THE NUMERICAL SOLUTION OF NONLINEAR OPERATOR EQUATIONS BY IMBEDDING METHODS

THE NUMERICAL SOLUTION OF NONLINEAR OPERATOR EQUATIONS BY IMBEDDING METHODS

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN
DOCTOR IN DE WISKUNDE EN NATUURWETENSCHAPPEN
AAN DE RIJKSUNIVERSITEIT TE LEIDEN, OP GEZAG VAN DE
RECTOR MAGNIFICUS DR. D.J. KUENEN,
HOOGLERAAR IN DE FACULTEIT DER WISKUNDE EN NATUURWETENSCHAPPEN,
VOLGENS BESLUIT VAN HET COLLEGE VAN DEKANEN
TE VERDEDIGEN OP WOENSDAG 14 FEBRUARI 1979
TE KLOKKE 16.15 UUR

DOOR

CORNELIS DEN HEIJER

GEBOREN TE 'S-GRAVENHAGE IN 1951

1979
MATHEMATISCH CENTRUM, AMSTERDAM

PROMOTOR: PROF.DR. M.N. SPIJKER

Aan mijn ouders
Aan Patricia

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to the governing board of the "Stichting Mathematisch Centrum" for giving him the opportunity to carry out these investigations. He also thanks them for publishing this monograph.

He is grateful to Prof. A. van der Sluis and Prof. W.W.E. Wetterling for their useful comments on an early draft of the paper.

The author also thanks Mr. M. Bakker and Mr. F.J. Burger for their programming assistance in the computational experiments.

Furthermore, he is grateful to Mrs. K.M.H. Boet-Herbert for making the necessary corrections in the English text.

Finally, he would like to thank Miss L. Brown (for typing the manuscript), D. Zwarst and his group (for the printing and binding) and T. Baanders (for designing the front cover and drawing the figures).

CONTENTS

PART I

PART II

CHAPTER 1

INTRODUCTION

1.1. IMBEDDING METHODS: A DESCRIPTION

Let E be a (real) Hilbert space, $D \subset E$ and let $F: D \to E$ be a nonlinear operator. In this monograph we shall be concerned with the numerical solution of the equation

$$(1.1.1) \qquad F(x) = 0.$$

Until further notice we assume that $D = E$.

Suppose that $x^* \in E$ is a solution of (1.1.1). A well-known method for approximating x^* is *Newton's method*. It consists of the calculation of a sequence of approximations $\{x_k\}$ where

$$(1.1.2) \qquad x_{k+1} = x_k - \Gamma(x_k)F(x_k) \qquad (k = 0,1,2,\ldots).$$

In (1.1.2) $x_0 \in E$ is a given approximation to x^* and $\Gamma(x) \equiv [F'(x)]^{-1}$ where $F'(x)$ denotes the Fréchet-derivative of F at x. Newton's method has the drawback that when the starting point x_0 is remote from x^*, the sequence $\{x_k\}$ defined in (1.1.2) generally will not converge to x^*. In many such cases *imbedding methods* - or *continuation methods*, as they are also called - appear to be able to generate a sequence $\{x_k\}$ that converges to x^*. In these methods a mapping $H: [\tau_0,\tau_1] \times E \to E$ is introduced, for which $H(t,x)$ depends continuously on t, such that

$$(1.1.3) \qquad \begin{array}{ll} H(\tau_0,x) = 0 & \text{is easily solvable} \\ \text{and} & \\ H(\tau_1,x) \equiv F(x). & \end{array}$$

$[\tau_0,\tau_1]$ is a closed interval in $\mathbb{R}$. Thus the operator F is imbedded in the family of operators $\{H(t,\cdot) \mid t \in [\tau_0,\tau_1]\}$.

We observe that if F does not depend naturally on a suitable parameter t, it is still always possible to define an H, satisfying (1.1.3), in several ways. For example, let $x_0 \in E$ be given, then

$$\text{(1.1.4)} \qquad \begin{aligned} &H\colon [0,1] \times E \to E, \\ &H(t,x) = (1-t)\{F(x) - F(x_0)\} + tF(x) \qquad (\text{for all } t \in [0,1] \text{ and } x \in E) \end{aligned}$$

satisfies the conditions (1.1.3). More generally, let $K\colon E \times E \to E$, where K may depend on F, satisfy $K(x,x) = 0$ (for all $x \in E$). Then, for any $x_0 \in E$, the operator

$$\text{(1.1.5)} \qquad \begin{aligned} &H\colon [0,1] \times E \to E, \\ &H(t,x) = (1-t)K(x,x_0) + tF(x) \qquad (\text{for all } t \in [0,1] \text{ and } x \in E) \end{aligned}$$

meets the conditions (1.1.3).

We return to our original problem and suppose that a mapping $H\colon [\tau_0,\tau_1] \times E \to E$ has been introduced satisfying (1.1.3). Instead of the single problem (1.1.1), the entire family of problems

$$\text{(1.1.6)} \qquad H(t,x) = 0 \qquad (t \in [\tau_0,\tau_1])$$

is considered. Let $u_0 \in E$ be the solution of $H(\tau_0,x) = 0$. Suppose that (1.1.6) has, for each $t \in [\tau_0,\tau_1]$, a unique solution $x = U(t)$, which depends continuously on t. (See e.g. [MEYER, 1968] for the restrictions on H which ensure that such a curve U exists. We shall not go into this type of problem.) We have

$$\text{(1.1.7)} \qquad H(t,U(t)) = 0 \qquad (\text{for all } t \in [\tau_0,\tau_1])$$

and

$$\text{(1.1.8)} \qquad U(\tau_0) = u_0, \quad U(\tau_1) = x^*.$$

Thus U defines a curve in E with starting point u_0 and with end point equal to the solution x^* of (1.1.1). We shall describe some ways in which the imbedding can be used in the numerical solution of equation (1.1.1).

a. *Discrete imbedding*

As a first possibility for approximating $x^* = U(\tau_1)$ one may approximate successively the solutions of

$$(1.1.9) \qquad H(t_i,x) = 0 \qquad (i = 0,1,\ldots,N)$$

by some numerical process (e.g. Newton's method). Here N is an integer and $\{t_0,t_1,\ldots,t_N\}$ is a partition of $[\tau_0,\tau_1]$, that is, $\tau_0 = t_0 < t_1 < \ldots < t_N = \tau_1$ holds. As starting point for the iterative process for approximating $U(t_i)$ the last iterate of the iterative process for approximating $U(t_{i-1})$ is often used ($i \geq 1$). If this approximation is close to $U(t_{i-1})$ and if $t_i - t_{i-1}$ is sufficiently small then hopefully a sequence may be generated that converges to $U(t_i)$.

This way of approximating x^* is called *discrete imbedding.*

b. *Transformation to an initial value problem*

We next consider a somewhat different way of aprpoximating $x^* = U(\tau_1)$. Assume that the mapping U, satisfying (1.1.7) is continuously differentiable on $[\tau_0,\tau_1]$ and that H has continuous partial Fréchet-derivatives. Differentiating the identity (1.1.7) with respect to t, we obtain

$$(1.1.10) \qquad \partial_1 H(t,U(t)) + \partial_2 H(t,U(t))\dot{U}(t) = 0 \qquad (\text{for all } t \in [\tau_0,\tau_1]),$$

where $\partial_1 H$ and $\partial_2 H$ are the partial Fréchet-derivatives of H with respect to t and x respectively and $\dot{U}(t)$ denotes $\frac{d}{dt}U(t)$. If we assume that $\partial_2 H(t,U(t))$ is invertible (for all $t \in [\tau_0,\tau_1]$) then (cf. (1.1.10)) U satisfies the following initial value problem:

$$(1.1.11a) \qquad \dot{U}(t) = -[\partial_2 H(t,U(t))]^{-1}\{\partial_1 H(t,U(t))\} \qquad (t \in [\tau_0,\tau_1]),$$

$$(1.1.11b) \qquad U(0) = u_0.$$

Thus $x^* = U(\tau_1)$ may be approximated by applying a numerical integration procedure to (1.1.11). The approximation of x^* thus obtained can then be used as starting point for an iterative process for approximating x^* more closely (e.g. Newton's method).

Transforming (1.1.1) into the initial value problem (1.1.11) is often called *Davidenko's method*.

c. *Iterative imbedding*

We finally notice that imbeddings of type (1.1.5) may also be used to construct iterative methods for approximating x^*. To that end, let $K: E \times E \to E$, such that $K(x,x) = 0$ (for all $x \in E$), be given. Let $x_0 \in E$ and let $H: [0,1] \times E \to E$ be of type (1.1.5). In this case the initial value problem (1.1.11) reduces to

$$\text{(1.1.12)} \quad \begin{aligned} &\dot{X}(t) = -[(1-t)\partial_1 K(X(t),x_0) + tF'(X(t))]^{-1}\{-K(X(t),x_0) + F(X(t))\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad (t \in [0,1]), \\ &X(0) = x_0. \end{aligned}$$

Computing the solution $X(t)$ of (1.1.12) at $t = 1$ be means of a given numerical integration procedure, we obtain an approximation, say $x_1 \approx X(1)$, which is uniquely determined by x_0. We thus have $x_1 = G(x_0)$, where the operator G depends only on F, K and the given numerical integration procedure. Solving (1.1.12) once more by the same numerical integration procedure, with x_0 replaced by x_1, we obtain an approximation $x_2 \approx X(1)$, which is related to x_1 by $x_2 = G(x_1)$. In this way we arrive at the iterative process

$$\text{(1.1.13)} \qquad x_{k+1} = G(x_k) \qquad (k = 0,1,2,\ldots).$$

We call this way of constructing iterative methods, *iterative imbedding*. Hereafter we shall use for iterative imbedding initial value problems that are of a rather more general type than problem (1.1.12) (cf. section 2.6).

1.2. IMBEDDING METHODS: HISTORICAL SURVEY AND CURRENT STATUS

Originally, imbedding methods were only used as tools to demonstrate the existence of solutions to operator equations. In [FICKEN, 1951] a review is given of such applications, which date back at least to the last century.

The idea of using the *discrete imbedding method* for the numerical solution of nonlinear equations seems to date back to [LAHAYE, 1934, 1935] (see also [ORTEGA & RHEINBOLDT, 1970; pp. 234-235] for a bibliography on early numerical applications). More recent investigations have been performed by e.g. [AVILA, 1974], [LAASONEN, 1970], [POMENTALE, 1974], [WACKER, 1971, 1974, 1977(a,b)] and [RIBARIČ & SELIŠKAR, 1974]. In [LEDER, 1974] and [RHEINBOLDT, 1975, 1976] adaptive methods are proposed for determining the partition $\{t_0,t_1,\ldots,t_N\}$.

The idea of *transforming* problem (1.1.1) *into the initial value problem* (1.1.11) is usually attributed to Davidenko ([DAVIDENKO, 1953]). Davidenko applied this method to a variety of problems including integral equations and matrix inversion (e.g. [DAVIDENKO, 1965(a,b), 1975]). In [RALL, 1968] an exposition of Davidenko's work is given (it also contains some translations and a bibliography). A review of still more applications (among which two-point boundary value problems) is given in [WASSERSTROM, 1973]. In [MEYER, 1968] and [BOSARGE, 1971] the method is considered, with the initial value problem (1.1.11) solved by Runge-Kutta methods.

In [GAVURIN, 1958] a somewhat different way of transforming problem (1.1.1) into an initial value problem is used. Any iterative process of type

$$x_{k+1} = x_k + P(x_k) \qquad (k = 0,1,2,\ldots), \tag{1.2.1}$$

where $x_0 \in E$ is given, can be conceived as an application of Euler's method (see e.g. [LAMBERT, 1973; p. 13]) with stepsize $h = 1$ to the initial value problem

$$\begin{aligned} &\dot{Y}(t) = P(Y(t)) \qquad (t \in [0,\infty)),\\ &Y(0) = x_0. \end{aligned} \tag{1.2.2}$$

The curve Y is called the *continuous analogue of the iterative process* (1.2.1) (see also [ROSENBLOOM, 1956], [BITTNER, 1967] and [UEBERHUBER, 1976]).

This approach is closely related to Davidenko's method (cf. [MEYER, 1968]). For example, let

$$\begin{aligned} &H\colon [0,\infty) \times E \to E,\\ &H(t,x) = F(x) - e^{-t}F(x_0) \qquad (\text{for all } t \in [0,\infty) \text{ and } x \in E). \end{aligned} \tag{1.2.3}$$

Suppose that for all $t \in [0,\infty)$ the problem $H(t,x) = 0$ has a unique solution $x = U(t)$, which depends continuously on t. We have

$$F(U(t)) - e^{-t}F(x_0) = 0, \qquad (t \in [0,\infty)). \tag{1.2.4}$$

If we assume that $F'(U(t))$ exists and is invertible (for all $t \in [0,\infty)$) then differentiating (1.2.4) with respect to t and using the identity (1.2.4) yield

$$\begin{aligned} \dot{U}(t) &= -\Gamma(U(t))F(U(t)) \qquad (t \in [0,\infty)), \\ U(0) &= x_0. \end{aligned} \tag{1.2.5}$$

Thus the curve $Y = U$, satisfying (1.2.5), is the continuous analogue of Newton's method. See also [BOGGS, 1971], [BOGGS & DENNIS, 1974], [BITTNER, 1967] and [DI LENA & TRIGIANTE, 1976].

We also mention the possibility of *combining the discrete imbedding method with Davidenko's method*. For example, let a partition $\{t_0,t_1,\ldots,t_N\}$ of $[\tau_0,\tau_1]$ be given. Then the first step of the k-th stage ($k = 1,2,\ldots,N$) of this process consists of computing an approximation u_k of $U(t_k)$ by applying a numerical integration procedure on (1.1.11a) at $[t_{k-1},t_k]$. The second step of the k-th stage consists of solving the equation (1.1.9), for $i = k$ by means of an iterative process with starting point u_k. See e.g. [DEIST & SEFOR, 1967] and [BROYDEN, 1969]. More recently in [FEILMEIER, 1972], [KUBIČEK, 1976], [MENZEL & SCHWETLICK, 1976], [SCHWETLICK, 1975, 1976], [DEUFLHARD, PESCH & RENTROP, 1976] and [DEUFLHARD, 1976] these types of methods have been investigated.

The method of *iterative imbedding* has been investigated in [KIZNER, 1964] for the case $E = \mathbb{R}$. In this paper the relation is given between the order of accuracy of a numerical integration method for solving the initial value problem (1.1.12) and the order of convergence of the corresponding iterative process (1.1.13). In [DAVIDENKO, 1966], [BITTNER, 1967], [KLEINMICHEL, 1968] and [PETRY, 1971] the iterative imbedding method is considered for the case that E is $\mathbb{R}^n$ or an arbitrary Banach space. In these papers the emphasis lies on constructing high order iterative methods of type (1.1.13). In the papers of Kleinmichel and Petry theorems on these iterative methods are given which are similar to the famous Newton-Kantorovich theorem (cf. [KANTOROWITSCH & AKILOW, 1964; Theorem 6 (1.XVIII)]).

1.3. SCOPE OF THE STUDY

In this monograph we shall be concerned mainly with *iterative imbedding*. As described in section 1.1, in this way iterative methods (1.1.13) for solving (1.1.1) can be constructed. These iterative methods are based on the operator K (determining the initial value problem (1.1.12)) and on a numerical integration procedure. We shall restrict our attention to integration procedures of (generalized) Runge-Kutta type.

In the last chapter we shall consider some *algorithms* based on *discrete imbedding*, for the solution of problem (1.1.1). In some of them *Davidenko's method* is used also.

In chapter 2 we introduce the basic concepts to be used throughout this study. In Section 2.4 we introduce the concept of a radius of convergence $r(M;\mathcal{F})$ of an iterative method M with respect to some (given) class $\mathcal{F}$ of operators F. Let x^* denote the solution of $F(x) = 0$ where F belongs to $\mathcal{F}$. Then $r(M;\mathcal{F})$ indicates how close to x^* an initial guess x_0 should be, for the iterative process to yield a sequence of approximations x_k that converges to x^*. Section 2.5 is concerned with the local convergence behaviour of an iterative process (i.e. its convergence behaviour near the solution x^* of $F(x) = 0$). We introduce two types of local convergence behaviour that lie between the well-known concepts of local and quadratic convergence. We derive both necessary and sufficient conditions for these intermediate types of local convergence - generalizing the results of [OSTROWSKI, 1960; Theorems 22.1 and 22.2] and [KITCHEN, 1966]. In section 2.6 we present the imbedding on which the iterative methods with which we shall be concerned will be based.

In chapter 3 (section 3.2) we introduce the Runge-Kutta methods that will be used for constructing the iterative processes (1.1.13). In section 3.3 we shall indicate that the problem of the nonconvergence of Newton's method is closely related to a certain type of instability of Euler's method.

In chapter 4 we derive general formulae for iterative methods which are constructed by means of iterative imbedding.

In chapter 5 we investigate the local convergence behaviour of the iterative methods that were consrtucted in chapter 4. We derive both necessary and sufficient conditions for quadratic convergence and for the two intermediate types of convergence introduced in section 2.5. Our main results in this chapter are formulated in nine theorems, that are given in sections 5.1 and 5.2. The most obvious conditions on the operator K, for which the iterative process (1.1.13) is locally convergent, are given in the

Theorems 5.2.3 - 8 (cf. subsections 5.2.2 - 3). Some of these theorems may be viewed as generalizations of the results of [KIZNER, 1964].

In chapter 6 we determine the radii of convergence of iterative methods of the type described in section 4.1. Our main results in this chapter are formulated in four theorems (Theorems 6.2.1, 6.3.1, 6.5.1 and 6.6.1). Part I of this chapter is concerned with $\mathcal{F}\langle\sigma,\beta,\gamma\rangle$, a class of operators F which is defined in section 6.1. In section 6.2 we determine the radius of convergence of Newton's method with respect to $\mathcal{F}\langle\alpha,\beta,\gamma\rangle$. A related result is given in [RHEINBOLDT, 1975] although in that paper it was only shown that the value obtained is a lower bound of the radius of convergence. In section 6.3 we prove that a class of iterative methods that are closely related to the so-called *damped Newton methods*, all have a greater radius of convergence than Newton's method. Part II of chapter 6 is concerned with $\mathcal{F}\langle\sigma,\alpha\rangle$, a class of operators F which is defined in section 6.4. In section 6.5 we determine the radius of convergence of Newton's method with respect to this class. Finally, in section 6.6 we are able to give an explicit expression of the radii of convergence (with respect to $\mathcal{F}\langle\sigma,\alpha\rangle$) of the iterative methods which were considered in section 6.3.

In chapter 7 we present numerical experiments with iterative methods which were dealt with in the preceding chapters. Iterative methods based on a generalized Runge-Kutta method related to the Backward Euler method, appear to be more successful in solving the testproblems than the other methods tested (including Newton's method).

In chapter 8 we present some algorithms for solving problem (1.1.1). In section 8.1 we describe an algorithm which is proposed in [RHEINBOLDT, 1975]. This algorithm is based on discrete imbedding and has an adaptive step strategy for determining the partition $\{t_0,t_1,\dots,t_N\}$ (cf. section 1.1). In the sections 8.2, 8.3 and 8.4 we present some variants of this algorithm, in some of which Davidenko's method is used also (cf. section 1.1). These variants also use the results of section 6.5. From the numerical results given in section 8.5, it appears that all these algorithms are very reliable, but the algorithms in which Davidenko's method is used require the least amount of work for the solution of a problem.

CHAPTER 2

PRELIMINARIES

In this chapter we introduce some well-known concepts such as the spectral radius of a linear operator (section 2.2). In section 2.3 we give a definition of an iterative method. In section 2.4 the concept of the radius of convergence of an iterative method is introduced. Let x^* be the solution of (1.1.1). Then this concept indicates how close to x^* an initial guess x_0 must be in order that the application of the iterative method to problem (1.1.1) may yield a sequence of approximations x_k that converges to x^*. Section 2.5 is concerned with the local convergence behaviour of iterative processes. We introduce two types of local convergence behaviour that lie between the well-known concepts of local and quadratic convergence. In section 2.6 we present the imbedding on which the iterative methods to be investigated in the following chapters will be based.

2.1. CONVENTIONS AND NOTATIONS

From now on the following conventions hold.

If F is an operator, then D(F) denotes its *domain*.
Let Z_1 and Z_2 be Banach spaces. Then

$$L[Z_1,Z_2] = \{C \mid C\colon Z_1 \to Z_2;\ C \text{ is linear and bounded}\}.$$

Furthermore we define the sets $L^{(n)}[Z_1,Z_2]$ by

$$L^{(1)}[Z_1,Z_2] = L[Z_1,Z_2] \quad \text{and} \quad L^{(n)}[Z_1,Z_2] = L[Z_1,L^{(n-1)}[Z_1,Z_2]]$$
$$(n = 2,3,\ldots).$$

Let ℓ and n be the positive integers with $\ell \le n$. Let $Q \in L^{(n)}[Z_1,Z_2]$ and $y_j \in Z_1$ $(j = 1,2,\ldots,\ell)$. We use the notation

$$Qy_1y_2\dots y_\ell = (\dots((Qy_1)y_2)\dots)y_\ell.$$

Let F: $D(F) \to Z_2$ with $D(F) \subset Z_1$ and let $x \in \text{interior}(D(F))$. Then $F'(x)$ (or $F^{(1)}(x)$) denotes the *(Fréchet-)derivative* of F at x ($F'(x) \in L[Z_1,Z_2]$). $F''(x)$ (or $(F^{(2)}(x))$ denotes the *second (Fréchet-)derivative* of F at x (it is the Fréchet-derivative of the operator F' at x, $F''(x) \in L^{(2)}[Z_1,Z_2]$). The *n-th (Fréchet-)derivative* of F at x ($n = 3,4,\dots$) is denoted by $F^{(n)}(x)$.

If $Z_1 = Z_2 = Z$ we set $L[Z_1,Z_2] = L[Z]$ and $L^{(n)}[Z_1,Z_2] = L^{(n)}[Z]$ ($n = 1,2,\dots$).

For $i = 1,2,\dots,n+1$ let Z_i be a given Banach space with norm $\|\cdot\|_i$. Let $Z_0 = Z_1 \times Z_2 \times \dots \times Z_n$ be the product space. We shall always assume that the norm in Z_0 is $\|\cdot\|_0$ where $\|x\|_0 = \max\{\|x_i\|_i \mid 1 \le i \le n\}$ (for all $x = (x_1,x_2,\dots,x_n) \in Z_0$). Let P: $D(P) \to Z_{n+1}$ with $D(P) \subset Z_0$. Let $x = (x_1,x_2,\dots,x_n) \in Z_0$. For $i = 1,2,\dots,n$ we define the operator $P(x_1,\dots,x_{i-1},\cdot,x_{i+1},\dots,x_n)$ by

$$P(x_1,\dots,x_{i-1},\cdot,x_{i+1},\dots,x_n) = Q$$

with

$$D(Q) = \{v \mid v \in Z_i;\ (x_1,\dots,x_{i-1},v,x_{i+1},\dots,x_n) \in D(P)\},$$

$$Q(v) = P(x_1,\dots,x_{i-1},v,x_{i+1},\dots,x_n) \qquad \text{(for all } v \in D(Q)).$$

For $x_i \in \text{interior}(D(P(x_1,\dots,x_{i-1},\cdot,x_{i+1},\dots,x_n)))$ we denote by $\partial_i P(x)$ the *partial (Fréchet-)derivative* of P with respect to x_i at x ($i = 1,2,\dots,n$). For $x_j \in \text{interior}(D(\partial_i P(x_1,\dots,x_{j-1},\cdot,x_{j+1},\dots,x_n)))$ we denote be $\partial_{ij}P(x)$ the derivative $\{\partial_j[\partial_i P]\}(x)$ ($i,j = 1,2,\dots,n$).

E denotes a real Hilbert space, with inner product $(\cdot,\cdot)$ and norm $\|\cdot\| = (\cdot,\cdot)^{\frac{1}{2}}$. We shall always assume that $E \neq \{0\}$. Let $x \in E$ and $\sigma \in (0,\infty]$. We set

$$B(x,\sigma) = \{y \mid y \in E;\ \|y-x\| < \sigma\}.$$

Furthermore if $V \subset E$ is a subset of E, then $\bar{V}$ denotes the *closure* of V.

For a detailed definition of the above concepts we refer to [KANTOROWITSCH & AKILOW, 1964].

If $C \in L[E]$, then C is said to be *invertible* if a $T \in L[E]$ exists such that $CT = TC = I$, the identity (we write $T = [C]^{-1}$).

Let $F\colon D(F) \to E$ with $D(F) \subset E$ and let $x \in \text{interior}(D(F))$. Suppose F is Fréchet-differentiable at x, and suppose $F'(x)$ is invertible. Then $\Gamma(x)$ denotes $[F'(x)]^{-1}$.

Let $X\colon [0,1] \to E$, then $\dot{X}(t)$ denotes $\frac{d}{dt} X(t)$ $(t \in [0,1])$.

Let $\phi\colon D(\phi) \to \mathbb{R}$ with $D(\phi) \subset \mathbb{R}$. ϕ is *isotone (antitone)* on $D(\phi)$ if $\phi(\xi_1) \leq \phi(\xi_2)$ $(\phi(\xi_1)) \geq \phi(\xi))$ whenever $\xi_1,\xi_2 \in D(\phi)$ and $\xi_1 \leq \xi_2$. ϕ is *strictly isotone (anitone)* on $D(\phi)$ if $\phi(\xi_1) < \phi(\xi_2)$ $(\phi(\xi_1) > \phi(\xi_2))$ whenever $\xi_1,\xi_2 \in D(\phi)$ and $\xi_1 < \xi_2$.

Let m and n be integers. We shall always use the conventions

$$\sum_{i=m}^{n} \ldots = 0 \quad \text{and} \quad \prod_{i=m}^{n} \ldots = 1 \qquad (\text{if } m > n).$$

$\mathbb{R}$ denotes the set of real numbers, $\mathbb{C}$ denotes the set of complex numbers. $\mathbb{R}^n$ and $\mathbb{C}^n$ denote the real and the complex n-dimensional vector space, respectively. Finally, $\mathbb{N} = \{1,2,3,\ldots\}$.

2.2. THE SPECTRUM OF A LINEAR OPERATOR

Let $E_{\mathbb{C}}$ denote the *complex extension* of E (see [KANTOROWITSCH & AKILOW, 1964; section 2(XIII)]). $E_{\mathbb{C}}$ is a complex Hilbert space. We denote its innerproduct by $(\cdot,\cdot)_{\mathbb{C}}$. The norm in $E_{\mathbb{C}}$ is $\|\cdot\|_{\mathbb{C}} = (\cdot,\cdot)_{\mathbb{C}}^{\frac{1}{2}}$. We note that E is a subspace of $E_{\mathbb{C}}$. Any element z of $E_{\mathbb{C}}$ can be written in the form $z = x + iy$ where $x,y \in E$ are uniquely determined. Let $z_1,z_2 \in E_{\mathbb{C}}$. Then

$$(z_1,z_2)_{\mathbb{C}} = (x_1,x_2) + (y_1,y_2) + i\{(y_1,x_2) - (x_1,y_2)\}$$

where $x_j,y_j \in E$ and $z_j = x_j + iy_j$ $(j = 1,2)$.

Let $T \in L(E_{\mathbb{C}})$.

<u>DEFINITION 2.2.1.</u> The *spectrum* of T is the set

$$(2.2.1) \qquad sp(T) = \{\alpha \mid \alpha \in \mathbb{C};\ [T-\alpha I] \text{ is not invertible}\}.$$

sp(T) is non-empty (cf. [RUDIN, 1973; Theorem 10.13(a)]).

<u>DEFINITION 2.2.2.</u> The number

(2.2.2) $$\mathrm{sr}(T) = \sup\{|\alpha| \mid \alpha \in \mathrm{sp}(T)\}$$

is called the *spectral radius* of T.

We have

THEOREM 2.2.1.

$$\mathrm{sr}(T) = \lim_{n\to\infty} \|T^n\|^{\frac{1}{n}} = \inf_{n\geq 1} \|T^n\|^{\frac{1}{n}}.$$

PROOF. cf. [RUDIN, 1973; Theorem 10.13(b)]. □

Let $C \in L(E)$, and let

(2.2.3) $$C_{\mathbb{C}}: E_{\mathbb{C}} \to E_{\mathbb{C}}, \qquad C_{\mathbb{C}}z = Cx + iCy \qquad (z \in E_{\mathbb{C}},\ z = x+iy,\ x,y \in E).$$

Then $C_{\mathbb{C}} \in L(E_{\mathbb{C}})$ and it follows that

(2.2.4) $$\|C_{\mathbb{C}}\|_{\mathbb{C}} = \|C\|.$$

$C_{\mathbb{C}}$ is called the *extension of* C *in* $E_{\mathbb{C}}$. The *spectrum of* C is defined by $\mathrm{sp}(C) = \mathrm{sp}(C_{\mathbb{C}})$. Analogously, the *spectral radius of* C is defined by $\mathrm{sr}(C) = \mathrm{sr}(C_{\mathbb{C}})$. From (2.2.4) and Theorem 2.2.1 it follows that

(2.2.5) $$\mathrm{sr}(C) = \lim_{n\to\infty} \|C^n\|^{\frac{1}{n}} = \inf_{n\geq 1} \|C^n\|^{\frac{1}{n}}.$$

Let ρ be a rational function with real coefficients, i.e. $\rho(z) \equiv [q(z)]^{-1}p(z)$ where p and q are polynomials with real coefficients. We define

(2.2.6) $$D_{\mathbb{C}}(\rho) = \{z \mid z \in \mathbb{C};\ q(z) \neq 0\}$$

and

(2.2.7) $$D_E(\rho) = \{C \mid C \in L(E);\ q(C) \text{ is invertible}\}.$$

Let $C \in D_E(\rho)$. Then $\rho(C)$ denotes the linear operator $[q(C)]^{-1}p(C)$.

THEOREM 2.2.2. *Let ρ: $D(\rho) \to \mathbb{C}$ be a rational function with real coefficients. Let $C \in L(E)$. If $sp(C) \subset D_{\mathbb{C}}(\rho)$ then*

$$C \in D_E(\rho) \quad \text{and} \quad sp(\rho(C)) = \rho(sp(C)).$$

PROOF. Let $\rho(z) \equiv [q(z)]^{-1}p(z)$.

1. It follows from [RUDIN, 1973; Theorem 10.28(a)] that $q(C_{\mathbb{C}})$ is invertible. Since $q(C_{\mathbb{C}}) = [q(C)]_{\mathbb{C}}$ we have $C \in D_E(\rho)$.
2. Since $q(C_{\mathbb{C}}) = [q(C)]_{\mathbb{C}}$, $p(C_{\mathbb{C}}) = [p(C)]_{\mathbb{C}}$ we have $\rho(C_{\mathbb{C}}) = [\rho(C)]_{\mathbb{C}}$. From [RUDIN, 1973; Theorem 10.28(b)] follows that

$$sp(\rho(C_{\mathbb{C}})) = \rho(sp(C_{\mathbb{C}})).$$

Consequently $sp(\rho(C)) = \rho(sp(C_{\mathbb{C}})) = \rho(sp(C))$, which completes the proof of the theorem. □

2.3. ITERATIVE METHODS

In this section we introduce some concepts that will play an important role in our investigations. Throughout,

$$(2.3.1) \qquad G = \{G \mid G: D(G) \to E;\ D(G) \subset E\}$$

and

$$(2.3.2) \qquad F = \{F \mid F: D(F) \to E;\ D(F) \subset E;\ F(x) = 0 \text{ has a unique solution}\}.$$

Let $F \in F$ be given. Then x^* will always denote the solution of $F(x) = 0$.

Let $\hat{F} \subset F$.

DEFINITION 2.3.1. Any operator with domain $\hat{F}$ and range in G is called an *iterative method* for $\hat{F}$.

We note that this concept of an iterative method is much wider than what is intuitively understood by the term. However, for ease of presentation, we shall work with the above concept of an iterative method.

Let M be an iterative method for $\hat{F}$ and let $F \in \hat{F}$. Set

(2.3.3) $D(M,F) = \{x_0 \mid \text{there exists a sequence } \{x_k\} \text{ such that } x_k \in D(G), \text{ where } G = M(F), \text{ and } x_{k+1} = G(x_k)\ (k = 0,1,2,\ldots)\}.$

DEFINITION 2.3.2. The *iterative process*, corresponding to the iterative method M and the function $F \in \hat{F}$ will be denoted by [M,F], and consists in computing vectors x_k from

(2.3.4a) $x_{k+1} = G(x_k) \qquad (k = 0,1,2,\ldots)$

where

(2.3.4b) $G = M(F).$

In this connection, the operator G in (2.3.4b) is called an *iteration function*. The *starting point* x_0 of (2.3.4a) should be an element of D(M,F) in order to prevent the iterative process from breaking off prematurely.

DEFINITION 2.3.3. The set D(M,F) defined in (2.3.3) is called the *domain* of the iterative process [M,F].

Let $x_0 \in D(M,F)$. Then the *sequence* $\{x_k\}$ *generated by* x_0 and the iterative process [M,F] is, of course, defined by (2.3.4).

DEFINITION 2.3.4. The set

(2.3.5) $S(M,F) = \{x_0 \mid x_0 \in D(M,F) \text{ and the sequence } \{x_k\} \text{ generated by } x_0 \text{ and } [M,F] \text{ converges to } x^*\}$

is called the *region of convergence* of the iterative process [M,F].

2.4. THE RADIUS OF CONVERGENCE

Let $F \in F$. As pointed out in the introduction, we are interested in iterative methods M such that the related iterative processes [M,F] generate sequences $\{x_k\}$ that converge to x^*, even if the starting point x_0 is remote from x^*.

In this section we introduce the concept of the radius of convergence of an iterative method M. This concept indicates how close to x^* a starting point x_0 should be for the iterative process [M,F] to generate a sequence

$\{x_k\}$ that converges to x^*.

Let $\hat{F}$ be a non-empty subset of F.

DEFINITION 2.4.1. Let M be an iterative method (for $\hat{F}$) and $F \in \hat{F}$. We define

$$r(M,F) = \begin{cases} 0 & (\text{if } x^* \notin \text{interior}(S(M,F))), \\ \sup\{\sigma \mid \sigma > 0 \text{ and } B(x\ ,\sigma) \subset S(M,F)\} & (\text{otherwise}). \end{cases}$$

r(M,F) is called *radius of convergence of the iterative process* [M,F].

DEFINITION 2.4.2. Let M be an iterative method (for $\hat{F}$) and let F_0 be a non-empty subset of $\hat{F}$. Then

$$r(M;F_0) = \inf\{r(M,F) \mid F \in F_0\}$$

is called the *radius of convergence of the iterative method* M *with respect to* F_0.

It is clear that the larger $r(M;F_0)$ is for an iterative method M, the less sensitive with regard to starting points will be the iterative processes [M,F] ($F \in F_0$) generated by it.

2.5. LOCAL CONVERGENCE BEHAVIOUR OF ITERATIVE PROCESSES

Let $\hat{F} \subset F$ and let M be an iterative method for $\hat{F}$. Let $F \in \hat{F}$.

DEFINITION 2.5.1. The iterative process [M,F] is called *locally convergent* (LC) is a neighbourhood V of x^* exists such that $V \subset S(M,F)$.

Obviously the iterative process will have a positive radius of convergence if and only if it is locally convergent. We give three other definitions concerning the convergence behaviour of [M,F] near x^*.

DEFINITION 2.5.2. The iterative process [M,F] is called *stably convergent* (SC) if a constant $\theta > 0$, an $N_0 \in \mathbb{N}$ and a neighbourhood V of x^* exist such that $V \subset S(M,F)$, and for all $x_0 \in V$, the sequence $\{x_k\}$ generated by x_0 and [M,F] satisfies

$$\|x_k - x^*\| \le \theta \|x_0 - x^*\| \qquad (\text{for all } k \ge N_0).$$

DEFINITION 2.5.3. The iterative process $[M,F]$ is called *regularly convergent* (RC) if a neighbourhood V of x^* exists such that $V \subset S(M,F)$ and for all $x_0 \in V$, the sequence $\{x_k\}$ generated by x_0 and $[M,F]$ satisfies

$$\|x_k-x^*\| \le \|x_0-x^*\| \qquad (k = 1,2,\dots).$$

DEFINITION 2.5.4. The iterative process $[M,F]$ is called *quadratically convergent* (QC) if a neighbourhood V of x^* and a number $\delta > 0$ exist such that $V \subset S(M,F)$ and for all $x_0 \in V$, the sequence $\{x_k\}$ generated by x_0 and $[M,F]$ satisfies

$$\|x_{k+1}-x^*\| \le \delta\|x_k-x^*\|^2 \qquad (k = 0,1,2,\dots).$$

It is clear that the following relations hold.

$$QC \Rightarrow RC \Rightarrow SC \Rightarrow LC.$$

If $[M,F]$ is locally convergent and $G = M(F)$ is continuous in x^* then obviously $G(x^*) = x^*$. This means that x^* is a *fixed point* of G.

Conversely, the following theorem holds.

THEOREM 2.5.1. *Suppose* x^* *is a fixed point of* $G = M(F)$. *Let* $x^* \in \text{interior}(D(G))$ *and let* G *be differentiable at* x^*. *The following four statements* (i) - (iv) *hold.*

(i) *If* $\text{sr}(G'(x^*)) < 1$ *then* $[M,F]$ *is stably convergent.*

(ii) *If* $\text{sr}(G'(x^*)) > 1$ *then* $[M,F]$ *is not stably convergent.*

(iii) *If* $\|G'(x^*)\| < 1$ *then* $[M,F]$ *is regularly convergent.*

(iv) *If* $\|G'(x^*)\| > 1$ *then* $[M,F]$ *is not regularly convergent.*

In order to prove Theorem 2.5.1 we need two lemmata.

LEMMA 2.5.2. *Let* $P\colon D(P) \to E$ *with* $D(P) \subset E$. *Suppose*

(a) *The operator* P *has a fixed point* y^* *with* $y^* \in \text{interior}(D(P))$ *and* P *is* differentiable at y^*.

(b) $\|P'(y^*)\| \le \alpha$ *and* $\|[P'(y^*)]^{n_0}\|^{1/n_0} \le \delta < 1$ *for some* $n_0 \ge 1$ *and real numbers* α *and* δ.

Then for any $\varepsilon > 0$ *with* $\delta^{n_0} + \varepsilon < 1$ *there exists a number* $\sigma > 0$ *such that for all* $x \in B(y^*,\sigma)$ *and all* $n \ge 1$ *we have*

$$x \in D(P^n) \quad \textit{and} \quad \|P^n(x) - y^*\| \le c\lambda^{\ell}\|x-y^*\|.$$

Here $c = \max\{\alpha^j + \varepsilon \mid 1 \le j \le n_0\}$, $\lambda = \delta^{n_0} + \varepsilon$ *and* $\ell = \text{entier}(\frac{n-1}{n_0})$.

PROOF. Let $\varepsilon > 0$. Set $\lambda = \delta^{n_0} + \varepsilon$, suppose $\lambda < 1$. Let $j \in \mathbb{N}$ and $j \le n_0$. The operator P^j is defined in a neighbourhood of y^*, $P^j(y^*) = y^*$ and P^j is continuous in y^*. From the *chainrule* (cf. [KANTOROWITSCH & AKILOW, 1964; section 1.2(XVII)] it follows that P^j is differentiable at y^* and

$$[P^j]'(y^*) = [P'(y^*)]^j. \tag{2.5.1}$$

Consequently, a number $\sigma_1 > 0$ exists such that for all $x \in B(y^*,\sigma_1)$ and all $j \le n_0$ we have $x \in D(P^j)$ and

$$\|P^j(x) - P^j(y^*) - [P^j]'(y^*)(x-y^*)\| \le \varepsilon\|x-y^*\|.$$

Hence for all $x \in B(y^*,\sigma_1)$ we have

$$\|P^j(x) - y^*\| \le (\alpha^j + \varepsilon)\|x-y^*\| \qquad \text{(for all } j \le n_0)$$

and

$$\|P^{n_0}(x) - y^*\| \le \lambda\|x-y^*\|.$$

Let $c = \max\{\alpha^j + \varepsilon \mid 1 \le j \le n_0\}$. Set $\sigma = \min\{\sigma_1, \frac{\sigma_1}{c}\}$. Let $n \ge 1$. Then $n = \ell n_0 + k$ where $\ell \ge 0$ and $1 \le k \le n_0$. Then for all $x \in B(y^*,\sigma)$ we have $x \in D(P^n)$ and

$$\|P^n(x) - y^*\| = \|[P^{n_0}]^{\ell}(P^k(x)) - y^*\| \le \lambda^{\ell}\|P^k(x) - y^*\| \le c\lambda^{\ell}\|x-y^*\|.$$

This proves the lemma. □

LEMMA 2.5.3. *Let* $P\colon D(P) \to E$ *with* $D(P) \subset E$. *Let the assumption* (a) *of lemma* 2.5.2 *be fulfilled. Suppose* $\|[P'(y^*)]^{n_0}\|^{1/n_0} \ge \delta > 1$ *for some* $n_0 \ge 1$. *Then for all positive numbers* ε *and* σ *there exists an* $\tilde{x} \in B(y^*,\sigma)$ *with* $\tilde{x} \ne y^*$ *such that* $\tilde{x} \in D(P^{n_0})$ *and*

$$\|P^{n_0}(\tilde{x}) - y^*\| \ge (\delta^{n_0} - \varepsilon)\|\tilde{x}-y^*\|.$$

PROOF. Let ε and δ be positive numbers. From the chainrule it follows that $[P^{n_0}]'(y^*)$ exists. Consequently a number $\tilde{\sigma} \in (0,\sigma]$ exists such that for all $x \in B(y^*,\tilde{\sigma})$ we have $x \in D(P^{n_0})$ and

$$(2.5.2) \qquad \|P^{n_0}(x) - P^{n_0}(y^*) - [P^{n_0}]'(y^*)(x-y^*)\| \leq \frac{\varepsilon}{2}\|x-y^*\|.$$

Moreover an element $y \in E$ with $\|y\| = 1$ exists such that

$$(2.5.3) \qquad \|[P'(y^*)]^{n_0}y\| \geq (\delta^{n_0} - \frac{\varepsilon}{2})\|y\|.$$

Let $\tilde{x} = y^* + \frac{\tilde{\sigma}}{2}y$. Then (2.5.1), (2.5.2) and (2.5.3) yield

$$\|P^{n_0}(\tilde{x}) - y^*\| \geq (\delta^{n_0} - \frac{\varepsilon}{2})\|\tilde{x}-y^*\| - \frac{\varepsilon}{2}\|\tilde{x}-y^*\|.$$

This proves the lemma. □

We now turn to the proof of Theorem 2.5.1.

(i) If $sr(G'(x^*)) < 1$ then from (2.2.5) it follows that an $n_0 \geq 1$ exists such that

$$\|[G'(x^*)]^{n_0}\|^{\frac{1}{n_0}} \leq \delta < 1.$$

Let $\varepsilon > 0$ with $\delta^{n_0} + \varepsilon < 1$. According to Lemma 2.5.2 a positive number σ exists such that for all $x \in B(x^*,\sigma)$ and $n \geq 1$ we have $x \in D(G^n)$ and

$$\|G^n(x) - x^*\| \leq c\lambda^{\ell}\|x-x^*\|.$$

Here $c = \max\{\alpha^j + \varepsilon \mid 1 \leq j \leq n_0\}$, $\alpha = \|G'(x^*)\|$, $\lambda = \delta^{n_0} + \varepsilon$ and $\ell = \text{entier}(\frac{n-1}{n_0})$. Hence [M,F] is stably convergent.

(ii) Suppose $sr(G'(x^*)) > 1$ and [M,F] is stably convergent. Thus positive numbers $\tilde{\sigma}$ and θ and an integer $N_0 \geq 1$ exist such that for all $x \in B(x^*,\tilde{\sigma})$ and all $n \geq N_0$ we have $x \in D(G^n)$ and

$$\|G^n(x) - x^*\| \leq \theta\|x-x^*\|.$$

Furthermore from (2.2.5) it follows that an $\tilde{N} \geq N_0$ and a $\delta > 1$ exist such that

$$\|[G'(x^*)]^n\|^{\frac{1}{n}} \geq \delta \qquad (\text{for all } n \geq \tilde{N}).$$

Hence Lemma 2.5.3 applies where $\sigma \in (0,\tilde{\sigma}]$, $\varepsilon > 0$ and $n_0 \geq N$ are such that $\delta^{n_0} - \varepsilon > \theta$. This yields a contradiction.

(iii) If $\|G'(x^*)\| < 1$, Lemma 2.5.2 applies with $n_0 = 1$ and $\alpha = \delta < 1$. This proves the result.

(iv) Suppose $\|G'(x^*)\| = \delta > 1$. If [M,F] is regularly convergent then a positive number σ exists such that $B(x^*,\sigma) \subset D(G)$ and

$$\|G(x) - x^*\| \leq \|x-x^*\| \qquad (\text{for all } x \in B(x^*,\sigma)).$$

However Lemma 2.5.3 applies with $n_0 = 1$ and $\varepsilon > 0$ such that $\delta - \varepsilon > 1$. This yields a contradiction. Hence [M,F] is not regularly convergent. □

REMARK 2.5.1. We note that Theorem 2.5.1 remains valid if E is a Banach space.

The sufficiency of the condition $sr(G'(x^*)) < 1$ for [M,F] to be locally convergent where $E = \mathbb{R}$, was proved by Schröder (cf. [SCHRÖDER, 1870]). Ostrowski proved it if $E = \mathbb{R}^n$ (cf. [OSTROWSKI, 1960; Theorem 22.1]). In [KITCHEN, 1966] it is proved if E is an arbitrary Banach space. □

We end this section with a simple result with regard to quadratically convergent iterative processes.

THEOREM 2.5.4. *Suppose* [M,F] *is quadratically convergent. Then with* $G = M(F)$ *it follows that* $x^* \in \text{interior}(D(G))$, $G'(x^*)$ *exists and* $G'(x^*) = 0$.

PROOF. If [M,F] is quadratically convergent then positive constants σ and δ exist such that $B(x^*,\sigma) \subset D(G)$ and

$$\|G(x) - x^*\| \leq \delta\|x-x^*\|^2 \qquad (\text{for all } x \in B(x^*,\sigma)).$$

Hence $G'(x^*)$ exists and $G'(x^*) = 0$. □

2.6. THE IMBEDDING AND THE DIFFERENTIAL EQUATION

In this section we present the imbedding and the differential equation which we use to construct iterative methods by means of iterative imbedding (cf. section 1.1).

We first define some classes of operators that we shall need subsequently. Let

(2.6.1) $\quad F_1 = \{F \mid F \in F;\ D(F)$ is open; F is twice continuously differentiable on $D(F)$; there exists a constant μ such that $\|F''(x)\| \le \mu$ (for all $x \in D(F)$); $F'(x^*)$ is invertible$\}$

(see also (2.3.2)). We shall restrict our attention to problems $F(x) = 0$ where $F \in F_1$. In order to describe the type of imbedding to be used, we define the following two classes of operators. Let

(2.6.2) $\quad K = \{K \mid K\colon D(K) \to E,\ D(K) = W(K) \times W(K)$ where $W(K)$ is an open subset of E; $K(x,x) = 0$ (for all $x \in W(K)$); $\partial_{11}K(y,x)$, $\partial_{12}K(y,x)$ and $\partial_{21}K(y,x)$ exist (for all $x,y \in W(K)$); $\partial_{11}K$ is continuous on $D(K)$; there exist constants μ_1, μ_2 and μ_3 such that $\|\partial_{11}K(y,x)\| \le \mu_1$, $\|\partial_{12}K(y,x)\| \le \mu_2$ and $\|\partial_{21}K(y,x)\| \le \mu_3$ (for all $x,y \in W(K)$)$\}$.

Let

(2.6.3) $\quad A = \{A \mid A\colon F_1 \to K;$ with $K = A(F)$ it holds $W(K) = D(F)$ (for all $F \in F_1$)$\}$.

Examples of elements of A. Let $F \in F_1$.

$$[A(F)](y,x) \equiv F(y) - F(x).$$
$$[A(F)](y,x) \equiv F'(y)(y-x).$$
$$[A(F)](y,x) \equiv y-x.$$

We give three lemmata. The first lemma is a general result that will often be used subsequently.

LEMMA 2.6.1. *Let* $P\colon D(P) \to Z_2$ *with* $D(P) \subset Z_1$ *where* Z_1 *and* Z_2 *are Banach spaces. Assume that a positive number* δ *and* $V \subset D(P)$ *with* V *open and convex exist such that for all* $x \in V$ *the derivative* $P'(x)$ *exists and* $\|P'(x)\| \le \delta$. *Then*

$$\|P(x) - P(y)\| \le \delta\|x-y\| \qquad \text{(for all } x,y \in V).$$

PROOF. See [KANTOROWITSCH & AKILOW, 1964; section 1.3(XVII)]. □

LEMMA 2.6.2. *Let* $Q\colon D(Q) \to Z$ *with* $D(Q) = V \times V$ *where* V *is an open subset of* E *and* Z *is a Banach space. Suppose* $\partial_1 Q(y,x)$ *and* $\partial_2 Q(y,x)$ *exist (for all* $x,y \in V$*) and suppose that* $\partial_1 Q$ *is continuous on* $D(Q)$*. Then* Q *is continuous on* $D(Q)$.

PROOF. Let $x_0, y_0 \in V$. Let $\varepsilon > 0$. Then a number $\delta > 0$ exists such that $x,y \in V$, $\|\partial_1 Q(y,x) - \partial_1 Q(y_0,x_0)\| \le \varepsilon$ and

$$\|Q(y_0,x) - Q(y_0,x_0) - \partial_2 Q(y_0,x_0)(x-x_0)\| \le \varepsilon\|x-x_0\|$$

whenever $x \in B(x_0,\delta)$ and $y \in B(y_0,\delta)$. Let $x \in B(x_0,\delta)$ and $y \in B(y_0,\delta)$. Let

$$P(z) = Q(z,x) - \partial_1 Q(y_0,x_0)(z-y_0) \qquad (z \in D(P))$$

with $D(P) = B(y_0,\delta)$. From Lemma 2.6.1 it follows that

$$\|Q(y,x) - Q(y_0,x) - \partial_1 Q(y_0,x_0)(y-y_0)\| \le \varepsilon\|y-y_0\|.$$

Therefore

$$\|Q(y,x) - Q(y_0,x_0)\| \le \{\|\partial_1 Q(y_0,x_0)\| + \varepsilon\}\|y-y_0\|$$
$$+ \{\|\partial_2 Q(y_0,x_0)\| + \varepsilon\}\|x-x_0\|.$$

This proves the lemma. □

The next lemma is a consequence of Lemma 2.6.2.

LEMMA 2.6.3. *If* $K \in \mathcal{K}$ *then both the operators* $\partial_1 K$ *and* K *are continuous on* $D(K)$.

PROOF. From (2.6.2) it follows that Lemma 2.6.2 applies for $Q = \partial_1 K$. Hence $\partial_1 K$ is continuous on $D(K)$. Consequently, Lemma 2.6.2 also applies for $Q = K$. This completes the proof. □

Let $F \in \mathcal{F}_1$ and $A \in \mathcal{A}$. Set $K = A(F)$. Let $x_0 \in D(F)$. Consider the imbedding

$$(2.6.4)\qquad \begin{aligned}&H\colon [0,1] \times D(F) \to E,\\ &H(t,x) = (1-t)K(x,x_0) + tF(x) \qquad (t \in [0,1],\ x \in D(F)).\end{aligned}$$

Thus $H(0,x_0) = 0$ and $H(1,x) \equiv F(x)$.

Suppose $H(t,x) = 0$ has a unique solution $x = X(t)$ for all $t \in [0,1]$. Then

$$(2.6.5)\qquad H(t,X(t)) = 0 \qquad (\text{for all } t \in [0,1]).$$

We have

$$(2.6.6)\qquad X(0) = x_0 \quad \text{and} \quad X(1) = x^*.$$

We also note that for all $t \in [0,1]$ and $x \in D(F)$ the derivatives $\partial_1 H(t,x)$ and $\partial_2 H(t,x)$ exist and are continuous (in t and x) and satisfy

$$\partial_1 H(t,x) = -K(x,x_0) + F(x)$$

and

$$\partial_2 H(t,x) = (1-t)\partial_1 K(x,x_0) + tF'(x).$$

The derivatives $\partial_1 H(0,x)$ and $\partial_1 H(1,x)$ should be considered respectively as the right and left partial derivatives of H with respect to t.

Let

$$(2.6.7)\qquad D(A,F) = \{(t,y,z) \mid t \in [0,1];\ y,z \in D(F);\ \text{with } K = A(F), \text{ the operator } [(1-t)\partial_1 K(y,z) + tF'(y)] \text{ is invertible}\}.$$

Suppose $(t,X(t),x_0) \in D(A,F)$ for all $t \in [0,1]$. Then $\dot{X}(t)$ exists (see Lemma 2.6.3 and [KANTOROWITSCH & AKILOW, 1964; Theorem 3(XVII)]). Differentiation of (2.6.5) with respect to t yields

$$(2.6.8)\qquad \partial_1 H(t,X(t)) + \partial_2 H(t,X(t))\dot{X}(t) = 0 \qquad (\text{for all } t \in [0,1]).$$

Before we give the differential equation from which we shall derive the iterative methods, we introduce another class of functions, with which we can significantly enlarge the number of iterative methods to be constructed in chapter 4 (see also section 6.3). We set

(2.6.9) $$\mathcal{S} = \{h \mid h\colon [0,1] \to \mathbb{R};\ h \text{ is continuous on } [0,1)\}.$$

Let $g \in \mathcal{S}$. It is obvious (see (2.6.5) and (2.6.8)) that X satisfies

(2.6.10) $$\partial_1 H(t,X(t)) + \partial_2 H(t,X(t))\dot{X}(t) + g(t)H(t,X(t)) = 0 \qquad \text{(for all } t \in [0,1]).$$

Since by assumption $(t,X(t),x_0) \in D(A,F)$, it follows that

(2.6.11) $$\begin{aligned} &\dot{X}(t) = -[\partial_2 H(t,X(t))]^{-1}\{\partial_1 H(t,X(t)) + g(t)H(t,X(t))\} \qquad \text{(for all } t \in [0,1]),\\ &X(0) = x_0. \end{aligned}$$

The relation (2.6.11) is equivalent to

(2.6.12a) $$\begin{aligned} &\dot{X}(t) = \Phi(t,X(t),x_0) \qquad (t \in [0,1]),\\ &X(0) = x_0 \end{aligned}$$

where

(2.6.12b) $$\begin{aligned} &\Phi\colon D(\Phi) \to E,\ D(\Phi) = D(A,F) \quad \text{(cf. (2.6.7))},\\ &\Phi(t,y,z) = -[(1-t)\partial_1 K(y,z) + tF'(y)]^{-1} \times\\ &\qquad \{-K(y,z) + F(y) + g(t)[(1-t)K(y,z) + tF(y)]\} \qquad ((t,y,z) \in D(\Phi)). \end{aligned}$$

Conversely, if Y is a solution of

(2.6.13) $$\begin{aligned} &\dot{Y}(t) = \Phi(t,Y(t),x_0) \qquad (t \in [0,1]),\\ &Y(0) = x_0, \end{aligned}$$

then

$$\frac{d}{dt} H(t,Y(t)) + g(t)H(t,Y(t)) = 0 \qquad \text{(for all } t \in [0,1]\text{)},$$

$$H(0,Y(0)) = 0.$$

So that (cf. [BERGER, 1977; Theorem 3.1.23])

$$H(t,Y(t)) = 0 \qquad \text{(for all } t \in [0,1)\text{)}.$$

Since $K \in \mathcal{K}$ and $F \in \mathcal{F}_1$, and since Y is left-continuous in t = 1, it follows that

$$H(1,Y(1)) = \lim_{t\uparrow 1} H(t,Y(t)) = 0.$$

We summarize the above results in a theorem.

THEOREM 2.6.4. *Assume* $F \in \mathcal{F}_1$, $A \in \mathcal{A}$ *and* $g \in \mathcal{S}$. *Set* $K = A(F)$. *Let* $x_0 \in D(F)$. *Let* H *be defined in* (2.6.4).

(i) *If* $H(t,x) = 0$ *has a unique solution* $x = X(t)$ *and* $[(1-t)\partial_1 K(X(t),x_0) + tF'(X(t))]$ *is invertible for all* $t \in [0,1]$, *then* (2.6.12) *holds.*

(ii) *If* X *is a solution of the initial value problem* (2.6.12a), *then* (2.6.5) *holds.*

The initial value problem (2.6.12a) will be used in chapter 4 to construct iterative methods by means of iterative imbedding.

CHAPTER 3

RUNGE-KUTTA METHODS

In this chapter we present the Runge-Kutta methods that will be used for the construction of iterative methods by means of iterative imbedding. We shall deal with two types of Runge-Kutta methods, which will be presented in the subsections 3.2.1 and 3.2.2, respectively.

In section 3.3 we shall indicate that the problem of nonconvergence of Newton's method is closely related to a certain type of unstable behaviour of Euler's method.

3.1. ONE-STEP METHODS

Consider the initial value problem

$$(3.1.1)\qquad \begin{aligned} &\dot{Y}(t) = f(t,Y(t)) \qquad (t \in [0,1]),\\ &Y(0) = y_0, \end{aligned}$$

where $f\colon [0,1] \times V \to E$ with $V \subset E$ and $y_0 \in V$ are given. We assume that (3.1.1) has a unique solution Y. Most of the computational methods for solving (3.1.1) approximate the true solution Y of (3.1.1) on a discrete point set $\{t_0,t_1,\ldots,t_{N+1}\}$ where $0 = t_0 < t_1 < \ldots < t_{N+1} = 1$.

Runge-Kutta methods, which we shall use in the construction of iterative methods, are *one-step methods*. This means that, starting from y_0 and t_0, approximations y_n of $Y(t_n)$ $(n = 1,2,\ldots,N+1)$ are obtained by

$$(3.1.2a)\qquad y_{i+1} = y_i + h_i\Psi(t_i,y_i;h_i,f),$$

where

$$(3.1.2b)\qquad h_i = t_{i+1} - t_i \qquad (i = 0,1,\ldots,N).$$

The function ψ is characteristic for the method. We shall therefore define a Runge-Kutta method in terms of Ψ. By H we shall denote the sequence of stepsizes

$$(3.1.3) \qquad H = \{h_0, h_1, \ldots, h_N\}.$$

Together with the one-step method, H determines the numerical integration procedure. Let a sequence of stepsizes $H = \{h_0, h_1, \ldots, h_N\}$ be given. Then the numbers $t_1, t_2, \ldots, t_{N+1}$ are supposed to satisfy (3.1.2b) where $t_0 = 0$.

In the next section we describe two types of Runge-Kutta methods for solving numerically problem (3.1.1). We shall confine our considerations to so-called explicit Runge-Kutta methods and generalized Runge-Kutta methods (see e.g. [VAN DER HOUWEN, 1977; sections 2.2 - 3]).

3.2. DESCRIPTION OF THE RUNGE-KUTTA METHODS

3.2.1. Runge-Kutta methods with scalar coefficients

DEFINITION 3.2.1. Let $L = (\lambda_{i,j})$ be a real strictly lower triangular $(m+1) \times (m+1)$ matrix ($m \in \mathbb{N}$). Then the general m-*stage Runge-Kutta method with scalar coefficients* is defined by

$$(3.2.1a) \qquad \Psi = \sum_{\ell=1}^{m} \lambda_{m+1,\ell} k_\ell$$

where

$$(3.2.1b) \qquad \begin{aligned} &k_1(t,y;h,f) \equiv f(t,y), \\ &k_\ell(t,y;h,f) \equiv f(t + \nu_\ell h, y + h \sum_{j=1}^{\ell-1} \lambda_{\ell,j} k_j(t,y;h,f)) \quad (\ell = 2,3,\ldots,m) \end{aligned}$$

and

$$(3.2.1c) \qquad \nu_\ell = \sum_{j=1}^{\ell-1} \lambda_{\ell,j} \qquad (\ell = 1,2,\ldots,m+1).$$

The matrix L is called the *generating matrix* of the Runge-Kutta method, which, obviously, completely determines the method.

For the sake of shortness we shall use the phrase "Runge-Kutta method with scalar coefficients L "to mean" Runge-Kutta method with scalar coefficients with generating matrix L". Moreover, given a Runge-Kutta method with scalar coefficients $L = (\lambda_{i,j})$, we shall always assume that ν_ℓ ($\ell = 1,2,\ldots,m+1$) satisfies (3.2.1c). In addition, we shall always assume that

(3.2.2a) $\nu_{m+1} = 1$

and

(3.2.2b) $\nu_\ell \in [0,1] \qquad (\ell = 2,3,\ldots,m).$

The Runge-Kutta methods with scalar coefficients defined here are of the so-called explicit type. For a more detailed description of Runge-Kutta methods see [VAN DER HOUWEN, 1977], [LAMBERT, 1973] and [STETTER, 1973].

3.2.2. Runge-Kutta methods with operator coefficients

We assume that $\partial_2 f(t,y)$ exists (for all $t \in [0,1]$ and $y \in \text{interior}(V)$). Let $J(t,y) \equiv \partial_2 f(t,y)$.

DEFINITION 3.2.2. Let

(3.2.3a)
$$R\colon \mathbb{C} \to L(\mathbb{C}^{m+1}),$$
$$R(z) = (\rho_{i,j}(z)) \qquad (z \in \mathbb{C}).$$

Here $R(z)$ is a strictly lower triangular $(m+1) \times (m+1)$-matrix and $\rho_{i,j}$ is a rational function with real coefficients for which

(3.2.3b) $0 \in D_{\mathbb{C}}(\rho_{i,j}) \qquad (i = 2,3,\ldots,m+1 \text{ and } j = 1,2,\ldots,i-1).$

Then the general m-*stage Runge-Kutta method with operator coefficients* is defined by

(3.2.4a) $\Psi = \sum_{\ell=1}^{m} \Lambda_{m+1,\ell} k_\ell$

where

(3.2.4b)
$$k_1(t,y;h,f) \equiv f(t,y),$$
$$k_\ell(t,y;h,f) \equiv f(t+\nu_\ell h, y+h\sum_{j=1}^{\ell-1}\Lambda_{\ell,j}k_j(t,y;h,f)) \quad (\ell = 2,3,\ldots,m),$$

(3.2.4c) $$\Lambda_{i,j} = \rho_{i,j}(hJ(t,y)) \qquad (i = 2,3,\ldots,m+1 \text{ and } j = 1,2,\ldots,i-1)$$

and

(3.2.4d) $$\nu_\ell = \sum_{j=1}^{\ell-1}\rho_{\ell,j}(0) \qquad (\ell = 1,2,\ldots,m+1).$$

The operator R completely determines the method. For the sake of shortness we shall use the phrase "Runge-Kutta method with operator coefficients R" to mean "Runge-Kutta method with operator coefficients defined in (3.2.4) where R is of the type (3.2.3)". Analogous to the case of scalar coefficients, for a given Runge-Kutta method with operator coefficients R, ν_ℓ ($\ell = 1,2,\ldots,m+1$) is always supposed to satisfy (3.2.4d). Furthermore, we shall always assume that

(3.2.5a) $$\nu_{m+1} = 1$$

and

(3.2.5b) $$\nu_\ell \in [0,1] \qquad (\ell = 2,3,\ldots,m).$$

The Runge-Kutta methods with operator coefficients that are defined here, require one evaluation of the operator J per step. A more detailed description of Runge-Kutta methods with operator coefficients, including methods that require several evaluations of the operator J per step, can be found in [VAN DER HOUWEN, 1977; section 2.3].

3.3. NONCONVERGENCE OF NEWTON'S METHOD

3.3.1. Strong stability of discretization methods

Let $F \in \mathcal{F}_1$ and $x_0 \in D(F)$ (cf. (2.6.1)). Let $K \in \mathcal{K}$ where $K(y,x) \equiv F(y) - F(x)$ and $g \in \mathcal{S}$ where $g = 0$ (cf. (2.6.2), (2.6.9)).

Suppose the initial value problem (2.6.12a) has a unique solution X. Hence

$$(3.3.1)\qquad \begin{aligned} \dot{X}(t) &= -\Gamma(X(t))F(x_0) \qquad (t \in [0,1]), \\ X(0) &= x_0. \end{aligned}$$

In (3.3.1) $\Gamma(X(t))$ denotes $[F'(X(t))]^{-1}$. It follows that $X(1) = x^*$ (cf. Theorem 2.6.4(ii)).

Suppose we solve (3.3.1) numerically, using a Runge-Kutta method of the type described in section 3.2. Let x_1 denote the approximation to $X(1) = x^*$ thus obtained.

If $x_0 = x^*$ it is easily verified that $x_1 = x^*$. If $x_0 = x^* + \delta$ where $\delta \in E$ with $\delta \neq 0$, then

$$\|X(1) - x^*\| < \|x_0 - x^*\|.$$

We wish x_1 to be closer to x^* than x_0 is, i.e.

$$(3.3.2)\qquad \|x_1 - x^*\| < \|x_0 - x^*\|.$$

In other words, we want the Runge-Kutta method to damp out the pertubations δ, just like the true solution does. This requirement is closely related to the concept of *strong stability of discretization methods for initial value problems* (cf. [STETTER, 1973; section 1.5.3]).

Suppose we solve (3.3.1) numerically, using Euler's method with stepsize $h = 1$. We then obtain an approximation x_1 to $X(1) = x^*$ that satisfies

$$(3.3.3)\qquad x_1 = x_0 - \Gamma(x_0)F(x_0).$$

Hence, in this case, x_1 is the first Newton iterate. Therefore, *the iterative process* [M,F] *where* M *is Newton's method is of type* (1.1.13) *where Euler's method with stepsize* $h = 1$ *is used.*

Consequently, *nonconvergence* of Newton's method might be conceived as a case in which Euler's method with stepsize $h = 1$ is *not strongly stable* with respect to perturbations in the vector x_0 occurring in the initial value problem (3.3.1).

3.3.2. Class of strongly stable Runge-Kutta methods

Consider the initial value problem

$$(3.3.4a) \qquad \dot{Y}(t) = CY(t) \qquad (t \geq 0),$$

$$(3.3.4b) \qquad Y(0) = y_0$$

where $y_0 \in \mathbb{R}^n$, $C \in L(\mathbb{R}^n)$ and all eigenvalues μ of C satisfy $\mathrm{Re}(\mu) < 0$. This type of differential equation has the property that perturbations that are introduced at $t = 0$ are damped out as t grows (cf. [STETTER, 1973; Theorem 2.3.4]).

Obviously, problem (3.3.4) is a rather poor model for problems of the type (3.3.1). This last type of problem, however, turns out to be too complicated in the analysis of strong stability of Runge-Kutta methods.

Let $\theta \in [0,\frac{1}{2}]$ and let R_θ be the Runge-Kutta method with operator coefficients for which

$$(3.3.5) \qquad R_\theta(z) \equiv \begin{pmatrix} 0 & 0 \\ [1-(1-\theta)z]^{-1} & 0 \end{pmatrix}.$$

Suppose we solve the initial value problem (3.3.4a) where $Y(0) = y_0 + \delta$ with $\delta \in \mathbb{R}^n$. It can be shown that if the numerical integration procedure uses R_θ with stepsize $h > 0$, the effect of δ is damped out in the approximations y_n of $Y(t_n)$ if t_n grows (cf. [LAMBERT, 1973; pp. 240-241], where methods are described that, when applied to (3.3.4), are equivalent to the methods R_θ ($\theta \in [0,\frac{1}{2}]$)). On the other hand, if we use Euler's method with stepsize $h > 0$, this need *not* be the case (cf. [LAMBERT, 1973; p. 227]).

Such phenomena may also occur in problems of the type (3.3.4) where $t \in [0,T]$ with $T < \infty$ (cf. [STETTER, 1973; section 2.3.7]).

Let $\bar{M}$ denote an iterative method for which the iterative process $[\bar{M},F]$ is of the type (1.1.13) and which is based on the numerical integration of (3.3.1), using R_θ with $\theta \in [0,\frac{1}{2}]$. In view of the above considerations one might expect $\bar{M}$ to be more successful than Newton's method in some cases where the latter fails (at least if $E = \mathbb{R}^n$).

In chapter 7 we shall therefore apply iterative methods of the type $\bar{M}$ to problems for which Newton's method fails. In that chapter we shall also consider iterative processes of the type (1.1.13) that are based on the numerical integration of (3.3.1), using the Runge-Kutta methods with scalar coefficients L_1 or L_2. L_1 and L_2 are defined by

$$(3.3.6) \qquad L_1 = \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{8} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad L_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{64} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{20} & 0 & 0 & 0 \\ 0 & 0 & \frac{5}{32} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} .$$

With respect to problem (3.3.4) these methods have a strong stability behaviour that is better than that of Euler's method when all eigenvalues μ of the linear operator C in (3.3.4) satisfy $\mu \in \mathbb{R}$ and $\mu < 0$ (cf. [VAN DER HOUWEN, 1977; pp. 89-90]).

For the sake of comparison, we shall also consider iterative processes of the type (1.1.13) that are based on the numerical integration of (3.3.1), using Euler's method with stepsizes $h < 1$.

CHAPTER 4

CONSTRUCTION OF ITERATIVE METHODS

In this chapter we derive general formulae for iterative methods which are constructed by means of iterative imbedding (cf. p. 4).

4.1. ITERATIVE METHODS BASED ON RUNGE-KUTTA METHODS WITH SCALAR COEFFICIENTS

Let $A \in \mathcal{A}$ and $g \in \mathcal{S}$ be given (cf. (2.6.3) and (2.6.9)). Let $L = (\lambda_{i,j})$ be an m-stage Runge-Kutta method with scalar coefficients and let $\mathcal{H} = \{h_0, h_1, \ldots, h_N\}$ be a sequence of stepsizes.

Let $F \in \mathcal{F}_1$ (cf. (2.6.1)) and set $K = A(F)$.

The iterative method to be described here will be constructed by means of iterative imbedding. More specifically, it will be based on the numerical integration of the initial value problem

$$\dot{X}(t) = -[(1-t)\partial_1 K(X(t),x_0) + tF'(X(t))]^{-1} \times$$

$$\{-K(X(t),x_0) + F(X(t)) + g(t)[(1-t)K(X(t),x_0) + tF(X(t))]\}$$

$$(t \in [0,1]),$$

(4.1.1)

$$X(0) = x_0$$

(cf. (2.6.12)). In short,

$$\dot{X}(t) = f(t,X(t)) \qquad (t \in [0,1]),$$

(4.1.2)

$$X(0) = x_0 ,$$

where $f(t,y) \equiv \Phi(t,y,x_0)$ (cf. (2.6.12b)). In (4.1.1) (or, equivalently, (4.1.2)) $x_0 \in D(F)$ is given. The numerical integration procedure is based on the Runge-Kutta method L and the sequence of stepsizes $\mathcal{H}$.

Suppose the initial value problem (4.1.1) has a unique solution X. Then according to Theorem 2.6.4(ii) it follows that $X(1) = x^*$. The Runge-Kutta approximation x_1 of $X(1) = x^*$ is given by

$$x_1 = y_{N+1}$$

where

$$y_0 = x_0$$

and

$$y_{n+1} = z^n_{m+1} \qquad (n = 0,1,\ldots,N).$$

Herewith we define for $n = 0,1,\ldots,N$ the quantities z^n_ℓ $(\ell = 1,2,\ldots,m+1)$ by

$$k^n_{\ell-1} = f(t_n + \nu_{\ell-1}h_n, z^n_{\ell-1})$$

(if $\ell > 1$) and

$$z^n_\ell = y_n + h_n \sum_{j=1}^{\ell-1} \lambda_{\ell,j}k^n_j.$$

If we repeat this process in the way described on page 4 we arrive at an iterative process [M,F] of type (1.1.13). M is an iterative method for $\mathcal{F}_1$, depending on A, g, L and $\mathcal{H}$. For the sake of clarity we shall sometimes denote this iterative method by $\mathbb{M}(A,g,L,\mathcal{H})$ in order to emphasize its dependence on A, g, L and $\mathcal{H}$. Hence

$$\text{(4.1.3)} \qquad M = \mathbb{M}(A,g,L,\mathcal{H}).$$

We shall now give, for $F \in \mathcal{F}_1$, an expression of $G = M(F)$ in terms of K, g, L and $\mathcal{H}$, where $K = A(F)$.

$$\text{(4.1.4a)} \qquad \begin{aligned} &G: D(G) \to E, \\ &G = \eta_{N+1} \end{aligned}$$

where

$$\eta_0 : D(\eta_0) \to E,$$

(4.1.4b) $$D(\eta_0) = D(F),$$

$$\eta_0(x) = x \qquad (x \in D(\eta_0))$$

and

(4.1.4c) $$\eta_{n+1} = \zeta^n_{m+1} \qquad (n = 0,1,\dots,N).$$

Herewith we define for $n = 0,1,\dots,N$ the functions ζ^n_ℓ ($\ell = 1,2,\dots,m+1$) by

$$\kappa^n_{\ell-1} : D(\kappa^n_{\ell-1}) \to E,$$

(4.1.4d) $$D(\kappa^n_{\ell-1}) = \{x \mid x \in D(\zeta^n_{\ell-1});\ (t_n + \nu_{\ell-1}h_n, \zeta^n_{\ell-1}(x), x) \in D(\Phi)\},$$

$$\kappa^n_{\ell-1}(x) = \Phi(t_n + \nu_{\ell-1}h_n, \zeta^n_{\ell-1}(x), x) \qquad (x \in D(\kappa^n_{\ell-1}))$$

(if $\ell > 1$) and

$$\zeta^n_\ell : D(\zeta^n_\ell) \to E,$$

(4.1.4e) $$D(\zeta^n_\ell) = \{x \mid x \in D(\eta_n);\ x \in D(\kappa^n_j)\ (j = 1,2,\dots,\ell-1),\ \text{if } \ell > 1\},$$

$$\zeta^n_\ell(x) = \eta_n(x) + h_n \sum_{j=1}^{\ell-1} \lambda_{\ell,j}\kappa^n_j(x) \qquad (x \in D(\zeta^n)).$$

4.2. ITERATIVE METHODS BASED ON RUNGE-KUTTA METHODS WITH OPERATOR COEFFICIENTS

Again, iterative methods that will be described in this section are based on the numerical integration of (4.1.1) (or (4.1.2)), where K, g and x_0 are given. However, the numerical integration procedure will now be based on a Runge-Kutta method with operator coefficients.

4.2.1. Preliminaries

When using a Runge-Kutta method with operator coefficients to solve the initial value problem (4.1.2), an expression for $\partial_2 f(t,y)$ is required.

In this subsection we give a lemma which will be used in the next subsection to derive a formula for $\partial_2 f(t,y)$.

We first give a lemma that will often be used subsequently.

LEMMA 4.2.1. *Let* $C, \Delta \in L(E)$. *Suppose* C *is invertible and*

$$\|\Delta\| < \frac{1}{\|[C]^{-1}\|}\,.$$

Then $C + \Delta$ *is invertible and*

$$\|[C+\Delta]^{-1}\| \le \frac{\|[C]^{-1}\|}{1-\|[C]^{-1}\|\cdot\|\Delta\|}\,.$$

PROOF. see [KANTOROWITSCH & AKILOW, 1964; Theorem 4(2.V)]. □

Let $P\colon U \to E$ and $Q\colon U \to L(E)$ where U is an open subset of E. If the derivative of Q at $x \in U$ exists, we have $Q'(x) \in L^{(2)}(E)$, (see section 2.1).

LEMMA 4.2.2. *Assume that* $Q(x)$ *is invertible for all* $x \in U$. *Define*

$$W\colon U \to E,$$

$$W(x) = [Q(x)]^{-1}P(x) \qquad (x \in U).$$

Let $y \in U$. *Assume that* $Q'(y)$ *and* $P'(y)$ *exist. Then* $W'(y)$ *exists and*

$$W'(y)z = [Q(y)]^{-1}P'(y)z - [Q(y)]^{-1}Q'(y)z[Q(y)]^{-1}P(y)$$

$$(\text{for all } z \in E).$$

PROOF. 1. Let $I(E)$ denote the subset of $L(E)$ that consists of all invertible linear operators on E. From Lemma 4.2.1 it follows that this subset is open in $L(E)$. Let

$$R_1\colon I(E) \to L(E),$$

$$R_1(C) = [C]^{-1} \qquad (C \in I(E)).$$

From [RALL, 1969; p. 96] it follows that for all $C \in I(E)$ the derivative $R_1'(C)$ exists and $[R_1'(C)](\tilde{C}) = -[C]^{-1}\tilde{C}[C]^{-1}$ (for all $\tilde{C} \in L(E)$). Let

$$R_2 \colon U \to L(E),$$

$$R_2(x) = R_1(Q(x)) \qquad (x \in U).$$

From the chainrule it follows that R_2 is differentiable at y and

$$R_2'(y)z = R_1'(Q(y))Q'(y)z = -[Q(y)]^{-1}Q'(y)z[Q(y)]^{-1} \tag{4.2.1}$$

$$\text{(for all } z \in E).$$

2. Obviously $W(x) \equiv R_2(x)P(x)$. Using the relation (4.2.1) it follows that to prove the lemma, it is sufficient to prove that $W'(y)$ exists and

$$W'(y)z = R_2'(y)zP(y) + R_2(y)P'(y)z \qquad \text{(for all } z \in E).$$

Choose $\sigma > 0$ such that $B(y,\sigma) \subset U$. For $h \in E$ with $\|h\| < \sigma$ let

$$\varepsilon_1(h) = R_2(y+h) - R_2(y) - R_2'(y)h$$

and

$$\varepsilon_2(h) = P(y+h) - P(y) - P'(y)h.$$

Then

$$W(y+h) - W(y) = R_2(y)[P'(y)h + \varepsilon_2(h)]$$

$$+ [R_2'(y)h + \varepsilon_1(h)][P(y) + P'(y)h + \varepsilon_2(h)].$$

Since, for i = 1,2,

$$\frac{\|\varepsilon_i(h)\|}{\|h\|} \to 0 \quad \text{when} \quad h \to 0 \ (h \neq 0),$$

this completes the proof. □

4.2.2. Formula for iterative methods based on Runge-Kutta methods with operator coefficients

Let $A \in \mathcal{A}$ and $g \in \mathcal{S}$. Let $R = (\rho_{i,j})$ be an m-stage Runge-Kutta method with operator coefficients and let $\mathcal{H} = \{h_0, h_1, \ldots, h_N\}$ be a sequence of stepsizes.

Let $F \in \mathcal{F}_1$ and set $K = A(F)$. Let Φ be defined by (2.6.12b).

LEMMA 4.2.3. *Let* $(t,y,x) \in D(\Phi)$. *Then* $y \in \text{interior}(D(\Phi(t,\cdot,x)))$, $x \in \text{interior}(D(\Phi(t,y,\cdot)))$ *and the partial derivatives* $\partial_2\Phi(t,y,x)$ *and* $\partial_3\Phi(t,y,x)$ *exist. We have*

$$\partial_2\Phi(t,y,x) =$$

$$-g(t)I - [(1-t)\partial_1K(y,x) + tF'(y)]^{-1} \times \tag{4.2.2}$$

$$\{-\partial_1K(y,x) + F'(y) + [(1-t)\partial_{11}K(y,x) + tF''(y)]\Phi(t,y,x)\}$$

and

$$\partial_3\Phi(t,y,x) =$$

$$[(1-t)\partial_1K(y,x) + tF'(y)]^{-1}\{[1-g(t)(1-t)]\partial_2K(y,x) \tag{4.2.3}$$

$$- (1-t)\partial_{21}K(y,x)\Phi(t,y,x)\}.$$

PROOF. Since $(t,y,x) \in D(\Phi)$, the operator $[(1-t)\partial_1K(y,x) + tF'(y)]$ is invertible. Set $\alpha = \|[(1-t)\partial_1K(y,x) + tF'(y)]^{-1}\|$. We recall that $D(F)$ is an open subset of E. From Lemma 2.6.3 it follows that ∂_1K is continuous at (y,x). Further $F''(y)$ exists. Consequently, a number $\sigma > 0$ exists such that $B(x,\sigma) \subset D(F)$, $B(y,\sigma) \subset D(F)$ and

$$\|[(1-t)\partial_1K(\bar{y},\bar{x}) + tF'(\bar{y})] - [(1-t)\partial_1K(y,x) + tF'(y)]\| < \frac{1}{\alpha}$$

(for all $\bar{x} \in B(x,\sigma)$ and $\bar{y} \in B(y,\sigma)$).

From Lemma 4.2.1 it follows that for all $\bar{y} \in B(y,\sigma)$ and $\bar{x} \in B(x,\sigma)$ we have $(t,\bar{y},\bar{x}) \in D(\Phi)$ so that $B(y,\sigma) \subset D(\Phi(t,\cdot,x))$ and $B(x,\sigma) \subset D(\Phi(t,y,\cdot))$.

From [RALL, 1969; Theorem 18.1] it follows that $Az_1z_2 = Az_2z_1$ (for all z_1 and z_2 in E) with $A = [(1-t)\partial_{11}K(y,x) + tF''(y)]$. Consequently, from Lemma 4.2.2 it follows that $\partial_2\Phi(t,y,x)$ exists and that (4.2.2) holds.

Since $\partial_{12}K(y,x)$ and $\partial_{21}K(y,x)$ exist, and $\partial_{12}K(y,x)z_1z_2 = \partial_{21}K(y,x)z_2z_1$ (for all z_1 and z_2 in E) (cf. [RALL, 1969; p. 116]), from Lemma 4.2.2 it follows that $\partial_3\Phi(t,y,x)$ exists and that (4.2.3) holds. □

Let $x_0 \in D(F)$ be given. Set $f(t,y) \equiv \Phi(t,y,x_0)$. From Lemma 4.2.3 it follows that whenever $(t,y,x_0) \in D(\Phi)$, the partial derivative $\partial_2\Phi(t,y,x_0)$ exists. Thus $\partial_2 f(t,y)$ exists and $\partial_2 f(t,y) = \partial_2\Phi(t,y,x_0)$ (for all $(t,y) \in D(f)$).

Consider the initial value problem (4.1.1) and suppose it has a unique solution X. Then according to Theorem 2.6.4(ii) it follows that $X(1) = x^*$. The Runge-Kutta approximation x_1 of $X(1) = x^*$ is given by

$$x_1 = y_{N+1}$$

where

$$y_0 = x_0$$

and

$$y_{n+1} = z^n_{m+1} \qquad (n = 0,1,\dots,N).$$

Herewith we define for $n = 0,1,\dots,N$ the quantities z^n_ℓ $(\ell = 1,2,\dots,m+1)$ by

$$k^n_{\ell-1} = f(t_n + \nu_{\ell-1}h_n, z^n_{\ell-1})$$

(if $\ell > 1$),

$$\Lambda^{(n)}_{\ell,j} = \rho_{\ell,j}(h_n\partial_2 f(t_n,y_n)) \qquad (j = 1,\dots,\ell-1)$$

(if $\ell > 1$) and

$$z^n_\ell = y_n + h_n \sum_{j=1}^{\ell-1} \Lambda^{(n)}_{\ell,j}k^n_j \ .$$

If we repeat this process in the way described on page 4 we arrive at an iterative process [M,F] of type (1.1.13). M is an iterative method for F_1, depending on A, g, R and H. Analogous to (4.1.3) we shall sometimes denote this iterative method by $\mathbb{M}(A,g,R,\mathcal{H})$. Hence

$$M = \mathbb{M}(A,g,R,\mathcal{H}). \tag{4.2.4}$$

We recall that for a rational function with real coefficients ρ, the set $D_E(\rho)$ is defined by (2.2.7). We shall now give, for $F \in \mathcal{F}_1$, an expression of $G = M(F)$ in terms of K, g, R and $\mathcal{H}$, where $K = A(F)$.

$$\begin{aligned} &G: D(G) \to E, \\ &G = \eta_{N+1} \end{aligned} \tag{4.2.5a}$$

where

$$\begin{aligned} &\eta_0 : D(\eta_0) \to E, \\ &D(\eta_0) = D(F), \\ &\eta_0(x) = x \qquad (x \in D(\eta_0)) \end{aligned} \tag{4.2.5b}$$

and

$$\eta_{n+1} = \zeta^n_{m+1} \qquad (n = 0,1,\ldots,N). \tag{4.2.5c}$$

Herewith we define for $n = 0,1,\ldots,N$ the functions ζ^n_ℓ ($\ell = 1,2,\ldots,m+1$) by

$$\begin{aligned} &\kappa^n_{\ell-1} : D(\kappa^n_{\ell-1}) \to E, \\ &D(\kappa^n_{\ell-1}) = \{x \mid x \in D(\zeta^n_{\ell-1});\ (t_n + \nu_{\ell-1}h_n, \zeta^n_{\ell-1}(x), x) \in D(\Phi)\}, \\ &\kappa^n_{\ell-1}(x) = \Phi(t_n + \nu_{\ell-1}h_n, \zeta^n_{\ell-1}(x), x) \qquad (x \in D(\kappa^n_{\ell-1})) \end{aligned} \tag{4.2.5d}$$

(if $\ell > 1$) and

$$\zeta_{\ell}^{n}\colon D(\zeta_{\ell}^{n}) \to E,$$

(4.2.5e)
$$D(\zeta_{\ell}^{n}) = \{x \mid x \in D(\eta_n);\ x \in D(\kappa_j^n) \text{ and } h_n \partial_2 \Phi(t_n, \eta_n(x), x) \in D_E(\rho_{\ell,j}) (j = 1,2,\ldots,\ell-1),\text{ if } \ell > 1\},$$

$$\zeta_{\ell}^{n}(x) = \eta_n(x) + h_n \sum_{j=1}^{\ell-1} \Lambda_{\ell,j}^{(n)} \kappa_j^n(x) \qquad (x \in D(\zeta^n)),$$

where, for $\ell > 1$,

(4.2.5f)
$$\Lambda_{\ell,j}^{(n)} = \rho_{\ell,j}(h_n \partial_2 \Phi(t_n, \eta_n(x), x)) \qquad (j = 1,2,\ldots,\ell-1).$$

CHAPTER 5

LOCAL CONVERGENCE

As already noted in section 2.5, an iterative process will have a positive radius of convergence if and only if it is locally convergent.

Throughout this chapter we denote with A and g given (fixed) elements belonging to $\mathcal{A}$ and $\mathcal{S}$ respectively (see (2.6.3) and (2.6.9)). Further, F denotes a given (fixed) operator, $F \in \mathcal{F}_1$ (see (2.6.1)) and $K = A(F)$.

Let R be a Runge-Kutta method with operator coefficients and let $\mathcal{H}$ be a sequence of stepsizes. In this chapter we investigate the conditions which should be imposed on A, g, R and $\mathcal{H}$ for the iterative process [M,F] where $M = \mathbb{M}(A,g,R,\mathcal{H})$ (cf. (4.2.4)) to be locally convergent. In particular we shall investigate under what conditions on A, g, R and $\mathcal{H}$ the iterative process [M,F] exhibits one of the three types of convergence behaviour that were introduced in the Definitions 2.5.2 - 4. The most obvious conditions are presented in subsection 5.2.2, in which R is supposed to be a Runge-Kutta method with scalar coefficients, and in subsection 5.2.3.

5.1. NECESSARY CONDITIONS AND SUFFICIENT CONDITIONS

Throughout section 5.1 we denote with $R = (\rho_{i,j})$ and $\mathcal{H} = \{h_0, h_1, \ldots, h_N\}$ respectively a given (fixed) m-stage Runge-Kutta method with operator coefficients and a given (fixed) sequence of stepsizes. In particular, R may be a Runge-Kutta method with scalar coefficients.

Let $M = \mathbb{M}(A,g,R,\mathcal{H})$ (cf. (4.2.4)). In this section we present a theorem that gives necessary conditions and sufficient conditions for the iterative process [M,F] to have a local convergence behaviour of one of the types that were introduced in the Definitions 2.5.2 - 4.

To that end we introduce the following functions.

For any $t \in [0,1]$, let

(5.1.1) $$D_t = \{z \mid z \in \mathbb{C};\ 1 - t + tz \neq 0\}.$$

For i = 1,2, let

(5.1.2a) $$u_i\colon D(u_i) \to \mathbb{C}$$

where

(5.1.2b) $$D(u_i) = \{(t,z) \mid z \in D_t,\ t \in [0,1]\}$$

and for all $(t,z) \in D(u_i)$,

(5.1.2c) $$u_i(t,z) = \begin{cases} [1-t+tz]^{-1}[1-z] - g(t) & (\text{if } i = 1), \\ -[1-g(t)(1-t)][1-t+tz]^{-1} & (\text{if } i = 2). \end{cases}$$

Let

(5.1.3a) $$\gamma\colon D(\gamma) \to \mathbb{C},$$ $$\gamma = \tau_{N+1}$$

where

(5.1.3b) $$\tau_0\colon \mathbb{C} \to \mathbb{C},$$ $$\tau_0(z) \equiv 1$$

and

(5.1.3c) $$\tau_{n+1} = \alpha^n_{m+1} \qquad (n = 0,1,\ldots,N).$$

Herewith we define for n = 0,1,...,N the functions α^n_ℓ ($\ell = 1,2,\ldots,m+1$) by

(5.1.3d) $$\pi^n_{\ell-1}\colon D(\pi^n_{\ell-1}) \to \mathbb{C},$$ $$D(\pi^n_{\ell-1}) = D(\alpha^n_{\ell-1}) \cap D_{t_n+\nu_{\ell-1}h_n},$$ $$\pi^n_{\ell-1}(z) = u_1(t_n+\nu_{\ell-1}h_n,z)\alpha^n_{\ell-1}(z) + u_2(t_n+\nu_{\ell-1}h_n,z) \quad (z \in D(\pi^n_{\ell-1}))$$

(if $\ell > 1$) and

$$\alpha^n_\ell : D(\alpha^n_\ell) \to \mathbb{C},$$

(5.1.3e)
$$D(\alpha^n_\ell) = \{z \mid z \in D(\tau_n);\ z \in D(\pi^n_j) \text{ and } h_n u_1(t_n,z) \in D_{\mathbb{C}}(\rho_{\ell,j})\ (j = 1,2,\dots,\ell-1),\ \text{if } \ell > 1\},$$

$$\alpha^n_\ell(z) = \tau_n(z) + h_n \sum_{j=1}^{\ell-1} \rho_{\ell,j}(h_n u_1(t_n,z))\pi^n_j(z) \qquad (z \in D(\alpha^n_\ell)).$$

Consider the following conditions (we note that $sr(T)$ with $T \in L(E)$ is defined by (2.2.2); the set $D_E(\gamma)$ is defined by (2.2.7)).

<u>CONDITION 0</u>. $\partial_1 K(x^*,x^*)$ is invertible and $C \in D_E(\gamma)$ where $C = [\partial_1 K(x^*,x^*)]^{-1}F'(x^*)$.

<u>CONDITION 1</u>. Condition 0 holds and $sr(\gamma(C)) < 1$.

<u>CONDITION 2</u>. Condition 0 holds and $sr(\gamma(C)) \le 1$.

<u>CONDITION 3</u>. Condition 0 holds and $\|\gamma(C)\| < 1$.

<u>CONDITION 4</u>. Condition 0 holds and $\|\gamma(C)\| \le 1$.

<u>CONDITION 5</u>. Condition 0 holds and $\gamma(C) = 0$.

Then the following theorem holds.

<u>THEOREM 5.1.1</u>. *The following propositions* (i) - (v) *are true.*

(i) *Condition* 1 *is a <u>sufficient</u> condition for the iterative process* [M,F] *to be <u>stably</u> convergent.*

(ii) *Condition* 2 *is a <u>necessary</u> condition for the iterative process* [M,F] *to be <u>stably</u> convergent.*

(iii) *Condition* 3 *is a <u>sufficient</u> condition for the iterative process* [M,F] *to be <u>regularly</u> convergent.*

(iv) *Condition* 4 *is a <u>necessary</u> condition for the iterative process* [M,F] *to be <u>regularly</u> convergent.*

(v) *Condition* 5 *is a <u>necessary</u> and <u>sufficient</u> condition for the iterative process* [M,F] *to be <u>quadratically</u> convergent.*

The proof of this theorem will be given in section 5.3.

In many cases it might be very complicated to verify whether or not one of the Conditions 1 - 5 is satisfied. In the next section we shall give, amongst other things, less complicated stipulations under which one of the Conditions 1 - 5 is satisfied.

5.2. FURTHER CONDITIONS ON THE ITERATIVE METHODS

This section is divided into three subsections. The first contains two theorems that give sufficient conditions under which the iterative process $[M,F]$ where $M = \mathbb{M}(A,g,R,\mathcal{H})$ is stably convergent and quadratically convergent, respectively. Here R is a given Runge-Kutta method with operator coefficients and $\mathcal{H}$ is a given sequence of stepsizes. These conditions are simpler than those given in Theorem 5.1.1. The last two subsections give still simpler conditions under which $[M,F]$ exhibits one of the types of convergence behaviour that were considered in section 5.1.

5.2.1. Simpler sufficient conditions

Throughout subsection 5.2.1 we denote with $R = (\rho_{i,j})$ and $\mathcal{H} = \{h_0, h_1, \ldots, h_N\}$ respectively a given (fixed) m-stage Runge-Kutta method with operator coefficients and a given (fixed) sequence of stepsizes. Let $M = \mathbb{M}(A,g,R,\mathcal{H})$. Consider the following condition:

$$\text{for } n = 0,1,\ldots,N, \tag{5.2.1}$$
$$-h_n g(t_n) \in D_{\mathbb{C}}(\rho_{\ell,j}) \text{ and } \rho_{\ell,j}(-h_n g(t_n)) = \rho_{\ell,j}(0)$$
$$(\ell = 2,3,\ldots,m+1;\ j = 1,2,\ldots,\ell-1).$$

The following two theorems hold.

<u>THEOREM 5.2.1</u>. *Assume that* (5.2.1) *holds and that* $\mathrm{sp}([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) = \{1\}$. *Then the iterative process* $[M,F]$ *is stably convergent.*

<u>THEOREM 5.2.2</u>. *Assume that* (5.2.1) *holds and that* $\partial_1 K(x^*,x^*) = F'(x^*)$. *Then the iterative process* $[M,F]$ *is quadratically convergent.*

We shall prove these theorems in section 5.3.

<u>REMARK 5.2.1</u>. If $E = \mathbb{R}$ then, obviously, the assumptions of Theorem 5.2.1 and Theorem 5.2.2 are equivalent.

Suppose $E \neq \mathbb{R}$, then this need not be the case, as the following example shows.

Choose $u_1, u_2 \in E$ such that $(u_1, u_2) = 0$ and $\|u_1\| = \|u_2\| = 1$. Define $T \in L(E)$ by

$$Tx = (x, u_1)u_2 \qquad (x \in E).$$

Thus, $T^2 = 0$ so that $\mathrm{sp}(T) = \{0\}$. Suppose that, for the operator $A \in A$,

$$[A(\tilde{F})](y,x) \equiv \tilde{F}(y) - \tilde{F}(x) + T(y-x) \qquad (\tilde{F} \in F_1).$$

Suppose that $F(x) \equiv x$. Hence $F'(x^*) = I$ and $\partial_1 K(x^*,x^*) = I + T$. Consequently, $\partial_1 K(x^*,x^*) \neq F'(x^*)$. On the other hand, from Theorem 2.2.2 we have $\mathrm{sp}([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) = \mathrm{sp}(I+T) = \{1\}$. □

REMARK 5.2.2. $\partial_1 K(x^*,x^*) = F'(x^*)$ if, for example

$$[A(F)](y,x) \equiv F(y) - F(x), \quad \text{or}$$

$$[A(F)](y,x) \equiv F'(x)(y-x).$$

The imbedding (1.1.5) with $K = A(F)$ where

$$[A(F)](y,x) \equiv y - x$$

is sometimes used in the discrete imbedding method (see e.g. [MENZEL & SCHWETLICK, 1976]). Since in general $\partial_1 K(x^*,x^*) \neq F'(x^*)$, this choice of A does not seem to be suitable for iterative imbedding (see also the Theorems 5.2.3 - 8). □

REMARK 5.2.3. The condition (5.2.1) holds if $R = L$ where L is a Runge-Kutta method with scalar coefficients. This case will be considered in the next subsection.

If R is a Runge-Kutta method with operator coefficients, the condition (5.2.1) holds if $g(t_n) = 0$ $(n = 0,1,\ldots,N)$. This case will be considered in subsection 5.2.3. □

5.2.2. Iterative methods based on Runge-Kutta methods with scalar coefficients

We recall that throughout this subsection A, g and F are given (fixed) elements belonging to $\mathcal{A}$, $\mathcal{S}$ and $\mathcal{F}_1$ respectively.

The iterative methods to be considered in this subsection are of the type described in section 4.1. We shall give three theorems that give necessary and sufficient conditions on A in order that for any Runge-Kutta method with scalar coefficients L and any sequence of stepsizes $\mathcal{H}$, the iterative process [M,F] where $M = \mathbb{M}(A,g,L,\mathcal{H})$ may have one of the three types of convergence behaviour with which Theorem 5.1.1 was concerned.

The proofs of these theorems will be given in section 5.3.

THEOREM 5.2.3. *The following propositions* (i) *and* (ii) *are equivalent.*

(i) *For any Runge-Kutta method with scalar coefficients* L *and any sequence of stepsizes* $\mathcal{H}$ *the iterative process* [M,F] *where* $M = \mathbb{M}(A,g,L,\mathcal{H})$ *is stably convergent.*

(ii) $sp([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) = \{1\}$.

This theorem is a consequence of Theorem 5.2.1 and Theorem 5.1.1(ii). The next two theorems are consequent on the last three propositions of Theorem 5.1.1 and of Theorem 5.2.2.

THEOREM 5.2.4. *The following propositions* (iii) *and* (iv) *are equivalent.*

(iii) *For any Runge-Kutta method with scalar coefficients* L *and any sequence of stepsizes* $\mathcal{H}$ *the iterative process* [M,F] *where* $M = \mathbb{M}(A,g,L,\mathcal{H})$ *is regularly convergent.*

(iv) $\partial_1 K(x^*,x^*) = F'(x^*)$.

THEOREM 5.2.5. *The following propositions* (v) *and* (vi) *are equivalent.*

(v) *For any Runge-Kutta method with scalar coefficients* L *and any sequence of stepsizes* $\mathcal{H}$ *the iterative process* [M,F] *where* $M = \mathbb{M}(A,g,L,\mathcal{H})$ *is quadratically convergent.*

(vi) $\partial_1 K(x^*,x^*) = F'(x^*)$.

An immediate consequence of these theorems is, of course, that the propositions (iii) and (v) are equivalent.

5.2.3. Iterative methods based on Runge-Kutta methods with operator coefficients

We recall that throughout this subsection A, g and F are given (fixed) elements belonging to A, S and F_1 respectively.

In this subsection we consider iterative methods of the type described in section 4.2. Analogous to the previous subsection we shall give three theorems that give necessary and sufficient conditions on A and g in order that, for any Runge-Kutta method with operator coefficients R and any sequence of stepsizes H, the iterative process [M,F] where $M = \mathbb{M}(A,g,R,H)$ may have one of the three types of convergence behaviour with which Theorem 5.1.1 was concerned.

The proofs of these theorems will be given in the next section.

THEOREM 5.2.6. *The following propositions* (i) *and* (ii) *are equivalent.*

(i) *For any Runge-Kutta method with operator coefficients* R *and any sequence of stepsizes* H *the iterative process* [M,F] *where* $M = \mathbb{M}(A,g,R,H)$ *is stably convergent.*

(ii) $\mathrm{sp}([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) = \{1\}$ *and* $g(t) = 0$ *(for all* $t \in [0,1)$*).*

Theorem 5.2.6 is similar to Theorem 5.2.3. As may be expected we can also obtain results similar to Theorem 5.2.4 and Theorem 5.2.5.

THEOREM 5.2.7. *The following propositions* (iii) *and* (iv) *are equivalent.*

(iii) *For any Runge-Kutta method with operator coefficients* R *and any sequence of stepsizes* H *the iterative process* [M,F] *where* $M = \mathbb{M}(A,g,R,H)$ *is regularly convergent.*

(iv) $\partial_1 K(x^*,x^*) = F'(x^*)$ *and* $g(t) = 0$ *(for all* $t \in [0,1)$*).*

THEOREM 5.2.8. *The following propositions* (v) *and* (vi) *are equivalent.*

(v) *For any Runge-Kutta method with operator coefficients* R *and any sequence of stepsizes* H *the iterative process* [M,F] *where* $M = \mathbb{M}(A,g,R,H)$ *is quadratically convergent.*

(vi) $\partial_1 K(x^*,x^*) = F'(x^*)$ *and* $g(t) = 0$ *(for all* $t \in [0,1)$*).*

An immediate consequence of these theorems is that the propositions (iii) and (v) are equivalent.

5.3. PROOF OF THE THEOREMS

As may be expected, the proofs of the theorems listed above exhibit similarities. Therefore in the next subsection we have collected some lemmata that will be useful for proving the theorems of both section 5.1 and section 5.2. Lemma 5.3.8 (statement (5.3.13d)), Lemma 5.3.10 (statement (5.3.14d)) and the Lemmata 5.3.11 and 5.3.12 will play an important role in the proofs of the theorems.

5.3.1. Preliminary lemmata

Throughout subsection 5.3.1 we denote with $R = (\rho_{i,j})$ and $H = \{h_0, h_1, \ldots, h_N\}$ respectively a given (fixed) m-stage Runge-Kutta method with operator coefficients and a given (fixed) sequence of stepsizes.

Furthermore, $G = M(F)$ where $M = \mathbb{M}(A,g,R,H)$. Since $F \in F_1$ and $K = A(F)$ where $A \in A$, there exist positive constants σ^*, β^*, μ, μ_1, μ_2 and μ_3 such that $B(x^*,\sigma^*) \subset D(F)$, and such that

(5.3.1a) $\|[F'(x^*)]^{-1}\| \leq \beta^*$,

(5.3.1b) $\|F''(x)\| \leq \mu$,

(5.3.1c) $\|\partial_{11}K(y,x)\| \leq \mu_1$, $\|\partial_{12}K(y,x)\| \leq \mu_2$ and $\|\partial_{21}K(y,x)\| \leq \mu_3$,

(5.3.1d) $K(x,x) = 0$

for all $x,y \in B(x^*,\sigma^*)$ (see (2.6.1), (2.6.2) and (2.6.3)). As a consequence of Lemma 2.6.1 and (5.3.1) we have for all $x,y,z \in B(x^*,\sigma^*)$

(5.3.2a) $\|F'(x) - F'(y)\| \leq \mu\|x-y\|$,

(5.3.2b) $\|\partial_1K(y,x) - \partial_1K(z,x)\| \leq \mu_1\|y-z\|$,

(5.3.2c) $\|\partial_1K(y,x) - \partial_1K(y,z)\| \leq \mu_2\|x-z\|$,

(5.3.2d) $\|\partial_2K(y,x) - \partial_2K(z,x)\| \leq \mu_3\|y-z\|$.

We note that

$$\partial_1 K(y,x) - \partial_1 K(x^*,x^*)$$

$$= \partial_1 K(y,x) - \partial_1 K(y,x^*) + \partial_1 K(y,x^*) - \partial_1 K(x^*,x^*).$$

Hence, for all $x,y \in B(x^*,\sigma^*)$ we have

$$\|\partial_1 K(y,x) - \partial_1 K(x^*,x^*)\| \le \mu_1\|y-x^*\| + \mu_2\|x-x^*\| \tag{5.3.3a}$$

and

$$\|[(1-t)\partial_1 K(y,x) + tF'(y)] - [(1-t)\partial_1 K(x^*,x^*) + tF'(x^*)]\| \tag{5.3.3b}$$

$$\le (1-t)\{\mu_1\|y-x^*\| + \mu_2\|x-x^*\|\} + t\mu\|y-x^*\| \qquad (\text{for all } t \in [0,1]).$$

The following lemma is a general result that will often be used subsequently.

LEMMA 5.3.1. *Let* $P\colon D(P) \to E$ *with* $D(P) \subset E$. *Let* $x,y \in D(P)$. *Suppose* $V = \{z \mid z = x+t(y-x)$ *with* $t \in [0,1]\} \subset \text{interior}(D(P))$. *Let* $P'(z)$ *exist (for all* $z \in V$*) and suppose a constant* $\delta > 0$ *exists such that* $\|P'(z)-P'(x)\| \le \delta\|z-x\|$ *(for all* $z \in V$*). Then*

$$\|P(y) - P(x) - P'(x)(y-x)\| \le \frac{\delta}{2}\|x-y\|^2.$$

PROOF. see [SPIJKER, 1972; Lemma 1]. □

We return to operator K.

LEMMA 5.3.2.

$$\partial_2 K(x,x) = -\partial_1 K(x,x) \qquad (\text{for all } x \in B(x^*,\sigma^*)).$$

PROOF. Let $x,y \in B(x^*,\sigma^*)$. Using (5.3.1d), (5.3.2b,c) and Lemma 5.3.1,

$$\|K(x,y) - K(x,x) + \partial_1 K(x,x)(y-x)\|$$

$$\le \|K(y,y) - K(x,y) - \partial_1 K(x,y)(y-x)\| + \|\partial_1 K(x,y) - \partial_1 K(x,x)\|\|x-y\|$$

$$\le \frac{\mu_1}{2}\|x-y\|^2 + \mu_2\|x-y\|^2.$$

This proves the lemma. □

Let $x,y \in B(x^*,\sigma^*)$. Then

$$\partial_2K(y,x) - \partial_2K(x^*,x^*) = \partial_2K(y,x) - \partial_2K(x^*,x) + \partial_2K(x^*,x)$$
$$-\partial_2K(x,x) + \partial_2K(x,x) - \partial_2K(x^*,x^*).$$

Therefore, as a consequence of (5.3.2d), Lemma 5.3.2 and (5.3.3a) we have

$$(5.3.4) \qquad \|\partial_2K(y,x) - \partial_2K(x^*,x^*)\| \le \mu_3\|y-x^*\| + (\mu_1+\mu_2+\mu_3)\|x-x^*\|$$
$$\text{(for all } x,y \in B(x^*,\sigma^*)).$$

We recall that the operator Φ is defined in (2.6.12b). Each iterative method described in chapter 4 is based on the numerical integration of the initial value problem (2.6.12a). Therefore, Φ will play an important role in the proofs of the theorems. The next five lemmata are concerned with Φ.

LEMMA 5.3.3. *Let* $t \in [0,1]$ *and suppose that* $(t,x^*,x^*) \in D(\Phi)$. *Assume that* $\partial_1K(x^*,x^*)$ *is invertible and set*

$$C = [\partial_1K(x^*,x^*)]^{-1}F'(x^*).$$

Then the operator $[(1-t)I + tC]$ *is invertible. Further*

$$x^* \in \text{interior}(D(\Phi(t,\cdot,x^*)) \cap D(\Phi(t,x^*,\cdot)))$$

and the derivatives $\partial_i\Phi(t,x^*,x^*)$ $(i = 2,3)$ exist and satisfy

$$\partial_2\Phi(t,x^*,x^*) = -g(t)I + [(1-t)I + tC]^{-1}[I-C],$$

$$\partial_3\Phi(t,x^*,x^*) = -\{1 - g(t)(1-t)\}[(1-t)I + tC]^{-1}.$$

PROOF. The result follows from Lemma 4.2.3 and Lemma 5.3.2. □

LEMMA 5.3.4. *Let* $t \in [0,1]$. *If* $(t,x^*,x^*) \in D(\Phi)$ *then positive constants* β *and* σ *exist such that* $(t,y,x) \in D(\Phi)$ *and*

$$(5.3.5) \qquad \|[(1-t)\partial_1 K(y,x) + tF'(y)]^{-1}\| \le \beta$$

whenever $x,y \in B(x^*,\sigma)$.

PROOF. Set $\beta = 2\|[(1-t)\partial_1 K(x^*,x^*) + tF'(x^*)]^{-1}\|$. Set $\hat{\mu} = \max\{\mu_1,\mu_2,\mu\}$ and $\sigma = \min\{\sigma^*, \frac{1}{2\beta\hat{\mu}}\}$. From Lemma 4.2.1 and (5.3.3b) the result follows. □

LEMMA 5.3.5. *Let* $t \in [0,1]$. *If* $(t,x^*,x^*) \in D(\Phi)$ *then positive constants* c_1, c_2 *and* σ *exist such that* $(t,y,x) \in D(\Phi)$ *and*

$$(5.3.6) \qquad \|\Phi(t,y,x)\| \le c_1\|x-x^*\| + c_2\|y-x^*\|$$

whenever $x,y \in B(x^*,\sigma)$.

PROOF. Set $\theta_1 = \|F'(x^*)\|$ and $\delta_1 = \theta_1 + \mu\sigma^*$. From (5.3.2a) it follows that $\|F'(z)\| \le \delta_1$ (for all $z \in B(x^*,\sigma^*)$). From Lemma 2.6.1 it now follows that

$$(5.3.7) \qquad \|F(y)\| \le \delta_1\|y-x^*\| \qquad \text{(for all } y \in B(x^*,\sigma^*)).$$

Set $\theta_2 = \|\partial_1 K(x^*,x^*)\|$ and $\delta_2 = \theta_2 + (\mu_1+\mu_2+\mu_3)\sigma^*$. From (5.3.3a) it follows that

$$\|\partial_1 K(z,x)\| \le \theta_2 + (\mu_1+\mu_2)\sigma^* \qquad \text{(for all } z,x \in B(x^*,\sigma^*)).$$

From (5.3.4) and Lemma 5.3.2 it follows that

$$\|\partial_2 K(x^*,z)\| \le \theta_1 + (\mu_1+\mu_2+\mu_3)\sigma^* \qquad \text{(for all } z \in B(x^*,\sigma^*)).$$

Let $x,y \in B(x^*,\sigma^*)$. Then

$$K(y,x) = K(y,x) - K(x^*,x) + K(x^*,x) - K(x^*,x^*).$$

Using Lemma 2.6.1 it follows that

$$(5.3.8) \qquad \|K(y,x)\| \le \delta_2\{\|y-x^*\| + \|x-x^*\|\} \qquad \text{(for all } x,y \in B(x^*,\sigma^*)).$$

From Lemma 5.3.4 it follows that numbers $\beta > 0$ and $\sigma \in (0,\sigma^*]$ exist such that for all $x,y \in B(x^*,\sigma)$ we have $(t,y,x) \in D(\Phi)$ and (5.3.5). Together with (5.3.7) and (5.3.8) this yields the result with $c_1 = \beta[1 + |g(t)|(1-t)]\delta_2$

and $c_2 = c_1 + \beta[1 + t|g(t)|]\delta_1$. □

LEMMA 5.3.6. *Let* $t \in [0,1]$. *Suppose* $(t,x^*,x^*) \in D(\Phi)$. *Then positive constants* $c_{1,i}$, $c_{2,i}$ $(i = 2,3)$ *and* σ *exist such that* $y \in \text{interior}(D(\Phi(t,\cdot,x)))$, $x \in \text{interior}(D(\Phi(t,y,\cdot)))$ *and the derivatives* $\partial_i\Phi(t,y,x)$ *exist and satisfy*

$$(5.3.9) \qquad \|\partial_i\Phi(t,y,x) - \partial_i\Phi(t,x^*,x^*)\| \le c_{1,i}\|x-x^*\| + c_{2,i}\|y-x^*\| \quad (i = 2,3)$$

whenever $x,y \in B(x^*,\sigma)$.

PROOF. 1. From Lemma 5.3.4 and Lemma 5.3.5 it follows that positive constants c_1, c_2, β and a number $\sigma \in (0,\sigma^*]$ exist such that for all $x,y \in B(x^*,\sigma)$ we have $(t,y,x) \in D(\Phi)$ and (5.3.5), (5.3.6). From Lemma 4.2.3 it follows that $x \in \text{interior}(D(\Phi(t,y,\cdot)))$, $y \in \text{interior}(D(\Phi(t,\cdot,x)))$ and the derivatives $\partial_i\Phi(t,y,x)$ $(i = 2,3)$ exist whenever $x,y \in B(x^*,\sigma)$.

Let $x,y \in B(x^*,\sigma)$, set $T = [(1-t)\partial_1K(y,x) + tF'(y)]$ and $T^* = [(1-t)\partial_1K(x^*,x^*) + tF'(x^*)]$.

2. From Lemma 4.2.3 it follows that

$$\begin{aligned}
&\partial_2\Phi(t,y,x) - \partial_2\Phi(t,x^*,x^*)\\
&= -[T]^{-1}\{-\partial_1K(y,x) + F'(y) + [(1-t)\partial_{11}K(y,x) + tF''(y)]\Phi(t,y,x)\}\\
&+ [T^*]^{-1}\{-\partial_1K(x^*,x^*) + F'(x^*)\}.
\end{aligned}$$

Note that

$$[T^*]^{-1} - [T]^{-1} = [T]^{-1}\{T - T^*\}[T^*]^{-1}.$$

Using this relation we obtain

$$\begin{aligned}
&\|\partial_2\Phi(t,y,x) - \partial_2\Phi(t,x^*,x^*)\|\\
&\le \|[T]^{-1}\|\{\|\partial_1K(y,x) - \partial_1K(x^*,x^*) - F'(y) + F'(x^*)\|\\
&+ [(1-t)\|\partial_{11}K(y,x)\| + t\|F''(y)\|]\|\Phi(t,y,x)\|\}\\
&+ \|[T]^{-1}\|\|[T^*]^{-1}\|\|T - T^*\|\{\|\partial_1K(x^*,x^*)\| + \|F'(x^*)\|\}.
\end{aligned}$$

Using (5.3.5), (5.3.3), (5.3.2a), (5.3.1b,c) and (5.3.6) the result follows for i = 2.

3. From Lemma 4.2.3 it follows that

$$\partial_3\Phi(t,y,x) - \partial_3\Phi(t,x^*,x^*)$$

$$= (1 - g(t)(1-t))\{[T]^{-1}\partial_2K(y,x) - [T^*]^{-1}\partial_2K(x^*,x^*)\}$$

$$- (1-t)[T]^{-1}\partial_{21}K(y,x)\Phi(t,y,x).$$

Hence

$$\|\partial_3\Phi(t,y,x) - \partial_3\Phi(t,x^*,x^*)\|$$

$$\leq |1 - g(t)(1-t)|\{\|[T]^{-1}\|\|\partial_2K(y,x) - \partial_2K(x^*,x^*)\| +$$

$$+ \|[T]^{-1}\|\|[T^*]^{-1}\|\|T - T^*\|\|\partial_2K(x^*,x^*)\|\}$$

$$+ (1-t)\|[T]^{-1}\|\|\partial_{21}K(y,x)\|\|\Phi(t,y,x)\|.$$

Using (5.3.5), (5.3.4), (5.3.3b), (5.3.1c) and (5.3.6) the result follows for i = 3. This completes the proof. □

LEMMA 5.3.7. *Let* $t \in [0,1]$ *and suppose* $(t,x^*,x^*) \in D(\Phi)$. *Let*

$$f: D(f) \to E,$$

$$f(z) = \Phi(t,y,x) \qquad (z = (y,x),\ z \in D(f))$$

with

$$D(f) = \{(y,x) \mid x,y \in E;\ (t,y,x) \in D(\Phi)\} \subset E \times E.$$

Then a constant $\sigma > 0$ *exists such that*

$$B(x^*,\sigma) \times B(x^*,\sigma) \subset D(f),\ \{t\} \times B(x^*,\sigma) \times B(x^*,\sigma) \subset D(\Phi),$$

$$f'((y,x)),\ \partial_2\Phi(t,y,x) \quad \textit{and} \quad \partial_3\Phi(t,y,x) \quad \textit{exist}$$

and

$$f'((y,x))(h_1,h_2) = \partial_2\Phi(t,y,x)h_1 + \partial_3\Phi(t,y,x)h_2 \quad ((h_1,h_2) \in E \times E) \tag{5.3.10}$$

whenever $x,y \in B(x^*,\sigma)$.

PROOF. Notice that the operators K, $\partial_1 K$ and $\partial_{11}K$ are continuous on $D(F) \times D(F)$ (cf. Lemma 2.6.3 and (2.6.2,3)). Further the operators F, F' and F'' are continuous on $D(F)$ (cf. (2.6.1)). From Lemma 5.3.6 it follows that a constant $\sigma > 0$ exists such that $t \times B(x^*,\sigma) \times B(x^*,\sigma) \subset D(\Phi)$ and the derivatives $\partial_i\Phi(t,y,x)$ $(i = 2,3)$ exist whenever $x,y \in B(x^*,\sigma)$. Consequently $B(x^*,\sigma) \times B(x^*,\sigma) \subset D(f)$. It is easily verified that the mapping $(y,x) \to \partial_2\Phi(t,y,x)$ and the operator f are continuous on $B(x^*,\sigma) \times B(x^*,\sigma)$. The result now follows from [BROWN & PAGE, 1970; pp. 284-285 and Theorem 7.4.2]. □

In the next lemmata Condition 0 will play an important role. We first introduce the function G^*.

Suppose that, for $n = 0,1,\ldots,N$,

$$x^* \in \text{interior}(D(\Phi(t_n,\cdot,x^*))), \text{ the derivative } \partial_2\Phi(t_n,x^*,x^*)$$

$$\text{exists and} \tag{5.3.11}$$

$$h_n\partial_2\Phi(t_n,x^*,x^*) \in D_E(\rho_{\ell,j}) \qquad (\ell = 2,3,\ldots,m+1;\ j = 1,2,\ldots,\ell-1)$$

(cf. (2.2.7)). Let

$$G^* = G \tag{5.3.12a}$$

where G is defined in (4.2.5a - d) in which the functions ζ^n_ℓ are defined by

$$\zeta^n_\ell\colon D(\zeta^n_\ell) \to E,$$

$$D(\zeta^n_\ell) = \{x \mid x \in D(\eta_n);\ x \in D(\kappa^n_j)\,(j = 1,2,\ldots,\ell-1),\text{ if } \ell > 1\}, \tag{5.2.12b}$$

$$\zeta^n_\ell(x) = \eta_n(x) + h_n \sum_{j=1}^{\ell-1} \Lambda^{*(n)}_{\ell,j}\,\kappa^n_j(x) \qquad (x \in D(\zeta^n_\ell))$$

$(n = 0,1,\ldots,N;\ \ell = 1,2,\ldots,m+1)$. In (5.3.12b)

(5.3.12c) $\quad \overset{*}{\Lambda}{}^{(n)}_{\ell,j} = \rho_{\ell,j}(h_n\partial_2\Phi(t_n,x^*,x^*)) \quad (n = 0,1,\dots,N;\ \ell = 2,3,\dots,m+1;$
$j = 1,2,\dots,\ell-1).$

For $n = 0,1,\dots,N$ we set

(5.3.12d) $\quad \overset{*}{\eta}_n = \eta_n,$

(5.3.12e) $\quad \overset{*}{\zeta}{}^n_\ell = \zeta^n_\ell$ and, if $\ell > 1$, $\overset{*}{\kappa}{}^n_{\ell-1} = \kappa^n_{\ell-1} \quad (\ell = 1,2,\dots,m+1).$

REMARK 5.3.1. Obviously, if $R = L$, where L is an m-stage Runge-Kutta method with scalar coefficients, then $G^* = G$ where $G = M(F)$ with $M = \mathbb{M}(A,g,L,\mathcal{H})$.
□

The operator G^* will be a useful tool in investigating the local convergence behaviour of $[M,F]$ where $M = \mathbb{M}(A,g,R,\mathcal{H})$. To that end we also need an expression of $[G^*]'(x^*)$.

Let the functions γ, τ_n $(n = 0,1,\dots,N+1)$, α^n_ℓ and, for $\ell > 1$, $\pi^n_{\ell-1}$ $(n = 0,1,\dots,N;\ \ell = 1,2,\dots,m+1)$ be defined in (5.1.3).

LEMMA 5.3.8. *Let Condition* 0 *be satisfied. Then* (5.3.11) *holds. Let* G^* *be defined in* (5.3.12). *Then for* $n = 0,1,\dots,N$ *the propositions* (5.3.13a), (5.3.13b) *and* (5.3.13c) *are true.*

(5.3.13a)
Positive constants σ_n *and* δ_n *exist such that* $B(x^*,\sigma_n) \subset D(\overset{*}{\eta}_n)$, $C \in D_E(\tau_n)$,
the derivative $\overset{*}{\eta}_n{}'(x)$ *exists and* $\|\overset{*}{\eta}_n{}'(x)-\tau_n(C)\| \le \delta_n\|x-x^*\|$ *whenever* $x \in B(x^*,\sigma_n)$.

For $\ell = 1,2,\dots,m+1$ *positive constants* $\sigma_{n,\ell}$, $\delta_{2,n,\ell}$ *and, if* $\ell > 1$, $\delta_{1,n,\ell-1}$ *exist such that*

(5.3.13b)
$B(x^*,\sigma_{n,\ell}) \subset D(\overset{*}{\kappa}{}^n_{\ell-1})$, $C \in D_E(\pi^n_{\ell-1})$, *the derivative* $[\overset{*}{\kappa}{}^n_{\ell-1}]'(x)$ *exists,*
$\|[\overset{*}{\kappa}{}^n_{\ell-1}]'(x) - \pi^n_{\ell-1}(C)\| \le \delta_{1,n,\ell-1}\|x-x^*\|$

(if $\ell > 1$) and

$$B(x^*,\sigma_{n,\ell}) \subset D(\overset{*}{\zeta}{}_{\ell}^{n}),\ C \in D_E(\alpha_{\ell}^{n}),\ \textit{the derivative}\ [\overset{*}{\zeta}{}_{\ell}^{n}]'(x)\ \textit{exists},$$

(5.3.13c)

$$\|[\overset{*}{\zeta}{}_{\ell}^{n}]'(x) - \alpha_{\ell}^{n}(C)\| \le \delta_{2,n,\ell}\|x-x^*\|$$

whenever $x \in B(x^*,\sigma_{n,\ell})$. *In particular, since* $G^* = \overset{*}{\eta}_{N+1} = \overset{*}{\zeta}{}_{m+1}^{N}$ *and* $\gamma = \tau_{N+1} = \alpha_{m+1}^{N}$,

$$B(x^*,\sigma_{N,m+1}) \subset D(G^*),\ \textit{the derivative}\ [G^*]'(x^*)\ \textit{exists and}$$

(5.3.13d)

$$\|[G^*]'(x) - \gamma(C)\| \le \delta_{2,N,m+1}\|x-x^*\|$$

whenever $x \in B(x^*,\sigma_{N,m+1})$.

PROOF. Since Condition 0 is satisfied, (5.3.11) holds (cf. Lemma 5.3.3, (5.1.2) and (5.1.3e)).

1. Let $n = 0$. Since $\overset{*}{\eta}_0(x) = x$ (for all $x \in D(F)$) and $\tau_0(z) \equiv 1$, the proposition (5.3.13a) is true for $n = 0$.

2. Suppose (5.3.13a) is true for some $n = n_0$ with $0 \le n_0 \le N$.

2.1. Since $\overset{*}{\zeta}{}_{1}^{n_0} = \overset{*}{\eta}_{n_0}$ and $\alpha_{1}^{n_0} = \tau_{n_0}$, the propositions (5.3.13b) and (5.3.13c) are true for $n = n_0$ and $\ell = 1$.

2.2. Suppose (5.3.13b) and (5.3.13c) are true for $n = n_0$ and all $\ell \le \ell_0-1$ where $2 \le \ell_0 \le m+1$. Set $t_{n_0,\ell_0-1} = t_{n_0} + \nu_{\ell_0-1}h_{n_0}$. Then positive constants $\bar{\sigma}_1$, δ_{n_0}, $\delta_{1,n_0,\ell-1}$ ($\ell = 2,3,\dots,\ell_0-1$) and δ_{2,n_0,ℓ_0-1} exist such that for $n = n_0$ we have (5.3.13a) is true, (5.3.13b) is true (for $\ell = 2,3,\dots,\ell_0-1$) and (5.3.13c) is true (for $\ell = \ell_0-1$), if $\sigma_{n_0} = \bar{\sigma}_1$ and $\sigma_{n_0,\ell} = \bar{\sigma}_1$ ($\ell = 1,2,\dots,\ell_0-1$).

Since $[\overset{*}{\zeta}{}_{\ell_0-1}^{n_0}]'(x^*)$ exists and $\overset{*}{\zeta}{}_{\ell_0-1}^{n_0}(x^*) = x^*$, a constant $\theta > 0$ and a constant $\bar{\sigma}_2 \in (0,\bar{\sigma}_1]$ exist such that $B(x^*,\bar{\sigma}_2) \subset D(\overset{*}{\zeta}{}_{\ell_0-1}^{n_0})$ and

$$\|\overset{*}{\zeta}{}_{\ell_0-1}^{n_0}(x) - x^*\| \le \theta\|x-x^*\| \qquad \text{(for all } x \in B(x^*,\bar{\sigma}_2)\text{)}.$$

From Lemma 5.3.6 and Lemma 5.3.7 it follows that positive constants $c_{1,i}$, $c_{2,i}$ ($i = 2,3$) and a constant $\bar{\sigma}_3 > 0$ exist such that (5.3.9) and (5.3.10) hold for $t = t_{n_0,\ell_0-1}$ with $\sigma = \bar{\sigma}_3$. Set $\bar{\sigma}_4 = \min\{\bar{\sigma}_2,\bar{\sigma}_3/(1+\theta)\}$. Then $B(x^*,\bar{\sigma}_4) \subset D(\overset{*}{\kappa}{}_{\ell_0-1}^{n_0})$. Let $x \in B(x^*,\bar{\sigma}_4)$. As a consequence of the chainrule and (5.3.10) it follows that $[\kappa^{*n_0}_{\ell_0-1}]'(x)$ exists and

$$[\kappa^{*n_0}_{\ell_0-1}]'(x) = \partial_2\Phi(t_{n_0,\ell_0-1},\zeta^{*n_0}_{\ell_0-1}(x),x)[\zeta^{*n_0}_{\ell_0-1}]'(x)$$

$$+ \partial_3\Phi(t_{n_0,\ell_0-1},\zeta^{*n_0}_{\ell_0-1}(x),x).$$

Hence, (cf. Lemma 5.3.3),

$$[\kappa^{*n_0}_{\ell_0-1}]'(x^*) = u_1(t_{n_0,\ell_0-1},C)\alpha^{n_0}_{\ell_0-1}(C) + u_2(t_{n_0,\ell_0-1},C)$$

$$= \pi^{n_0}_{\ell_0-1}(C).$$

Further

$$\|[\kappa^{*n_0}_{\ell_0-1}]'(x) - [\kappa^{*n_0}_{\ell_0-1}]'(x^*)\|$$

$$\le \|\partial_2\Phi(t_{n_0,\ell_0-1},\zeta^{*n_0}_{\ell_0-1}(x),x) - \partial_2\Phi(t_{n_0,\ell_0-1},x^*,x^*)\|\|[\zeta^{*n_0}_{\ell_0-1}]'(x)\|$$

$$+ \|\partial_2\Phi(t_{n_0,\ell_0-1},x^*,x^*)\|\|[\zeta^{*n_0}_{\ell_0-1}]'(x) - [\zeta^{*n_0}_{\ell_0-1}]'(x^*)\|$$

$$+ \|\partial_3\Phi(t_{n_0,\ell_0-1},\zeta^{*n_0}_{\ell_0-1}(x),x) - \partial_3\Phi(t_{n_0,\ell_0-1},x^*,x^*)\|$$

$$\le (c_{1,2} + c_{2,2}\theta)\|x-x^*\|\{\|\alpha^{n_0}_{\ell_0-1}(C)\| + \delta_{2,n_0,\ell_0-1}\bar{\sigma}_4\}$$

$$+ \|\partial_2\Phi(t_{n_0,\ell_0-1},x^*,x^*)\|\delta_{2,n_0,\ell_0-1}\|x-x^*\| + (c_{1,3}+c_{2,3}\theta)\|x-x^*\|.$$

Set

$$\delta_{1,n_0,\ell_0-1} = (c_{1,2}+c_{2,2}\theta)\{\|\alpha^{n_0}_{\ell_0-1}(C)\| + \delta_{2,n_0,\ell_0-1}\bar{\sigma}_4\} +$$

$$+ \|\partial_2\Phi(t_{n_0,\ell_0-1},x^*,x^*)\|\delta_{2,n_0,\ell_0-1} + c_{1,3} + c_{2,3}\theta.$$

Then (5.3.13b) holds for $n = n_0$ and $\ell = \ell_0$ when $\sigma_{n_0,\ell_0} = \bar{\sigma}_4$. Consequently $B(x^*,\bar{\sigma}_4) \subset D(\zeta^{*n_0}_{\ell_0})$ and for any $x \in B(x^*,\bar{\sigma}_4)$ the derivative $[\zeta^{*n_0}_{\ell_0}]'(x)$ exists and

$$[\zeta^{*n_0}_{\ell_0}]'(x) = \eta^{*\prime}_{n_0}(x) + h_n \sum_{j=1}^{\ell_0-1} \Lambda^{*(n_0)}_{\ell_0,j}[\kappa^{*n_0}_j]'(x).$$

Hence,

$$[\zeta^{*n_0}_{\ell_0}]'(x^*) = \tau_{n_0}(C) + h_{n_0}\sum_{j=1}^{\ell_0-1} \rho_{\ell_0,j}(h_{n_0}u_1(t_{n_0},C))\pi_j^{n_0}(C)$$
$$= \alpha_{\ell_0}^{n_0}(C).$$

Further

$$\|[\zeta^{*n_0}_{\ell_0}]'(x) - [\zeta^{*n_0}_{\ell_0}]'(x^*)\|$$
$$\le \{\delta_{n_0} + h_{n_0}\sum_{j=1}^{\ell_0-1} \|\Lambda^{*(n_0)}_{\ell_0,j}\|\delta_{1,n_0,j}\}\|x-x^*\|.$$

Hence (5.3.13c) is true for $n = n_0$ and $\ell = \ell_0$. Therefore (5.3.13b), (5.3.13c) are true for $n = n_0$ and all $\ell \le m+1$

2.3. Suppose $n_0 < N$. Since $\eta^*_{n_0+1} = \zeta^{*n_0}_{m+1}$ and $\tau_{n_0+1} = \alpha^{n_0}_{m+1}$, the proposition (5.3.13a) is true for $n = n_0+1$. This proves the lemma. □

Lemma 5.3.10 describes the similarity between G and G^* near x^*. For the proof of Lemma 5.3.10 we need the following lemma.

<u>LEMMA 5.3.9</u>. *Let* ρ *be a rational function with real coefficients. If* $C_0 \in D_E(\rho)$, *then positive numbers* ε *and* δ *exist such that all* $C \in L(E)$ *with* $\|C - C_0\| < \varepsilon$ *belong to* $D_E(\rho)$ *and satisfy the inequality*

$$\|\rho(C) - \rho(C_0)\| \le \delta\|C - C_0\|.$$

<u>PROOF</u>. Let $p(z) \equiv [q(z)]^{-1}p(z)$ whereby p and q are polynomials with real coefficients. Then $\bar{\varepsilon},\delta_1,\delta_2 > 0$ exist such that all $C \in L(E)$ with $\|C - C_0\| < \bar{\varepsilon}$ satisfy

$$\|q(C) - q(C_0)\| \le \delta_1\|C - C_0\|$$

and

$$\|p(C) - p(C_0)\| \le \delta_2\|C - C_0\|.$$

Further $[C_1]^{-1} - [C_2]^{-1} = [C_2]^{-1}[C_2 - C_1][C_1]^{-1}$ (for all $C_1, C_2 \in L(E)$ which are invertible). Together with Lemma 4.2.1 this yields the result. □

LEMMA 5.3.10. *Let Condition* 0 *be satisfied. Then* (5.3.11) *holds. Let* G^* *be defined in* (5.3.12). *Then for* $n = 0,1,\dots,N$ *the propositions* (5.3.14a), (5.3.14b) and (5.3.14c) *are true.*

Positive constants σ_n *and* δ_n *exist such that*

(5.3.14a)
$$B(x^*,\sigma_n) \subset D(\eta_n) \cap D(\eta_n^*),$$
$$\|\eta_n(x) - \eta_n^*(x)\| \le \delta_n\|x-x^*\|^2 \text{ whenever } x \in B(x^*,\sigma_n).$$

For $\ell = 1,2,\dots,m+1$ *positive constants* $\sigma_{n,\ell}$, $\delta_{2,n,\ell}$ *and, if* $\ell > 1$, $\delta_{1,n,\ell-1}$ *exist such that*

(5.3.14b)
$$B(x^*,\sigma_{n,\ell}) \subset D(\kappa_{\ell-1}^n) \cap D(\kappa_{\ell-1}^{*n}),$$
$$\|\kappa_{\ell-1}^n(x) - \kappa_{\ell-1}^{*n}(x)\| \le \delta_{1,n,\ell-1}\|x-x^*\|^2$$

(if $\ell > 1$*) and*

(5.3.14c)
$$B(x^*,\sigma_{n,\ell}) \subset D(\zeta_\ell^n) \cap D(\zeta_\ell^{*n}),$$
$$\|\zeta_\ell^n(x) - \zeta_\ell^{*n}(x)\| \le \delta_{2,n,\ell}\|x-x^*\|^2$$

whenever $x \in B(x^*,\sigma_{n,\ell})$. *In particular, since* $G = \zeta_{m+1}^N$ *and* $G^* = \zeta_{m+1}^{*N}$,

(5.3.14d)
$$B(x^*,\sigma_{N,m+1}) \subset D(G) \cap D(G^*) \text{ and}$$
$$\|G(x) - G^*(x)\| \le \delta_{2,N,m+1}\|x-x^*\|^2$$

whenever $x \in B(x^*,\sigma_{N,m+1})$.

PROOF. Since Condition 0 is satisfied, (5.3.11) holds.

1. Let $n = 0$. Since $D(F) = D(\eta_0) = D(\eta_0^*)$ and $\eta_0 = \eta_0^*$, the proposition (5.3.14a) is true for $n = 0$.

2. Suppose (5.3.14a) is true for some $n = n_0$ with $0 \le n_0 \le N$.

2.1. Since $\zeta_1^{n_0} = \eta_{n_0}$ and $\zeta_1^{*n_0} = \eta_{n_0}^*$, the propositions (5.3.14b), (5.3.14c) are true for $n = n_0$ and $\ell = 1$.

2.2. Suppose (5.3.14b), (5.3.14c) are true for $n = n_0$ and all $\ell \le \ell_0 - 1$

where $2 \le \ell_0 \le m+1$. Set $t_{n_0,\ell_0-1} = t_{n_0} + \nu_{\ell_0-1}h_{n_0}$. Then positive constants $\bar{\sigma}_1$, δ_{n_0}, δ_{2,n_0,ℓ_0-1} and $\delta_{1,n_0,\ell-1}$ ($\ell = 2,3,\dots,\ell_0-1$) exist such that for $n = n_0$ we have (5.3.14a) is true, (5.3.14b) is true (for $\ell = 2,3,\dots,\ell_0-1$) and (5.3.14c) is true (for $\ell = \ell_0-1$), if $\sigma_{n_0} = \bar{\sigma}_1$ and $\sigma_{n_0,\ell} = \bar{\sigma}_1$ ($\ell = 1,2,\dots,\ell_0-1$).

From Lemma 5.3.8 it follows that constants $\bar{\sigma}_2 \in (0,\bar{\sigma}_1]$ and $\bar{\delta}_1 > 0$ exist such that $B(x^*,\bar{\sigma}_2) \subset D(\eta^*_{n_0}) \cap D(\zeta^{*n_0}_{\ell_0-1})$,

$$(5.3.15a) \qquad \|\eta^*_{n_0}(x) - x^*\| \le \bar{\delta}_1\|x-x^*\|$$

and

$$(5.3.15b) \qquad \|\zeta^{*n_0}_{\ell_0-1}(x) - x^*\| \le \bar{\delta}_1\|x-x^*\| \qquad (\text{for all } x \in B(x^*,\bar{\sigma}_2)).$$

Moreover, for $\ell = 2,3,\dots,\ell_0$, $B(x^*,\bar{\sigma}_2) \subset D(\kappa^{*n_0}_{\ell-1})$ and

$$(5.3.15c) \qquad \|\kappa^{*n_0}_{\ell-1}(x)\| \le \bar{\delta}_1\|x-x^*\| \qquad (\text{for all } x \in B(x^*,\bar{\sigma}_2)).$$

From (5.3.14c) ($n = n_0$, $\ell = \ell_0 - 1$) and (5.3.15b) it follows that

$$(5.3.15d) \qquad B(x^*,\bar{\sigma}_2) \subset D(\zeta^{n_0}_{\ell_0-1}) \text{ and } \|\zeta^{n_0}_{\ell_0-1}(x) - x^*\| \le \bar{\delta}_2\|x-x^*\|$$
$$(\text{for all } x \in B(x^*,\bar{\sigma}_2))$$

where $\bar{\delta}_2 = \bar{\delta}_1 + \delta_{2,n_0,\ell_0-1}\bar{\sigma}_2$. Since $(t,x^*,x^*) \in D(\Phi)$ (for all $t \in \{t_{n_0}, t_{n_0,\ell_0-1}\}$), because of Lemma 5.3.6 and the relations (5.3.15d), (5.3.15c) ($\ell = \ell_0$), there exist positive constants c_1, c_2 and a constant $\bar{\sigma}_3 \in (0,\bar{\sigma}_2]$ such that $B(x^*,\bar{\sigma}_3) \subset D(\kappa^{n_0}_{\ell_0-1}) \cap D(\kappa^{*n_0}_{\ell_0-1})$ and

$$(5.3.16) \qquad \|\partial_2\Phi(t,y,x) - \partial_2\Phi(t,x^*,x^*)\| \le c_1\|x-x^*\| + c_2\|y-x^*\|$$
$$(\text{for all } t \in \{t_{n_0},t_{n_0,\ell_0-1}\} \text{ and all } x,y \in B(x^*,\bar{\sigma}_3)).$$

Set $\bar{\sigma}_4 = [1+\bar{\delta}_2]^{-1}\bar{\sigma}_3$. Then, from Lemma 2.6.1 it follows that for all $x \in B(x^*,\bar{\sigma}_4)$

$$\|\kappa_{\ell_0-1}^{n_0}(x) - \kappa_{\ell_0-1}^{*n_0}(x)\|$$

$$= \|\Phi(t_{n_0,\ell_0-1},\zeta_{\ell_0-1}^{n_0}(x),x) - \Phi(t_{n_0,\ell_0-1},\zeta_{\ell_0-1}^{*n_0}(x),x)\|$$

$$\leq \{\|\partial_2\Phi(t_{n_0,\ell_0-1},x^*,x^*)\| + c_1 + c_2\bar{\delta}_2\bar{\sigma}_4\}\delta_{2,n_0,\ell_0-1}\|x-x^*\|^2.$$

Therefore, a constant δ_{1,n_0,ℓ_0-1} exists such that (5.3.14b) is true for $n = n_0$ and $\ell = \ell_0$ with $\sigma_{n_0,\ell_0} \in (0,\bar{\sigma}_4]$.

Observe that

$$h_{n_0}\partial_2\Phi(t_{n_0},x^*,x^*) \in D_E(\rho_{\ell_0,j}) \qquad (j = 1,2,\ldots,\ell_0-1).$$

Hence, from Lemma 5.3.9 and relation (5.3.16) it follows that constants $\bar{\sigma}_5 \in (0,\bar{\sigma}_4]$ and $\bar{\delta}_3 > 0$ exist such that for all $x,y \in B(x^*,\bar{\sigma}_5)$

$$h_{n_0}\partial_2\Phi(t_{n_0},y,x) \in D_E(\rho_{\ell_0,j})$$

and

$$\|\rho_{\ell_0,j}(h_{n_0}\partial_2\Phi(t_{n_0},y,x)) - \rho_{\ell_0,j}(h_{n_0}\partial_2\Phi(t_{n_0},x^*,x^*))\|$$

$$\leq \bar{\delta}_3\{c_1\|x-x^*\| + c_2\|y-x^*\|\} \qquad (j = 1,2,\ldots,\ell_0-1).$$

Let $\bar{\delta}_4 = \bar{\delta}_1 + \delta_{n_0}\bar{\sigma}_5$. Then $B(x^*,\bar{\sigma}_5) \subset D(\eta_{n_0})$ and $\|\eta_{n_0}(x) - x^*\| \leq \bar{\delta}_4\|x-x^*\|$ (for all $x \in B(x^*,\bar{\sigma}_5)$). Set $\bar{\sigma}_6 = [1+\bar{\delta}_4]^{-1}\bar{\sigma}_5$. Then $B(x^*,\bar{\sigma}_6) \subset D(\zeta_{\ell_0}^{n_0}) \cap D(\zeta_{\ell_0}^{*n_0})$ and for all $x \in B(x^*,\bar{\sigma}_6)$

$$\|\zeta_{\ell_0}^{n_0}(x) - \zeta_{\ell_0}^{*n_0}(x)\|$$

$$= \|\eta_{n_0}(x) - \eta_{n_0}^*(x) + h_{n_0}\sum_{j=1}^{\ell_0-1}\{\Lambda_{\ell_0,j}^{(n_0)}\kappa_j^{n_0}(x) - \Lambda_{\ell_0,j}^{*(n_0)}\kappa_j^{*n_0}(x)\}\|$$

$$\leq \delta_{n_0}\|x-x^*\|^2 + h_{n_0}\sum_{j=1}^{\ell_0-1}\{\|\Lambda_{\ell_0,j}^{(n_0)}\|\|\kappa_j^{n_0}(x) - \kappa_j^{*n_0}(x)\| +$$

$$+\|\Lambda_{\ell_0,j}^{(n_0)} - \Lambda_{\ell_0,j}^{*(n_0)}\|\|\kappa_j^{*n_0}(x)\|\} \leq$$

$$\leq \{\delta_{n_0} + h_{n_0} \sum_{j=1}^{\ell_0-1} \{[\| \Lambda^{*(n_0)}_{\ell_0,j} \| + \bar{\delta}_3(c_1 + c_2\bar{\delta}_4)\bar{\sigma}_6]\delta_{1,n_0,j}$$

$$+ \bar{\delta}_3(c_1 + c_2\bar{\delta}_4)\bar{\delta}_1\}\}\| x-x^*\|^2.$$

Hence (5.3.14b), (5.3.14c) hold for $n = n_0$ and all $\ell \leq m+1$.
2.3. If $n_0 < N$ then, since $\eta_{n_0+1} = \zeta^{n_0}_{m+1}$ and $\eta^*_{n_0+1} = \zeta^{*n_0}_{m+1}$, (5.3.14a) holds for $n = n_0 + 1$ as well. This proves the lemma. □

From Lemma 5.3.10 it follows that $x^* \in \text{interior}(D(G))$ whenever Condition 0 is satisfied. The next lemma shows that the reverse is also true.

LEMMA 5.3.11. *Condition* 0 *is satisfied if and only if* $x^* \in \text{interior}(D(G))$.

PROOF. 1. Suppose $x^* \in \text{interior}(D(G))$. Obviously, $(t,x^*,x^*) \in D(\Phi)$ for all $t = t_n + \nu_{\ell-1}h_n$ $(n = 0,1,\ldots,N;\ \ell = 2,3,\ldots,m+1)$. In particular $\partial_1 K(x^*,x^*)$ is invertible. Using Lemma 5.3.3, it is easily verified that Condition 0 is satisfied.
2. If Condition 0 is satisfied, from Lemma 5.3.10 it follows that $x^* \in \text{interior}(D(G))$. □

The following important lemma is a consequence of Lemma 5.3.8 and Lemma 5.3.10.

LEMMA 5.3.12. *If Condition* 0 *is satisfied, then* $x^* \in \text{interior}(D(G))$, *the derivative* $G'(x^*)$ *exists and* $G'(x^*) = \gamma(C)$.

PROOF. From Lemma 5.3.8 it follows that $x^* \in \text{interior}(D(G^*))$, the derivative $[G^*]'(x^*)$ exists and $[G^*]'(x^*) = \gamma(C)$. From Lemma 5.3.10 it follows that $x^* \in \text{interior}(D(G))$. From (5.3.14d) it follows that $G'(x^*)$ exists and $G'(x^*) = [G^*]'(x^*)$. □

The next three lemmata will be useful in the proofs of the Theorems 5.2.1 and 5.2.2.

LEMMA 5.3.13. *Suppose* $-h_n g(t_n) \in D_{\mathbb{C}}(\rho_{\ell,j})$ *and* $\rho_{\ell,j}(-h_n g(t_n)) = \rho_{\ell,j}(0)$ $(n = 0,1,\ldots,N;\ \ell = 2,3,\ldots,m+1;\ j = 1,2,\ldots,\ell-1)$. *Then*

(5.3.17a) $\quad 1 \in D_{\mathbb{C}}(\pi^n_{\ell-1}), \quad \pi^n_{\ell-1}(1) = -1$

(if $\ell > 1$) and

(5.3.17b) $\quad 1 \in D_{\mathbb{C}}(\alpha_{\ell}^{n}), \quad \alpha_{\ell}^{n}(1) = 1 - t_n - \nu_{\ell} h_n$

($n = 0,1,\dots,N$; $\ell = 1,2,\dots,m+1$). *In particular we have*

(5.3.17c) $\quad 1 \in D_{\mathbb{C}}(\gamma)$ *and* $\gamma(1) = 0$.

PROOF. Notice that $u_1(t,1) \equiv -g(t)$ and $u_2(t,1) \equiv -(1 - g(t)(1-t))$.

1. $\alpha_1^0 = \tau_0$ and $\tau_0(z) \equiv 1$ hold. Hence (5.3.17a), (5.3.17b) are true for $n = 0$ and $\ell = 1$.

2. Suppose (5.3.17a), (5.3.17b) hold for $n = n_0 - 1$ with $0 < n_0 \le N$ and $\ell = 1,2,\dots,m+1$.

2.1. Since $\alpha_1^{n_0} = \tau_{n_0} = \alpha_{m+1}^{n_0-1}$, we have $1 \in D_{\mathbb{C}}(\alpha_1^{n_0})$ and $\alpha_1^{n_0}(1) = 1 - t_{n_0-1} - \nu_{m+1} h_{n_0-1} = 1 - t_{n_0}$. Thus (5.3.17a), (5.3.17b) hold for $n = n_0$ and $\ell = 1$.

2.2. Suppose (5.3.17a), (5.3.17b) are true for $n = n_0$ and all $\ell \le \ell_0 - 1$ where $2 \le \ell_0 \le m+1$. Set $t_{n_0,\ell_0-1} = t_{n_0} + \nu_{\ell_0-1} h_{n_0}$. Then $1 \in D_{\mathbb{C}}(\pi_{\ell_0-1}^{n_0})$ and

$$\pi_{\ell_0-1}^{n_0}(1) = -g(t_{n_0,\ell_0-1})(1 - t_{n_0,\ell_0-1})$$

$$- \{1 - g(t_{n_0,\ell_0-1})(1 - t_{n_0,\ell_0-1})\} = -1.$$

Further

$$1 \in D_{\mathbb{C}}(\alpha_{\ell_0}^{n_0})$$

and

$$\alpha_{\ell_0}^{n_0}(1) = \alpha_1^{n_0}(1) + h_{n_0} \sum_{j=1}^{\ell_0-1} \rho_{\ell_0,j}(0)\pi_j^{n_0}(1) = 1 - t_{n_0} - \nu_{\ell_0} h_{n_0}.$$

This proves the lemma. □

The next two lemmata are direct consequences of Lemma 5.3.13.

LEMMA 5.3.14. *Let the assumptions of Lemma* 5.3.13 *hold. Suppose* $\mathrm{sp}([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) = \{1\}$. *Then Condition* 0 *is satisfied and* $\mathrm{sp}(\gamma(C)) = \{0\}$.

PROOF. Set $T = [F'(x^*)]^{-1}\partial_1 K(x^*,x^*)$. According to Theorem 2.2.2 the operator T is invertible. It is easily verified that $\partial_1 K(x^*,x^*)$ is invertible and $[\partial_1 K(x^*,x^*)]^{-1} = [T]^{-1}[F'(x^*)]^{-1}$. Set $C = [\partial_1 K(x^*,x^*)]^{-1}F'(x^*)$. Since $C = [T]^{-1}$, from Theorem 2.2.2 it follows that $sp(C) = \{1\}$. Consequently, from Lemma 5.3.13 and Theorem 2.2.2 it follows that the statement of the lemma is true. □

LEMMA 5.3.15. *Let the assumptions of Lemma* 5.3.13 *hold. Suppose* $\partial_1 K(x^*,x^*) = F'(x^*)$. *Then Condition* 0 *is satisfied and* $\gamma(C) = 0$.

PROOF, According to Lemma 5.3.14 Condition 0 is satisfied. In this case $C = I$. It is easily verified that $\gamma(I) = 0$. □

5.3.2 Proof of Theorem 5.1.1

a. The propositions (i), (ii), (iii) and (iv) are immediate consequences of Lemma 5.3.11, Lemma 5.3.12 and Theorem 2.5.1.

b. Suppose that Condition 5 is satisfied. Due to Lemma 5.3.8 and Lemma 5.3.10 positive constants σ, δ_1 and δ_2 exist such that $B(x^*,\sigma) \subset D(G) \cap D(G^*)$ and

$$\|[G^*]'(x)\| \le \delta_1 \|x-x^*\|,$$

$$\|G(x) - G^*(x)\| \le \delta_2 \|x-x^*\|^2$$

whenever $x \in B(x^*,\sigma)$. Since $G^*(x^*) = x^*$, from Lemma 5.3.1 it follows that

$$\|G^*(x) - x^*\| \le \frac{\delta_1}{2} \|x-x^*\|^2 \qquad \text{(for all } x \in B(x^*,\sigma)).$$

Hence

$$\|G(x) - x^*\| \le (\delta_2 + \frac{\delta_1}{2}) \|x-x^*\|^2 \qquad \text{(for all } x \in B(x^*,\sigma)).$$

Thus [M,F] is quadratically convergent.

c. Suppose [M,F] is quadratically convergent. From Lemma 5.3.11 it follows that Condition 0 is satisfied. From Lemma 5.3.12 and Theorem 2.5.4 it follows that Condition 5 is satisfied. This completes the proof. □

5.3.3. Proof of the Theorems 5.2.1 and 5.2.2

Suppose that (5.2.1) holds.

1. PROOF OF THEOREM 5.2.1. Suppose $\mathrm{sp}([F'(x^*)]^{-1}\partial_1K(x^*,x^*)) = \{1\}$. Due to Lemma 5.3.14 Condition 1 holds. Theorem 5.1.1(i) yields the result. □

2. PROOF OF THEOREM 5.2.2. Suppose $F'(x^*) = \partial_1K(x^*,x^*)$. Due to Lemma 5.3.15 Condition 5 holds. Theorem 5.1.1(v) yields the result. □

5.3.4. Proof of the Theorems 5.2.3 - 5

1. PROOF OF THEOREM 5.2.3. a. *We show that* (i) *implies* (ii). To that end we need the following two lemmata. For $\theta \in (0,1]$ and $\delta \in \mathbb{R}$, define

$$(5.3.18) \qquad \hat{L}(\theta,\delta) = \begin{pmatrix} 0 & 0 & 0 \\ \theta & 0 & 0 \\ \delta & 1-\delta & 0 \end{pmatrix} .$$

LEMMA 5.3.16. *Let* $\theta \in (0,1]$ *and* $\delta \in \mathbb{R}$. *Let* $L = \hat{L}(\theta,\delta)$ *and* $H = \{1\}$. *Set* $G = M(F)$ *where* $M = \mathbb{M}(A,g,L,H)$. *Let the rational function* γ *be defined in* (5.1.3) *where* $R = L$. *Suppose* $x^* \in \mathrm{interior}(D(G))$. *Then Condition* 0 *holds,* $\partial_1K(x^*,x^*)$ *is invertible and* $[(1-\theta)I + \theta C]$ *is invertible where*

$$(5.3.19) \qquad C = [\partial_1K(x^*,x^*)]^{-1}F'(x^*).$$

Further $C \in D_E(\gamma)$, *the derivative* $G'(x^*)$ *exists,* $G'(x^*) = \gamma(C)$ *and* $\gamma(C) = C_1 + \delta C_2$ *where*

$$(5.3.20) \qquad C_1 = I + \{[(1-\theta)I + \theta C]^{-1}[I-C] - g(\theta)I\}[I-\theta C] - \{1 - g(\theta)(1-\theta)\}[(1-\theta)I + \theta C]^{-1}$$

and

$$(5.3.21) \qquad C_2 = \theta\{2 + \theta g(\theta)\}[(1-\theta)I + \theta C]^{-1}C[I-C].$$

PROOF. According to Lemma 5.3.11 Condition 0 is satisfied (where $R = L$). Hence $\partial_1 K(x^*,x^*)$ is invertible, $[(1-\theta)I + \theta C]$ is invertible and $C \in D_E(\gamma)$. From Lemma 5.3.12 it follows that $G'(x^*)$ exists and $G'(x^*) = \gamma(C)$. A little calculation shows that

$$(5.3.22) \qquad \gamma(C) = I + \delta\{-C\} + (1-\delta)\{\{[(1-\theta)I + \theta C]^{-1}[I-C] - g(\theta)I\}[I-\theta C]$$

$$- [1 - g(\theta)(1-\theta)][(1-\theta)I + \theta C]^{-1}\}.$$

Resolving relation (5.3.22) one obtains

$$\gamma(C) = C_1 + \delta C_2.$$

This proves the lemma. □

LEMMA 5.3.17. *Let* $\theta \in (0,1]$ *and let* C_1 *and* C_2 *be defined in* (5.3.20) *and* (5.3.21) *respectively. Then for any* $\delta \in \mathbb{R}$

$$sr(C_1 + \delta C_2) \geq |\delta| sr(C_2) - sr(C_1).$$

PROOF. Let $\delta \in \mathbb{R}$. It is easily verified that C_1 and C_2 commute. Hence $C_1 + \delta C_2$ and $-C_1$ commute. Therefore (cf. [RUDIN, 1973; Theorem 11.23])

$$sr(C_1 + \delta C_2 - C_1) \leq sr(C_1 + \delta C_2) + sr(-C_1).$$

Consequently,

$$sr(C_1 + \delta C_2) \geq sr(\delta C_2) - sr(-C_1) = |\delta| sr(C_2) - sr(C_1). \qquad □$$

Suppose (i) holds and $sp([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) \neq \{1\}$. There follows that $(0,x^*,x^*) \in D(\Phi)$ so that $\partial_1 K(x^*,x^*)$ is invertible. Choose $\theta \in (0,1]$ such that $2 + \theta g(\theta) \neq 0$. (This is possible since $g \in S$, cf. (2.6.9).) Let $\delta \in \mathbb{R}$. Since (i) holds, Lemma 5.3.16 applies. Let C_1 and C_2 be defined in (5.3.20) and (5.3.21), respectively. From Theorem 2.2.2 it follows that $sp(C_2) \neq \{0\}$. Choose $\delta_1 \in \mathbb{R}$ such that $|\delta_1| sr(C_2) - sr(C_1) > 1$. Let $L = \hat{L}(\theta,\delta_1)$ and $H = \{1\}$. Let the rational function γ be defined in (5.1.3) where $R = L$. From Lemma 5.3.16 and Lemma 5.3.17 it follows that

$$\mathrm{sr}(\gamma(C)) \geq |\delta_1|\mathrm{sr}(C_2) - \mathrm{sr}(C_1) > 1.$$

Using Theorem 5.1.1(ii) this yields a contradiction.

b. From Theorem 5.2.1 it follows that (ii) implies (i). □

2. PROOF OF THE THEOREMS 5.2.4, 5.2.5. We shall show that (vi) implies (v), and that (iii) implies (iv). This is of course sufficient to prove both Theorem 5.2.4 and Theorem 5.2.5.

a. From Theorem 5.2.2 it follows that (vi) implies (v).

b. *We show that* (iii) *implies* (iv). Assume (iii) holds and suppose $\partial_1 K(x^*,x^*) \neq F'(x^*)$. Choose $\theta \in (0,1]$ such that $2+\theta g(\theta) \neq 0$. Let $\delta \in \mathbb{R}$. Lemma 5.3.16 applies. Let C, C_1 and C_2 be defined in (5.3.19), (5.3.20) and (5.3.21), respectively. Since $I-C \neq 0$ and C is invertible, we have $\|C_2\| \neq 0$. Choose $\delta_1 \in \mathbb{R}$ such that $|\delta_1|\|C_2 - C_1\| > 1$. Let $L = \hat{L}(\theta,\delta_1)$ (cf. (5.3.18)) and $H = \{1\}$. Let the rational function γ be defined in (5.1.3) where R = L. From Lemma 5.3.16 it follows that $C \in D_E(\gamma)$ and

$$\|\gamma(C)\| = \|\delta_1 C_2 + C_1\| \geq |\delta_1|\|C_2\| - \|C_1\| > 1.$$

Using Theorem 5.1.1(iv) this yields a contradiction. □

5.3.5. Proof of the Theorems 5.2.6 - 8

1. PROOF OF THEOREM 5.2.6. a. *We show that* (i) *implies* (ii). We shall use the following lemma.

LEMMA 5.3.18. *Suppose* $\mathrm{sp}([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) = \{1\}$ *and* $g(\theta) \neq 0$ *for some* $\theta \in [0,1)$. *Then a Runge-Kutta method with operator coefficients* R *and a sequence of stepsizes* H *exist such that* $x^* \notin D(G)$ *where* $G = M(F)$ *and* $M = \mathbb{M}(A,g,R,H)$.

PROOF. Since $\mathrm{sp}([F'(x^*)]^{-1}\partial_1 K(x^*,x^*)) = \{1\}$, the operator $\partial_1 K(x^*,x^*)$ is invertible. Set $C = [\partial_1 K(x^*,x^*)]^{-1}F'(x^*)$. Since $\mathrm{sp}(C) = \{1\}$ it follows that the operator $[(1-\theta)I + \theta C]$ is invertible. From Lemma 5.3.3 it follows that $\partial_2\Phi(\theta,x^*,x^*)$ exists and

$$\partial_2\Phi(\theta,x^*,x^*) = -g(\theta) + [(1-\theta)I + \theta C]^{-1}[I-C].$$

From Theorem 2.2.2 it then follows that

$$sp((1-\theta)\partial_2\Phi(\theta,x^*,x^*)) = \{-(1-\theta)g(\theta)\} \neq \{0\}.$$

Let $R = (\rho_{i,j})$ be a one-stage Runge-Kutta method with operator coefficient where $\rho_{2,1}(z) \equiv [1 + \frac{1}{g(\theta)(1-\theta)} z]^{-1}$. Let $H = \{h_0,\ldots,h_N\}$, where $H = \{1\}$ if $\theta = 0$ and $H = \{\theta,(1-\theta)\}$ if $\theta \neq 0$. Let $G = M(F)$ where $M = \mathbb{M}(A,g,R,H)$. Then $h_N\partial_2\Phi(t_N,x^*,x^*) \notin D_E(\rho_{2,1})$. Hence $x^* \notin D(G)$ (cf. (4.2.5e)). □

Since any Runge-Kutta method with scalar coefficients is a Runge-Kutta method with operator coefficients, it follows from Theorem 5.2.3 and Lemma 5.3.18 that (i) implies (ii).

b. From Theorem 5.2.1 it follows that (ii) implies (i).

2. PROOF OF THE THEOREMS 5.2.7, 5.2.8. We shall show that (vi) implies (v), and that (iii) implies (iv).

a. From Theorem 5.2.2 it follows that (vi) implies (v).

b. Suppose (iii) holds. From Theorem 5.2.4 it follows that $\partial_1K(x^*,x^*) = F'(x^*)$. From Theorem 5.2.6 it follows that $g(t) = 0$ (for all $t \in [0,1)$). This completes the proof. □

CHAPTER 6

RADIUS OF CONVERGENCE

In this chapter we determine the radii of convergence of some of the iterative methods described in section 4.1.

In section 6.1 we introduce $\mathcal{F}\langle\sigma,\beta,\gamma\rangle$, the subclass of $\mathcal{F}_1$ (see (2.6.1)) to which the radii of convergence to be considered in Part I of this chapter will be related. In section 6.2 we determine the radius of convergence of Newton's method. In section 6.3 a class of iterative methods, which are denoted by $M_{\bar{\omega}}$, is introduced. These have greater radii of convergence than Newton's method. Any method $M_{\bar{\omega}}$, which is a kind of damped Newton method, is of a type considered in subsection 5.2.2. We also give a result which shows that certain damped Newton methods have greater radii of convergence than Newton's method.

Part II of this chapter is concerned with $\mathcal{F}\langle\sigma,\alpha\rangle$, the subclass of $\mathcal{F}_1$ that is introduced in section 6.4. In section 6.5 we determine the radius of convergence of Newton's method with respect to $\mathcal{F}\langle\sigma,\alpha\rangle$. As in Part I of this chapter, we investigate in section 6.6 the convergence behaviour of the iterative methods $M_{\bar{\omega}}$ (introduced in section 6.3) with respect to $\mathcal{F}\langle\sigma,\alpha\rangle$. In this case, we are able to give an explicit expression for the radii of convergence.

We finally note that we base the determination of the radius of convergence of an iterative method M with respect to a subclass $\mathcal{F}_0$ of $\mathcal{F}_1$ on the following principle. We first give a lower bound, say $\hat{r}$, of $r(M;\mathcal{F}_0)$. Then we consider the case $E = \mathbb{R}$ and construct an $f \in \mathcal{F}_0$ for which $r(M,f) = \hat{r}$. When E is an arbitrary Hilbert space we "extend" f to an operator F on E such that $F \in \mathcal{F}_0$ and $r(M,F) = \hat{r}$. The extension of f to F is described in subsection 6.2.3.

PART I

6.1. THE CLASS $F\langle\sigma,\beta,\gamma\rangle$

In the first part of this chapter we shall be concerned with the following subclass of F_1 (see 2.6.1)).

For given $\sigma \in (0,\infty]$ and $\beta,\gamma > 0$

$$(6.1.1) \qquad F\langle\sigma,\beta,\gamma\rangle = \{F \mid F \in F_1;\ B(x^*,\sigma) \subset D(F);\ \|[F'(x^*)]^{-1}\| \le \beta;$$
$$\|F''(x)\| \le \gamma \quad (\text{for all } x \in B(x^*,\sigma))\}.$$

We notice that for any $F \in F_1$, numbers $\sigma \in (0,\infty]$ and $\beta,\gamma > 0$ exist such that $F \in F\langle\sigma,\beta,\gamma\rangle$. Hence

$$(6.1.2) \qquad F_1 = \bigcup_{\substack{\sigma\in(0,\infty] \\ \beta,\gamma>0}} F\langle\sigma,\beta,\gamma\rangle.$$

Let $\beta > 0$ and let $F \in L(E)$ with $F = \frac{1}{\beta} I$ (I is the identity). Obviously, $F \in F\langle\sigma,\beta,\gamma\rangle$ (for all $\sigma \in (0,\infty]$, $\gamma > 0$). Consequently, for each $\sigma \in (0,\infty]$ and all $\beta,\gamma > 0$, the set $F\langle\sigma,\beta,\gamma\rangle$ is not empty.

6.2. THE RADIUS OF CONVERGENCE OF NEWTON'S METHOD WITH RESPECT TO $F\langle\sigma,\beta,\gamma\rangle$

Let $\sigma \in (0,\infty]$ and $\beta,\gamma > 0$. Let M be Newton's method, which means that for $F \in F_1$, the function $G = M(F)$ is defined by

$$G: D(G) \to E,$$
$$(6.2.1) \qquad D(G) = \{x \mid x \in D(F);\ F'(x) \text{ is invertible}\},$$
$$G(x) = x - \Gamma(x)F(x) \qquad (x \in D(G)).$$

We recall that $\Gamma(x)$ denotes $[F'(x)]^{-1}$ (for $x \in D(G)$) .

<u>THEOREM 6.2.1</u>. *Let* M *be the Newton's method, then*

$$(6.2.2) \qquad r(M;F\langle\sigma,\beta,\gamma\rangle) = \min\{\sigma, \frac{2}{3\beta\gamma}\}.$$

We shall prove Theorem 6.2.1 in the next subsections.

6.2.1 Proof of Theorem 6.2.1

In order to prove Theorem 6.2.1 we need the following two lemmata.

LEMMA 6.2.2. *If* $F \in F\langle\sigma,\beta,\gamma\rangle$ *and* $x \in B(x^*,\sigma) \cap B(x^*, \frac{1}{\beta\gamma})$ *then* $F'(x)$ *is invertible and*

$$\|[F'(x)]^{-1}\| \leq \frac{\beta}{1-\beta\gamma\|x-x^*\|} .$$

PROOF. Since $F \in F\langle\sigma,\beta,\gamma\rangle$, it follows that $\|[F'(x^*)]^{-1}\| \leq \beta$ and $\|F'(x) - F'(x^*)\| \leq \gamma\|x-x^*\|$ (for all $x \in B(x^*,\sigma) \cap B(x^*, \frac{1}{\beta\gamma})$), (cf. Lemma 2.6.1). Therefore, Lemma 4.2.1 applies, thus proving this lemma. □

LEMMA 6.2.3. *For* $F \in F\langle\sigma,\beta,\gamma\rangle$ *let* G *be defined by* (6.2.1). *Then*

$$B(x^*,\sigma) \cap B(x^*, \frac{1}{\beta\gamma}) \subset D(G)$$

and

$$\|G(x) - x^*\| \leq \frac{\beta\gamma\|x-x^*\|^2}{2(1-\beta\gamma\|x-x^*\|)}$$

whenever $x \in B(x^*,\sigma) \cap B(x^*, \frac{1}{\beta\gamma})$.

PROOF. The first part of the conclusion is a consequence of Lemma 6.2.2. Let $x \in B(x^*,\sigma) \cap B(x^*, \frac{1}{\beta\gamma})$. From Lemma 5.3.1 it follows that $0 = F(x^*) = F(x) + F'(x)(x^*-x) + r(x)$, where $\|r(x)\| \leq \frac{\gamma}{2}\|x-x^*\|^2$. Thus

$$\|x - \Gamma(x)F(x) - x^*\| = \|\Gamma(x)r(x)\| .$$

Using Lemma 6.2.2, we obtain

$$\|\Gamma(x)r(x)\| \leq \frac{\beta\gamma\|x-x^*\|^2}{2(1-\beta\gamma\|x-x^*\|)} .$$

This completes the proof. □

We are now able to prove the following lemma.

LEMMA 6.2.4. *Let* M *be the Newton's method, then*

$$r(M;\mathcal{F}\langle\sigma,\beta,\gamma\rangle) \geq \min\{\sigma, \frac{2}{3\beta\gamma}\}.$$

PROOF. Let $F \in \mathcal{F}\langle\sigma,\beta,\gamma\rangle$, and $G = M(F)$. Thus G is defined by (6.2.1). For $\varepsilon \in (0,\frac{2}{3})$, set $\alpha(\varepsilon) = [\frac{2}{3}+2\varepsilon]^{-1}[\frac{2}{3}-\varepsilon]$. Note that $0 < \alpha(\varepsilon) < 1$. From Lemma 6.2.3 it follows that for any $x \in B(x^*,\sigma) \cap \bar{B}(x^*,(\frac{2}{3}-\varepsilon)\frac{1}{\beta\gamma})$ we have $x \in D(G)$ and

$$(6.2.3) \qquad \|G(x) - x^*\| \leq \frac{\beta\gamma\|x-x^*\|^2}{2(1-\beta\gamma\|x-x^*\|)} \leq \frac{\frac{2}{3}-\varepsilon}{2(1-(\frac{2}{3}-\varepsilon))}\|x-x^*\| = \alpha(\varepsilon)\|x-x^*\|.$$

If $x_0 \in B(x^*,\sigma) \cap B(x^*, \frac{2}{3\beta\gamma})$ then an $\varepsilon \in (0,\frac{2}{3})$ exists such that $x_0 \in B(x^*,\sigma) \cap \bar{B}(x^*,(\frac{2}{3}-\varepsilon)\frac{1}{\beta\gamma})$. The relation (6.2.3) shows that $x_0 \in D(M,F)$ (cf. (2.3.3)) and that the sequence $\{x_k\}$ generated by x_0 and $[M,F]$ satisfies

$$\|x_k - x^*\| \leq [\alpha(\varepsilon)]^k\|x_0 - x^*\| \to 0 \qquad (k \to \infty).$$

Hence $x_0 \in S(M,F)$ (cf. (2.3.5)), so that $r(M,F) \geq \min\{\sigma, \frac{2}{3\beta\gamma}\}$. Since F was an arbitrary element of $\mathcal{F}\langle\sigma,\beta,\gamma\rangle$ this proves the lemma. □

Results similar to Lemmata 6.2.3 and 6.2.4 can also be found in [RHEINBOLDT, 1975].

REMARK 6.2.1. Notice that, if $F \in \mathcal{F}\langle\sigma,\beta,\gamma\rangle$, then at the same time, $F_\sigma \in \mathcal{F}\langle\sigma,\beta,\gamma\rangle$ where F_σ is the restriction of F to $B(x^*,\sigma)$, $(D(F_\sigma) = B(x^*,\sigma))$. Obviously $r(M,F_\sigma) \leq \sigma$, so that $r(M;\mathcal{F}\langle\sigma,\beta,\gamma\rangle) \leq \sigma$. □

REMARK 6.2.2. Suppose $E = \mathbb{R}$ and let $\sigma > \frac{2}{3\beta\gamma}$. In view of Lemma 6.2.4 and Remark 6.2.1 the proof of Theorem 6.2.1 is completed if we can show that an $f \in \mathcal{F}\langle\sigma,\beta,\gamma\rangle$ and an $x_0 \in D(f)$ with $\|x_0 - x^*\| = \frac{2}{3\beta\gamma}$ exist such that $x_0 \notin S(M,f)$. Consider

$$f_0\colon \mathbb{R} \to \mathbb{R},$$

$$f_0(\xi) = \begin{cases} \dfrac{1}{2\beta^2\gamma} & (\text{if } \xi \geq \frac{1}{\beta\gamma}), \\[2ex] \dfrac{1}{\beta}\xi - \dfrac{\gamma}{2}\xi^2 & (\text{if } \xi \in [0, \frac{1}{\beta\gamma})), \\[2ex] \dfrac{1}{\beta}\xi + \dfrac{\gamma}{2}\xi^2 & (\text{if } \xi \in (-\frac{1}{\beta\gamma}, 0)), \\[2ex] -\dfrac{1}{2\beta^2\gamma} & (\text{if } \xi \leq -\frac{1}{\beta\gamma}). \end{cases}$$

It is easily verified that

a) f_0 is continuously differentiable on $\mathbb{R}$, $f_0''(\xi)$ exists and $|f_0''(\xi)| \leq \gamma$ (for all $\xi \in \mathbb{R}$ with $\xi \notin \{0, -\frac{1}{\beta\gamma}, \frac{1}{\beta\gamma}\}$);

b) the equation $f_0(\xi) = 0$ has a unique solution at $\xi^* = 0$ and $|[f_0'(\xi^*)]^{-1}| = \beta$;

c) with $\xi_0 = \frac{2}{3\beta\gamma}$ we have

$$\xi_0 - \frac{f_0(\xi_0)}{f_0'(\xi_0)} = -\xi_0 \quad \text{and} \quad -\xi_0 - \frac{f_0(-\xi_0)}{f_0'(-\xi_0)} = \xi_0.$$

Consequently, $|\xi_0 - \xi^*| = \frac{2}{3\beta\gamma}$ and $\xi_0 \notin S(M,f_0)$. However, since $f_0 \notin F\langle\sigma,\beta,\gamma\rangle$ (f_0 is not twice-differentiable at $\xi = 0$, $\xi = \frac{1}{\beta\gamma}$ and $\xi = -\frac{1}{\beta\gamma}$), this example does *not* complete the proof of Theorem 6.2.1. The proof can be completed by using the next lemma. It states that an f with the desired properties exists, which is not only twice but infinitely differentiable on D(f). □

LEMMA 6.2.5. *If* $\sigma > \frac{2}{3\beta\gamma}$ *then for any* $\varepsilon > 0$ *there exists an* $F \in F\langle\sigma,\beta,\gamma\rangle$ *which is infinitely differentiable on* D(F), *for which*

$$r(M,F) < \frac{2}{3\beta\gamma} + \varepsilon.$$

For the case $E = \mathbb{R}$ we prove this lemma, with the function f_0 of Remark 6.2.2 in mind, in the next subsection. If $E \neq \mathbb{R}$ then Lemma 6.2.5 is proved in the subsection 6.2.3 (cf. p. 86).

In view of Lemma 6.2.4 and Remark 6.2.1 the proof of Theorem 6.2.1 is easily completed by application of Lemma 6.2.5.

6.2.2. Proof of Lemma 6.2.5 where $E = \mathbb{R}$

In this section we shall prove Lemma 6.2.5 if $E = \mathbb{R}$ with innerproduct $(\xi_1,\xi_2) = \xi_1\,\xi_2$ (for all $\xi_1,\xi_2 \in \mathbb{R}$).

PART A

We start with a lemma which will be used subsequently.

LEMMA 6.2.6. *Let ϕ: $D(\phi) \to \mathbb{R}$ with $D(\phi) \subset \mathbb{R}$. Assume that $D(\phi) \supset (-\varepsilon,\varepsilon)$ for some $\varepsilon > 0$. Suppose that ϕ has a fixed point $\xi = 0$. Assume that ϕ is continuous on $D(\phi)$ and that $|\phi(\xi)| \le |\xi|$ (for all $\xi \in \mathbb{R}$ with $0 \le |\xi| < \varepsilon$). If some $\lambda > 0$ satisfies $[-\lambda,\lambda] \subset D(\phi)$, $|\phi(\lambda)| \ge \lambda$ and $|\phi(-\lambda)| \ge \lambda$, then a number $\xi_0 \in [-\lambda,\lambda]$ with $\xi_0 \ne 0$ exists for which $\phi(\xi_0) \in D(\phi)$ and $\phi[\phi(\xi_0)] = \xi_0$.*

PROOF. If $\phi(\lambda) \ge \lambda$ or $\phi(-\lambda) \le -\lambda$ then $\phi(\xi_0) = \xi_0$ (for some $\xi_0 \in [-\lambda,\lambda]$ with $\xi_0 \ne 0$), so that the statement is true. Suppose $\phi(\lambda) \le -\lambda$ and $\phi(-\lambda) \ge \lambda$. Then a number $\delta_1 \in [-\lambda,0)$ exists such that $\phi(\delta_1) = \lambda$ and a number $\delta_2 \in (0,\lambda]$ exists for which $\phi(\delta_2) = -\lambda$. We assume that $|\phi(\xi)| \le \lambda$ (for all $\xi \in [\delta_1,\delta_2]$). This is no restriction. Hence $\phi(\xi) \in D(\phi)$ (for all $\xi \in [\delta_1,\delta_2]$). Since $\phi[\phi(\delta_2)] = \phi(-\lambda) \ge \lambda$ and $|\phi[\phi(\xi)]| \le |\xi|$ (for all $\xi \in (0,\varepsilon)$), a number $\xi_0 \in (0,\delta_2]$ exists such that $\phi[\phi(\xi_0)] = \xi_0$. This proves the lemma. □

PART B

We note that in Remark 6.2.2 we were not able to complete the proof of Theorem 6.2.1. But with the function f_0 in mind, we can construct an F that satisfies the proposition of Lemma 6.2.5.

We introduce some function classes. Let $\tau \in (0,\infty]$. We define

$$(6.2.4a)\qquad C^{\infty}(-\tau,\tau) = \{f \mid f\colon (-\tau,\tau) \to \mathbb{R};\ f \text{ is infinitely differentiable on } (-\tau,\tau)\}.$$

If $\tau = \infty$, we set

$$(6.2.4b)\qquad C^{\infty}(\mathbb{R}) = C^{\infty}(-\tau,\tau).$$

We give two well-known results (see also [COURANT, 1961; p. 172]).

LEMMA 6.2.7. *There exists a function $\psi \in C^{\infty}(\mathbb{R})$ such that*

$$\psi(\xi) = 0 \qquad (\text{if } \xi \le 0),$$

$$\psi(\xi) > 0 \qquad (\text{if } \xi > 0).$$

PROOF. Let $\psi\colon \mathbb{R} \to \mathbb{R}$,

$$\psi(\xi) = \begin{cases} 0 & (\text{if } \xi \le 0), \\ e^{-1/\xi} & (\text{if } \xi > 0). \end{cases} \tag{6.2.5}$$

It is easily verified that the statement of the lemma holds. □

LEMMA 6.2.8. *Let the numbers γ_1, γ_2, δ_1 and δ_2 be given. Suppose $\delta_1 < \delta_2$. Set $\bar{d} = (\gamma_1,\gamma_2,\delta_1,\delta_2)$. Then a monotone function $h_{\bar{d}} \in C^{\infty}(\mathbb{R})$ exists such that*

$$h_{\bar{d}}(\xi) = \begin{cases} \gamma_1 & (\text{if } \xi \le \delta_1), \\ \gamma_2 & (\text{if } \xi \ge \delta_2). \end{cases} \tag{6.2.6}$$

PROOF. We assume that $\gamma_1 \ne \gamma_2$ ($\gamma_1 = \gamma_2$ is a trivial case). Let $\phi(\xi) \equiv \psi(\delta_2 - \xi)\psi(\xi - \delta_1)$ where ψ is defined in (6.2.5). Set

$$\begin{aligned} &h_{\bar{d}}\colon \mathbb{R} \to \mathbb{R}, \\ &h_{\bar{d}}(\xi) = \gamma_1 + \Big[\int_{-\infty}^{\xi} \phi(\tau)d\tau\Big] \frac{\gamma_2 - \gamma_1}{\Big[\int_{\delta_1}^{\delta_2} \phi(\tau)d\tau\Big]} . \end{aligned} \tag{6.2.7}$$

It is easily verified that $h_{\bar{d}}$ has the properties stated. □

PART C

Let σ, β and γ denote the constants introduced at the beginning of section 6.2, which also appear in Lemma 6.2.5. Let $\delta \in (0,\frac{1}{9})$ and set

$$\bar{d}_1 = (0,\gamma,-\frac{1}{\beta\gamma},(-1+\delta)\frac{1}{\beta\gamma}),$$

$$\bar{d}_2 = (\gamma,-\gamma,-\frac{\delta}{\beta\gamma},\frac{\delta}{\beta\gamma}),$$

$$\bar{d}_3 = (-\gamma, 0, (1-\delta)\frac{1}{\beta\gamma}, \frac{1}{\beta\gamma}).$$

Let

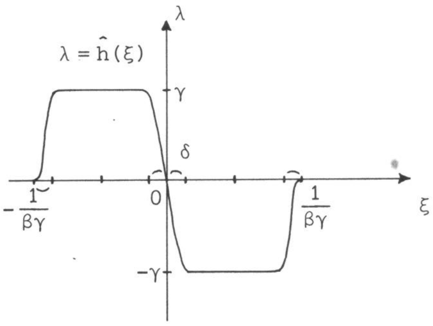
$$\hat{h}: \quad \mathbb{R} \to \mathbb{R},$$

(6.2.8)
$$\hat{h}(\xi) = \begin{cases} h_{\bar{d}_1}(\xi) & (\text{if } \xi \le -\frac{1}{2\beta\gamma}), \\ h_{\bar{d}_2}(\xi) & (\text{if } \xi \in (-\frac{1}{2\beta\gamma}, \frac{1}{2\beta\gamma})), \\ h_{\bar{d}_3}(\xi) & (\text{if } \xi \ge \frac{1}{2\beta\gamma}). \end{cases}$$

We define the function f by

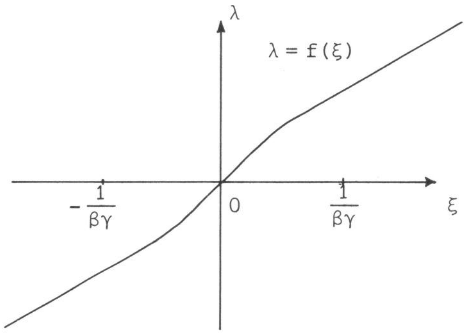
$$f: \mathbb{R} \to \mathbb{R},$$

(6.2.9)
$$f(\xi) = \frac{1}{\beta}\xi + \int_0^{\xi} (\xi-\tau)\hat{h}(\tau)d\tau \qquad (\xi \in \mathbb{R}).$$



graph of ĥ

Fig. 6.2.1



graph of f

Fig. 6.2.2

The following lemma holds.

LEMMA 6.2.9. *The function* f *defined in* (6.2.9) *has the following properties:*

$$f(0) = 0,\ f \in C^{\infty}(\mathbb{R}),$$

$$f'(\xi) > 0 \text{ (for all } \xi \in \mathbb{R}) \textit{ and}$$

$$f \in F\langle\sigma,\beta,\gamma\rangle.$$

PROOF. Notice that $\hat{h} \in C^{\infty}(\mathbb{R})$. Furthermore, for all $\xi \in \mathbb{R}$ we have $f'(\xi) = \frac{1}{\beta} + \int_0^{\xi} \hat{h}(\tau)d\tau$, $f''(\xi) = \hat{h}(\xi)$ and $f^{(k)}(\xi) = \hat{h}^{(k-2)}(\xi)$ $(k = 3,4,\ldots)$. Consequently, $f \in C^{\infty}(\mathbb{R})$ and $f'(0) = \frac{1}{\beta}$. Furthermore, for all $\xi \in \mathbb{R}$ we have

a) $|f''(\xi)| = |\hat{h}(\xi)| \le \gamma$;

b) if $\xi > 0$, $f'(\xi) \ge \frac{1}{\beta} + \int_0^{1/\beta\gamma} \hat{h}(\tau)d\tau > \frac{1}{\beta} - \frac{1}{\beta} = 0$;

if $\xi < 0$, $f'(\xi) \ge \frac{1}{\beta} + \int_0^{-1/\beta\gamma} \hat{h}(\tau)d\tau > \frac{1}{\beta} - \frac{1}{\beta} = 0$.

Since $f(0) = 0$ it follows that $\xi = 0$ is the unique solution of $f(\xi) = 0$. This proves the lemma. □

LEMMA 6.2.10. *Let* $\delta \in (0,\frac{1}{9})$. *Let* M *be Newton's method and let* f *be defined in* (6.2.9). *Set* $\phi = M(f)$. *Then* $D(\phi) = \mathbb{R}$. *Further, a number* $\xi_0 \in [-(\frac{2}{3}+2\delta)\frac{1}{\beta\gamma}, (\frac{2}{3}+2\delta)\frac{1}{\beta\gamma}]$ *exists such that* $\phi[\phi(\xi_0)] = \xi_0$ *and* $\xi_0 \notin S(M,f)$.

PROOF. From Lemma 6.2.9 and (6.2.1) it follows that $D(\phi) = \mathbb{R}$. Set $\lambda = (\frac{2}{3}+2\delta)\frac{1}{\beta\gamma}$. (Thus $0 < \lambda < (\frac{2}{3}+\frac{2}{9})\frac{1}{\beta\gamma} < (1-\delta)\frac{1}{\beta\gamma}$.) Then $f(\lambda) > \frac{1}{\beta}\lambda - \frac{\gamma}{2}\lambda^2$ and $f'(\lambda) < \frac{1}{\beta} - \gamma(\lambda - \frac{\delta}{\beta\gamma})$. Hence

$$\lambda - \frac{f(\lambda)}{f'(\lambda)} < \lambda - \frac{\frac{1}{\beta}\lambda - \frac{\gamma}{2}\lambda^2}{\frac{1}{\beta} - \gamma(\lambda - \frac{\delta}{\beta\gamma})}$$

$$= \frac{(\frac{\beta\gamma}{2}\lambda - \delta)}{1-\beta\gamma\lambda+\delta}(-\lambda) = \frac{(\frac{1}{3}+\delta-\delta)}{\frac{1}{3}-\delta}(-\lambda) < -\lambda.$$

Analogously we can show that

$$-\lambda - \frac{f(-\lambda)}{f'(-\lambda)} > \lambda.$$

From Lemma 6.2.3 and Lemma 6.2.6 it follows that a number $\xi_0 \in [-\lambda,\lambda]$ exists such that $\phi[\phi(\xi_0)] = \xi_0$ and $\xi_0 \notin S(M,f)$. This proves the lemma. □

By virtue of this last lemma, for all $\varepsilon > 0$ we are able to construct an F that satisfies the statement of Lemma 6.2.5 for the case $E = \mathbb{R}$.

In the next subsection we shall prove Lemma 6.2.5 where E is an arbitrary Hilbert space.

6.2.3. The E-extension of a real function

From [BROWN & PAGE, 1970; Theorem 9.2.13, 9.2.16] it follows that a subset $\mathcal{B}$ of E exists such that the following three statements (i), (ii) and (iii) hold.

(i) *All* $u,v \in \mathcal{B}$ *satisfy* $(u,v) = 0$ *(if* $u \neq v$*) and* $(u,v) = 1$ *(if* $u = v$*).*

(ii) *For any* $x \in E$*, the set* $\dot{\mathcal{B}}_x = \{u \mid u \in \mathcal{B};\ (x,u) \neq 0\}$ *is countable.*

(iii) *Let* $x \in E$ *and let* $\dot{\mathcal{B}}_x = \{u \mid u = u_n \text{ with } n \in \mathbb{N}_0\}$. *Here* $\mathbb{N}_0 \subset \mathbb{N}$ *is a set of consecutive integers with* $1 \in \mathbb{N}_0$. *In the case that* $\mathbb{N}_0$ *contains only* $\ell < \infty$ *numbers, put* $u_n = 0$ $(n = \ell+1, \ell+2, \ell+3, \ldots)$. *Put* $\mathcal{B}_x = \{u \mid u = u_n \text{ with } n \in \mathbb{N}\}$. *Then*

$$x = \sum_{n=1}^{\infty} (x,u_n)u_n$$

and

$$\|x\|^2 = \sum_{n=1}^{\infty} (x,u_n)^2.$$

From now on, $\mathcal{B}$ denotes a fixed subset of E for which the above mentioned properties hold.

We call a sequence $\{j_n\}$ in $\mathbb{N}$ in which every positive integer appears once and only once, a *reordering* of $\mathbb{N}$.

Let $\{u_n\} \subset \mathcal{B} \cup \{0\}$ satisfy $(u_i,u_j) = 0$ $(i,j \in \mathbb{N},\ i \neq j)$.

<u>LEMMA 6.2.11</u>. *Let the sequence* $\{\eta_n\}$ *in* $\mathbb{R}$ *satisfy* $\eta_n = 0$ *if* $u_n = 0$ $(n = 1,2,\ldots)$. *Suppose* $\Sigma_{n=1}^{\infty} \eta_n^2 < \infty$. *Then the following four statements* (i) - (iv) *hold.*

(i) *The sequence* $\{\Sigma_{n=1}^{N} \eta_n u_n\}$ *converges. Put* $y = \Sigma_{n=1}^{\infty} \eta_n u_n$.

(ii) *Let* $\{j_n\}$ *be a reordering of* $\mathbb{N}$. *Then* $\{\Sigma_{n=1}^{N} \eta_{j_n} u_{j_n}\}$ *converges and* $\Sigma_{n=1}^{\infty} \eta_{j_n} u_{j_n} = y$.

(iii) $\|y\|^2 = \Sigma_{n=1}^{\infty} \eta_n^2$.

(iv) *Let* $c > 0$. *Suppose* $\{\alpha_n\}$ *is a sequence in* $\mathbb{R}$ *that satisfies* $|\alpha_n| \leq c|\eta_n|$ $(n = 1,2,\ldots)$. *Then* $\{\Sigma_{n=1}^{N} \alpha_n u_n\}$ *converges. Put* $z = \Sigma_{n=1}^{\infty} \alpha_n u_n$. *Then* $\|z\| \leq c\|y\|$.

PROOF. (i) Let $\varepsilon > 0$. Then an $N_1 \in \mathbb{N}$ exists such that

$$\sum_{n=N+1}^{N+m} \eta_n^2 < \varepsilon \qquad \text{(for all } N,m \in \mathbb{N} \text{ with } N \geq N_1). \tag{6.2.10}$$

Consequently,

$$\| \sum_{n=1}^{N+m} \eta_n u_n - \sum_{n=1}^{N} \eta_n u_n \|^2 = \sum_{n=N+1}^{N+m} \eta_n^2 < \varepsilon \qquad \text{(for all } N,m \in \mathbb{N} \text{ with } N \geq N_1).$$

This proves (i).

(ii) Let $\varepsilon > 0$. Then an $N_1 \in \mathbb{N}$ exists such that (6.2.10) holds. Choose $N_2 \geq N_1$ such that $n \in \{j_1, j_2, \ldots, j_{N_2}\}$ $(n = 1,2,\ldots,N_1)$. Then $\| y - \Sigma_{n=1}^{N} \eta_{j_n} u_{j_n} \|^2 \leq \varepsilon$ (for all $N \geq N_2$). Hence (ii) holds.

(iii) Suppose $\mathcal{B}_y = \{u \mid u = v_n \text{ with } n \in \mathbb{N}\}$ and $\zeta_n = (y, v_n)$ $(n = 1,2,\ldots)$. It is easily verified that a reordering $\{j_n\}$ of $\mathbb{N}$ exists, such that $\zeta_{j_n} = \eta_n$ $(n = 1,2,\ldots)$. Consequently,

$$\|y\|^2 = \sum_{n=1}^{\infty} \zeta_n^2 = \sum_{n=1}^{\infty} \zeta_{j_n}^2 = \sum_{n=1}^{\infty} \eta_n^2$$

(cf. [RUDIN, 1953; Theorem 3.56]).

(iv) Let $N,m \geq 0$. Then

$$\| \sum_{n=1}^{N+m} \alpha_n u_n - \sum_{n=1}^{N} \alpha_n u_n \|^2 = \sum_{n=N+1}^{N+m} \alpha_n^2 \leq c^2 \sum_{n=N+1}^{N+m} |\eta_n|^2.$$

Hence $\{\Sigma_{n=1}^{N} \alpha_n u_n\}$ is a Cauchy sequence and $\|z\| \leq c\|y\|$. □

We shall now define the E-extension of a real function. Let $\tau \in (0,\infty]$. Let

$$\phi \in C^{\infty}(-\tau,\tau). \tag{6.2.11a}$$

Suppose

$$\phi(0) = 0. \tag{6.2.11b}$$

Let $x \in B(0,\tau)$ and $\mathcal{B}_x = \{u \mid u = u_n \text{ with } n \in \mathbb{N}\}$. Let the constant $c > 0$ satisfy

$$(6.2.12a) \qquad c \geq |\phi'(\xi)| \qquad (\text{for all } \xi \in \mathbb{R} \text{ with } |\xi| \leq \|x\|).$$

Set $\xi_n = (x,u_n)$ $(n = 1,2,\ldots)$. Thus $x = \Sigma_{n=1}^{\infty} \xi_n u_n$ and $|\xi_n| \leq \|x\|$ $(n = 1,2,\ldots)$. From Lemma 2.6.1 it follows that

$$(6.2.12b) \qquad |\phi(\xi_n)| \leq c|\xi_n| \qquad (n = 1,2,\ldots).$$

Lemma 6.2.11(iv) implies that $\{\Sigma_{n=1}^{N} \phi(\xi_n)u_n\}$ is a Cauchy sequence. We set

$$(6.2.13a) \qquad F(x) = \sum_{n=1}^{\infty} \phi(\xi_n)u_n$$

for

$$(6.2.13b) \qquad x = \sum_{n=1}^{\infty} \xi_n u_n.$$

Let $\{\eta_n\},\{\zeta_n\} \subset \mathbb{R}$. Let $\Sigma_{n=1}^{\infty} \eta_n^2 < \infty$ and $\Sigma_{n=1}^{\infty} \zeta_n^2 < \infty$. Suppose $\Sigma_{n=1}^{\infty} \eta_n v_n = \Sigma_{n=1}^{\infty} \zeta_n w_n = x$ with $v_n, w_n \subset \mathcal{B} \cup \{0\}$ $(n = 1,2,\ldots)$ and $(v_i,v_j) = (w_i,w_j) = 0$ $(i,j \in \mathbb{N}, i \neq j)$. Then $F(x) = \Sigma_{n=1}^{\infty} \phi(\eta_n)v_n = \Sigma_{n=1}^{\infty} \phi(\zeta_n)w_n$ (cf. Lemma 6.2.11(ii)).

For given ϕ satisfying (6.2.11) we call

$$(6.2.14) \qquad \begin{array}{l} F\colon B(0,\tau) \to E, \quad \text{where} \\ F(x) \text{ is defined in } (6.2.13) \ (\text{for all } x \in B(0,\tau)), \end{array}$$

the E*-extension of* ϕ.

<u>THEOREM 6.2.12</u>. *Let* ϕ *satisfy* (6.2.11) *and suppose* $\phi'(\xi) \neq 0$ *(for all* $\xi \in (-\tau,\tau)$*). Let* F *be the* E*-extension of* ϕ*. Then*

(i) $x = 0$ *is the unique solution of* $F(x) = 0$.

(ii) F *is infinitely differentiable on* D(F)*. Let* $k \in \mathbb{N}$*. For* $x \in D(F)$ *and* $h_j \in E$ $(j = 1,2,\ldots,k)$ *Let*

$$\mathcal{B}_x \cup \{\bigcup_{j=1}^{k} \mathcal{B}_{h_j}\} = \{u \mid u = v_n \textit{ with } n \in \mathbb{N}\}.$$

For $n = 1,2,\ldots,$ *let* $\xi_n = (x,v_n)$ *and* $h_{j,n} = (h_j,v_n)$ $(j = 1,2,\ldots,k)$. *Then*

$$(6.2.15a) \qquad F^{(k)}(x)h_1h_2\ldots h_k = \sum_{n=1}^{\infty} \phi^{(k)}(\xi_n)h_{1,n}h_{2,n}\ldots h_{k,n}v_n.$$

Suppose the number c_k *satisfies*

$$(6.2.15b) \qquad c_k \geq |\phi^{(k)}(\xi)| \qquad \textit{(for all } \xi \in \mathbb{R} \textit{ with } |\xi| \leq \|x\|).$$

Then

$$(6.2.15c) \qquad \|F^{(k)}(x)\| \leq c_k.$$

(iii) $F'(x)$ *is invertible (for all* $x \in D(F)$*). Let* $x \in D(F)$ *and* $y \in E$*, and let* $\mathcal{B}_x \cup \mathcal{B}_y = \{u \mid u = u_n$ *with* $n \in \mathbb{N}\}$*. Let* $\xi_n = (x,u_n)$ *and* $\eta_n = (y,u_n)$ $(n = 1,2,\ldots)$*. Then*

$$[F'(x)]^{-1}y = \sum_{n=1}^{\infty} [\phi'(\xi_n)]^{-1}\eta_n u_n.$$

In particular

$$(6.2.16) \qquad [F'(x)]^{-1}F(x) = \sum_{n=1}^{\infty} [\phi'(\xi_n)]^{-1}\phi(\xi_n)u_n.$$

<u>PROOF</u>. (i) Obviously, $x = 0$ is a solution of $F(x) = 0$. Suppose $F(x) = 0$ has a solution $x = y^*$. Let $\mathcal{B}_{y*} = \{u \mid u = v_n$ with $n \in \mathbb{N}\}$ and $\eta_n = (y^*,v_n)$ $(n = 1,2,\ldots)$. Then $F(y^*) = \Sigma_{n=1}^{\infty} \phi(\eta_n)v_n$. Consequent on Lemma 6.2.11(iii), $0 = \|F(y^*)\|^2 = \Sigma_{n=1}^{\infty} [\phi(\eta_n)]^2$. Hence $\phi(\eta_n) = 0$, so that $\eta_n = 0$ $(n = 1,2,\ldots)$. This implies $y^* = 0$.
(ii) Suppose $F^{(k)}(x)$ exists and satisfies (6.3.15a,c) (for all $x \in D(F)$), with $k \geq 0$ (if $k = 0$ then (6.3.15a) should read $F(x) = \Sigma_{n=1}^{\infty} \phi(\xi_n)v_n$).

Let $x \in D(F)$. Let the positive numbers δ and ε satisfy $\|x\| + \varepsilon \leq \delta < \tau$. We notice that a constant $c > 0$ exists such that

$$(6.2.17) \qquad |\phi^{(k)}(\xi+\eta) - \phi^{(k)}(\xi) - \phi^{(k+1)}(\xi)\eta| \leq c\eta^2 \qquad \text{(for all } \xi,\eta \in \mathbb{R} \text{ with } |\xi| + |\eta| \leq \delta).$$

Let the number c_{k+1} satisfy

$$(6.2.18) \qquad |\phi^{(k+1)}(\xi)| \leq c_{k+1} \qquad \text{(for all } \xi \in \mathbb{R} \text{ with } |\xi| \leq \|x\|).$$

Let $y,h_1,h_2,\ldots,h_k \in E$. Let $\mathcal{B}_x \cup \mathcal{B}_y \cup \{\cup_{i=1}^{k} \mathcal{B}_{h_i}\} = \{u \mid u = v_n$ with $n \in \mathbb{N}\}$. For $n = 1,2,\ldots,$ let

$$\xi_n = (x, v_n), \quad \eta_n = (y, v_n) \quad \text{and} \quad h_{i,n} = (h_i, v_n) \quad (i = 1,2,\dots,k). \tag{6.2.19}$$

Set $\alpha_n = \phi^{(k+1)}(\xi_n)\eta_n h_{1,n} h_{2,n} \dots h_{k,n}$, then $|\alpha_n| \le c_{k+1}\|h_1\|\|h_2\|\dots\|h_k\||\eta_n|$ $(n = 1,2,\dots)$. From Lemma 6.2.11(iv) it follows that $\{\Sigma_{n=1}^{N} \alpha_n v_n\}$ converges. Set $z = \Sigma_{n=1}^{\infty} \alpha_n v_n$. Then

$$z = \sum_{n=1}^{\infty} \phi^{(k+1)}(\xi_n)\eta_n h_{1,n} h_{2,n} \dots h_{k,n} v_n \tag{6.2.20a}$$

and

$$\|z\| \le c_{k+1}\|y\|\|h_1\|\dots\|h_k\|. \tag{6.2.20b}$$

Suppose $\mathcal{B}_x \cup \mathcal{B}_y \cup \{\cup_{i=1}^{k} \mathcal{B}_{h_i}\} = \{u \mid u = \tilde{v}_n \text{ with } n \in \mathbb{N}\}$, and for $n = 1,2,\dots$, $\tilde{\xi}_n = (x,\tilde{v}_n)$, $\tilde{\eta}_n = (y,\tilde{v}_n)$ and $\tilde{h}_{i,n} = (h_i,\tilde{v}_n)$ $(i = 1,2,\dots,k)$. Then a reordering $\{j_n\}$ of $\mathbb{N}$ exists such that for $n = 1,2,\dots$ we have $\tilde{v}_n = v_{j_n}$, $\tilde{\xi}_n = \xi_{j_n}$, $\tilde{\eta}_n = \eta_{j_n}$ and $\tilde{h}_{i,n} = h_{i,j_n}$ $(i = 1,2,\dots,k)$. Let $\tilde{\alpha}_n = \phi^{(k+1)}(\tilde{\xi}_n)\tilde{\eta}_n\tilde{h}_{1,n}\tilde{h}_{2,n}\dots \dots\tilde{h}_{k,n}$ $(n = 1,2,\dots)$. By virtue of Lemma 6.2.11(ii), the sequence $\{\Sigma_{n=1}^{N} \tilde{\alpha}_n\tilde{v}_n\}$ converges and $\Sigma_{n=1}^{\infty} \tilde{\alpha}_n\tilde{v}_n = z$.

Let

$$Q \in L^{(k+1)}(E),$$

$$Qyh_1h_2\dots h_k = z \quad \text{(for all } y,h_1,h_2,\dots,h_k \in E\text{. Here } z \text{ is defined in (6.2.20a) with } y,h_1,h_2,\dots,h_k \text{ satisfying (6.2.19)).} \tag{6.2.21}$$

In view of (6.2.20b), $\|Q\| \le c_{k+1}$.

We show that $Q = F^{(k+1)}(x)$. Let $y,h_1,h_2,\dots,h_k \in E$ satisfy (6.2.19). Suppose $\|y\| \le \varepsilon$ and $\|h_1\| = \|h_2\| = \dots\|h_k\| = 1$. (Notice that $\{\Sigma_{n=1}^{\infty} \eta_n^2\}^2 \ge \{\Sigma_{n=1}^{N} \eta_n^2\}^2 \ge \Sigma_{n=1}^{N} \eta_n^4$ (for all $N \in \mathbb{N}$).) Then (cf. (6.2.17))

$$\begin{aligned}
&\|F^{(k)}(x+y)h_1h_2\dots h_k - F^{(k)}(x)h_1h_2\dots h_k - Qyh_1h_2\dots h_k\|^2 \\
&= \sum_{n=1}^{\infty} [\phi^{(k)}(\xi_n + \eta_n) - \phi^{(k)}(\xi_n) - \phi^{(k+1)}(\xi_n)\eta_n]^2 h_{1,n}^2 \dots h_{k,n}^2 \\
&\le c^2 \sum_{n=1}^{\infty} \eta_n^4 \le c^2\|y\|^4.
\end{aligned}$$

Hence,

$$\|F^{(k)}(x+y) - F^{(k)}(x) - Qy\| \le c\|y\|^2 \quad (\text{for all } y \in E \text{ with } \|y\| \le \varepsilon).$$

This implies that $Q = F^{(k+1)}(x)$. Therefore (6.2.15) holds for k+1. Consequently, (6.2.15) holds for all $k \in \mathbb{N}$.

(iii) Let $x \in D(F)$. Then a constant c exists such that

(6.2.22) $\quad |[\phi'(\xi)]^{-1}| \le c \quad (\text{for all } \xi \in \mathbb{R} \text{ with } |\xi| \le \|x\|).$

For $y \in E$ let $\mathcal{B}_x \cup \mathcal{B}_y = \{u \mid u = v_n \text{ with } n \in \mathbb{N}\}$. Let

(6.2.23) $\quad \xi_n = (x,v_n) \quad$ and $\quad \eta_n = (y,v_n) \qquad (n = 1,2,\ldots).$

According to (6.2.22) and Lemma 6.2.11(iv), the sequence $\{\Sigma_{n=1}^{N} [\phi'(\xi_n)]^{-1}\eta_n v_n\}$ converges, and with

(6.2.24a) $\quad z = \sum_{n=1}^{\infty} [\phi'(\xi_n)]^{-1}\eta_n v_n$

we have

(6.2.24b) $\quad \|z\| \le c\|y\|.$

Suppose $\mathcal{B}_x \cup \mathcal{B}_y = \{u \mid u = \tilde{v}_n \text{ with } n \in \mathbb{N}\}$. Let $\tilde{\xi}_n = (x,\tilde{v}_n)$ and $\tilde{\eta}_n = (y,\tilde{v}_n)$ $(n = 1,2,\ldots)$. From Lemma 6.2.11(ii) it follows that $\Sigma_{n=1}^{\infty} [\phi'(\tilde{\xi}_n)]^{-1}\tilde{\eta}_n\tilde{v}_n = z$.

Let

(6.2.25)
$$C: E \to E,$$
$$Cy = z \quad (z \text{ is defined in (6.2.24a) with } y \in E \text{ satisfying (6.2.23)}).$$

Obviously C is linear in y and from (6.2.24b) it follows that C is bounded. Moreover

$$F'(x)Cy = CF'(x)y = y \quad (\text{for all } y \in E).$$

This completes the proof. □

Let the constants ϕ, β and γ denote the constants appearing in Lemma 6.2.5, which were introduced at the beginning of section 6.2.

PROOF OF LEMMA 6.2.5. Let $\delta \in (0,\frac{1}{9})$. Let f be defined in (6.2.9). F is the E-extension of f. From Lemma 6.2.9 and Theorem 6.2.12 it follows that the equation $F(x) = 0$ has a unique solution $x = 0$, that F is infinitely differentiable on E, that $F'(0)$ is invertible and $[F'(0)]^{-1} = \beta I$, and that $\|F''(x)\| \leq \gamma$ (for all $x \in E$). Hence $F \in \mathcal{F}\langle\sigma,\beta,\gamma\rangle$. Let $u \in \mathcal{B}$. Then $\Gamma(\lambda u)F(\lambda u) = [f'(\lambda)]^{-1}f(\lambda)u$ (for all $\lambda \in \mathbb{R}$), (cf. (6.2.16)). Set $G = M(F)$. Let ϕ be defined in Lemma 6.2.10. It follows that $G(\lambda u) = \phi(\lambda)u$ (for all $\lambda \in \mathbb{R}$). According to Lemma 6.2.10 a number $\xi_0 \in [-(\frac{2}{3}+2\delta)\frac{1}{\beta\gamma},(\frac{2}{3}+2\delta)\frac{1}{\beta\gamma}]$ with $\xi_0 \neq 0$ exists for which $\phi[\phi(\xi_0)] = \xi_0$. Consequently, $G(G(\xi_0 u)) = \xi_0 u$. Hence $\xi_0 u \notin S(M,F)$. This proves Lemma 6.2.5. □

6.3. ITERATIVE METHODS WITH GREATER RADII OF CONVERGENCE

Let $\sigma \in (0,\infty]$ and $\beta,\gamma > 0$. In this section we present a class of iterative methods (for $\mathcal{F}_1$) which all have greater radii of convergence with respect to $\mathcal{F}\langle\sigma,\beta,\gamma\rangle$ than Newton's method when $\sigma > \frac{2}{3\beta\gamma}$.

Let

(6.3.1a) $\quad A \in \mathcal{A}$ with $[A(F)](y,x) \equiv F(y) - F(x)$ (for all $F \in \mathcal{F}_1$),

let

(6.3.1b) $\quad g \in \mathcal{S}$ with $g(t) = \begin{cases} \frac{1}{1-t} & (\text{if } t \in [0,1)), \\ 1 & (\text{if } t = 1) \end{cases}$

and let

(6.3.1c) L be Euler's method

(cf. (2.6.1), (2.6.3) and (2.6.9)).

In this section we shall be concerned, among other things, with the iterative methods M that satisfy

(6.3.2) $\quad M = \mathbb{M}(A,g,L,\mathcal{H})$

(cf. (4.1.3)). Here $\mathcal{H} = \{h_0,h_1,\ldots,h_N\}$ with $N \in \mathbb{N}$.

Let M satisfy (6.3.2) and let $F \in \mathcal{F}_1$. We notice that proposition (vi) of Theorem 5.2.5 with $K = A(F)$ is true. Consequently, the iterative process [M,F] is quadratically convergent.

For $F \in \mathcal{F}_1$, the function $G = M(F)$ is defined by (cf. (4.1.4))

$$G: D(G) \to E, \qquad G = \eta_{N+1}. \tag{6.3.3a}$$

In (6.3.3a) the function η_{N+1} is defined as follows:

$$\begin{aligned} &\eta_0: D(\eta_0) \to E,\\ &D(\eta_0) = D(F),\\ &\eta_0(x) = x \quad (\text{for all } x \in D(\eta_0)) \end{aligned} \tag{6.3.3b}$$

and for n = 0,1,...,N the functions η_{n+1} are defined by

$$\begin{aligned} &\eta_{n+1}: D(\eta_{n+1}) \to E,\\ &D(\eta_{n+1}) = \{x \mid x \in D(\eta_n),\ \eta_n(x) \in D(F) \text{ and } F'(\eta_n(x)) \text{ is invertible}\},\\ &\eta_{n+1}(x) = \eta_n(x) - \omega_n \Gamma(\eta_n(x))F(\eta_n(x)) \quad (\text{for all } x \in D(\eta_{n+1})). \end{aligned} \tag{6.3.3c}$$

In (6.3.3c)

$$\omega_n = \frac{h_n}{1-t_n} \qquad (n = 0,1,\ldots,N). \tag{6.3.4}$$

Let

$$\Omega = \{\bar{\omega} \mid \bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N) \text{ with } N \in \mathbb{N};\ \omega_n \in (0,1)\ (n = 0,1,\ldots,N-1) \text{ and } \omega_N = 1\}. \tag{6.3.5}$$

DEFINITION 6.3.1. Let $\bar{\omega} \in \Omega$ with $\bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N)$ and $N \in \mathbb{N}$. By $M_{\bar{\omega}}$ we denote the iterative method for which, for all $F \in \mathcal{F}_1$, the function $G = M_{\bar{\omega}}(F)$ is defined in (6.3.3).

Let $H = \{h_0,h_1,\ldots,h_N\}$ with $N \in \mathbb{N}$. Since $0 < h_n < 1-t_n$

$(n = 0,1,\ldots,N-1)$ and $h_N = 1 - t_N$, it follows that $\mathbb{M}(A,g,L,H) = M_{\bar{\omega}}$. Here $\bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N)$ and ω_n $(n = 0,1,\ldots,N)$ is defined by (6.3.4).

Conversely, let $\bar{\omega} \in \Omega$ with $\bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N)$. Then the sequence $H = \{h_0,h_1,\ldots,h_N\}$ where h_n $(n = 0,1,\ldots,N)$ satisfies (6.3.4), is uniquely determined. Consequently, $M_{\bar{\omega}} = \mathbb{M}(A,g,L,H)$.

For notational convenience, instead of iterative methods M of the type (6.3.2) we shall consider the iterative methods $M_{\bar{\omega}}$ $(\bar{\omega} \in \Omega)$. In view of the above considerations, this is no restriction.

The following theorem holds.

THEOREM 6.3.1. *Let* $\bar{\omega} \in \Omega$. *Then* $M_{\bar{\omega}}$ *is quadratically convergent. Further*

$$r(M_{\bar{\omega}};\ F\langle\sigma,\beta,\gamma\rangle) = \sigma \qquad (\text{if } \sigma \le \frac{2}{3\beta\gamma}),$$

$$r(M_{\bar{\omega}};\ F\langle\sigma,\beta,\gamma\rangle) > \frac{2}{3\beta\gamma} \qquad (\text{if } \sigma > \frac{2}{3\beta\gamma}).$$

We shall prove this theorem in the next subsection.

Let $F \in F_1$ (cf. (2.6.1)). A well-known class of numerical methods for solving $F(x) = 0$ are the so-called *damped Newton methods*. For given $x_0 \in D(F)$ these methods generate a sequence $\{x_n\}$ for which

$$(6.3.6) \qquad x_{n+1} = x_n - \omega_n\Gamma(x_n)F(x_n) \qquad (n = 0,1,2,\ldots)$$

(cf. [ORTEGA & RHEINBOLDT, 1970; chapter 14.4]). Here $0 < \omega_n \le 1$ $(n = 0,1,2,\ldots)$ and $\omega_k < 1$ for at least one k. Obviously, the methods $M_{\bar{\omega}}$ $(\bar{\omega} \in \Omega)$ are damped Newton methods with periodic coefficients ω_n and with $\omega_0 < 1$, $\omega_1 < 1$, $\ldots,\omega_{N-1} < 1$, $\omega_N = 1$ for some $N \ge 1$.

If $\sigma > \frac{2}{3\beta\gamma}$, one may conjecture that the general damped Newton methods have a greater radius of convergence with respect to $F\langle\sigma,\beta,\gamma\rangle$ than Newton's method (cf. [ORTEGA & RHEINBOLDT, 1970; Theorem 14.4.4], [DEUFLHARD, 1974]). This means that, for an $F \in F\langle\sigma,\beta,\gamma\rangle$, these methods are expected to generate sequences $\{x_n\}$ that converge to x^*, even if $\|x_0-x^*\| \ge \frac{2}{3\beta\gamma}$. Theorem 6.3.1 shows that this holds for the damped Newton methods $M_{\bar{\omega}}$. The next theorem shows that the conjecture is also true for all damped Newton methods with $\inf\{\omega_n \mid n \ge 0\} > 0$.

THEOREM 6.3.2. *Let* $0 < \omega_n \le 1$ $(n = 0,1,2,\ldots)$ *and* $\omega_k < 1$ *for at least one* k. *Let* $\inf\{\omega_n \mid n \ge 0\} > 0$. *Then a constant* $\hat{\mu}$ *exists such that the following* *propositions hold.*

(i) $\hat{\mu} = \sigma$ *(if* $\hat{\mu} \leq \frac{2}{3\beta\gamma}$*) and* $\hat{\mu} > \frac{2}{3\beta\gamma}$ *(if* $\sigma > \frac{2}{3\beta\gamma}$*).*

(ii) *Let* $F \in \hat{F}<\sigma,\beta,\gamma>$. *Then for any* $x_0 \in B(x^*,\hat{\mu})$ *a sequence* $\{x_n\}$ *exists such that for* $n = 0,1,2,\ldots$

$$x_n \in D(F), \quad F'(x_n) \textit{ is invertible,}$$

$$x_{n+1} = x_n - \omega_n\Gamma(x_n)F(x_n),$$

and $\lim_{n\to\infty} x_n = x^*$.

We shall prove this theorem in the next subsection.

6.3.1. Proof of the Theorems 6.3.1 and 6.3.2

In order to prove the Theorems 6.3.1 and 6.3.2 we require the following lemmata.

LEMMA 6.3.3. *Let* $P: D(P) \to E$ *with* $D(P)$ *an open subset of* E. *Let* $x_0 \in D(P)$. *Suppose that* P *is twice differentiable on* $D(P)$ *and that* $P'(x_0)$ *is invertible. Suppose positive constants* α_0, β_0, γ_0 *and* τ *exist such that*

(i) $\quad \|[P'(x_0)]^{-1}\| \leq \beta_0, \quad \|[P'(x_0)]^{-1}P(x_0)\| \leq \alpha_0,$

(ii) $\quad h < \frac{1}{2}$ *where* $h = \alpha_0\beta_0\gamma_0,$

(iii) $\quad B(x_0,\tau) \subset D(P)$ *and* $\|P''(x)\| \leq \gamma_0$ *(for all* $x \in B(x_0,\tau)$*),*

(iv) $\quad r_1 < \tau < r_2$ *with* $r_i = [1 + (-1)^i\sqrt{1-2h}]/(\beta\gamma)$ $(i = 1,2)$.

Under these assumptions the equation $P(x) = 0$ *has a solution* $\tilde{x}$ *in* $\bar{B}(x_0,r_1)$ *and* $\tilde{x}$ *is the unique solution of* $P(x) = 0$ *in* $B(x_0,\tau)$.

PROOF. The conclusion is a direct consequence of the well-known Newton-Kantorovich theorem (cf. [KANTOROWITSCH & AKILOW, 1964; Theorem 6(1.XVIII)], [GRAGG & TAPIA, 1974]). □

LEMMA 6.3.4. *If* $u,x,y,z \in E$ *and* $z = \omega x + (1-\omega)y$ *with* $\omega \in \mathbb{R}$, *then*

$$\|z-u\|^2 = \omega\|x-u\|^2 + (1-\omega)\|y-u\|^2 - \omega(1-\omega)\|x-y\|^2.$$

PROOF. Observe that the following relations (a) and (b) hold.

(a) $\|z-u\|^2 = \|\omega(x-u)+(1-\omega)(y-u)\|^2 = \omega^2\|x-u\|^2+(1-\omega)^2\|y-u\|^2+2\omega(1-\omega)(x-u,y-u)$.

(b) $\|x-y\|^2 = \|(x-u)-(y-u)\|^2 = \|x-u\|^2+\|y-u\|^2-2(x-u,y-u)$.

Therefore,

$$2(x-u,y-u) = \|x-u\|^2 + \|y-u\|^2 - \|x-y\|^2.$$

Together with relation (a), this proves the lemma. □

Throughout this subsection

(6.3.7) $$\hat{\sigma} = \min\{\sigma,\frac{1}{\beta\gamma}\}.$$

Let $\omega \in (0,1]$. Define

(6.3.8)
$$\delta_\omega\colon [0,\hat{\sigma}) \to [0,\infty),$$
$$\delta_\omega(\varepsilon) = (1-\omega) + \omega\left[\frac{\beta\gamma\varepsilon}{2(1-\beta\gamma\varepsilon)}\right]^2 - \omega(1-\omega)\left[\frac{(2-\beta\gamma\varepsilon)}{2(1+\beta\gamma\varepsilon)}\right]^2 \qquad (\varepsilon \in [0,\hat{\sigma})).$$

The following lemma holds.

LEMMA 6.3.5. *δ_ω is continuous and strictly isotone on $[0,\hat{\sigma})$, and $\delta_\omega(0) = (1-\omega)^2$.*

PROOF. The proof of this lemma is straightforward. □

As a consequence of this lemma we have

(6.3.9) $$\delta_\omega(\varepsilon) \geq 0 \qquad (\text{for all } \varepsilon \in [0,\hat{\sigma})).$$

Let the function θ_ω be defined by

(6.3.10)
$$\theta_\omega\colon [0,\hat{\sigma}) \to [0,\infty),$$
$$\theta_\omega(\varepsilon) = \sqrt{\delta_\omega(\varepsilon)} \qquad (\varepsilon \in [0,\hat{\sigma})).$$

Let

$$(6.3.11a)\qquad \mu_\omega = \sup\{\varepsilon \mid \varepsilon \in [0,\hat{\sigma});\ \theta_\omega(\varepsilon) < 1\}$$

and

$$(6.3.11b)\qquad \pi_\omega = \sup\{\varepsilon \mid \varepsilon \in [0,\hat{\sigma});\ \theta_\omega(\varepsilon)\varepsilon < \hat{\sigma}\}.$$

Obviously, $\mu_\omega \le \pi_\omega \le \hat{\sigma}$ (see also Fig. 6.3.1).

<u>LEMMA 6.3.6</u>. *Let* $\omega \in (0,1]$. *Let* θ_ω *be defined by* (6.3.10). *The following statements* (i), (ii) *and* (iii) *hold.*

(i) $\theta_\omega(0) = 1-\omega$ *and* θ_ω *is continuous and strictly isotone on* $[0,\hat{\sigma})$.

(ii) *If* $\hat{\sigma} > \frac{2}{3\beta\gamma}$ *then* $\theta_\omega(\frac{2}{3\beta\gamma}) < 1$ *(if* $\omega \in (0,1)$*) and* $\theta_\omega(\frac{2}{3\beta\gamma}) = 1$ *(if* $\omega = 1$*).* *If* $\hat{\sigma} \le \frac{2}{3\beta\gamma}$ *then* $\lim_{\varepsilon\uparrow\hat{\sigma}} \theta_\omega(\varepsilon) < 1$ *(if* $\omega \in (0,1)$*) and* $\lim_{\varepsilon\uparrow\hat{\sigma}} \theta_\omega(\varepsilon) \le 1$ *(if* $\omega = 1$*).*

(iii) $\mu_\omega > \frac{2}{3\beta\gamma}$ *(if* $\hat{\sigma} > \frac{2}{3\beta\gamma}$*) and* $\mu_\omega = \hat{\sigma}$ *(if* $\hat{\sigma} \le \frac{2}{3\beta\gamma}$*).*

<u>PROOF</u>. By virtue of Lemma 6.3.5 it follows that $\theta_\omega(0) = 1-\omega$ and that θ_ω is continuous and strictly isotone on $[0,\hat{\sigma})$. If $\hat{\sigma} > \frac{2}{3\beta\gamma}$ then it is easily verified that $\theta_\omega(\frac{2}{3\beta\gamma}) < 1$ (if $\omega \in (0,1)$) and $\theta_\omega(\frac{2}{3\beta\gamma}) = 1$ (if $\omega = 1$). This completes the proof. □

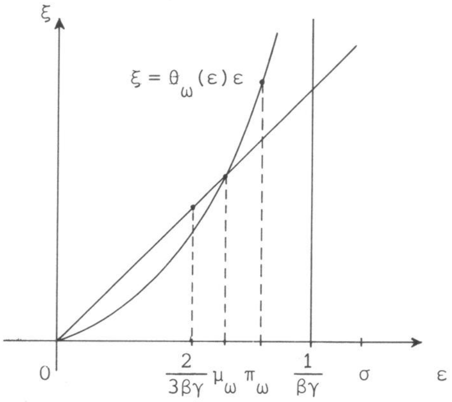


Fig. 6.3.1.

$(\hat{\sigma} = \frac{1}{\beta\gamma};\ \omega \in (0,1))$.

The following important lemma holds.

LEMMA 6.3.7. *Let* $F \in \mathcal{F}\langle\sigma,\beta,\gamma\rangle$ *and* $\omega \in (0,1]$. *Let* θ_ω *be defined by* (6.3.10). *Let* $y \in B(x^*,\hat{\sigma})$. *Then* $F'(y)$ *is invertible. Set*

$$z = y - \omega\Gamma(y)F(y).$$

Then the following estimate holds.

$$\|z-x^*\| \le \theta_\omega(\|y-x^*\|)\|y-x^*\|.$$

Further, if $\|y-x^*\| \le \rho < \mu_\omega$ *then*

$$\|z-x^*\| \le \kappa\|y-x^*\|.$$

Here $\kappa = \theta_\omega(\rho)$ *and it follows that* $\kappa \in [0,1)$.

PROOF. 1. According to Lemma 6.2.2 the derivative $F'(y)$ is invertible. Set $\varepsilon = \|y-x^*\|$ and let

$$v = y - \Gamma(y)F(y).$$

Then $z = \omega v + (1-\omega)y$. According to Lemma 6.3.4 we have

$$\|z-x^*\|^2 = (1-\omega)\|y-x^*\|^2 + \omega\|v-x^*\|^2 - \omega(1-\omega)\|\Gamma(y)F(y)\|^2 \tag{6.3.12}$$

Using Lemma 6.2.3 we thus obtain

$$\|z-x^*\|^2 \le (1-\omega)\varepsilon^2 + \omega\left[\frac{\beta\gamma\varepsilon^2}{2(1-\beta\gamma\varepsilon)}\right]^2 - \omega(1-\omega)\Delta^2 \tag{6.3.13a}$$

where

$$\Delta = \|\Gamma(y)F(y)\|. \tag{6.3.13b}$$

In order to show that $\|z-x^*\| \le \theta_\omega(\varepsilon)\varepsilon$ it is sufficient to prove that

$$\Delta \ge \frac{(2-\beta\gamma\varepsilon)\varepsilon}{2(1+\beta\gamma\varepsilon)} \qquad \text{(cf. (6.3.8), (6.2.10))}. \tag{6.3.14}$$

2. Suppose

$$(6.3.15) \qquad 0 < 2\Delta\beta\gamma(1+\beta\gamma\varepsilon) < \beta\gamma\varepsilon(2-\beta\gamma\varepsilon).$$

Consider

$$(6.3.16) \qquad \begin{aligned} &P: D(F) \to E, \\ &P(x) = F(x) - F(y) \qquad (x \in D(F)). \end{aligned}$$

Then $P'(x) = F'(x)$ (for all $x \in D(F)$). Consequently, $P'(x^*)$ is invertible and

$$\begin{aligned} [P'(x^*)]^{-1}P(x^*) &= -\Gamma(x^*)F(y) \\ &= \Gamma(x^*)[F'(x^*) - F'(y)]\Gamma(y)F(y) - \Gamma(y)F(y). \end{aligned}$$

Hence, with $\alpha = \Delta(1+\beta\gamma\varepsilon)$, we have

$$(6.3.17) \qquad \|[P'(x^*)]^{-1}P(x^*)\| \le \alpha.$$

Set $h = \alpha\beta\gamma$. According to (6.3.15) we have

$$(6.3.18) \qquad 0 < h < \frac{1}{2}.$$

Since $2h < \beta\gamma\varepsilon(2-\beta\gamma\varepsilon)$ it follows that

$$1 - \sqrt{1-2h} < 1 - \sqrt{1-2\beta\gamma\varepsilon+(\beta\gamma\varepsilon)^2} = \beta\gamma\varepsilon.$$

Consequently,

$$(6.3.19) \qquad \frac{1-\sqrt{1-2h}}{\beta\gamma} < \varepsilon < \hat{\sigma} < \frac{1+\sqrt{1-2h}}{\beta\gamma}.$$

From (6.3.18) and (6.3.19) it follows that Lemma 6.3.3 applies (with $\tau = \hat{\sigma}$). Hence $P(x) = 0$ has a unique solution $\tilde{x}$ in $B(x^*,\hat{\sigma})$ and $\tilde{x} \in \bar{B}(x^*,r_1)$. Since $P(y) = 0$ and $y \in B(x^*,\hat{\sigma})$, it follows that $\tilde{x} = y$, so that $y \in \bar{B}(x^*,r_1)$. Therefore

$$\|x^*-y\| \le r_1 = \frac{1-\sqrt{1-2h}}{\beta\gamma} < \varepsilon \qquad \text{(cf. (6.3.19))}.$$

This yields a contradiction. Consequently (6.3.14) holds. Hence $\|z-x^*\| \le \theta_\omega(\varepsilon)\varepsilon$.

If $\varepsilon \le \rho < \mu_\omega$, then $\theta_\omega(\varepsilon) \le \theta_\omega(\rho) < 1$ (cf. Lemma 6.3.6(i)). This completes the proof. □

PROOF OF THEOREM 6.3.2.

PART A. A number $k \ge 0$ exists for which $\omega_k < 1$. For $n = 0,1,\ldots,k$ set $\theta_n = \theta_{\omega_n}$ and $\pi_n = \pi_{\omega_n}$ where θ_{ω_n} and π_{ω_n} are defined by (6.3.10) and (6.3.11b) respectively. For $n = 0,1,\ldots,k$ let

$$\phi_n\colon [0,\hat{\sigma}] \to [0,\hat{\sigma}],$$

(6.3.20)
$$\phi_n(\varepsilon) = \begin{cases} \theta_n(\varepsilon)\varepsilon & (\text{if } \varepsilon \in [0,\pi_j)), \\ \hat{\sigma} & (\text{if } \varepsilon \in [\pi_j,\hat{\sigma}]). \end{cases}$$

Consider the following function.

$$\psi\colon [0,\hat{\sigma}] \to [0,\hat{\sigma}],$$

(6.3.21a)
$$\psi(\varepsilon) = \varepsilon_{k+1}.$$

Herewith we define the quantities ε_n $(n = 0,1,\ldots,k+1)$ by

(6.3.21b) $\varepsilon_0 = \varepsilon$

and

(6.3.21c) $\varepsilon_{n+1} = \phi_n(\varepsilon_n) \qquad (n = 0,1,\ldots,k).$

The following lemma holds.

LEMMA 6.3.8.

(i) *The function ψ defined by (6.3.21) is isotone on $[0,\hat{\sigma}]$ and continuous on $[0,\hat{\sigma})$.*

(ii) *A constant $\hat{\mu}$ exists such that $\hat{\mu} = \hat{\sigma}$ (if $\hat{\sigma} \le \frac{2}{3\beta\gamma}$), $\frac{2}{3\beta\gamma} < \hat{\mu} < \hat{\sigma}$ (if $\hat{\sigma} > \frac{2}{3\beta\gamma}$) and $\lim_{\varepsilon\uparrow\hat{\mu}} \psi(\varepsilon) < \min\{\hat{\sigma},\frac{2}{3\beta\gamma}\}$.*

(iii) *Let* $\varepsilon \in [0,\hat{\mu})$. *Let* ε_n $(n = 0,1,\ldots,k)$ *be defined by* (6.3.21b,c). *Then*
$\varepsilon_n < \pi_n \le \hat{\sigma}$ $(n = 0,1,\ldots,k)$.

PROOF. From Lemma 6.3.6(i) it follows that ψ is isotone on $[0,\hat{\sigma}]$ and continuous on $[0,\hat{\sigma})$. It is easily verified that $\psi(\frac{2}{3\beta\gamma}) < \frac{2}{3\beta\gamma}$ (if $\hat{\sigma} > \frac{2}{3\beta\gamma}$) and $\lim_{\varepsilon\uparrow\hat{\sigma}} \psi(\varepsilon) < \hat{\sigma}$ (if $\hat{\sigma} \le \frac{2}{3\beta\gamma}$), (cf. Lemma 6.3.6(ii)). Consequently, a constant $\hat{\mu}$ exists such that statement (ii) is true.

Let $\varepsilon \in [0,\hat{\mu})$ and let ε_n $(n = 0,1,\ldots,k)$ be defined by (6.3.21b,c). Suppose $\varepsilon_{n_0} \ge \pi_{n_0}$ for some n_0 with $0 \le n_0 \le k$. Then $\phi_{n_0}(\varepsilon_{n_0}) = \hat{\sigma}$ and it is easily verified that $\psi(\varepsilon) = \hat{\sigma}$. This yields a contradiction. □

PART B. Let $\hat{\mu}$ satisfy statement (ii) of Lemma 6.3.8. Let $F \in \mathcal{F}\langle\sigma,\beta,\gamma\rangle$.

Let $x_0 \in B(x^*,\hat{\mu})$ and set $\varepsilon = \|x_0-x^*\|$. Let ε_n $(n = 0,1,\ldots,k+1)$ be defined by (6.3.21b,c). It is easily verified that for $n = 0,1,\ldots,k$ $\|x -x^*\| \le \varepsilon_n < \pi_n < \hat{\sigma}$, $F'(x_n)$ is invertible, and, with

$$x_{n+1} = x_n - \omega_n\Gamma(x_n)F(x_n)$$

we have

$$\|x_{n+1}-x^*\| \le \theta_n(\|x_n-x^*\|)\|x_n-x^*\| \le \theta_n(\varepsilon_n)\varepsilon_n = \phi_n(\varepsilon_n) = \varepsilon_{n+1}$$

(cf. Lemma 6.3.7, Lemma 6.3.6(i) and Lemma 6.3.8(iii)). Hence

$$\|x_{k+1}-x^*\| \le \psi(\|x_0-x^*\|) < \min\{\hat{\sigma},\frac{2}{3\beta\gamma}\}.$$

PART C. From part B of the proof it follows that a number $\rho \in (0,\frac{2}{3\beta\gamma})$ with $\rho < \hat{\sigma}$ exists such that $x_{k+1} \in B(x^*,\rho)$. Set

$$\hat{\theta} = \sup\{\theta_{\omega_n}(\rho) \mid n \ge k+1\}.$$

Since $\rho < \mu_\omega$ (for all $\omega \in (0,1]$), (cf. Lemma 6.3.6(iii)), it follows that $\theta_\omega(\rho) < 1$ (for all $\omega \in (0,1]$). Further, $\theta_\omega(\rho)$ is continuous in ω (for all $\omega \in [\lambda,1]$ with $\lambda = \inf\{\omega_n \mid n \ge k+1\}$). Hence $\hat{\theta} < 1$. From Lemma 6.3.7 it follows that for $n = k+1,k+2,\ldots$

$$\|y-\omega_n\Gamma(y)F(y)-x^*\| \le \hat{\theta}\|y-x^*\| \qquad \text{(for all } y \in B(x^*,\rho)).$$

Consequently, the statement of the theorem holds. □

PROOF OF THEOREM 6.3.1. Theorem 6.3.1 is a direct consequence of Theorem 6.3.2 and Remark 6.2.1. □

PART II

Let M denote Newton's method and let the method $M_{\bar{\varpi}}$ (with $\bar{\omega} \in \Omega$) be defined by Definition 6.3.1. In part I of this chapter we were able to prove that $r(M_{\bar{\varpi}};\ \mathcal{F}\langle\sigma,\beta,\gamma\rangle) \geq r(M;\ \mathcal{F}\langle\sigma,\beta,\gamma\rangle)$ (for all $\sigma \in (0,\infty]$ and $\beta,\gamma > 0$). However, we did not manage to determine $r(M_{\bar{\varpi}};\ \mathcal{F}\langle\sigma,\beta,\gamma\rangle)$. In this second part of chapter 6 we shall present the subclass $\mathcal{F}\langle\sigma,\alpha\rangle$ ($\sigma \in (0,\infty]$ and $\alpha \in [-\frac{1}{2},\frac{1}{2}]$) of $\mathcal{F}_1$. It will appear that in this case we can determine not only $r(M;\ \mathcal{F}\langle\sigma,\alpha\rangle)$ but also $r(M_{\bar{\varpi}};\ \mathcal{F}\langle\sigma,\alpha\rangle)$.

Before we introduce $\mathcal{F}\langle\sigma,\alpha\rangle$ we observe the following. Let $F \in \mathcal{F}_1$ with $F \in \mathcal{L}_1(E)$ where

$$(6.4.0.1) \qquad \mathcal{L}_1(E) = \{F \mid F\colon E \to E, \text{ and } F(x) \equiv b + L(x) \text{ with } b \in E,\ L \in \mathcal{L}(E)\}.$$

Let M be Newton's method. It is easily verified that, with $G = M(F)$, we have $G(x) = x^*$ (for all $x \in E$). Consequently, for any starting point $x_0 \in E$, Newton's method requires only one step to solve $F(x) = 0$. More generally speaking, let $\sigma \in (0,\infty]$, let $F \in \mathcal{F}_1$ and suppose that the following holds:

$$(6.4.0.2) \qquad \begin{aligned} &B(x^*,\sigma) \subset D(F), \quad F'(x) \text{ is invertible and} \\ &(x-x^*,\Gamma(x)F(x)) = \tfrac{1}{2}\{\|x-x^*\|^2 + \|\Gamma(x)F(x)\|^2\} \quad (\text{for all } x \in B(x^*,\sigma)). \end{aligned}$$

A little calculation shows that (6.4.0.2) implies $\|G(x)-x^*\|^2 = 0$ (for all $x \in B(x^*,\sigma)$). Hence Newton's method requires only one step to solve $F(x) = 0$ whenever $x_0 \in B(x^*,\sigma)$. The subclass $\mathcal{F}\langle\sigma,\frac{1}{2}\rangle$ consists of all functions F for which (6.4.0.2) is true. Further $\mathcal{F}\langle\sigma,\frac{1}{2}\rangle \subset \mathcal{F}\langle\sigma,\alpha_2\rangle \subset \mathcal{F}\langle\sigma,\alpha_1\rangle \subset \mathcal{F}\langle\sigma,-\frac{1}{2}\rangle = \mathcal{F}_1$ (for all $\alpha_1,\alpha_2 \in [-\frac{1}{2},\frac{1}{2}]$ with $\alpha_1 < \alpha_2$). See also Fig. 6.4.1 on page 99. Consequently, $r(M;\ \mathcal{F}\langle\sigma,\cdot\rangle)$ is isotone on $[-\frac{1}{2},\frac{1}{2}]$, $r(M;\ \mathcal{F}\langle\sigma,-\frac{1}{2}\rangle) = 0$ (cf. (6.1.2) and Theorem 6.2.1) and $r(M;\ \mathcal{F}\langle\sigma,\frac{1}{2}\rangle) = \sigma$.

6.4. THE CLASS $F\langle\sigma,\alpha\rangle$

Before we present the class $F\langle\sigma,\alpha\rangle$ we give two lemmata.

LEMMA 6.4.1. *Let* $F \in F_1$. *Then for any* $x \in D(F)$ *for which* $F'(x)$ *is invertible we have*

$$(6.4.1) \qquad -\tfrac{1}{2}\{\|x-x^*\|^2+\|\Gamma(x)F(x)\|^2\} \le (x-x^*,\Gamma(x)F(x)) \le \tfrac{1}{2}\{\|x-x^*\|^2+\|\Gamma(x)F(x)\|^2\}.$$

Further, positive numbers ρ *and* δ *exist such that* $B(x^*,\rho) \subset D(F)$, *and* $F'(x)$ *is invertible and*

$$(6.4.2) \qquad (x-x^*,\Gamma(x)F(x)) \ge [\tfrac{1}{2}-\delta\|x-x^*\|^2]\{\|x-x^*\|^2 + \|\Gamma(x)F(x)\|^2\}$$

$$\text{(for all } x \in B(x^*,\rho)).$$

PROOF. 1. Let $x \in D(F)$ and suppose $F'(x)$ is invertible. Then

$$(x-x^*,\Gamma(x)F(x)) - \tfrac{1}{2}\{\|x-x^*\| - \|\Gamma(x)F(x)\|\}^2 \le (x-x^*,\Gamma(x)F(x))$$

$$\le (x-x^*,\Gamma(x)F(x)) + \tfrac{1}{2}\{\|x-x^*\| - \|\Gamma(x)F(x)\|\}^2.$$

From Schwarz's inequality (see [KANTOROWITSCH & AKILOW, 1964; section 7.1 (II)]) it follows that

$$-\|x-x^*\|\|\Gamma(x)F(x)\| \le (x-x^*,\Gamma(x)F(x)) \le \|x-x^*\|\|\Gamma(x)F(x)\|.$$

Hence (6.4.1) holds.

2. Let M denote Newton's method. From Lemma 6.2.3 and (6.1.2) it follows that [M,F] is quadratically convergent. Hence $\rho,\delta_1 > 0$ exist such that $B(x^*,\rho) \subset D(F)$, and $F'(x)$ is invertible and

$$\|x-\Gamma(x)F(x)-x^*\|^2 \le \delta_1^2\|x-x^*\|^4 \qquad \text{(for all } x \in B(x^*,\rho)).$$

Consequently,

$$2(x-x^*,\Gamma(x)F(x)) \ge \|x-x^*\|^2 + \|\Gamma(x)F(x)\|^2 - \delta_1^2\|x-x^*\|^4$$

$$\text{(for all } x \in B(x^*,\rho)).$$

Therefore, with $\delta = \frac{1}{2}\delta_1^2$, (6.4.2) is true. □

Let

$$\chi\colon \mathcal{F}_1 \times (0,\infty] \times \mathbb{R} \to \mathbb{R},$$

$$(6.4.3)\qquad \chi(F;\sigma,\alpha) = \begin{cases} \inf\limits_{0<\|x-x^*\|<\sigma} \{\dfrac{(x-x^*,\Gamma(x)F(x))}{\|x-x^*\|^2+\|\Gamma(x)F(x)\|^2} - \alpha\} \\ \qquad (\text{if } B(x^*,\sigma) \subset D(F) \text{ and } F'(x) \text{ is invertible} \\ \qquad \text{for all } x \in B(x^*,\sigma)), \\ -\frac{1}{2} - \alpha \quad (\text{otherwise}). \end{cases}$$

The following lemma holds.

<u>LEMMA 6.4.2</u>. *For any* $F \in \mathcal{F}_1$ *we have*

$$(6.4.4)\qquad -\frac{1}{2} - \alpha \le \chi(F;\sigma,\alpha) \le \frac{1}{2} - \alpha \qquad (\text{for all } \sigma \in (0,\infty] \text{ and } \alpha \in \mathbb{R})$$

and

$$(6.4.5)\qquad \lim_{\sigma\downarrow 0} \chi(F;\sigma,\alpha) = \frac{1}{2} - \alpha \qquad (\text{for all } \alpha \in \mathbb{R}).$$

<u>PROOF</u>. This lemma is a direct consequence of Lemma 6.4.1. □

Let $F \in \mathcal{F}_1$ and $\sigma \in (0,\infty]$. From (6.4.4) it follows that

$$(6.4.6)\qquad \chi(F;\sigma,\alpha) \ge 0 \qquad (\text{for all } \alpha \le -\tfrac{1}{2}),$$

$$(6.4.7)\qquad \chi(F;\sigma,\alpha) < 0 \qquad (\text{for all } \alpha > \tfrac{1}{2}).$$

For $\sigma \in (0,\infty]$ and $\alpha \in [-\frac{1}{2},\frac{1}{2}]$ we define

$$(6.4.8)\qquad \mathcal{F}\langle\sigma,\alpha\rangle = \{F \mid F \in \mathcal{F}_1;\ \chi(F;\sigma,\alpha) \ge 0\}.$$

When $\sigma \in (0,\infty]$ and $\alpha \in (-\frac{1}{2},\frac{1}{2}]$ then the following relation holds.

(6.4.9) $\mathcal{F}\langle\sigma,\alpha\rangle = \{F \mid F \in \mathcal{F}_1;\ B(x^*,\sigma) \subset D(F);\ F'(x)$ is invertible and $(x-x^*,\Gamma(x)F(x)) \geq \alpha\{\|x-x^*\|^2+\|\Gamma(x)F(x)\|^2\}$ (for all $x \in B(x^*,\sigma))\}$.

Further

(6.4.10) $\mathcal{F}\langle\sigma,-\frac{1}{2}\rangle = \mathcal{F}_1$ (for all $\sigma \in (0,\infty]$).

Let $\sigma \in (0,\infty]$. We notice that

(6.4.11) $\mathcal{F}\langle\sigma,\alpha_1\rangle \supset \mathcal{F}\langle\sigma,\alpha_2\rangle$ (for all numbers α_1,α_2 with $-\frac{1}{2} \leq \alpha_1 \leq \alpha_2 \leq \frac{1}{2}$).

Let $\mathcal{L}_1(E)$ be defined by (6.4.0.1). It is easily verified that

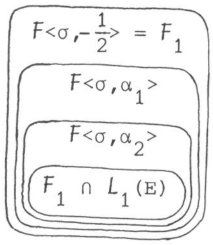
(6.4.12) $\mathcal{F}_1 \cap \mathcal{L}_1(E) \subset \mathcal{F}\langle\sigma,\frac{1}{2}\rangle$ (for all $\sigma \in (0,\infty]$).

Thus for each $\sigma \in (0,\infty]$ and $\alpha \in [-\frac{1}{2},\frac{1}{2}]$, the set $\mathcal{F}\langle\sigma,\alpha\rangle$ is not empty.

Let $\sigma \in (0,\infty)$ and $\alpha \in (-\frac{1}{2},\frac{1}{2}]$. Let $F \in \mathcal{F}\langle\sigma,\alpha\rangle$ and let $F_{\sigma/2}$ denote the restriction of F to $B(x^*,\sigma/2)$. Then $F_{\sigma/2} \in \mathcal{F}_1$ but $F_{\sigma/2} \notin \mathcal{F}\langle\sigma,\alpha\rangle$ (cf. (6.4.9)). Consequently

(6.4.13) $\emptyset \not\subset \mathcal{F}\langle\sigma,\alpha\rangle \not\subset \mathcal{F}_1$ (for all $\sigma \in (0,\infty]$ and $\alpha \in (-\frac{1}{2},\frac{1}{2}]$).

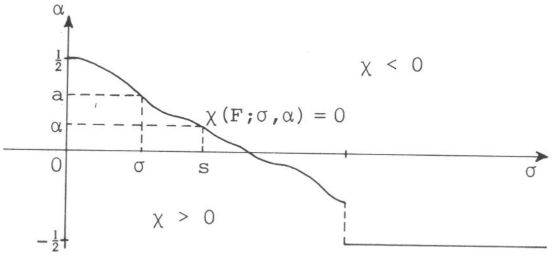
Fig. 6.4.1 illustrates the properties of $\mathcal{F}\langle\sigma,\alpha\rangle$ listed above.



$\mathcal{F}\langle\sigma,-\frac{1}{2}\rangle = \mathcal{F}_1$
$\mathcal{F}\langle\sigma,\alpha_1\rangle$
$\mathcal{F}\langle\sigma,\alpha_2\rangle$
$\mathcal{F}_1 \cap \mathcal{L}_1(E)$

$\sigma \in (0,\infty];\ -\frac{1}{2} < \alpha_1 < \alpha_2 \leq \frac{1}{2}$.

Fig. 6.4.1



$F \in \mathcal{F}_1$.

Fig. 6.4.2

REMARK 6.4.1.

1. Let $\sigma \in (0,\infty]$. It can be shown that $\mathcal{F}\langle\sigma,\alpha_1\rangle \neq \mathcal{F}\langle\sigma,\alpha_2\rangle$ whenever $\alpha_1,\alpha_2 \in [0,\frac{1}{2}]$ and $\alpha_1 \neq \alpha_2$ (this is a consequence of the Lemmata 6.6.8 and 6.6.11).

2. If $E = \mathbb{R}$, then it is easily verified that $\mathcal{F}\langle\sigma,\alpha\rangle = \mathcal{F}\langle\sigma,0\rangle$ whenever $\alpha \in (-\frac{1}{2},0]$ and $\sigma \in (0,\infty]$.

If $E \neq \mathbb{R}$, then this need not be the case. We shall illustrate this for the case $E = \mathbb{R}^2$ with innerproduct $(x,y) \equiv \xi_1\eta_1 + \xi_2\eta_2$ (for $x = (\xi_1,\xi_2)$, $y = (\eta_1,\eta_2) \in \mathbb{R}^2$). Let

$$\text{(6.4.14)}\qquad \begin{aligned} &F\colon \mathbb{R}^2 \to \mathbb{R}^2, \\ &F(x) = (\xi_2^4(1+\xi_1^2)^4 + \xi_1(1+\xi_1^2), \xi_2(1+\xi_1^2)) \quad (\text{for all } x \in \mathbb{R}^2, \\ &\hspace{20em} \text{where } x = (\xi_1,\xi_2)). \end{aligned}$$

It is easily verified that $F(0) = 0$, $F \in \mathcal{F}_1$ and that $F'(x)$ is invertible (for all $x \in \mathbb{R}^2$). With $x_0 = (2,1)$ it can be shown that $-\frac{1}{2} < \alpha_0 < 0$, where

$$\alpha_0 = (x_0,\Gamma(x_0)F(x_0))/\{\|x_0\|^2 + \|\Gamma(x_0)F(x_0)\|^2\}.$$

Consequently, since $\|x_0\| = \sqrt{5}$, we have $\chi(F;\sqrt{5},\alpha_0) \leq 0$. From Lemma 6.4.2 it follows that $\lim_{\sigma\downarrow 0} \chi(F;\sigma,\alpha_0) = \frac{1}{2} - \alpha_0 > 0$. It can be shown that in this case $\chi(F;\sigma,\alpha_0)$ is continuous on $(0,\infty]$. Consequently, a number $\sigma_0 \in (0,\sqrt{5}]$ exists such that $F \in \mathcal{F}\langle\sigma_0,\alpha_0\rangle$ and $F \notin \mathcal{F}\langle\sigma_0,0\rangle$. Hence $\mathcal{F}\langle\sigma_0,\alpha_0\rangle \neq \mathcal{F}\langle\sigma_0,0\rangle$.

Even if E is an arbitrary Hilbertspace with $E \neq \mathbb{R}$, using the above example it can be shown that numbers α and σ with $-\frac{1}{2} < \alpha < 0$ and $\sigma > 0$ exist such that $\mathcal{F}\langle\sigma,\alpha\rangle \neq \mathcal{F}\langle\sigma,0\rangle$. □

Let $\sigma \in (0,\infty]$ and $F \in \mathcal{F}_1$. Notice that (cf. (6.4.3) and (6.4.4))

$$\text{(6.4.15)}\qquad \begin{aligned} &\chi(F;\sigma,\cdot) \text{ is strictly antitone and continuous on } [-\tfrac{1}{2},\tfrac{1}{2}], \text{ and} \\ &\qquad\qquad \chi(F;\sigma,-\tfrac{1}{2}) \geq 0. \end{aligned}$$

We define

$$\text{(6.4.16)}\qquad \begin{aligned} &a\colon \mathcal{F}_1 \times (0,\infty] \to [-\tfrac{1}{2},\tfrac{1}{2}], \\ &a(F;\sigma) = \sup\{\alpha \mid \alpha \in [-\tfrac{1}{2},\tfrac{1}{2}];\ \chi(F;\sigma,\alpha) \geq 0\} \quad (F \in \mathcal{F}_1,\ \sigma \in (0,\infty]). \end{aligned}$$

Let $F \in F_1$. It is easily verified that (cf. Fig. 6.4.2 and (6.4.5))

(6.4.17) $\quad a(F;\cdot)$ is antitone on $(0,\infty]$ and $\lim\limits_{\sigma\downarrow 0} a(F;\sigma) = \frac{1}{2}$.

From (6.4.15) and (6.4.16) it follows that

(6.4.18) $\quad F \in F\langle\sigma,a(F;\sigma)\rangle \qquad$ (for all $\sigma \in (0,\infty]$).

Let $\alpha \in [-\frac{1}{2},\frac{1}{2})$ and $F \in F_1$. Notice that (cf. (6.4.3) and (6.4.5))

(6.4.19) $\quad \chi(F;\cdot,\alpha)$ is antitone on $(0,\infty]$, left-continuous on $(0,\infty)$, and

$$\lim_{\sigma\downarrow 0} \chi(F;\sigma,\alpha) = \frac{1}{2} - \alpha > 0.$$

We define

$$\text{(6.4.20)} \quad \begin{aligned} &s\colon F_1 \times [-\tfrac{1}{2},\tfrac{1}{2}) \to (0,\infty], \\ &s(F;\alpha) = \sup\{\sigma \mid \sigma \in (0,\infty];\ \chi(F;\sigma,\alpha) \geq 0\} \qquad (F \in F_1,\ \alpha \in [-\tfrac{1}{2},\tfrac{1}{2})). \end{aligned}$$

Let $F \in F_1$. It is easily verified that (cf. Fig. 6.4.2)

(6.4.21) $\quad s(F;\cdot)$ is antitone on $[-\frac{1}{2},\frac{1}{2})$.

From (6.4.19) and (6.4.20) it follows that

(6.4.22) $\quad F \in F\langle s(F;\alpha),\alpha\rangle \qquad$ (for all $\alpha \in [-\frac{1}{2},\frac{1}{2})$).

6.5. THE RADIUS OF CONVERGENCE OF NEWTON'S METHOD WITH RESPECT TO $F\langle\sigma,\alpha\rangle$

Let $\sigma \in (0,\infty]$ and $\alpha \in [-\frac{1}{2},\frac{1}{2}]$.

THEOREM 6.5.1. *Let M be the Newton's method. Then*

$$r(M;\ F\langle\sigma,\alpha\rangle) = \begin{cases} 0 & (\text{if } \alpha \leq \frac{2}{5}), \\ \sigma & (\text{if } \alpha > \frac{2}{5}). \end{cases}$$

We shall prove this theorem in the next subsection.

6.5.1. Proof of Theorem 6.5.1

First, we introduce an auxiliary function, which will be used frequently in this chapter. Let

$$(6.5.1)\qquad \begin{aligned} &\theta\colon (0,\tfrac{1}{2}] \times \{1,2\} \to \mathbb{R},\\ &\theta(\xi,k) = \frac{1+(-1)^k\sqrt{1-4\xi^2}}{2\xi} \qquad ((\xi,k) \in D(\theta)). \end{aligned}$$

Direct computation shows that the following relations hold.

$$(6.5.2)\qquad \xi\{1+\theta(\xi,k)^2\} = \theta(\xi,k) \qquad (k = 1,2)$$

and

$$(6,5,3)\qquad \theta(\xi,1) = [\theta(\xi,2)]^{-1} \qquad (\text{for all } \xi \in (0,\tfrac{1}{2}]).$$

The following relations, that are easily verified, will be used subsequently.

$$(6.5.4)\qquad \begin{aligned} &\theta(\cdot,2) \text{ is strictly antitone on } (0,\tfrac{1}{2}],\ \lim_{\xi\downarrow 0}\theta(\xi,2) = \infty,\\ &\theta(\tfrac{2}{5},2) = 2,\ \theta(\tfrac{1}{2},2) = 1. \end{aligned}$$

$$(6.5.5)\qquad \theta(\cdot,1) \text{ is strictly isotone on } (0,\tfrac{1}{2}],\ \theta(\tfrac{2}{5},1) = \tfrac{1}{2},\ \theta(\tfrac{1}{2},1) = 1.$$

The following lemma holds.

LEMMA 6.5.2. *Suppose* $\alpha \in (0,\frac{1}{2}]$. *Let* $F \in F\langle\sigma,\alpha\rangle$. *Then* $B(x^*,\sigma) \subset D(F)$, *and* $F'(x)$ *is invertible and*

$$(6.5.6)\qquad \theta(\alpha,1)\|x-x^*\| \le \|\Gamma(x)F(x)\| \le \theta(\alpha,2)\|x-x^*\| \qquad (\text{for all } x \in B(x^*,\sigma)).$$

PROOF. In view of (6.4.9) we have $B(x^*,\sigma) \subset D(F)$. Let $x \in B(x^*,\sigma)$. Then $F'(x)$ is invertible and (cf. (6.4.9))

$$(x-x^*,\Gamma(x)F(x)) \ge \alpha\{\|x-x^*\|^2 + \|\Gamma(x)F(x)\|^2\}.$$

Using Schwarz's inequality, we obtain

$$\alpha\|\Gamma(x)F(x)\|^2 - \|x-x^*\|\|\Gamma(x)F(x)\| + \alpha\|x-x^*\|^2 \leq 0.$$

This implies (6.5.6). □

Let

$$(6.5.7)\qquad \begin{aligned} &\psi\colon (0,\tfrac{1}{2}] \to \mathbb{R}, \\ &\psi(\xi) = \theta(\xi,2) - 1 \qquad (\xi \in (0,\tfrac{1}{2}]). \end{aligned}$$

From (6.5.4) it follows that

(6.5.8) ψ is strictly antitone on $(0,\frac{1}{2}]$ and $\psi(\frac{2}{5}) = 1$.

The following lemma holds.

LEMMA 6.5.3. *Suppose* $\alpha \in (0,\frac{1}{2}]$. *Let* $F \in F\langle\sigma,\alpha\rangle$. *Set* $G = M(F)$ *where* M *is Newton's method. Then* $B(x^*,\sigma) \subset D(G)$ *and*

$$(6.5.9)\qquad \|G(x)-x^*\| \leq \psi(\alpha)\|x-x^*\| \qquad (\text{for all } x \in B(x^*,\sigma)).$$

PROOF. In view of (6.4.9) we have $B(x^*,\sigma) \subset D(G)$. Let $x \in B(x^*,\sigma)$. It follows that (cf. (6.4.9))

$$\begin{aligned} \|G(x)-x^*\|^2 &= \|x-\Gamma(x)F(x)-x^*\|^2 \\ &= \|x-x^*\|^2 + \|\Gamma(x)F(x)\|^2 - 2(x-x^*,\Gamma(x)F(x)) \\ &\leq (1-2\alpha)\{\|x-x^*\|^2 + \|\Gamma(x)F(x)\|^2\}. \end{aligned}$$

Hence, using Lemma 6.5.2,

$$(6.5.10)\qquad \|G(x)-x^*\|^2 \leq (1-2\alpha)\{1+\theta(\alpha,2)^2\}\|x-x^*\|^2.$$

Since $[\psi(\alpha)]^2 = 1 + \theta(\alpha,2)^2 - 2\theta(\alpha,2)$, from (6.5.2) and (6.5.10) it follows that (6.5.9) is true. □

A consequence of Lemma 6.5.3 is the following lemma.

LEMMA 6.5.4. *Let* M *be the Newton's method. Then*

$$r(M;\mathcal{F}<\sigma,\alpha>) \geq \begin{cases} 0 & (\text{if } \alpha \in [-\frac{1}{2},\frac{2}{5}]), \\ \sigma & (\text{if } \alpha \in (\frac{2}{5},\frac{1}{2}]). \end{cases}$$

PROOF. Suppose $\alpha \in (\frac{2}{5},\frac{1}{2}]$ and let $F \in \mathcal{F}<\sigma,\alpha>$. Notice that ψ is strictly antitone on $(0,\frac{1}{2}]$ and that $\psi(\frac{2}{5}) = 1$ (cf. (6.5.8)). Hence, in view of Lemma 6.5.3, we have $B(x^*,\sigma) \subset S(M,F)$. This proves the lemma. □

REMARK 6.5.1. Suppose $E = \mathbb{R}$ and suppose $\alpha \in (0,\frac{2}{5}]$. In view of Lemma 6.5.4 and Remark 6.2.1 the proof of Theorem 6.5.1 is completed if we can show that for all $\varepsilon > 0$ there exist an $f \in \mathcal{F}<\sigma,\alpha>$ and an $x_0 \in D(f)$ with $\|x - x^*\| < \varepsilon$, such that $x_0 \notin S(M,f)$. Consider

$$f: \mathbb{R} \to \mathbb{R},$$

$$f(\xi) = \begin{cases} \xi^{\frac{1}{\theta(\alpha,2)}} & (\text{if } \xi > 0), \\ 0 & (\text{if } \xi = 0), \\ -f(-\xi) & (\text{if } \xi < 0). \end{cases}$$

It is easily verified that f is continuous on $\mathbb{R}$, $\xi = 0$ is the unique solution of $f(\xi) = 0$, and, for all $\xi \neq 0$, f is twice continuously differentiable and $f'(\xi) \neq 0$. Further

$$[f'(\xi)]^{-1}f(\xi) = \theta(\alpha,2)\xi \qquad (\text{for all } \xi \neq 0).$$

Consequently (cf. (6.5.2)),

$$\xi\ [f'(\xi)]^{-1}f(\xi) = \xi^2\theta(\alpha,2) = \xi^2\alpha\{1+\theta(\alpha,2)^2\}$$

$$= \alpha[\xi^2 + \{[f'(\xi)]^{-1}f(\xi)\}^2] \qquad (\text{for all } \xi \neq 0).$$

If $\xi \neq 0$ then (cf. (6.5.8))

$$\xi - [f'(\xi)]^{-1}f(\xi) = [1 - \theta(\alpha,2)]\xi = -\psi(\alpha)\xi \le -\xi.$$

It is easily verified that $\xi \notin S(M,F)$.

However, since $f \notin F\langle\sigma,\alpha\rangle$ (f is not differentiable at $\xi = 0$), with this example we have *not* completed the proof of Theorem 6.5.1. With the next lemma, however, we can complete the proof of Theorem 6.5.1 by showing that an $f \in F\langle\sigma,\alpha\rangle$ with the desired properties exists, which is not only twice but infinitely differentiable on $D(f)$. □

LEMMA 6.5.5. *Let* M *be the Newton's method. Suppose* $\alpha \in [-\frac{1}{2},\frac{2}{5}]$. *Then for any* $\varepsilon > 0$ *there exists an* $F \in F\langle\sigma,\alpha\rangle$ *which is infinitely differentiable on* $D(F)$ *for which* $r(M,F) \le \varepsilon$.

We shall prove this lemma, with the function f of Remark 6.5.1 in mind, in a more general setting in subsection 6.6.2 (cf. p.119).

6.6. THE RADIUS OF CONVERGENCE OF $M_{\bar{\omega}}$ WITH RESPECT TO $F\langle\sigma,\alpha\rangle$

Let Ω be defined in (6.3.5); for $\bar{\omega} \in \Omega$ the iterative method $M_{\bar{\omega}}$ is defined by Definition 6.3.1.

In this section we give a theorem concerning the radius of convergence of $M_{\bar{\omega}}$ with respect to $F\langle\sigma,\alpha\rangle$, where $\sigma \in (0,\infty]$ and $\alpha \in [-\frac{1}{2},\frac{1}{2}]$. To that end let

$$\psi_1 : (0,1] \times (0,\tfrac{1}{2}] \to \mathbb{R},$$

(6.6.1)
$$\psi_1(\omega,\xi) = \begin{cases} 1 - \omega\theta(\xi,1) & (\text{if } 0 < \omega \le 2\xi \le 1), \\ \omega\theta(\xi,2) - 1 & (\text{if } 0 < 2\xi < \omega \le 1). \end{cases}$$

Here θ is defined in (6.5.1). Let $\omega \in (0,1]$. In view of (6.5.2 - 5) the following relations (6.6.2a,b,c) hold.

(6.6.2a) $\psi_1(\omega,\cdot)$ is continuous and strictly antitone on $(0,\frac{1}{2}]$,

$$\lim_{\xi \downarrow 0} \psi_1(\omega,\xi) = \infty, \quad \psi_1(\omega,\tfrac{1}{2}) = 1 - \omega.$$

(6.6.2b) If $\omega = 1$ then $\psi_1(\omega,\frac{2}{5}) = 1$, if $\omega \in (0,1)$ then $\psi_1(\omega,\xi) < 1$

(for all $\xi \in [\frac{2}{5},\frac{1}{2}]$).

(6.6.2c) $\quad \psi_1(\omega,\xi) > \psi_1(2\xi,\xi) = \sqrt{1-4\xi^2} > 0 \quad$ (for all $\xi \in (0,\frac{1}{2}]$ with $2\xi \neq \omega$).

Consider the functions

$$\phi_1\colon \Omega \times (0,\tfrac{1}{2}] \to \mathbb{R},$$

(6.6.3)

$$\phi_1(\bar{\omega},\xi) = \max\{1, \prod_{j=0}^{n} \psi_1(\omega_j,\xi) \ (n = 0,1,\ldots,N-1)\}$$

$$(\xi \in (0,\tfrac{1}{2}] \text{ and } \bar{\omega} \in \Omega \text{ with } \bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N))$$

and

$$\phi_2\colon \Omega \times (0,\tfrac{1}{2}] \to \mathbb{R},$$

(6.6.4)

$$\phi_2(\bar{\omega},\xi) = \prod_{j=0}^{N} \psi_1(\omega_j,\xi)$$

$$(\xi \in (0,\tfrac{1}{2}] \text{ and } \bar{\omega} \in \Omega \text{ with } \bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N)).$$

Let $\bar{\omega} \in \Omega$. Note that $\omega_N = 1$. In view of (6.6.2a,b) we have

(6.6.5) $\quad \phi_1(\bar{\omega},\cdot)$ is continuous and antitone on $(0,\frac{1}{2}]$, $\phi_1(\bar{\omega},\xi) = 1$

(for all $\xi \in [\hat{\alpha},\frac{1}{2}]$), where $0 < \hat{\alpha} < \frac{2}{5}$,

and

(6.6.6) $\quad \phi_2(\bar{\omega},\cdot)$ is continuous and strictly antitone on $(0,\frac{1}{2}]$,

$\lim_{\xi \downarrow 0} \phi_2(\bar{\omega},\xi) = \infty$ and $\phi_2(\bar{\omega},\frac{2}{5}) < 1$.

The following theorem holds.

<u>THEOREM 6.6.1</u>. *Let $\sigma \in (0,\infty]$, $\alpha \in [-\frac{1}{2},\frac{1}{2}]$ and $\bar{\omega} \in \Omega$. Then the equation $\phi_2(\bar{\omega},\xi) = 1$ has a unique solution $\xi = \alpha_{\bar{\omega}}$ in $(0,\frac{1}{2}]$ and $\alpha_{\bar{\omega}} \in (0,\frac{2}{5})$. We have*

$$(6.6.7) \qquad r(M_{\varpi};\ F\langle\sigma,\alpha\rangle) = \begin{cases} \dfrac{\sigma}{\phi_1(\bar{\omega},\alpha)} & (\text{if } \alpha_{\varpi} < \alpha \le \frac{1}{2}), \\ \delta & (\text{if } \alpha = \alpha_{\varpi}), \\ 0 & (\text{if } -\frac{1}{2} \le \alpha < \alpha_{\varpi}). \end{cases}$$

Here δ satisfies

$$(6.6.8) \qquad 0 \le \delta \le \frac{\sigma}{\phi_1(\bar{\omega},\alpha_{\varpi})} .$$

The proof of Theorem 6.6.1 will be given in the next two subsections. It depends on the following principle. Let $\sigma \in (0,\infty]$, $\alpha \in (0,\frac{1}{2}]$, $\bar{\omega} \in \Omega$ and $\omega \in (0,1]$. For $F \in F\langle\sigma,\alpha\rangle$ and $y \in B(x^*,\sigma)$ set

$$z = y - \omega\Gamma(y)F(y).$$

It can be shown (see Lemma 6.6.5) that

$$\|z-x^*\| \le \psi_1(\omega,\alpha)\|y-x^*\| .$$

From this relation it can be shown that

$$r(M_{\varpi};\ F\langle\sigma,\alpha\rangle) \ge \begin{cases} \dfrac{\sigma}{\phi_1(\bar{\omega},\alpha)} & (\text{if } \phi_2(\bar{\omega},\alpha) < 1), \\ 0 & (\text{otherwise}). \end{cases}$$

Using specially chosen iterative processes $[M_{\varpi},F]$ it can be shown that (6.6.7) holds.

We give some illustrations.

1. Graph of ψ_1; $\xi \in (0,\frac{1}{2}]$ is fixed.

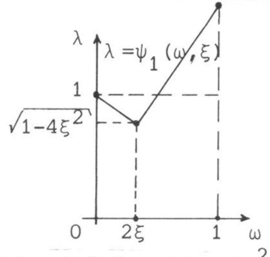


Fig. 6.6.1a $(\xi \in (0,\frac{2}{5}))$

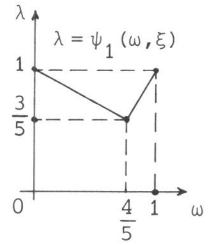


Fig. 6.6.1b $(\xi = \frac{2}{5})$

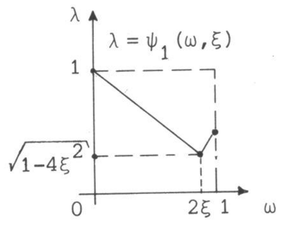


Fig. 6.6.1c $(\xi \in (\frac{2}{5},\frac{1}{2}))$

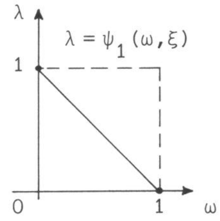
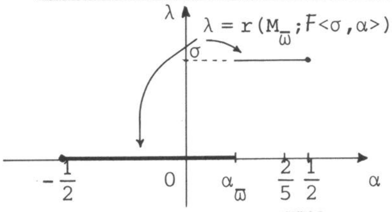


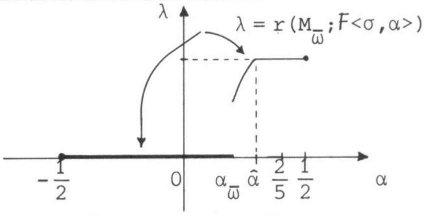
Fig. 6.6.1d $(\xi = \frac{1}{2})$

2. Graph of $r(M_{\bar{\omega}};\ F\langle\sigma,\alpha\rangle)$; $\sigma \in (0,\infty]$ and $\bar{\omega} \in \Omega$ are fixed.



$\phi_2(\bar{\omega},\alpha_{\bar{\omega}}) = 1$ $\phi_1(\bar{\omega},\alpha_{\bar{\omega}}) = 1$

Fig. 6.6.2a



$\phi_2(\bar{\omega},\alpha_{\bar{\omega}}) = 1$ $\phi_1(\bar{\omega},\alpha_{\bar{\omega}}) > 1$

Fig. 6.6.2b

3. Graph of $r(M;\ F\langle\sigma,\alpha\rangle)$, where M is Newton's method; $\sigma \in (0,\infty]$ is fixed.

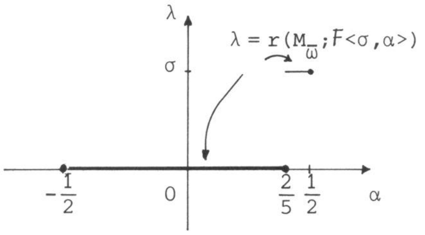


Fig. 6.6.3

4. Graph of $r(M_{\bar{\omega}};\ F\langle\sigma,\alpha\rangle)$; $\alpha \in [-\frac{1}{2},\frac{1}{2}]$ and $\bar{\omega} \in \Omega$ are fixed $(\alpha \neq \alpha_{\bar{\omega}})$.

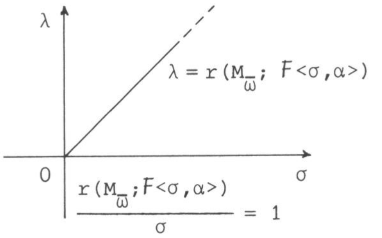


Fig. 6.6.4a

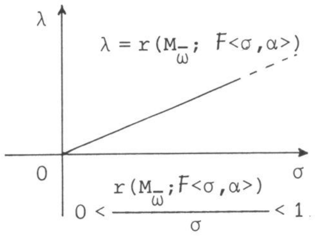


Fig. 6.6.4b

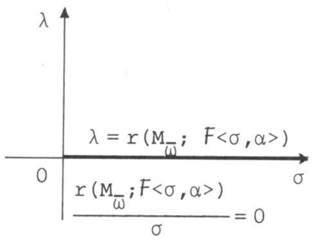


Fig. 6.6.4c

We give some corollaries of Theorem 6.6.1.

<u>COROLLARY 6.6.2</u>. *Let $\sigma \in (0,\infty]$ and $\bar\omega \in \Omega$. Let M be Newton's method. Then*

$$(6.6.9)\qquad r(M_{\bar\omega};\ \mathcal{F}\langle\sigma,\alpha\rangle) \geq r(M;\ \mathcal{F}\langle\sigma,\alpha\rangle) \qquad (\text{for all } \alpha \in [-\tfrac{1}{2},\tfrac{1}{2}]).$$

More specifically,

$$(6.6.10)\qquad r(M_{\bar\omega};\ \mathcal{F}\langle\sigma,\alpha\rangle) = r(M;\ \mathcal{F}\langle\sigma,\alpha\rangle) = \sigma \qquad (\text{for all } \alpha \in (\tfrac{2}{5},\tfrac{1}{2}])$$

and there exist constants $\alpha_{\bar\omega}$ and $\hat\alpha_{\bar\omega}$ with $0 < \alpha_{\bar\omega} \leq \hat\alpha_{\bar\omega} < \frac{2}{5}$ such that

$$(6.6.11a)\qquad 0 < r(M_{\bar\omega};\ \mathcal{F}\langle\sigma,\alpha\rangle) \leq \sigma \quad \text{and} \quad r(M;\ \mathcal{F}\langle\sigma,\alpha\rangle) = 0$$
$$(\text{for all } \alpha \in (\alpha_{\bar\omega},\tfrac{2}{5}])$$

and

$$(6.6.11b)\qquad r(M_{\bar\omega};\ \mathcal{F}\langle\sigma,\alpha\rangle) = \sigma \qquad (\text{for all } \alpha \in (\hat\alpha_{\bar\omega},\tfrac{2}{5}]).$$

<u>PROOF</u>. According to (6.6.5) and (6.6.6) it follows that numbers $\alpha_{\bar\omega},\tilde\alpha \in (0,\frac{2}{5})$ exist such that $\phi_1(\bar\omega,\alpha) = 1$ (for all $\alpha \in [\tilde\alpha,\frac{1}{2}]$), $\phi_2(\bar\omega,\alpha) < 1$ (for all $\alpha \in [\alpha_{\bar\omega},\frac{1}{2}]$) and $\phi_2(\bar\omega,\alpha_{\bar\omega}) = 1$. Let $\hat\alpha_{\bar\omega}' = \max\{\tilde\alpha,\alpha_{\bar\omega}\}$. From Theorem 6.6.1 and Theorem 6.5.1 it follows that the statement is true. □

Let $\sigma \in (0,\infty]$, $\alpha \in [-\frac{1}{2},\frac{1}{2}]$ and $\bar\omega \in \Omega$. Corollary 6.6.2 shows that $M_{\bar\omega}$ has a radius of convergence with respect to $\mathcal{F}\langle\sigma,\alpha\rangle$ that is not smaller and, for certain values of α is even greater, than the radius of convergence of Newton's method. See also Remark 6.3.1.

Corollary 6.6.2 can also be interpreted as follows. Let $F \in \mathcal{F}_1$. Then $F \in \mathcal{F}\langle s(F;\alpha),\alpha\rangle$ (for all $\alpha \in [-\frac{1}{2},\frac{1}{2})$) (cf. (6.4.20), (6.4.22)). Consequently, $r(M,F) \geq \rho_0 = \sup\{s(F;\alpha)\,|\,\alpha \in (\frac{2}{5},\frac{1}{2}]\}$ and $r(M_{\bar\omega},F) \geq \rho_{\bar\omega} = \sup\{s(F;\alpha)\,|\,\alpha \in (\hat\alpha_{\bar\omega},\frac{1}{2}]\}$. In view of (6.4.21) it follows that $\rho_{\bar\omega} \geq \rho_0$. Corollary 6.6.2 guarantees that $[M,F]$ generates a sequence $\{x_k\}$ that converges to x^* whenever the starting point $x_0 \in B(x^*,\rho_0)$. Furthermore, $[M_{\bar\omega},F]$ generates a sequence $\{x_k\}$ that converges to x^* whenever $x_0 \in B(x^*,\rho_{\bar\omega}) \supset B(x^*,\rho_0)$.

Let $N_0 \in \mathbb{N}$ be given. Let $\xi_{N_0} \in (0,\frac{2}{5})$ satisfy

$$(6.6.12)\qquad (\sqrt{1-4\xi_{N_0}^2})^{N_0}(\theta(\xi_{N_0},2)-1) = 1.$$

ξ_{N_0} is unique (cf. (6.5.4)). Define

$$(6.6.13)\qquad \bar\pi[N_0] = (\pi_0[N_0],\pi_1[N_0],\ldots,\pi_{N_0}[N_0]) \quad \text{where}$$

$$\pi_j[N_0] = 2\xi_{N_0} \quad (j = 0,1,\ldots,N_0-1) \quad \text{and} \quad \pi_{N_0}[N_0] = 1.$$

Hence, $\bar\pi[N_0] \in \Omega$ and the following result is a consequence of Theorem 6.6.1.

COROLLARY 6.6.3. *Let $\sigma \in (0,\infty]$. Let $\bar\omega \in \Omega$ with $\bar\omega = (\omega_0,\omega_1,\ldots,\omega_N)$, $N \leq N_0$ and $\bar\omega \neq \bar\pi[N_0]$. Then*

$$(6.6.14)\qquad r(M_{\bar\pi[N_0]};\ \mathcal{F}\langle\sigma\ \alpha\rangle) \geq r(M_{\bar\omega};\ \mathcal{F}\langle\sigma,\alpha\rangle) \quad (\text{for all } \alpha \in [-\tfrac{1}{2},\tfrac{1}{2}]).$$

More specificallly,

$$(6.6.15)\qquad r(M_{\bar\pi[N_0]};\ \mathcal{F}\langle\sigma,\alpha\rangle) = \sigma \qquad (\text{for all } \alpha \in (\xi_{N_0},\tfrac{1}{2}])$$

and there exists a constant $\mu_{\bar\omega}$ with $\xi_{N_0} < \mu_{\bar\omega} < \frac{1}{2}$ such that

$$(6.6.16)\qquad r(M_{\bar\omega};\ \mathcal{F}\langle\sigma,\alpha\rangle) = 0 \qquad (\text{for all } \alpha \in [-\tfrac{1}{2},\mu_{\bar\omega})).$$

PROOF. From (6.6.2c) it follows that for $j = 0,1,\ldots,N_0-1$ we have

$$\psi_1(\pi_j[N_0],\xi_{N_0}) < \psi_1(\omega,\xi_{N_0}) \quad (\text{for all } \omega \in (0,1] \text{ with } \omega \neq \pi_j[N_0])$$

and

$$\psi_1(\pi_j[N_0],\xi_{N_0}) < 1.$$

Consequently,

$$1 = \phi_2(\bar{\pi}[N_0],\xi_{N_0}) < \phi_2(\bar{\omega},\xi_{N_0})$$

and

$$1 = \phi_1(\bar{\pi}[N_0],\xi_{N_0}) \le \phi_1(\bar{\omega},\xi_{N_0}).$$

Hence, from Theorem 6.6.1, (6.6.5) and (6.6.6) it follows that the statements (6.6.14 - 16) are true. □

Corollary 6.6.3 may be interpreted as follows. Of all iterative methods $M_{\bar{\omega}}$ for which $\bar{\omega} \in \Omega$ with $\bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N)$ and $N \le N_0$, *the iterative method* $M_{\bar{\pi}[N_0]}$ *is optimal* in the following sense. Let $\sigma \in (0,\infty]$ and $\alpha \in [-\frac{1}{2},\frac{1}{2}]$. Any iterative method $M_{\bar{\omega}}$ for which $\bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N)$ with $N \le N_0$ and $\bar{\omega} \ne \bar{\pi}[N_0]$, possesses a radius of convergence with respect to $F\langle\sigma,\alpha\rangle$ that is never greater and, for certain values of α is even smaller, than the radius of convergence of $M_{\bar{\pi}[N_0]}$.

We conclude this section by presenting another consequence of Theorem 6.6.1.

COROLLARY 6.6.4. *For any* $\alpha \in (0,\frac{1}{2}]$ *an* $\bar{\omega} \in \Omega$ *exists for which*

$$r(M_{\bar{\omega}};\ F\langle\sigma,\alpha\rangle) = \sigma \qquad (\text{for all } \sigma \in (0,\infty]).$$

PROOF. It is easily verified that $\xi_{N_0} \downarrow 0$ (if $N_0 \to \infty$), (cf. (6.6.12)). Hence, Corollary 6.6.4 follows from (6.6.15). □

6.6.1. Proof of Theorem 6.6.1

Throughout this subsection, σ and α denote the constants that appear in Theorem 6.6.1 and $\bar{\omega}$ the element of Ω. We start with a lemma which was alluded to in the previous section (page 107).

LEMMA 6.6.5. *Let* $\omega \in (0,1]$. *Suppose* $\alpha \in (0,\frac{1}{2}]$. *Let* $F \in F\langle\sigma,\alpha\rangle$. *Then* $B(x^*,\sigma) \subset D(F)$. *Let* $y \in B(x^*,\sigma)$. *Then* $F'(y)$ *is invertible. Set* $z = y - \omega\Gamma(y)F(y)$. *Then*

$$(6.6.17) \qquad \|z-x^*\| \le \psi_1(\omega,\alpha)\|y-x^*\|.$$

Here ψ_1 is defined in (6.6.1).

PROOF. Notice that

$$(y-x^*,\Gamma(y)F(y)) \geq \alpha\{\|y-x^*\|^2 + \|\Gamma(y)F(y)\|^2\}.$$

Consequently,

$$\|z-x^*\|^2 = \|y-x^*\|^2 + \omega^2\|\Gamma(y)F(y)\|^2 - 2\omega(y-x^*,\Gamma(y)F(y))$$

$$\leq (1-2\omega\alpha)\|y-x^*\|^2 + (\omega^2-2\omega\alpha)\|\Gamma(y)F(y)\|^2.$$

Using Lemma 6.5.2 it follows that

$$\|z-x^*\|^2 \leq \begin{cases}\{1-2\omega\alpha + (\omega^2-2\omega\alpha)[\theta(\alpha,1)]^2\}\|y-x^*\|^2 & (\text{if } \omega \leq 2\alpha),\\ \{1-2\omega\alpha + (\omega^2-2\omega\alpha)[\theta(\alpha,2)]^2\}\|y-x^*\|^2 & (\text{if } \omega > 2\alpha).\end{cases}$$

From (6.5.2) it thus follows that

$$\|z-x^*\|^2 \leq \begin{cases}\{\omega^2[\theta(\alpha,1)]^2 - 2\omega\theta(\alpha,1) + 1\}\|y-x^*\|^2 & (\text{if } \omega \leq 2\alpha),\\ \{\omega^2[\theta(\alpha,2)]^2 - 2\omega\theta(\alpha,2) + 1\}\|y-x^*\|^2 & (\text{if } \omega > 2\alpha).\end{cases}$$

Thus (cf. (6.6.1) and (6.6.2c)) relation (6.6.17) is true. □

Let ϕ_1 and ϕ_2 be defined in (6.6.3) and (6.6.4) respectively. A consequence of the previous lemma is

LEMMA 6.6.6. *Suppose* $\alpha \in (0,\frac{1}{2}]$. *Let* $F \in F\langle\sigma,\alpha\rangle$ *and set* $G = M_{\overline{\omega}}(F)$. *Then for all* $x \in B(x^*,\sigma/\phi_1(\bar{\omega},\alpha))$ *we have* $x \in D(G)$ *and*

(6.6.18) $\|G(x)-x^*\| \leq \phi_2(\bar{\omega},\alpha)\|x-x^*\|.$

PROOF. Let $x \in B(x^*,\sigma/\phi_1(\bar{\omega},\alpha))$. Using Lemma 6.6.5 we have, with $y_0 = x$,

$$\|y_{n+1} - x^*\| \leq \psi_1(\omega_n,\alpha)\|y_n-x^*\| \leq \prod_{j=0}^{n} \psi_1(\omega_j,\alpha)\|y_0-x^*\| < \sigma$$

where

$$y_{n+1} = y_n - \omega_n\Gamma(y_n)F(y_n) \qquad (n = 0,1,\ldots,N-1).$$

Since $G(x) = y_N - \Gamma(y_N)F(y_N)$ it follows that

$$\|G(x)-x^*\| \le \psi_1(1,\alpha)\|y_N-x^*\| \le \prod_{j=0}^{N} \psi_1(\omega_j,\alpha)\|x-x^*\|.$$

This proves the statement. □

We are now in a position to give a lower bound on $r(M_{\bar{\omega}}; F\langle\sigma,\alpha\rangle)$.

LEMMA 6.6.7.

$$(6.6.19) \qquad r(M_{\bar{\omega}}; F\langle\sigma,\alpha\rangle) \ge \begin{cases} \dfrac{\sigma}{\phi_1(\bar{\omega},\alpha)} & (\text{if } \alpha \in (0,\tfrac{1}{2}] \text{ and } \phi_2(\bar{\omega},\alpha) < 1), \\ 0 & (\text{otherwise}). \end{cases}$$

PROOF. Relation (6.6.19) is a direct consequence of Lemma 6.6.6. □

The next subsection will be devoted to the proof of the following lemma, which may be conceived as a generalization of Lemma 6.5.5.

LEMMA 6.6.8. *Suppose* $\alpha \in (0,\frac{1}{2}]$. *Then the following propositions hold.*

(i) *If* $\phi_1(\bar{\omega},\alpha) > 1$ *then, for any* $\tau \in \mathbb{R}$ *for which* $\phi_1(\bar{\omega},\alpha)\tau > \sigma$, *an* $F \in F\langle\sigma,\alpha\rangle$ *which is infinitely differentiable on* $D(F)$ *exists for which*

$$(6.6.20) \qquad r(M_{\bar{\omega}},F) \le \tau.$$

(ii) *If* $\phi_2(\bar{\omega},\alpha) > 1$ *then, for any* $\varepsilon > 0$, *an* $F \in F\langle\sigma,\alpha\rangle$ *which is infinitely differentiable on* $D(F)$ *exists for which*

$$(6.6.21) \qquad r(M_{\bar{\omega}},F) \le \varepsilon.$$

Notice that

1. $r(M_{\bar{\omega}}; F\langle\sigma,\alpha\rangle) \le \sigma$ (cf. Remark 6.2.1). Hence (6.6.7) follows from Lemma 6.6.7, (6.6.6) and Lemma 6.6.8.

2. $F\langle\sigma,\alpha_1\rangle \supset F\langle\sigma,\alpha_2\rangle$ (if $-\frac{1}{2} \le \alpha_1 \le \alpha_2 \le \frac{1}{2}$). Hence $r(M_{\bar{\omega}}; F\langle\sigma,\alpha\rangle)$ is isotone on $[-\frac{1}{2},\frac{1}{2}]$ so that relation (6.6.8) holds.

Therefore, the statement of Theorem 6.6.1 is true.

6.6.2. Proof of the Lemmata 6.5.5 and 6.6.8

PART A.

We assume in this part of the proof that $E = \mathbb{R}$.

LEMMA 6.6.9. *Let the integer* $n \geq 0$ *and let* $\sigma \in (0,\infty]$. *Let* $\alpha_j \in (0,\frac{1}{2}]$, $\tau_j \in (0,\sigma)$ *and* $k_j \in \{1,2\}$ $(j = 0,1,\dots,n)$. *Suppose* $\tau_i \neq \tau_j$ *if* $i \neq j$ $(i,j = 0,1,\dots,n)$. *Then for any* $\hat{\alpha}$, *with* $0 < \hat{\alpha} \leq \min\{\alpha_j \mid 0 \leq j \leq n\}$, *an* $f \in F\langle\sigma,\alpha\rangle \cap C^{\infty}(-\sigma,\sigma)$ *(see (6.2.4)) exists such that*

$$\text{(i)} \quad x^* = 0 \textit{ is the unique solution of the equation } f(x) = 0,$$

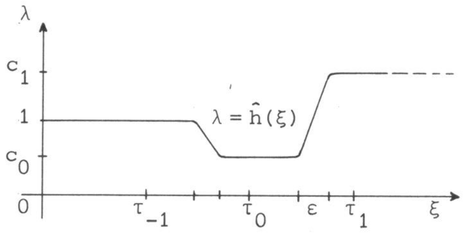
$$\text{(6.6.22)} \quad \text{(ii)} \quad f'(\xi) \neq 0 \textit{ (for all } \xi \in (-\sigma,\sigma)),$$

$$\text{(iii)} \quad [f'(\xi)]^{-1}f(\xi) = \theta(\alpha_j,k_j)\xi \textit{ (if } |\xi| = \tau_j \ (j = 0,1,\dots,n), \textit{ or}$$
$$|\xi| \in [\tau_j,\sigma) \textit{ and } \tau_j \geq \tau_i \ (i = 0,1,\dots,n)).$$

PROOF. We assume that $0 < \tau_0 < \tau_1 < \dots < \tau_n$, this is no restriction. Let $\hat{\alpha}$ satisfy $0 < \hat{\alpha} \leq \min\{\alpha_j \mid 0 \leq j \leq n\}$. Set $\tau_{n+1} = \sigma$ and let $\tau_{-1} \in (0,\tau_0)$. Let $\varepsilon > 0$ with $\varepsilon < \frac{1}{3}\min\{\tau_j - \tau_{j-1} \mid 0 \leq j \leq n+1\}$. Set $c_j = \theta(\alpha_j,k_j)$ $(j = 0,1,\dots,n)$, $c_{-1} = 1$. Let $\bar{d}_j = (c_{j-1},c_j,\tau_j-2\varepsilon,\tau_j-\varepsilon)$ $(j = 0,1,\dots,n)$ and let $\tilde{h}_j = h_{\bar{d}_j}$ (cf. (6.2.6)) $(j = 0,1,\dots,n)$. Let

$$\hat{h}\colon [0,\sigma) \to (0,\infty),$$

$$\text{(6.6.23)} \qquad \hat{h}(\xi) = \begin{cases} 1 & (\text{if } \xi \in [0,\tau_{-1}+\varepsilon]), \\ \tilde{h}_j(\xi) & (\text{if } \xi \in (\tau_{j-1}+\varepsilon,\tau_j+\varepsilon] \ (j = 0,1,\dots,n)), \\ \tilde{h}_n(\xi) & (\text{if } \xi \in (\tau_n+\varepsilon,\sigma)). \end{cases}$$



graph of $\hat{h}$

Fig. 6.6.5.

Define

$$f\colon (-\sigma,\sigma) \to \mathbb{R},$$

(6.6.24)
$$f(\xi) = \begin{cases} \exp[\int_{\tau_{-1}}^{\xi} \frac{1}{\hat{h}(\lambda)\lambda}\, d\lambda] & (\text{if } \xi \in (0,\sigma)), \\ 0 & (\text{if } \xi = 0), \\ -f(-\xi) & (\text{if } \xi \in (-\sigma,0)). \end{cases}$$

1. We show that $f \in C^{\infty}(-\sigma,\sigma)$.

a. f is continuous on $(0,\sigma)$ and

(6.6.25)
$$f'(\xi) = \frac{1}{\hat{h}(\xi)\xi} f(\xi) > 0 \qquad (\text{for all } \xi \in (0,\sigma)).$$

In general, for $k \geq 2$

$$f^{(k)}(\xi) = \frac{f(\xi)}{[\hat{h}(\xi)\xi]^{k}} p_k(\xi) \qquad (\text{for all } \xi \in (0,\sigma))$$

where p_k is a differentiable function that is composed of the functions $\hat{h},\hat{h}',\ldots,\hat{h}^{(k-1)}$.

b. If $\xi \in (0,\tau_{-1})$, then

$$f(\xi) = \exp[\int_{\tau_{-1}}^{\xi} \frac{1}{\lambda}\, d\lambda] = \frac{\xi}{\tau_{-1}} .$$

Hence f is continuous on $(-\tau_{-1},\tau_{-1})$ and $f^{(k)}(\xi)$ exists (for all $k \in \mathbb{N}$ and all $\xi \in (-\tau_{-1},\tau_{-1})$).

c. For $\xi \in (-\sigma,0)$ we have $f(\xi) = -\phi(-\xi)$ where ϕ denotes the restriction of f to $(0,\sigma)$. Since ϕ is infinitely differentiable on $(0,\sigma)$, the function f is infinitely differentiable on $(-\sigma,0)$ with

(6.6.26)
$$f^{(k)}(\xi) = (-1)^{k+1} f^{(k)}(-\xi) \qquad (\text{for all } k \in \mathbb{N} \text{ and all } \xi \in (-\sigma,0)).$$

Hence $f \in C^{\infty}(-\sigma,\sigma)$.

2. We prove that $f \in F\langle\sigma,\hat{\alpha}\rangle$. Obviously (cf. (6.6.25), (6.6.26)), $f'(\xi) > 0$ (for all $\xi \in (-\sigma,\sigma)$). Hence $\xi = 0$ is the unique solution of $f(\xi) = 0$.

Notice that, since $\hat{\alpha} > 0$ and $\hat{\alpha} \leq \min\{\alpha_j \mid 0 \leq j \leq n\}$ it follows that

$$\frac{1-\sqrt{1-4\hat{\alpha}^2}}{2\hat{\alpha}} = \theta(\hat{\alpha},1) \le \min\{\theta(\alpha_j,1) \mid 0 \le j \le n\} \le \hat{h}(\xi)$$

$$\le \max\{\theta(\alpha_j,2) \mid 0 \le j \le n\} \le \theta(\hat{\alpha},2) = \frac{1+\sqrt{1-4\hat{\alpha}^2}}{2\hat{\alpha}} \quad \text{(for all } \xi \in [0,\sigma)),$$

(see (6.5.4) and (6.5.5)). Hence

$$\hat{h}(\xi) \ge \hat{\alpha}\{1+[\hat{h}(\xi)]^2\} \qquad \text{(for all } \xi \in [0,\sigma)).$$

Thus

$$\hat{h}(\xi)\xi^2 \ge \hat{\alpha}\{\xi^2+[\hat{h}(\xi)\xi]^2\} \qquad \text{(for all } \xi \in [0,\sigma)).$$

Therefore (cf. (6.6.25)) it follows that

$$(6.6.27) \qquad \xi\,\frac{f(\xi)}{f'(\xi)} \ge \hat{\alpha}\{\xi^2+[\frac{f(\xi)}{f'(\xi)}]^2\} \qquad \text{(for all } \xi \in [0,\sigma)).$$

For $\xi \in (-\sigma,0)$ we have (cf. (6.6.26))

$$(6.6.28) \qquad \frac{f(\xi)}{f'(\xi)} = -\frac{f(-\xi)}{f'(-\xi)}\ .$$

Hence (6.6.27) holds for $\xi \in (-\sigma,0)$ as well. Thus $f \in F\langle\sigma,\hat{\alpha}\rangle$.

3. Relation (iii) follows from (6.6.23), (6.6.25) and (6.6.28). □

LEMMA 6.6.10. *Let $\sigma \in (0,\infty]$ and $n \in \mathbb{N}$. Let the numbers α_j and τ_j ($j = 0,1,\ldots,n$) satisfy the conditions of Lemma 6.6.9. Let $\tau_{n+1} > 0$ be given. Furthermore, let $\bar{\omega} \in \Omega$ with $\bar{\omega} = (\omega_0,\omega_1,\ldots,\omega_N)$ and $N \ge n$. Assume that $\tau_{j+1} = \psi_1(\omega_j,\alpha_j)\tau_j$ ($j = 0,1,\ldots,n$). Let $k_j = 1$ (if $\omega_j \le 2\alpha_j$) and $k_j = 2$ (if $\omega_j > 2\alpha_j$) ($j = 0,1,\ldots,n$). Let $f \in F\langle\sigma,\hat{\alpha}\rangle \cap C^\infty(-\sigma,\sigma)$ satisfy (6.6.22) where $0 < \hat{\alpha} \le \min\{\alpha_j \mid 0 \le j \le n\}$. Then for any $\xi \in (-\sigma,\sigma)$ with $|\xi| = \tau_0$ we have*

$$(6.6.29a) \qquad |y_{j+1}| = \tau_{j+1} \qquad (j = 0,1,\ldots,n).$$

Here

$$(6.6.29b) \qquad y_{j+1} = y_j - \omega_j[f'(y_j)]^{-1}f(y_j) \qquad (j = 0,1,\ldots,n)$$

and

(6.6.29c) $\quad y_0 = \xi$.

PROOF. Let $\xi \in (-\sigma,\sigma)$ satisfy $|\xi| = \tau_0$. Assume ℓ is an integer with $0 \le \ell \le n$. Suppose (6.6.29) holds for all j with $0 \le j \le \ell-1$. Thus $y_\ell \in (-\sigma,\sigma)$, and

$$y_{\ell+1} = y_\ell - \omega_\ell[f'(y_\ell)]^{-1}f(y_\ell) = (1 - \omega_\ell\theta(\alpha_\ell,k_\ell))y_\ell.$$

Hence $|y_{\ell+1}| = \tau_{\ell+1}$ (cf. (6.6.1), (6.6.2c)). This proves the lemma. □

The following lemma shows that the Lemmata 6.5.5, 6.6.8 hold if $E = \mathbb{R}$.

LEMMA 6.6.11. *Let* $\alpha \in (0,\frac{1}{2}]$ *and* $\bar{\omega} \in \Omega$.

(i) *Suppose* $\phi_1(\bar{\omega},\alpha) > 1$. *Let* $\sigma \in (0,\infty)$. *Then for all* $\tau \in \mathbb{R}$ *with* $\phi_1(\bar{\omega},\alpha)\tau > \sigma$ *an* $f \in \mathcal{F}<\sigma,\alpha> \cap C^\infty(-\sigma,\sigma)$ *and a number* $\xi_0 \in D(f)$ *with* $|\xi_0| \le \tau$ *exist such that* $f(0) = 0$, $f'(\xi) \ne 0$ *(for all* $\xi \in (-\sigma,\sigma)$*)* *and* $\xi_0 \notin D(M_{\bar{\omega}},f)$.

(ii) *Suppose* $\phi_2(\bar{\omega},\alpha) > 1$. *Then for any* $\varepsilon > 0$ *an* $f \in \mathcal{F}<\infty,\alpha> \cap C^\infty(\mathbb{R})$ *and a number* $\xi_0 \in D(f)$ *with* $|\xi_0| \le \varepsilon$ *exist such that* $f(0) = 0$, $f'(\xi) \ne 0$ *(for all* $\xi \in \mathbb{R}$*),* $\phi(\xi_0) \in D(\phi)$, $\phi[\phi(\xi_0)] = \xi_0$ *and* $\xi_0 \notin S(M_{\bar{\omega}},f)$. *Here* $\phi = M_{\bar{\omega}}(f)$.

(iii) *Suppose* $\alpha = \frac{2}{5}$. *Then for any* $\varepsilon > 0$ *an* $f \in \mathcal{F}<\infty,\alpha> \cap C^\infty(\mathbb{R})$ *and a number* $\xi_0 \in D(f)$ *with* $|\xi_0| = \varepsilon$ *exist such that* $f(0) = 0$, $f'(\xi) \ne 0$ *(for all* $\xi \in \mathbb{R}$*),* $\phi(\xi_0) \in D(\phi)$, $\phi[\phi(\xi_0)] = \xi_0$ *and* $\xi_0 \notin S(M,f)$. *Here M is Newton's method and* $\phi = M(f)$.

PROOF. (i) Suppose $\phi_1(\bar{\omega},\alpha) > 1$. Consequently $\alpha < \frac{1}{2}$ and an $n < N$ exists such that $\Pi_{j=0}^{n} \psi_1(\omega_j,\alpha) > 1$. Let $\tau \in (0,\sigma)$ such that $\Pi_{j=0}^{n} \psi_1(\omega_j,\alpha) > \frac{\sigma}{\tau}$. We assume that $\Pi_{j=0}^{i} \psi_1(\omega_j,\alpha) \le \frac{\sigma}{\tau}$ (for all $i < n$).

Due to (6.6.2a) an $\varepsilon > 0$ exists with $\alpha + \varepsilon < \frac{1}{2}$ such that $\Pi_{j=0}^{i} \psi_1(\omega_j,\alpha+\varepsilon_j) < \frac{\sigma}{\tau}$ (for all $i < n$) and $\Pi_{j=0}^{n} \psi_1(\omega_j,\alpha+\varepsilon_j) > \frac{\sigma}{\tau}$ (for any set $\{\varepsilon_0,\varepsilon_1,\ldots,\varepsilon_n\} \subset (0,\varepsilon]$). Set $\tau_0 = \tau$. We construct the numbers α_j and τ_{j+1} ($j = 0,1,\ldots,n$) as follows. Choose $\varepsilon_j \in (0,\varepsilon]$ such that $\psi_1(\omega_j,\alpha+\varepsilon_j)\tau_j \ne \tau_i$ (for all $i \le j$). Set $\alpha_j = \alpha+\varepsilon_j$ and $\tau_{j+1} = \psi_1(\omega_j,\alpha_j)\tau_j$. Thus $\{\tau_0,\ldots,\tau_n\} \subset (0,\sigma)$ and $\tau_{n+1} > \sigma$. According to Lemma 6.6.10 an $f \in \mathcal{F}<\sigma,\alpha> \cap C^\infty(-\sigma,\sigma)$ exists such that for all $\xi_0 \in \mathbb{R}$ with $|\xi_0| = \tau$ the relation (6.6.29) holds. Hence $\xi_0 \notin D(M_{\bar{\omega}}f)$. This proves statement (i).

(ii) Suppose $\phi_2(\bar{\omega},\alpha) > 1$. Hence $\alpha < \frac{1}{2}$. Let $\varepsilon > 0$. As in the first part of the proof we can construct numbers α_j ($j = 0,1,\ldots,N$) and τ_j

$(j = 0,1,\dots,N+1)$, where $\tau_0 = \varepsilon$ and $\tau_{N+1} > \tau_0$, such that $\alpha \le \min\{\alpha_j \mid 0 \le j \le N\}$, and $\tau_{j+1} = \psi(\omega_j,\alpha_j)\tau_j$ and $\tau_i \ne \tau_j$ for $i < j$ $(j = 0,1,\dots,N)$. From Lemma 6.6.10 it follows that an $f \in F\langle\infty,\alpha\rangle \cap C^\infty(\mathbb{R})$ exists such that, for all $\xi \in \mathbb{R}$ with $|\xi| = \varepsilon$, we have $\phi(\xi) = \tau_{N+1} > \tau_0 = |\xi|$. Here $\phi = M_{\overline{\omega}}(f)$. Since ϕ is continuous on $\mathbb{R}$ and $[M_{\overline{\omega}},f]$ is quadratically convergent (cf. p.87), from Lemma 6.2.6 it follows that a number $\xi_0 \in [-\varepsilon,\varepsilon]$ with $\xi_0 \ne 0$ exists such that $\phi[\phi(\xi_0)] = \xi_0$. This proves (ii).

(iii) Suppose $\alpha = \frac{2}{5}$. Let $\varepsilon > 0$ and let $\tau_0 = \varepsilon$. From Lemma 6.6.9 and (6.5.4) it follows that an $f \in F\langle\infty,\alpha\rangle \cap C^\infty(\mathbb{R})$ exists such that $[f'(\xi)]^{-1}f(\xi) = 2\xi$ (for all $\xi \in \mathbb{R}$ with $|\xi| = \tau_0$). This proves (iii). □

PART B.

We are now in a position to prove the Lemmata 6.5.5 and 6.6.8 where E is an arbitrary Hilbert space, as will be clear from the next lemma.

LEMMA 6.6.12. *Let* $\sigma \in (0,\infty]$ *and* $\alpha \in (0,\frac{1}{2}]$. *Let* $f \in C^\infty(-\sigma,\sigma)$ *and suppose that* $f'(\xi) \ne 0$ *(for all* $\xi \in (-\sigma,\sigma)$*) and* $f(0) = 0$. *Moreover, let*

$$(6.6.30) \qquad \xi\,\frac{f(\xi)}{f'(\xi)} \ge \alpha\{\xi^2 + [\frac{f(\xi)}{f'(\xi)}]^2\} \qquad (\text{for all } \xi \in (-\sigma,\sigma)).$$

Let F *be the* E*-extension of* f *(cf.* (6.2.14)*). Then* $F \in F\langle\sigma,\alpha\rangle$ *and* $x^* = 0$.

PROOF. In view of Theorem 6.2.12 and (6.4.9) it is sufficient to prove that

$$(6.6.31) \qquad (x,\Gamma(x)F(x)) \ge \alpha\{\|x\|^2 + \|\Gamma(x)F(x)\|^2\} \qquad (\text{for all } x \in B(0,\sigma)).$$

Let $x \in B(0,\sigma)$. Suppose that $\mathcal{B}_x = \{u \mid u = u_n \text{ with } n \in \mathbb{N}\}$ and $\xi_n = (x,u_n)$ $(n = 1,2,\dots)$. Hence $|\xi_n| < \sigma$ $(n = 1,2,\dots)$. From Theorem 6.2.12(iii) it follows that

$$(x,\Gamma(x)F(x)) = \sum_{n=1}^{\infty} \xi_n \frac{f(\xi_n)}{f'(\xi_n)}\,.$$

Since

$$\|x\|^2 + \|\Gamma(x)F(x)\|^2 = \sum_{n=1}^{\infty} \{\xi_n^2 + [\frac{f(\xi_n)}{f'(\xi_n)}]^2\},$$

relation (6.6.30) shows that (6.6.31) holds. □

PROOF OF THE LEMMATA 6.5.5, 6.6.8.

1. Suppose $\phi_1(\bar{\omega},\alpha) > 1$. Let $\tau \in \mathbb{R}$ satisfy $\phi_1(\bar{\omega},\alpha)\tau > \sigma$. Let $f \in C^\infty(-\sigma,\sigma)$ and ξ_0 satisfy statement (i) of Lemma 6.6.11. Let F be the E-extension of f. Thus $F \in \mathcal{F}\langle\sigma,\alpha\rangle$ and there exists an element u of E with $\|u\| = 1$ such that

$$\Gamma(\lambda u)F(\lambda u) = \frac{f(\lambda)}{f'(\lambda)}\,u \qquad (\text{for all } \lambda \in (-\sigma,\sigma)).$$

Set $x_0 = \xi_0 u$. Then $x_0 \notin D(M_{\bar{\omega}},F)$. Finally, from Theorem 6.2.12 it follows that F is infinitely differentiable on D(F). This proves statement (i) of Lemma 6.6.8.

2. By a similar argument to the above, (using Lemma 6.6.11(ii)) one can show that statement (ii) of Lemma 6.6.8 is true, and (using Lemma 6.6.11 (iii)) prove Lemma 6.5.5. □

CHAPTER 7

NUMERICAL EXPERIMENTS

In this chapter we present some numerical results. The test examples will all be finite dimensional problems. Examples of the application of the imbedding method to problems in infinite dimensional (Banach) spaces can be found, amongst others, in the following three references: [WACKER, 1972] (where the discrete imbedding method is used), [BOSARGE, 1971] (where Davidenko's method is used), [KLEINMICHEL, 1968] (where iterative imbedding is used), (cf. section 1.1).

We present some numerical results with iterative methods of the type investigated in Chapters 5 and 6.

All computations have been carried out on a CDC Cyber 73/173-28 computer (accuracy: 48 binary digits in the mantissa).

7.1. ITERATIVE METHODS TO BE TESTED

We have divided the iterative methods to be tested into three classes.

Let $A \in \mathcal{A}$ (cf. (2.6.3)) where $[A(F)](y,x) \equiv F(y) - F(x)$ (for all $F \in \mathcal{F}_1$ (cf. (2.6.1)). Let $F \in \mathcal{F}_1$.

CLASS 1. Let $N \geq 0$ and $\mathcal{H} = \{h_0, h_1, \ldots, h_N\}$ with $h_i = \frac{1}{N+1}$ $(i = 0,1,\ldots,N)$. Let $g = 0$. Let R_θ (for $\theta \in \mathbb{R}$) be the one-stage Runge-Kutta method with operator coefficient for which

$$\rho_{2,1}(z) \equiv [1-(1-\theta)z]^{-1}$$

(see also subsection 3.3.2). We notice that, when solving a linear differential equation with constant coefficients, the method R_θ with $\theta = \frac{1}{2}$ is equivalent to the *Trapezoidal rule,* and the method R_θ with $\theta = 0$ is equivalent to the *Backward Euler method*. The Trapezoidal rule is more accurate than the Backward Euler method. However the latter method has better stability properties than the former one (cf. [LAMBERT, 1973; pp. 235-236]). If $\theta = 1$ then R_θ reduces to *Euler's method.*

Set $M_{N,\theta} = \mathbb{M}(A,g,R_\theta,H)$. Then $G = M_{N,\theta}(F)$ satisfies (cf. section 4.2)

$$G: D(G) \to E, \qquad G(x) = \eta_{N+1}(x) \quad (x \in D(G)). \tag{7.1.1a}$$

In (7.1.1a) $\eta_j(x)$ $(j = 0,1,\dots,N+1)$ is defined by

$$\eta_0(x) = x$$
and
$$\eta_{j+1}(x) = \eta_j(x) - \frac{1}{N+1}\left[I - \frac{(1-\theta)}{N+1}\Gamma(\eta_j(x))F''(\eta_j(x))\Gamma(\eta_j(x))F(x)\right]^{-1} \Gamma(\eta_j(x))F(x) \quad (j = 0,1,\dots,N). \tag{7.1.1b}$$

We notice that for $\theta = \frac{1}{2}$ and $N = 0$ the method $M_{N,\theta}$ is known in the literature as the *method of tangent hyperbolas* (cf. [ORTEGA & RHEINBOLDT, 1970; p. 188] for a bibliography on this method). Furthermore for $\theta = 0$ and $N = 0$ the method $M_{N,\theta}$ has been investigated in [DI LENA & TRIGIANTE, 1976]. In that paper it was supposed that $E = \mathbb{R}$. In the computations that were performed on some problems in $\mathbb{R}$, the method exhibited better convergence behaviour than the convergence behaviour of Newton's method, especially when the starting points were not close to the desired solution.

From Theorem 5.2.8 it follows that the iterative processes $[M_{N,\theta},F]$ $(N \geq 0,\ \theta \in \mathbb{R})$ are all quadratically convergent.

In higher dimensional vector spaces computation of $F''(z)$ requires in general an exorbitant amount of work. Consequently, for $\theta \neq 1$, the iterative methods $M_{N,\theta}$ are rather cumbersome from the computational point of view. We shall therefore modify $M_{N,\theta}$ in such a way that $F''(\eta_j(x))$ need not be computed. Let $y,x \in D(F)$. Suppose $F'(y)$ is invertible. For any $\varepsilon > 0$ there exists a number $\rho > 0$ such that for all τ with $0 < |\tau| < \rho$

$$\frac{1}{\tau}\,\| F'(y+\tau\Gamma(y)F(x)) - F'(y) - \tau F''(y)\Gamma(y)F(x)\| < \varepsilon.$$

When τ is small, the operator $\frac{1}{\tau}\{F'(y+\tau\Gamma(y)F(x)) - F'(y)\}$ is therefore approximately equal to $F''(y)\Gamma(y)F(x)$. Thus we can approximate the iteration function G, defined in (7.1.1) by an iteration function $\tilde{G}$,

$$\tilde{G}: D(\tilde{G}) \to E,$$

(7.1.2a)

$$\tilde{G}(x) = \tilde{\eta}_{N+1}(x) \qquad (x \in D(\tilde{G})).$$

In (7.1.2a) $\tilde{\eta}_j(x)$ $(j = 0,1,\ldots,N+1)$ is defined by

$$\tilde{\eta}_0(x) = x$$

(7.1.2b) and

$$\tilde{\eta}_{j+1}(x) = \tilde{\eta}_j(x) - \frac{1}{N+1}[F'(\tilde{\eta}_j(x)) - \frac{(1-\theta)}{(N+1)\tau}\{F'(\tilde{\eta}_j(x)+\tau\Gamma(\tilde{\eta}_j(x))F(x)) - F'(\tilde{\eta}_j(x))\}]^{-1}F(x) \qquad (j = 0,1,\ldots,N).$$

$D(\tilde{G})$ is defined in a way similar to $D(G)$ (cf. section 4.2.2). $\tilde{G}$ can be conceived as an iteration function of an iterative method, which we denote by $\tilde{M}_{N,\theta}$. Hence $\tilde{G} = \tilde{M}_{N,\theta}(F)$.

We notice that for $N = 0$, iterative methods $\tilde{M}_{N,\theta}$ have been investigated by several authors (cf. [TRAUB, 1964; p. 164], where it is assumed that $E = \mathbb{R}$; for the case that E is an arbitrary Banach space, an example is given, for instance, in [KOGAN, 1967]). Just as with the method of tangent hyperbolas, the main purpose of investigating these methods was their local convergence behaviour (*near* x^*). Our main interest, however is in what happens when the starting point x_0 is *remote* from x^*.

<u>CLASS 2</u>. We shall also consider the optimal methods $M_{\bar{\omega}}$ defined in Definition 6.3.1, where $\bar{\omega} = \pi[N]$ (cf. (6.6.13)) with $N \in \mathbb{N}$ given. We notice that the iterative processes $[M_{\overline{\pi[N]}},F]$ $(N \in \mathbb{N})$ are all quadratically convergent (cf. p. 87).

<u>CLASS 3</u>. Let $N \geq 0$ and let $H = \{h_0,h_1,\ldots,h_N\}$ with $h_i = \frac{1}{N+1}$ $(i = 0,1,\ldots,N)$. Let $g_1,g_2 \in S$ with

(7.1.3a) $$g_1 = 0$$

and

(7.1.3b) $$g_2(t) = \begin{cases} \frac{1}{1-t} & (\text{if } t \in [0,1)), \\ 1 & (\text{if } t = 1). \end{cases}$$

We shall also consider the methods $M_{i,j,N}$ where

(7.1.4) $\quad M_{i,j,N} = \mathbb{M}(A,g_i,L_j,H) \qquad (i = 1,2;\ j = 1,2,3).$

Here, L_1 and L_2 are defined in (3.3.6). The Runge-Kutta method L_3 is the so-called *Modified Euler method* (cf. [LAMBERT, 1973; p. 118]) and is defined by

(7.1.5) $$L_3 = \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} .$$

From Theorem 5.2.5 it follows that the iterative processes $[M_{i,j,N},F]$ $(i = 1,2;\ j = 1,2,3)$ are quadratically convergent.

We shall test iterative methods $\tilde{M}_{N,\theta}$ of Class 1 with $\theta = 0,\frac{1}{2},1$. The choice of these methods has been motivated in section 3.3.

Methods of classes 2 and 3 have the following features, which motivated their choice.

(a) The methods of Class 2, for which we could determine the radius of convergence, are all optimal according to the theory given in Part II of Chapter 6.

(b) The methods $M_{i,j,N}$ $(i,j = 1,2)$ are based on Runge-Kutta methods that are superior to Euler's method for certain types of differential equations (see section 3.3).

(c) The methods $M_{i,j,N}$ $(i,j = 1,2)$ are based on *first order* Runge-Kutta methods which require per step *at least two* evaluations of the function f (appearing on the right-hand side of the differential equation). The methods $M_{i,3,N}$ $(i = 1,2)$ are based on a *second order* Runge-Kutta method which requires *two* such function evaluations per step.

(d) The methods $M_{\bar{\pi}[N]}$ and $M_{2,j,N}$ $(j = 1,2,3)$ are both of the type (4.1.3), with the same A and g. We thus have $M_{\bar{\pi}[N]} = \mathbb{M}(A,g_2,L_0,\tilde{H})$ and $M_{2,j,N} = \mathbb{M}(A,g_2,L_j,H)$ where A, g_2, H and L_j $(j = 1,2,3)$ are defined above. L_0 denotes Euler's method and $\tilde{H} = (\tilde{h}_0,\tilde{h}_1,\ldots,\tilde{h}_N)$ denotes the sequence of stepsizes which corresponds to the optimal $\bar{\omega} = (\tilde{\omega}_0,\tilde{\omega}_1,\ldots,\tilde{\omega}_N) = \bar{\pi}[N]$ by means of (6.3.4).

(e) With $j \in \{1,2,3\}$ and $N \geq 0$ both fixed, the iterative methods $M_{i,j,N}$

(i = 1,2) differ only with regard to function g_i (i = 1,2).

7.2. NUMERICAL RESULTS WITH THE METHODS DESCRIBED IN SECTION 7.1

Iterative methods of the Classes 1, 2 and 3 have been applied to two problems.

PROBLEM 1. This problem arises from a finite-difference approach to the one-dimensional two-point boundary value problem

$$\frac{d}{ds}\{s^2 \frac{d}{ds} V(s)\} - s^2 f(V(s)) = 0 \qquad (0 < s < 1),$$

(7.2.1)

$$V'(0) = 0, \quad V(1) = 1$$

where $f(v) \equiv \varepsilon^{-1} \frac{v}{v+\lambda}$ (see [KELLER, 1968; p. 162 et seq.]). On physical grounds the solution of (7.2.1) looked for should be positive and continuous. It can be shown that a unique positive solution exists, which is strictly isotone. In [MURRAY, 1968] it is shown that $V(0) \sim 2[\varepsilon\lambda]^{-\frac{1}{2}}\exp\{-[\varepsilon\lambda]^{-\frac{1}{2}}\}$ (if $\varepsilon\lambda << 1$).

Let $m \in \mathbb{N}$ and let $\varepsilon,\lambda > 0$. Consider the (m+1)-dimensional problem $F(x) = 0$, where for $x = (\xi_0,\xi_1,\ldots,\xi_m)$ and $F(x) = (\phi_0(x),\phi_1(x),\ldots,\phi_m(x))$

$$\phi_0(x) = [s_{\frac{1}{2}}]^2(\xi_0 - \xi_1),$$

$$\phi_j(x) = -[s_{j-\frac{1}{2}}]^2\xi_{j-1} + ([s_{j-\frac{1}{2}}]^2 + [s_{j+\frac{1}{2}}]^2)\xi_j - [s_{j+\frac{1}{2}}]^2\xi_{j+1}$$

(7.2.2)
$$+ \Delta^2[s_j]^2 f(\xi_j) \qquad (j = 1,2,\ldots,m-1),$$

$$\phi_m(x) = -[s_{m-\frac{1}{2}}]^2\xi_{m-1} + ([s_{m-\frac{1}{2}}]^2 + [s_{m+\frac{1}{2}}]^2)\xi_m - [s_{m+\frac{1}{2}}]^2$$

$$+ \Delta^2[s_m]^2 f(\xi_m).$$

Here, $\Delta = \frac{1}{m+1}$ and $s_i = i\,\Delta$ $(i = \frac{1}{2},1,\ldots,m+\frac{1}{2})$. Set

(7.2.3) $$D(F) = \{x \mid x \in \mathbb{R}^{m+1};\ x = (\xi_0,\xi_1,\ldots,\xi_m) \text{ where } \xi_j > 0$$

$$(j = 0,1,\ldots,m)\}.$$

It can be shown that $F'(x)$ is invertible (for all $x \in D(F)$) and that $F(x) = 0$ has a solution that is unique in $D(F)$. Consequently, $F \in \mathcal{F}_1$.

The starting point x_0 was chosen to be $x_0 = (\xi_0^0, \xi_1^0, \ldots, \xi_m^0)$ where $\xi_j^0 = (1-\varepsilon\lambda)[s_j]^2 + \varepsilon\lambda$ $(j = 0,1,\ldots,m)$. The computations were performed for $m = 99$ and $\lambda = 0.1$. For ε we took $\varepsilon = 0.05$, 0.01 and 0.001.

In all these three cases, Newton's method failed. More specifically, it generated sequences $\{x_k\}$ that converged to a "solution" outside $D(F)$. □

Any iterative process was considered to yield a sequence $\{x_k\}$ converging to a solution of the equation $F(x) = 0$ whenever, for some k with $1 \le k \le 20$,

$$(7.2.4) \qquad \|x_k - x_{k-1}\| \le \delta_1(1 + \|x_k\|) \quad \text{or} \quad \|F(x_k)\| \le \delta_2$$

(see e.g. [BUS, 1975]). For both δ_1 and δ_2 we took 10^{-6}. The norm used was the Euclidean norm. The number of iteration steps required for the stopping criterion to be satisfied is only given if a process succeeded in generating a sequence $\{x_k\}$ that converged to the desired solution. In all other cases we assume that the iterative process failed and indicate this by FAILURE.

TABLE 7.2.1.

Method Class 1 (cf. (7.1.2); $\tau = 10^{-4}$).

Problem 1 ($\lambda = 0.1$, $m = 99$).

method	N	0	0	0	1	1	1
Problem	θ	1	0.5	0	1	0.5	0
$\varepsilon = 0.05$		FAILURE	3	4	4	2	3
$\varepsilon = 0.01$		FAILURE	FAILURE	5	FAILURE	3	4
$\varepsilon = 0.001$		FAILURE	FAILURE	7	FAILURE	4	5

method	N	3	3	3	7
Problem	θ	1	0.5	0	1
$\varepsilon = 0.05$		3	2	3	3
$\varepsilon = 0.01$		4	2	4	3
$\varepsilon = 0.001$		FAILURE	2	4	FAILURE

TABLE 7.2.2.

Method Class 2 (cf. (6.3.3), (6.6.13)).

Problem 1 ($\lambda = 0.1$, $m = 99$).

method N / Problem	1	3	7
$\varepsilon = 0.05$	3	2	1
$\varepsilon = 0.01$	FAILURE	FAILURE	FAILURE
$\varepsilon = 0.001$	FAILURE	FAILURE	FAILURE

TABLE 7.2.3.

Method Class 3, $M_{1,j,N}$ (cf. (7.1.4), (7.1.3a), (3.3.6) and (7.1.5)).

Problem 1 ($\lambda = 0.1$, $m = 99$).

method N	0	0	0	1	1	1	3	3
Problem j	1	2	3	1	2	3	1	3
$\varepsilon = 0.05$	FAILURE	8	4	3	3	2	3	2
$\varepsilon = 0.01$	FAILURE	FAILURE	7	FAILURE	FAILURE	4	4	3
$\varepsilon = 0.001$	FAILURE	FAILURE	FAILURE	FAILURE	FAILURE	7	FAILURE	3

TABLE 7.2.4.

Method Class 3, $M_{2,j,N}$ (cf. (7.1.4), (7.1.3b), (3.3.6) and (7.1.5)).

Problem 1 ($\lambda = 0.1$, $m = 99$).

method N	0	0	0	1	1	1	3	3
Problem j	1	2	3	1	2	3	1	3
$\varepsilon = 0.05$	FAILURE	7	8	3	3	3	2	2
$\varepsilon = 0.01$	FAILURE	FAILURE	FAILURE	FAILURE	FAILURE	FAILURE	2	3
$\varepsilon = 0.001$	FAILURE	FAILURE	FAILURE	FAILURE	FAILURE	FAILURE	FAILURE	FAILURE

PROBLEM 2. As a second example we consider the problem $F(x) = 0$, where for $x = (\xi_1,\xi_2)$ and $F(x) = (\phi_1(x),\phi_2(x))$

$$\begin{aligned}
\phi_1(x) &= \tfrac{1}{2}\{\sin(\xi_1\xi_2) - \tfrac{1}{2\pi}\,\xi_2 - \xi_1\}, \\
\phi_2(x) &= (1 - \tfrac{1}{4\pi})(e^{2\xi_1} - e) + \tfrac{e}{\pi}\,\xi_2 - 2e\xi_1
\end{aligned} \tag{7.2.5}$$

(cf. [CARNAHAN, LUTHER & WILKES, 1969; p.319]). This problem has a solution $x^* \approx (0.30,\ 2.84)$. Set

(7.2.6) $$D(F) = \{y_0 \mid y_0 \in \mathbb{R}^2;\ \text{a continuous curve } X \text{ on } [0,1] \text{ with } X(0) = y_0 \text{ and } X(1) = x^* \text{ exists, such that } F'(X(t)) \text{ exists and is invertible and } F(X(t)) - (1-t)F(y_0) = 0 \text{ (for all } t \in [0,1])\}.$$

It can be shown that $D(F)$ is open. The equation $F(x) = 0$ also has two "solutions", y^* and z^*, outside $D(F)$. $y^* = (0.5, \pi)$ and $z^* \approx (-0.26,\ 0.62)$. The starting point was chosen to be $x_0 = (0.4,\ 3)$. It appears that $x_0 \in D(F)$ (cf. [BOGGS, 1971]). In this case Newton's method failed in the sense that it generated a sequence $\{x_k\}$ converging to z^* instead of x^*. □

TABLE 7.2.5.

Method Class 1 (cf. (7.1.2); $\tau = 10^{-4}$).

Problem 2.

method N	0	0	0	1	1	1
θ	1	0.5	0	1	0.5	0
	FAILURE	5	7	FAILURE	3	5

method N	3	3	3	7
θ	1	0.5	0	1
	4	3	4	3

TABLE 7.2.6.

Method Class 2 (cf. (6.3.3), (6.6.13)).

Problem 2.

method N	1	3	7
	FAILURE	FAILURE	2

TABLE 7.2.7.

Method Class 3 (cf. (7.1.4), (7.1.3), (3.3.6) and (7.1.5)).

Problem 2.

	N	0	0	0	1	1	1	3	3
method	j	1	2	3	1	2	3	1	3
	$M_{1,j,N}$	4	3	3	3	3	3	3	3
	$M_{2,j,N}$	FAILURE	3	FAILURE	3	2	FAILURE	2	3

CONCLUSION. The methods $\tilde{M}_{N,\theta}$ of Class 1 with $\theta = 0$ and $\theta = \frac{1}{2}$, and the methods $M_{1,3,N}$ of Class 3 appear to be more reliable than the other methods whose test results have been presented above (including Newton's method $\tilde{M}_{0,1}$, which failed).

The work per step for a method $\tilde{M}_{N,\theta}$ (with $\theta \neq 1$) and for a method $M_{1,3,N}$ is roughly twice the work for a method $\tilde{M}_{N,\theta}$ (with $\theta = 1$) and for a method $M_{\overline{\pi}[N]}$ - with the same $N \geq 0$. Even if we compare methods $\tilde{M}_{N,\theta}$ (with $\theta = 0,\frac{1}{2}$) and $M_{1,3,N}$ with the methods $\tilde{M}_{N,\theta}$ (with $\theta = 1$) and $M_{\overline{\pi}[N]}$ that require the same amount of work per iteration step, the former (especially $\tilde{M}_{N,\theta}$ with $\theta = 0$) turn out to be better.

Among the methods $\tilde{M}_{N,\theta}$ ($\theta = 0,\frac{1}{2}$) and $M_{1,3,N}$, the methods $\tilde{M}_{N,\theta}$ with $\theta = 0$ (in which case the Backward Euler method underlies the iterative methods) appear to be more reliable than the methods $\tilde{M}_{N,\theta}$ with $\theta = \frac{1}{2}$ (in which case the Trapezoidal rule underlies the iterative methods) and $M_{1,3,N}$ (in which case the Modified Euler method underlies the iterative methods). However, if small stepsizes are taken (i.e. N large) then the last two types of methods generate sequences that converge faster to the solution than the first ones.

If we restrict our attention to the methods of Classes 2 and 3, methods $M_{1,3,N}$ of Class 3 appear to be more reliable than the other ones. The methods $M_{\overline{\pi}[N]}$ and the methods $M_{2,j,N}$, which are closely related to one another, appear to have about the same (non-) convergence behaviour.

If we compare the methods of Class 3 with one another with regard to the function g, the iterative methods that are based on g_1 (cf. (7.1.3a)) appear to be superior to the iterative methods based on g_2 (cf. (7.1.3b)). This indicates that the choice of the function g may significantly affect the convergence behaviour of an iterative method of the type (4.1.3b). It therefore seems worth while to investigate in future research the role of function g in the convergence behaviour of iterative processes [M,F] where

M is of the type (4.1.3b) and x_0 is remote from x^*.

In conclusion, the methods $\tilde{M}_{N,\theta}$ appear to be very reliable for solving the problem $F(x) = 0$. The investigation of these methods in future research would therefore seem to be desirable.

CHAPTER 8

PREDICTOR-CORRECTOR CONTINUATION ALGORITHMS

Let $F \in \mathcal{F}_1$. So far, we have only been interested in *iterative processes* for solving

(8.1.0) $F(x) = 0.$

In [RHEINBOLDT, 1975] an *algorithm* for solving (8.1.0) is proposed which is based on discrete imbedding. It has an adaptive step strategy for determining the partition $\{t_0, t_1, \ldots, t_N\}$ (cf. section 1.1).

In section 8.1 we give an outline of this algorithm. In sections 8.2, 8.3 and 8.4 we go into more detail, and present some variants of the algorithm, in some of which Davidenko's method is used also (cf. section 1.1). These variants also use the results given in section 6.5. In some of them Newton's method is used, while in the others the iterative method $\tilde{M}_{0,0}$ is used also (cf. (7.1.2)). In section 8.5 we present some numerical results obtained with the algorithms described here.

In this chapter we assume that E is finite-dimensional. Throughout the sections 8.1 - 4 the operator F denotes a given (fixed) element of $\mathcal{F}_1$.

8.1. INTRODUCTION

Let $u_0 \in D(F)$ and let

(8.1.1)
$$Q\colon [0,1] \times D(F) \to E,$$
$$Q(t,x) \equiv F(x) - (1-t)F(u_0)$$

(cf. (1.1.4)). Suppose that for all $t \in [0,1]$ the equation

(8.1.2) $Q(t,x) = 0$

has a unique solution $x = U(t)$, which depends continuously on t.

We recall that the method of discrete imbedding consists of selecting a partition $P = \{t_0, t_1, \ldots, t_N\}$ with $0 = t_0 < t_1 < \ldots < t_N = 1$ and of solving (8.1.2) successively for $t = t_i$ $(i = 1,2,\ldots,N)$, (cf. (1.1.9)).

Let M be the iterative method (the *local method*) that is used for approximating the solution $U(t_i)$ of (8.1.2) $(i = 1,2,\ldots,N)$. By $r(t)$ we denote the radius of convergence (see Definition 2.4.1) of the iterative process $[M, Q(t,\cdot)]$ $(t \in [0,1])$.

We shall now describe how a partition P and starting points $y_{i,0}$ for $[M, Q(t_i,\cdot)]$(the *i-th local process*) are selected in the algorithms to be described here.

Suppose the algorithm has progressed through $t_1, t_2, \ldots, t_k$ with $t_k < 1$ and $k \geq 1$, so that close approximations u_i of $U(t_i)$ $(i = 1,2,\ldots,k)$ have been obtained. We indicate how t_{k+1} and $y_{k+1,0}$ are selected (we shall specify what follows further on).

1. An approximation U_k of U is constructed. Thus

$$(8.1.3) \qquad U_k(t) \approx U(t) \qquad (t \in [0,1]).$$

2. $\|U_k(t) - U(t)\|$ is estimated by $\varepsilon_k(t)$ (for all $t \geq t_k$). Thus

$$(8.1.4) \qquad \varepsilon_k(t) \approx \|U_k(t) - U(t)\| \qquad (t \geq t_k).$$

3. $r(t)$ is estimated by $r_k(t)$ (for all $t \geq t_k$). Thus

$$(8.1.5) \qquad r_k(t) \approx r(t) \qquad (t \geq t_k).$$

4. Let h_k be the solution of

$$(8.1.6) \qquad \varepsilon_k(t_k + h) = \gamma_1 r_k(t_k + h) \qquad (h > 0),$$

where γ_1, $0 < \gamma_1 \leq 1$, is a given number which reflects the uncertainties of the estimates. Then

$$t_{k+1} = t_k + h_k.$$

5. Set

(8.1.7) $\quad y_{k+1,0} = U_k(t_{k+1})$.

$y_{k+1,0}$ is the starting point of the local process $[M,Q(t_{k+1},\cdot)]$.

Observe that, when $r_k(t) \equiv r(t)$ and $\varepsilon_k(t) \equiv \|U_k(t) - U(t)\|$, for any $\gamma_1 \in (0,1)$ the (k+1)-th local process yields a sequence $\{y_{k+1,j}\}_{j=0,1,2,...}$ that converges to $U(t_{k+1})$.

In the algorithms to be presented here, either $\varepsilon_k = 0$, or ε_k is strictly isotone on $[t_k,\infty)$, $\varepsilon_k(t_k) = 0$ and $\lim_{t\to\infty} \varepsilon_k(t) = \infty$. In the first case we set $t_{k+1} = 1$, in the last case the equation (8.1.6) has a unique solution.

Following Rheinboldt ([RHEINBOLDT, 1976]) we call an algorithm of the type described above a *predictor-corrector continuation* (PCC) *algorithm* (U_k is the predictor and M is the corrector).

In the following survey of the algorithm, we also give a more detailed description of the way in which the stepsize is selected.

```
1. k := 0; $t_k$ := 0; $h_k$ := $h_{start}$; go to 3.
2. $h_k$ is the solution of (8.1.6).

3.   $t_{k+1} := \begin{cases} t_k + h_{min} & (\text{if } h_k < h_{min}), \\ t_k + h_k & (\text{if } h_{min} \le h_k \le h_{max}), \\ t_k + h_{max} & (\text{if } h_k > h_{max}). \end{cases}$

   If $t_{k+1} > 1$ then $t_{k+1}$ := 1; $y_{k+1,0}$ := $U_k(t_{k+1})$.
4. Perform $[M,Q(t_{k+1},\cdot)]$ with starting point $y_{k+1,0}$
   If convergence criterion is met for some $y_{k+1,\ell}$, with $\ell \le \ell_{max}$,
      then ($u_{k+1}$ := $y_{k+1,\ell}$;
            if $t_{k+1}$ = 1 then STOP
            else (k := k+1; go to 2))
      else
      if $t_{k+1} - t_k = h_{min}$ then FAILURE
      else ($h_k$ := $\gamma_2 h_k$; go to 3).
   The constant $\gamma_2 \in (0,1)$ is given.
```

In sections 8.2 - 4 we specify how U_k (k = 0,1,...), and ε_k and r_k (k = 1,2,...) are determined. We assume that for all $t \in [0,1]$ the operator F'(x) exists and is invertible (for all x in some open neighbourhood of U(t)).

Hence, from Theorem 2.6.4(i) it follows that

$$(8.1.8)\qquad \begin{aligned} &\dot{U}(t) = -\Gamma(U(t))F(u_0) \qquad (t \in [0,1]),\\ &U(0) = u_0. \end{aligned}$$

In sections 8.2 - 4 we suppose the algorithm has progressed through $t_0,t_1,\dots,t_k$ with $k \geq 0$. Hence close approximations $u_0,u_1,\dots,u_k$ of respectively $U(t_0),U(t_1),\dots,U(t_k)$ are available.

8.2. DETERMINATION OF U_k

We shall describe three types of approximations U_k which we tested. We suppose $k \geq 0$.

a. *Lagrange interpolation*. Let an integer p with $0 \leq p \leq k$ be given. Choose $U_k = U_{k,p}$, where

$$(8.2.1a)\qquad U_{k,p}(t) \equiv \sum_{j=0}^{p} \pi_{k,p,j}(t)u_{k-j}$$

and

$$(8.2.1b)\qquad \pi_{k,p,j}(t) \equiv \prod_{\substack{\ell=0\\ \ell\neq j}}^{p} \left[\frac{t-t_{k-\ell}}{t_{k-j}-t_{k-\ell}}\right] \qquad (j = 0,1,\dots,p).$$

It follows that

$$(8.2.2)\qquad U_{k,p}(t_{k-j}) = u_{k-j} \qquad (j = 0,1,\dots,p).$$

In [RHEINBOLDT, 1975] U_k is of this type.

b. *Interpolation, using relation (8.1.8)*. Let an integer p with $0 \leq p \leq k$ be given. Set

$$(8.2.3)\qquad \dot{u}_k = -\Gamma(u_k)F(u_0).$$

Choose $U_k = \bar{U}_{k,p}$, where

$$(8.2.4a)\qquad \bar{U}_{k,p}(t) \equiv u_0 \qquad (\text{if } k = 0)$$

and

$$(8.2.4b) \qquad \bar{U}_{k,p}(t) \equiv \sum_{j=1}^{p} \frac{t-t_k}{t_{k-j}-t_k} \pi_{k,p,j}(t)u_{k-j}$$

$$+ \pi_{k,p,0}(t)[(1 - \dot{\pi}_{k,p,0}(t_k)(t-t_k))u_k + (t-t_k)\dot{u}_k]$$

$$(\text{if } k \geq 1).$$

Here $\pi_{k,p,j}$ is defined in (8.2.1b). It is easily verified that, when $k \geq 1$,

$$(8.2.5a) \qquad \bar{U}_{k,p}(t_{k-j}) = u_{k-j} \qquad (j = 0,1,\dots,p)$$

and

$$(8.2.5b) \qquad \dot{\bar{U}}_{k,p}(t_k) = \dot{u}_k.$$

This type of interpolation curve is, for example, used in [BYRNE & HINDMARSH, 1975] as a predictor for implicit backward differentiation formulae for the numerical solution of ordinary differential equations.

c. *Approximation, using a Runge-Kutta method with operator coefficient.* Consider the one-stage Runge-Kutta method with operator coefficient $R = (\rho_{i,j})$ where $\rho_{2,1}(z) \equiv [1-z]^{-1}$ (see section 3.3 and p.121); R is related to the Backward Euler method.

Let $\hat{U}_k(t)$ be the approximation through u_k of $U(t)$, which is obtained by performing one integration step of R on problem (8.1.8). Thus

$$\hat{U}_k(t) \equiv u_k - (t-t_k)[I - (t-t_k)\Gamma(u_k)F''(u_k)\Gamma(u_k)F(u_0)]^{-1}\Gamma(u_k)F(u_0).$$

As in (7.1.2) we approximate $F''(u_k)\Gamma(u_k)F(u_0)$ by

$$\frac{1}{\gamma 3}\{F'(u_k + \gamma_3\Gamma(u_k)F(u_0)) - F'(u_k)\}.$$

Here, $\gamma_3 > 0$ is a given small number (cf. p.122). We then obtain the approximation $\check{U}_k$ to U,

$$(8.2.6a) \qquad \check{U}_k(t) \equiv u_k - (t-t_k)[F'(u_k) - \frac{(t-t_k)}{\gamma_3}\{F'(u_k+\gamma_3\Gamma(u_k)F(u_0))-F'(u_k)\}]^{-1}$$

$$F(u_0).$$

Compared to the predictor $U_{k,p}$, the determination of $\bar{U}_{k,p}$ with $k \geq 1$, requires approximately one Newton-step extra work. For $k \geq 1$ we therefore approximate $\dot{u}_k$ by $\Gamma(y_{k,\ell-1})F(u_0)$, where $y_{k,\ell-1}$ is the last but one iterate of the k-th local process (the L-U decomposition of $F'(y_{k,\ell-1}) = \partial_2 Q(t_k, y_{k,\ell-1})$ is available from the local process).

Similarly, if $k \geq 1$, instead of $\check{U}_k$, defined in (8.2.6a), we use

$$\text{(8.2.6b)} \qquad \tilde{U}_k(t) \equiv u_k - (t-t_k)[F'(y_{k,\ell-1}) - \frac{t-t_k}{\gamma_3}\{F'(y_{k,\ell-1} + \gamma_3 \Gamma(y_{k,\ell-1})F(u_0)) - F'(y_{k,\ell-1})\}]^{-1} F(u_0).$$

When compared to $U_{k,p}$, the amount of extra work required for $\tilde{U}_k$ is approximately equal to one Newton-step. For ease of reference, we put $\tilde{U}_0 = \check{U}_0$.

8.3. ESTIMATION OF $\|U_k(t)-U(t)\|$

In this section we present the estimates of $\|U_k(t)-U(t)\|$ we use in the algorithms. We suppose $k \geq 1$. We shall use the following lemma.

<u>LEMMA 8.3.1</u>. *Let* p *and* q *be integers with* $-1 \leq q \leq p \leq k$ *and* $p \geq 0$. *Set* $p+q+1 = n$. *Let* $Y\colon [0,1] \to E$ *be* $(n+1)$-*times continuously differentiable on* $[0,1]$ *and suppose*

$$\text{(8.3.1)} \qquad \max_{t \in [0,1]} \left\| \frac{d^{n+1}}{dt^{n+1}} Y(t) \right\| \leq \delta < \infty.$$

Let $Y_n\colon [0,1] \to E$ *be a polynomial of degree* $\leq n$ *satisfying*

$$\text{(8.3.2a)} \qquad Y_n(t_{k-j}) = Y(t_{k-j}) \qquad (j = 0,1,\ldots,p)$$

and

$$\text{(8.3.2b)} \qquad \dot{Y}_n(t_{k-j}) = \dot{Y}(t_{k-j}) \qquad (j = 0,1,\ldots,q).$$

Then

$$\text{(8.3.3)} \qquad \|Y_n(t)-Y(t)\| \leq \frac{\delta}{(n+1)!} \prod_{j=0}^{p} |t-t_{k-j}| \prod_{j=0}^{q} |t-t_{k-j}| \qquad \text{(for all } t \in [0,1]).$$

PROOF. 1. If $E = \mathbb{R}$ then this result can be found in e.g. [STUMMEL & HAINER, 1971; section 3.2.1, 3.2.2].

2. Suppose $E = \mathbb{R}^m$ (with $m > 1$). Let $\tau_0 \in [0,1]$. Suppose $\Delta(\tau_0) \neq 0$ where $\Delta(\tau_0) = Y_n(\tau_0) - Y(\tau_0)$. Then a linear functional $L\colon E \to \mathbb{R}$ exists with $\|L\| = 1$ such that $L\Delta(\tau_0) = \|\Delta(\tau_0)\|$.

Let $LY = \eta$ and $LY_n = \eta_n$. It is easily verified that η is $(n+1)$-times continuously differentiable on $[0,1]$ and that $\max_{t\in[0,1]} |d^{n+1}/dt^{n+1}\ \eta(t)| \le \delta$. Further $\eta_n(t_{k-j}) = \eta(t_{k-j})$ $(j = 0,1,\ldots,p)$ and $\dot{\eta}_n(t_{k-j}) = \dot{\eta}(t_{k-j})$ $(j = 0,1,\ldots,q)$. Consequently (cf. part 1 of this proof)

$$|\eta_n(t)-\eta(t)| \le \frac{\delta}{(n+1)!} \prod_{j=0}^{p} |t-t_{k-j}| \prod_{j=0}^{q} |t-t_{k-j}|$$

$$\text{(for all } t \in [0,1]).$$

In particular, since $\|\Delta(\tau_0)\| = |L\Delta(\tau_0)| = |\eta_n(\tau_0)-\eta(\tau_0)|$, relation (8.3.3) is true for $t = \tau_0$.

Consequently, (8.3.3) is true for all $t \in [0,1]$. This proves the lemma. □

a. Let $p \ge 0$ (with $p \le k-1$) be given. Suppose that U is $(p+2)$-times continuously differentiable, so that a constant $\delta_1 > 0$ exists such that (8.3.1) holds with $Y = U$, $n = p+1$ and $\delta = \delta_1$. Let $U_k = U_{k,p}$ (cf. (8.2.1)). If $u_{k-j} = U(t_{k-j})$ $(j = 0,1,\ldots,p)$ then

$$\|U_{k,p}(t)-U(t)\| \le \|U_{k,p}(t)-U_{k,p+1}(t)\| + \|U_{k,p+1}(t)-U(t)\|$$

$$\le \|U_{k,p}(t)-U_{k,p+1}(t)\| + \frac{\delta_1}{(p+2)!} \prod_{j=0}^{p+1} |t-t_{k-j}|$$

$$\text{(for all } t \in [0,1])$$

(cf. (8.3.3)). It is easily verified that

$$U_{k,p}(t) - U_{k,p+1}(t) \equiv [\prod_{j=0}^{p} (t-t_{k-j})]v, \quad \text{where } v \in E. \tag{8.3.4}$$

In this case we therefore set $\varepsilon_k = \varepsilon_{k,p}$, where

$$\varepsilon_{k,p}(t) \equiv \|U_{k,p}(t)-U_{k,p+1}(t)\| \tag{8.3.5}$$

or, equivalently,

$$
(8.3.6)\qquad
\begin{aligned}
&\varepsilon_{k,p}(t) \equiv \|v\| \prod_{j=0}^{p} |t-t_{k-j}|, \quad \text{where}\\
&v = \{\prod_{j=0}^{p} (t-t_{k-j})\}^{-1}[U_{k,p}(t)-U_{k,p+1}(t)]
\end{aligned}
$$

(for any $t \in [0,1]$ with $t \neq t_{k-j}$ $(j = 0,1,\ldots,p)$).

This kind of estimate of $\|U_k(t)-U(t)\|$ is also used in [RHEINBOLDT, 1975].

<u>REMARK 8.3.1</u>. Determination of p. The - assumed to be unique - solution of (8.1.6) depends on p and we might choose $p = p_k$ such that the solution h of (8.1.6) is maximum. However, in general, too many changes in p are inadvisable. Following [RHEINBOLDT, 1975] we only tested on values of h corresponding to $p = p_{k-1} + \delta$ $(\delta = -1,0,-1)$. Moreover, whenever p was changed we kept it fixed for at least one further step. □

b. Let $p \geq 0$ be given $(p \leq k-1)$. Let $U_k = \bar{U}_{k,p}$ (cf. (8.2.4)). Like case a, we estimate $\|\bar{U}_{k,p}(t)-U(t)\|$ by $\|\bar{U}_{k,p}(t)-\bar{U}_{k,p+1}(t)\|$. As in (8.3.4) it can be shown that

$$\bar{U}_{k,p}(t) - \bar{U}_{k,p+1}(t) \equiv [\prod_{j=0}^{p} (t-t_{k-j})](t-t_k)\bar{v}, \quad \text{where } \bar{v} \in E.$$

We thus set $\varepsilon_k = \bar{\varepsilon}_{k,p}$, where

$$(8.3.7)\qquad \bar{\varepsilon}_{k,p}(t) \equiv \|\bar{U}_{k,p}(t) - \bar{U}_{k,p+1}(t)\|$$

or, equivalently,

$$
(8.3.8)\qquad
\begin{aligned}
&\bar{\varepsilon}_{k,p}(t) \equiv \|\bar{v}\|[\prod_{j=0}^{p} |t-t_{k-j}|]|t-t_k|, \quad \text{where}\\
&\bar{v} = \{[\prod_{j=0}^{p} (t-t_{k-j})](t-t_k)\}^{-1}[\bar{U}_{k,p}(t)-\bar{U}_{k,p+1}(t)]
\end{aligned}
$$

(for any $t \in [0,1]$ with $t \neq t_{k-j}$ $(j = 0,1,\ldots,p)$).

The integer $p = p_k$ is determined in the way described in Remark 8.3.1.

c. Let $U_k = \hat{U}_k$. Suppose U is three-times continuously differentiable. Let $f(u) \equiv -\Gamma(u)F(u_0)$. Thus

$$\dot{U}(t) = f(U(t)) \qquad (t \in [0,1]). \tag{8.3.9}$$

It is easily verified that (cf. p.135)

$$\hat{U}_k(t) \equiv u_k + (t-t_k)[I - (t-t_k)f'(u_k)]^{-1}f(u_k).$$

Define d_1 and d_2 such that

$$\hat{U}_k(t) \equiv u_k + (t-t_k)[I + (t-t_k)f'(u_k)]f(u_k) + d_1(t)$$

and

$$U(t) \equiv U(t_k) + (t-t_k)\dot{U}(t_k) + \frac{1}{2}(t-t_k)^2 \frac{d^2}{dt^2} U(t_k) + d_2(t).$$

It is easily verified that a constant $\delta_1 \in (0,\infty)$ exists such that, when $|t-t_k|$ is sufficiently small,

$$\|d_1(t)\| \leq \delta_1|t-t_k|^3.$$

Suppose

$$\max_{t\in[0,1]} \|\frac{d^3}{dt^3} U(t)\| \leq \delta_2 < \infty.$$

Then

$$\|d_2(t)\| \leq \frac{\delta_2}{6} |t-t_k|^3 \qquad (\text{for all } t \in [0,1]).$$

From (8.3.9) it follows that $d^2/dt^2\, U(t) \equiv f'(U(t))f(U(t))$ $(t \in [0,1])$. Hence, if $u_k = U(t_k)$,

$$\|\hat{U}_k(t)-U(t)\| \leq \|f'(u_k)f(u_k)\|\frac{|t-t_k|^2}{2} + (\delta_1 + \frac{\delta_2}{6})|t-t_k|^3$$

$$(|t-t_k| \text{ sufficiently small}).$$

We can therefore estimate $\|\hat{U}_k(t) - U(t)\|$ by $\hat{\varepsilon}_k(t)$, where

$$\hat{\varepsilon}_k(t) \equiv \|f'(u_k)f(u_k)\| \frac{|t-t_k|^2}{2}$$

or, equivalently,

$$\hat{\varepsilon}_k(t) \equiv \|\Gamma(u_k)F''(u_k)[\Gamma(u_k)F(u_0)]^2\| \frac{|t-t_k|^2}{2} .$$

Similarly, we estimate $\|\check{U}_k(t)-U(t)\|$ (cf. (8.2.6a)) by $\check{\varepsilon}_k(t)$, where

$$(8.3.10a) \quad \check{\varepsilon}_k(t) \equiv \|\Gamma(u_k)\frac{1}{\gamma_3}\{F'(u_k+\gamma_3\Gamma(u_k)F(u_0))-F'(u_k)\}\Gamma(u_k)F(u_0)\|\frac{|t-t_k|^2}{2} .$$

Obviously, $\|\tilde{U}_k(t)-U(t)\|$ (cf. (8.2.6b)) is estimated by $\tilde{\varepsilon}_k(t)$, where

$$(8.3.10b) \quad \tilde{\varepsilon}_k(t) \equiv \|\Gamma(y_{k,\ell-1})\frac{1}{\gamma_3}\{F'(y_{k,\ell-1}+\gamma_3\Gamma(y_{k,\ell-1})F(u_0))-F'(y_{k,\ell-1})\} \Gamma(y_{k,\ell-1})F(u_0)\|\frac{|t-t_k|^2}{2} .$$

8.4. LOCAL PROCESSES; ESTIMATION OF THE RADII OF CONVERGENCE

In this section we present the two types of local processes that are used in the algorithms whose test results are presented in section 8.5.

The first type uses Newton's method, which requires the least amount of work per iteration step of all the methods considered in Chapter 7. Since Newton's method is quadratically convergent, this local process will, in general, only require a few steps. The second type of local process uses a combination of Newton's method and the method $\tilde{M}_{0,0}$ (cf. (7.1.2)), which appears to have very good convergence behaviour (cf. Tables 7.2.1, 7.2.5).

We also tested algorithms in which the methods $M_{\overline{\pi}[N]}$ (cf. (6.3.3), (6.6.13)) are used as local method. (The radius of convergence of these methods was estimated by using the theory of chapter 6, Part II.) However, in contradistinction to the iterative methods in chapter 7, almost all PCC algorithms which were tested, managed to deliver the desired solution (within the required accuracy). Consequently, it is only sensible to compare the PCC algorithms with respect to the total amount of work, which is roughly the total number of local steps required to solve the problem. It appeared that, when using a method $M_{\overline{\pi}[N]}$ or a combination of $M_{\overline{\pi}[N]}$ and Newton's method, the total amount of work was significantly higher than in the case in which Newton's method only was used.

We shall now specify the local process that uses Newton's method, and the one that uses $\tilde{M}_{0,0}$ also. We assume $k \geq 1$. Both local processes contain the following control mechanism (for the k-th local process):

```
          ITER: $y_{k,j+1} = y_{k,j} - d_{k,j}$;
(8.4.1a)  if ($\|d_{k,j}\| < (1+\|y_{k,j}\|)\delta_1$ or $\|Q(t_k,y_{k,j})\| < \delta_2$)
             then "convergence criterion is met"
(8.4.1b)  else if $\|d_{k,j}\| > \gamma_4 r_{k-1}(t_k)$
             then "convergence criterion is not met"
(8.4.1c)  else if (($j \geq 7$ and $\|d_{k,j}\| < \delta_3$ and $\|d_{k,j}\| < \delta_4\|d_{k,j-1}\|$) or $j < 7$)
             then ($j := j+1$; go to ITER)
(8.4.1d)  else
             "convergence criterion is not met".
```

Three remarks on this control mechanism.

1. We shall use $\delta_1 = \delta_2 = 10^{-6}$ in (8.4.1a). Other types of stopping criteria than requirement (8.4.1a) are possible (see e.g. [SCHWETLICK, 1975]). However, it is not yet clear which stopping criterion is optimal (in the sense that it minimizes the total number of local iteration steps), (see also [RIBARIČ & SELIŠKAR, 1974] and [WACKER, 1977b]).

2. In requirement (8.4.1b), the constant $\gamma_4 \geq 2$ is given. If $r_{k-1}(t_k) = r(t_k)$ and if $\|d_{k,j}\| > \gamma_4 r_{k-1}(t_k)$ (for some $j \geq 0$) it seems reasonable to assume that $y_{k,j} \notin B(U(t_k),r(t_k))$. In general $r_{k-1}(t_k) \approx r(t_k)$. In practice it appeared that in cases where requirement (8.4.1b) was met, in general no convergence occurred (we chose $\gamma_4 = 10$).

3. We shall use $\delta_3 = \delta_4 = 10^{-2}$ in (8.4.1c). The requirements (8.4.1a,c) then guarantee that the number of steps in a local process is always less than 10. If after the 7^{th} step no convergence is to be expected, the process is broken off.

In the subsections 8.4.1 and 8.4.2 we assume that the k-th local process has been performed successfully, so that a sequence $\{y_{k,j}\}_{j=0,1,\dots,\ell}$ has been generated where $y_{k,\ell} = u_k$. In these subsections we show how $\{y_{k,j}\}_{j=0,1,\dots,\ell}$ has been obtained and how $r_k(t)$ ($\approx r(t)$) is constructed. r_k will be of the type

$$(8.4.2) \qquad r_k(t) \equiv \rho_k,$$

where ρ_k is an estimate of $r(t_k)$. Set

$$(8.4.3) \qquad P(x) \equiv Q(t_k,x).$$

8.4.1. Newton's method

If Newton's method (subsequently denoted by M_0) is used as local method, then

$$(8.4.4a) \qquad y_{k,j+1} = y_{k,j} - d^0_{k,j}$$

where

$$(8.4.4b) \qquad d^0_{k,j} = [P'(y_{k,j})]^{-1}P(y_{k,j}) \qquad (j = 0,1,\dots,\ell-1).$$

It is easily verified that $P \in F_1$ (cf. pp. 131, 133, and (2.6.1)).

Consider the function a, defined in (6.4.16). For all $\sigma \in (0,\infty]$ we have

$$(8.4.5a) \qquad a(P;\sigma) = \begin{cases} \inf\limits_{0<\|U(t_k)-x\|<\sigma} \left\{\dfrac{(x-U(t_k),[P'(x)]^{-1}P(x))}{\|x-U(t_k)\|^2+\|[P'(x)]^{-1}P(x)\|^2}\right\} \\ \qquad \text{(if } B(U(t_k),\sigma) \subset D(P) \text{ and } P'(x) \text{ is} \\ \qquad \text{invertible on } B(U(t_k),\sigma)\text{)}, \\ -\frac{1}{2} \qquad \text{(otherwise)}. \end{cases}$$

We set

$$(8.4.5b) \qquad a(P;0) = \frac{1}{2}.$$

It follows that (cf. (6.4.18)) $P \in F\langle\sigma, a(P;\sigma)\rangle$ (for all $\sigma \in (0,\infty]$). Let

$$(8.4.6) \qquad s_k = \sup\{\sigma \mid a(P;\sigma) - \frac{2}{5} > 0\}.$$

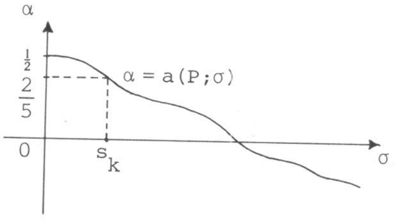


Fig. 8.4.1.

By virtue of Theorem 6.5.1 we have $r(M_0; F\langle\sigma, a(P;\sigma)\rangle) = \sigma$ whenever σ satisfies $a(P;\sigma) - \frac{2}{5} > 0$. Consequently

$$(8.4.7) \qquad s_k \le r(t_k).$$

ρ_k will be an estimate of s_k,

$$(8.4.8) \qquad \rho_k \approx s_k.$$

s_k is determined by $a(P;\cdot)$. However, in general, we do not know $a(P;\cdot)$ explicitly. We therefore estimate $a(P;\cdot)$, using $y_{k,0}, y_{k,1}, \ldots, y_{k,\ell}$.

From Lemma 6.4.1, (6.4.17) and (8.4.5b) it follows that $a(P;\cdot)$ is antitone on $[0,\infty]$, the right-derivative of $a(P;\cdot)$ exists at $\sigma = 0$ and

$$(8.4.9) \qquad \frac{d}{d\sigma} a(P;\sigma) = 0 \qquad (\text{for } \sigma = 0).$$

We therefore estimate s_k as follows. For $j = 0,1,\ldots,\ell-1$ let

$$(8.4.10a) \qquad \sigma_{k,j} = \|y_{k,j} - u_k\|$$

and

$$(8.4.10b) \qquad \alpha_{k,j} = \frac{(y_{k,j} - u_k, d_{k,j})}{\|y_{k,j} - u_k\|^2 + \|d_{k,j}\|^2}$$

where $d_{k,j} = d^0_{k,j}$ is defined by (8.4.4b).

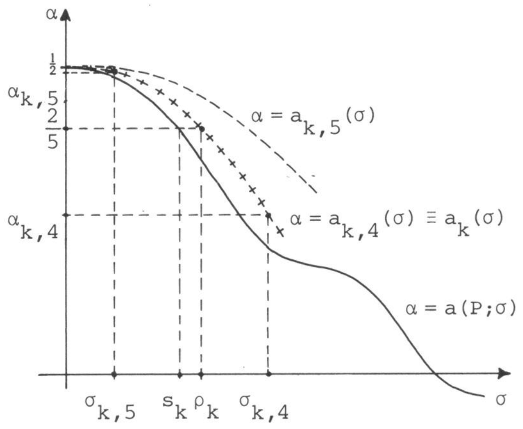


Fig. 8.4.2.

Define

$$(8.4.11a) \qquad a_{k,j}(\sigma) \equiv \frac{1}{2} + c_{k,j}\sigma^2$$

where $c_{k,j}$ is such that

$$(8.4.11b) \qquad a_{k,j}(\sigma_{k,j}) = \alpha_{k,j} \qquad (j = 0,1,\ldots,\ell-1).$$

Set

$$(8.4.12) \qquad a_k(\sigma) \equiv a_{k,j_0}(\sigma), \text{ where } c_{k,j_0} \le c_{k,j} \qquad (j = 0,1,\ldots,\ell-1).$$

Then ρ_k is the solution of $a_k(\sigma) = \frac{2}{5}$ $(\sigma > 0)$. Thus

$$(8.4.13) \qquad \rho_k = [\sqrt{-10c_{k,j_0}}]^{-1}$$

(see also Fig. 8.4.2).

We notice that in [RHEINBOLDT, 1975] also, Newton's method is used as local method. However, in that paper, the estimate $r_k(t)$ of $r(t)$ is based on a result similar to Theorem 6.2.1.

REMARK 8.4.1. It is easily verified that, when $\ell = \ell_k = 1$ and $\|y_{k,0}-y_{k,1}\| \neq 0$, then $c_{k,j_0} = 0$. Hence ρ_k cannot be determined. (In fact $\rho_k = \infty$ so that $t_{k+1} - t_k = h_{max}$ (cf. p.133).) In such a case, either $r(t_k)$ is large (so that $\rho_k =$ is a "good" estimate of $r(t_k)$) or $r(t_k)$ is small and it is only by chance that the k-th local process required only one step. In this last case, the stepsize $h_k = h_{max}$ may be too large. Unfortunately too little information is available to show us which is the case. Consequently, in order to prevent the (k+1)-th local process from failing we put (if $\ell_k = 1$)

$$(8.4.14) \qquad h_k = \sqrt{(t_k - t_{k-1}) \times h_{max}}$$

so that the new step h_k will not differ too much from $t_k - t_{k-1}$. □

REMARK 8.4.2. For k = 1, the control mechanism of the k-th local process should contain an estimate $r_0(t)$ of $r(t)$, (cf. (8.4.1b)). However, r_0 cannot be determined in the way described above without making extra computations. For k = 1 we therefore drop the requirement (8.4.1b) (or, equivalently, we put $r_0(t) = \infty$). □

8.4.2. A mixed method

The method $\tilde{M}_{0,0}$ (cf. (7.1.2) with $N = \theta = 0$) exhibited good convergence behaviour on the test problems of section 7.2. Suppose $\tilde{M}_{0,0}$ is the local method, then $\{y_{k,j}\}_{j=0,1,\ldots,\ell}$ satisfies

$$(8.4.15a) \qquad y_{k,j+1} = y_{k,j} - \tilde{d}_{k,j}$$

where

$$(8.4.15b) \qquad \tilde{d}_{k,j} = [P'(y_{k,j}) - \frac{1}{\tau}\{P'(y_{k,j} + \tau d^0_{k,j}) - P'(y_{k,j})\}]^{-1} P(y_{k,j})$$

$$(j = 0,1,\ldots,\ell-1).$$

In (8.4.15b) the number $\tau > 0$ is a given (small) constant ($\tau = 10^{-4}$) and $d^0_{k,j}$ is defined by (8.4.4b). However, the work per step required for $\tilde{M}_{0,0}$ is roughly twice the work required for Newton's method. In order to reduce the amount of work, we therefore use a mixed method which only uses $\tilde{M}_{0,0}$ when the Newton-step may be expected to be bad. Using this mixed method

$\{y_{k,j}\}_{j=0,1,\ldots,\ell}$ satisfies

$$(8.4.16a) \qquad y_{k,j+1} = y_{k,j} - \bar{d}_{k,j}$$

where

$$(8.4.16b) \qquad \bar{d}_{k,j} = \begin{cases} d^0_{k,j} & \text{(cf. (8.4.4b)) (if } \|d^0_{k,j}\| < \gamma_5 2r^0_{k-1}(t_k)), \\ \tilde{d}_{k,j} & \text{(cf. (8.4.15b) (otherwise)} \end{cases} \qquad (j = 0,1,\ldots,\ell-1).$$

In our test examples we took for the constant γ_5 the values 0, $\frac{1}{2}$ and 1 (cf. section 8.5). If $\gamma_5 = 0$, then the mixed method reduces to $\tilde{M}_{0,0}$. In (8.4.16b) $r^0_{k-1}(t_k)$ is an estimate of the radius of convergence of Newton's method for the k-th local problem. We have assumed the k-th local process has already been performed. In order to perform (8.4.16) for k+1 we should estimate the radius of convergence of Newton's method for the (k+1)-th local problem ($r^0_k(t_{k+1})$). Since in the k-th local process $d^0_{k,j}$ is evaluated in each iteration step, we set $r^0_k(t_{k+1}) = \rho_k$, where ρ_k is defined by (8.4.13).

Let $\bar{r}(t_k)$ denote the radius of convergence of the mixed process (solving the k-th local problem). We shall only use the mixed method if $r(M_0,P) < r(\tilde{M}_{0,0},P)$ is to be expected (cf. (8.4.21)). If this is the case and if γ_5 is not large (cf. (8.4.16b) and (8.5.3)), it may be expected that $\bar{r}(t_k) = r(\tilde{M}_{0,0},P)$. We therefore use an estimate of $r(\tilde{M}_{0,0},P)$ as estimate of $\bar{r}(t_k)$.

We notice that $\tilde{M}_{0,0}$ has been derived from $M_{0,0}$ (cf. (7.1.1)). $[M_{0,0},P]$ is equivalent to $[M_0,\hat{P}]$, where

$$\hat{P}(x) \equiv [P'(x)]^{-1}P(x).$$

(i) Suppose $\{y_{k,j}\}_{j=0,1,\ldots,\ell}$ has been obtained by $[M_{0,0},P]$. Thus

$$y_{k,j+1} = y_{k,j} - \hat{d}_{k,j}$$

where

$$\hat{d}_{k,j} = [\hat{P}'(y_{k,j})]^{-1}\hat{P}(y_{k,j}) \qquad (j = 0,1,\ldots,\ell-1).$$

Similarly to (8.4.13) we can estimate $r(M_0,\hat{P})$, which is equal to $r(M_{0,0},P)$, by

$$\hat{\rho}_k = [\sqrt{-10\hat{c}_{k,j_0}}]^{-1},$$

where $\hat{c}_{k,j_0} = c_{k,j_0}$, and j_0 and c_{k,j_0} are defined by (8.4.10 - 12) with $d_{k,j} = \hat{d}_{k,j}$ $(j = 0,1,\ldots,\ell-1)$.

(ii) Suppose $\{y_{k,j}\}_{j=0,1,\ldots,\ell}$ has been obtained by $[\tilde{M}_{0,0},P]$ (cf. (8.4.15)). Then, similarly, we estimate $r(\tilde{M}_{0,0},P)$ by $\tilde{\rho}_k$, defined by

(8.4.17) $$\tilde{\rho}_k = [\sqrt{-10\tilde{c}_{k,j_0}}]^{-1}.$$

Here, $\tilde{c}_{k,j_0} = c_{k,j_0}$, and j_0 and c_{k,j_0} are defined by (8.4.10 - 12) with $d_{k,j} = \tilde{d}_{k,j}$ $(j = 0,1,\ldots,\ell-1)$.

(iii) Suppose $\{y_{k,j}\}_{j=0,1,\ldots,\ell}$ has been obtained by the mixed method. Hence $\{y_{k,j}\}_{j=0,1,\ldots,\ell}$ satisfies (8.4.16). We would like to estimate the radius of convergence of the (k+1)-th local process by $\tilde{\rho}_k$ (cf. (8.4.17)). However, it may occur that $\bar{d}_{k,j} = d^0_{k,j}$ for some j, so that $\tilde{d}_{k,j}$ is not available. In general it is too expensive to calculate $\tilde{d}_{k,j}$ only for the purpose of estimating $\tilde{\rho}_k$. However, if $\|d^0_{k,j}\|$ is not large, then

$$\begin{aligned} d^0_{k,j} + d^0_{k,j+1} &= \hat{P}(y_{k,j}) + \hat{P}(y_{k,j} - \hat{P}(y_{k,j})) \approx \\ &\approx [I + (I - \hat{P}'(y_{k,j}))]\hat{P}(y_{k,j}) \approx [I - (I - \hat{P}'(y_{k,j}))]^{-1}\hat{P}(y_{k,j}) \\ &= \hat{d}_{k,j} \approx \tilde{d}_{k,j}. \end{aligned}$$

Hence if $\bar{d}_{k,j} \neq \tilde{d}_{k,j}$ and $0 \leq j \leq \ell-2$ we approximate $\tilde{d}_{k,j}$ by $d^0_{k,j} + d^0_{k,j+1}$. Thus, similarly to (8.4.17) we estimate $r(\tilde{M}_{0,0},P)$ by $\bar{\rho}_k$, defined by

(8.4.18) $$\bar{\rho}_k = \gamma_6[\sqrt{-10\bar{c}_{k,j_0}}]^{-1}.$$

Here, $\bar{c}_{k,j_0} = c_{k,j_0}$, and j_0 and c_{k,j_0} are defined by (8.4.10 - 12) with

(8.4.19) $$d_{k,j} = \begin{cases} \tilde{d}_{k,j} & (\text{if } \bar{d}_{k,j} = \tilde{d}_{k,j}), \\ d^0_{k,j} + d^0_{k,j+1} & (\text{if } \bar{d}_{k,j} = d^0_{k,j}). \end{cases}$$

The factor γ_6 reflects the uncertainties in the estimate.

Let $\bar{M}$ denote the mixed method. Suppose the estimate $\varepsilon_k(t)$ of $\|U(t)-U_k(t)\|$ satisfies $\varepsilon_k(t_k+h) = c|h|^2$ (with $c > 0$), (see e.g. (8.3.8) with $p = 0$, and (8.3.10)). Suppose furthermore that $\bar{M} = \tilde{M}_{0,0}$. The method $\tilde{M}_{0,0}$ requires twice as much work per iteration step as Newton's method. It therefore seems reasonable to use $\bar{M}$ $(=\tilde{M}_{0,0})$ only if this implies that the stepsize h_k is at least twice as large as it would be if Newton's method was used. Consequently, in cases where

$$(8.4.20) \qquad r_k^0(t_{k+1}) = \rho_k \geq \gamma_7\bar{\rho}_k = \gamma_7\bar{r}_k(t_{k+1}),$$

where $\gamma_7 = \frac{1}{4}$, it seems advisable to use Newton's method instead of method $\bar{M}$ (cf. (8.1.6)). Obviously, in general $\bar{M} \neq \tilde{M}_{0,0}$. We therefore use Newton's method instead of $\bar{M}$ whenever the (weaker) requirement (8.4.20) with $\gamma_7 = \frac{1}{2}$ (cf. (8.5.3)) is satisfied. We shall denote this combination of Newton's method and the mixed method by $\bar{M}_0$. Hence

$$(8.4.21) \qquad \bar{M}_0 = \begin{cases} \text{Newton's method} & \text{(if (8.4.20) is true),} \\ \text{the mixed method (cf. (8.4.16))} & \text{(otherwise).} \end{cases}$$

<u>REMARK 8.4.3</u>. When $\ell \leq 2$ and Newton's method only has been used in the k-th local process, $\bar{\rho}_k$ cannot be determined (when $\ell = 2$, then $\bar{c}_{k,j_0} = 0$). In that case we use Newton's method only in the (k+1)-th local process. □

<u>REMARK 8.4.4</u>. In the first local process we use Newton's method only. See also Remark 8.4.2. □

8.5. NUMERICAL RESULTS WITH THE PCC ALGORITHMS

In all the test examples that are to be presented, we use

$$(8.5.1) \qquad \begin{aligned} \delta_1 &= \delta_2 = 10^{-6} \text{ (cf. (8.4.1a),} \\ \delta_3 &= \delta_4 = 10^{-2} \text{ (cf. (8.4.1c)),} \end{aligned}$$

in the control mechanisms for the local methods.

1. As a first example, we consider the behaviour of several PCC algorithms when applied to Problem 1 (cf. p.125). The control parameters were chosen as follows:

$$\begin{aligned} &\gamma_1 = 1 \quad (\text{cf. } (8.1.6)), \qquad \gamma_2 = \frac{1}{2} \quad (\text{cf. p.133}), \\ (8.5.2) \qquad & \\ &\gamma_3 = 10^{-4} \quad (\text{cf. } (8.2.6\text{b})), \qquad \gamma_4 = 10 \quad (\text{cf. } (8.4.1\text{b})). \end{aligned}$$

Furthermore, we set

$$h_{start} = h_{min} = 10^{-2}, \quad h_{max} = 1 \qquad (\text{cf. p.133}).$$

In Table 8.5.1 results are given where Newton's method is used as local method.

TABLE 8.5.1.

PCC algorithm.

Local method: M_0.

Problem 1 ($\lambda = 0.1$, $m = 99$).

ε	predictor	total number of t-steps	total number of local steps	total amount of work
0.01	$U_{k,p}$	6	25	25
0.01	$\bar{U}_{k,p}$	5	20	20
0.01	$\tilde{U}_k$	4	16	21
0.001	$U_{k,p}$	8	37	37
0.001	$\bar{U}_{k,p}$	6	27	27
0.001	$\tilde{U}_k$	6	23	30

The total amount of work has been expressed in the number of Newton-steps required. (We assumed that the evaluation of $\tilde{U}_k(t_{k+1})$ $(k \geq 1)$ was equivalent to one Newton-step.) As an example, we specify the performance of one of the algorithms.

TABLE 8.5.2.

PCC algorithm.

Predictor: $U_{k,p}$.

Local method: M_0.

Problem 1 ($\varepsilon = 0.001$, $\lambda = 0.1$, $m = 99$).

k	t_k	p	$r_{k-1}(t_k)$ (estimate of the radius of conv.)	number of local steps
1	0.010	0	-	3
2	0.143	1	$0.278_{10}1$	6
3	0.227	1	$0.111_{10}1$	5
4	0.364	2	$0.886_{10}0$	5
5	0.524	2	$0.576_{10}0$	5
6	0.715	2	$0.400_{10}0$	5
7	0.944	2	$0.278_{10}0$	5
8	1	2	$0.191_{10}0$	3

We also used the combination of the mixed method and Newton's method as local method. Its control parameters were chosen as follows:

$$\gamma_5 = \frac{1}{2} \quad (\text{cf. } (8.4.16b)),$$

$$\gamma_6 = 1 \quad (\text{cf. } (8.4.18)), \tag{8.5.3}$$

$$\gamma_7 = \frac{1}{2} \quad (\text{cf. } (8.4.20)).$$

TABLE 8.5.3.

PCC algorithm.

Predictor: $\tilde{U}_k$.

Local method: $\bar{M}_0$ ($\tau = 10^{-4}$, cf. (8.4.21), (8.4.15b)).

Problem 1 ($\lambda = 0.1$, $m = 99$).

ε	total number of t-steps	total number of		total amount of work
		M_0-steps	$\tilde{M}_{0,0}$-steps	
0.01	3	12	0	16
0.001	4	15	1	22

When both an iterative method and a PCC algorithm manage to solve a problem $F(x) = 0$, both starting from the same $u_0 = x_0$, the latter, in general, requires much more work than the former (see e.g. Table 7.2.1 (for $\tilde{M}_{0,0}$) and Table 8.5.3). This phenomenon is due to the fact that PCC algorithms contain many control mechanisms which prevent divergence. They are therefore reliable but also rather expensive.

2. Secondly, we consider the behaviour of PCC algorithms when applied to Problem 3, a problem given in [RHEINBOLDT, 1975].

<u>PROBLEM 3</u>. This problem arises from a finite element approach to the two-dimensional boundary value problem

$$\text{(8.5.4a)} \qquad \begin{aligned} &\frac{\partial}{\partial v}[f(T;v,w)\frac{\partial}{\partial v}T(v,w)] + \frac{\partial}{\partial w}[f(T;v,w)\frac{\partial}{\partial w}T(v,w)] = c \\ &\qquad\qquad\qquad\qquad\qquad\qquad ((v,w) \in \Theta = [0,1]\times[0,1]), \\ &T(v,w) = 0 \quad (\text{if } (v,w) \in \partial\Theta, \text{ the boundary of } \Theta). \end{aligned}$$

In (8.5.4a)

$$\text{(8.5.4b)} \qquad f(T;v,w) \equiv q([\frac{\partial}{\partial v}T(v,w)]^2 + [\frac{\partial}{\partial w}T(v,w)]^2)$$

and

$$\text{(8.5.4c)} \qquad q(s) = \begin{cases} q_0 & (\text{if } s \le 0.15), \\ \frac{1}{2}(q_0+q_1) + \frac{1}{4}(q_1-q_0)(3\bar{s}-\bar{s}^3) & (\text{if } 0.15 < s < 0.5,\ \bar{s} = \frac{40s-13}{7}), \\ q_1 & (\text{if } s \ge 0.5). \end{cases}$$

Let $m \ge 1$. Consider the $m\times m$-dimensional problem $F(x) = 0$

$$\text{(8.5.5a)} \qquad \phi_{i,j}(x) = 0$$

where

$$\text{(8.5.5b)} \qquad \begin{aligned} \phi_{i,j}(x) = {} & Q^C_{i,j}(x)\xi_{i,j} - Q^N_{i,j}(x)\xi_{i,j+1} - Q^W_{i,j}(x)\xi_{i-1,j} \\ & - Q^S_{i,j}(x)\xi_{i,j-1} - Q^E_{i,j}(x)\xi_{i+1,j} + \Delta^2 c \\ & \qquad\qquad\qquad\qquad (i,j = 1,2,\dots,m). \end{aligned}$$

Here, $\xi_{i,j}$ and $\phi_{i,j}(x)$ are the $((i-1)m+j)$-th components of respectively x and F(x). Moreover,

$$(8.5.5c) \qquad \Delta = \frac{1}{m+1} \,,$$

$$(8.5.5d) \qquad \xi_{k,\ell} = 0 \quad (\text{if } (k\Delta,\ell\Delta) \in \partial\Theta \ (k,\ell = 0,1,\ldots,m+1)).$$

Further,

$$(8.5.5e) \qquad \begin{aligned}
Q^{N}_{i,j}(x) &= \tfrac{1}{2}\{q(\delta_v(i,j+1)^2 + \delta_w(i,j)^2) + q(\delta_v(i-1,j)^2 + \delta_w(i,j)^2)\},\\
Q^{W}_{i,j}(x) &= \tfrac{1}{2}\{q(\delta_v(i-1,j)^2 + \delta_w(i,j)^2) + q(\delta_v(i-1,j)^2 + \delta_w(i-1,j-1)^2)\},\\
Q^{S}_{i,j}(x) &= \tfrac{1}{2}\{q(\delta_v(i-1,j-1)^2 + \delta_w(i,j-1)^2)\\
&\qquad\qquad + q(\delta_v(i,j)^2 + \delta_w(i,j-1)^2)\},\\
Q^{E}_{i,j}(x) &= \tfrac{1}{2}\{q(\delta_v(i,j)^2 + \delta_w(i,j-1)^2) + q(\delta_v(i,j)^2 + \delta_w(i+1,j)^2)\},\\
Q^{C}_{i,j}(x) &= Q^{N}_{i,j}(x) + Q^{W}_{i,j}(x) + Q^{S}_{i,j}(x) + Q^{E}_{i,j}(x) \quad (i,j = 1,2,\ldots,m)
\end{aligned}$$

and

$$(8.5.5f) \qquad \delta_v(k,\ell) = \frac{1}{\Delta}\,(\xi_{k+1,\ell} - \xi_{k,\ell}), \quad \delta_w(\ell,k) = \frac{1}{\Delta}\,(\xi_{\ell,k+1} - \xi_{\ell,k})$$
$$(k = 0,1,\ldots,m;\ \ell = 0,1,\ldots,m+1).$$

The starting point is $u_0 = 0$. For the problems we solved, (8.1.2) (or, equivalently, (8.1.8)) has a unique continuous solution. Setting $x^* = U(1)$, we define D(F) in a similar way as in Problem 2 (cf. (7.2.6)). Although $F \notin \mathcal{F}_1$ (F need not be twice differentiable on D(F) if $q_0 \neq q_1$) we tested PCC algorithms on (8.5.5). The computations were performed for m = 5 and 11, c = 15 and $q_0 = 1$; for q_1 we took $q_1 = 10$, 25 and 50. □

In the examples to be given, the control parameters γ_1, γ_2, γ_3 and γ_4 satisfy (8.5.2). We chose

$$(8.5.6) \qquad h_{start} = 7.10^{-2}, \quad h_{min} = 2.10^{-2}, \quad h_{max} = 5.10^{-1} \quad (\text{cf. p. } 133).$$

Furthermore, following Rheinboldt, we let $h_1 = h_{start}$.

In Table 8.5.4 results are given where Newton's method is used as local method.

TABLE 8.5.4.

PCC algorithm.

Local method: M_0.

Problem 3 ($c = 15$, $m = 5$).

q_0	q_1	predictor	total number of t-steps	total number of local steps	total amount of work
1	10	$U_{k,p}$	11	53	53
1	10	$\bar{U}_{k,p}$	8	37	37
1	10	$\tilde{U}_k$	7	28	36
1	50	$U_{k,p}$	14	72	72
1	50	$\bar{U}_{k,p}$	11	56	56
1	50	$\tilde{U}_k$	8	34	43

We also used the combination of the mixed method and Newton's method as local method. Its control parameters were chosen as follows:

$$(8.5.7) \qquad \gamma_6 = \frac{1}{4} \text{ (cf. (8.4.18))}, \qquad \gamma_7 = \frac{1}{2} \text{ (cf. (8.4.20))}.$$

For γ_5 (cf. (8.4.16b)) we took $\gamma_5 = 0$ and $\gamma_5 = 1$.

TABLE 8.5.5.

PCC algorithm.

Predictor: $\tilde{U}_k$.

Local method: $\bar{M}_0$ ($\tau = 10^{-4}$, cf. (8.4.21), (8.4.15b)).

Problem 3 ($c = 15$, $m = 5$).

	Problem		total number	total number of		total amount
γ_5	q_0	q_1	of t-steps	M_0-steps	$\tilde{M}_{0,0}$-steps	of work
0	1	10	8	16	14	53
1	1	10	6	25	1	34
0	1	50	8	18	16	59
1	1	50	6	29	1	38

As an example, we specify the performance of a PCC algorithm, using $\bar{M}_0$ as local method.

TABLE 8.5.6.

PCC algorithm.

Predictor: $\tilde{U}_k$.

Local method: $\bar{M}_0$ ($\gamma_5 = 1$, $\tau = 10^{-4}$, cf. (8.4.21), (8.4.15b)).

Problem 3 ($c = 15$, $q_0 = 1$, $q_1 = 50$, $m = 5$).

		estimate $r_{k-1}(t_k)$ of the radius of conv. of		number of local steps
k	t_k	M_0	$\tilde{M}_{0,0}$	
1	0.070	-	-	1
2	0.140	-	-	7
3	0.279	$0.165_{10}-1$	$0.544_{10}-1$	7^*
4	0.366	$0.102_{10}-1$	$0.217_{10}-1$	5
5	0.588	$0.538_{10}-2$	$0.133_{10}-1$	6
6	1	$0.394_{10}-2$	$0.142_{10}-1$	4

* The local process included one $\tilde{M}_{0,0}$-step.

We also tested PCC algorithms on Problem 3 where $q_0 = 1$, $q_1 = 25$ and $m = 11$. This case was also considered in [RHEINBOLDT, 1975]. We present results for the following two PCC algorithms:

PCCI, which uses $\tilde{U}_k$ as predictor and $\bar{M}_0$ (with control parameters $\gamma_5 = 1$, $\gamma_6 = \frac{1}{4}$ and $\gamma_7 = \frac{1}{2}$) as local method.

PCCII, which uses $U_{k,p}$ as predictor and Newton's method as local method.

In both algorithms the control parameters γ_1, γ_2, γ_3 and γ_4 satisfy (8.5.2), and h_{start}, h_{min} and h_{max} satisfy (8.5.6). Further, $h_1 = h_{start}$.

TABLE 8.5.7.

PCC algorithm.

Problem 3 ($c = 15$, $q_0 = 1$, $q_1 = 25$, $m = 11$).

Algorithm	total number of t-steps	total number of		total amount of work
		M_0-steps	$\tilde{M}_{0,0}$-steps	
RHEINBOLDT*	18	82	-	82
PCCI	10	47	1	60
PCCII	FAILURE**			

* The numbers have been obtained from [RHEINBOLDT, 1975].

** The 5th local process broke off at $t_5 \approx 0.17$ while $t_5 - t_4 = h_{min}$.

CONCLUSION. From tables 8.5.1 and 8.5.4 it appears that the predictors $\bar{U}_{k,p}$ (see (8.2.4)) and $\tilde{U}_k$ (see (8.2.6b)) allow greater t-steps than the predictor $U_{k,p}$ (see (8.2.1)). Furthermore, the total amount of work required to solve the problem is significantly less when one of the first two predictors is used rather than the last one.

The predictor $\tilde{U}_k$ requires the least t-steps to solve the problem. As an evaluation of $\tilde{U}_k(t)$ is more expensive than an evaluation of $\bar{U}_{k,p}(t)$, algorithms using $\bar{U}_{k,p}$ or $\tilde{U}_k$ require about the same amount of work.

The algorithms using the method $\bar{M}_0$ in which the mixed method with $\gamma_5 > 0$ is used (see (8.4.21) and (8.4.16b)) required less work than the ones using Newton's method only as local method.

With regard to the total amount of work, from Table 8.5.5 it appears that the method $\bar{M}_0$ in which the mixed method with $\gamma_5 = 0$ is used is inferior to the method $\bar{M}_0$ in which the mixed method with $\gamma_5 = 1$ is used. (We recall that the mixed method with $\gamma_5 = 0$ is equivalent to the method $\tilde{M}_{0,0}$ (cf. (7.1.2)).)

Obviously, more numerical experiments are required in order to assess this combination of the mixed method and Newton's method.

We note that, although the algorithm in [RHEINBOLDT, 1975] and the algorithm PCCII both use the predictor $U_{k,p}$ and Newton's method, the latter failed to solve the problem with which Table 8.5.7 is concerned. This may be due to a better choice of control parameters, which are unknown to us, in the former. It may also be due to the different ways in which the radii of convergence of the local processes are estimated.

We finally notice that the differential equation (8.1.8) is of the type (2.6.11) where $H = Q$ and $g = 0$. In view of the computational results of chapter 7 it also seems worth while to investigate PCC algorithms that use the differential equation (2.6.11) with $H = Q$ and $g \neq 0$.

In conclusion, PCC algorithms appear to be very reliable for solving the problem $F(x) = 0$. Among the algorithms tested, the ones in which Davidenko's method is used, require the least amount of work. With suitably chosen control parameters, the combination of the mixed method and Newton's method appears to require less work as local method than Newton's method alone. It is clear, however, that many questions remain open for future research.

REFERENCES

AVILA, J.H. JR., 1974, *The feasibility of continuation methods for nonlinear equations,* SIAM J. Num. Anal. 11, 102-122.

BERGER, M.S., 1977, *"Nonlinearity and Functional Analysis",* Academic Press, New York.

BITTNER, L., 1967, *Einige kontinuierliche Analogien von Iterationsverfahren,* in: *"Funktionalanalysis, Approximationstheorie, Numerische Mathematik",* ISNM 7, pp. 114-135, Birkhauser, Basel.

BOGGS, P.T., 1971, *The solution of nonlinear systems of equations by A-Stable integration techniques,* SIAM J. Num. Anal. 8, 767-785.

BOGGS, P.T. & J.E. DENNIS JR., 1974, *A continuous analogue analysis of nonlinear iterative methods,* Cornell University, Dept. of Computer Science, Techn. Rep. TR 74-200.

BOSARGE, W.E., 1971, *Iterative continuation and the solution of nonlinear two-point boundary-value problems,* Num. Math. 17, 268-283.

BROWN, A.L. & A. PAGE, 1970, *"Elements of Functional Analysis",* Van Nostrand Reinhold Company, London.

BROYDEN, C., 1969, *A new method of solving nonlinear simultaneous equations,* Comp. J. 12, 94-99.

BUS, J.C.P., 1975, *A comparitive study of programs for solving nonlinear equations,* Mathematisch Centrum, Amsterdam, Rep. NW 25-75.

BYRNE, G.D. & A.C. HINDMARSH, 1975, *A polyalgorithm for the numerical solution of ordinary differential equations,* A.C.M. Trans. on Math. Software 1, 71-96.

CARNAHAN, B., H.A. LUTHER & J.O. WILKES, 1969, *"Applied Numerical Methods",* Wiley, New York.

COURANT, R., 1961, *"Vorlesungen über Differential- und Integralrechnung",* Erster Band; third (improved) print, Springer, Berlin.

DAHLQUIST, G., 1963, *A special stability problem for linear multistep methods,* BIT 3, 27-43.

DAVIDENKO, D.F., 1953, *On a new method of numerically integrating a system of nonlinear equations,* (Russian), Dokl. Akad. Nauk. SSSR 88, 601-604.

DAVIDENKO, D.F., 1965a, *An application of the method of variation of parameters to the construction of iterative formulas of higher accuracy for the determination of the elements of the inverse matrix,* Dokl. Akad. Nauk. SSSR 162, 743-746; Soviet Math. Dokl. 6, 738-742.

DAVIDENKO, D.F., 1965b, *An application of the method of variation of paramaters to the construction of iterative formulas of increased accuracy for numerical solutions of nonlinear integral equations,* Dokl. Akad. Nauk. SSSR 162, 499-502; Soviet Math. Dokl. 6, 702-706.

DAVIDENKO, D.F., 1966, *On the construction of iterative processes of increased accuracy by the method of variation of a parameter,* (Russian), Abstracts of Session 14, p. 31, International Congress of Mathematicians, Moscow, 1966.

DAVIDENKO, D.F., 1975, *An iterative method of parameter variation for inverting linear operators,* USSR Comp. Math. Phys. 15(1), 27-43.

DEIST, F. & L. SEFOR, 1967, *Solution of systems of nonlinear equations by parameter variation,* Comp. J. 10, 78-82.

DEUFLHARD, P., 1974, *A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting,* Num. Math. 22, 289-315.

DEUFLHARD, P., 1976, *A stepsize control for continuation methods with special application to multiple shooting techniques,* Technische Universität München, Institut für Mathematik, Techn. Rep. TUM-MATH 7627.

DEUFLHARD, P., H.J. PESCH & P. RENTROP, 1976, *A modified continuation method for the numerical solution of nonlinear two-point boundary value problems by shooting techniques,* Num. Math. 26, 327-343.

DI LENA, G. & D. TRIGIANTE, 1976, *Metodo di Euler e ricerca delle radici di una equazione,* Calcollo 13, 377-396.

FEILMEIER, M., 1972, *Numerische Aspekte bei der Einbettung nichtlineare Probleme,* Computing 9, 355-364.

FICKEN, F., 1951, *The continuation method for functional equations,* Comm. Pure Appl. Math. 4, 435-456.

GAVURIN, M., 1958, *Nonlinear functional equations and continuous analogues of iterative methods,* (Russian), Izv. Vysš. Učebn. Zaved. Matematica 6, 18-31.

GEAR, C.W., 1971, *"Numerical Initial Value Problems in Ordinary Differential Equations",* Prentice-Hall, Englewood Cliffs, New Jersey.

GRAGG, W.B. & R.A. TAPIA, 1974, *Optimal error bounds for the Newton-Kantorovich theorem,* SIAM J. Num. Anal. 11, 10-13.

HOUWEN, P.J. VAN DER, 1977, *"Construction of integration formulas for initial value problems",* North-Holland, Amsterdam.

KANTOROWITSCH, L.W. & G.P. AKILOW, 1964, *"Funktional Analysis in Normierten Raümen",* Akademie-Verlag, Berlin.

KELLER, H.B., 1968, *"Numerical Methods for Two-Point Boundary-Value Problems",* Blaisdell, London.

KITCHEN, J.W., 1966, *Concerning the convergence of iterates to fixed points,* Studia Math. 27, 247-249.

KIZNER, W., 1964, *A numerical method for finding solutions of nonlinear equations,* SIAM J. Appl. Math. 12, 424-428.

KLEINMICHEL, H., 1968, *Stetige Analoga und Iterationsverfahren für nichtlineare Gleichungen in Banachraüme,* Math. Nachr. 37, 313-343.

KOGAN, T.I., 1967, *An iteration process for functional equations,* (Russian), Sibirsk. Mat. Ž. 8, 958-960.

KUBIČEK, M., 1976, *Dependence of solution of nonlinear systems on a parameter,* A.C.M. Trans. on Math. Software 2, 98-107.

LAASONEN, P., 1970, *An imbedding method of iteration with global convergence,* Computing 5, 253-258.

LAHAYE, E., 1934, *Une méthode de résolution d'une catégorie d'équations transcendantes,* C.R. Acad. Sci. Paris 198, 1840-1842.

LAHAYE, E., 1935, *Sur la représentation des racines systèmes d'équations transcendantes,* Deuxième Congrès National des Sciences 1, 141-146.

LAMBERT, J.D., 1973, *"Computational Methods in Ordinary Differential Equations"*, Wiley, London.

LEDER, D., 1974, *Automatische Schrittweitensteuerung bei global konvergenten Einbettungsmethoden,* ZAMM 54, 319-324.

MENZEL, R. & H. SCHWETLICK, 1976, *Über einen Ordnungsbegriff bei Einbettungsalgorithmen zur Lösung nichtlinearer Gleichungen,* Computing 16, 187-199.

MEYER, G.H., 1968, *On solving nonlinear equations with a one-parameter operator imbedding,* SIAM J. Num. Anal. 5, 739-752.

MURRAY, J.D., 1968, *A simple method for obtaining approximate solutions for a class of diffusion-kinetics enzyme problems,* Math. Biosciences 2, 379-411.

ORTEGA, J.M. & W.C. RHEINBOLDT, 1970, *"Iterative Solution of Nonlinear Equations in Several Variables"*, Academic Press, New York.

OSTROWSKI, A.M., 1960, *"Solution of Equations and Systems of Equations"*, Academic Press, New York; third edition 1973.

PETRY, W., 1971, *Eine Verallgemeinerung des Newtonsche Iterationsverfahren,* Computing 7, 25-45.

POMENTALE, T., 1974, *Homotopy iterative methods for polynomial equations,* J. Inst. Math. Applic. 13, 201-213.

RALL, L.B., 1968, *Davidenko's method for the solution of nonlinear operator equations,* University of Wisconsin, Mathematics Research Center, MRC Techn. Summary Rep. 948.

RALL, L.B., 1969, *"Computational Solution of Nonlinear Operator Equations"*, Wiley, New York.

RHEINBOLDT, W.C., 1975, *An adaptive continuation process for solving systems of nonlinear equations,* University of Maryland, Computer Science Techn. Rep., TR - 393.

RHEINBOLDT, W.C., 1976, *Numerical continuation methods for finite element applications,* University of Maryland, Computer Science Techn. Rep., TR - 454.

RIBARIČ, M. & M. SELIŠKAR, 1974, *An optimization of stepsize in the continuation method,* Mat. Balkanica 4, 517-521.

ROSENBLOOM, P., 1956, *The method of steepest descent,* Sixth Symp. Appl. Math., pp. 126-176, Amer. Math. Soc., Providence, Rhode Island.

RUDIN, W., 1953, *"Principles of Mathematical Analysis"*, McGraw-Hill, New York.

RUDIN, W., 1973, *"Functional Analysis"*, McGraw-Hill, New York.

SCHRÖDER, E., 1870, *Über unendlich viele Algorithmen zur Auflosung der Gleichungen,* Math. Ann. 2, 317-365.

SCHWETLICK, H., 1975, *Ein neues Prinzip zur Konstruktion implementierbarer, global konvergenter Einbettungsverfahren,* Beitr. Num. Math. 4, 215-228.

SCHWETLICK, H., 1976, *Ein neues Prinzip zur Konstruktion implementierbarer, global konvergenter Einbettungsalgorithmen (Testbeispiele),* Beitr. Num. Math. 5, 201-206.

SPIJKER, M.N., 1972, *Equivalence theorems for nonlinear finite-difference methods,* in: *"Numerisch Lösung nichtlinearer partieller Differential- und Integrodifferentialgleichungen"*, Lecture Notes 267, pp. 233-264, Springer, Berlin.

STETTER, H.J., 1973, *"Analysis of Discretization Methods for Ordinary Differential Equations"*, Springer, Berlin.

STUMMEL, F. & K. HAINER, 1971, *"Praktische Mathematik"*, Teubner, Stuttgart.

TRAUB, J.F., 1964, *"Iterative Methods for the Solution of Equations"*, Pentrice-Hall, Englewood Cliffs, New Jersey.

UEBERHUBER, C.W., 1976, *Optimierung mittels kontinuierlicher Quasi-Newton-Verfahren und deren Diskretisierungen,* Technische Universität Wien, Institut für Numerische Mathematik, Bericht nr. 19/76.

WACKER, H., 1971, *Eine Lösungsmethode zur Behandlung nichtlinearer Randwertprobleme,* in: *"Iterationsverfahren, Numerische Mathematik, Approxmationstheorie"*, ISNM 15, pp. 245-257, Birkhauser, Basel.

WACKER, H., 1974, *Übergangsmöglichkeiten zwischen verschiedene Iterationsverfahren,* ZAMM 54, T236 - T237.

WACKER, H., 1977a, *Globalisierung und Aufwandminimierung für ein Iterationsverfahren von Kivistik,* ZAMM 57, T306 - T308.

WACKER, H., 1977b, *Minimierung des Rechenaufwands bei Globalisierungen spezieller Iterationsverfahren vom Typ Minimales Residuum,* Computing 18, 209-224.

WASSERSTROM, E., 1973, *Numerical solution by the continuation method,* SIAM Review 15, 89-119.

SUMMARY

In this monograph we are concerned with the numerical solution of the equation

$$(*) \qquad F(x) = 0.$$

Here F is a nonlinear operator from D into E, where E is a (real) Hilbert space and D a subset of E. Let x^* be the (unknown) solution of (*).

A well-known iterative method for approximating x^* is Newton's method. However, if the starting point $x_0 \in D$ is remote from x^*, then in general x^* cannot be approximated by this method. In many such cases, so-called imbedding methods succeed in approximating x^* numerically. We restrict our attention to the imbedding

$$H(t,x) = (1-t)K(x,x_0) + tF(x) \qquad (0 \le t \le 1,\ x \in D)$$

where K is an operator from D×D into E for which K(x,x) = 0 (for all $x \in D$). Let X be the solution of the initial value problem

$$\binom{*}{*} \qquad \begin{aligned} \dot{X}(t) &= -\left[\frac{\partial}{\partial x} H(t,X(t))\right]^{-1} \frac{\partial}{\partial t} H(t,X(t)) \qquad (0 \le t \le 1), \\ X(0) &= x_0. \end{aligned}$$

With certain restrictions on K and F it follows that $x = X(t)$ is the solution of

$$(*^{*}*) \qquad H(t,x) = 0 \quad (t \in [0,1]),$$

so that $X(1) = x^*$.

In this thesis iterative methods are investigated that are based on the successive numerical integration of $\binom{*}{*}$ with changing initial value x_0. The numerical integration is performed by means of (generalized) Runge-Kutta methods. Hence these iterative methods depend on K and the Runge-Kutta method used.

The investigations are performed using the concept "radius of convergence of an iterative method" (cf. section 2.4).

In chapter 5 the restrictions on K and the Runge-Kutta method which are sufficient for the iterative process to have a positive radius of convergence are given.

In chapter 6 the radii of convergence of some iterative methods of this type are determined. Iterative methods are presented that have a greater radius of convergence than Newton's method.

In chapter 7 numerical results are presented. In particular, an iterative method based on a generalized Runge-Kutta method related to the Backward Euler method appears to be very reliable when x_0 is a remote from x^*.

In chapter 8 algorithms are described that solve successively $(*{}^{*}*)$ for $t = t_1, t_2, \ldots, t_N$ (with $0 = t_0 < t_1 < \ldots < t_N = 1$) using an iterative method. In some of these algorithms the differential equation $(\overset{*}{*})$ is also used. The numbers $t_1, t_2, \ldots, t_N$ are determined during the process itself. Some numerical results are presented. The algorithms appear to be very reliable. The algorithms in which differential equation $(\overset{*}{*})$ is used, appear to require the least amount of work in order to solve $(*)$.

SAMENVATTING

In dit proefschrift houden we ons bezig met het numeriek oplossen van de vergelijking

$$(*)\qquad F(x) = 0.$$

Hierbij is F een niet-lineaire afbeelding van D in E, waarbij E een (reële) Hilbertruimte is en D een deelverzameling van E. x^* is de (onbekende) oplossing van (*).

Een bekende iteratieve methode om x^* te benaderen is de methode van Newton. Als echter het startpunt $x_0 \in D$ niet dichtbij x^* ligt, dan kan x^* meestal niet benaderd worden met deze methode. Het blijkt dat in veel van dergelijke gevallen x^* wèl kan worden benaderd met behulp van z.g. inbeddingsmethoden. Wij beperken ons tot de inbedding

$$H(t,x) = (1-t)K(x,x_0) + tF(x) \qquad (0 \le t \le 1,\ x \in D)$$

waarbij K een afbeelding is van D×D in E waarvoor $K(x,x) = 0$ (voor alle $x \in D$). X is de oplossing van het beginwaardeprobleem

$$\binom{*}{*}\qquad \begin{aligned} &\dot{X}(t) = -\Big[\frac{\partial}{\partial x} H(t,X(t))\Big]^{-1} \frac{\partial}{\partial t} H(t,X(t)) \qquad (0 \le t \le 1), \\ &X(0) = x_0. \end{aligned}$$

Onder bepaalde aannamen voor K en F geldt dat $x = X(t)$ de oplossing is van

$$(*\,{}^{*}\,*)\qquad H(t,x) = 0 \qquad (t \in [0,1]),$$

zodat $X(1) = x^*$.

In dit proefschrift worden iteratieve methoden onderzocht die gebaseerd zijn op successief numeriek integreren van $\binom{*}{*}$ met wisselende beginwaarde x_0. De numerieke integratie vindt plaats met (gegeneraliseerde) Runge-Kutta methoden. De iteratieve methoden zijn dus gebaseerd op K en de gebruikte Runge-Kutta methode.

Het onderzoek wordt gedaan aan de hand van het begrip "convergentiestraal van een iteratieve methode" (zie sectie 2.4).

In hoofdstuk 5 worden voorwaarden voor K en de Runge-Kutta methode opgesteld waaronder het iteratieve proces een positieve convergentiestraal bezit.

In hoofdstuk 6 wordt voor enige van dergelijke iteratieve methoden de convergentiestraal berekend. Er worden methoden gepresenteerd met een convergentiestraal die groter is dan die van de methode van Newton.

In hoofdstuk 7 worden enige numerieke resultaten gepresenteerd. Vooral een iteratieve methode die gebaseerd is op een gegeneraliseerde Runge-Kutta methode die verwant is aan de achterwaartse methode van Euler, blijkt bijzonder betrouwbaar als x_0 niet dichtbij x^* ligt.

In hoofdstuk 8 worden algoritmen beschreven die de vergelijking $(*^{*}*)$ achtereenvolgens voor $t = t_1, t_2, \ldots, t_N$ (waarbij $0 = t_0 < t_1 < \ldots < t_N = 1$) met een iteratieve methode oplossen. In enige van deze algoritmen wordt ook gebruik gemaakt van de differentiaalvergelijking $(^{*}_{*})$. De getallen $t_1, t_2, \ldots, t_N$ worden tijdens het proces bepaald. Er worden enige numerieke resultaten gepresenteerd. De algoritmen blijken bijzonder betrouwbaar. De algoritmen waarin ook van de differentiaalvergelijking $(^{*}_{*})$ gebruik wordt gemaakt, blijken de minste hoeveelheid werk nodig te hebben om $(*)$ op te lossen.

CURRICULUM VITAE

De schrijver van dit proefschrift werd geboren te 's-Gravenhage, op 12 januari 1951.

In 1968 legde hij het examen H.B.S.-B af aan het Christelijk Lyceum Populierstraat te 's-Gravenhage. In hetzelfde jaar begon hij zijn studie in de wiskunde aan de Rijksuniversiteit te Leiden. Hij volgde colleges in de wiskunde bij de hoogleraren dr. C. Visser, dr. A.C. Zaanen, dr. G. Zoutendijk, dr. W.R. van Zwet, dr. A.A. Verrijn Stuart, dr. M.N. Spijker en dr. J. Fabius, alsmede bij dr. A. Ollongren. In 1974 legde hij het doctoraalexamen met hoofdvak wiskunde af.

Van 1970 tot 1974 was hij als candidaats-assistent werkzaam op het Centraal Reken-Instituut van de Leidse universiteit. In 1974 trad hij, als wetenschappelijk medewerker, in dienst van het Mathematisch Centrum te Amsterdam, waar hij de gelegenheid kreeg het onderzoek te verrichten, dat in dit proefschrift is beschreven.

STELLINGEN

bij het proefschrift

THE NUMERICAL SOLUTION OF
NONLINEAR OPERATOR EQUATIONS
BY IMBEDDING METHODS

van

C. DEN HEIJER

14 februari 1979

I

De in [1] genoemde stelling kan gegeneraliseerd worden tot de volgende stelling.

Laat E een complexe Hilbert ruimte zijn. Zij $L > 0$ en laat $T: E \to E$ een niet-lineaire afbeelding zijn waarvoor geldt dat $\|T(u)-T(v)\| \leq L\|u-v\|$ en $\mathrm{Re}(T(u)-T(v),u-v) \geq 0$ (voor alle $u,v \in E$). Zij $f \in E$, en $S: E \to E$ gedefinieerd door $S(u) \equiv -T(u)+f$. $t_0,t_1,t_2,\ldots$ zijn niet-negatieve reële getallen waarvoor geldt dat

$$\sum_{n=0}^{\infty} t_n\left(1-\frac{t_n}{2}(1+L^2)\right) = \infty.$$

Onder deze aannamen convergeert het iteratieve proces

$$v_{n+1} = (1-t_n)v_n + t_n S(v_n) \qquad (n = 0,1,2,\ldots)$$

voor elke $v_0 \in E$ naar de unieke oplossing van $u+T(u) = f$.

[1] DOTSON, W.G. JR., *An iterative process for nonlinear monotonic nonexpansive operators in Hilbert space,* Math. Comp. 32 (1978), pp. 223-225.

II

Bij het bewijs van stelling 22.1 in [2] over een z.g. "point of attraction" van een interatief proces in $\mathbb{R}^n$, wordt gebruik gemaakt van de vorm van Jordan van een matrix. Dit laatste is niet noodzakelijk: door gebruik te maken van de formule

$$(*) \qquad \rho(C) = \lim_{n\to\infty} \|C^n\|^{\frac{1}{n}},$$

kan men het bewijs vereenvoudigen en zelfs generaliseren, zodat het resultaat ook geldt voor een iteratief proces in een Banach ruimte X. In (*) is $\rho(C)$ de spectraal straal van een lineaire operator C in X. Het bewijs kan nog meer vereenvoudigd worden door gebruik te maken van de relatie

$$\rho(C) = \inf\{|C| \mid |\cdot| \in N\}$$

waarbij N de verzameling is van alle normen op X die equivalent zijn met de gegeven norm $\|\cdot\|$. Zie I(1.4) in [3] en 16.2 in [4].

[2] OSTROWSKI, A.M., *"Solution of Equations in Euclidean and Banach Spaces"*, Academic Press, New York, 1973.

[3] KRASNOSEL'SKII, M.A., e.a., *"Approximate Solution of Operator Equations"*, (D. Louvish (vert.)), Wolters-Noordhoff, Groningen, 1972.

[4] NASHED, M.Z., *Differentiability and related properties of non-linear operators*, in: *"Nonlinear Functional Analysis and Applications"*, L.B. Rall (red.), Academic Press, New York, 1971.

III

Het gebruik van de term "point of repulsion" in stelling 22.2 van [2] is verwarrend.

IV

Het bewijs van propositie 4.2 in [5] is niet correct.

[5] TAPIA, R.A., *Differentiation and integration*, in: *"Nonlinear Functional Analysis and Applications"*, L.B. Rall (red.), Academic Press, New York, 1971.

V

De in [6] bewezen stelling leidt tot een minder scherp resultaat dan stelling 6.2.1 van dit proefschrift.

[6] RALL, L.B., *A note on the convergence of Newton's method*, SIAM J. Num. Anal. 11 (1974), pp. 34-36.

VI

Het bewijs van stelling 26.1 in [7] kan aanzienlijk vereenvoudigd en bekort worden. Bovendien kan een scherper resultaat worden verkregen.

[7] RALL, L.B., *"Computational Solution of Nonlinear Operator Equations"*, Wiley, New York, 1969.

VII

Dat een slordig geformuleerde definitie makkelijk tot fouten leidt, wordt geïllustreerd door de bewering (2.1.11), die niet correct is, en het bewijs van stelling 2.1.19, dat fout is, in [8]. Deze fouten zijn een gevolg van de niet-precieze formulering in definitie 2.1.4 in [8].

[8] BERGER, M.S., *"Nonlinearity and Functional Analysis"*, Academic Press, New York, 1977.

VIII

Zij

$$a_n = 0 \qquad (n = 1,2,3,\ldots).$$

Indien men de begrippen "limiet" en "convergente reeks" hanteert zoals die in [9] omschreven zijn, dan is

(i) 0 niet de limiet van de rij $\{a_n\}$;

(ii) $a_1 + a_2 + a_3 + \ldots$ geen convergente reeks.

[9] VAN DALE, *"Groot Woordenboek der Nederlandse Taal"*, 10^e druk, C. Kruyskamp (red.), Martinus Nijhoff, 's-Gravenhage, 1976

IX

Het verdient overweging, ter bevordering van het gebruik van de fiets in het woon-werkverkeer, douches te installeren in kantoorgebouwen.

Volkskrant, 23/12/1978.