_i \to \hat{\alpha}_i$ $(1 \le i \le 11)$, where

$$(6.2.34)\qquad \hat{\alpha}_1 = \max \mu[f'(\square\, \overline{Y}_1)] ,$$

$$(6.2.35)\qquad \hat{\alpha}_2 = \max \mu[-f'(\square\, \overline{Y}_1)],$$

$$(6.2.36)\qquad \hat{\alpha}_3 = \begin{cases} M\, \|f'(\square\, \overline{Y}_1)\| + \|f'(\overline{Y}_1)\| & (\text{for } k = 2), \\ 0 & (\text{for } k > 2), \end{cases}$$

$$(6.2.37)\qquad \hat{\alpha}_4 = \hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3,$$

$$(6.2.38)\qquad \hat{\alpha}_5 = 1 + e^{\hat{\alpha}_4 T},$$

$$(6.2.39)\qquad \hat{\alpha}_6 = \frac{1}{2}\, \hat{\alpha}_5 L_0(\square\, \overline{Y}_1),$$

$$(6.2.40)\qquad \hat{\alpha}_7 = 0,$$

$$(6.2.41)\qquad \hat{\alpha}_8 = \hat{\alpha}_1,$$

$$(6.2.42)\qquad \hat{\alpha}_9 = \hat{\alpha}_6\, e^{2\hat{\alpha}_1^{+} T},$$

$$\hat{\alpha}_{10} = \frac{\hat{\alpha}_5}{2k!} e^{\hat{\alpha}_1^+ T} L(\square \bar{Y}_1), \tag{6.2.43}$$

$$\hat{\alpha}_{11} = \frac{\hat{\alpha}_5}{2k!} L(\square \bar{Y}_1) \operatorname{diam}([0,1]\, \bar{f}_0(\square \bar{Y}_1)). \tag{6.2.44}$$

Further we have $\Delta \to 0$, $\beta_1 \to \hat{\beta}_1$ and $\beta_2 \to \hat{\beta}_2$, where

$$\hat{\beta}_1 = T\, \omega(\hat{\alpha}_8 T)(\hat{\alpha}_9 + \tfrac{1}{2}\, \hat{\alpha}_{10}), \tag{6.2.45}$$

$$\hat{\beta}_2 = \begin{cases} T\, \omega(\hat{\alpha}_8 T)(\hat{\alpha}_{11} + \frac{1}{2}\, \hat{\alpha}_{10}) & (\text{for } k = 2), \\ T\, \omega(\hat{\alpha}_8 T)\hat{\alpha}_{11} & (\text{for } k > 2). \end{cases} \tag{6.2.46}$$

A further simplification of the estimate (6.1.7) is possible if we assume

$$\delta = 0. \tag{6.2.47}$$

Comparing (6.1.60) and (6.1.61) in the proof of theorem 6.1.1, we see that we can then replace (6.1.28) by

$$\beta_2 = e^{\alpha_8 \bar{H}}\, T\, \omega(\alpha_8 T)\, .\, \alpha_{11}. \tag{6.2.48}$$

Therefore we can also replace (6.2.46) by

$$\hat{\beta}_2 = T\, \omega(\hat{\alpha}_8 T)\hat{\alpha}_{11}. \tag{6.2.49}$$

Thus we have

$$\gamma_n \le \beta_2 (h_{max})^k \le \beta_2 (H_{max})^k, \tag{6.2.50}$$

where $\beta_2 \to \hat{\beta}_2$ for $H_{max} \to 0$,

$$\hat{\beta}_2 = T\, \omega(\hat{\alpha}_1 T)\hat{\alpha}_{11}, \tag{6.2.51}$$

and $\hat{\alpha}_1$ and $\hat{\alpha}_{11}$ are defined by (6.2.34) - (6.2.38) and (6.2.44).

6.3. COMPARISON WITH OTHER METHODS

In section 4.5.1 we have considered some alternatives of Algorithm I. We have shown their disadvantages, which obviously affect the global method. If the n'th step of the global method contains a never ending process or produces a zero step size, then the global method is not applicable on the whole prescribed interval [0,T]. Further, if the step size is unnecessarily small, then the global method is of course less efficient than is possible.

In section 5.5 we have considered the local error for some alternatives of our method. We will not analyse the global error for these methods, but it is plausible that large local errors cause the global error to be large as well.

HUNGER [1971] and MARCOWITZ [1973,1975] give methods for the inclusion of $U(T,\overline{y}_0)$ which are based on the choice

$$(6.3.1) \qquad \mathbb{Y} = \mathbb{I}\,\mathbb{R}^M$$

instead of (3.5).

MOORE [1966] showed with an example that this choice may cause any method with constant step size to behave badly (in the sense that it produces rough inclusions), especially for large t. We will treat a slightly modified version of this example in our own notation, and for variable step size. The example will be analysed in greater detail. The behaviour of the error γ_n for methods based on the choice (6.3.1) will be compared with the behaviour of γ_n for our method.

EXAMPLE 6.3.1. Let $M = 2$, $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $f(x) = Ax$, $\overline{y}_0 = \begin{pmatrix} [-\alpha,\alpha] \\ [-\alpha,\alpha] \end{pmatrix}$. We have

$$(6.3.2) \qquad U(t,x) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} . x.$$

Assume that for any set of grid-points $0 = t_0 < t_1 < \ldots < t_n = T$, a method has produced for $n = 1,2,\ldots,N$ vector intervals $\overline{y}_n$ with the property $\overline{y}_n \supset U(h_n,\overline{y}_{n-1})$, where $h_n = t_n - t_{n-1}$.

Using (6.3.2) we find

$$(6.3.3)\qquad \bar{y}_n \supset \square \begin{pmatrix} \cos h_n & -\sin h_n \\ \sin h_n & \cos h_n \end{pmatrix} \bar{y}_{n-1}.$$

Assume

$$(6.3.4)\qquad h_n \leq \frac{\pi}{2} \quad (1 \leq n \leq N).$$

We prove by induction

$$(6.3.5)\qquad \bar{y}_n \supset \left[\prod_{i=1}^{n} (\cos h_i + \sin h_i) \right] \bar{y}_0 \qquad (0 \leq n \leq N).$$

It is true for $n = 0$. Let it be true if n is replaced by $n-1$. Then

$$\begin{aligned}
\bar{y}_n &\supset \square \begin{pmatrix} \cos h_n & -\sin h_n \\ \sin h_n & \cos h_n \end{pmatrix} \left[\prod_{i=1}^{n-1} (\cos h_i + \sin h_i) \right] \bar{y}_0 \\
&= \left[\prod_{i=1}^{n-1} (\cos h_i + \sin h_i) \right] \square \begin{pmatrix} \cos h_n & -\sin h_n \\ \sin h_n & \cos h_n \end{pmatrix} \begin{pmatrix} [-\alpha,\alpha] \\ [-\alpha,\alpha] \end{pmatrix} \\
&= \left[\prod_{i=1}^{n-1} (\cos h_i + \sin h_i) \right] \begin{pmatrix} (\cos h_n)[-\alpha,\alpha] - (\sin h_n)[-\alpha,\alpha] \\ (\sin h_n)[-\alpha,\alpha] + (\cos h_n)[-\alpha,\alpha] \end{pmatrix} \\
&= \left[\prod_{i=1}^{n-1} (\cos h_i + \sin h_i) \right] (\cos h_n + \sin h_n)\bar{y}_0 \\
&= \left[\prod_{i=1}^{n} (\cos h_i + \sin h_i) \right] \bar{y}_0,
\end{aligned}$$

which had to be proved.

Assume $\frac{2T}{\pi} \in \mathbb{Z}$, then $U(T,\bar{y}_0) = \bar{y}_0$ and

$$\gamma_N = q(\bar{y}_N,\bar{y}_0) \geq \left[\prod_{i=1}^{N} (\cos h_i + \sin h_i) - 1 \right] . \alpha.$$

Write $h_{max} = \max_{1\leq n\leq N} h_n$ and observe that $\cos h_i + \sin h_i = e^{h_i(1+\mathcal{O}(h_{max}))}$ ($h_{max} \to 0$, uniform in i). For δ, defined in (6.1.4), we have $\delta = 2\alpha$.

For fixed T we may now conclude

$$(6.3.6)\qquad \gamma_N \geq \frac{1}{2} [e^T - 1 + \mathcal{O}(h_{max})] . \delta \qquad (h_{max} \to 0),$$

i.e., for arbitrary, sufficiently small step sizes, γ_N is arbitrarily close to, or greater than, $\frac{1}{2}(e^T - 1) . \delta$.

Define

$$(6.3.7) \qquad \hat{\gamma}_n = q(\square\, U(t_n,\bar{y}_0),\ U(t_n,\bar{y}_0)) \qquad (0 \le n \le N).$$

We have

$$(6.3.8) \qquad \gamma_n \ge \hat{\gamma}_N \qquad (0 \le n \le N)$$

and $\hat{\gamma}_n$ can be interpreted as the part of the error γ_n which is unavoidable due to the demand $\bar{y}_n \in \mathbb{Y} = \amalg\, \mathbb{R}^M$.

By virtue of (2.3.47) we have

$$(6.3.9) \qquad \hat{\gamma}_n \le \operatorname{diam} U(t_n,\bar{y}_0) = O(\delta)$$

and in general $\hat{\gamma}_n$ is not smaller than $O(\delta)$.

In view of (6.3.8), γ_n can therefore not be expected to satisfy

$$\gamma_n = O(\delta^2 + h_{max})$$

as we have for our method, described in the sections 4.2, 5.2 and 6.1 (see (6.1.7)).

Let us now consider the value

$$[\gamma_N - \max_{0\le n\le N} \hat{\gamma}_n].$$

Using (6.3.9) and (6.3.2) we obtain

$$\hat{\gamma}_n \le \operatorname{diam}\left[\begin{pmatrix} \cos t_n & -\sin t_n \\ \sin t_n & \cos t_n \end{pmatrix} \bar{y}_0\right] = (|\cos t_n| + |\sin t_n|) \cdot 2\alpha \le \delta\sqrt{2}$$

$$(0 \le n \le N).$$

Combining this with (6.3.6) yields

$$(6.3.10) \qquad \gamma_N - \max_{0\le n\le N} \hat{\gamma}_n \ge [\tfrac{1}{2}(e^T - 1) - \sqrt{2} + O(h_{max})] \cdot \delta.$$

Thus we see that not only γ_N itself, but even

$$\gamma_N - \max_{0\le n\le N} \hat{\gamma}_n$$

is, for T large enough and h_{max} small enough, at least proportional to δ' and hence not of order $O(\delta^2 + h_{max})$. Furthermore, the coefficient of δ is increasing at least exponentially in T. □

Summarizing the above, we have seen that for any method based on the choice (6.3.1) the behaviour of the error γ_n may be less favourable than the general behaviour of the error for our method, at least asymptotically as $\delta \to 0$ and meanwhile $h_{max} \to 0$ fast enough.

A comparison for the case $\delta = 0$ is more complicated. Note however that even in this case it is useful that the error depends in a favourable way on δ, because any enclosing set $\overline{y}_n$ may be considered as the initial value set $\overline{y}_0$ of a new integration process, for which in general $\delta = \text{diam } \overline{y}_n > 0$.

6.4. THE NECESSITY OF THE ERROR TERM OF ORDER δ^2

Consider the error estimate

$$\gamma_n \leq \beta_1 \delta^2 + \beta_2 (h_{max})^k \qquad (0 \leq n \leq N)$$

(see (6.1.7)).

For $\beta_1 > 0$ and fixed $\delta > 0$ the right-hand side of this inequality does not tend to zero as $h_{max} \to 0$. Indeed, this cannot be expected, because in general $U(t_n, \bar{y}_0) \notin \mathbb{Y}$. One might wonder whether the error term $\beta_1 \delta^2$ can be improved, and for instance be replaced by a term of order δ^3.

With an example we will show that a term of order δ^2 in the total error cannot be avoided, neither in our method, nor in any other method using $\mathbb{Y}$ as defined in (3.5).

EXAMPLE 6.4.2. Consider the differential system

$$(6.4.1) \qquad U'(t) = \begin{pmatrix} 0 \\ [U(t)]_1^2 \end{pmatrix} \qquad (0 \leq t \leq T)$$

with the initial condition

$$(6.4.2) \qquad U(0) \in \begin{pmatrix} [-\lambda, \lambda] \\ 0 \end{pmatrix}$$

where $0 \leq \lambda \leq \lambda_0$.

We have

$$(6.4.3) \qquad U\left(t, \begin{pmatrix} \mu \\ 0 \end{pmatrix}\right) = \begin{pmatrix} \mu \\ t\mu^2 \end{pmatrix} \qquad (0 \leq t \leq T, |\mu| \leq \lambda).$$

Let $\bar{y}_N$ be an inclusion of $U(T, \bar{y}_0)$ produced by any method using $\mathbb{Y}$ as defined in (3.5). The set $\bar{y}_N$ is convex and contains the vectors

$$U\left(T, \begin{pmatrix} -\lambda \\ 0 \end{pmatrix}\right) = \begin{pmatrix} -\lambda \\ T\lambda^2 \end{pmatrix} \text{ and } U\left(T, \begin{pmatrix} \lambda \\ 0 \end{pmatrix}\right) = \begin{pmatrix} \lambda \\ T\lambda^2 \end{pmatrix}.$$

Hence

$$\begin{pmatrix} 0 \\ T\lambda^2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} -\lambda \\ T\lambda^2 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \lambda \\ T\lambda^2 \end{pmatrix} \in \bar{y}_N.$$

Therefore

$$\gamma_N = q(\bar{y}_N, U(T,\bar{y}_0))$$

$$\geq \min_{u\in U(T,\bar{y}_0)} \left\| \begin{pmatrix} 0 \\ T\lambda^2 \end{pmatrix} - u \right\| .$$

Combining this with (6.4.2) and (6.4.3) yields

$$\gamma_N \geq \min_{|\mu| \leq \lambda} \left\| \begin{pmatrix} -\mu \\ T\lambda^2 - T\mu^2 \end{pmatrix} \right\|$$

$$= \min_{0\leq\mu\leq\lambda} \max(\mu, T\lambda^2 - T\mu^2)$$

$$= \frac{\sqrt{1 + 4T^2\lambda^2} - 1}{2T}$$

$$= \frac{2T\lambda^2}{1 + \sqrt{1 + 4T^2\lambda^2}} .$$

For δ, defined by (6.1.4), we have $\delta = 2\lambda$, hence, using $\lambda \leq \lambda_0$, we find

$$\gamma_N \geq \frac{1}{2(1 + \sqrt{1 + 4T^2\lambda_0^2})} \cdot T\delta^2 \qquad \square$$

This shows that γ_n is in general not smaller than of order δ^2.

Note that this result is still valid if $\mathcal{Y}$, instead of defined by (3.1), is any other class of convex sets, provided that $\begin{pmatrix} [-\lambda,\lambda] \\ 0 \end{pmatrix} \in \mathcal{Y}$.

6.5. STEP SIZE AND ORDER CONTROL

6.5.1. Introduction

In section 6.1 we described the global method, which contains a fixed parameter value $k \geq 2$. In view of error bound (6.1.7) we call this k the *order* of the method.

Let k_{max} be a given integer, $k_{max} \geq 2$. Let the function f be $(k_{max}-1)$ times continuously differentiable and let the functions $\bar{f}_i (1 \leq i \leq k_{max}-1)$ and $\bar{g}_i (0 \leq i \leq k_{max}-2)$ be given and satisfy (5.1.7) and (5.1.8) for all k with $2 \leq k \leq k_{max}$.

Although section 6.1 treats the global method for a fixed choice of the order, we can also vary the order step-by-step without any difficulty. In this section we will consider this generalization of the global method. Further we will treat the choice of the parameters $H_n (1 \leq n \leq N)$, which control the step size.

For the n'th step we try to choose the order k_n and the parameter H_n such that the local error ε_n (see (5.4.1)) satisfies

$$(6.5.1) \qquad \varepsilon_n \approx h_n [E_a + E_r . \| U([t_{n-1},t_n],\bar{y}_0) \|].$$

Here E_a and E_r are prescribed non-negative real numbers, which we call the *absolute error parameter* and the *relative error parameter*, respectively. Thus we use, like e.g. SHAMPINE & GORDON [1975] (see pp. 97, 98 and 164) a mixed absolute-relative error criterion for the local error per unit step. A discussion on the choice to control the local error per unit step rather than the local error per step can be found in GEAR [1971], p. 79.

In (6.5.1) we approximate the value $\| U([t_{n-1},t_n],\bar{y}_0) \|$ by $\| \bar{b}_{n-1} \|$, where $\bar{b}_{n-1}$ is defined by Algorithm I, described in section 4.2. Thus we try to choose H_n and k_n such that

$$(6.5.2) \qquad \varepsilon_n \approx h_n E,$$

where E is defined by

$$(6.5.3) \qquad E = E_a + E_r . \| \bar{b}_{n-1} \|.$$

6.5.2. The step size for a given order

Let k be the order we use in the n'th step.

From the proof of theorem 5.4.1 we easily obtain

$$\text{(6.5.4)} \qquad \operatorname{diam} \bar{c} \le \sum_{i=1}^{k-1} \frac{h_n^i}{i!} L_{i-1}(\square \bar{y}_{n-1})\delta_{n-1}^2 + \frac{h_n^k}{k!} \operatorname{diam}[\bar{f}_{k-1}(\bar{b}_n)],$$

where $\bar{c}$ is defined by (5.2.4).

Assume that cond $A_n \approx 1$ and that δ_{n-1} is small enough. Then in view of (6.5.4) and (6.2.5) we have

$$\text{(6.5.5)} \qquad \varepsilon_n \lesssim \frac{h_n^k}{k!} \operatorname{diam}[\bar{f}_{k-1}(\bar{b}_n)].$$

It is not unreasonable to assume that $\operatorname{diam}[\bar{f}_{k-1}(\bar{b}_n)]$ is approximately proportional to $\operatorname{diam} \bar{b}_n$. In view of (4.2.11) $\operatorname{diam} \bar{b}_n$ is approximately proportional to h_n, where we use again the assumption that δ_{n-1} is small. Thus we obtain

$$\text{(6.5.6)} \qquad \varepsilon_n \lesssim \frac{h_n^{k+1}}{k!} \cdot \frac{\operatorname{diam}[\bar{f}_{k-1}(\bar{b}_n)]}{h_n},$$

where the quotient

$$\frac{\operatorname{diam}[f_{k-1}(\bar{b}_n)]}{h_n}$$

is approximately constant as h_n varies. For $n > 1$ we approximate this quotient by

$$\frac{\operatorname{diam}[f_{k-1}(\bar{b}_{n-1})]}{h_{n-1}}.$$

Thus in view of (6.5.2) and (6.5.6) a suitable step size seems to be

$$\text{(6.5.7)} \qquad h'(k) = \sqrt[k]{\frac{k!h_{n-1} \cdot E}{\operatorname{diam}[f_{k-1}(\bar{b}_{n-1})]}}.$$

However, Algorithm I may also limit the step size, independently of the limitation of the step size related to the local error parameters. If the prescribed value H_n is too large, the value of $\hat{H}$, initially equal to H_n, is halved according to (4.2.8). If this halving is actually performed

in some step, it is useful to take this fact into account in the following steps by limiting the value of H_n. This is done by the demand

$$(6.5.8) \qquad H_n \leq \theta \cdot H'_{n-1},$$

where H'_{n-1} is the final value of $\hat{H}$ in Algorithm I for the (n-1)'st step, and where $\theta > 1$ is some chosen factor.

If this factor is chosen too small it may slow down a possible and useful growth of the step sizes. If it is too large the limitation (6.5.8) has little effect and halvings of $\hat{H}$ may occur in every step. Both would be inefficient. As a compromise we use the choice

$$(6.5.9) \qquad \theta = (1.1)^{n-j},$$

where j is the number of the last step for which an actual halving of $\hat{H}$ in Algorithm I was necessary.

The value of H_n is the step size Algorithm I aims at. Therefore, in order to obtain a step size h_n satisfying $h_n \approx h(k)$ we might choose $H_n = h(k)$. However, in view of a possible systematic deviation between H_n and the corresponding h_n, we put

$$(6.5.10) \qquad H_n = \frac{H'_{n-1}}{h_{n-1}} \cdot h(k).$$

In order to ensure that (6.5.8) is fulfilled we will compute H_n from (6.5.10) with a value of h(k) satisfying

$$(6.5.11) \qquad h(k) \leq \theta h_{n-1}.$$

To that end we put

$$(6.5.12) \qquad h(k) = \min(\theta h_{n-1}, h'(k)),$$

where h'(k) is defined by (6.5.7) and θ by (6.5.9).

This h(k), as a function of k, is the basis for the choice of the order, as described in the next subsection.

REMARK 6.5.1. For n = 1 the quotient

$$\frac{\operatorname{diam}[f_{k-1}(\bar{b}_n)]}{h_n}$$

in (6.5.6) is approximated as follows.

$$\begin{aligned}\frac{1}{h_1}\operatorname{diam}[f_{k-1}(\bar{b}_1)] &\approx \frac{1}{h_1}\|f'_{k-1}(\bar{b}_1)\| \cdot \operatorname{diam}\bar{b}_1\\ &\approx \|f'_{k-1}(\bar{b}_1)\| \cdot \operatorname{diam}([0,1]\bar{f}_0(\bar{b}^{(i_1-1)}))\\ &\approx \|f'_{k-1}(\bar{b}_1)\| \cdot \|\bar{f}_0(\bar{b}^{(i_1-1)})\|\\ &\approx \|\bar{f}_k(\square\, \bar{y}_0)\|.\end{aligned}$$

Thus for n = 1 we obtain instead of (6.5.7)

$$h'(k) = \sqrt[k]{\frac{k!E}{\|\bar{f}_k(\square\, \bar{y}_0)\|}}. \tag{6.5.13}$$

Since this is not defined for $k = k_{max}$, we do not use the maximum order for the first step. □

REMARK 6.5.2. For h_{n-1} in (6.5.7) and (6.5.10) we have to take the value produced by Algorithm I, even if that value was reduced by Algorithm II. □

6.5.3. The choice of the order

For the n'th step (where $n \geq 2$) we consider two possible orders, namely k_{n-1} and k'_n. For k_1 and k'_2 values are prescribed with $|k_1 - k'_2| = 1$. If $n \geq 3$ we choose for k'_n either $k_{n-1} - 1$ or $k_{n-1} + 1$, namely the value that was not considered as possible order for the (n-1)'st step.

Thus if $k_{n-1} > k_{n-2}$ then $k_{n-1} = k'_{n-1} = k_{n-2} + 1$ and $k'_n = k_{n-1} + 1$, i.e., a further increase of the order is considered. Similarly, if $k_{n-1} < k_{n-2}$ then $k_{n-1} = k'_{n-1} = k_{n-2} - 1$ and $k'_n = k_{n-1} - 1$, i.e., a further decrease of the order is considered. Finally, if $k_{n-1} = k_{n-2}$ then an increase of the order is considered if for the (n-1)'st step a decrease was considered, and vice versa.

If $k_n' < 2$ or $k_n' > k_{max}$ then we choose $k_n = k_{n-1}$. Otherwise we choose k_n to be the value of $k \in \{k_{n-1}, k_n'\}$ for which $h(k)/w_k$ is maximal, where $h(k)$ is computed according to (6.5.12) and w_k indicates an estimate of the relative amount of work involved in the performance of one integration step with order k. Thus we maximize the step length we proceed per unit of work, or in other words, we minimize the work per unit step (cf. GEAR [1971], pp. 75,79).

6.6. THE EFFECT OF ROUNDING ERRORS

In practice it is in general impossible to perform all interval arithmetic operations exactly. Therefore we have to use rounded-interval arithmetic instead of exact-interval arithmetic (see MOORE [1966]; see also section 7.1). This section will deal with the effect on the global error of the rounding errors thus occurring.

Assume that, due to rounding errors, (6.1.55) has to be replaced by

$$\varepsilon_n \leq \alpha_7 h_n \gamma_{n-1} + \alpha_9 \delta^2 h_n + \alpha_{10} \delta\, h_n^k + \alpha_{11} h_n^{k+1} + \xi \tag{6.6.1}$$

for some fixed $\xi > 0$. This value ξ will in practice be related to the machine accuracy and to $\|U([0,T],\bar{y}_0)\|$. Then (6.1.56) has to be replaced by

$$\gamma_n \leq (e^{h_n \alpha_1} + \alpha_7 h_n)\gamma_{n-1} + \alpha_9 \delta^2 h_n + \alpha_{10} \delta\, h_n^k + \alpha_{11} h_n^{k+1} + \xi \tag{6.6.2}$$

and (6.1.29) by

$$\gamma_i \leq e^{h_i \alpha_8}\, \gamma_{i-1} + \alpha_9 \delta^2 h_i + \alpha_{10} \delta\, h_i^k + \alpha_{11} h_i^{k+1} + \xi. \tag{6.6.3}$$

It is easy to see that we therefore have to replace (6.1.7) by

$$\gamma_n \leq \beta_1 \delta^2 + \beta_2 (h_{max})^k + \eta \leq \beta_1 \delta^2 + \beta_2 (H_{max})^k + \eta, \tag{6.6.4}$$

where we can choose for the added term η either

$$\eta = N \cdot e^{\alpha_8^+ T} \cdot \xi \tag{6.6.5}$$

or

$$\eta = e^{\alpha_8^- H}\, T\, \omega(\alpha_8 T) \cdot \frac{1}{\min\limits_{1 \leq n \leq N} h_n} \cdot \xi, \tag{6.6.6}$$

or, of course, the minimum of these values.
The same term η has to be added in the left hand side of (6.1.26). For fixed ξ it may be impossible to choose H and Δ such that the new version of (6.1.26) holds. However, it is possible if ξ is small enough.

The formulas (6.6.4) - (6.6.6) suggest that the effect of rounding errors can become very large if the step sizes are chosen very small. This phenomenon is not surprising and similar to that occurring in the approximative solution of differential equations (see e.g. HENRICI [1962]).

CHAPTER 7

COMPUTER PROGRAM

7.1. INTRODUCTION

In this chapter we will give a computer program for the numerical method described in this monograph. We will use the programming languages Algol 60 and Triplex-Algol 60.

The language Triplex-Algol 60 is described in WIPPERMANN [1968]. It has been designed to program calculations on real intervals as easily as similar calculations on real numbers. For that purpose a new kind of variable of type "triplex" has been added to the variables of type "integer", "real", "boolean", etc. A triplex variable is a triple $(\xi_1,\hat{\xi},\xi_2)$ of real numbers ξ_1, $\hat{\xi}$ and ξ_2 satisfying $\xi_1 \leq \hat{\xi} \leq \xi_2$. It denotes the real interval $[\xi_1,\xi_2]$ and is supplied with a so-called "main value" $\hat{\xi} \in [\xi_1,\xi_2]$. This main value is of little importance for us. In our description we will therefore often identify the triplex number $(\xi_1,\hat{\xi},\xi_2)$ with the real interval $[\xi_1,\xi_2]$.

An interval $[\xi_1,\xi_2]$ is representable in a computer only if the real numbers ξ_1 and ξ_2 are representable. Therefore it often occurs that an interval resulting from an operation has to be rounded to a representable interval. This rounding is always performed outwards so as to guarantee that any interval actually computed in the course of the calculations encloses the corresponding theoretical interval.

We use the actual realization of Triplex-Algol 60 described in EIJGENRAAM, VAN DE GRIEND & STATEMA [1976]. In this realization the text of a program written in Triplex-Algol 60 is translated into an Algol 60 text. In the latter calls occur of auxiliary subroutines written in assembler language. These subroutines perform basic operations such as computing the upwards rounded sum of two real numbers.

Since, basically, Triplex-Algol 60 is an extension of Algol 60, we could have written the whole program in Triplex-Algol 60. However, several

parts of the program have been written directly in Algol 60. These parts are not subjected to the translation into Algol 60, but inserted into the Algol 60 text resulting from the translation of the parts of the program written in Triplex-Algol 60. Although this procedure saves computer time for the translation of Triplex-Algol 60 statements, its main advantage is that it increases the efficiency of the final Algol 60 program. This increase is caused by the following. The organisation of the translation program makes it unavoidable that ordinary Algol 60 statements are often transformed into less efficient statements. For instance, an addition of two integers is transformed into a procedure call.

The parts of the program which are directly written in Algol 60 are those between "@ALG" and "@TRI". In these parts the three values of a triplex T are referred to by T(/1/), T(/2/) and T(/3/). If S(/I/) is an element of a triplex array its three values are referred to by S(/I,1/), S(/I,2/) and S(/I,3/), etc. For simplicity in describing the program we will identify numbers of type triplex with real intervals and arrays of type triplex with vector intervals and matrix intervals, depending on the number of subscripts. If a triplex number (ξ_1,ξ_2,ξ_3) has to be supplied by the user of the program, or is produced by the program, the main value ξ_2 is irrelevant, except that in the first case it has to satisfy $\xi_1 \leq \xi_2 \leq \xi_3$.

Our computer program has been written in the form of the procedures TAYL, SOC and SOLVE. Procedure TAYL calls the procedures HB and INVI, which are declared within it. Procedure SOLVE calls TAYL and SOC. For a user it is most convenient to call SOLVE. However, if more flexibility is required, it is recommended to call TAYL and SOC directly.

Procedures TAYL, SOC and SOLVE call a number of (mostly simple) auxiliary procedures. The complete texts of TAYL, SOC and SOLVE as well as code declarations (i.e., procedure headings followed by 'CODE' or 'TRICODE') of the auxiliary procedures are given in section 7.9. Sections 7.2 - 7.8 deal with the description and explanation of the procedures.

7.2. DESCRIPTION OF AUXILIARY PROCEDURES

Procedures TAYL, SOC and SOLVE call a number of auxiliary procedures. The headings of the latter are given in section 7.9. These procedures have the following meaning.

CM(X) - "compose mean", has the value {mean X};

CMV(M,T,X) - "compose mean of vector", X := {mean T} for $T \in \amalg \mathbb{R}^M$;

CP(X) - "compose", has the value {X};

CPM(M,AR,A) - "compose matrix", A := {AR} for $AR \in \mathbb{R}^{M,M}$;

DIAM(M,T) - "diameter", has the value diam T for $T \in \amalg \mathbb{R}^M$;

DIAMM(M,A) - "diameter of a matrix interval", has the value diam A for $A \in \amalg \mathbb{R}^{M,M}$;

DUPMI(M,X,Y) - "duplicate matrix interval", Y := X for $X \in \amalg \mathbb{R}^{M,M}$;

DUPVI(M,X,Y) - "duplicate vector interval", Y := X for $X \in \amalg \mathbb{R}^M$;

FO4AEA(A,B,N,M,C,IFAIL) - $C := A^{-1}B$ for a regular matrix $A \in \mathbb{R}^{N,N}$ and a real N×M matrix B;
IFAIL is an error indicator;
see NAG Library Manual [1974];

MAI(M,A,B,C) - "matrix addition for intervals",
$C := A+B$ for $A,B \in \amalg \mathbb{R}^{M,M}$;

MMMI(M,A,B,C) - "matrix matrix multiplication for intervals",
$C := \square AB$ for $A,B \in \amalg \mathbb{R}^{M,M}$;

MSI(M,A,B,C) - "matrix subtraction for intervals",
$C := A-B$ for $A,B \in \amalg \mathbb{R}^{M,M}$;

MVMI(M,A,X,Y) - "matrix vector multiplication for intervals",
$Y := \square A.X$ for $A \in \amalg \mathbb{R}^{M,M}$, $X \in \amalg \mathbb{R}^M$;

NORM(M,T) - has the value $\|T\|$ for $T \in \amalg \mathbb{R}^M$;

NORMM(M,A) - "norm of a matrix", has the value $\|A\|$ for $A \in \amalg \mathbb{R}^{M,M}$;

SMMI(M,C,A,B) - "scalar matrix multiplication for intervals", $B := \square\, C.A$ for $C \in \amalg \mathbb{R}$, $A \in \amalg \mathbb{R}^{M,M}$;

SVMI(M,C,X,Y) - "scalar vector multiplication for intervals", $Y := \square\, C.X$ for $C \in \amalg \mathbb{R}$, $X \in \amalg \mathbb{R}^{M}$;

UQTAD(A,B,C) - has the value A/B, for C = -1,0,1 rounded downwards, ordinarily rounded and rounded upwards, respectively;

UQTAO(A,B,C) - has the value A+B, rounded as in UQTAD;

UQTAV(A,B,C) - has the value A×B, rounded as in UQTAD;

VAI(M,A,B,C) - "vector addition for intervals", C := A+B for $A,B \in \amalg \mathbb{R}^{M}$;

VSI(M,A,B,C) - "vector subtraction for intervals, C := A-B for $A,B \in \amalg \mathbb{R}^{M}$.

7.3. DESCRIPTION OF PROCEDURE TAYL

Procedure TAYL ("method based on a Taylor series") performs one step of the method described in sections 4.2, 5.2 and 6.1 to solve an initial value problem for an autonomous system of first order differential equations. It has the following procedure heading.

```
'PROCEDURE' TAYL(M,K,TMAX,F,G,T,A,X,Y,HN,H1,HR,B,FB,DHH);
'INTEGER' M,K;
'REAL' HN,H1,HR;
'TRIPLEX' T,MAX;
'TRIPLEX''ARRAY' A,X,Y,B,FB;
'BOOLEAN' DHH;
'PROCEDURE' F,G;
```

Calling TAYL to perform the n'th step of the method, the parameters have the following meaning.

'INTEGER'	M	-	the number of equations in the differential system;
'INTEGER'	K	-	the "order" k of the method;
'TRIPLEX'	TMAX	-	the end T of the integration interval [0,T]; if the number T is a non-representable number, then the triplex TMAX must represent a small interval containing T;
'PROCEDURE'	F	-	an externally declared subroutine to evaluate the functions $\overline{f}_0$, $\overline{f}_{k-1}$ and f_i $(1 \leq i \leq k-2)$ (see below);
'PROCEDURE'	G	-	an externally declared subroutine to evaluate the functions $\overline{g}_i$ $(0 \leq i \leq k-2)$ (see below);
'TRIPLEX'	T	-	on input SUP(T) must have the value t_{n-1}; on output T has the value t_n; if t_n = TMAX this triplex number may represent an interval with small but positive diameter; in other cases this diameter is zero;

'TRIPLEX''ARRAY' A (dimension (/1 : M, 1 : M/)) -
on input the real matrix A_{n-1},
on output the real matrix A_n
(we could have used a 'REAL''ARRAY', but it is
found to be convenient to use a 'TRIPLEX''ARRAY',
where every element is a 'TRIPLEX' consisting
of three equal real numbers);
'TRIPLEX''ARRAY' X (dimension (/1 : M/)) -
on input the vector interval $\overline{x}_{n-1}$,
on output the vector interval $\overline{x}_n$;
'TRIPLEX''ARRAY' Y (dimension (/1 : M/)) -
on input the vector interval $\square\ \overline{y}_{n-1}$,
on output the vector interval $\square\ \overline{y}_n$;
'REAL' HN - on input the value H_n,
on output the final value of the variable $\hat{H}$
in Algorithm I;
'REAL' H1 - on output the output value of parameter H of
procedure HB, called within the body of TAYL;
'REAL' HR - on output the final step size h_n;
'TRIPLEX''ARRAY' B (dimension (/1 : M/)) -
on output the vector interval $\overline{b}_n$;
'TRIPLEX''ARRAY' FB (dimension (/1 : M/)) -
on output the vector interval $\overline{f}_{k-1}(\overline{b}_n)$;
'BOOLEAN' DHH - on output the value 'TRUE' if and only if the
output value of parameter HN is less than its
input value.

Procedure F must have a procedure heading of the form

```
'PROCEDURE' F(I,X,Y); 'VALUE' I;
'INTEGER' I;
'TRIPLEX''ARRAY' X,Y;
```

X and Y must be of dimension (/1 : M/). The statement F(I,X,Y) must result in the calculation of $\overline{f}_I(X)$, for I = 0 (see (4.1.2)) and I = k-1 (see (5.1.7)), and in the calculation of a vector interval enclosing $f_I(X)$ (see (5.1.6)), for $1 \leq I \leq k-2$ and for an X representing a set of only one vector. This result is assigned to Y.

Procedure G must have a procedure heading of the form

```
'PROCEDURE' G(I,X,Y); 'VALUE' I;
'INTEGER' I;
'TRIPLEX''ARRAY' X,Y;
```

X and Y must be of dimension (/1 : M/) and (/1 : M, 1 : M/), respectively. The statement G(I,X,Y) must result in the calculation of $\bar{g}_I(X)$, for $0 \leq I \leq k-2$ (see (5.1.8)). This result is assigned to Y.

7.4. DESCRIPTION AND EXPLANATION OF PROCEDURE HB

Procedure HB ("calculation of H and B") performs Algorithm I, described in section 4.2. It has the following procedure heading.

```
'PROCEDURE' HB(M,F,Y,L,HN,H,B,DHH);
'INTEGER' M;
'REAL' L,HN,H;
'TRIPLEX''ARRAY' Y,B;
'BOOLEAN' DHH;
'   'PROCEDURE' F;
```

Using the notation of Algorithm I, the parameters have the following meaning.

'INTEGER' M - as in Algorithm I, i.e., the number of equations in the differential system;

'PROCEDURE' F - an externally declared subroutine to evaluate the function $\overline{f}_0$;
we use the same procedure F which is parameter of procedure TAYL;

'TRIPLEX''ARRAY' Y (dimension (/1 : M/)) - the vector interval $\square\ \overline{y}_{n-1}$;

'REAL' L - $\|\overline{g}_0(\square\ \overline{y}_{n-1})\|$;

'REAL' HN - as in procedure TAYL;

'REAL' H - on output the value $\min(\hat{h},H_n)$, where $\hat{h}$ is as in (4.2.10);

'TRIPLEX''ARRAY' B (dimension (/1 : M/)) - as in procedure TAYL;

'BOOLEAN' DHH - as in procedure TAYL.

NOTE. It is convenient to deal with the requirement $t_n \leq T$ within procedure TAYL and not in procedure HB. Therefore the output value H is not necessarily equal to the value of h_n, defined in (4.2.10). □

The procedure body of HB is a straight-forward translation of Algorithm I and therefore needs no further explanation.

7.5. DESCRIPTION AND EXPLANATION OF PROCEDURE INVI

For a given matrix interval $\overline{A} \in \mathbb{II}\,\mathbb{R}^{M,M}$ procedure INVI ("inversion of a matrix interval") enables us to enclose the set $\overline{B} = \{A^{-1} \mid A \in \overline{A}\}$. The result is given in the form of a matrix interval $\overline{S} \in \mathbb{II}\,\mathbb{R}^{M,M}$ and a real number $\alpha \geq 0$, such that for all $\hat{B} \in \overline{B}$ there is an $S \in \overline{S}$ with $\|\hat{B}-S\| \leq \alpha$. Thus we have

$$(7.5.1) \qquad p_{ij}(\overline{B}) \in p_{ij}(\overline{S}) + \alpha.[-1,1] \qquad (1 \leq i \leq M,\ 1 \leq j \leq M),$$

and for all $\overline{b} \in \mathbb{II}\,\mathbb{R}^{M}$ we have

$$(7.5.2) \qquad p_{i}(\overline{B}\,\overline{b}) \in p_{i}(\overline{S}\,\overline{b}) + \alpha\|\overline{b}\|.[-1,1] \qquad (1 \leq i \leq M).$$

The procedure heading is as follows.

```
'PROCEDURE' INVI(M,A,S,R,FAIL);
'INTEGER' M;
'REAL' R;
'TRIPLEX''ARRAY' A,S;
'LABEL' FAIL;
```

The parameters have the following meaning.

```
'INTEGER' M              - M;
'TRIPLEX''ARRAY' A (dimension (/1 : M, 1 : M/)) - the matrix interval Ā;
'TRIPLEX''ARRAY' S (dimension (/1 : M, 1 : M/)) - on output the matrix
                           interval S̄;
'REAL' R                 - α;
'LABEL' FAIL             - label to which is jumped on failure.
```

The procedure uses essentially the method of HANSEN [1965]. Using the auxiliary procedure F04AEA we first compute a matrix $B \in \mathbb{R}^{M,M}$ with $B \approx [\text{mean } \overline{A}]^{-1}$. Further we compute

$$\overline{C} := I - \square\, \overline{A}B,$$

$$\overline{S}_0 := \{B\},$$

$$\overline{S}_j := B + \square\, \overline{S}_{j-1}\overline{C} \qquad (1 \le j \le \ell),$$

$$\overline{S} := \overline{S}_\ell,$$

$$\alpha := \frac{\|\overline{C}\|^{\ell+1}}{1 - \|\overline{C}\|} \cdot \|\overline{B}\|.$$

It is easy to prove that

$$\operatorname{diam}\{\hat{B} \mid \hat{B} \in \mathbb{R}^{M,M},\ \exists S \in \overline{S} : \|\hat{B}-S\| \le \alpha\} = \operatorname{diam} \overline{S} + 2\alpha.$$

Therefore we increase ℓ until [diam $\overline{S}$ + 2α] does not decrease anymore.

7.6. EXPLANATION OF PROCEDURE TAYL

In the procedure body of TAYL first Algorithm I, described in section 4.2, is performed by a call of procedure HB. For the obtained step size h_n the value $t_{n-1} + h_n$ is in general not a representable number. We could enclose this value in an interval $\overline{t}_n$, but it is more convenient to use grid-points t_n which are exact real numbers. Therefore we define the grid-point t_n to be the downwards rounded sum $t_{n-1} + h_n$, and replace the real number h_n by the interval we obtain by computing $t_n - t_{n-1}$ in interval arithmetic.

We allow the end T of the integration interval [0,T] to be a non-representable number. Therefore this number is represented by an interval $\overline{T}$ with a small diameter that may be non-zero.

Let t_n be computed as explained above. If $t_n \geq \max \overline{T}$ then t_n is replaced by $\overline{T}$, and the last step size h_n is the result of the rounded interval subtraction $\overline{T} - t_{n-1}$. However, if $\min \overline{T} < t_n < \max \overline{T}$ then t_n is replaced by $\min \overline{T}$ (this is a rare case and even impossible if $\overline{T}$ is the smallest representable interval containing the number T). In this case the next step will be a very small one to complete the interval [0,T].

Further Algorithm II, described in section 5.2, is performed. The computation of A_n, according to (5.2.2) and (5.2.3), has to be performed in interval arithmetic and results in a matrix interval $\overline{A}_n$. However, similar to our wish to use an exact grid-point t_n, we want the matrix A_n to be an exactly given real matrix. Therefore we put

$$A_n := \text{mean } \overline{A}_n. \tag{7.6.1}$$

Let $\overline{x}$ be the vector interval $\overline{x}_n$ computed according to Algorithm II, on the understanding that the set $A_n^{-1}\,\overline{c}$ in (5.2.6) has been interpreted as the set $\{A^{-1} c \mid A \in \overline{A}_n,\ c \in \overline{c}\}$. Then (5.3.1) yields

$$U(h_n, \overline{y}_{n-1}) \subset \overline{A}_n\,\overline{x}. \tag{7.6.2}$$

In order to ensure that we have

$$U(h_n, \overline{y}_{n-1}) \subset A_n\,\overline{x}_n \tag{7.6.3}$$

we compute $\overline{x}_n$ such that

(7.6.4) $\quad \overline{A}_n\,\overline{x} \subset A_n\,\overline{x}_n.$

Since

$$A_n^{-1}\,\overline{A}_n\,\overline{x} = A_n^{-1}[A_n + (\overline{A}_n - A_n)]\overline{x} = [I + A_n^{-1}(\overline{A}_n - A_n)]\overline{x},$$

(7.6.4) holds for $\overline{x}_n$, defined by

(7.6.5) $\quad \overline{x}_n := \square\,[I + \square\,\overline{D}\,(\overline{A}_n - A_n)]\overline{x},$

where $\overline{D} \in \mathbb{IR}^{M,M}$ is such that

(7.6.6) $\quad A_n^{-1} \in \overline{D}.$

After a call of procedure INVI, we can compute in view of (7.5.1) and (7.5.2) a suitable $\overline{D}$, as well as an inclusion of $\{A^{-1}c \mid A \in \overline{A}_n,\ c \in \overline{c}\}$, which we need, according to (5.2.6), to compute $\overline{x}$.

7.7. DESCRIPTION AND EXPLANATION OF PROCEDURE SOC

Procedure SOC ("step size and order control") computes for given values of the absolute and relative error parameter a suitable step size and order for the next step of the integration process. This is done according to the method described in section 6.5. It has the following procedure heading.

```
'PROCEDURE' SOC(M,KMAX,ABSERR,RELERR,WPST,F,K,K1,H,B,FB,HH,HHF);
'INTEGER' M,KMAX,K,K1;
'REAL' ABSERR,RELERR,H,HH,HHF;
'REAL''ARRAY' WPST;
'TRIPLEX''ARRAY' B,FB;
'PROCEDURE' F;
```

Calling SOC in preparation for the performance of the n'th step of the integration process ($n \geq 2$), the parameters have the following meaning.

'INTEGER' M - as in procedure TAYL;

'INTEGER' KMAX - the maximum order k_{max} (see section 6.5);

'kEAL' ABSERR - the absolute error parameter;

'REAL' RELERR - the relative error parameter;

'ARRAY' WPST (dimension (/2 : KMAX/)) - for $2 \leq K \leq KMAX$ WPST(/K/) denotes an estimate of the relative amount of work to perform a step of the integration process with order K;

'PROCEDURE' F - an externally declared subroutine to evaluate the functions $\overline{f}_i (0 \leq i \leq k_{max}-1)$ (see below);

'INTEGER' K - on input the value of parameter K of procedure TAYL in the (n-1)'st step;
on output the value of parameter K of procedure TAYL to be used in the n'th step;

'INTEGER' K1 - on input the alternative value of K considered, either K-1 or K+1;
on output the value of this parameter to be used in the next call of SOC;

'REAL' H - the output value of parameter H1 of procedure TAYL in the (n-1)'st step;

'TRIPLEX''ARRAY' B (dimension (/1 : M/)) - the output value of parameter B of procedure TAYL in the (n-1)'st step;

'TRIPLEX''ARRAY' FB (dimension (/1 : M/)) - the output value of parameter FB of procedure TAYL in the (n-1)'st step;

'REAL' HH - on input the output value of parameter HN of procedure TAYL in the (n-1)'st step; on output the input value of parameter HN of procedure TAYL to be used in the n'th step;

'REAL' HHF - the factor θ indicating the maximally allowed increase of HH, i.e., the maximum quotient of output and input value of parameter HH (see (6.5.8) and (6.5.9)).

As in procedure TAYL, procedure F must have a procedure heading of the form

```
'PROCEDURE' F(I,X,Y); 'VALUE' I;
'INTEGER' I;
'TRIPLEX''ARRAY' X,Y;
```

X and Y must be of dimension (/1 : M/). For $0 \leq I \leq k_{max}-1$ the statement F(I,X,Y) must result in the calculation of $\overline{f}_I(X)$ (see (5.1.7)) and the assignment of this vector interval to Y. Note that for $2 \leq k \leq k_{max}$ these requirements imply those for procedure F in TAYL.

The procedure body of SOC is a straight-forward translation of the method of varying the step size and order described in section 6.5 and therefore needs no further explanation.

We only mention that it uses an integer or real array FAC with dimension (/2 : KMAX/) with values FAC(/K/) = K! This array has to be declared and assigned its values in the main program before calling SOC. Thus we avoid computing these values in every step. For KMAX > 12 the array must be of type real since integers must be smaller than 2^{31} for the computer implementation we use.

7.8. DESCRIPTION AND EXPLANATION OF PROCEDURE SOLVE

Procedure SOLVE performs the method described in sections 4.2, 5.2 and 6.1 for solving an initial value problem for an autonomous system of first order differential equations. The order and step size are varied automatically, according to the method described in section 6.5. The procedure heading is as follows.

```
'PROCEDURE' SOLVE (M,KMAX,N,TT,ABSERR,RELERR,WORK,YY,F,G);
'INTEGER' M,KMAX,N;
'REAL' ABSERR,RELERR;
'REAL''ARRAY' WORK;
'TRIPLEX''ARRAY' TT,YY;
'PROCEDURE' F,G;
```

The parameters have the following meaning.

'INTEGER' M - the number of equations in the differential system;

'INTEGER' KMAX - the value k_{max} of section 6.5, i.e., the maximally allowed order;

'INTEGER' N - on input the number of grid-points T_j for which an inclusion of the solution $U(T_j)$ is required; on output the number of grid-points T_j for which such an inclusion has been obtained;

'TRIPLEX''ARRAY' TT (dimension (/1 : N/)) - the grid-points T_j = TT(/j/), satisfying $0 < T_1 < T_2 < \ldots < T_N$, for which an inclusion of the solution $U(T_j)$ is required; if, for any j, T_j is a non-representable number, then the triplex TT(/j/) must represent a small interval containing T_j;

'REAL' ABSERR - the value E_a of section 6.5, i.e., the absolute error parameter;

'REAL' RELERR - the value E_r of section 6.5, i.e., the relative error parameter;

'REAL''ARRAY' WORK (dimension (/0 : KMAX-2/)) - WORK(/k/) indicates an estimate of the relative amount of work involved in the evaluation of the function $\bar{g}_k$;

'TRIPLEX''ARRAY' YY (dimension (/0 : N, 1 : M/)) - the values YY (/0,I/) ($1 \leq I \leq M$) denote the given vector interval $\overline{y}_0$ of initial values;
for $1 \leq J \leq N$ the values YY(/J,I/) ($1 \leq I \leq M$) denote on output the vector interval $\square\, \overline{y}_n$ (where $t_n = T_J$), which encloses the solution $U(T_J)$;

'PROCEDURE' F - an externally declared subroutine to evaluate the functions $\overline{f}_i$ ($0 \leq i \leq$ KMAX-1), see parameter F of procedure SOC (section 7.7);

'PROCEDURE' G - an externally declared subroutine to evaluate the functions $\overline{g}_i$ ($0 \leq i \leq$ KMAX-2), see (5.1.8), similar to parameter F.

The matrix A_0 has been chosen to be the unit matrix. Further we have chosen k_1 = KMAX-1, k_2' = KMAX-2 (see section 6.5).

The procedure body of SOLVE consists mainly of the procedure calls of TAYL and SOC.

If, for any n, the step size h_n is zero, then the process is stopped and the output value of N is smaller than its input value. This may for instance occur if, for any j, the diameter of the interval represented by the triplex TT(/j/) is too large.

We observe that the initial value set used for the subinterval $[T_j, T_{j+1}]$ is not the vector interval $\square\, \overline{y}_n$ (where $t_n = T_j$), but the vector set $\overline{y}_n$, represented by the matrix A_n and the vector interval $\overline{x}_n$ (see section 5.2). If one is interested in A_n and $\overline{x}_n$, or if one wishes to choose other values of A_0, k_1 or k_2', one can call the procedures TAYL and SOC directly.

Finally we mention that the relative amount of work to perform a step of the integration process with order K is estimated by 2.WORK(/0/) + $\sum_{I=1}^{K-2}$ WORK(/I/). The factor 2 is used to account for the work involved in the evaluations of $\overline{f}_0$. This amount of work is assumed to be related to WORK(/0/).

7.9. TEXT OF THE PROCEDURES

PROCEDURE TAYL

```
'PROCEDURE' TAYL(M,K,TMAX,F,G,T,A,X,Y,HN,H1,HR,B,FB,DHH);
'INTEGER' M,K; 'REAL' HN,H1,HR; 'TRIPLEX' T,TMAX;
'TRIPLEX' 'ARRAY' A,X,Y,B,FB; 'BOOLEAN' DHH; 'PROCEDURE' F,G;

'BEGIN'



    'PROCEDURE' HB(M,F,Y,L,HN,H,B,DHH); 'INTEGER' M;
    'REAL' L,HN,H; 'TRIPLEX' 'ARRAY' Y,B; 'BOOLEAN' DHH;
    'PROCEDURE' F;

    'BEGIN' 'INTEGER' I,J;
        'REAL' AL,BB,HH,HOLD,H1,RØ,R1,R2,S,W,WRI;
        'TRIPLEX' 'ARRAY' BOLD,B1,FB,FBOLD,FY(/1:M/);
        F(Ø,Y,FY); RØ:=NORM(M,FY); HH:=HN; @ALG;
    NEWHH:
        AL:=HH*L;
        'IF' AL<.1 'THEN' AL:=.1 'ELSE'
        'IF' AL>.5 'THEN' AL:=.5;
        R1:=RØ*AL/(1-AL);
        'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
        'BEGIN'
            R2:=FY(/I,1/);
            'IF' R2>Ø 'THEN' R2:=Ø;
            B1(/I,1/):=UQTAO(R2,-R1,-1); R2:=FY(/I,3/);
            'IF' R2<Ø 'THEN' R2:=Ø;
            B1(/I,3/):=UQTAO(R2,R1,1);
        'END' ;
        'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
        'BEGIN'
            B(/I,1/):=UQTAO(Y(/I,1/),UQTAV(HH,B1(/I,1/),-1),
            -1);
            B(/I,3/):=UQTAO(Y(/I,3/),UQTAV(HH,B1(/I,3/),1),1);
            B(/I,2/):=(B(/I,1/)+B(/I,3/))/2;
        'END' ;
        F(Ø,B,FB); H:=2*HN/HH;
        'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
        'BEGIN'
            'IF' FB(/I,1/)<Ø 'THEN'
            'BEGIN'
                H1:=UQTAD(B1(/I,1/),FB(/I,1/),-1);
                'IF' H1<H 'THEN' H:=H1;
            'END' ;
            'IF' FB(/I,3/)>Ø 'THEN'
            'BEGIN'
                H1:=UQTAD(B1(/I,3/),FB(/I,3/),-1);
```

```
            'IF' H1<H 'THEN' H:=H1;
        'END' ;
    'END' ;
    H:=UQTAV(H,HH,-1);
    'IF' H>HN 'THEN' H:=HN;
    'IF' (HH=HN&H<HN/2)!(HOLD<H&H<HH&HH<HN) 'THEN'
    'BEGIN'
        HH:=HH/2; HOLD:=H; @TRI; DUPVI(M,B,BOLD);
        DUPVI(M,FB,FBOLD); @ALG; 'GOTO' NEWHH;
    'END' ;
    'IF' HH<HN&HOLD>=H 'THEN'
    'BEGIN'
        H:=HOLD; HH:=2*HH; @TRI; DUPVI(M,BOLD,B);
        DUPVI(M,FBOLD,FB); @ALG;
    'END' ;
    DHH:=HH<.9*HN; HN:=HH;
    J:=Ø;
NEWB:
    J:=J+1; W:=1;
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
    'BEGIN'
        S:=B(/I,3/)-B(/I,1/); R2:=FB(/I,1/);
        'IF' R2>Ø 'THEN' R2:=Ø;
        BB:=UQTAO(Y(/I,1/),UQTAV(H,R2,-1),-1);
        'IF' BB>B(/I,1/) 'THEN' B(/I,1/):=BB;
        R2:=FB(/I,3/);
        'IF' R2<Ø 'THEN' R2:=Ø;
        BB:=UQTAO(Y(/I,3/),UQTAV(H,R2,1),1);
        'IF' BB<B(/I,3/) 'THEN' B(/I,3/):=BB;
        'IF' S=Ø 'THEN' WRI:=1 'ELSE'
        WRI:=(B(/I,3/)-B(/I,1/))/S;
        'IF' WRI<W 'THEN' W:=WRI;
    'END' ;
    @TRI;
    'IF' W<.9&J<1Ø 'THEN'
    'BEGIN'
        F(Ø,B,FB); 'GOTO' NEWB
    'END' ;
'END' HB;



'PROCEDURE' INVI(M,A,S,R,FAIL); 'INTEGER' M; 'REAL' R;
'TRIPLEX' 'ARRAY' A,S; 'LABEL' FAIL;

'BEGIN' 'INTEGER' I,IFAIL,J,L; 'REAL' NB,NC,P,PO,Q;
    'ARRAY' AR,BR,ER(/1:M,1:M/);
    'TRIPLEX' 'ARRAY' B,C,E(/1:M,1:M/);
    @ALG;
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
    'FOR' J:=1 'STEP' 1 'UNTIL' M 'DO'
    'BEGIN'
        AR(/I,J/):=(A(/I,J,1/)+A(/I,J,3/))/2;
        ER(/I,J/):= 'IF' I=J 'THEN' 1 'ELSE' Ø;
```

```
        'END' ;
        @TRI; FØ4AEA(AR,ER,M,M,BR,IFAIL); CPM(M,BR,B);
        MMMI(M,A,B,C);
        'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
        'FOR' J:=1 'STEP' 1 'UNTIL' M 'DO'
            C(/I,J/):=CP(ER(/I,J/))-C(/I,J/);
        NC:=NORMM(M,C);
        'IF' NC>=1 'THEN' 'GOTO' FAIL;
        NB:=NORMM(M,B); DUPMI(M,B,S); P:=DIAMM(M,S); @ALG;
        Q:=2*NB/(1-NC); P:=P+NC*Q; PO:=P+1; L:=Ø;
    NEWL:
        L:=L+1; @TRI; MMMI(M,S,C,E); MAI(M,B,E,S); PO:=P;
        P:=DIAMM(M,S); @ALG; P:=P+NC**(L+1)*Q;
        'IF' P<PO 'THEN' 'GOTO' NEWL;
        @TRI; R:=SUP(CP(NB)*CP(NC)**(L+1)/(1-CP(NC)));
    'END' INVI;



    'INTEGER' I,J;
    'REAL' C,L;
    'TRIPLEX' C1,C2,H,R,R1,T1,U;
    'TRIPLEX' 'ARRAY' FYØ,VØ,V1,XØ,X1,YØ,Z(/1:M/),A1,GYØ,MØ,
    M1,M2,M3(/1:M,1:M/);

    G(Ø,Y,M3); L:=NORMM(M,M3); HB(M,F,Y,L,HN,HR,B,DHH);
    H1:=HR; T:=CP(SUP(T));
NEWH:
    T1:=CP(INF(T+CP(HR)));
    'IF' T1>=SUP(TMAX) 'THEN' T1:=TMAX 'ELSE'
    'IF' T1>INF(TMAX) 'THEN' T1:=CP(INF(TMAX));
    H:=T1-T; HR:=MAIN(H); CMV(M,X,XØ); MVMI(M,A,XØ,YØ);
    C2:=H/K;
    'FOR' I:=K-1 'STEP' -1 'UNTIL' 1 'DO'
    'BEGIN'
        F(I-1,YØ,FYØ); G(I-1,YØ,GYØ); MVMI(M,GYØ,YØ,V1);
        VSI(M,FYØ,V1,V1);
        'IF' I>1 'THEN'
        'BEGIN'
            G(I-1,Y,M2); MSI(M,M2,GYØ,M2)
        'END' 'ELSE' MSI(M,M3,GYØ,M2);
        'IF' I<K-1 'THEN'
        'BEGIN'
            VAI(M,V1,VØ,V1); MAI(M,GYØ,MØ,GYØ);
            MAI(M,M2,M1,M2);
        'END' ;
        C1:=H/I; C2:=C2*C1; SVMI(M,C1,V1,VØ);
        SMMI(M,C1,GYØ,MØ); SMMI(M,C1,M2,M1);
    'END' I-LOOP;
    VSI(M,Y,YØ,Z); MVMI(M,M1,Z,V1); F(K-1,B,FB);
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
    'BEGIN'
        VØ(/I/):=VØ(/I/)+V1(/I/)+C2*FB(/I/);
        MØ(/I,I/):=MØ(/I,I/)+1;
```

```
    'END' ;
    MMMI(M,MØ,A,A1); INVI(M,A1,M1,C,FAIL); MVMI(M,M1,VØ,V1);
    R1:=CP(C)*(/-1,Ø,1/); R:=R1*CP(NORM(M,VØ));
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
        V1(/I/):=V1(/I/)+R;
    VAI(M,X,V1,X);
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
    'FOR' J:=1 'STEP' 1 'UNTIL' M 'DO'
    'BEGIN'
        M1(/I,J/):=M1(/I,J/)+R1; U:=A1(/I,J/);
        A1(/I,J/):=CM(U); M2(/I,J/):=U-A1(/I,J/);
    'END' ;
    MMMI(M,M1,M2,M3);
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
        M3(/I,I/):=M3(/I,I/)+1;
    MVMI(M,M3,X,X1); MVMI(M,A1,X1,Y); 'GOTO' END;
FAIL:
    HR:=HR/2; 'GOTO' NEWH;
END:T:=T1; DUPVI(M,X1,X); DUPMI(M,A1,A);
'END' TAYL;
```

PROCEDURE SOC

```
'PROCEDURE' SOC(M,KMAX,ABSERR,RELERR,WPST,F,K,K1,H,B,FB,HH,
HHF); 'INTEGER' M,KMAX,K,K1; 'REAL' ABSERR,RELERR,H,HH,HHF;
'REAL' 'ARRAY' WPST; 'TRIPLEX' 'ARRAY' B,FB; 'PROCEDURE' F;

'BEGIN' 'INTEGER' AUX; 'REAL' EAO,ERR,ESO,HAO,HSO,H1;
    ERR:=ABSERR+RELERR*NORM(M,B); H1:=HHF*H;
    HSO:=(ERR*FAC(/K/)*H/DIAM(M,FB))**(1/K);
    'IF' HSO>H1 'THEN' HSO:=H1;
    ESO:=HSO/WPST(/K/);
    'IF' K1>=2 & K1<=KMAX 'THEN'
    'BEGIN'
        F(K1-1,B,FB);
        HAO:=(ERR*FAC(/K1/)*H/DIAM(M,FB))**(1/K1);
        'IF' HAO>H1 'THEN' HAO:=H1;
        EAO:=HAO/WPST(/K1/);
    'END' 'ELSE' EAO:=Ø;
    'IF' EAO>ESO 'THEN'
    'BEGIN'
        AUX:=K; K:=K1; K1:=2*K-AUX; HSO:=HAO;
    'END' 'ELSE' K1:=2*K-K1;
    HH:=(HH/H)*HSO;
'END' SOC;
```

PROCEDURE SOLVE

```
'PROCEDURE' SOLVE(M,KMAX,N,TT,ABSERR,RELERR,WORK,YY,F,G);
'INTEGER' M,KMAX,N; 'REAL' ABSERR,RELERR; 'REAL' 'ARRAY' WORK;
'TRIPLEX' 'ARRAY' TT,YY; 'PROCEDURE' F,G;

'BEGIN' 'INTEGER' I,J,K,K1,TTI; 'REAL' ERR,H,HH,HHF,H1;
    'TRIPLEX' T,TMAX,TOLD; 'REAL' 'ARRAY' WPST(/2:KMAX/);
    'TRIPLEX' 'ARRAY' B,FB,X,Y(/1:M/),A(/1:M,1:M/);
    'BOOLEAN' DHH,START;
    WPST(/2/):=2*WORK(/0/);
    'FOR' K:=3 'STEP' 1 'UNTIL' KMAX 'DO'
        WPST(/K/):=WPST(/K-1/)+WORK(/K-2/);
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
    'FOR' J:=1 'STEP' 1 'UNTIL' M 'DO'
        A(/I,J/):= 'IF' I=J 'THEN' 1 'ELSE' 0;
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
    'BEGIN'
        X(/I/):=YY(/0,I/); Y(/I/):=YY(/0,I/);
    'END' ;
    T:=0; START:= 'TRUE' ; TTI:=0;
NEWTT:
    TTI:=TTI+1; TMAX:=TT(/TTI/);
NEWSTP:
    'IF' START 'THEN'
    'BEGIN'
        START:= 'FALSE' ; K:=KMAX-1; K1:=KMAX-2; HHF:=2;
        F(K,Y,FB); ERR:=ABSERR+RELERR*NORM(M,Y);
        HH:=(ERR*FAC(/K/)/NORM(M,FB))**(1/K);
    'END' 'ELSE'
    'BEGIN'
        'IF' DHH 'THEN' HHF:=1;
        HHF:=1.1*HHF;
        SOC(M,KMAX,ABSERR,RELERR,WPST,F,K,K1,H1,B,FB,HH,HHF);
    'END' ;
    TOLD:=CP(SUP(T));
    TAYL(M,K,TMAX,F,G,T,A,X,Y,HH,H1,H,B,FB,DHH);
    'IF' T=TOLD 'THEN'
    'BEGIN'
        N:=TTI-1; 'GOTO' END;
    'END' ;
    'IF' T<SUP(TMAX) 'THEN' 'GOTO' NEWSTP;
    'FOR' I:=1 'STEP' 1 'UNTIL' M 'DO'
        YY(/TTI,I/):=Y(/I/);
    'IF' TTI<N 'THEN' 'GOTO' NEWTT;
END:
'END' SOLVE;
```

AUXILIARY PROCEDURES

```
'TRIPLEX' 'PROCEDURE' CM(X); 'VALUE' X; 'TRIPLEX' X;
'TRICODE' ;

'PROCEDURE' CMV(M,T,X); 'INTEGER' M; 'TRIPLEX' 'ARRAY' T,X;
'TRICODE' ;

'TRIPLEX' 'PROCEDURE' CP(X); 'VALUE' X; 'REAL' X; 'TRICODE' ;

'PROCEDURE' CPM(M,AR,A); 'INTEGER' M; 'ARRAY' AR;
'TRIPLEX' 'ARRAY' A; 'TRICODE' ;

'REAL' 'PROCEDURE' DIAM(M,T); 'INTEGER' M;
'TRIPLEX' 'ARRAY' T; 'TRICODE' ;

'REAL' 'PROCEDURE' DIAMM(M,A); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A; 'TRICODE' ;

'PROCEDURE' DUPMI(M,X,Y); 'INTEGER' M; 'TRIPLEX' 'ARRAY' X,Y;
'TRICODE' ;

'PROCEDURE' DUPVI(M,X,Y); 'INTEGER' M; 'TRIPLEX' 'ARRAY' X,Y;
'TRICODE' ;

'PROCEDURE' FØ4AEA(A,B,N,M,C,IFAIL); 'VALUE' N,M;
'INTEGER' N,M,IFAIL; 'REAL' 'ARRAY' A,B,C; 'CODE' ;

'PROCEDURE' MAI(M,A,B,C); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A,B,C; 'TRICODE' ;

'PROCEDURE' MMMI(M,A,B,C); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A,B,C; 'TRICODE' ;

'PROCEDURE' MSI(M,A,B,C); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A,B,C; 'TRICODE' ;

'PROCEDURE' MVMI(M,A,X,Y); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A,X,Y; 'TRICODE' ;

'REAL' 'PROCEDURE' NORM(M,T); 'INTEGER' M;
'TRIPLEX' 'ARRAY' T; 'TRICODE' ;

'REAL' 'PROCEDURE' NORMM(M,A); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A; 'TRICODE' ;

'PROCEDURE' SMMI(M,C,A,B); 'VALUE' C; 'INTEGER' M;
'TRIPLEX' C; 'TRIPLEX' 'ARRAY' A,B; 'TRICODE' ;

'PROCEDURE' SVMI(M,C,X,Y); 'VALUE' M,C; 'INTEGER' M;
'TRIPLEX' C; 'TRIPLEX' 'ARRAY' X,Y; 'TRICODE' ;

'REAL' 'PROCEDURE' UQTAD(A,B,C); 'VALUE' A,B,C; 'INTEGER' C;
'REAL' A,B; 'CODE' ;
```

```
'REAL' 'PROCEDURE' UQTAO(A,B,C); 'VALUE' A,B,C; 'INTEGER' C;
'REAL' A,B; 'CODE' ;

'REAL' 'PROCEDURE' UQTAV(A,B,C); 'VALUE' A,B,C; 'INTEGER' C;
'REAL' A,B; 'CODE' ;

'PROCEDURE' VAI(M,A,B,C); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A,B,C; 'TRICODE' ;

'PROCEDURE' VSI(M,A,B,C); 'INTEGER' M;
'TRIPLEX' 'ARRAY' A,B,C; 'TRICODE' ;
```

CHAPTER 8

NUMERICAL EXPERIMENTS

In this chapter we will give the numerical results of applying procedure SOLVE, described in chapter 7, to a number of initial value problems. We consider 4 differential systems, each with both an initial value set consisting of only one element and an initial value set which is an interval of non-zero diameter.

The computations have been performed on the AMDAHL V7-B computer of the Centraal Rekeninstituut (Central Computing Institute) of the University of Leiden.

DIFFERENTIAL SYSTEM 1.

We consider the differential equation

$$U'(t) = -[U(t)]^2 \qquad (0 \leq t \leq 10^5). \tag{8.1}$$

For problem 1a and problem 1b we choose as initial condition

$$U(0) = 1 \tag{8.2}$$

and

$$U(0) \in [0.999, 1.001], \tag{8.3}$$

respectively. To (8.1) corresponds, for $x > 0$, the solution function

$$U(t,x) = \frac{1}{x^{-1} + t} \qquad (0 \leq t \leq 10^5). \tag{8.4}$$

The maximum order k_{max} is chosen to be 20. The procedure parameters F and G of procedure SOLVE are declared as follows.

```
'PROCEDURE' F(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN'
    Y(/1/):=(-1)**(I+1)*FAC(/I+1/)*X(/1/)**(I+2);
'END' F;


'PROCEDURE' G(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN'
    Y(/1,1/):=(-1)**(I+1)*FAC(/I+2/)*X(/1/)**(I+1);
'END' G;
```

Further we choose

$$\text{ABSERR} = 0,$$
$$\text{RELERR} = 10^{-16},$$
$$\text{WORK}(/I/) = 1 \qquad (0 \leq I \leq 18).$$

With these parameter values procedure SOLVE produces the following results.

PROBLEM 1a (83 steps performed, computation time 44 sec).

TT(/J/)	YY(/J,I/)		DIAM
Ø	+1.00000000000000'+00	+1.00000000000000'+00	Ø
+1'+01	+9.09090909090907'-02	+9.09090909090912'-02	+5'-16
+1'+02	+9.90099009900987'-03	+9.90099009900994'-03	+7'-17
+1'+03	+9.99000999000994'-04	+9.99000999001003'-04	+9'-18
+1'+04	+9.99900009998986'-05	+9.99900009999114'-05	+2'-17
+1'+05	+9.99990000099362'-06	+9.99990000101101'-06	+2'-17

PROBLEM 1b (83 steps performed, computation time 45 sec).

TT(/J/)	YY(/J,I/)		DIAM
Ø	+9.98999999999999'-Ø1	+1.ØØ1ØØØØØØØØØØ1'+ØØ	+3'-Ø3
+1'+Ø1	+9.Ø9ØØ7936115Ø82'-Ø2	+9.Ø9173882Ø66737'-Ø2	+2'-Ø5
+1'+Ø2	+9.9ØØ89164Ø79777'-Ø3	+9.9Ø1Ø8855722Ø4'-Ø3	+2'-Ø7
+1'+Ø3	+9.98999996586873'-Ø4	+9.99ØØ2ØØ1415124'-Ø4	+3'-Ø9
+1'+Ø4	+9.998999Ø9576525'-Ø5	+9.99900110421575'-Ø5	+3'-11
+1'+Ø5	+9.99989990055304'-Ø6	+9.999900101451Ø9'-Ø6	+3'-13

Some general remarks on all tables of this chapter:

1) A'B denotes $A.10^B$.
2) Under the heading YY(/J,I/) the tables show for each J the following numbers:

$$\begin{array}{ll} \min YY(/J,1/) & \max YY(/J,1/) \\ \vdots & \vdots \\ \min YY(/J,M/) & \max YY(/J,M/) \end{array}$$

The "main values" (see section 7.1) of the produced triplex numbers YY(/J,I/) are irrelevant and have therefore been left out.

3) Under the heading "diam" the tables show the diameters of the triplex numbers YY(/J,I/), rounded upwards.

We see that the diameter of the solution set $\bar{y}_n$ of problem 1b decreases considerably as n increases, according to the nature of the differential equation. Further we remark that the variable step size is very useful for problems 1a and 1b, because of the long integration interval $[0,10^5]$.

DIFFERENTIAL SYSTEM 2.

We consider the differential system

$$(8.5) \qquad \left\{ \begin{array}{l} U_1'(t) = -U_2(t), \\ U_2'(t) = U_1(t) \end{array} \right\} \quad (0 \le t \le 8\pi).$$

For problem 2a and problem 2b we choose as initial conditions

$$(8.6) \qquad \begin{cases} U_1(0) = 1, \\ U_2(0) = 0 \end{cases}$$

and

$$(8.7) \qquad \begin{cases} U_1(0) \in [\ 0.999, 1.001], \\ U_2(0) \in [-0.001, 0.001], \end{cases}$$

respectively. To (8.5) corresponds the solution function

$$(8.8) \qquad U(t,x) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \cdot x \qquad (0 \le t \le 8\pi).$$

Problems 2a and 2b, especially the latter, are of interest since they can illustrate the effectiveness of the choice of the class Y , introduced in chapter 3. MOORE [1966] showed that the choice $Y = \Pi\, \mathbb{R}^M$ would cause the diameter of the solution set to increase exponentially in t (compare also example 6.3.1).

One "revolution" of the solution set, that is, an increase of t by 2π, would increase the diameter by a factor $e^{2\pi} \approx 500$.

The maximum order k_{max} is chosen to be 20. The procedure parameters F and G of procedure SOLVE are declared as follows.

```
'PROCEDURE' F(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN' I:=I 'MOD' 4;
    'IF' I=Ø 'THEN'
    'BEGIN'
        Y(/1/):=-X(/2/); Y(/2/):=X(/1/)
    'END' 'ELSE'
    'IF' I=1 'THEN'
    'BEGIN'
        Y(/1/):=-X(/1/); Y(/2/):=-X(/2/)
    'END' 'ELSE'
    'IF' I=2 'THEN'
    'BEGIN'
        Y(/1/):=X(/2/); Y(/2/):=-X(/1/)
    'END' 'ELSE'
    'BEGIN'
        Y(/1/):=X(/1/); Y(/2/):=X(/2/)
    'END' ;
'END' F;
```

```
'PROCEDURE' G(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN' I:=I 'MOD' 4;
    'IF' I=Ø 'THEN'
    'BEGIN'
        Y(/1,1/):=Ø; Y(/1,2/):=-1; Y(/2,1/):=1; Y(/2,2/):=Ø
    'END' 'ELSE'
    'IF' I=1 'THEN'
    'BEGIN'
        Y(/1,1/):=-1; Y(/1,2/):=Ø; Y(/2,1/):=Ø; Y(/2,2/):=-1
    'END' 'ELSE'
    'IF' I=2 'THEN'
    'BEGIN'
        Y(/1,1/):=Ø; Y(/1,2/):=1; Y(/2,1/):=-1; Y(/2,2/):=Ø
    'END' 'ELSE'
    'BEGIN'
        Y(/1,1/):=1; Y(/1,2/):=Ø; Y(/2,1/):=Ø; Y(/2,2/):=1
    'END' ;
'END' G;
```

Further we choose

$$\text{ABSERR} = 10^{-16},$$

$$\text{RELERR} = 0,$$

$$\text{WORK(/I/)} = 1 \qquad (0 \leq I \leq 18).$$

PROBLEM 2a (48 steps performed, computation time 34 sec).

TT(/J/)	YY(/J,I/)		DIAM
0	+1.000000000000000	+1.000000000000000	ø
	.000000000000000	.000000000000000	ø
$\frac{1}{2}\pi$	-.000000000000001	+.000000000000001	+2'-15
	+.999999999999999	+1.000000000000004	+5'-15
π	-1.000000000000006	-.999999999999999	+7'-15
	-.000000000000001	+.000000000000001	+2'-15
$1\frac{1}{2}\pi$	-.000000000000002	+.000000000000001	+3'-15
	-1.000000000000009	-.999999999999998	+2'-14
2π	+.999999999999998	+1.000000000000012	+2'-14
	-.000000000000002	+.000000000000002	+4'-15
$2\frac{1}{2}\pi$	-.000000000000002	+.000000000000003	+5'-15
	+.999999999999997	+1.000000000000015	+2'-14
3π	-1.000000000000018	-.999999999999997	+3'-14
	-.000000000000003	+.000000000000004	+7'-15
$3\frac{1}{2}\pi$	-.000000000000005	+.000000000000003	+8'-15
	-1.000000000000021	-.999999999999996	+3'-14
4π	+.999999999999996	+1.000000000000024	+3'-14
	-.000000000000006	+.000000000000003	+9'-15
$4\frac{1}{2}\pi$	-.000000000000004	+.000000000000006	+1'-14
	+.999999999999995	+1.000000000000027	+4'-14
5π	-1.000000000000030	-.999999999999995	+4'-14
	-.000000000000004	+.000000000000007	+2'-14
$5\frac{1}{2}\pi$	-.000000000000011	+.000000000000005	+2'-14
	-1.000000000000034	-.999999999999993	+5'-14
6π	+.999999999999991	+1.000000000000039	+5'-14
	-.000000000000013	+.000000000000007	+2'-14
$6\frac{1}{2}\pi$	-.000000000000006	+.000000000000019	+3'-14
	+.999999999999989	+1.000000000000043	+6'-14
7π	-1.000000000000048	-.999999999999987	+7'-14
	-.000000000000008	+.000000000000021	+3'-14
$7\frac{1}{2}\pi$	-.000000000000027	+.000000000000007	+4'-14
	-1.000000000000052	-.999999999999985	+7'-14
8π	+.999999999999984	+1.000000000000056	+8'-14
	-.000000000000029	+.000000000000009	+4'-14

PROBLEM 2b (48 steps performed, computation time 34 sec).

TT(/J/)	YY(/J,I/)		DIAM
0	+.998999999999999	+1.001000000000001	+3'-03
	-.001000000000001	+.001000000000001	+3'-03
$\frac{1}{2}\pi$	-.001000000000001	+.001000000000001	+3'-03
	+.998999999999999	+1.001000000000004	+3'-03
π	-1.001000000000007	-.998999999999999	+3'-03
	-.001000000000001	+.001000000000001	+3'-03
$1\frac{1}{2}\pi$	-.001000000000002	+.001000000000001	+3'-03
	-1.001000000000009	-.998999999999998	+3'-03
2π	+.998999999999998	+1.001000000000012	+3'-03
	-.001000000000002	+.001000000000002	+3'-03
$2\frac{1}{2}\pi$	-.001000000000002	+.001000000000003	+3'-03
	+.998999999999997	+1.001000000000015	+3'-03
3π	-1.001000000000018	-.998999999999997	+3'-03
	-.001000000000002	+.001000000000004	+3'-03
$3\frac{1}{2}\pi$	-.001000000000005	+.001000000000003	+3'-03
	-1.001000000000021	-.998999999999996	+3'-03
4π	+.998999999999996	+1.001000000000024	+3'-03
	-.001000000000005	+.001000000000003	+3'-03
$4\frac{1}{2}\pi$	-.001000000000004	+.001000000000006	+3'-03
	+.998999999999995	+1.001000000000027	+3'-03
5π	-1.001000000000030	-.998999999999995	+3'-03
	-.001000000000004	+.001000000000007	+3'-03
$5\frac{1}{2}\pi$	-.001000000000011	+.001000000000005	+3'-03
	-1.001000000000035	-.998999999999993	+3'-03
6π	+.998999999999991	+1.001000000000039	+3'-03
	-.001000000000013	+.001000000000007	+3'-03
$6\frac{1}{2}\pi$	-.001000000000006	+.001000000000019	+3'-03
	+.998999999999989	+1.001000000000043	+3'-03
7π	-1.001000000000048	-.998999999999987	+3'-03
	-.001000000000008	+.001000000000021	+3'-03
$7\frac{1}{2}\pi$	-.001000000000027	+.001000000000007	+3'-03
	-1.001000000000052	-.998999999999985	+3'-03
8π	+.998999999999984	+1.001000000000056	+3'-03
	-.001000000000029	+.001000000000009	+3'-03

We see that the diameter of the solution set of problem 2b almost does not increase for increasing t, while it would have increased enormously (i.e., by a factor of approximately $500^4 \approx 6.10^{10}$) if we would have chosen $\mathbb{X} = \mathbb{I}\mathbb{R}^M$).

DIFFERENTIAL SYSTEM 3.

We consider the differential system

$$(8.9) \qquad \left\{ \begin{array}{l} U_1'(t) = U_1(t) \cdot U_2(t), \\ U_2'(t) = U_1(t) - [U_2(t)]^2 \end{array} \right\} \qquad (0 \leq t \leq 2).$$

For problem 3a and problem 3b we choose as initial conditions

$$(8.10) \qquad \left\{ \begin{array}{l} U_1(0) = 1, \\ U_2(0) = 0 \end{array} \right.$$

and

$$(8.11) \qquad \left\{ \begin{array}{l} U_1(0) \in [\ 0.9999, 1.0001], \\ U_2(0) \in [-0.0001, 0.0001] \end{array} \right.$$

respectively.

The maximum order k_{max} is chosen to be 4. The procedure parameters F and G of procedure SOLVE are declared as follows.

```
'PROCEDURE' F(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN' 'TRIPLEX' U,V;
    U:=X(/1/); V:=X(/2/);
    'IF' I=Ø 'THEN'
    'BEGIN'
        Y(/1/):=U*V; Y(/2/):=U-V**2;
    'END' 'ELSE'
    'IF' I=1 'THEN'
    'BEGIN'
        Y(/1/):=U**2; Y(/2/):=-U*V+2*V**3;
    'END' 'ELSE'
    'IF' I=2 'THEN'
    'BEGIN'
        Y(/1/):=2*U**2*V; Y(/2/):=-U**2+6*U*V**2-6*V**4;
    'END' 'ELSE'
    'BEGIN'
        Y(/1/):=2*U**3+2*(U*V)**2;
        Y(/2/):=1Ø*U**2*V-3Ø*U*V**3+24*V**5;
    'END' ;
'END' F;
```

```
'PROCEDURE' G(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN' 'TRIPLEX' U,V;
    U:=X(/1/); V:=X(/2/);
    'IF' I=Ø 'THEN'
    'BEGIN'
        Y(/1,1/):=V; Y(/1,2/):=U; Y(/2,1/):=1; Y(/2,2/):=-2*V;
    'END' 'ELSE'
    'IF' I=1 'THEN'
    'BEGIN'
        Y(/1,1/):=2*U; Y(/1,2/):=Ø; Y(/2,1/):=-V;
        Y(/2,2/):=-U+6*V**2;
    'END' 'ELSE'
    'BEGIN'
        Y(/1,1/):=4*U*V; Y(/1,2/):=2*U**2;
        Y(/2,1/):=-2*U+6*V**2; Y(/2,2/):=12*V*(U-2*V**2);
    'END' ;
'END' G;
```

Further we choose

$$\text{ABSERR} = 0,$$

$$\text{RELERR} = 10^{-7},$$

$$\text{WORK(/I/)} = I + 1 \qquad (I = 0,1,2).$$

With these parameter values procedure SOLVE produces the following results.

PROBLEM 3a (196 steps performed, computation time 54 sec).

TT(/J/)	YY(/J,I/)		DIAM
.Ø	+1.00000000	+1.00000000	Ø
	.00000000	.00000000	Ø
+.2	+1.02013422	+1.02013424	+2'-Ø8
	+.19869300	+.19869303	+3'-Ø8
+.4	+1.08219153	+1.08219158	+5'-Ø8
	+.39014663	+.39014669	+6'-Ø8
+.6	+1.19148303	+1.19148314	+2'-Ø7
	+.56983919	+.56983929	+1'-Ø7
+.8	+1.35813933	+1.35813959	+3'-Ø7
	+.73756333	+.73756350	+2'-Ø7
+1.Ø	+1.59952386	+1.59952449	+7'-Ø7
	+.89765113	+.89765143	+3'-Ø7
+1.2	+1.94491190	+1.94491354	+2'-Ø6
	+1.05847353	+1.05847413	+6'-Ø7
+1.4	+2.44475395	+2.44475871	+5'-Ø6
	+1.23219075	+1.23219215	+2'-Ø6
+1.6	+3.19012706	+3.19014311	+2'-Ø5
	+1.43570321	+1.43570709	+4'-Ø6
+1.8	+4.35724463	+4.35731072	+7'-Ø5
	+1.69402347	+1.69403671	+2'-Ø5
+2.Ø	+6.32181080	+6.32216318	+4'-Ø4
	+2.04886788	+2.04892556	+6'-Ø5

PROBLEM 3b (198 steps performed, computation time 54 sec).

TT(/J/)	YY(/J,I/)		DIAM
.Ø	+.99989999	+1.00010001	+3'-Ø4
	-.00010001	+.00010001	+3'-Ø4
+.2	+1.02000990	+1.02025855	+3'-Ø4
	+.19857518	+.19881084	+3'-Ø4
+.4	+1.08203262	+1.08235049	+4'-Ø4
	+.39001551	+.39027781	+3'-Ø4
+.6	+1.19127555	+1.19169061	+5'-Ø4
	+.56969798	+.56998051	+3'-Ø4
+.8	+1.35786312	+1.35841580	+6'-Ø4
	+.73741239	+.73771444	+4'-Ø4
+1.Ø	+1.59914814	+1.59990021	+8'-Ø4
	+.89748701	+.89781555	+4'-Ø4
+1.2	+1.94438696	+1.94543847	+2'-Ø3
	+1.05828830	+1.05865936	+4'-Ø4
+1.4	+2.44399411	+2.44551854	+2'-Ø3
	+1.23197069	+1.23241221	+5'-Ø4
+1.6	+3.18897121	+3.19129895	+3'-Ø3
	+1.43542523	+1.43598507	+6'-Ø4
+1.8	+4.35534170	+4.35921365	+4'-Ø3
	+1.69364318	+1.69441700	+8'-Ø4
+2.Ø	+6.31802738	+6.32594659	+8'-Ø3
	+2.04824889	+2.04954456	+2'-Ø3

DIFFERENTIAL SYSTEM 4.

We adopt from HUNGER [1971] the 4'th order differential equation

$$V''''(t) = 6V(t) \cdot [2[V'(t)]^2 + V(t)V''(t)] \qquad (0 \le t \le \tfrac{1}{10}). \tag{8.12}$$

For problem 4a and problem 4b we choose as initial conditions

$$\begin{cases} V(0) = 1, \\ V'(0) = -1, \\ V''(0) = 2, \\ V'''(0) = -6 \end{cases} \tag{8.13}$$

and

$$\begin{cases} V(0) \in [0.999,1.001], \\ V'(0) \in -[0.999,1.001], \\ V''(0) \in 2.[0.999,1.001], \\ V'''(0) \in -6.[0.999,1.001], \end{cases} \tag{8.14}$$

respectively. Transforming (8.12) into a system of 4 first order differential equations we obtain

$$\left.\begin{cases} U_1'(t) = U_2(t), \\ U_2'(t) = U_3(t), \\ U_3'(t) = U_4(t), \\ U_4'(t) = 6U_1(t) \cdot [2[U_2(t)]^2 + U_1(t)U_3(t)] \end{cases}\right\} \qquad (0 \le t \le \tfrac{1}{10}). \tag{8.15}$$

Of course the initial conditions are transformed correspondingly.

Problem 4a has the solution

$$\begin{cases} U_1(t) = V(t) = (1+t)^{-1}, \\ U_2(t) = -(1+t)^{-2}, \\ U_3(t) = 2(1+t)^{-3}, \\ U_4(t) = -6(1+t)^{-4}. \end{cases} \tag{8.16}$$

The maximum order k_{max} is chosen to be 6. The procedure parameters F and G of procedure SOLVE are declared as follows.

```
'PROCEDURE' F(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN' 'INTEGER' J,L,P,Q; 'TRIPLEX' SUM,PROD;
    'FOR' J:=1,2,3,4 'DO'
    'BEGIN'
        P:=I+J; SUM:=Ø;
        'FOR' Q:=1 'STEP' 1 'UNTIL' NT(/P/) 'DO'
        'BEGIN'
            PROD:=FPAR(/P,Q,Ø/);
            'FOR' L:=1,2,3,4 'DO'
                PROD:=PROD*X(/L/)**FPAR(/P,Q,L/);
            SUM:=SUM+PROD;
        'END' Q-LOOP;
        Y(/J/):=SUM;
    'END' J-LOOP;
'END' F;


'PROCEDURE' G(I,X,Y); 'VALUE' I; 'INTEGER' I;
'TRIPLEX' 'ARRAY' X,Y;
'BEGIN' 'INTEGER' J,K,L,P,Q; 'TRIPLEX' SUM,PROD;
    'FOR' J:=1,2,3,4 'DO'
    'BEGIN'
        P:=I+J;
        'FOR' K:=1,2,3,4 'DO'
        'BEGIN'
            SUM:=Ø;
            'FOR' Q:=1 'STEP' 1 'UNTIL' NT(/P/) 'DO'
                'IF' FPAR(/P,Q,K/)>Ø 'THEN'
            'BEGIN'
                PROD:=FPAR(/P,Q,Ø/);
                'FOR' L:=1,2,3,4 'DO'
                    'IF' L]=K 'THEN' PROD:=PROD*X(/L/)**
                FPAR(/P,Q,L/);
                PROD:=PROD*X(/K/)**(FPAR(/P,Q,K/)-1)*
                FPAR(/P,Q,K/); SUM:=SUM+PROD;
            'END' Q-LOOP;
            Y(/J,K/):=SUM;
        'END' K-LOOP;
    'END' J-LOOP;
'END' G;
```

The values of the integer arrays NT(/1 : 9/) and FPAR (/1 : 9, 1 : 10, 0 : 4/) are read with the statements

```
'FOR' P:=1 'STEP' 1 'UNTIL' 9 'DO'
'BEGIN'
    ININTEGER(0,NT(/P/));
    'FOR' Q:=1 'STEP' 1 'UNTIL' NT(/P/) 'DO'
    'FOR' S:=0,1,2,3,4 'DO'
        ININTEGER(0,FPAR(/P,Q,S/));
'END' P-LOOP;
```

from the following dataset.

```
 1          1   0   1   0   0

 1          1   0   0   1   0

 1          1   0   0   0   1

 2          6   2   0   1   0        12   1   2   0   0

 3          6   2   0   0   1        36   1   1   1   0
           12   0   3   0   0

 5         36   4   0   1   0        72   3   2   0   0
           36   1   0   2   0        48   1   1   0   1
           72   0   2   1   0

 6         36   4   0   0   1       576   3   1   1   0
          792   2   3   0   0       120   1   0   1   1
          120   0   2   0   1       180   0   1   2   0

 9        216   6   0   1   0       432   5   2   0   0
          720   3   1   0   1      1296   3   0   2   0
         6264   2   2   1   0      3024   1   4   0   0
          120   1   0   0   2       720   0   1   1   1
          180   0   0   3   0

10        216   6   0   0   1      6480   5   1   1   0
        10800   4   3   0   0      4752   3   0   1   1
        11304   2   2   0   1     20736   2   1   2   0
        33264   1   3   1   0      3024   0   5   0   0
          840   0   1   0   2      1260   0   0   2   1
```

Further we choose

$$\text{ABSERR} = 10^{-10},$$
$$\text{RELERR} = 10^{-10},$$
$$\text{WORK}(/I/) = I + 1 \qquad (0 \le I \le 4).$$

With these parameter values procedure SOLVE produces the following results.

PROBLEM 4a (15 steps performed, computation time 46 sec).

TT(/J/)	YY(/J,I/)		DIAM
.00	+1.0000000000000	+1.0000000000000	0
	-1.0000000000000	-1.0000000000000	0
	+2.0000000000000	+2.0000000000000	0
	-6.0000000000000	-6.0000000000000	0
+.05	+.9523809523808	+.9523809523811	+3'-13
	-.9070294784584	-.9070294784576	+8'-13
	+1.7276751970594	+1.7276751970651	+6'-12
	-4.9362148487601	-4.9362148487267	+4'-11
+.10	+.9090909090907	+.9090909090911	+4'-13
	-.8264462809928	-.8264462809904	+3'-12
	+1.5026296017946	+1.5026296018097	+2'-11
	-4.0980807322075	-4.0980807321407	+7'-11

PROBLEM 4b (16 steps performed, computation time 49 sec).

TT(/J/)	YY(/J,I/)		DIAM
.00	+.9989999999999	+1.0010000000001	+3'-03
	-1.0010000000001	-.9989999999999	+3'-03
	+1.9979999999999	+2.0020000000001	+5'-03
	-6.0060000000001	-5.9939999999999	+2'-02
+.05	+.9513283205366	+.9534335842253	+3'-03
	-.9081374873833	-.9059214695326	+3'-03
	+1.7252901146652	+1.7300602794596	+5'-03
	-4.9454981258636	-4.9269315716257	+2'-02
+.10	+.9079798025295	+.9102020156523	+3'-03
	-.8276802271459	-.8252123348373	+3'-03
	+1.4997082227064	+1.5055509808982	+6'-03
	-4.1101182275942	-4.0860432367594	+3'-02

REFERENCES

ALEFELD, G. & J. HERZBERGER [1974], *Einführung in die Intervallrechnung*, Reihe Informatik 12, Bibliographisches Institut, Mannheim.

APOSTOLATOS, N. & U. KULISCH [1968], *Grundzüge einer Intervallrechnung für Matrizen und einige Anwendungen*, Elektr. Rechenanlagen 10, 73-83.

BAUCH, H. [1977], *Zur Lösungseinschliessung bei Anfangswertaufgaben gewöhnlicher Differentialgleichungen nach der Defektmethode*, ZAMM 57, 387-396.

CONRADT, J. [1980], *Ein Intervallverfahren zur Einschliessung des Fehlers einer Näherungslösung bei Anfangswertaufgaben für Systeme von gewöhnlichen Differentialgleichungen*, Freiburger Intervallberichte 80/1, Institut für Angewandte Mathematik, Universität Freiburg i. Br.

COPPEL, W.A. [1965], *Stability and asymptotic behaviour of differential equations*, D.C. Heath and Co., Boston.

DAHLQUIST, G. [1959], *Stability and error bounds in the numerical integration of ordinary differential equations*, Kungl. Tekn. Högsk. Handl. Stockholm, no. 130.

EIJGENRAAM, P., J.A. van de GRIEND & L.S.C. STATEMA [1976], *Triplex-Algol 60 on the IBM 370/158 of the "Centraal Rekeninstituut"*, report no. 76-1, Institute of Applied Mathematics and Computer Science, University of Leiden.

GEAR, C.W. [1971], *Numerical initial value problems in ordinary differential equations*, Prentice-Hall Inc., Englewood Cliffs, New Jersey.

HANSEN, E. [1965], *Interval arithmetic in matrix computations*, part I, S.I.A.M. J. Numer. Anal. 2, 308-320.

HAUSDORFF, F. [1914], *Grundzüge der Mengenlehre*, Von Veit & Comp., Leipzig.

HENRICI, P. [1962], *Discrete variable methods in ordinary differential equations*, J. Wiley & Sons, New York.

HUNGER, S. [1971], *Intervallanalytische Defektabschätzung bei Anfangswertaufgaben für Systeme gewöhnlicher Differentialgleichungen*, Berichte der Gesellschaft für Mathematik und Datenverarbeitung, nr. 41, Bonn.

JACKSON, L.W. [1975], *Interval arithmetic error-bounding algorithms*, S.I.A.M. J. Numer. Anal. 12, 223-238.

KRÜCKEBERG, F. [1969], *Ordinary differential equations*, "Topics in interval analysis", ed. E. Hansen, Oxford University Press, 91-97.

KULISCH, U. [1976], *Grundlagen des numerisches Rechnens*, Reihe Informatik 19, Bibliographisches Institut Mannheim.

LOHNER, R. & E. ADAMS [1978], *On initial value problems in $\mathbb{R}^n$ with intervals for both initial data and a parameter in the differential equations*, Technical Report 8, Center for Applied Mathematics, University of Georgia.

LOZINSKIĬ, S.M. [1958], *Error estimates for the numerical integration of ordinary differential equations I (Russian)*, Izv. Vyssš. Učebn. Zaved. Matematika, no. 5 (6) 52-90.

MARCOWITZ, U. [1973], *Fehlerabschätzung bei Anfangswertaufgaben für Systeme von gewöhnlichen Differentialgleichungen mit Anwendung auf das Problem des Wiedereintritts eines Raumfahrzeugs in die Lufthülle der Erde*, Dissertation, Universität Köln.

MARCOWITZ, U. [1975], *Fehlerabschätzung bei Anfangswertaufgaben für Systeme von gewöhnlichen Differentialgleichungen mit Anwendung auf das Reentryproblem*, Numer. Math. 24, 249-275.

MOORE, R.E. [1966], *Interval analysis*, Prentice Hall, Inc., Englewood Cliffs, New Jersey.

MOORE, R.E. [1979], *Methods and applications of interval analysis*, S.I.A.M., Philadelphia.

NAG Library Manual [1974], Mark 4, Nottingham Algorithm Group, Oxford.

NICKEL, K. [1979], *Schranken für die Lösungsmengen von Funktional-Differential-Gleichungen*, Freiburger Intervallberichte 79/4, Institut für Angewandte Mathematik, Universität Freiburg i. Br.

SHAMPINE, L.F. & M.K. GORDON [1975], *Computer solution of ordinary differential equations, the initial value problem*, W.H. Freeman & Co., San Francisco.

SMART, D.R. [1974], *Fixed point theorems*, Cambridge University Press.

STERN, K. [1980], *Fehlerabschätzungen von Anfangswertaufgaben bei Systemen von gewöhnlichen Differentialgleichungen*, Diplomarbeit, Institut für Angewandte Mathematik, Universität Freiburg i. Br.

STOER, J. [1972], *Einführung in die numerische Mathematik I*, Heidelberger Taschenbücher 105, Springer Verlag, Berlin.

WIPPERMANN, H.-W. [1968] (ed.), N. APOSTOLATOS, R. KRAWCZYK, U. KULISCH, B. LORTZ, K. NICKEL & H.-W. WIPPERMANN, *The algorithmic language Triplex-Algol 60*, Numer. Math. 11, 175-180.

YOUNG, R.C. [1931], *The algebra of many-valued quantities*, Math. Ann. 104, 260-290.

SUBJECT INDEX

SYMBOL INDEX

TITLES IN THE SERIES MATHEMATICAL CENTRE TRACTS

(An asterisk before the MCT number indicates that the tract is under preparation).

A leaflet containing an order form and abstracts of all publications mentioned below is available at the Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Orders should be sent to the same address.

MCT 1 T. VAN DER WALT, *Fixed and almost fixed points*, 1963. ISBN 90 6196 002 9.

MCT 2 A.R. BLOEMENA, *Sampling from a graph*, 1964. ISBN 90 6196 003 7.

MCT 3 G. DE LEVE, *Generalized Markovian decision processes, part I: Model and method*, 1964. ISBN 90 6196 004 5.

MCT 4 G. DE LEVE, *Generalized Markovian decision processes, part II: Probabilistic background*, 1964. ISBN 90 6196 005 3.

MCT 5 G. DE LEVE, H.C. TIJMS & P.J. WEEDA, *Generalized Markovian decision processes, Applications*, 1970. ISBN 90 6196 051 7.

MCT 6 M.A. MAURICE, *Compact ordered spaces*, 1964. ISBN 90 6196 006 1.

MCT 7 W.R. VAN ZWET, *Convex transformations of random variables*, 1964. ISBN 90 6196 007 X.

MCT 8 J.A. ZONNEVELD, *Automatic numerical integration*, 1964. ISBN 90 6196 008 8.

MCT 9 P.C. BAAYEN, *Universal morphisms*, 1964. ISBN 90 6196 009 6.

MCT 10 E.M. DE JAGER, *Applications of distrubutions in mathematical physics*, 1964. ISBN 90 6196 010 X.

MCT 11 A.B. PAALMAN-DE MIRANDA, *Topological semigroups*, 1964. ISBN 90 6196 011 8.

MCT 12 J.A.Th.M. VAN BERCKEL, H. BRANDT CORSTIUS, R.J. MOKKEN & A. VAN WIJNGAARDEN, *Formal properties of newspaper Dutch*, 1965. ISBN 90 6196 013 4.

MCT 13 H.A. LAUWERIER, *Asymptotic expansions*, 1966, out of print; replaced by MCT 54.

MCT 14 H.A. LAUWERIER, *Calculus of variations in mathematical physics*, 1966. ISBN 90 6196 020 7.

MCT 15 R. DOORNBOS, *Slippage tests*, 1966. ISBN 90 6196 021 5.

MCT 16 J.W. DE BAKKER, *Formal definition of programming languages with an application to the definition of ALGOL 60*, 1967. ISBN 90 6196 022 3.

MCT 17 R.P. VAN DE RIET, *Formula manipulation in ALGOL 60, part 1*, 1968. ISBN 90 6196 025 8.

MCT 18 R.P. VAN DE RIET, *Formula manipulation in ALGOL 60, part 2*, 1968. ISBN 90 6196 038 X.

MCT 19 J. VAN DER SLOT, *Some properties related to compactness*, 1968. ISBN 90 6196 026 6.

MCT 20 P.J. VAN DER HOUWEN, *Finite difference methods for solving partial differential equations*, 1968. ISBN 90 6196 027 4.

MCT 21 E. WATTEL, *The compactness operator in set theory and topology*, 1968. ISBN 90 6196 028 2.

MCT 22 T.J. DEKKER, *ALGOL 60 procedures in numerical algebra, part 1*, 1968. ISBN 90 6196 029 0.

MCT 23 T.J. DEKKER & W. HOFFMANN, *ALGOL 60 procedures in numerical algebra, part 2*, 1968. ISBN 90 6196 030 4.

MCT 24 J.W. DE BAKKER, *Recursive procedures*, 1971. ISBN 90 6196 060 6.

MCT 25 E.R. PAËRL, *Representations of the Lorentz group and projective geometry*, 1969. ISBN 90 6196 039 8.

MCT 26 EUROPEAN MEETING 1968, *Selected statistical papers, part I*, 1968. ISBN 90 6196 031 2.

MCT 27 EUROPEAN MEETING 1968, *Selected statistical papers, part II*, 1969. ISBN 90 6196 040 1.

MCT 28 J. OOSTERHOFF, *Combination of one-sided statistical tests*, 1969. ISBN 90 6196 041 X.

MCT 29 J. VERHOEFF, *Error detecting decimal codes*, 1969. ISBN 90 6196 042 8.

MCT 30 H. BRANDT CORSTIUS, *Exercises in computational linguistics*, 1970. ISBN 90 6196 052 5.

MCT 31 W. MOLENAAR, *Approximations to the Poisson, binomial and hypergeometric distribution functions*, 1970. ISBN 90 6196 053 3.

MCT 32 L. DE HAAN, *On regular variation and its application to the weak convergence of sample extremes*, 1970. ISBN 90 6196 054 1.

MCT 33 F.W. STEUTEL, *Preservation of infinite divisibility under mixing and related topics*, 1970. ISBN 90 6196 061 4.

MCT 34 I. JUHÁSZ, A. VERBEEK & N.S. KROONENBERG, *Cardinal functions in topology*, 1971. ISBN 90 6196 062 2.

MCT 35 M.H. VAN EMDEN, *An analysis of complexity*, 1971. ISBN 90 6196 063 0.

MCT 36 J. GRASMAN, *On the birth of boundary layers*, 1971. ISBN 90 6196 064 9.

MCT 37 J.W. DE BAKKER, G.A. BLAAUW, A.J.W. DUIJVESTIJN, E.W. DIJKSTRA, P.J. VAN DER HOUWEN, G.A.M. KAMSTEEG-KEMPER, F.E.J. KRUSEMAN ARETZ, W.L. VAN DER POEL, J.P. SCHAAP-KRUSEMAN, M.V. WILKES & G. ZOUTENDIJK, *MC-25 Informatica Symposium* 1971. ISBN 90 6196 065 7.

MCT 38 W.A. VERLOREN VAN THEMAAT, *Automatic analysis of Dutch compound words*, 1971. ISBN 90 6196 073 8.

MCT 39 H. BAVINCK, *Jacobi series and approximation*, 1972. ISBN 90 6196 074 6.

MCT 40 H.C. TIJMS, *Analysis of (s,S) inventory models*, 1972. ISBN 90 6196 075 4.

MCT 41 A. VERBEEK, *Superextensions of topological spaces*, 1972. ISBN 90 6196 076 2.

MCT 42 W. VERVAAT, *Success epochs in Bernoulli trials (with applications in number theory)*, 1972. ISBN 90 6196 077 0.

MCT 43 F.H. RUYMGAART, *Asymptotic theory of rank tests for independence*, 1973. ISBN 90 6196 081 9.

MCT 44 H. BART, *Meromorphic operator valued functions*, 1973. ISBN 90 6196 082 7.

MCT 45 A.A. BALKEMA, *Monotone transformations and limit laws* 1973. ISBN 90 6196 083 5.

MCT 46 R.P. VAN DE RIET, *ABC ALGOL, A portable language for formula manipulation systems, part 1: The language*, 1973. ISBN 90 6196 084 3.

MCT 47 R.P. VAN DE RIET, *ABC ALGOL, A portable language for formula manipulation systems, part 2: The compiler*, 1973. ISBN 90 6196 085 1.

MCT 48 F.E.J. KRUSEMAN ARETZ, P.J.W. TEN HAGEN & H.L. OUDSHOORN, *An ALGOL 60 compiler in ALGOL 60, Text of the MC-compiler for the EL-X8*, 1973. ISBN 90 6196 086 X.

MCT 49 H. KOK, *Connected orderable spaces*, 1974. ISBN 90 6196 088 6.

MCT 50 A. VAN WIJNGAARDEN, B.J. MAILLOUX, J.E.L. PECK, C.H.A. KOSTER, M. SINTZOFF, C.H. LINDSEY, L.G.L.T. MEERTENS & R.G. FISKER (eds), *Revised report on the algorithmic language ALGOL 68*, 1976. ISBN 90 6196 089 4.

MCT 51 A. HORDIJK, *Dynamic programming and Markov potential theory*, 1974. ISBN 90 6196 095 9.

MCT 52 P.C. BAAYEN (ed.), *Topological structures*, 1974. ISBN 90 6196 096 7.

MCT 53 M.J. FABER, *Metrizability in generalized ordered spaces*, 1974. ISBN 90 6196 097 5.

MCT 54 H.A. LAUWERIER, *Asymptotic analysis, part 1*, 1974. ISBN 90 6196 098 3.

MCT 55 M. HALL JR. & J.H. VAN LINT (eds), *Combinatorics, part 1: Theory of designs, finite geometry and coding theory*, 1974. ISBN 90 6196 099 1.

MCT 56 M. HALL JR. & J.H. VAN LINT (eds), *Combinatorics, part 2: Graph theory, foundations, partitions and combinatorial geometry*, 1974. ISBN 90 6196 100 9.

MCT 57 M. HALL JR. & J.H. VAN LINT (eds), *Combinatorics, part 3: Combinatorial group theory*, 1974. ISBN 90 6196 101 7.

MCT 58 W. ALBERS, *Asymptotic expansions and the deficiency concept in statistics*, 1975. ISBN 90 6196 102 5.

MCT 59 J.L. MIJNHEER, *Sample path properties of stable processes*, 1975. ISBN 90 6196 107 6.

MCT 60 F. GÖBEL, *Queueing models involving buffers*, 1975. ISBN 90 6196 108 4.

*MCT 61 P. VAN EMDE BOAS, *Abstract resource-bound classes, part 1*, ISBN 90 6196 109 2.

*MCT 62 P. VAN EMDE BOAS, *Abstract resource-bound classes, part 2*, ISBN 90 6196 110 6.

MCT 63 J.W. DE BAKKER (ed.), *Foundations of computer science*, 1975. ISBN 90 6196 111 4.

MCT 64 W.J. DE SCHIPPER, *Symmetric closed categories*, 1975. ISBN 90 6196 112 2.

MCT 65 J. DE VRIES, *Topological transformation groups 1 A categorical approach*, 1975. ISBN 90 6196 113 0.

MCT 66 H.G.J. PIJLS, *Locally convex algebras in spectral theory and eigenfunction expansions*, 1976. ISBN 90 6196 114 9.

*MCT 67 H.A. LAUWERIER, *Asymptotic analysis, part 2*, ISBN 90 6196 119 X.

MCT 68 P.P.N. DE GROEN, *Singularly perturbed differential operators of second order*, 1976. ISBN 90 6196 120 3.

MCT 69 J.K. LENSTRA, *Sequencing by enumerative methods*, 1977. ISBN 90 6196 125 4.

MCT 70 W.P. DE ROEVER JR., *Recursive program schemes: Semantics and proof theory*, 1976. ISBN 90 6196 127 0.

MCT 71 J.A.E.E. VAN NUNEN, *Contracting Markov decision processes*, 1976. ISBN 90 6196 129 7.

MCT 72 J.K.M. JANSEN, *Simple periodic and nonperiodic Lamé functions and their applications in the theory of conical waveguides*, 1977. ISBN 90 6196 130 0.

MCT 73 D.M.R. LEIVANT, *Absoluteness of intuitionistic logic*, 1979. ISBN 90 6196 122 X.

MCT 74 H.J.J. TE RIELE, *A theoretical and computational study of generalized aliquot sequences*, 1976. ISBN 90 6196 131 9.

MCT 75 A.E. BROUWER, *Treelike spaces and related connected topological spaces*, 1977. ISBN 90 6196 132 7.

MCT 76 M. REM, *Associations and the closure statement*, 1976. ISBN 90 6196 135 1.

MCT 77 W.C.M. KALLENBERG, *Asymptotic optimality of likelihood ratio tests in exponential families*, 1977. ISBN 90 6196 134 3.

MCT 78 E. DE JONGE & A.C.M. VAN ROOIJ, *Introduction to Riesz spaces*, 1977. ISBN 90 6196 133 5.

MCT 79 M.C.A. VAN ZUIJLEN, *Empirical distributions and rank statistics*, 1977. ISBN 90 6196 145 9.

MCT 80 P.W. HEMKER, *A numerical study of stiff two-point boundary problems*, 1977. ISBN 90 6196 146 7.

MCT 81 K.R. APT & J.W. DE BAKKER (eds), *Foundations of computer science II*, part 1, 1976. ISBN 90 6196 140 8.

MCT 82 K.R. APT & J.W. DE BAKKER (eds), *Foundations of computer science II*, part 2, 1976. ISBN 90 6196 141 6.

MCT 83 L.S. BENTHEM JUTTING, *Checking Landau's "Grundlagen" in the AUTOMATH system*, 1979. ISBN 90 6196 147 5.

MCT 84 H.L.L. BUSARD, *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?) books vii-xii*, 1977. ISBN 90 6196 148 3.

MCT 85 J. VAN MILL, *Supercompactness and Wallman spaces*, 1977. ISBN 90 6196 151 3.

MCT 86 S.G. VAN DER MEULEN & M. VELDHORST, *Torrix I, A programming system for operations on vectors and matrices over arbitrary fields and of variable size*. 1978. ISBN 90 6196 152 1.

*MCT 87 S.G. VAN DER MEULEN & M. VELDHORST, *Torrix II*, ISBN 90 6196 153 X.

MCT 88 A. SCHRIJVER, *Matroids and linking systems*, 1977. ISBN 90 6196 154 8.

MCT 89 J.W. DE ROEVER, *Complex Fourier transformation and analytic functionals with unbounded carriers*, 1978. ISBN 90 6196 155 6.

MCT 90 L.P.J. GROENEWEGEN, *Characterization of optimal strategies in dynamic games*, 1981 . ISBN 90 6196 156 4.

MCT 91 J.M. GEYSEL, *Transcendence in fields of positive characteristic*, 1979. ISBN 90 6196 157 2.

MCT 92 P.J. WEEDA, *Finite generalized Markov programming*, 1979. ISBN 90 6196 158 0.

MCT 93 H.C. TIJMS & J. WESSELS (eds), *Markov decision theory*, 1977. ISBN 90 6196 160 2.

MCT 94 A. BIJLSMA, *Simultaneous approximations in transcendental number theory*, 1978. ISBN 90 6196 162 9.

MCT 95 K.M. VAN HEE, *Bayesian control of Markov chains*, 1978. ISBN 90 6196 163 7.

MCT 96 P.M.B. VITÁNYI, *Lindenmayer systems: Structure, languages, and growth functions*, 1980. ISBN 90 6196 164 5.

*MCT 97 A. FEDERGRUEN, *Markovian control problems; functional equations and algorithms*, . ISBN 90 6196 165 3.

MCT 98 R. GEEL, *Singular perturbations of hyperbolic type*, 1978. ISBN 90 6196 166 1.

MCT 99 J.K. LENSTRA, A.H.G. RINNOOY KAN & P. VAN EMDE BOAS, *Interfaces between computer science and operations research*, 1978. ISBN 90 6196 170 X.

MCT 100 P.C. BAAYEN, D. VAN DULST & J. OOSTERHOFF (eds), *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1*, 1979. ISBN 90 6196 168 8.

MCT 101 P.C. BAAYEN, D. VAN DULST & J. OOSTERHOFF (eds), *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2*, 1979. ISBN 90 6196 169 6.

MCT 102 D. VAN DULST, *Reflexive and superreflexive Banach spaces*, 1978. ISBN 90 6196 171 8.

MCT 103 K. VAN HARN, *Classifying infinitely divisible distributions by functional equations*, 1978. ISBN 90 6196 172 6.

MCT 104 J.M. VAN WOUWE, *Go-spaces and generalizations of metrizability*, 1979. ISBN 90 6196 173 4.

*MCT 105 R. HELMERS, *Edgeworth expansions for linear combinations of order statistics*, . ISBN 90 6196 174 2.

MCT 106 A. SCHRIJVER (ed.), *Packing and covering in combinatorics*, 1979. ISBN 90 6196 180 7.

MCT 107 C. DEN HEIJER, *The numerical solution of nonlinear operator equations by imbedding methods*, 1979. ISBN 90 6196 175 0.

MCT 108 J.W. DE BAKKER & J. VAN LEEUWEN (eds), *Foundations of computer science III*, part 1, 1979. ISBN 90 6196 176 9.

MCT 109 J.W. DE BAKKER & J. VAN LEEUWEN (eds), *Foundations of computer science III*, part 2, 1979. ISBN 90 6196 177 7.

MCT 110 J.C. VAN VLIET, *ALGOL 68 transput*, part I: *Historical review and discussion of the implementation model*, 1979. ISBN 90 6196 178 5.

MCT 111 J.C. VAN VLIET, *ALGOL 68 transput*, part II: *An implementation model*, 1979. ISBN 90 6196 179 3.

MCT 112 H.C.P. BERBEE, *Random walks with stationary increments and renewal theory*, 1979. ISBN 90 6196 182 3.

MCT 113 T.A.B. SNIJDERS, *Asymptotic optimality theory for testing problems with restricted alternatives*, 1979. ISBN 90 6196 183 1.

MCT 114 A.J.E.M. JANSSEN, *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes*, 1979. ISBN 90 6196 184 X.

MCT 115 P.C. BAAYEN & J. VAN MILL (eds), *Topological Structures II*, part 1, 1979. ISBN 90 6196 185 5.

MCT 116 P.C. BAAYEN & J. VAN MILL (eds), *Topological Structures II*, part 2, 1979. ISBN 90 6196 186 6.

MCT 117 P.J.M. KALLENBERG, *Branching processes with continuous state space*, 1979. ISBN 90 6196 188 2.

MCT 118 P. GROENEBOOM, *Large deviations and asymptotic efficiencies*, 1980. ISBN 90 6196 190 4.

MCT 119 F. J. PETERS, *Sparse matrices and substructures, with a novel implementation of finite element algorithms*, 1980. ISBN 90 6196 192 0.

MCT 120 W.P.M. DE RUYTER, *On the asymptotic analysis of large-scale ocean circulation*, 1980. ISBN 90 6196 192 9.

MCT 121 W.H. HAEMERS, *Eigenvalue techniques in design and graph theory*, 1980. ISBN 90 6196 194 7.

MCT 122 J.C.P. BUS, *Numerical solution of systems of nonlinear equations*, 1980. ISBN 90 6196 195 5.

MCT 123 I. YUHÁSZ, *Cardinal functions in topology - ten years later*, 1980. ISBN 90 6196 196 3.

MCT 124 R.D. GILL, *Censoring and stochastic integrals*, 1980. ISBN 90 6196 197 1.

MCT 125 R. EISING, *2-D systems, an algebraic approach*, 1980. ISBN 90 6196 198 X.

MCT 126 G. VAN DER HOEK, *Reduction methods in nonlinear programming*, 1980. ISBN 90 6196 199 8.

MCT 127 J.W. KLOP, *Combinatory reduction systems*, 1980. ISBN 90 6196 200 5.

MCT 128 A.J.J. TALMAN, *Variable dimension fixed point algorithms and triangulations*, 1980. ISBN 90 6196 201 3.

MCT 129 G. VAN DER LAAN, *Simplicial fixed point algorithms*, 1980. ISBN 90 6196 202 1.

MCT 130 P.J.W. TEN HAGEN et al., *ILP Intermediate language for pictures*, 1980. ISBN 90 6196 204 8.

MCT 131 R.J.R. BACK, *Correctness preserving program refinements: Proof theory and applications*, 1980. ISBN 90 6196 207 2.

MCT 132 H.M. MULDER, *The interval function of a graph*, 1980. ISBN 90 6196 208 0.

MCT 133 C.A.J. KLAASSEN, *Statistical performance of location estimators*, 1981. ISBN 90 6196 209 9.

MCT 134 J.C. VAN VLIET & H. WUPPER (eds), *Proceedings international conference on ALGOL 68*, 1981. ISBN 90 6196 210 2.

MCT 135 J.A.G. GROENENDIJK, T.M.V. JANSSEN & M.J.B. STOKHOF (eds), *Formal methods in the study of language*, part I, 1981. ISBN 90 6196 211 0.

MCT 136 J.A.G. GROENENDIJK, T.M.V. JANSSEN & M.J.B. STOKHOF (eds), *Formal methods in the study of language*, part II, 1981. ISBN 90 6196 213 7.

MCT 137 J. TELGEN, *Redundancy and linear programs*, 1981. ISBN 90 6196 215 3.

MCT 138 H.A. LAUWERIER, *Mathematical models of epidemics*, 1981. ISBN 90 6196 216 1.

MCT 139 J. VAN DER WAL, *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games*, 1980. ISBN 90 6196 218 8.

MCT 140 J.H. VAN GELDROP, *A mathematical theory of pure exchange economies without the no-critical-point hypothesis*, 1981.
ISBN 90 6196 219 6.

MCT 141 G.E. WELTERS, *Abel-Jacobi isogenies for certain types of Fano threefolds*, 1981.
ISBN 90 6196 227 7.

MCT 142 H.R. BENNETT & D.J. LUTZER (eds), *Topology and order structures*, part 1, 1981.
ISBN 90 6196 228 5.

MCT 143 H. SCHUMACHER, *Dynamic feedback in finite- and infinite-dimensional linear systems*, 1981.
ISBN 90 6196 229 3.

MCT 144 P. EIJGENRAAM, *The solution of initial value problems using interval arithmetic. Formulation and analysis of an algorithm*, 1981.
ISBN 90 6196 230 7.

MCT 145 A.J. BRENTJES, *Multi-dimensional continued fraction algorithms*, 1981. ISBN 90 6196 231 5.

MCT 146 C. VAN DER MEE, *Semigroup and factorization methods in transport theory*, 1982. ISBN 90 6196 233 1.

An asterisk before the number means "to appear".