ERRATA

Numerical Continuation Methods for Nonlinear Equations and Bifurcation Problems

by James P. Abbott

page 3, line 10: delete first comma.

page 13, lines 1 and 12: for "$\eta(\partial_x G(x^*,h))$" read "$\|\partial_x G(x^*,h)\|$".

page 16, line 4: for "(x^*,h^*)" read "x^*".

page 18, line 14: for "$\|e_{im}\|$" read "$B_1\|e_{ki}\|$".

page 27, line 16: for "$\mathrm{Det}(J(x_{i+1})) \neq \mathrm{Det}(J(x_0))$"
read "$\mathrm{sgn}(\mathrm{Det}(J(x_{i+1}))) \neq \mathrm{sgn}(\mathrm{Det}(J(x_0)))$".

page 42, line 12: insert "are" at end of line.

page 53, line -7: for "aften" read "often".

page 66, line 13: insert comma before "equal to 1".

page 74, line -2: for "section 5.5" read "section 6.4".

page 86, line 6: delete "a system of".

page 95, line -3: for "then" read "than".

page 101, line 6: for "m^2" read "$(m-1)^2$".

page 123, line -2: for "$(x-1)^2$" read "$(x_1-1)^2$".

page 124, line 3: replace the sentence beginning "This is ..."
by "This is in contrast to Branin's method
which is convergent for any $x_0 \notin \{x | x_1=1, x_1=0 \text{ or } x_1=-0.75\}$
i.e. the solution trajectory of (6.2.1) passes
through all three zeros for any such x_0".

page 131, line 14: for "euqations" read "equations".

NUMERICAL CONTINUATION METHODS FOR NONLINEAR EQUATIONS AND BIFURCATION PROBLEMS

by

James P. Abbott

A thesis submitted to the

Australian National University

for the degree of Doctor of Philosophy

June, 1977

ACKNOWLEDGEMENTS

The work for this thesis was undertaken at the Computer Centre, the Australian National University, and I gratefully acknowledge the financial assistance of the Australian National University during this time.

I am indebted to Richard Brent for his help and supervision of the work and for his constructive comments on a draft of this thesis. I am grateful to Mike Osborne and Bob Watts for their continual support and for many clarifying discussions. I also thank Professor W.C. Rheinboldt for his prompt replies to my requests for advice and for his comments on a précis of Chapter 5.

My thanks are also due to the typist Mrs Barbara Geary whose care and skill are clearly evident in the thesis and to Erin Brent for her help with the graph plotting.

Finally I thank my wife, Peta, for her careful reading of both the draft and final thesis. Without her general encouragement and support this work would not have been possible and for this I owe a great debt of gratitude.

PREFACE

Some of the work of this thesis was carried out in collaboration with Dr Richard Brent. In particular, Chapters 2, 3 and 4 contain results which were established jointly. Also, Chapters 2 and 3 have been published as Abbott and Brent [2].

Elsewhere in the thesis, unless otherwise stated, the work described is my own.

James P. Abbott

ABSTRACT

This thesis investigates some aspects of the continuation method for the solution of a system of nonlinear equations, $f(x) = 0$, $f : D \subset R^n \to R^n$. This approach is useful for generating methods which do not rely on a good initial estimate of a solution and the problem is converted to one of following the solution trajectory $x(t)$ of a problem of the form $H(x(t),t) = 0$, $H : D \subset R^n \times R \to R^n$, from the starting guess $x_0 = x(0)$, hopefully to the solution x^* .

In Chapter 1 we give a brief introduction and note that $x(t)$ also satisfies

$$\dot{x}(t) = -\partial_x H(x,t)^{-1} \partial_t H(x,t) , \qquad x(0) = x_0 ,$$

and so we can follow $x(t)$ by applying methods traditionally used for the solution of ordinary differential equations. In Chapter 2 we consider general single-step methods and, in particular, Runge-Kutta methods, for following $x(t)$. We also give conditions on the methods to attain rapid convergence to x^* and, as a result, for a particular choice of $H(x,t)$ we are able to derive methods which have improved rates of convergence to x^* . We apply similar arguments in Chapter 3 to the class of linear multistep methods and again generate methods which follow $x(t)$ accurately and then give rapid final convergence to x^* .

In Chapter 4 we consider Newton-like methods for finding $x(t_i)$ for a sequence of values $\{t_i\}$, and discuss the accuracy and computational efficiency of the methods. We use the results of Chapter 2 to derive a method which changes in a continuous way from one which follows $x(t)$ accurately to one which converges rapidly to x^* .

Chapter 5 is concerned with problems where the need to follow the

solution of $H(x(t),t) = 0$ arises naturally. We consider, in particular, the difficulties associated with certain critical points, i.e. points on the solution branch $(x(t),t)$ at which $\partial_x H(x,t)$ is singular. We describe an efficient method for following a branch through a simple turning point and present an efficient method for determining such turning points accurately. This method is also useful for finding certain simple bifurcation points.

Finally, in Chapter 6, we consider the problem of finding several solutions of the equation $f(x) = 0$. We consider two recent approaches and show that the two methods are essentially the same. A reformulation of one of the methods indicates a technique which is, in some sense, more efficient than the other methods.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In this thesis we consider some aspects of numerical methods for the solution of nonlinear equations in several variables. We are interested in methods which do not rely on the availability of a good estimate of a solution. Such methods can be derived by embedding the given problem in a class of problems formulated so that the method of solution becomes one of following a trajectory in R^n, $n > 1$. Some of the theory developed for these methods is also relevant in two related applications. The first is in problems where the need to follow a trajectory arises naturally, often called bifurcation problems, and the second is in the problem of finding several solutions of a system of nonlinear equations. We consider each of these problem areas in this work.

Recently various different, but related, methods have been proposed for the solution of a system of nonlinear equations when only a poor initial estimate of a zero is known. These methods all use the continuation approach which, in principle, goes back to the last century but appears to have been used as a numerical tool for the first time by Lahaye [43], [44]. Historical surveys can be found in Ficken [25] and Avila [4]. Suppose we wish to find a zero x^* of the function $f : D \subset R^n \to R^n$. We embed this problem in a family of problems of the form

$$H(x(t),t) = 0 \tag{1.1}$$

where $t \in [0,\tau)$, for some $\tau > 0$. (τ may be infinite but, for brevity, we do not specifically distinguish this case.) The embedding is chosen so that, for $t = 0$, the solution $x(t)$ of (1.1) is known to be x_0, i.e. $x(0) = x_0$, and $x(\tau)$ is the required solution x^*. For the general problem (1.1), Rheinboldt [60] gives sufficient conditions on $H(x,t)$ for

$x(t)$ to exist for each $t \in [0,\tau)$. Also, in section 2.1, we give a theorem which, for a particular choice of $H(x,t)$, gives sufficient conditions for $x(\tau)$ to equal x^* . Similar results for particular choices of $H(x,t)$ can be found in e.g. [23], [28], [50] and [72]. Then the problem of solving

$$f(x) = 0 \tag{1.2}$$

becomes one of following the solution trajectory from $x(0) = x_0$ to $x(\tau) = x^*$.

The most common choice for $H(x,t)$ is

$$H(x,t) = f(x) - (1-t)f(x_0) \tag{1.3}$$

for which $x(0) = x_0$ and $x(1) = x^*$. Another example is to transform the embedding parameter of (1.3) to the infinite interval to give

$$H(x,t) = f(x) - e^{-t}f(x_0) \ , \tag{1.4}$$

where e is the base of the natural logarithm, and then $x^* = \lim_{t\to\infty} x(t)$.

It appears to have been Davidenko [19] who first considered converting (1.1) to an ordinary differential equation. By application of the chain-rule, it follows that the solution of (1.1) satisfies the initial value problem

$$\dot{x}(t) = -\partial_x H(x,t)^{-1}\partial_t H(x,t) \ , \qquad x(0) = x_0 \ , \tag{1.5}$$

where $\partial_x H(x,t)$ represents the Frechet partial derivative of $H(x,t)$ with respect to x . For (1.3) and (1.4) this gives

$$\dot{x}(t) = -J(x)^{-1}f(x_0) \ , \qquad x(0) = x_0 \ , \tag{1.6}$$

and

$$\dot{x}(t) = -J(x)^{-1}f(x) \ , \qquad x(0) = x_0 \ , \tag{1.7}$$

respectively, where $J(x)$ is the Jacobian of f at x . Note that the solution trajectories of (1.3) and (1.4), and therefore of (1.6) and (1.7), are essentially the same, the difference is only in the choice of parameter-

isation. Subsequent to Davidenko's original work, various authors have suggested integrating (1.3) or (1.6) (e.g. [10], [15], [40], [50], [53], [72]) and (1.4) or (1.7) (e.g. [9], [11], [16], [28]) whilst others have suggested less general choices of $H(x,t)$, usually dependent upon the form of f (e.g. [20], [22], [24], [27], [41], [50], [72]).

The differential equation (1.7) was also derived, using an alternative approach, by Gavurin [28]. He considered a general iterative process of the form

$$x_{i+1} = x_i + hg(x_i) \quad , \tag{1.8}$$

where h represents a steplength, and, by taking the limit as $h \to 0$, generated the continuous analogue of (1.8),

$$\dot{x}(t) = g(x) \quad . \tag{1.9}$$

Then (1.7) represents the continuous analogue of Newton's method. In a recent application of continuation, Kellogg, Li and Yorke [39] used the continuous analogue of a combination of the Newton and direct iteration methods, in a constructive proof of the Brower fixed point theorem. Their differential equation is

$$\dot{x}(t) = -\left(J(x)+\mu(x)^{-1}I\right)^{-1}f(x) \quad , \tag{1.10}$$

where I is the unit matrix and $\mu : R^n \to R$ is such that $\mu(x) \to -\infty$ as x approaches a solution of (1.2). Equation (1.10) gives an approach somewhat in the style of the Levenberg/Marquardt method for optimisation [47], [48].

Gavurin notes that each zero of $f(x)$ is a stable node of the autonomous differential equation (1.7), i.e. stable in the sense of Liapunov [45], and so difference formulae used to integrate (1.7) should enjoy a similar stability. This is not necessarily the case, since equations of the form (1.9) can be Liapunov stable but also be stiff [18] in which case standard difference formulae may not be stable. This is actually not the case for (1.7), at least close to a solution, although it was the concern of Boggs [8], [9]. In Chapter 2 we discuss a suggestion of Boggs that the most

suitable methods for the solution of (1.7) are the A-stable techniques of Dahlquist [18]. Also Boggs noted that integrating (1.6) requires a greater concern for accuracy than is required when integrating (1.7). This is because, under reasonable conditions, all solutions of $\dot{x}(t) = -J(x)^{-1}f(x)$ converge locally to x^*, which is a consequence of the Liapunov stability, and this is not the case for (1.6). Thus we concern ourselves primarily with the use of (1.4) and (1.7).

When the solution $x(t)$ of (1.7) converges to x^*, any method which, because of small steps or high accuracy, follows the trajectory sufficiently closely will surely converge to x^* also. However this convergence will be slow since $x(t)$ converges to x^* only linearly. This follows because, from (1.4), $f(x(t)) = e^{-t}f(x_0)$. Therefore, for an algorithm to be efficient, there must be a change of emphasis at some stage from accurate representation of $x(t)$ to rapid convergence to x^*. In Chapters 2-4 we consider methods for the solution of the differential equation (1.7) which can, by suitable step length control, be induced to give rapid final convergence to x^*. In Chapter 2 we present some general results on the convergence of one step methods with variable step size and use these results to derive Runge-Kutta methods suitable for integrating (1.7) and which can give rapid final convergence to x^*. In Chapter 3 we present general results on the convergence of multistep methods and use the results to generate methods which can give high order accuracy in following the solution of (1.7) and then give rapid final convergence to x^*. We also discuss the stability problems involved with such methods if the step size is varied. Then in Chapter 4 we direct attention to methods of solving (1.1) for a sequence of values of t, using Newton-like methods. We consider their orders of accuracy in following the solution of (1.1) and also their computational efficiency. We apply these results to the cases when $H(x,t)$ is given by (1.3) and (1.4). We also derive a method, which

has certain desirable order and convergence properties, for integrating (1.7).

Problems of the form given in (1.1) often arise naturally in a form where it is necessary to find the value of $x(t)$ for sufficient values of t to define the solution $(x(t),t)$. The formulation describes how the state vector $x(t)$ depends upon the control parameter t . There is a large literature on the theoretical and numerical analysis of such problems, much of it being in the theory of elasticity where $x(t)$ represents the position of a structure and t represents a physical load. See for example [3], [6], [17], [33], [36], [38], [66], [69] and the references therein. Much of the analysis is involved with critical points on the solution $(x(t),t)$ of (1.1), i.e. points at which $\partial_x H(x(t),t)$ is singular, and the behaviour of the solution in the region of such critical points. As mentioned above, some methods are described in Chapter 4 which are suitable for following solutions of (1.1). In Chapter 5 we develop these methods for the specific problem of following a solution through a certain kind of critical point, known as a turning point. We suggest an improved technique, similar to the methods suggested by Riks [66] and Menzel and Schwetlick [49]. Turning points represent the boundary between stability and instability of a system and, as such, are of special interest. For example, Simpson [69] gives a numerical method for finding such points. In Chapter 5 we also consider this problem and present some methods which are more efficient than Simpson's method. It happens that the derived methods are also useful for finding certain simple bifurcation points, which are another example of critical points. One of the methods provides information useful for finding points on a secondary solution which emanates from a simple bifurcation point [37], [64].

Methods for following a solution of (1.1) are also of interest in the problem of finding several solutions of (1.2) and this is the concern of

Chapter 6. The usual approach is to solve (1.2) using a standard iterative procedure with several starting guesses. However, this method often has the failing that it continually converges to a solution which is already known. In Chapter 6 we consider two suggestions, the first by Branin [11] and the second, a deflation method by Brown and Gearhardt [14], for overcoming this problem. Branin uses the continuation principle by integrating (1.7) both forwards and backwards and tries to find all the solutions on a particular trajectory. Whilst Branin's method can only be guaranteed to find all the zeros of f under special circumstances (see e.g. [16]) the general approach appears to be the best currently available. We consider a reformulation of the Brown and Gearhardt method which indicates that it is essentially the same as Branin's method. This reformulation also indicates a possible improvement to the deflation technique giving a method which proves to be, in some sense, more efficient than the other two methods.

CHAPTER 2

CONTINUATION WITH SINGLE-STEP METHODS

2.1. Introduction

As a preliminary to the main results of this chapter we present, in section 2.2, a convergence result for the continuation methods introduced in Chapter 1 for solving $f(x) = 0$, where $f : D \subset R^n \to R^n$. This result is not new in principle, but it specifies the type of conditions required on f before convergence to x^* can be guaranteed. It also indicates that the continuation method is not a panacea for problems with a poor starting guess, but that it can often widen the region of convergence. The theorem gives conditions on f and x_0 for the solution of

$$\dot{x}(t) = -J(x)^{-1}f(x) , \qquad x(0) = x_0 , \tag{2.1.1}$$

to converge to x^*.

Following section 2.2, we consider the application of single-step methods to the problem of integrating (2.1.1) and, in particular, we are interested in the use of explicit Runge-Kutta schemes. Our purpose is to find methods which can follow the solution of (2.1.1) accurately, in some sense, and can also give rapid local convergence to x^*. In section 2.3 we generalise the local convergence theory of Ostrowski [58] to single-step methods involving a variable steplength and, in section 2.4, we apply these results to Runge-Kutta schemes for integrating (2.1.1). The resulting theory shows that, with odd-order Runge-Kutta methods, it is possible to gain rapid convergence to x^* by suitable choice of the step size. Also, in section 2.4, we challenge a suggestion of Boggs [9] that the most suitable methods for the solution of (2.1.1) are the A-stable methods of Dahlquist [18]. Finally, in section 2.5, we give the results of some numerical experiments

and compare the methods suggested by the theory with some existing methods.

2.2. A Convergence Result

In this section we consider the differential equation (2.1.1), where $f(x)$ is assumed to be continuously differentiable for all $x \in D$. There are a great many theorems on the existence and uniqueness of solutions of (2.1.1) (see e.g. [4], [8], [50], [53], [60], [72] and the references therein) but most are local in nature. Since the differential equation approach is concerned with wider convergence we present a theorem which is not local. The theorem is not new, having been proved with marginally greater assumptions on f by Gavurin [28], Deuflhard [23] and Ortega and Rheinboldt [53], but is given for clarity and as motivation for the overall approach. Its purpose is to characterise a region in which solutions of (2.1.1) are guaranteed to converge to a zero of f. First we give some definitions.

DEFINITION 2.2.1. $P \subset D$ is *a region of stability* of (2.1.1) if, for any $x_0 \in P$, the solution $x(t)$ of (2.1.1) is defined and unique for all $t \geq 0$, $x(t) \in P$ for all $t \geq 0$ and $\lim_{t\to\infty} x(t) = x^* \in P$, where x^* is a zero of f.

For any nonsingular $n \times n$ matrix A define $\phi_A : D \subset R^n \to R$ by

$$\phi_A(x) = f(x)^T A^T A f(x)$$

and, for any $\alpha > 0$, define $P_\alpha(A)$ by

$$P_\alpha(A) = \{x \mid x \in D,\ \phi_A(x) \leq \alpha\} .$$

$P_\alpha(A)$ is a level set of $\phi_A(x)$, (see [23], [53]). Let $L = \{x \mid x \in D,\ \mathrm{Det}(J(x)) = 0\}$. Then, for some $\alpha > 0$ and $P^*_\alpha(A)$, a path connected component of $P_\alpha(A)$, condition A will be

A : $P_\alpha^*(A) \cap L$ and $P_\alpha^*(A) \cap \partial D$ are empty, $P_\alpha^*(A)$ is bounded.

Under these conditions $P_\alpha^*(A)$ is compact and contains one and only one zero of f.

THEOREM 2.2.1. *Assume* $f : D \subset R^n \to R^n$ *is continuously differentiable on* D *and* $\alpha > 0$ *is such that condition* A *holds. If, in addition,* $J(x)^{-1}f(x)$ *is Lipschitz continuous on* $\mathrm{Int}\left(P_\alpha^*(A)\right)$ *then* $\mathrm{Int}\left(P_\alpha^*(A)\right)$ *is a region of stability of* (2.1.1).

Proof. Standard theorems on ordinary differential equations (e.g. [32, Chapter 1]) show that, for any $x_0 \in \mathrm{Int}\left(P_\alpha^*(A)\right)$, there exists a $\tau > 0$ such that (2.1.1) has a solution which is unique in $\mathrm{Int}\left(P_\alpha^*(A)\right)$ for each $t \in [0,\tau)$. Also, if the maximal such τ is not ∞ and $\{x(t) \mid 0 \le t < \tau\}$ has limit point x_τ, then $x_\tau \in \partial P_\alpha^*(A)$.

When the solution $x(t)$ of (2.1.1) exists it satisfies

$$f\left(x(t)\right) = e^{-t}f(x_0) = e^{-t}f_0 , \tag{2.2.1}$$

say, because (2.1.1) is equivalent to the initial value problem $df/dt = -f, f(0) = f_0$. Thus

$$\phi_A\left(x(t)\right) = e^{-2t}\phi_A(x_0) , \qquad t \in [0,\tau) ,$$

and so $\phi_A\left(x(t)\right)$ is a decreasing function of t. Thus

$$\phi_A(x_\tau) = \lim_{t\to\tau-} \phi_A\left(x(t)\right) < \alpha .$$

Now suppose, if possible, that $x_\tau \in \partial P_\alpha^*(A)$. Since $P_\alpha(A)$ is closed and $P_\alpha^*(A) \cap \partial D$ is empty there exists an $\varepsilon > 0$ such that $S(x_\tau,\varepsilon) \subset D$ and $S(x_\tau,\varepsilon) \cap \{P_\alpha(A) \backslash P_\alpha^*(A)\}$ is empty, where $S(x,\varepsilon)$ is the open ball with centre x and radius ε. Let $\varepsilon_i = \varepsilon/i$, then because $x_\tau \in \partial P_\alpha^*(A)$, for each $i > 0$ there exists a $y_i \in S(x_\tau,\varepsilon_i)$ such that $\phi_A(y_i) > \alpha$. Now

$\lim_{i\to\infty} y_i = x_\tau$ and, by continuity of $\phi_A(x)$, $\lim_{i\to\infty} \phi_A(y_i) = \phi_A(x_\tau) \geq \alpha$, which is a contradiction. Thus $x_\tau \in \mathrm{Int}\left(P_\alpha^*(A)\right)$ and it follows that $\tau = \infty$, so $x(t)$ is defined and $x(t) \in \mathrm{Int}\left(P_\alpha^*(A)\right)$ for all $t \geq 0$. Also, from (2.2.1), if x_∞ is a limit point of $\{x(t)\}$, then $f(x_\infty) = 0$. Since a zero of f is unique in $P_\alpha^*(A)$ it follows that $x_\infty = x^* = \lim_{t\to\infty} x(t)$. This completes the proof. □

We note that a sufficient condition for $J(x)^{-1}f(x)$ to be Lipschitz continuous on $\mathrm{Int}\left(P_\alpha^*(A)\right)$ is that, in addition to condition A, $J(x)$ be Lipschitz continuous on $\mathrm{Int}\left(P_\alpha^*(A)\right)$. This follows from the fact that $\|J(x)^{-1}\|$ and $\|f(x)\|$ are bounded on $P_\alpha^*(A)$ and $f(x)$ is continuously differentiable (and hence Lipschitz continuous) on $P_\alpha^*(A)$.

Whilst Theorem 2.2.1 is not practically useful, it shows that around each zero at which $J(x)$ is nonsingular there is a region of stability of (2.1.1). Also this region will generally be larger than that predicted by the local existence theorems. We emphasise that if x_0 is not in such a region then convergence to a root is unpredictable. We discuss this case further in Chapter 6.

In Chapters 1, 2 and 3 we assume that x_0 is contained in a region of stability and that the solution trajectory of (2.1.1) converges to a zero x^*. If this is the case then, by following the trajectory closely enough, we can guarantee convergence to x^*. For this purpose several of the standard methods for solving initial value problems may be employed and, for sufficiently small steps, convergence to x^* is certain. In practice however, we would like to take large steps. Far from the zero this entails using a sophisticated step size estimator which will adapt the step according to the function behaviour and choose it to be as large as possible

consistent with sufficient accuracy. Obviously the lower the accuracy the less work will be involved but the higher the probability of leaving the correct trajectory and diverging or finding the wrong solution.

Close to the solution, however, we can make use of the special characteristics of the problem to give rapid final convergence, using methods which are also suitable for following the trajectory far from the solution. In this and the following chapter we consider single and multistep methods, traditionally used for the standard initial value problem, which are adapted to give rapid convergence close to the zero x^* .

2.3. General Theory

In this section we give some general results on iterative processes of the form

$$x_{i+1} = G(x_i, h_i) , \qquad i = 0,1,\ldots , \tag{2.3.1}$$

where $G : D \times D_h \subset R^n \times R \to R^n$, and in the following sections we apply these results to particular iterations. We use the results of Ostrowski [58] and Ortega and Rockoff [54] on processes of the form $x_{i+1} = G(x_i)$, $G : D \subset R^n \to R^n$, and generalise the existing theory to include the extra variable. We quote the following definitions which can be found in [53], except that here suitable modification has been made to allow for the slight generalisation.

Let $C(I,x^*)$ denote the set of all sequences generated by an iterative process I with limit point x^* . Let $\{x_k\} \subset R^n$ be any sequence that converges to x^* . Then the *R-convergence factors of the sequence* are the numbers

$$R_p\{x_k\} = \begin{cases} \limsup_{k\to\infty} \|x_k - x^*\|^{1/k} & \text{, if } p = 1 \text{ ,} \\ \limsup_{k\to\infty} \|x_k - x^*\|^{1/p^k} & \text{, if } p > 1 \text{ .} \end{cases}$$

The *R-convergence factor of* I *at* x^* is defined by

$$R_p(I,x^*) = \sup\{R_p\{x_k\} \mid \{x_k\} \in C(I,x^*)\}$$

and the quantity

$$O_R(I,x^*) = \begin{cases} \infty & \text{if } R_p(I,x^*) = 0 \text{ for all } p \in [1,\infty) \text{ ,} \\ \inf\{p \in [1,\infty) \mid R_p(I,x^*) = 1\} & \text{otherwise} \end{cases}$$

is called the *R-order of* I *at* x^* . We say that the convergence of I at x^* is *superlinear* if $R_1(I,x^*) = 0$ and *linear* if $0 < R_1(I,x^*) < 1$.

Let $G : D \times D_h \subset R^n \times R \to R^n$, then x^* is *a point of attraction* of the iterative process (2.3.1) if there exists an open neighbourhood S of x^* and a set I , called the *h-domain of* I , such that $S \subset D$, $I \subset D_h$ and for any $x_0 \in S$ and any $\{h_i\} \subset I$ the sequence $\{x_i\}$ remains in D and converges to x^* . Also we say that x^* is a *fixed point* of the iteration (2.3.1) if $x^* = G(x^*,h)$ for all $h \in D_h$.

Finally, we say that $G(x,h)$ is *uniformly differentiable with respect to* x *at* $z \in D$ *on* $I \subset D_h$ if, for each $h \in I$, $G(x,h)$ is Frechet differentiable with respect to x at z and if, for any $\varepsilon > 0$, there exists a $\delta > 0$, independent of h , such that $S(z,\delta) \subset D$ and

$$\|G(x,h) - G(z,h) - \partial_x G(z,h)(x-z)\| \le \varepsilon\|x-z\|$$

for all $x \in S(z,\delta)$ and for all $h \in I$.

We can now give conditions on $G(x,h)$ which are sufficient for x^* to be a point of attraction of (2.3.1). In this chapter and the next, if A is a square matrix, $\eta(A)$ will denote the spectral radius of A .

THEOREM 2.3.1. *Suppose that* $G : D \times D_h \subset R^n \times R \to R^n$ *has a fixed*

point $x^* \in \mathrm{Int}(D)$. *Let* $I_\alpha \subset D_h$ *be such that* $\eta\left(\partial_x G(x^*,h)\right) \leq \alpha < 1$ *for all* $h \in I_\alpha$ *and suppose that* $G(x,h)$ *is uniformly differentiable with respect to* x *at* x^* *on* I_α . *Then, if* I_α *is non empty,* x^* *is a point of attraction of iteration* (2.3.1) *with h-domain* I_α .

Proof. The proof is almost identical to that given for the Generalised Ostrowski Theorem in [53] and so is omitted. □

Theorem 2.3.1 is rather more general than we require and so we present a corollary which is more suitable for our purposes.

COROLLARY 2.3.1. *Suppose* $G : D \times D_h \subset R^n \times R \to R^n$ *has a fixed point* $x^* \in \mathrm{Int}(D)$. *Suppose also that* $\partial_x G(x,h)$ *and* $\partial_h G(x,h)$ *are Lipschitz continuous on* $S \times I_\alpha$ *where* S *is an open convex neighbourhood of* x^* *and* I_α *is an interval such that* $\eta\left(\partial_x G(x^*,h)\right) \leq \alpha < 1$ *for all* $h \in I_\alpha$. *If* I_α *is nonempty then* x^* *is a point of attraction of iteration* (2.3.1) *with h-domain* I_α .

Proof. It follows from the Lipschitz continuity of $\partial_x G(x,h)$ and $\partial_h G(x,h)$ and from [53, Theorem 3.2.5] that, for all $(x,h) \in S \times I_\alpha$, there exists a constant K such that

$$\|G(x,h)-G(x^*,h)-\partial_x G(x^*,h)(x-x^*)\| \leq K\|x-x^*\|^2 .$$

This result is immediate if we assume that

$$\left\|\begin{matrix} a \\ \alpha \end{matrix}\right\| = \|a\| + |\alpha| ,$$

however the result follows anyway if we use the equivalence of norms. Now, given $\varepsilon > 0$, if $K\delta < \varepsilon$ then, for all $h \in I_\alpha$,

$$\|G(x,h)-G(x^*,h)-\partial_x G(x^*,h)(x-x^*)\| \leq \varepsilon\|x-x^*\|$$

for all $x \in S(x^*,\delta)$. Thus $G(x,h)$ is uniformly differentiable with

respect to x at x^* on I_α. The result now follows from Theorem 2.3.1.□

Corollary 2.3.1 gives sufficient conditions for local convergence of the iterative process (2.3.1) but gives no information on the rate of convergence. For this we require conditions on $\{h_i\}$. We begin by deriving a result on the assumption that $\lim_{i\to\infty} h_i$ exists.

THEOREM 2.3.2. *Suppose* $G : D \times D_h \subset R^n \times R \to R^n$ *has a fixed point* $x^* \in \mathrm{Int}(D)$, *and that* $\partial_x G(x,h)$ *and* $\partial_h G(x,h)$ *are Lipschitz continuous in a neighbourhood of* (x^*,h^*), *where* $\lim_{i\to\infty} h_i = h^* \in \mathrm{Int}(D_h)$. *If* $\alpha = \eta(\partial_x G(x^*,h^*)) < 1$ *then* x^* *is a point of attraction of the iterative process* I *given by* (2.3.1). *Moreover*

$$R_1(I,x^*) = \alpha$$

and if $\alpha > 0$ *then* $O_R(I,x^*) = 1$.

Proof. Define $u(x,h)$ by

$$G(x,h) = G(x^*,h) + \partial_x G(x^*,h)(x-x^*) + u(x,h) . \tag{2.3.2}$$

Then, as in the proof of Corollary 2.3.1, there exist positive constants K_1, δ and δ_2 such that

$$\|u(x,h)\| \le K_1\|x-x^*\|^2 \tag{2.3.3}$$

for all $x \in S(x^*,\delta)$, $h \in (h^*-\delta_2, h^*+\delta_2) = I_2$ say. Furthermore, from the Lipschitz continuity of $\partial_x G(x,h)$, with $D(h)$ defined by

$$D(h) = \partial_x G(x^*,h) - \partial_x G(x^*,h^*) , \tag{2.3.4}$$

there is a constant $K_2 > 0$ such that

$$\|D(h)\| \le K_2|h-h^*| \tag{2.3.5}$$

for all $h \in I_2$.

Now

$$G(x,h) - x^* = G(x^*,h) + \partial_x G(x^*,h)(x-x^*) + u(x,h) - x^*$$

and so

(2.3.6) $$G(x,h) - x^* = D(h)(x-x^*) + \partial_x G(x^*,h^*)(x-x^*) + u(x,h) .$$

For arbitrary $\varepsilon > 0$ there exists a norm on R^n such that $\|\partial_x G(x^*,h^*)\| \le \alpha + \varepsilon$ [53, Theorem 2.2.8] and in this norm, if δ satisfies $K_1\delta < \varepsilon$ it follows from (2.3.6) that, for any $h \in I_2$,

$$\|G(x,h)-x^*\| \le \left(K_2|h-h^*|+\alpha+2\varepsilon\right)\|x-x^*\|$$

for all $x \in S(x^*,\delta)$. Also since $\{h_i\}$ converges to h^*, there is an i_0 such that $K_2|h_i-h^*| \le \varepsilon$ for all $i \ge i_0$. If ε is chosen so that $\varepsilon/K_2 < \delta_2$, then $h_i \in I_2$ for all $i \ge i_0$, and if $x_i \in S(x^*,\delta)$, it follows that

$$\|x_{i+1}-x^*\| \le (\alpha+3\varepsilon)\|x_i-x^*\| .$$

Since $\alpha < 1$ and ε may be chosen so that $\alpha + 3\varepsilon < 1$, it follows from [53, Theorem 10.1.2] that $R_1(I,x^*) \le \alpha$. If $\alpha = 0$ this completes the proof.

From (2.3.6) we also have, for all $(x,h) \in S(x^*,\delta) \times I_2$,

$$\|G(x,h)-G(x^*,h^*)-\partial_x G(x^*,h^*)(x-x^*)\| = \|D(h)(x-x^*)+u(x,h)\|$$

$$\le \left(K_2|h-h^*|+K_1\|x-x^*\|\right)\|x-x^*\| .$$

Now if $K_2\delta_2 < \varepsilon/2$ and $K_1\delta < \varepsilon/2$, we have

(2.3.7) $$\|G(x,h)-G(x^*,h^*)-\partial_x G(x^*,h^*)(x-x^*)\| \le \varepsilon\|x-x^*\|$$

for all $x \in S(x^*,\delta)$ and for all $h \in I_2$. The remainder of the proof is almost identical to the proof of the Linear Convergence Theorem given in [53] with (2.3.7) replacing equation (10.1.7) in [53] and $\partial_x G(x^*,h^*)$ replacing $G'(x^*)$. □

To complete the theoretical background we consider the possibility of

faster convergence in the case when $\eta(\partial_x G(x^*,h^*)) = 0$. For this case we require further knowledge of the sequence $\{h_i\}$.

THEOREM 2.3.3. *Suppose* $G : D \times D_h \subset R^n \times R \to R^n$ *satisfies the conditions of Theorem* 2.3.2 *and that* $\eta(\partial_x G(x^*,h^*)) = 0$. *Then* (x^*,h^*) *is a point of attraction of the iterative process* I *given by* (2.3.1) *and* $R_1(I,x^*) = 0$. *If, in addition,* $\{h_i\}$ *converges to* h^* *with R-order* $r \geq 1$ *then* $O_R(I,x^*) \geq \min(2^{1/k},r)$, *where* k *is the unique integer such that* $\partial_x G(x^*,h^*)^k = 0$ *and* $\partial_x G(x^*,h^*)^{k-1} \neq 0$.

Proof. Theorem 2.3.2 shows that x^* is a point of attraction of I and that $R_1(I,x^*) = 0$ so we may assume that $\{x_i\}$ converges to x^* .

Let $A = \partial_x G(x^*,h^*)$. Then $\eta(A) = 0$ and there is an integer $k \leq n$ such that $A^{k-1} \neq 0$ and $A^k = 0$. With the definition of $u(x,h)$ and $D(h)$ given in (2.3.2) and (2.3.4), let $D_i = D(h_i)$ and $u_i = u(x_i,h_i)$. Then, if we write $e_i = x_i - x^*$ it follows from (2.3.6) that

$$e_{i+1} = Ae_i + D_i e_i + u_i$$

and, by induction, for $j \geq 0$,

$$e_i = A^j e_{i-j} + A^{j-1} D_{i-j} e_{i-j} + \ldots + AD_{i-2} e_{i-2} + D_{i-1} e_{i-1} + A^{j-1} u_{i-j} + \ldots + Au_{i-2} + u_{i-1} . \tag{2.3.8}$$

Since $\{x_i\}$ and $\{h_i\}$ converge to x^* and h^* respectively, it follows from (2.3.3) and (2.3.5) that, for all sufficiently large i , $\|u_i\| \leq K_1 e_i^2$ and $\|D_i\| \leq K_2 \varepsilon_i$, where $\varepsilon_i = |h_i - h^*|$. Since $A^k = 0$, it now follows from (2.3.8), with $j = k$, that

$$\|e_i\| \le K_1\left(\gamma^{k-1}\|e_{i-k}\|^2 + \ldots + \gamma\|e_{i-2}\|^2 + \|e_{i-1}\|^2\right)$$

$$+ K_2\left(\gamma^{k-1}\|e_{i-k}\|\varepsilon_{i-k} + \ldots + \gamma\|e_{i-2}\|\varepsilon_{i-2} + \|e_{i-1}\|\varepsilon_{i-1}\right)$$

where $\gamma = \|A\|$.

Since $\{x_i\}$ converges to x^* , it follows that there exists an $i_0 > 0$ and constants B_1, B_2 such that, for each $i \ge i_0$,

$$\|e_i\| \le B_1\|e_{i-k}\|^2 + B_2\|e_{i-k}\|\varepsilon_{i-k} .$$

Replacing i by ki and writing $\alpha_i = B_1\|e_{ki}\|$ and $\beta_i = B_2\varepsilon_{ki}$ we have

$$\alpha_i \le \alpha_{i-1}^2 + \alpha_{i-1}\beta_{i-1} , \tag{2.3.9}$$

for all sufficiently large i .

We now require the result that, if $1 < p < \min(2,r^k)$, then there exists a constant $c > 0$ and a $j > 0$ such that

$$\alpha_i \le e^{-cp^i} \tag{2.3.10}$$

for all $i \ge j$. To prove this, suppose that s satisfies $p < s < \min(2,r^k)$. Then, because $\{\beta_i\}$ converges to zero with R-order r^k ,

$$\beta_i \le e^{-s^i} \tag{2.3.11}$$

for all i sufficiently large. Let c be some constant, yet to be determined, then because $p < s$, it follows that, for i sufficiently large,

$$e^{-s^i} \le e^{-cp^i} . \tag{2.3.12}$$

Also, since $\{\alpha_i\}$ converges to zero and $p < 2$,

$$\alpha_i \le e^{-\ln 2/(2-p)} \tag{2.3.13}$$

for all sufficiently large i . Let j be such that (2.3.11), (2.3.12) and

(2.3.13) are all satisfied for all $i \geq j$ and suppose that $\alpha_i \leq e^{-cp^i}$ for some $i \geq j$. Then, from (2.3.9), $\alpha_{i+1} \leq e^{-2cp^i} + e^{-s^i} e^{-cp^i}$ and, from (2.3.12), $\alpha_{i+1} \leq 2e^{-2cp^i}$. We wish to deduce that $\alpha_{i+1} \leq e^{-cp^{i+1}}$ and this will be so if $2e^{-2cp^i} \leq e^{-cp^{i+1}}$. Some simple algebra shows this to be the case if

$$c \geq \frac{\ln 2}{p^i(2-p)} \tag{2.3.14}$$

and a suitable choice for c is

$$c = \frac{\ln 2}{p^j(2-p)}$$

for then (2.3.14) is satisfied for each $i \geq j$. Now, by choice of c, $\alpha_j < e^{-cp^j}$ and we have shown that, assuming (2.3.10) for some $i \geq j$, then (2.3.10) follows with i replaced by $i + 1$. So, by induction, (2.3.10) is true for all $i \geq j$ as we required.

It now follows from (2.3.10) that the R-order of the sequence $\{\alpha_i\}$ is at least p. Since $\alpha_i = \|e_{im}\|$ and p is arbitrarily close to $\min(2, r^k)$, it follows that $O_R(I, x^*) \geq \min(2^{1/k}, r)$. □

2.4. Runge-Kutta Methods

Consider the general class of explicit Runge-Kutta methods for solving the differential equation

$$\dot{x}(t) = q(x), \qquad x(0) = x_0, \tag{2.4.1}$$

given by

$$x_{m+1} = x_m + h_m \sum_{i=1}^{r} \alpha_i k_i(x_m, h_m), \qquad m = 0, 1, \ldots, \tag{2.4.2a}$$

where x_m is an approximation to $x(h_0 + h_1 + \ldots + h_{m-1})$,

$$(2.4.2b) \qquad k_i(x,h) = q\left(x+h \sum_{j=1}^{i-1} \beta_{ij} k_j(x,h)\right) , \qquad i = 1,\ldots,r ,$$

and h_m is the step length. A discussion of stability for this method is usually based upon consideration of the linear differential equation

$$(2.4.3) \qquad \dot{x}(t) = Ax , \qquad x(0) = x_0 ,$$

where A is a fixed matrix whose eigenvalues have negative real part. The true solution of (2.4.3) is

$$x(t+h_m) = \exp(h_m A)x(t)$$

whereas the solution given by (2.4.2) is

$$(2.4.4) \qquad x_{m+1} = p(h_m A)x_m ,$$

where $p(z)$ is a polynomial of degree r whose coefficients depend upon choice of the α's and β's in (2.4.2). The usual practice is to choose these parameters so that $p(z)$ is a good approximation to $\exp(z)$. We note that, since the true solution of (2.4.3) is decreasing, a requirement on the step length h_m is that the condition

$$(2.4.5) \qquad \eta(p(h_m A)) < 1 , \qquad m = 0,1,\ldots ,$$

be satisfied so that the iterates in (2.4.4) also decrease. However, in the nonlinear case, (2.4.5) is of little practical use in controlling the stepsize.

In this section we consider (2.4.2) not only as a means of approximating the solution of (2.1.1) but also as a one-step method for finding a zero of f . For the former the theory is well known [34] and for the latter we use the results of section 2.3. In this case we have

$$x_{m+1} = G(x_m,h_m) , \qquad m = 0,1,\ldots ,$$

where

(2.4.6) $$G(x,h) = x + h \sum_{i=1}^{r} \alpha_i k_i(x,h)$$

and $k_i(x,h)$ is given in (2.4.2b) for $i = 1,\ldots,r$. We apply this process to the case when $q(x)$ is given by

(2.4.7) $$q(x) = -J(x)^{-1} f(x) \ .$$

Then, if I represents the unit matrix,

$$\partial_x G(x,h) = I + h \sum_{i=1}^{r} \alpha_i \partial_x k_i(x,h) \ .$$

If x^* is a zero of $f(x)$ then x^* is a fixed point of (2.4.2) and also, from (2.4.7), we have

$$q'(x^*) = -I \ ,$$

where the prime denotes differentiation with respect to x . It then follows by some simple algebra that

(2.4.8) $$\partial_x G(x^*,h) = p(-h)I$$

where $p(z)$ is the same polynomial as appeared in (2.4.4). Rather than proving this result here, for the sake of continuity we present it in the appendix to this chapter as Theorem 2.4.1. It now follows from Corollary 2.3.1 that a sufficient condition for x^* to be a point of attraction of (2.4.2) is that, for some $\alpha < 1$,

(2.4.9) $$\eta\left(p(-h_m)I\right) = |p(-h_m)| \leq \alpha \ , \qquad m = 1,2,\ldots \ ,$$

which, unlike (2.4.5), provides an *explicit* bound on each h_m for ultimate convergence to x^* . We note that the region of the complex plane defined by

$$|p(z)| < 1$$

is called *the region of absolute stability* of the method (see Gear [29]) and so the condition for convergence to x^* is that, for each m , $-h_m$ lies in this region. It also follows from Theorem 2.3.2 that, if $\lim_{i\to\infty} h_i = h^*$,

the iterative process can give superlinear convergence to x^* only if h^* satisfies

(2.4.10) $$p(-h^*) = 0 .$$

Therefore, when $f(x)$ is three times continuously differentiable it follows from Theorem 2.3.3 that if $\{h_m\}$ converges to h^* with R-order ≥ 2, then the iterative process (2.4.2) has R-order at least 2.

In the application of (2.4.2) it is of benefit to choose the parameters so that the resulting method will follow the solution of (2.1.1) well enough to inhibit divergence but will also provide a fast rate of final convergence to x^*. This means choosing a method which allows h^* to be chosen so that (2.4.10) is satisfied. We note here that for the well-known 4th-order Runge-Kutta process $p(z)$ is defined by

$$p(z) = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!}$$

and $p(-z)$ has no real root. Thus no choice of h^* can furnish superlinear convergence. Also Heun's predictor-corrector method [34] may be written

(2.4.11) $$x_{m+1} = x_m + \frac{h_m}{2}\left[q(x_m)+q(x_m+h_m q(x_m))\right] .$$

This is of the class (2.4.2) and has $p(z)$ defined by

$$p(z) = 1 + z + \frac{z^2}{2} .$$

This is simply a Runge-Kutta method of order 2 and again $p(-z)$ has no real root, so no choice of h^* can give superlinear convergence to x^*.† In attempting to solve (2.1.1), Boggs [9] used this method as an explicit approximation to the trapezoidal rule.

We note that for these two methods we can use Theorem 2.3.2 to show that

---

† Note that "order" is a term related to the accuracy of single and multi-step methods in following the trajectory $x(t)$ (see [34] and the definition of H-order in section 4.2), while the term "R-order" is related to the speed of convergence of a sequence to its limit (see section 3.2 and [53]).

$$O_R(I,x^*) = 1$$

and

$$R_1(I,x^*) = |p(-h^*)| .$$

So assuming (2.4.9) is satisfied, convergence is at best linear and the fastest convergence is achieved by choosing h^* to minimise $|p(-h^*)|$. For Heun's method this is $h^* = 1.0$ when $R_1(I,x^*) = \frac{1}{2}$ and then convergence to x^* is rather slow. If the sequence $\{h_m\}$ does not satisfy (2.4.9), then the method will not generally converge.

Boggs [9] in his paper suggested there is a difficulty of stiffness involved in integrating (2.1.1). Stiffness is a problem which occurs when solving the differential equation

$$\dot{x}(t) = q(x)$$

when $q'(x)$ has eigenvalues with widely separated negative real parts. Their numerical solution requires the generation of special methods which are A-stable [18] or at least stiffly stable (see [29] for a full description of these concepts). One characteristic of an unsuitable method applied to a stiff system of differential equations is for the iterates to oscillate about the true solution and possibly diverge. In our problem, however, $q'(x^*) = -I$ and so, close to x^* at least, (2.1.1) is most certainly not a stiff system. The symptoms of instability which Boggs ascribes to stiffness appear identical to the behaviour observed if the sequence $\{h_m\}$ contravenes (2.4.9). If we attempt to solve the differential equation (2.1.1), the standard methods tend to allow $\{h_m\}$ to increase as the zero is approached, since the rate of change in direction of the solution trajectory is decreasing. If this happens then oscillation and divergence of the sequence $\{x_i\}$ may occur if h_m becomes too large, as would be the case, for example, when using Newton's method with a steplength greater than 2 . When the step is suitably controlled no problems of instability occur

and, indeed, as long as h_m satisfies (2.4.9) for each m, close to the zero the problem is extremely stable, simply because any zero of f is an asymptotically stable node of the autonomous differential equation (2.1.1) [45].

The foregoing theory shows that any method giving a polynomial $p(z)$ such that $p(-h)$ has a positive real root will be effective for producing rapid final convergence if $\{h_m\}$ is suitably chosen. For example, we consider briefly Runge-Kutta methods of orders one, three and five.

The simplest first-order method is Euler's method. In this case $p(z)$ is given by

$$p(z) = 1 + z$$

and, from (2.4.9), we see that x^* is a point of attraction with h-domain $[\delta, 2-\delta]$, for δ arbitrarily small, i.e. local convergence is guaranteed if $0 < \delta \leq h_m \leq 2-\delta$ for each m. Also, from (2.4.10) and Theorem 2.3.3, the R-order of convergence to x^* can be ≥ 2 if $\{h_m\}$ converges to 1 with R-order at least 2. This is essentially Newton's method.

There is a class of third-order Runge-Kutta methods and, for each, $p(z)$ is defined by

$$p(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} .$$

Now $|p(-h)| < 1$ if and only if $0 < h < h_u$, where $h_u = 2.5127 \ldots$, and so, from (2.4.9), each of these third-order methods converges locally to x^* with h-domain $[\delta, h_u-\delta]$, for arbitrarily small δ. Also, the R-order of convergence to x^* can be two if $\{h_m\}$ converges sufficiently fast to $h_r = 1.596 \ldots$, where h_r is the only real root of $p(-z)$.

Finally, there exists a class of six stage fifth-order methods described by Lawson [46]. For one which he recommends, $p(z)$ is defined by

$$p(z) = \sum_{j=0}^{5} \frac{z^j}{j!} + 0.5625\,\frac{z^6}{6!} .$$

In this case $|p(-h)| < 1$ if and only if $0 < h < h_u$, where $h_u = 5.6039 \ldots$, and so x^* is a point of attraction with h-domain $[\delta, h_u - \delta]$, for δ arbitrarily small. Again, convergence to x^* has R-order 2 if $\{h_m\}$ converges sufficiently fast to $h^* = 2.6299 \ldots$, where h^* is a real root of $p(-z)$.

The conclusion of this section is that there exist single-step methods which can follow the solution trajectory of (2.1.1) sufficiently accurately and which, by suitable control of the step length, can furnish rapid convergence to x^*. In section 2.5 numerical details are given for a third-order method which adapts the step length until it reaches a maximum of $h_r = 1.596 \ldots$, after which it is not allowed to increase further.

For completeness, we note here that the principles described in this chapter can be extended to implicit Runge-Kutta methods and to the predictor-corrector methods based on them. As an example we describe Heun's approximation to the trapezium rule with an extra correction, since this method was used by Boggs in [9]. Using standard notation (see [9], [29]), Heun's method, given in (2.4.11), can be considered as a predictor-corrector method of the form

$$\text{P}: \qquad p_m = x_m + h_m q_m ,$$

$$\text{E}: \qquad \hat{q}_m = q(p_m) ,$$

$$\text{C}: \qquad x_{m+1} = x_m + \frac{h_m}{2}\left[q_m + \hat{q}_m\right] ,$$

$$\text{E}: \qquad q_{m+1} = q(x_{m+1}) .$$

With an extra correction, the process becomes

$$\text{P}: \qquad p_m = x_m + h_m q_m ,$$

$$\text{E}: \qquad \hat{q}_m = q(p_m) ,$$

$$C: \quad y_{m+1} = x_m + \frac{h_m}{2}\left[q_m + \hat{q}_m\right] ,$$

$$E: \quad q_{m+1} = q\left(y_{m+1}\right) ,$$

$$C: \quad x_{m+1} = x_m + \frac{h_m}{2}\left[q_m + q_{m+1}\right] ,$$

which can be written in the iterative form

$$(2.4.12a) \quad y_{m+1} = x_m + \frac{h_m}{2}\left[q\left(y_m\right) + q\left(x_m + h_m q\left(y_m\right)\right)\right] ,$$

$$(2.4.12b) \quad x_{m+1} = x_m + \frac{h_m}{2}\left[q\left(y_m\right) + q\left[x_m + \frac{h_m}{2}\left\{q\left(y_m\right) + q\left(x_m + h_m q\left(y_m\right)\right)\right\}\right]\right] .$$

Define z_m , $m = 0,1,2,\ldots$, and z^* by

$$z_m = \begin{bmatrix} y_m \\ x_m \end{bmatrix} , \quad z^* = \begin{bmatrix} x^* \\ x^* \end{bmatrix} .$$

Let I denote the iterative process (2.4.12), then I can be written as

$$z_{m+1} = G\left(z_m, h_m\right) ,$$

which is of the form (2.3.1). In the case that $q(x) = -J(x)^{-1}f(x)$ and $f(x^*) = 0$, some simple algebra shows that $\partial_z G(z^*,h)$ has two eigenvalues λ_1,λ_2 which satisfy

$$\lambda_1 + \lambda_2 = \frac{3}{4}h^2 - h + 1 = \theta(h)$$

say. Since $\theta(h)$ has no real roots and the minimum value $\theta(h)$ is $2/3$, it follows that

$$\eta\left(\partial_z G(z^*,h)\right) \geq 1/3$$

for all h . Theorem 2.3.2 shows that, like Heun's method, convergence of this process to x^* is at best linear and $R_1(I,x^*) \geq 1/3$.

2.5. Numerical Results

We begin by making some general comments on the effectiveness of solving (2.1.1) as a means of finding a zero of f . Although it has been

necessary to assume that x_0 is in a stability region of a zero x^*, for if this is not so then convergence is not guaranteed, there are applications where the approach will be effective. For example, where the usual methods diverge or continually converge to a zero which is known but where the user requires to find a different zero, which he knows to exist, and has a suitable starting point. However, one should realize that, whilst the number of evaluations required to follow the trajectory sufficiently accurately may seem reasonable to one used to solving ordinary differential equations, it may seem surprisingly large to one used to solving nonlinear equations.

Following the trajectory $x(t)$ is usually a simple matter if h can be chosen sufficiently small, but in practice an important part of solving (2.1.1) is in the step length control. Far from a zero of f all of the usual problems of step control occur and great care is required to maintain accuracy. Close to a zero of f this is not the case so long as h is controlled in a way which will guarantee convergence, i.e. so long as h_m satisfies (2.4.9) for each m. As x^* is approached we are less interested in accuracy in following the trajectory than in convergence to x^* and indeed, if we are to achieve fast ultimate convergence to x^*, we must relax our preoccupation with accurate representation of $x(t)$ which converges to x^* only linearly (see (2.2.1)). In the examples that follow we are interested only in demonstrating ways of achieving faster final convergence and so we look only at cases when x_0 is fairly close to x^*. In this case the criterion for varying h can be simpler than would be necessary in the general case.

The basic technique is based upon the fact that the solution of (2.1.1) satisfies

$$f\big(x(t)\big) = e^{-t} f(x_0) \ .$$

Let $f_i = f(x_i)$ and Z_i be given by

$$Z_i = I - \frac{f_i f_i^T}{f_i^T f_i} .$$

Then any point x, on $x(t)$, satisfies

$$Z_0 f(x) = 0 .$$

Suppose x_i is our current approximation to x^*, then the solution of

$$\dot{x}(t) = -J(x)^{-1} f(x) , \qquad x(0) = x_i ,$$

converges to x^* (under the conditions of Theorem 2.2.1) and $\|Z_i f_{i+1}\|$ gives a measure of the deviation of x_{i+1} from this trajectory. On this basis a suitable step change criterion was found to be $h_{i+1} = \min(h^*, \alpha h_i)$ where α is given by

(2.5.1)
$$\alpha = \begin{cases} 2 & \text{if } 0 \le \delta \le \varepsilon_1 , \\ 1 & \text{if } \varepsilon_1 < \delta \le \varepsilon_2 , \\ 0.5 & \text{if } \varepsilon_2 < \delta \le \varepsilon_3 , \end{cases}$$

$\delta = \|Z_i f_{i+1}\|$ and h^* is the step size necessary for the fastest convergence for the method. In addition, the point x_{i+1} was rejected and the step repeated with half the step length if either $\delta > \varepsilon_3$ or $\mathrm{Det}(J(x_{i+1})) \neq \mathrm{Det}(J(x_0))$, in which case the iterates had crossed a region of singularity of the Jacobian.

Various methods were tested on a variety of problems and the results of some of these tests are tabulated below. As an example of a method with rapid final convergence we chose a third-order Runge-Kutta method (RK3) for which $h^* = h_r = 1.596\ldots$. For comparison we tried Heun's method (HEUN) which was used by Boggs and is given in (2.4.11), for which $h^* = 1.0$.

Since we are advocating the use of (1.7) as opposed to (1.6), we also looked at a third-order Runge-Kutta method (K3) for solving equation (1.6) to find an estimate of the solution at $t = 1$. In this method a major

iteration consists of integrating

(2.5.2) $$\dot{x}(t) = -J(x)^{-1}f(x_i) , \qquad x(0) = x_i ,$$

giving a sequence $\{y_{i,j}\}$, $j = 1,\ldots,N_i$, such that $y_{i,j}$ is an approximation to $x(t_{i,j})$, where $t_{i,j} = \sum_{k=1}^{j-1} h_{i,k}$ and $t_{i,N_i} = 1$. Then $x_{i+1} = y_{i,N_i} = y_{i+1,1}$. It is proved by Kleinmichel [41] and Bittner [7] that, under general conditions, if the method uses step size $h^* = 1$ then the sequence $\{x_i\}$ converges to x^* with R-order 4 . Despite this high rate of convergence, the greater demand on accuracy required in following the solution trajectory of (2.5.2) when x_i is not close to x^* causes the algorithm to be less effective than those described in this chapter.

For a fair comparison of methods we used a similar step control to that described above. Since the solution of

$$\dot{x}(t) = -J(x)^{-1}f(y_{ij}) , \qquad x(0) = y_{ij} ,$$

does not generally converge to x^* and may, in practice, cross a region of singularity of $J(x)$, it is necessary that each y_{ij} be close to the solution trajectory of (2.5.2). In this case, therefore, the most suitable criterion is that $h_{i,j+1} = \min(\alpha h_{i,j}, 1-t_{i,j+1})$ where α is given by (2.5.1) and $\delta = \|Z_i f(y_{i,j+1})\|$. Also we took $h_{i+1,1} = \min(1, 2\max(h_{i,N_i}, h_{i,N_{i-1}}))$. The conditions for rejecting a step were the same as before.

In each algorithm $\varepsilon_3 = 0.5$, $\varepsilon_2 = 0.25$ and $\varepsilon_1 = 0.05$ were found to be suitable and the initial step, in each case, was taken as $h^*/8$. Each algorithm was applied to a variety of functions and the following eight problems gave results which were typical. In each case the solution given is the limit of the trajectory defined by (2.1.1) with the given value of

x_0 .

1. A function found in Boggs [9];

$$f_1 = x_1^2 - x_2 + 1 ,$$

$$f_2 = x_1 - \cos\left(\frac{\pi}{2} x_2\right) ,$$

with initial guess $x_0 = (1,0)$. The correct solution is $x^* = (0,1)$.

2. Problem 1 with initial guess $(-1,-1)$. The correct solution is $(0,1)$ and the solution trajectory passes close to a region where $J(x)$ is singular.

3. A function found in Broyden [15];

$$f_1 = \tfrac{1}{2} \sin(x_1 x_2) - x_2/(4\pi) - x_1/2 ,$$

$$f_2 = \left(1-1/(4\pi)\right)\left(e^{2x_1}-e\right) + ex_2/\pi - 2ex_1 ,$$

with initial guess $(.6,3.)$. The correct solution is $(\tfrac{1}{2},\pi)$.

4. The gradient of Rosenbrock's function;

$$f_1 = 400\, x_1\left(x_1^2-x_2\right) + 2\left(x_1-1\right) ,$$

$$f_2 = -200\left(x_1^2-x_2\right) ,$$

with initial guess $(-1.2,1.0)$. The correct solution is $(1,1)$ and this problem can be considered fairly difficult since the solution trajectory is always close to the region where $J(x)$ is singular (see [11]).

5. A function found in Branin [11];

$$f_1 = 2 \sin(2\pi x_1/5) \sin(2\pi x_3/5) - x_2 ,$$

$$f_2 = 2.5 - x_3 + 0.1\, x_2 \sin(2\pi x_3) - x_1 ,$$

$$f_3 = 1 + 0.1\, x_2 \sin(2\pi x_1) - x_3 ,$$

with initial guess $(0,0,0)$. The correct solution is $(1.5,1.809\ldots,1.0)$.

6. A function found in Deist and Sefor [22];

$$f_i = \sum_{\substack{j=1 \\ j\neq i}}^{6} \cot \beta_i x_j , \qquad i = 1,\ldots,6 ,$$

where $100\,\beta_i = 2.249, 2.166, 2.083, 2.0, 1.918, 1.835$, for $i = 1,\ldots,6$ respectively. With initial guess $x_i = 75.0$, $i = 1,\ldots,6$ the correct solution is approximately (121.9, 114.2, 93.6, 62.3, 41.3, 30.5) .

7. A discretisation of

$$3\ddot{y}y + \dot{y}^2 = 0$$

with boundary conditions $y(0) = 0$, $y(1) = 20$, gives rise to the equations

$$f_1 = 3x_1(x_2 - 2x_1) + x_2^2/4 ,$$

$$f_i = 3x_i(x_{i+1} - 2x_i + x_{i-1}) + (x_{i+1} - x_{i-1})^2/4 , \qquad i = 2,\ldots,n-1 ,$$

$$f_n = 3x_n(20 - 2x_n + x_{n-1}) + (20 - x_{n-1})^2/4 .$$

The true solution of the boundary value problem is $y = 20t^{3/4}$. As initial guess we chose $x_i = 10$, $i = 1,\ldots,n$ and set $n = 10$.

8. Same as problem 7 with $n = 20$.

Both of these problems have solution trajectories which pass close to a region of singularity.

Table 2.1 gives results on the effort required by the methods to reduce each component of f to less than 10^{-6} . For each method the first line gives the number of Jacobian evaluations, the second gives the number of function evaluations and the third the number of equivalent function evaluations counting a Jacobian evaluation as n function evaluations, except for problems 7 and 8 where the Jacobian is tridiagonal and its evaluation is counted as being equivalent to 3 function evaluations. Note that, because of the way steps were either accepted or rejected, the number of Jacobian and function evaluations are not necessarily the same.

TABLE 2.1

ALGORITHM	PROBLEM							
	1	2	3	4	5	6	7	8
	21	29	18	110	28	24	69	69
RK3	22	31	19	114	29	25	73	73
	64	89	55	334	113	169	280	280
	18	39	15		26	29		
K3	7	13	6	*	10	11	*	**
	43	91	36		88	185		
	44	52	38	109	46	44	86	88
HEUN	45	53	39	119	47	45	89	91
	133	157	115	337	185	309	347	355

* - h reduced to minimum allowed, viz. $2^{-13}h^*$.

** - terminated after 200 function evaluations.

We can draw a number of conclusions from the numerical results. The first is that the HEUN algorithm, which has only linear convergence to x^* , requires significantly more evaluations than the other methods. This is as we would expect. Because of the high rate of ultimate convergence, the K3 algorithm is generally superior when the problem is simple, i.e. when the solution trajectory is smooth and does not approach close to regions where the Jacobian is singular. However, where this is not the case RK3 appears more efficient and in particular we note that it is more reliable in that it always succeeded in finding the desired solution in a reasonable time. The need for the K3 method to always follow the same trajectory led to the greater number of function evaluations in these cases.

We note here that any comparison of routines is necessarily a comparison also of the step change criteria and that the criteria chosen were not necessarily the best for each routine. However we have deliberately adopted simple criteria for changing stepsize in the hope of demonstrating that the methods which use (1.7) are more robust than those which use (1.6).

APPENDIX TO CHAPTER 2

We now prove the result quoted in section 2.4. We assume the notation of that section and that $f(x)$ is sufficiently differentiable.

THEOREM 2.4.1. *The iterate* x_{m+1}, *given by* (2.4.2) *applied to the function* $q(x) = Ax$, *where* A *is a fixed matrix, satisfies*

$$x_{m+1} = p(h_m A)x_m ,$$

where $p(z)$ *is a polynomial of degree* r.

In addition, if $G(x,h)$ *is given by* (2.4.6) *and* (2.4.2b), *with the choice* $q(x) = -J(x)^{-1}f(x)$, $\partial_x G(x^*,h)$ *is given by*

$$\partial_x G(x^*,h) = p(-h)I .$$

Proof. Define the polynomials $p_i(z)$, $i = 1,\ldots,r$ by

(A2.1) $$p_1(z) = 1 ,$$

(A2.2) $$p_i(z) = 1 + z \sum_{j=1}^{i-1} \beta_{ij} p_j(z) , \qquad i = 2,3,\ldots,r ,$$

where the β_{ij} are as in (2.4.2b). Also define $p(z)$ by

(A2.3) $$p(z) = 1 + z \sum_{i=1}^{r} \alpha_i p_i(z) ,$$

where the α_i are as in (2.4.2a). We now show that, with $q(x) = Ax$, the $k_i(x,h)$ given in (2.4.2b) satisfy

(A2.4) $$k_j(x,h) = Ap_j(hA)x ,$$

$j = 1,2,\ldots,r$. Certainly $k_1(x,h) = Ap_1(hA)x$, since $k_1(x,h) = q(x)$ and $p_1(z) = 1$. Now suppose that (A2.4) is true for $j = 1,\ldots,i-1$. Then, from (2.4.2b) and the definition of $q(x)$, we have

$$k_i(x,h) = A\left[x + h \sum_{j=1}^{i-1} \beta_{ij} A p_j(hA)x\right]$$

$$= A\left[I + hA \sum_{j=1}^{i-1} \beta_{ij} p_j(hA)\right]x ,$$

and from (A2.2), this gives

$$k_i(x,h) = Ap_i(hA)x .$$

By induction, (A2.4) is true for $j = 1,\ldots,r$. Now, from (2.4.2a),

$$x_{m+1} = x_m + h_m \sum_{i=1}^{r} \alpha_i A p_i(h_m A)x_m$$

$$= \left[I + h_m A \sum_{i=1}^{r} \alpha_i p_i(h_m A)\right]x_m$$

$$= p(h_m A)x_m ,$$

from (A2.3). Also it is trivial to show that $p(z)$ is a polynomial of degree r . This completes the first part of the proof.

Now we consider $\partial_x G(x,h)$ with $q(x) = -J(x)^{-1}f(x)$. From (2.4.2a) we have

(A2.5) $$\partial_x G(x,h) = I + h \sum_{i=1}^{r} \alpha_i \partial_x k_i(x,h) .$$

Also, from (2.4.2b),

$$\partial_x k_i(x,h) = q'\left[x + h \sum_{j=1}^{i-1} \beta_{ij} k_j(x,h)\right]\left[I + h \sum_{j=1}^{i-1} \beta_{ij} \partial_x k_j(x,h)\right] .$$

Now $k_1(x^*,h) = q(x^*) = 0$ and suppose that $k_l(x^*,h) = 0$, $l = 1,\ldots,j-1$. Then from (2.4.2b), $k_j(x^*,h) = q(x^*) = 0$ and so, by induction, $k_j(x^*,h) = 0$, $j = 1,2,\ldots,r$. Thus

(A2.6) $$\partial_x k_i(x^*,h) = q'(x^*)\left[I + h \sum_{j=1}^{i-1} \beta_{ij} \partial_x k_j(x,h)\right] .$$

We now show that

(A2.7) $$\partial_x k_i(x^*,h) = -p_i(-h)I , \qquad i = 1,\ldots,r .$$

First, $\partial_x k_1(x^*,h) = q'(x^*)$. Also $q'(x^*) = -I$ and from (A2.1), $\partial_x k_1(x^*,h) = -p_1(-h)I$. Suppose that $\partial_x k_j(x^*,h) = -p_j(-h)I$, $j = 1,\ldots,i-1$. Then, from (A2.6),

$$\partial_x k_i(x^*,h) = -\left(1 - h \sum_{j=1}^{i-1} \beta_{ij} p_j(-h)\right) I$$

and, from (A2.2),

$$\partial_x k_i(x^*,h) = -p_i(-h)I \ .$$

So, by induction, (A2.7) follows. Finally, from (A2.5),

$$\partial_x G(x^*,h) = \left(1 - h \sum_{i=1}^{r} \alpha_i p_i(-h)\right) I$$

and from (A2.3) we have

$$\partial_x G(x^*,h) = p(-h)\, I$$

as required. □

CHAPTER 3

CONTINUATION WITH MULTISTEP METHODS

3.1. Introduction

In Chapter 2 we considered the application of standard single-step methods to solving (2.1.1) and in this chapter we develop the corresponding theory for multistep methods. The local convergence theory for multistep methods, which is essentially a generalisation of the single-step theory of Ostrowski [58], has been considered in detail by Voigt [71]. In section 3.2 we quote Voigt's main result and apply it to multistep methods which are also suitable for solving (2.1.1). In this way we develop multistep methods which can follow the solution of (2.1.1) accurately and also converge rapidly to x^* . In section 3.3 we restrict attention to explicit multistep methods and prove a result on the order of accuracy attainable by these rapidly convergent methods. Also we derive a lower bound on the R-order of convergence of the methods when considered as iterative schemes for finding x^* . An important feature of any method for solving (2.1.1) is that the step size be adaptive. In section 3.4 we consider the possibility of variable step methods, based upon the fixed step methods derived, and indicate that they are unstable. However we suggest variable formula and variable step methods based upon a combination of the Adams-Bashforth and the derived methods. That the resulting methods are stable follows from the theory developed by Gear and Tu [30] and Gear and Watanabe [31]. Finally, in section 3.5, we give numerical results on the efficiency of some of the resulting methods and compare them with the methods of Chapter 1.

3.2. General Theory

In this section we consider the solution of the differential equation (2.4.1) by means of a linear multistep method of the form

$$\rho(E)x_m - h\sigma(E)q(x_m) = 0 , \qquad m = 0, 1, \ldots , \tag{3.2.1}$$

where E is the displacement operator defined by

$$E^k(v(x)) = v(x+kh)$$

and $\rho(\lambda)$ and $\sigma(\lambda)$ are polynomials given by

$$\rho(\lambda) = \sum_{j=0}^{r} \alpha_j \lambda^j , \qquad \alpha_r \neq 0 , \tag{3.2.2}$$

and

$$\sigma(\lambda) = \sum_{j=0}^{r} \beta_j \lambda^j . \tag{3.2.3}$$

The process (3.2.1) can be considered as a (possibly implicit) multistep method of the form

$$G(x_{m+r},\ldots,x_m) = 0 , \qquad m = 0,1,\ldots \tag{3.2.4}$$

and we can use the following theorem, due to Voigt [71], to give conditions on the method which will guarantee local convergence to a zero of f when $q(x)$ is given by (2.4.7). In the following, $\partial_i G(x_1,\ldots,x_m)$ denotes the Frechet partial derivative of G with respect to x_i .

THEOREM 3.2.1. *Suppose that* $G : D^{r+1} \subset (R^n)^{r+1} \to R^n$ *is continuously differentiable on an open neighbourhood* $D_0^{r+1} \subset D^{r+1}$. *Assume that there is an* $x^* \in D_0$ *such that* $G(x^*,\ldots,x^*) = 0$, $\partial_1 G(x^*,\ldots,x^*)$ *is nonsingular and* $\eta = \eta(W) < 1$, *where* W *is given by*

$$W = \begin{bmatrix} W_2 & W_3 & \ldots & W_{r+1} \\ I & 0 & \ldots & 0 \\ 0 & I & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ 0 & \cdot & \ldots & I \quad 0 \end{bmatrix} \tag{3.2.5}$$

and

(3.2.6) $W_i = -\partial_1 G(x^*,\ldots,x^*)^{-1}\partial_i G(x^*,\ldots,x^*)$, $i = 2,\ldots,r+1$.

Then there is an open neighbourhood S *of* x^* *such that the sequence* $\{x_k\}$ *defined by the iterative process* I *given by* (3.2.4) *is well defined for any* $(x_0,x_1,\ldots,x_{r-1}) \in S^r$ *and converges to* x^* *with*

$$R_1(I,x^*) = \eta .$$

Proof. See Voigt [71]. □

In our application, from (3.2.1) - (3.2.3), we have

(3.2.7) $$G(y_1,\ldots,y_{r+1}) = \sum_{j=0}^{r} \alpha_j y_{r-j+1} - h \sum_{j=0}^{r} \beta_j q_{r-j+1} ,$$

where $q_k = q(y_k)$.

The first condition that Theorem 3.2.1 imposes is that

(3.2.8) $$G(x^*,\ldots,x^*) = 0$$

which, since $q(x^*) = 0$, gives

(3.2.9) $$\sum_{j=0}^{r} \alpha_j = 0$$

and this, in the usual notation, can be expressed as

(3.2.10) $$\rho(1) = 0 .$$

Also

$$\partial_i G(y_1,\ldots,y_{r+1}) = \alpha_{r-i+1} I - h\beta_{r-i+1} q'(y_i) , \qquad i = 1,\ldots,r+1$$

and since $q'(x^*) = -I$, it follows that

$$\partial_i G(x^*,\ldots,x^*) = (\alpha_{r-i+1} + h\beta_{r-i+1}) I , \qquad i = 1,\ldots,r+1 .$$

For application of Theorem 3.2.1 we require that $\partial_1 G(x^*,\ldots,x^*)$ be non-singular, i.e. that

(3.2.11) $$\alpha_r + h\beta_r \neq 0$$

and subsequently we assume this to be the case. In section 3.3 we assume (3.2.1) to be an explicit method, in which case $\alpha_r \neq 0$ and $\beta_r = 0$, so

(3.2.11) is automatically satisfied.

Define

$$\xi_i = \frac{\alpha_{r-i+1}+h\beta_{r-i+1}}{\alpha_r+h\beta_r} , \qquad i = 2,\ldots,r+1 , \tag{3.2.12}$$

so

$$W_i = -\xi_i I , \qquad i = 2,\ldots,r+1 . \tag{3.2.13}$$

To guarantee that the sequence $\{x_k\}$ generated by (3.2.4) converges to x^*, we look at $\eta(W)$ with W given by (3.2.5) and (3.2.6). Now an eigenvalue λ of W satisfies

$$\begin{bmatrix} W_2 & W_3 & \cdots & W_{r+1} \\ I & 0 & \cdots & 0 \\ 0 & I & \cdots & 0 \\ \cdot & & & \\ \cdot & & & \\ 0 & \cdot & \cdots & I \quad 0 \end{bmatrix} \begin{bmatrix} v_2 \\ v_3 \\ \cdot \\ \cdot \\ \cdot \\ v_{r+1} \end{bmatrix} = \lambda \begin{bmatrix} v_2 \\ v_3 \\ \cdot \\ \cdot \\ \cdot \\ v_{r+1} \end{bmatrix}$$

where $\left(v_2^T, v_3^T, \ldots, v_{r+1}^T\right)^T$ is an eigenvector. From this, it follows that

$$\sum_{j=2}^{r+1} W_j v_j = \lambda v_2$$

and

$$v_j = \lambda v_{j+1} , \qquad j = 2,\ldots,r . \tag{3.2.14}$$

From (3.2.14) we have $v_j = \lambda^{r+1-j} v_{r+1}$, $j = 2,\ldots,r$ and so

$$\left(\sum_{j=2}^{r+1} W_j \lambda^{r+1-j}\right) v_{r+1} = \lambda^r v_{r+1} .$$

Using (3.2.13), we now have

$$\lambda^r + \sum_{j=2}^{r+1} \xi_j \lambda^{r+1-j} = 0 .$$

Replacing each ξ_j using (3.2.12) gives immediately that λ is an eigenvalue of W if and only if λ satisfies

(3.2.15) $$\rho(\lambda) + h\sigma(\lambda) = 0 \ .$$

Thus, from Theorem 3.2.1, a sufficient condition for local convergence to x^* is that each root of (3.2.15) is less than 1 in magnitude. As in (2.4.9), this gives an *explicit* bound on h to ensure ultimate convergence. So this condition corresponds to (2.4.9) in the single-step case. The corresponding region of absolute stability of a multistep method is that part of the Complex plane for which the roots of

$$\rho(\lambda) - z\sigma(\lambda) = 0$$

are less than one in magnitude. Therefore, once again, the condition for local convergence to x^* is that $-h$ lies in the region of absolute stability.

We now consider the possibility of superlinear convergence of the sequence $\{x_k\}$ to x^* . Theorem 3.2.1 shows that this is possible only if

$$\eta(W) = 0 \ ,$$

i.e. if all the roots of (3.2.15) are zero. This is equivalent to the condition

$$\rho(\lambda) + h\sigma(\lambda) = \gamma\lambda^r$$

for some $\gamma \neq 0$. From (3.2.2) and (3.2.3) this is equivalent to

$$\alpha_r + h\beta_r = \gamma \ ,$$

and

$$\alpha_j + h\beta_j = 0 \ , \qquad j = 0,\ldots,r-1 \ .$$

We have therefore proved the following theorem.

THEOREM 3.2.2. *For superlinear convergence of a linear multistep method applied to* (2.1.1), *the general iterative process*

$$\sum_{j=0}^{r} \alpha_j x_{m+j} + h \sum_{j=0}^{r} \beta_j J(x_{m+j})^{-1} f(x_{m+j}) = 0$$

must be of the form

$$\sum_{j=0}^{r} \alpha_j x_{m+j} - \sum_{j=0}^{r-1} \alpha_j J(x_{m+j})^{-1} f(x_{m+j}) + h\beta_r J(x_{m+r})^{-1} f(x_{m+r}) = 0 \ ,$$

where

$$\sum_{j=0}^{r} \alpha_j = 0$$

and

$$\alpha_r + h\beta_r \neq 0 \ .$$

In the explicit case, when $\beta_r = 0$, this can be considered as a weighted Newton method where, at each step, x_{r+m} is taken to be a weighted sum of Newton steps, i.e.

$$x_{r+m} = \sum_{j=0}^{r-1} \hat{\alpha}_j \left(x_{j+m} - J(x_{j+m})^{-1} f(x_{j+m}) \right) \ ,$$

where $\hat{\alpha}_j = -\alpha_j/\alpha_r$ and $\sum_{j=0}^{r-1} \hat{\alpha}_j = 1$.

3.3. Explicit Methods

Since an implicit method requires, at each iteration, the solution of a system of nonlinear equations and since finding such a solution is our original problem, we regard implicit methods as inappropriate and do not consider them further. In this section we consider explicit multistep methods for solving (2.1.1) which have satisfactory stability and order properties. The results of the previous section show that, given h_0, any method for which $\rho(\lambda)$ satisfies (3.2.10) and

(3.3.1) $$\rho(\lambda) = \lambda^r - h_0\sigma(\lambda) \ ,$$

(where $\sigma(\lambda)$ is a polynomial of degree $r - 1$), is explicit and gives local superlinear convergence to x^* when $h = h_0$. Consider now the order, in the sense of Henrici [34], attainable by this method.

THEOREM 3.3.1. *Given any* h_0 *in* (3.3.1), *there exists a unique polynomial* $\sigma(\lambda)$ *of degree* $r - 1$ *such that the resulting method has order*

$r - 1$. *For any* r *there exists at most* r *values of* h_0 *such that the method has order* r .

Proof. The proof is an application of Lemma 5.3 of Henrici [34] which states that a method has exact order p if and only if the function

$$\phi(\zeta) = \frac{\rho(\zeta)}{\log\zeta} - \sigma(\zeta)$$

has a zero of exact order p at $\zeta = 1$. In this case, from (3.3.1), $\phi(\zeta)$ is given by

$$\phi(\zeta) = \frac{\rho(\zeta)}{\log\zeta} - \frac{\zeta^r - \rho(\zeta)}{h_0} .$$

Thus, a method defined by (3.3.1) has order p if and only if there exists a function $\psi_1(\zeta)$ such that $\psi_1(1) \neq 0$ and

$$\frac{\rho(\zeta)}{\log\zeta} - \frac{\zeta^r - \rho(\zeta)}{h_0} = (\zeta-1)^p \psi_1(\zeta) .$$

Letting $1 + \gamma = \zeta$ this is equivalent to the existence of a function $\psi_2(\gamma)$ such that $\psi_2(0) \neq 0$ and

$$\rho(1+\gamma) = \left[\frac{\log(1+\gamma)}{h_0+\log(1+\gamma)}\right]\left[(1+\gamma)^r + \gamma^p \psi_2(\gamma)\right]$$

i.e.

$$\rho(1+\gamma) = \frac{(1+\gamma)^r \log(1+\gamma)}{h_0+\log(1+\gamma)} + \frac{\gamma^p \log(1+\gamma)}{h_0+\log(1+\gamma)} \psi_2(\gamma) .$$

Expanding both terms on the right hand side in powers of γ , the condition that the method has order p is that there exist constants $\pi_1, \pi_2, \ldots$, such that $\pi_1 \neq 0$ and

$$(3.3.2) \quad \rho(1+\gamma) = \frac{a_1(h_0)}{h_0}\gamma + \frac{a_2(h_0)}{h_0^2}\gamma^2 + \ldots + \frac{a_r(h_0)}{h_0^r}\gamma^r + \ldots + \gamma^{p+1}\left(\pi_1 + \pi_2\gamma + \pi_3\gamma^2 + \ldots\right)$$

where, for each j , $a_j(h_0)$ is a polynomial of degree $j - 1$ in h_0 .

For $p = r - 1$ the coefficients π_j, $j = 1,2,\ldots$, can be chosen so that $\pi_1 + a_r(h_0)/h_0^r = 1$ and

$$\frac{a_{j+r-1}(h_0)}{h_0^{j+r-1}} + \pi_j = 0 , \qquad j \geq 2 ,$$

in which case the right hand side of (3.3.2) represents a polynomial of degree r with coefficient of γ^r equal to 1 as required. The derived method is obviously unique and has order $r - 1$.

If $p = r$, h_0 is such that

$$a_r(h_0)/h_0^r = 1 , \tag{3.3.3}$$

and π_j, $j \geq 1$, are chosen to satisfy

$$\frac{a_{j+r}(h_0)}{h_0^{j+r}} + \pi_j = 0 ,$$

then the method has order r. (3.3.3) can only be the case when h_0 is a root of the polynomial $a_r(h_0) - h_0^r$, which is of degree r. Thus there at most r values of h_0 for which a method satisfying (3.3.1) can be of order r. This completes the proof. □

Next we use Theorem 2.3.3 to give a lower bound on the local R-convergence rate of methods satisfying (3.2.10) and (3.3.1).

THEOREM 3.3.2. *Suppose that* $q(x) = -J(x)^{-1}f(x)$ *is continuous and there exists a* $\delta > 0$ *such that* $q''(x)$ *exists and is bounded in* $S(x^*,\delta)$. *Then any iterative process* I *defined by* (3.2.1) - (3.2.3), *for which* $\rho(\lambda)$ *satisfies* (3.2.10) *and* (3.3.1), *when applied to* (2.1.1) *converges locally to* x^* *and*

$$O_R(I,x^*) \geq 2^{1/r} .$$

Proof. Rewrite (3.2.1) - (3.2.3) in the explicit form

$$x_{m+r} = G(x_{m+r-1},\ldots,x_m)$$

and set $z_k = (x_k,\ldots,x_{k-r+1})$, for $k = m+r-1,m+r,\ldots$, and $z^* = (x^*,\ldots,x^*)$. Define $\hat{G} : D^r \subset (R^n)^r \to (R^n)^r$ by

$$\hat{G}(y_1,\ldots,y_m) = (G(y_1,\ldots,y_m),y_1,\ldots,y_{m-1}) .$$

Then $z_{k+1} = \hat{G}(z_k)$. Since G is differentiable at x^* , $\hat{G}$ is differentiable at z^* and $\hat{G}'(z^*) = W$, where W is given by (3.2.5). However, it follows from (3.3.1) that in (3.2.13), $W_i = 0$, $i = 2,\ldots,r+1$, and so $\eta(\hat{G}'(z^*)) = 0$. Also, from the form of (3.2.5), $\hat{G}'(z^*)^r = 0$ and $\hat{G}'(z^*)^{r-1} \neq 0$.

$\hat{G}(z)$ therefore satisfies the conditions of Theorem 2.3.3, z^* is a point of attraction of the iteration $I_2 : z_{k+1} = \hat{G}(z_k)$, and $O_R(I_2,z^*) \geq 2^{1/r}$.

Now there exists a norm such that $\|x_i-x^*\| \leq \|z_i-z^*\|$ for each i (see [71]) and so $O_R(I,x^*) \geq O_R(I_2,z^*) \geq 2^{1/r}$. This completes the proof. □

We can now look at methods suggested by Theorem 3.3.1 for various values of r . The relevant polynomials are

$$\text{(3.3.4a)}\quad \rho(\lambda) = \lambda^2 - \frac{2h_0-1}{h_0}\lambda - \frac{(1-h_0)}{h_0} , \text{ for } r = 2 ,$$

$$\text{(3.3.4b)}\quad \rho(\lambda) = \lambda^3 - \frac{(6h_0^2-5h_0+2)}{2h_0^2}\lambda^2 + \frac{(3h_0^2-4h_0+2)}{h_0^2}\lambda - \frac{(2h_0^2-3h_0+2)}{2h_0^2} ,$$

for $r = 3$,

and

$$\text{(3.3.4c)}\quad \rho(\lambda) = \lambda^4 - \frac{(12h_0^3-13h_0^2+9h_0-3)}{3h_0^3}\lambda^3 + \frac{(12h_0^3-19h_0^2+16h_0-6)}{2h_0^3}\lambda^2$$

$$- \frac{(4h_0^3-7h_0^2+7h_0-3)}{h_0^3}\lambda + \frac{(6h_0^3-11h_0^2+12h_0-6)}{6h_0^3} , \text{ for } r = 4 ,$$

and similar formulae, of increasing complexity, can be derived for larger values of r. The two-step method in (3.3.4a) is order 1, but if $h_0 = 1$ the method deflates to a one-step method, also of order 1. This is, of course, Newton's method, and is the one-step method of order 1 suggested by Theorem 3.3.1.

Similarly if h_0 in (3.3.4b) is chosen so that the constant term is zero then the resulting method would be two-step and of order 2. That the polynomial $2h_0^2 - 3h_0 + 2$ has no real root shows that there is no such method. However there exists one value of h_0 for which a three-step method of order 3 exists. This is the method obtained by setting the constant coefficient of $\rho(\lambda)$ in (3.3.4c) equal to zero. The equation

$$6h_0^3 - 11h_0^2 + 12h_0 - 6 = 0 \tag{3.3.5}$$

has only one real solution, which is approximately 0.8599, and on setting h_0 to this value (3.3.4c) deflates to a three-step method.

Theorem 3.3.2 gives information on the R-order of convergence of iterative processes specified by (3.3.4). For (3.3.4a) the R-order is $\geq 2^{\frac{1}{2}}$ unless $h_0 = 1$, in which case the method deflates to a one-step method and $\rho(\lambda)$ can be written

$$\rho(\lambda) = \lambda - 1 .$$

In this case Theorem 3.3.2 states that the resulting method has R-order ≥ 2, which is as expected since this is simply Newton's method. The theorem also shows that the method using (3.3.4c) has R-order $\geq 2^{\frac{1}{4}}$, unless h_0 is chosen as the real root of (3.3.5), in which case, since the method becomes three-step, the R-order is $\geq 2^{1/3}$. This is therefore the most efficient method of order 3, a fact which is borne out in practice (see section 3.5). We also conclude that, for multistep methods, increasing the order increases the accuracy in following $x(t)$ but decreases the

efficiency of final convergence to x^* .

Two further requirements on any practical method, for small h at least, are those of consistency and stability (see Henrici [34]). Consistency is equivalent to having order at least 1 , which is the case for the methods under discussion, and stability demands that no root of $\rho(\lambda)$ exceeds 1 in modulus and that the roots of modulus 1 be simple. In this case the stability condition depends upon h_0 and for $r = 2,3,4$ the methods are stable if

$$(3.3.6) \qquad \begin{cases} h_0 \geq 1/2 & \text{for } r = 2 , \\ h_0 \geq 2/3 & \text{for } r = 3 , \\ 2/3 \leq h_0 \leq 2.5147 \ldots & \text{for } r = 4 . \end{cases}$$

So, for each r considered, if h_0 is chosen to satisfy (3.3.6) the methods will be stable for small h . That this condition need not be strictly fulfilled is shown in the next section for the methods will not be used with small h but only with $h = h_0$.

3.4. Practical Numerical Methods

The methods discussed in the previous section were derived with the idea of initially using a small step size which, as the zero x^* is approached, could be increased and finally fixed at h_0 to give superlinear convergence to x^* . However the foregoing theory assumes h to be fixed throughout and so is not directly applicable to variable step size. We may generate methods based upon those described in section 3.3 with varying step size, in the style of Gear [29]. These can be either of the Nordsieck type [52], where instead of using approximations to $x(ih)$ and $\dot{x}(ih)$, $i = m,m+1,\ldots,m+r-1$, we use approximations to the derivatives $x^{(k)}(mh)$, $k = 0,1,\ldots,2r-1$, or of the variable step type where we start with $r+1$

unequally spaced points t_{m+r-i} , $r \geq i \geq 0$, and compute the coefficients of the explicit multistep formula

$$y_{m+r} = h_{m+r-1}\beta_{r-1,m}y_{m+r-1} + \ldots + h_m\beta_{0,m}y_m + h_{m+r-1}\beta_{r-1,m}q_{m+r-1} + \ldots + h_m\beta_{0,m}q_m$$

so that the order is $r - 1$, where $h_{j+1} = t_{j+1} - t_j$. This is the formula for variable steps (based upon (3.3.1)) which, if $h_j = h_0$ for $j = m,\ldots,m+r$, gives the formulae listed in (3.3.4).

Unfortunately these variable step methods are unstable with respect to changes in step size. When programmed the methods work well for fixed step but display obvious instability when step sizes are increased. This behaviour is explained in detail by the theory developed by Gear and Tu [30] and precludes the use of the methods with varying step. However, it is shown in [30] that the variable step methods based upon the Adams-Bashforth formulae are stable and so the methods of section 3.3 can be combined with these to give the required characteristics. If an Adams-Bashforth variable-step method with r steps is applied to (2.1.1) then, as x^* is approached, the step size can be increased. Because the Adams method cannot give super-linear convergence to x^* we finally hold the step fixed at some value h_0 and when enough steps of fixed size have been taken we can switch to a method which gives fast ultimate convergence. Should a premature change to the fixed step be made then it will be necessary to reduce h and revert again to the variable step Adams method. These composite methods are thus variable formula and possibly variable order and an application of the comprehensive theory of Gear and Watanabe [31], on stability of variable order multistep methods, shows that the derived methods are stable.

Since the Adams method is to be used with stepsizes up to h_0 , it would be preferable if the region of absolute stability of the method contained h_0 . In our numerical tests we chose methods of order 3 and

unfortunately, close to x^* , the Adams-Bashforth predictor of order 3 is absolutely stable only if

(3.4.1) $$0 < h < 6/11 .$$

So the root of (3.3.5) is not contained in this interval. In practice this did not prove to be a difficulty because steps which did not satisfy (3.4.1) were so few that stability was hardly affected.

To see if improvement was possible we also considered the predictor-corrector schemes based upon the Adams-Bashforth and Adams-Moulton formulae. For example, in standard notation [29] the Adams-Bashforth method is denoted as a PE scheme. For order 3 the PEC Scheme is absolutely stable if

$$0 < h < 2/7$$

and so, in this case, is less suitable than the simpler PE scheme. The PECE scheme has the disadvantage of requiring two evaluations of $q(x)$ per iteration although, close to x^* , it is absolutely stable if

$$0 < h < 1.728 \ldots .$$

In practice this extra stability has little effect whilst the extra evaluation at each iteration only reduces efficiency. We note that this is not the case when solving differential equations since, on the whole, at the cost of an extra evaluation the possible step size is more than doubled (see [42] for a discussion in this case). However, in this application, we are not so preoccupied with following $x(t)$ accurately and the PE scheme is adequate.

In the following section we describe some numerical experience with variable formula methods of this type. The third-order Adams-Bashforth method is coupled with methods of order three as given by (3.3.4c).

3.5. Numerical Results

We tried several multistep methods for solving (2.1.1) and present some results for an Adams-Bashforth variable-step method of order 3 coupled

with a rapidly convergent multistep method of order 3 (AB3) described in section 3.3. This method was tested for various values of h_0 and some results for $h_0 = 0.8598 \ldots$, which is a three-step method, and for $h_0 = 0.7$, which is a four-step method, are given in Table 3.1. In each case, the final step length, h^*, equalled h_0. The same step-change criteria were used as described for the single-step methods in section 2.5, except that here ε_1 was chosen to be 0.01 since, with $\varepsilon_1 = 0.05$, it was found that the methods occasionally made a premature change to stepsize h^*. The initial stepsize was again chosen to be $h^*/8$.

The algorithms were applied to the functions listed in section 2.5 and the effort required to reduce each component of f to 10^{-6} is given in Table 3.1. The format of Table 3.1 is the same as for Table 2.1.

TABLE 3.1

ALGORITHM		PROBLEM							
		1	2	3	4	5	6	7	8
AB3		23	31	14	99	27	18	54	56
	$h_0 = .859 \ldots$	25	33	15	101	28	19	59	61
		71	95	43	299	109	127	221	229
		26	35	16	95	33	21	55	58
	$h_0 = .7$	28	38	17	98	34	22	59	63
		80	108	49	288	133	148	224	237

The methods of this chapter frequently proved more efficient than the single-step methods of the previous chapter, particularly when many steps were required, for then these methods gained by requiring only one evaluation per step. This is borne out in the results shown in Table 3.1. Also the improvement in the R-order of these methods is now shown to be worthwhile.

The three-step version, with R-order $2^{1/3}$, was usually superior to the four-step method which has R-order $2^{\frac{1}{4}}$. We note again that these methods are significantly more efficient than the linearly convergent Heun method.

CHAPTER 4

CONTINUATION WITH NEWTON-LIKE METHODS

4.1. Introduction

In the previous chapters we have derived methods for solving

$$\dot{x}(t) = -J(x)^{-1}f(x) , \qquad x(0) = x_0 , \tag{4.1.1}$$

which also have rapid convergence to x^* . As described in Chapter 1, (4.1.1) can be derived in different ways, one of which is as a reformulation of

$$H\big(x(t),t\big) = 0 , \qquad x(0) = x_0 , \tag{4.1.2}$$

where $H : D \times D_t \subset R^n \times R \to R^n$ is given by

$$H(x,t) = f(x) - e^{-t}f(x_0) . \tag{4.1.3}$$

This follows because the solution of (4.1.2) also satisfies

$$\dot{x}(t) = -\partial_x H(x,t)^{-1}\partial_t H(x,t) , \qquad x(0) = x_0 . \tag{4.1.4}$$

We note that the methods of Chapters 2 and 3 integrate (4.1.1) and make no use of the fact that the solution also satisfies (4.1.2). In this chapter we discuss methods which make special use of this relation.

In section 4.2 we consider the more general problem of following the solution trajectory of (4.1.2) for a general function $H(x,t)$. We adopt this generality since it is relevant to the theory discussed in Chapter 5. We describe a well known adaptation of Newton's method for solving (4.1.2) for a sequence of values t_i , $i = 1,2,\ldots$, and we give results on the order of accuracy attained by this method. Furthermore, we discuss its computational efficiency and show how the parameters of the method can be chosen to minimise the work required to gain a certain accuracy. In section 4.3 we apply the results to the case when $H(x,t)$ is given by (4.1.3) or by

(4.1.5) $$H(x,t) = f(x) - (1-t)f(x_0) \ ,$$

in which case, (4.1.4) becomes

(4.1.6) $$\dot{x}(t) = -J(x)^{-1}f(x_0) \ , \qquad x(0) = x_0 \ .$$

In keeping with our suggestions of Chapter 1, we prefer to derive methods which have the same stability properties, close to x^* , as the methods derived in Chapters 2 and 3, for solving (4.1.1). We do this by modifying the method of section 4.2 for the case when $H(x,t)$ is given by (4.1.3). Then we give results on the accuracy of this modified method for following the solution of (4.1.1) and use Theorem 2.3.2 to deduce results on its R-order of convergence to x^* . A method essentially due to Branin [11] is described in section 4.4, since it is similar to the methods of sections 4.2 and 4.3 in that it uses the relationship given by (4.1.2) and (4.1.3). Then finally, in section 4.5, we give details of some numerical tests carried out with the methods described. The theory and numerical experience shows that the methods which use (4.1.2) directly are computationally more efficient than the methods described in Chapters 2 and 3.

4.2. Some Order Properties

In the previous chapters we considered several methods and freely discussed their orders of accuracy in following the solution of certain differential equations. This was possible because the definitions of order are well known, but here we formally define the term order so that we can compare methods which satisfy different order properties. We introduce the term H-order to emphasise that the definition is identical to that given in Henrici [34].

Consider the solution of the initial value problem

(4.2.1) $$\dot{x}(t) = g(x,t) \ , \qquad x(0) = x_0 \ ,$$

by the iterative process

(4.2.2) $$x_{i+1} = G(x_i, t_i, h_i) , \qquad i = 0,1,2,\ldots ,$$

for some $G : \hat{D} \times D_t \times D_h \subset R^n \times R \times R \to R^n$, where x_i is an approximation to $x(t_i)$ and $h_i = t_{i+1} - t_i$. For such a method we give the standard definition of order.

DEFINITION 4.2.1. Method (4.2.2) has *H-order* l *for* (4.2.1) if

$$G(x,t,h) = z(t+h) + O(h^{l+1})$$

where $z(u)$ is the solution of

$$\dot{z}(u) = g(z,u) , \qquad z(t) = x .$$

We frequently make use of the $O(\cdot)$ notation which is used in the sense that, if a and b satisfy

$$a = b + O(\delta) ,$$

then there is a constant K , independent of δ , such that $\|a-b\| \le K|\delta|$ for all sufficiently small δ .

Some methods for solving (4.2.1) cannot be described in terms of *H*-order and we give a different definition of order which is relevant to their case.

DEFINITION 4.2.2. Method (4.2.2) has *C-order* l *for* (4.2.1) if, whenever $x = x(t) + O(h)$, then

$$G(x,t,h) = x(t+h) + O(h^{l+1}) ,$$

where $x(t)$ is the solution of (4.2.1).

The term *C*-order is chosen to suggest that a method with positive *C*-order is *corrective*, in that it always tries to approximate $x(t)$. This is in contrast to methods with positive *H*-order which, at the $(i+1)$st step, try to follow the solution of

$$\dot{y}(t) = g(y,t) , \qquad y(t_i) = x_i ,$$

which is different from $x(t)$ if $x_i \neq x(t_i)$ and can be considered as an adjacent trajectory to $x(t)$. When solving a standard initial value

problem of the form (4.2.1) it would be of benefit to use a method with positive C-order since we are specifically interested in following $x(t)$. Unfortunately we cannot generate methods with positive C-order without further information about the solution trajectory and so we must be satisfied with methods possessing a positive H-order. It is ironic that, in the application of continuation methods to the solution of nonlinear equations, where it is not necessary to follow $x(t)$ accurately, we can generate methods of arbitrary C-order.

In this section we give a result on the C-order of a well known method for following the solution of (4.1.2). The method is straight forward and has been suggested for the continuation approach by several authors in the case when $H(x,t)$ is given by (4.1.5) (e.g. [4], [21], [50], [53]). However, in its basic form, it has also been used extensively for more general $H(x,t)$ (see e.g. [3], [5], [6], [21], [22], [27], [36]).

Consider (4.2.2) with $G(x,t,h)$ given by

(4.2.3a) $$G(x,t,h) = p_m(x,t,h) ,$$

where

(4.2.3b) $$p_0(x,t,h) = x$$

and

(4.2.3c) $$p_{j+1} = p_j - \partial_x H(p_j,t+h)^{-1} H(p_j,t+h) , \qquad j = 0,1,\ldots,m-1 .$$

(Note that, for brevity, we shall aften omit the arguments x,t and h from $p_j(x,t,h)$ as we have done in (4.2.3c).) This method is an obvious choice for following the solution of (4.1.2) since (4.2.3b,c) is simply Newton's method for solving $H(z,t+h) = 0$, using x as initial guess. The method, as a whole, consists of finding x_{i+1} as the solution of

$$H(z,t_{i+1}) = 0$$

by Newton's method, using x_i, the computed value of $x(t_i)$, as initial

estimate. We can prove that, under suitable conditions on $H(x,t)$, the method has C-order $2^m - 1$, but we derive this as a special case of a more general result. We can consider (4.2.2) as a sequence of major iterations, each consisting of m minor iterations given by (4.2.3c). Then it may be more efficient in practice to evaluate $\partial_x H(x,t)$ only once per major iteration or, more generally, once every r minor iterations. In this case (4.2.3) generalises to

$$G(x,t,h) = p_s(x,t,h) , \tag{4.2.4a}$$

$$p_0(x,t,h) = x , \tag{4.2.4b}$$

$$y_j^{(0)}(x,t,h) = p_j(x,t,h) , \qquad j = 0,1,\ldots,s-1 , \tag{4.2.4c}$$

$$y_j^{(i+1)} = y_j^{(i)} - \partial_x H\big(y_j^{(0)},t+h\big)^{-1} H\big(y_j^{(i)},t+h\big) , \tag{4.2.4d}$$

for $j = 0,1,\ldots,s-1$, $i = 0,1,\ldots,r-1$, and

$$p_{j+1} = y_j^{(r)} , \tag{4.2.4e}$$

and $m = rs$. As before we have omitted the arguments of p_j and $y_j^{(i)}$. We can now prove a theorem on the C-order of this method.

THEOREM 4.2.1. *Suppose that* $H : D \times D_t \subset R^n \times R \to R^n$ *is such that* $x(\tau) \in \mathrm{Int}(D)$ *and* $\tau \in \mathrm{Int}\big(D_t\big)$ *satisfy* $H\big(x(\tau),\tau\big) = 0$. *Suppose also that* $\partial_x H(x,t)$ *and* $\partial_t H(x,t)$ *are Lipschitz continuous in a neighbourhood* S *of* $\big(x(\tau),\tau\big)$. *Assume also that* $\partial_x H(x,t)^{-1}$ *exists and is bounded on* S. *Then* (4.2.2), *where* $G(x,t,h)$ *is given by* (4.2.4), *has* C-*order* $(r+1)^s - 1$ *for* (4.1.4).

Proof. With the given assumptions, the Implicit Function Theorem ensures the existence of a unique continuous solution, $x(\tau+h)$, of (4.1.2), and therefore of (4.1.4), in a neighbourhood $S \times S_t \subset S$, where $S = S\big(x(\tau),\delta\big)$, for some $\delta > 0$, and $S_t = (\tau-\gamma,\tau+\gamma)$, for some $\gamma > 0$.

We assume subsequently that $|h| < \gamma$.

To prove the theorem we require a bound on $\|x(\tau+h)-G(x,\tau,h)\|$, where $x = x(\tau) + O(h)$, in terms of h . We note that, from (4.2.4a),

$$\|x(\tau+h)-G(x,\tau,h)\| = \|x(\tau+h)-p_s(x,\tau,h)\|$$

and we derive the required result by induction. Define α_j and $\beta_j^{(i)}$, for $j = 0,1,\ldots,s$ and $i = 0,1,\ldots,r$, by

$$\alpha_j = \|x(\tau+h)-p_j\| ,$$

$$\beta_j^{(i)} = \left\|x(\tau+h)-y_j^{(i)}\right\| , \qquad j \neq s ,$$

where we have omitted the arguments (x,τ,h) from p_j and $y_j^{(i)}$. Now, from the given conditions, there exist constants K_0, K_1, K_2 such that

$$\left\|\partial_x H(x,t)^{-1}\right\| \leq K_0 , \tag{4.2.5}$$

$$\|\partial_t H(x,t)\| \leq K_1 \tag{4.2.6}$$

and

$$\|\partial_x H(x,t)-\partial_x H(y,t)\| \leq K_2\|x-y\| \tag{4.2.7}$$

for all $x,y \in S$ and for all $t \in S_t$. Furthermore, it follows from the Lipschitz continuity of $\partial_x H(x,t)$ and $\partial_t H(x,t)$ and from [53, Theorem 3.2.5] that, for any $t \in S_t$ and for any $x,y \in S$, if $u(x,y,t)$ is given by

$$H(x,t) = H(y,t) + \partial_x H(y,t)(x-y) + u(x,y,t) , \tag{4.2.8}$$

then

$$\|u(x,y,t)\| \leq K_3\|x-y\|^2 \tag{4.2.9}$$

for some constant K_3 . As in the proof of Corollary 2.3.1 we may assume that

$$\left\|\begin{matrix} a \\ \alpha \end{matrix}\right\| = \|a\| + |\alpha|$$

and (4.2.7) and (4.2.9) follow immediately. That both equations are true for any norm follows from the equivalence of norms.

Now, from [53, Theorem 3.2.3], (4.1.4), (4.2.5) and (4.2.6), it follows that the solution $x(\tau+h)$ of (4.1.4) satisfies

$$\|x(\tau+h)-x(\tau)\| \leq K_0 K_1 |h|$$

for any $h \in (-\gamma,\gamma)$. So, assuming $p_0 = x(\tau) + O(h)$, we have

(4.2.10) $$\alpha_0 = \|x(\tau+h)-p_0\| = O(h) .$$

Thus, for small enough h, p_0 is contained in an open sphere $S(h) \subset S$ centred at $x(\tau+h)$. We assume that p_j and $y_j^{(i)}$ are also in $S(h)$, for some i,j, and prove by induction that $p_j \in S(h)$ and $y_j^{(i)} \in S(h)$ for all i,j. Also, we assume that, for some i and j,

(4.2.11) $$\beta_j^{(i)} = O\left(h^{(i+1)(r+1)^j}\right)$$

and

(4.2.12) $$\alpha_j = O\left(h^{(r+1)^j}\right)$$

and we prove by induction that (4.2.11) and (4.2.12) are true for all i,j. The case when $i = j = 0$ is given in (4.2.10).

We begin by noting from (4.2.4d), that

$$\beta_j^{(i+1)} = \left\| x(\tau+h)-y_j^{(i)}+\partial_x H\left(y_j^{(0)},\tau+h\right)^{-1} H\left(y_j^{(i)},\tau+h\right)\right\|$$

and, because $H\left(x(\tau+h),\tau+h\right) = 0$, we have

$$\beta_j^{(i+1)} = \left\| x(\tau+h)-y_j^{(i)}+\partial_x H\left(y_j^{(0)},\tau+h\right)^{-1}\left[H\left(y_j^{(i)},\tau+h\right)-H\left(x(\tau+h),\tau+h\right)\right]\right\| .$$

Then, from (4.2.8), since $y_j^{(i)} \in S(h)$,

$$\beta_j^{(i+1)} = \left\| x(\tau+h)-y_j^{(i)}-\partial_x H\left(y_j^{(0)},\tau+h\right)^{-1}\left[\partial_x H\left(x(\tau+h),\tau+h\right)\left(x(\tau+h)-y_j^{(i)}\right)-u\right]\right\| ,$$

where $u = u\left(y_j^{(i)},x(\tau+h),\tau+h\right)$ and, since $y_j^{(i)} \in S(h)$ and $|h| < \gamma$,

(4.2.13) $$\|u\| \leq K_3 \left\| x(\tau+h) - y_j^{(i)} \right\|^2 = K_3 \beta_j^{(i)^2} .$$

Now we have

$$\beta_j^{(i+1)} = \left\| \partial_x H\left(y_j^{(0)}, \tau+h\right)^{-1} \left[\left(\partial_x H\left(y_j^{(0)}, \tau+h\right) - \partial_x H\left(x(\tau+h), \tau+h\right) \right) \left(x(\tau+h) - y_j^{(i)} \right) + u \right] \right\| .$$

Because $y_j^{(0)} \in S(h)$, it follows from (4.2.5), (4.2.7) and (4.2.13), together with (4.2.4c) that

(4.2.14) $$\beta_j^{(i+1)} \leq K_0 K_2 \alpha_j \beta_j^{(i)} + K_0 K_3 \beta_j^{(i)^2} .$$

Let $A = K_0 K_2$ and $B = K_0 K_3$. For small enough h , $A\alpha_j + B\beta_j^{(i)} < 1$ and so $\beta_j^{(i+1)} < \beta_j^{(i)}$, hence $y_j^{(i+1)} \in S(h)$. Now, by induction $y_j^{(i)} \in S(h)$ for $i = 0,1,\ldots,r$, and since $y_j^{(r)} = p_{j+1}$, it follows that $p_{j+1} \in S(h)$. Again by induction, $p_j \in S(h)$, $j = 0,1,\ldots,s$.

It also follows from (4.2.11), (4.2.12) and (4.2.14) that

$$\beta_j^{(i+1)} = O\left(h^{(i+2)(r+1)^j}\right)$$

and so we have derived (4.2.11) with i replaced by $i + 1$. Using the result that $y_j^{(i)} \in S(h)$ for each i,j , we apply the induction to give $\beta_j^{(r)} = O\left(h^{(r+1)^{j+1}}\right)$. Since $\alpha_{j+1} = \beta_j^{(r)}$ we have derived (4.2.12) with j replaced by $j + 1$. We have now completed the induction and, applying (4.2.12), we have

$$\alpha_s = O\left(h^{(r+1)^s}\right) .$$

Since $\alpha_s = \|x(\tau+h) - G(x,\tau,h)\|$ we have the desired result. □

As an example, (4.2.3) results from (4.2.4) by taking $r = 1$ and $s = m$ and the C-order is $2^m - 1$. Also, evaluating $\partial_x H(x,t)$ only once per major iteration corresponds to $r = m$ and $s = 1$, in which case the

C-order is m .

We may now compare the work required by the method of (4.2.4) to attain various possible C-orders. We assume that one evaluation of $\partial_x H(x,t)$ is equivalent to k evaluations of $H(x,t)$ (e.g. if $\partial_x H(x,t)$ is a full matrix then $k = n$ may be appropriate or, if $\partial_x H(x,t)$ is tridiagonal, then $k = 3$ may be more reasonable). One measure of the work per iteration is the number of equivalent function evaluations and so, for the method given by (4.2.4) to attain C-order l , N equivalent function evaluations are required, where

$$N = sk + sr$$

and $(r+1)^s - 1 = l$. Given a specific value of l we can now find the optimal values of r and s to minimise N . It is trivial to show that, for $l < 6$, the optimal choice is always $r = m = l$, $s = 1$ (assuming $k \geq 1$). For higher orders the choice depends upon k . For example, C-order 8 can be achieved by taking

$$r = 8 \ , \quad s = 1 \ , \text{ giving } \quad N = 8 + k$$

or

$$r = 2 \ , \quad s = 2 \ , \text{ giving } \quad N = 4 + 2k \ .$$

The optimal choice for $k = 3$ is $r = 2$, $s = 2$ and if $k > 4$, the optimal choice is $r = 8$, $s = 1$. These results show that it will often be more efficient, in this sense, to evaluate $\partial_x H(x,t)$ only once per major iteration. In practice of course the step sizes, h_i , required to maintain the desired accuracy when $s = 1$ may be smaller than for the case when $s > 1$ and so the number of major iterations may be greater. However, in section 4.5, we give some numerical results which indicate that it is often less efficient to evaluate $\partial_x H(x,t)$ at each minor iteration.

4.3. An Adaptive Newton Method

Methods with high C-order are suitable for solving the equation (4.1.2) when, as in Chapter 5, values of $x(t)$ are required for each t. However, in this application, only x^* is required. Demanding high accuracy along the trajectory gives greater reliability in finding x^* but if we wish to balance reliability and efficiency we can consider reducing our concern for high accuracy in following $x(t)$. With this in mind we consider some choices of the function $H(x,t)$. In order to be able to apply Theorem 4.2.1 to each choice of $H(x,t)$ we assume, unless stated otherwise, that $f : D \subset R^n \to R^n$ has a continuous second derivative on D and that $J(x)^{-1}$ exists and is bounded on D.

Consider first $H(x,t)$ given by (4.1.5), then (4.2.3) becomes

$$G(x,t,h) = p_m(x,t,h) , \tag{4.3.1a}$$

$$p_0(x,t,h) = x \tag{4.3.1b}$$

and

$$p_{j+1} = p_j - J(p_j)^{-1}[f(p_j)-(1-t-h)f(x_0)] , \tag{4.3.1c}$$

$j = 0,1,\ldots,m-1$, and the method has C-order $2^m - 1$ for (4.1.6). Note that if $t_M = 1$, then x_M is only an approximation to x^* and further refinement may be necessary. This is the method used by several authors (e.g. [4], [21], [50]) and, in particular, by Ortega and Rheinboldt [53], who make the obvious suggestion of setting $t_i = 1$, $i \geq M$, in (4.3.1), which gives Newton's method for solving $H(x,1) = 0$. Note that, for this to converge, we are assuming that x_M is in the region of convergence of Newton's method at x^*. This is a corrective method for solving (4.1.5) but, as mentioned previously, we consider it preferable to integrate (4.1.1), or equivalently (4.1.3), because of its Liapunov stability. In this case we do not require a positive C-order, since neighbouring trajectories all

converge to x^* , and a positive H-order is adequate. We now generate a method, similar to (4.2.4), with arbitrary H-order for (4.1.1).

With $H(x,t)$ given by (4.1.3), (4.2.3c) becomes

$$p_{j+1} = p_j - J(p_j)^{-1}\left[f(p_j) - e^{-t-h}f(x_0)\right]$$

which, by Theorem 4.2.1, gives a method with C-order $2^m - 1$ for (4.1.1). However we can modify this, using (4.1.3) to note that $x(t)$ satisfies

$$f(x(t+h)) = e^{-t-h}f(x_0) = e^{-h}f(x(t)) \ .$$

This suggests the iterative process given by

(4.3.2) $$x_{i+1} = G(x_i, h_i)$$

where

(4.3.3a) $$G(x,h) = p_m(x,h) \ ,$$

(4.3.3b) $$p_0(x,h) = x$$

and

(4.3.3c) $$p_{j+1} = p_j - J(p_j)^{-1}\left[f(p_j) - e^{-h}f(x)\right] \ .$$

Then, as in section 4.2, we can prove that the method has H-order $2^m - 1$ for (4.1.1). However, we again generalise the result to consider evaluating $J(x)$ once every r minor iterations. With $m = rs$, (4.3.3) generalises to

(4.3.4a) $$G(x,h) = p_s(x,h) \ ,$$

(4.3.4b) $$p_0(x,h) = x \ ,$$

(4.3.4c) $$y_j^{(0)}(x,h) = p_j(x,h) \ , \qquad j = 0,1,\ldots,s-1 \ ,$$

(4.3.4d) $$y_j^{(i+1)} = y_j^{(i)} - J(y_j^{(0)})^{-1}\left[f(y_j^{(i)}) - e^{-h}f(x)\right] \ ,$$

$j = 0,1,\ldots,s-1$ and $i = 0,1,\ldots,r-1$, and

(4.3.4e) $$p_{j+1} = y_j^{(r)} \ .$$

The theorem in its generality is now given.

THEOREM 4.3.1. *Suppose that* $f : D \subset R^n \to R^n$ *has a derivative* $J(x)$ *which is Lipschitz continuous in a neighbourhood* S *of a point* $x \in \text{Int}(D)$. *Assume also that* $J(x)^{-1}$ *exists and is bounded on* S . *Then the method given by* (4.3.2) *and* (4.3.4) *has* H*-order* $(r+1)^s - 1$ *for* (4.1.1).

This result is a special case of the following theorem, Theorem 4.3.2, and we postpone the proof until then.

In its present form, this method converges only linearly to x^* and so it is necessary to amend it in such a way as to maintain the H-order properties and to allow for rapid final convergence to x^* .

We consider now (4.3.2) as a single step iterative process discussed in section 2.3. Some algebraic manipulation shows that, with $G(x,h)$ defined in (4.3.4), $\partial_x G(x^*,h)$ is given by

$$\partial_x G(x^*,h) = e^{-h} I \ .$$

In this case, Theorem 2.3.2 shows that there is no value of h for which the convergence rate can be faster than linear. This is because $\eta\left(\partial_x G(x^*,h)\right) = e^{-h} > 0$ for all h . This is unsatisfactory and so we modify (4.3.4d) to be

$$y_j^{(i+1)} = y_j^{(i)} - J\left(y_j^{(0)}\right)^{-1}\left[f\left(y_j^{(i)}\right) - \phi(h) f(x)\right] , \tag{4.3.5}$$

for $j = 0,1,\ldots,s-1$, $i = 0,1,\ldots,r-1$ and $\phi : D_h \subset R \to R$, where D_h is the open interval $(-\gamma,\gamma)$ for some $\gamma > 0$, is a function which will be an approximation to e^{-h} . We first prove a theorem on the H-order of this method, noting that the case when $\phi(h) = e^{-h}$ gives the result in Theorem 4.3.1.

THEOREM 4.3.2. *Suppose* $\phi : (-\gamma,\gamma) \subset R \to R$ *is continuous, where*

$\gamma > 0$, *and* $\phi(h) = e^{-h} + O(h^{k+1})$, $k \geq 0$. *Then, under the conditions of Theorem* 4.3.1, *the method given by* (4.3.2), (4.3.4a,b,c,e) *and* (4.3.5) *has H-order* $\min\{(r+1)^s-1,k\}$ *for* (4.1.1).

Proof. The proof is similar to the proof of Theorem 4.2.1 and so here we give only an outline. As before, the Implicit Function Theorem ensures the existence of a unique continuous solution, $z(h)$, of

$$f(z(h)) - e^{-h}f(x) = 0 , \tag{4.3.6}$$

and therefore of

$$\dot{z}(h) = -J(z)^{-1}f(z) , \qquad z(0) = x ,$$

for $h \in (-\delta,\delta)$, for some δ such that $0 < \delta < \gamma$.

We require a bound on $\|z(h)-G(x,h)\|$ in terms of h . We define α_j and $\beta_j^{(i)}$, $j = 0,1,\ldots,r$ and $i = 0,1,\ldots,s$, by

$$\alpha_j = \|z(h)-p_j\| ,$$

$$\beta_j^{(i)} = \left\|z(h)-y_j^{(i)}\right\| , \qquad j \neq s ,$$

and, from (4.3.5), we have

$$\beta_j^{(i+1)} = \left\|z(h)-y_j^{(i)}+J\left(y_j^{(0)}\right)^{-1}\left[f\left(y_j^{(i)}\right)-\phi(h)f(x)\right]\right\| .$$

Let $\psi(h) = \phi(h) - e^{-h}$, then for each i and j ,

$$\beta_j^{(i+1)} = \left\|z(h)-y_j^{(i)}+J\left(y_j^{(0)}\right)^{-1}\left[f\left(y_j^{(i)}\right)-e^{-h}f(x)-\psi(h)f(x)\right]\right\|$$

and it follows from (4.3.6) that

$$\beta_j^{(i+1)} \leq \left\|z(h)-y_j^{(i)}+J\left(y_j^{(0)}\right)^{-1}\left[f\left(y_j^{(i)}\right)-f\left(z(h)\right)\right]\right\| + \left\|J\left(y_j^{(0)}\right)^{-1}f(x)\right\||\psi(h)| .$$

Now, using the assumptions on f and the method used in the proof of Theorem 4.3.1, we can show that, for h sufficiently small, there exist constants C_1,C_2,C_3 , independent of h such that

$$\beta_j^{(i+1)} \leq C_1\alpha_j\beta_j^{(i)} + C_2\beta_j^{(i)^2} + C_3|\psi(h)| \ .$$

By assumption, $\psi(h) = O(h^{k+1})$ and so, for small enough h ,

$$\beta_j^{(i+1)} \leq C_1\alpha_j\beta_j^{(i)} + C_2\beta_j^{(i)^2} + C_4|h|^{k+1} \ ,$$

for some constant C_4 . We can now apply an analogous induction argument to that used in the proof of Theorem 4.3.1 to show that, for small enough h ,

$$\alpha_s \leq K_1|h|^{(r+1)^s} + K_2|h|^{k+1}$$

for suitable constants K_1,K_2 . Since $\alpha_s = \|z(h)-G(x,h)\|$ we have the required result. □

We can now see that this modified method, given by (4.3.2), (4.3.4a,b,c), (4.3.5) and (4.3.4e), can give superlinear convergence to x^* if $\phi(h)$ is suitably chosen. Some simple algebra shows that

$$\partial_x G(x^*,h) = \phi(h)I \ .$$

If $\phi'(h)$ is Lipschitz continuous on $(-\gamma,\gamma)$ and $f''(x)$ is Lipschitz continuous on a neighbourhood of x^* , then the conditions of Theorems 2.3.2 and 2.3.3 are satisfied and it follows that the process converges locally if, for some δ , $|\phi(h_i)| \leq 1 - \delta < 1$ for each i . Also, the R-order is at least 2 if the sequence $\{h_i\}$ converges sufficiently fast to $h^* \in (-\gamma,\gamma)$, where h^* satisfies $\phi(h^*) = 0$. If we wish to gain rapid convergence to a root of f and maintain a certain H-order, l say, then a suitable choice for $\phi(h)$ is

$$\phi(h) = \sum_{j=0}^{k} \frac{(-h)^j}{j!}$$

where $k = l$ or $l + 1$ is chosen to be odd, for then $\phi(h)$ satisfies the conditions of Theorem 4.3.2 for any $\gamma > 0$ and has a unique positive root. A practical algorithm is therefore to allow the stepsizes, h_i , to increase,

subject to suitable step-length tests, and finally, to hold h_i fixed at h^*. If k is large enough, the H-order of the method will be $(r+1)^s - 1$ and the sequence $\{x_i\}$ will converge rapidly to x^*. The method therefore changes in a continuous way from one which follows the solution trajectory $x(t)$ accurately to one which converges rapidly to x^*.

We now look at some choices of r and s in (4.3.5) to show that, when $h_i = h^*$, the method becomes one of the well known methods for solving $f(x) = 0$. With $r = 1$, $s = m$, $h_i = h^*$, (4.3.4e) and (4.3.5) together become

$$p_{j+1} = p_j - J(p_j)^{-1} f(p_j) , \qquad j = 0,1,\ldots,m-1 ,$$

and we have Newton's method. The sequence $\{x_i\}$ converges to x^* with R-order 2^m. For general r and s, when $h_i = h^*$ for each i, the method becomes that of Shamanskii, with exact Jacobian, and $\{x_i\}$ converges to x^* with R-order $(r+1)^s$ (see [12], [68], [70] for further details on this method).

Methods of the type discussed in this section were tried on the test functions described in section 2.5 and some numerical results are presented in section 4.5. The short discussion in section 4.2 on the work required to attain a specific C-order applies equally well to the methods of this section, with C-order replaced by H-order. For that reason, in our numerical results we consider both the standard choice, with $r = 1$, $s = m$, and the choice $r = m$, $s = 1$, which the theory indicates may be more efficient.

4.4. Branin's Method

On the assumption that $f''(x)$ is Lipschitz continuous in a neighbourhood

of x^* we can apply Theorems 2.3.2 and 2.3.3 to a method essentially due to Branin [11]. We discuss it here since it is similar to the methods of section 4.3 in that it attempts to integrate (4.1.1) by making specific use of the relation

$$f(x(t)) = e^{-t}f(x_0) = e^{-t}f_0 \tag{4.4.1}$$

for all $t \geq 0$. In this method $x(t_{i+1})$ is estimated by the first order prediction

$$p_0 = x_i - h_i J(x_i)^{-1} f(x_i) .$$

Then the component, v, of $f(p_0)$ orthogonal to f_0 is

$$v = \left[I - \frac{f_0 f_0^T}{f_0^T f_0} \right] f(p_0) .$$

Since $f(x_{i+1})$ should be parallel to f_0, a new estimate p_1 of x_{i+1} is calculated from

$$p_1 = p_0 - J(p_0)^{-1} v ,$$

which is the first order attempt to annihilate v. This process is repeated a finite number of times until the derived estimate of x_{i+1} is close enough to satisfying (4.4.1). Again omitting the arguments of $p_j(x,h)$, the process can be written as

$$x_{i+1} = G(x_i, h_i) \tag{4.4.2}$$

where

$$G(x,h) = p_m(x,h) , \tag{4.4.3a}$$

$$p_0(x,h) = x - hJ(x)^{-1} f(x) \tag{4.4.3b}$$

and, for $j = 0,1,\ldots,m-1$,

$$p_{j+1} = p_j - J(p_j)^{-1} Z_0 f(p_j) , \tag{4.4.3c}$$

where $Z_0 = \left(I - f_0 f_0^T / f_0^T f_0\right)$.

We can now apply Corollary 2.3.1 to give conditions for x^* to be a point of attraction of this method. Using the fact that $f(x^*) = 0$, for each h ,

$$\begin{aligned} \partial_x G(x^*,h) &= \partial_x p_m(x^*,h) \\ &= \left[I - J(x^*)^{-1} Z_0 \; J(x^*)\right] \partial_x p_{m-1}(x^*,h) \\ &= \left[I - J(x^*)^{-1} Z_0 \; J(x^*)\right]^m \partial_x p_0(x^*,h) \; . \end{aligned}$$

The bracketed matrix is idempotent and, by differentiating (4.4.3b) and evaluating at (x^*,h) , we have

$$\partial_x G(x^*,h) = \left[I - J(x^*)^{-1} Z_0 \; J(x^*)\right](1-h) \; .$$

Now

$$I - J(x^*)^{-1} Z_0 \; J(x^*) = \frac{1}{f_0^T f_0} J(x^*)^{-1} f_0 f_0^T J(x^*)$$

which has one non zero eigenvalue equal to 1 . Thus

$$\eta\left(\partial_x G(x^*,h)\right) = |1-h| \tag{4.4.4}$$

and it follows from Corollary 2.3.1 that the process converges locally to x^* if $0 < \delta \le h_i \le 2 - \delta$, $i = 0,1,\ldots$, for some δ . Also, from Theorem 2.3.3 final convergence is superlinear if $\lim_{i\to\infty} h_i = 1$ and has R-order ≥ 2 if $\{h_i\}$ converges to 1 sufficiently fast. In fact, in general, the R-order is exactly 2 , since, as $\{x_i\}$ converges to x^* , the corrections given by (4.4.3c) do not improve on x_i as an estimate of x^* , and the method tends to Newton's method.

Although we cannot discuss Branin's method in terms of H- or C-orders for a particular differential equation, the sequence $\{p_j\}$ in (4.4.3c) converges to a point on the solution trajectory of (4.1.1) and we

can apply the ideas of section 4.2 to show that holding the Jacobian fixed over a minor iteration may improve efficiency. In this case it is straightforward to show that, under the conditions of Theorem 4.3.1, if

$\|Z_0 f(x)\| = O(h)$, then $\|Z_0 f(p_m)\| = O(h^{2^m})$ whereas if (4.4.3c) is replaced by

$$p_{j+1} = p_j - J(x)^{-1} Z_0 f(p_j) , \qquad j = 0,1,\ldots,m-1 , \tag{4.4.5}$$

then $\|Z_0 f(p_m)\| = O(h^{m+1})$. Equation (4.4.4) is unaffected by this change and so the R-order is unchanged.

Thus, a practical algorithm is to adapt h to maintain accuracy but, when h increases to 1 , hold the step fixed so that, as x^* is approached, the R-order becomes equal to 2 . The performance of two such algorithms, based upon both (4.4.3c) and (4.4.5), is discussed in the next section.

4.5. Numerical Results

To make some comparisons, we tested implementations of the three methods discussed in this chapter. First is the corrective method given by (4.2.2) and (4.3.1), which is described, for example, by Ortega and Rheinboldt [53], and denoted by OR/1. Next is the H-order method, derived in the previous section, with $r = 1$, $s = m$ and defined by (4.3.2) with $G(x,h)$ given by

$$G(x,h) = p_m(x,h) , \tag{4.5.1a}$$

$$p_0(x,h) = x \tag{4.5.1b}$$

and

$$p_{j+1} = p_j - J(p_j)^{-1}[f(p_j) - \phi(h) f(x)] , \tag{4.5.1c}$$

$j = 0,1,\ldots,m-1$. So that the method would be third-order, we chose $\phi(h)$

to be

$$\phi(h) = 1 - h + \frac{h^2}{2} - \frac{h^3}{6} .$$

This method is denoted by NEW/1. Finally, we implemented Branin's method, given by (4.4.2) and (4.4.3) and denote it by BRANIN/1. In each of these implementations the Jacobian is evaluated at each minor iteration and, for comparison, second versions were tested in which the Jacobian was evaluated only once per major iteration. Firstly, in the method OR/1, (4.3.1c) was replaced by

$$p_{j+1} = p_j - J(x)^{-1}[f(p_j)-(1-t-h)f(x_0)] ,$$

$j = 0,1,\ldots,m-1$, to give the method OR/2. Secondly, in the NEW/1 method, (4.5.1c) was replaced by

$$p_{j+1} = p_j - J(x)^{-1}[f(p_j)-\phi(h)f(x)] ,$$

$j = 0,1,\ldots,m-1$, to give the method NEW/2. Finally, in the BRANIN/1 method, (4.4.3c) was replaced by (4.4.5) to give BRANIN/2.

To facilitate comparison with those methods, of order 3 , tested in Chapters 2 and 3, each of the above methods was implemented with m fixed so that their orders were 3 . As might be expected, we found, in each case, an improvement in efficiency if we allowed m to vary over each iteration so that the methods became variable order with maximum order 3 . It is for these implementations that the results are given.

The success of the algorithms under discussion is, to some extent, dependent on the way in which the step sizes are chosen. Rheinboldt [61] has looked in more detail at efficient step adjustments on the basis of estimates of the local attraction domains but again, for the purpose of comparing different methods, we chose the step test which was described in section 2.5 for the single-step methods. Again, we emphasise that this is not necessarily the best way of choosing step sizes, however it proved adequate for our purposes. For clarity, we briefly describe the step test

again. Let $f(x_i) = f_i$ and Z_i be given by

$$Z_i = I - \frac{f_i f_i^T}{f_i^T f_i} .$$

In OR/1 and OR/2 the step size was varied according to

$$h_{i+1} = \min(1-t_{i+1}, \alpha h_i)$$

where α is given by

$$\alpha = \begin{cases} 2 & \text{if } 0 \le \delta \le \varepsilon_1 , \\ 1 & \text{if } \varepsilon_1 < \delta \le \varepsilon_2 , \\ 0.5 & \text{if } \varepsilon_2 < \delta \le \varepsilon_3 \end{cases} \tag{4.5.2}$$

and $\delta = \|Z_0 f_{i+1}\|$. In NEW/1 & 2 and BRANIN/1 & 2, h_{i+1} was given by

$$h_{i+1} = \min(h^*, \alpha h_i)$$

where $h^* = 1.0$ for BRANIN, $h^* = 1.596...$, which is the unique root of $\phi(h)$, for NEW and α is given by (4.5.2) with $\delta = \|Z_0 f_{i+1}\|$ for BRANIN and $\delta = \|Z_i f_{i+1}\|$ for NEW.

The estimate p_j was accepted as x_{i+1} on two conditions. Firstly, if j equalled 2 for OR/1, NEW/1 and BRANIN/1 or if j equalled 3 for OR/2, NEW/2 and BRANIN/2 , i.e. the maximum order in each case was 3 . Secondly, if $\|Z_0 f(p_j)\| \le \varepsilon_1$ for OR/1 & 2 and BRANIN/1 & 2 or if $\|Z_i f(p_j)\| \le \varepsilon_1$ for NEW/1 & 2, i.e. the demand for order 3 was relaxed whenever possible. Having found x_{i+1} , the conditions for rejecting the step and repeating it with half the step length were the same as for the methods described in section 2.5, with the appropriate δ . Finally, the values of ε_i , $i = 1,2,3$, were fixed at the values chosen in section 2.5 and the initial step was chosen as $h^*/8$ for BRANIN and NEW, where h^* is the final stepsize in each case, and 0.125 for OR, which is essentially one eighth of its final step size.

The methods were tested on the eight problems described in section 2.5 and we tabulate the results in Table 4.1. The format of the table is the same as for Table 2.1 and the stopping criterion was again that each component of f should be less than 10^{-6}.

TABLE 4.1

ALGORITHM	PROBLEM							
	1	2	3	4	5	6	7	8
	10	20	6	22	23	8	26	30
OR/1	11	22	7	26	26	9	31	36
	31	62	19	70	95	57	109	126
	8	6	6			7	10	12
OR/2	12	35	7	**	*	10	62	61
	28	47	19			52	92	97
	10	15	6	26	15	8	27	29
NEW/1	11	18	7	28	16	9	31	33
	31	48	19	80	61	57	112	120
	8	8	6	36	9	7	14	15
NEW/2	11	29	7	134	20	9	61	65
	27	45	19	206	56	51	113	110
	11		6		19	9	34	38
BRANIN/1	12	**	7	**	20	10	37	41
	34		19		77	64	139	155
	9	8	6		7	8	15	17
BRANIN/2	12	20	7	**	12	11	60	66
	30	36	19		33	59	105	117

* - h reduced to minimum allowed, viz. $2^{-13}h^*$.

** - failed to converge in 200 function evaluations.

The first conclusion from the results is that the three algorithms described here represent a significant improvement upon those described in

Chapters 1 and 2, and therefore upon those of e.g. [7], [9], [10], [40], [41], which do not make use of the special characteristics of the solution of (4.1.1). The important feature of the methods described in this chapter is that they are adaptive, in that they choose the order of accuracy required at each iteration, depending on the local errors. Thus, at times when a low order is sufficient, these methods save function evaluations by performing only as much work as is necessary to maintain the required accuracy. We surmise that if we used variable order Runge-Kutta or multi-step methods, in place of the fixed order methods described in Chapters 2 and 3, then we could make similar savings in work. However the resulting algorithms would be rather more complicated than the simple methods of this chapter.

We also conclude from the results that holding the Jacobian fixed over each major iteration was almost always more efficient. This is as we were led to expect by the theory of section 4.2. The only notable exceptions were when the solution trajectory ran close to a region where $J(x)$ was singular. Thus it seems reasonable to monitor the value of $\mathrm{Det}\big(J(x)\big)$, which can be done at little extra cost since $J(x)$ is factorised into triangular factors at each evaluation, and evaluate $J(x)$ more often only when $\mathrm{Det}\big(J(x)\big)$ becomes small.

Finally, it appears that OR and NEW are more effective than BRANIN. Any further conclusions about comparative behaviour can only come from practical trials, but our experience indicates that NEW/1 and NEW/2 are the more robust of the methods tested. In particular, for these methods the step control is easier and they have a greater chance of success in more difficult problems.

CHAPTER 5

TURNING POINTS IN BIFURCATION THEORY

5.1. Introduction

In many physical problems it is necessary to solve a system of nonlinear equations of the form given in (1.1). In order to conform more closely to the literature on such problems, in this chapter we prefer to replace the variable t by λ. Then, we are interested in the solution of equations of the form

(5.1.1) $$H(x,\lambda) = 0 ,$$

$H : D \subset R^n \times R \to R^n$, where the solution vector $x(\lambda)$ is a simple, continuously differentiable arc in R^n dependent upon the scalar parameter λ. Problems frequently occur where it is of interest to follow the trajectory and to find $x(\lambda)$ numerically for values of λ sufficient to define the curve $\big(x(\lambda),\lambda\big)$, which is called *a solution branch* of (5.1.1) in $R^n \times R$. This solution branch will often exhibit complex behaviour, but we recall from (1.5) that $x(\lambda)$ satisfies the differential equation

(5.1.2) $$\frac{dx}{d\lambda}(\lambda) = -A(x,\lambda)^{-1}d(x,\lambda) ,$$

where $A(x,\lambda) = \partial_x H(x,\lambda)$ and $d(x,\lambda) = \partial_\lambda H(x,\lambda)$. We assume throughout this chapter that $H(x,\lambda)$ is twice continuously differentiable with respect to x and λ on D. Thus, if $A\big(x(\lambda),\lambda\big)$ is nonsingular, then $x(\lambda)$ is continuous at λ. Points on $\big(x(\lambda),\lambda\big)$ at which $A(x,\lambda)$ is singular are called *critical points* (often bifurcation points, or singular points) and have received a large amount of attention in the literature (see for example, [3], [6], [17], [21], [33], [36], [37], [38], [49], [63], [64], [66], [69]). The primary purpose of this chapter is to describe some efficient methods for

the accurate determination of certain critical points. Firstly we consider the simplest type, which is a point (x^*,λ^*), where $x^* = x(\lambda^*)$, such that

(5.1.3a) $$\text{rank}[A(x^*,\lambda^*)] = n - 1 ,$$

(5.1.3b) $$\text{rank}[A(x^*,\lambda^*) \quad d(x^*,\lambda^*)] = n$$

and if $(x(\lambda),\lambda) \neq (x^*,\lambda^*)$ and λ is sufficiently close to λ^*, then $A(x(\lambda),\lambda)$ is nonsingular. Such a point is called a *limit point*. If the solution branch $(x(\lambda),\lambda)$ through (x^*,λ^*) exists for all λ in an open neighbourhood of λ^*, then (x^*,λ^*) is called a *point of inflexion* otherwise it is a *turning point*. In structural problems, a turning point represents the boundary between stability and instability of the system.

Prior to discussing methods for finding a turning point, in section 5.2 we consider the problem of following a solution branch through a turning point. We describe a method which is similar to that developed by Riks [66] and Menzel and Schwetlick [49] but involves less work per step. It appears that this new method has also been developed independently by Rheinboldt, (private communication) whose description is currently available only in manuscript form [65].

In section 5.3 we describe methods for the accurate determination of (x^*,λ^*). Simpson [69] described an iterative method which requires, at each iteration, the solution of (5.1.1) for some λ and the estimation of the smallest eigenvalue of $A(x(\lambda),\lambda)$. His method converges linearly to (x^*,λ^*) and is suitable only for symmetric $A(x,\lambda)$. Here we describe methods which require less work per iteration, have second order convergence to (x^*,λ^*) and do not require $A(x,\lambda)$ to be symmetric.

A *simple bifurcation point* on a solution branch is a critical point (x_B,λ_B), where $x_B = x(\lambda_B)$, which satisfies the same conditions as a turning point except that (5.1.3a) and (5.1.3b) are replaced by

(5.1.4a) $$\text{rank}[A(x_B,\lambda_B)] = n - 1$$

and

(5.1.4b) $$\operatorname{rank}[A(x_B,\lambda_B) \quad d(x_B,\lambda_B)] = n - 1 .$$

Given an additional condition on the second derivative of $H(x,\lambda)$, Crandall and Rabinowitz [17] have shown that, in a neighbourhood of (x_B,λ_B), the totality of solutions of (5.1.1) form two continuous curves intersecting only at (x_B,λ_B). In many applications it is necessary to follow one of these, often called the primary branch, and on detecting the presence of a secondary branch, to follow it (see Keller and Langford [37] and Rheinboldt [63], [64] for methods). In the case when the primary branch satisfies some symmetry relations it is often possible to generate methods which converge to (x_B,λ_B) with second order convergence and we discuss this in section 5.4. One of the methods also has the advantage of providing an approximation to the null vector of $A(x_B,\lambda_B)$, which is required by the methods in [37] and [64] for finding a point on the secondary branch.

Finally, in section 5.5, we describe some numerical experience with the methods.

5.2. Following Trajectories Through Turning Points

5.2.1. The Method of Riks, Menzel and Schwetlick

In this section we describe briefly the method due to Riks [66] and Menzel and Schwetlick [49] and in section 5.2.3 we describe our modification. In [49] the method was described as a means of extending the region of convergence of methods for the solution of nonlinear equations. It appears that such a method involves an unnecessary amount of work for that problem where the accurate determination of the solution trajectory is not required (see section 5.5 for comments on this). However the approach is effective when following a solution branch past a turning point. Earlier methods for

this problem, e.g. [6], [69], solved (5.1.1) by Newton's method for a sequence of values of λ, λ_i , $i = 1,2\ldots$, i.e. by the method described in (4.2.3). However, failure occurs when $\{(x(\lambda_i),\lambda_i)\}$ approaches a turning point. Once failure has occurred, the turning point can be passed by extrapolating over (x^*,λ^*) but the accuracy and efficiency of the method is impaired since $A(x,\lambda)$ is nearly singular close to (x^*,λ^*) . Anselone and Moore [3] suggested changing the scalar variable to overcome these difficulties but considered only particular cases. Recently Riks [66] and Menzel and Schwetlick [49] have employed an idea essentially due to Davis [21] and make a change of variable which is applicable generally.

For the remainder of this chapter we will frequently write

$$y = \begin{bmatrix} x \\ \lambda \end{bmatrix} ,$$

or, more conveniently, $y = (x,\lambda)$, and consider H as a mapping from $D \subset R^{n+1} \to R^n$. Then (5.1.1) becomes the under-determined system

(5.2.1) $$H(y) = 0 .$$

Define $y^* = (x^*,\lambda^*)$ and $B(y)$ by

$$B(y) = [A(y) \quad d(y)]$$

then, from (5.1.3), $\operatorname{rank}[B(y^*)] = n$. In fact, it follows from our assumptions that, for any y satisfying (5.2.1) in a neighbourhood of y^* ,

(5.2.2) $$\operatorname{rank}[B(y)] = n .$$

The technique described by Riks, Menzel and Schwetlick is to add, at each iteration, an auxiliary equation to (5.2.1). They chose a function $\beta(y)$, $\beta : D \subset R^{n+1} \to R$, such that the solution of

(5.2.3) $$g(y) = \begin{bmatrix} H(y) \\ \beta(y) \end{bmatrix} = 0$$

is well defined and is a required point on the solution branch.

Suppose $\hat{y}$ is a known solution of (5.2.1) and we wish to find a new point on the solution branch. We can define the branch in R^{n+1} by $y(s)$,

where s represents the arc length, and let $\hat{y} = y(\hat{s})$. Also, it is sufficient to restrict $\beta(y)$ to be a linear function of the form

$$\beta(y) = b^T(y-\hat{y}) - \sigma \ , \tag{5.2.4}$$

for some unit vector b and scalar σ . Denoting the derivative of $y(s)$ with respect to s by $\dot{y}(s)$, Riks, Menzel and Schwetlick make the choice

$$b = \dot{y}(\hat{s}) \ . \tag{5.2.5}$$

(Note that this notation differs from the use of $\cdot$ in other chapters.) We justify this choice of b in Theorem 5.2.1, but first we present some notation which will be useful in the remainder of this chapter.

First we note that, because s represents the arc length, $\dot{y}(s)$ is a unit vector tangent to the solution branch at $y(s)$ and is the unique solution of unit length (up to a choice of sign) of

$$B\big(y(s)\big)\dot{y}(s) = 0 \ . \tag{5.2.6}$$

We denote the Jacobian of $g(y)$ by $G(y)$ and define the $(n+1) \times n$ matrices P_j , $j = 1,2,\ldots,n+1$, by

$$P_{n+1} = \begin{bmatrix} I_n \\ 0 \end{bmatrix} , \quad P_j = P_{n+1} + \big(\tilde{e}_{n+1}-\tilde{e}_j\big)e_j^T \ , \tag{5.2.7}$$

where I_n is the $n \times n$ unit matrix with columns $e_1,\ldots,e_n$ and $\tilde{e}_1,\ldots,\tilde{e}_{n+1}$ are the columns of I_{n+1} . Also we denote the columns of $A(y)$ by $a_1(y),\ldots,a_n(y)$ and write $a_{n+1}(y) = d(y)$. It follows that

$$B(y) = \big[a_1(y)\ldots a_n(y) \quad d(y)\big] = [A(y) \quad d(y)] \ . \tag{5.2.8}$$

Finally, we note that we will frequently omit the argument y in each of the functions in (5.2.8) and write B, a_j, d and A in place of $B(y)$, $a_j(y)$, $d(y)$ and $A(y)$ respectively. It follows from (5.2.7) and (5.2.8) that

$$BP_j = \big[a_1 \quad a_2\ldots a_{j-1} \quad d \quad a_{j+1}\ldots a_n\big] \ ,$$

for $j = 1,2,\ldots,n$, and

$$BP_{n+1} = A \ .$$

These equations clarify our reasons for defining the matrices P_j , $j = 1,\ldots,n+1$. These matrices will be used to select out certain columns of B and these columns will be chosen to form a linearly independent set. Using the identity

$$P_j P_j^T + \tilde{e}_j \tilde{e}_j^T = I_{n+1} \ ,$$

$j = 1,2,\ldots,n+1$, it follows from (5.2.6) that, for any j ,

$$B\left[P_j P_j^T + \tilde{e}_j \tilde{e}_j^T\right]\dot{y}(s) = 0 \ .$$

Then, if we know an index r such that $B(y)P_r$ is nonsingular, we can write

$$\left(BP_r\right)P_r^T\dot{y}(s) = -B\tilde{e}_r\tilde{e}_r^T\dot{y}(s)$$

and since $B\tilde{e}_r = a_r$, we have

(5.2.9a) $$P_r^T\dot{y}(s) = -\left(BP_r\right)^{-1}a_r\alpha$$

and

(5.2.9b) $$\tilde{e}_r^T\dot{y}(s) = \alpha \ ,$$

for some α chosen to normalise $\dot{y}(s)$. With this notation we can present a theorem which indicates why the choice of b , given by (5.2.5), was made in [49] and [66]. A similar result is given by Riks [66], but the following theorem is more straightforward.

THEOREM 5.2.1. *Let* $G(y)$ *denote the Jacobian of* $g(y)$, *defined in* (5.2.3), *with* $\beta(y)$ *defined in* (5.2.4). *Then if* $\operatorname{rank}\left[B\left(y(s)\right)\right] = n$,

$$\operatorname{Det}\left(G\left(y(s)\right)\right) = \rho b^T\dot{y}(s)$$

where ρ *is nonzero and independent of* b .

Proof. Since $\operatorname{rank}\left[B\left(y(s)\right)\right] = n$, there is a k such that $B\left(y(s)\right)P_k$ is nonsingular. Also $G(y)$ is given by

$$G(y) = \begin{bmatrix} B(y) \\ b^T \end{bmatrix}$$

and so

$$G(y)\begin{bmatrix} P_k & \tilde{e}_k \end{bmatrix} = \begin{bmatrix} B \\ b^T \end{bmatrix}\begin{bmatrix} P_k & \tilde{e}_k \end{bmatrix} = \begin{bmatrix} BP_k & a_k \\ b^T P_k & b^T\tilde{e}_k \end{bmatrix} .$$

Hence

$$G(y)\begin{bmatrix} P_k & \tilde{e}_k \end{bmatrix} = \begin{bmatrix} BP_k & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} I & (BP_k)^{-1}a_k \\ b^T P_k & b^T\tilde{e}_k \end{bmatrix} ,$$

and using the identity

$$\text{Det}\begin{bmatrix} I & u \\ v^T & \gamma \end{bmatrix} = \gamma - v^T u , \tag{5.2.10}$$

we have

$$\text{Det}(G(y))\,\text{Det}\begin{bmatrix} P_k & \tilde{e}_k \end{bmatrix} = \text{Det}(BP_k)\left(b^T\tilde{e}_k - b^T P_k (BP_k)^{-1}a_k\right) . \tag{5.2.11}$$

Now, $\begin{bmatrix} P_k & \tilde{e}_k \end{bmatrix}$ is I_{n+1} with the kth and $(n+1)$st columns interchanged and so $\text{Det}\begin{bmatrix} P_k & \tilde{e}_k \end{bmatrix} = \xi$, where $\xi = 1$ if $k = n + 1$ and $\xi = -1$ otherwise. Moreover, it follows from (5.2.9a) that

$$P_k^T\dot{y}(s) = -(BP_k)^{-1}a_k\ \tilde{e}_k^T\dot{y}(s)$$

and so

$$b^T\tilde{e}_k - b^T P_k (BP_k)^{-1}a_k = b^T\begin{bmatrix} \tilde{e}_k + P_k P_k^T\dot{y}(s)\ \dfrac{1}{\tilde{e}_k^T\dot{y}(s)} \end{bmatrix} . \tag{5.2.12}$$

Note that $\tilde{e}_k^T\dot{y}(s) \neq 0$ for otherwise it would follow from (5.2.9) that $\dot{y}(s) = 0$ which cannot be the case. From (5.2.12) we have

$$b^T\tilde{e}_k - b^T P_k (BP_k)^{-1} a_k = b^T \left(\tilde{e}_k \tilde{e}_k^T + P_k P_k^T\right) \dot{y}(s) \frac{1}{\tilde{e}_k^T \dot{y}(s)}$$

$$= \frac{1}{\tilde{e}_k^T \dot{y}(s)} b^T \dot{y}(s) \ .$$

This equation, with (5.2.11), gives

$$\text{Det}\left(G(y(s))\right) = \frac{\xi \text{Det}\left(BP_k\right)}{\tilde{e}_k^T \dot{y}(s)} b^T \dot{y}(s) \tag{5.2.13}$$

as required. □

It now follows immediately that, to maximise $|\text{Det}(G(\hat{y}))|$, we must choose b to maximise $|b^T \dot{y}(\hat{s})|$. Since b is to be of unit length this leads to the choice made in (5.2.5). Thus, in some sense, this choice of b makes the equations in (5.2.3) as well conditioned as possible.

The first step in finding the next point on the trajectory is to find an initial estimate by calculating z, given by

$$z = \hat{y} + \sigma \dot{y}(\hat{s}) \ .$$

Then the new point is taken to be the solution of the system

$$\begin{bmatrix} H(y) \\ (y-\hat{y})^T \dot{y}(\hat{s}) \end{bmatrix} = \begin{bmatrix} 0 \\ \sigma \end{bmatrix} , \tag{5.2.14}$$

which is solved using Newton's method with z as starting guess. The whole process can now be repeated to follow the solution branch. That (5.2.14) has a well defined solution, for sufficiently small σ, follows from the nonsingularity of $G(\hat{y})$ and the Implicit Function Theorem. The basic idea of the method is expressed in Figure 5.1 for the scalar case.

5.2.2. Calculation of $\dot{y}(s)$

Neither of the papers [49] or [66] gave an indication of how they calculated $\dot{y}(\hat{s})$. One way is to note that, except at y^*, $\dot{y}(s)$ is given by

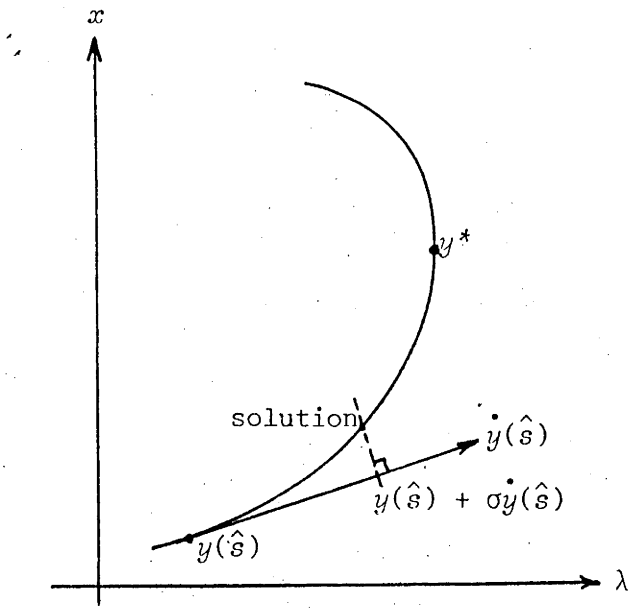

FIGURE 5.1: One step with $b = \dot{y}(\hat{s})$.

$$\dot{y}(s) = \begin{bmatrix} \dot{x}(s) \\ \dot{\lambda}(s) \end{bmatrix} = \begin{bmatrix} \frac{dx}{d\lambda}(\lambda) \\ 1 \end{bmatrix} \dot{\lambda}(s) ,$$

and we can use (5.1.2) for this calculation. However, close to a critical point, $A(y)$ is nearly singular and so it is better to use (5.2.9) for some r . Obviously the best r is that which gives the matrix BP_r which is, in some sense, the least singular and, to find this r , we use the following corollary of Theorem 5.2.1.

COROLLARY 5.2.1. *Suppose* $\operatorname{rank}[B(y(s))] = n$, *then for* $j = 1,2,\ldots,n+1$,

$$\operatorname{Det}(B(y(s))P_j) = \rho\tilde{e}_j^T\dot{y}(s) ,$$

where ρ *is non zero and, apart from a sign, is independent of* j .

Proof. Define G_j , $j = 1,\ldots,n+1$, by

(5.2.15) $$G_j(y) = \begin{bmatrix} B(y) \\ \tilde{e}_j^T \end{bmatrix} .$$

Then, if BP_j is nonsingular, substituting $b = \tilde{e}_j$ and $k = j$ in (5.2.13) shows that

(5.2.16) $$\mathrm{Det}\big(G_j(y)\big) = \pm\, \mathrm{Det}\big(BP_j\big) .$$

Also, trivially, (5.2.16) is true whenever BP_j is singular (expand $\mathrm{Det}\big(G_j(y)\big)$ about the nonzero element of its last row) and so (5.2.16) is true for $j = 1,\ldots,n+1$. It now follows from (5.2.13) that

$$\mathrm{Det}\big(G_j(y)\big) = \pm\, \mathrm{Det}\big(BP_j\big) = \pm\, \frac{\mathrm{Det}\big(BP_k\big)}{\tilde{e}_k^T \dot{y}(s)}\, \tilde{e}_j^T \dot{y}(s) ,$$

where k is such that BP_k is nonsingular. This gives the required result. □

Now we see that choosing r to be that j which maximises $\left|\tilde{e}_j^T \dot{y}(s)\right|$, $j = 1,2,\ldots,n+1$, will maximise $\left|\mathrm{Det}\big(BP_r\big)\right|$ and, in this sense, will give the least singular matrix for use in (5.2.9). Of course, to find this value for r we must already know $\dot{y}(s)$, however this choice of r is the most suitable for use in (5.2.9) at the next step and we assume that a prescribed value of r proves acceptable at the first step.

5.2.3. A New Method

The idea of this section is similar to methods used in [38] and [56] for problems in two dimensions. Equations (5.2.3) constitute $n + 1$ equations in $n + 1$ unknowns and whilst work can be saved by noting that one equation is linear, we prefer to reduce the number of variables in a direct way. If $\beta(y)$ is chosen as

$$\beta(y) = \tilde{e}_r^T (y - \hat{y}) - \sigma ,$$

for some r, σ , then (5.2.3) becomes

(5.2.17a) $$H(y) = 0$$

and

(5.2.17b) $$y_r = \hat{y}_r + \sigma \ ,$$

which, since y_r is specified, constitute n equations in n unknowns. The index r is chosen so that the determinant of $G(y)$ is as large as possible at $\hat{y}$. When $r = n + 1$ the method becomes the one of incrementing λ as described at the beginning of this section. However, close to a turning point some other element of y will be more suitable as the incremental variable. Since we have reduced the number of equations by one, the amount of work saved may be significant if n is small or if many points on the solution branch are required.

The Jacobian of the system at $\hat{y}$ is $B(\hat{y})P_r$. In fact, the Jacobian of the full system (5.2.17) is defined by $G_r(\hat{y})$ in (5.2.15), however y_r is known and in computations it is $B(\hat{y})P_r$ which is used. It is now obvious how to choose the index r . Again we wish to make $|\mathrm{Det}(B(\hat{y})P_r)|$ as large as possible and again, because of Corollary 5.2.1, we choose r to be that j which maximises $\left|\tilde{e}_j^T \dot{y}(\hat{s})\right|$, $j = 1,\ldots,n+1$. At this stage we know $\dot{y}(\hat{s})$ and the resulting value of r is the choice we make in (5.2.9) to calculate $\dot{y}(s)$ at the next step. We note that the angle, θ_j , between the solution branch at $\hat{y}$ and the jth coordinate direction satisfies

$$\mathrm{Cos}\theta_j = \tilde{e}_j^T \dot{y}(\hat{s}) \ .$$

(We have omitted a constant in this equation by assuming, for the moment, that $\dot{y}(\hat{s})$ has been normalised so that $\|\dot{y}(\hat{s})\|_2 = (\dot{y}(\hat{s})^T \dot{y}(\hat{s}))^{\frac{1}{2}} = 1$.) Thus our choice of r gives the variable, y_r , whose coordinate direction makes

the smallest angle with the solution branch. This is expressed in Figure 5.2 for the scalar case.

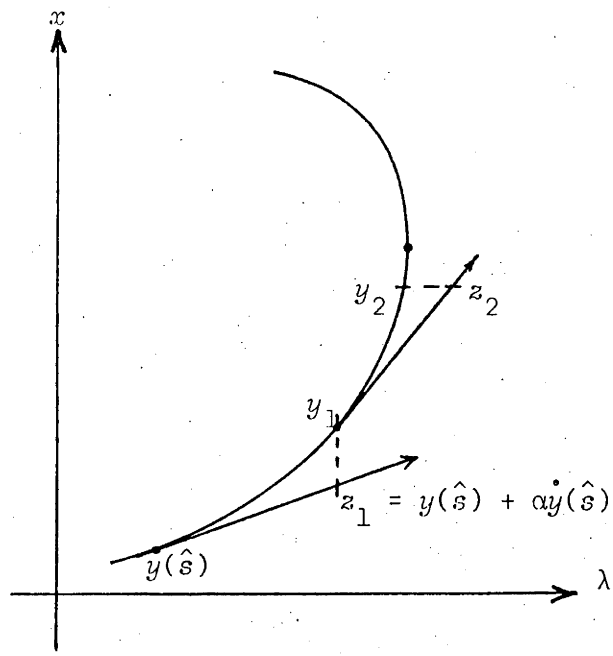

FIGURE 5.2: Two steps, with $b = e_1$ then $b = e_2$. (z_1, z_2 are initial estimates of y_1, y_2 .)

In practice the initial estimate of the solution of (5.2.17) is taken as the linear estimate, given by

(5.2.18) $$z = \hat{y} + \alpha\dot{y}(\hat{s}) \ ,$$

where $\alpha = \sigma/\tilde{e}_r^T\dot{y}(\hat{s})$. Then (5.2.17a) is solved by Newton's method. Apart from the initial estimate in (5.2.18), the resulting method is essentially the same as that described in (4.2.3) except that we periodically change the variable which is being incremented. Because of this, the discussion in section 4.2 regarding efficiency can equally be applied here. Thus our computational method would not evaluate $B(y)P_r$ at every iteration but only when necessary. If the value of $\mathrm{Det}\left(B(y)P_r\right)$ can be monitored easily then, when this becomes small, the Jacobian should be evaluated more frequently. For large sparse systems, where the determinant is not available, the number

of iterations required to solve (5.2.17a) serves as an indication of how effective the approximate Jacobian is. If the number of iterations increases, this suggests evaluating the Jacobian more frequently.

As a final remark we note that in a recent paper, Keller [36] made the choice $b = W\dot{y}(\hat{s})$ in place of (5.2.5), where $W = \text{diag}(\theta,\theta,\ldots,\theta,1-\theta)$, for some $\theta \in (0,1)$. Our reasons for modifying the choice in (5.2.5) apply equally well to Keller's choice.

5.3. The Determination of Turning Points

5.3.1. Introduction

Several methods, based upon interpolation, have been suggested for the accurate determination of a turning point, (x^*,λ^*), on a solution branch of (5.1.1). Notably Simpson [69] describes an iterative method which gives linear convergence to (x^*,λ^*) and which is suitable for problems with symmetric $A(x,\lambda)$. In this section we present some methods which, for less work per iteration, give second order convergence to (x^*,λ^*) and do not require $A(x,\lambda)$ to be symmetric.

We assume that a reasonable estimate, (x_0,λ_0), of (x^*,λ^*) is known as a consequence of following a solution branch using a method from section 5.2. In many problems the value of $\Gamma(\lambda) = \text{Det}\big(A\big(x(\lambda),\lambda\big)\big)$ determines whether or not the system is stable and, as $\Gamma(\lambda)$ changes sign, the branch passes through a turning point in or out of a region of stability. When $\Gamma(\lambda)$ can be easily evaluated it can be monitored to specify when two iterates straddle a turning point. But better than evaluating $\Gamma(\lambda)$ is to use the following theorem. We continue with the notation of the previous section.

THEOREM 5.3.1. *With* $B(x,\lambda) = [A(x,\lambda) \;\; d(x,\lambda)]$, *suppose for some* $r \le n$ *that* $B(x,\lambda)P_r$ *is nonsingular, where* P_r *is defined in* (5.2.7). *Then*

(5.3.1) $$\mathrm{Det}\big(A(x,\lambda)\big) = \mathrm{Det}\big(B(x,\lambda)P_r\big)\gamma(\lambda) ,$$

where $\gamma(\lambda)$ *is given by*

$$\gamma(\lambda) = e_r^T\big[B(x,\lambda)P_r\big]^{-1}a_r(x,\lambda) .$$

Proof. From the definition of P_{n+1} in (5.2.7), we have

$$A(x,\lambda) = B(x,\lambda)P_{n+1} .$$

Now, omitting the variables (x,λ) as arguments, we have from (5.2.7),

$$A = B\Big[P_r+\big(\tilde{e}_r-\tilde{e}_{n+1}\big)e_r^T\Big]$$

$$= BP_r + B\big(\tilde{e}_r-\tilde{e}_{n+1}\big)e_r^T .$$

Thus, since BP_r is nonsingular,

$$A = BP_r\Big[I+\big(BP_r\big)^{-1}B\big(\tilde{e}_r-\tilde{e}_{n+1}\big)e_r^T\Big] .$$

Using the identity,

(5.3.2) $$\mathrm{Det}\big(I+ab^T\big) = 1 + b^Ta$$

and noting from (5.2.8) that $B\tilde{e}_r = a_r$ and $B\tilde{e}_{n+1} = d$, we have

$$\mathrm{Det}(A) = \mathrm{Det}\big(BP_r\big)\Big(1+e_r^T\big(BP_r\big)^{-1}\big(a_r-d\big)\Big) .$$

But $\big(BP_r\big)^{-1}d = e_r$, since d is the rth column of BP_r , and so

$$\mathrm{Det}(A) = \mathrm{Det}\big(BP_r\big)e_r^T\big(BP_r\big)^{-1}a_r$$

as required. □

Evaluating (5.3.1) at $\big(x(s),\lambda(s)\big)$ gives

$$\Gamma\big(\lambda(s)\big) = \mathrm{Det}\big(B\big(x(s),\lambda(s)\big)P_r\big)\gamma\big(\lambda(s)\big)$$

and, assuming $\mathrm{Det}\big(BP_r\big) \neq 0$ in the neighbourhood of (x^*,λ^*) , $\Gamma\big(\lambda(s)\big)$ changes sign with $\gamma\big(\lambda(s)\big)$. But it is already necessary to calculate $\gamma\big(\lambda(s)\big)$, in (5.2.9) as part of the evaluation of $\dot{y}(s)$, and so the sign of $\Gamma(\lambda)$ can be monitored without extra work. We note that, since

$e_r^T P_r^T = \tilde{e}_{n+1}^T$, it follows from (5.2.9) that

$$\gamma\big(\lambda(s)\big) = \dot{\lambda}(s)/\dot{x}_r(s)$$

and, since $\dot{x}_r(s)$ $\left(= \tilde{e}_r^T \dot{y}(s)\right)$ is nonzero by choice of r , $\gamma\big(\lambda(s)\big)$ changing sign simply means that, with respect to the λ axis, the trajectory has changed direction, i.e. it has passed through a turning point.

To find the turning point we set up a system of equations which, in the region of interest, have a unique solution (x^*,λ^*) . These are of the form

(5.3.3a) $$H(x,\lambda) = 0$$

and

(5.3.3b) $$\phi(x,\lambda) = 0 \ ,$$

where $\phi : D \subset R^n \times R \to R^n$ is chosen so that

(5.3.3c) $\phi(x,\lambda) = 0$ iff $A(x,\lambda)$ is singular.

In section 5.3.3 we give some choices of $\phi(x,\lambda)$ which have proved successful in practice but are expensive to evaluate. For this reason we describe, in section 5.3.2, a method suitable for this case.

5.3.2. A Newton Like Method

In this section we describe a method which we will use for solving (5.3.3). Since it may be of interest in other cases, we describe it in some generality and apply it to (5.3.3) in the next section. We consider the general problem of solving the nonlinear equations

(5.3.4a) $$q(z,\mu) = 0$$

and

(5.3.4b) $$\psi(z,\mu) = 0 \ ,$$

$q : D \subset R^n \times R \to R^n$, $\psi : D \subset R^n \times R \to R$, where

(5.3.5) $$\partial_z q(z,\mu) = Q(z,\mu)$$

is nonsingular in the region of a solution (z^*,μ^*) of (5.3.4). We assume

that derivatives of $\psi(z,\mu)$ are not available and that $\psi(z,\mu)$ is expensive to evaluate. The method we describe is similar to those of Brown [13] and Brent [12] but is more suitable when $Q(z,\mu)$ is available analytically and when $Q(z,\mu)$ is large and sparse or easy to evaluate. We note that, for small problems, we have used Brent's method with success. (See [51] for an implementation and also [12] for Brent's comments on the suitability of his method for problems where the Jacobian is sparse.)

Suppose (z_i,μ_i) is an approximation to (z^*,μ^*), then we linearise (5.3.4a) about (z_i,μ_i) and define the subspace L_i to be the space where this linearisation is zero. That is, L_i is the set of points (z,μ) such that

$$q(z_i,\mu_i) + \left[Q(z_i,\mu_i) \quad u(z_i,\mu_i)\right]\begin{bmatrix} z-z_i \\ \mu-\mu_i \end{bmatrix} = 0 ,$$

where $u(z,\mu) = \partial_\mu q(z,\mu)$. Now, omitting the arguments (z_i,μ_i) and writing $q(z_i,\mu_i) = q_i$ etc., and assuming Q_i is nonsingular, L_i is defined by

$$L_i = \left\{(z,\mu) \mid z = \hat{z}_{i+1} - Q_i^{-1}u_i(\mu-\mu_i)\right\}$$

where

$$\hat{z}_{i+1} = z_i - Q_i^{-1}q_i . \tag{5.3.6}$$

Now we define $\Psi_i : D_i \subset R \to R$ as ψ, restricted to L_i, by

$$\Psi_i(\mu) = \psi\left(\hat{z}_{i+1} - Q_i^{-1}u_i(\mu-\mu_i),\mu\right) , \tag{5.3.7}$$

where $D_i = \left\{\mu \mid \left(\hat{z}_{i+1} - Q_i^{-1}u_i(\mu-\mu_i),\mu\right) \in D\right\}$. Then we can attempt to find a zero of $\psi(\mu)$ on L_i by linearising Ψ_i and applying a Newton step. Since we cannot evaluate $\frac{d\Psi_i}{d\mu}(\mu_i)$, we approximate it by

$$\frac{d\Psi_i}{d\mu}(\mu_i) \simeq \frac{\Psi_i(\mu_i+\delta_i)-\Psi_i(\mu_i)}{\delta_i} = \Delta_i \ , \tag{5.3.8}$$

for some $\delta_i \neq 0$, and generate the step

$$\mu_{i+1} = \mu_i - \frac{\Psi_i(\mu_i)}{\Delta_i} \ . \tag{5.3.9}$$

Then z_{i+1} is given by

$$z_{i+1} = \hat{z}_{i+1} - Q_i^{-1} u_i(\mu_{i+1}-\mu_i) \ . \tag{5.3.10}$$

The following theorem, which is proved in the appendix to Chapter 5, gives sufficient conditions for the sequence $\{(z_i,\mu_i)\}$ generated by (5.3.6) - (5.3.10) to converge to (z^*,μ^*) with R-order ≥ 2 . For the sake of continuity we prefer to postpone the proof since here we are primarily interested in the application of the method. The important feature of the method is that we can attain rapid convergence to (z^*,μ^*) with only two evaluations of $\psi(z,\mu)$ per iteration. Since we are assuming that the evaluation of ψ is the most expensive part of the process, this represents a considerable saving over standard methods for solving (5.3.4).

THEOREM 5.3.2. *Suppose* $q : D \subset R^n \times R \to R^n$ *and* $\psi : D \subset R^n \times R \to R$ *are Frechet differentiable on* D *and their derivatives satisfy a Lipschitz condition on an open neighbourhood* S *of the point* (z^*,μ^*) *which is a solution of* (5.3.4). *Suppose also that* $Q(z,\mu)$, *defined in* (5.3.5), *has a bounded inverse in* S *and that the inverse of*

$$R(z,\mu) = \begin{bmatrix} Q(z,\mu) & u(z,\mu) \\ \partial_z\psi(z,\mu)^T & \partial_\mu\psi(z,\mu) \end{bmatrix} \tag{5.3.11}$$

exists and is bounded on S , *where* $u(z,\mu) = \partial_\mu q(z,\mu)$. *Then there exists an* $\varepsilon > 0$ *such that, if*

$$\left\| \begin{matrix} z_0 - z^* \\ \mu_0 - \mu^* \end{matrix} \right\| \leq \varepsilon \ ,$$

the sequence $\{(z_i,\mu_i)\}$ *defined by* (5.3.6)-(5.3.10), *where* δ_i *is chosen as*

(5.3.12a) $$\delta_i = \begin{cases} |\psi(z_i,\mu_i)| \Big/ \left(1+\left\|Q(z_i,\mu_i)^{-1}u(z_i,\mu_i)\right\|\right) \ , & \text{if } \psi(z_i,\mu_i) \neq 0 \ , \\ \\ \text{(5.3.12b)} \quad \textit{sufficiently small otherwise,} \end{cases}$$

converges to (z^*,μ^*) *with R-order* ≥ 2 .

(Note that, in practice, to ensure that $\delta_i \neq 0$ we can choose $\delta_i = \tau$, where a stopping criterion for the iteration is $\left\| \begin{matrix} z_{i+1}-z_i \\ \mu_{i+1}-\mu_i \end{matrix} \right\| < \tau$, if (5.3.12a) gives a value less than τ .)

5.3.3. Solution of equation (5.3.3)

The equations we wish to solve are given in (5.3.3) and, to apply the method of section 5.3.2, we must put them into a form which satisfies the conditions of Theorem 5.3.2. To do this we note that, from (5.2.2), rank$[B(x,\lambda)] = n$ in the region of a turning point, where

(5.3.13) $$B(x,\lambda) = [A(x,\lambda) \quad d(x,\lambda)] \ .$$

Thus, $B(x,\lambda)$ has n linearly independent columns and we can choose an index r such that BP_r is nonsingular. We see below that the best choice of r is that which was chosen in section 5.2. Now we define (z,μ) and (z^*,μ^*) by

$$(5.3.14)\quad z = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{r-1} \\ \lambda \\ x_{r+1} \\ \vdots \\ x_n \end{bmatrix} = P_r^T \begin{bmatrix} x \\ \lambda \end{bmatrix}, \quad \mu = x_r \; ; \quad z^* = P_r^T \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix}, \quad \mu^* = x_r^* \; ,$$

and $q(z,\mu)$ by

$$q(z,\mu) = H(x,\lambda) \; .$$

Then $Q(z,\mu) = B(x,\lambda)P_r$, which is nonsingular in a neighbourhood of (z^*,μ^*) , as required in Theorem 5.3.2. Also $u(z,\mu) = \partial_\mu q(z,\mu)$, which equals $\partial_{x_r} H(x,\lambda)$, so $u(z,\mu) = a_r(x,\lambda)$. Since the method of section 5.3.2 requires the solution of two linear systems of equations with $Q(z,\mu)$ as coefficient matrix, we wish to choose r so that $Q(z,\mu)$ is, in some sense, the least singular choice. Thus, we use the value of r chosen in section 5.2 when following the trajectory through the turning point, and it follows from Corollary 5.2.1 that this choice maximises $\mathrm{Det}(Q)$ over all the possible choices of r .

Next we define $\psi(z,\mu)$ by

$$\psi(z,\mu) = \phi(x,\lambda) \; ,$$

where we will choose $\phi(x,\lambda)$ to satisfy (5.3.3c) and we will also require $\phi(x,\lambda)$ to have a Lipschitz continuous derivative in a neighbourhood of (x^*,λ^*) .

Finally, we require conditions that $R(z,\mu)$ in (5.3.11) has a bounded inverse in a neighbourhood of (z^*,μ^*) . Now $R(z,\mu)$ satisfies

$$R(z,\mu) = T(x,\lambda)\begin{bmatrix} P_r & \tilde{e}_r \end{bmatrix}$$

where

$$T(x,\lambda) = \begin{bmatrix} A(x,\lambda) & d(x,\lambda) \\ \partial_x \phi(x,\lambda)^T & \partial_\lambda \phi(x,\lambda) \end{bmatrix} .$$

It is then obvious that $R(z^*,\mu^*)$ is nonsingular if and only if $T(x^*,\lambda^*)$ is nonsingular. Also, by continuity of the derivatives of $H(x,\lambda)$ and $\phi(x,\lambda)$, this will imply $R(z,\mu)$ has a bounded inverse in a neighbourhood of (z^*,μ^*).

It follows from (5.2.6) that $T(x^*,\lambda^*)$ is singular if and only if

$$\begin{bmatrix} \partial_x \phi(x^*,\lambda^*)^T & \partial_\lambda \phi(x^*,\lambda^*) \end{bmatrix} \dot{y}(s^*) = 0 , \tag{5.3.15}$$

where $y(s^*) = (x^*,\lambda^*)$. If we define Φ to be ϕ restricted to the solution branch and write $\Phi(s) = \phi\big(x(s),\lambda(s)\big)$, then, by the chain rule

$$\dot{\Phi}(s) = \begin{bmatrix} \partial_x \phi(x,\lambda)^T & \partial_\lambda \phi(x,\lambda) \end{bmatrix} \dot{y}(s) .$$

It follows that $T(x^*,\lambda^*)$ is singular if and only if $\dot{\Phi}(s^*) = 0$. Furthermore, by choice of ϕ, $\Phi(s^*) = 0$ and so $T(x^*,\lambda^*)$ is singular if and only if $\Phi(s)$ has a double root at s^*. Below we will see that one choice of $\phi(x,\lambda)$ is given by $\phi(x,\lambda) = \mathrm{Det}\big(A(x,\lambda)\big)$ and, in this case, $\Phi(s) = \Gamma\big(\lambda(s)\big)$. It follows from the discussion following Theorem 5.3.1 that

$$\Phi(s) = \alpha(s)\dot{\lambda}(s)$$

where $\alpha(s)$ is nonzero in a neighbourhood of s^*. At the turning point $\dot{\lambda}(s^*) = 0$ and so

$$\dot{\Phi}(s^*) = \alpha(s^*)\ddot{\lambda}(s^*) ,$$

which implies that $\dot{\Phi}(s^*) = 0$ if and only if $\ddot{\lambda}(s^*) = 0$. This is the condition that (x^*,λ^*) is a point of inflexion, or a turning point at which $\dot{\lambda}(s)$ has a multiple zero. We can show that all the choices of $\phi(x,\lambda)$ discussed in the next section are of the form

$$\phi(x,\lambda) = \mathrm{Det}\big(A(x,\lambda)\big)\xi(x,\lambda)$$

for some function $\xi(x,\lambda)$ which is nonzero in a neighbourhood of (x^*,λ^*) and so the same argument applies to each choice. It follows that $R(z^*,\mu^*)$

will be singular if and only if $\dot{\lambda}(s)$ has a multiple zero at s^* and, in this case, the R-order of convergence of the method will be only one. Geometrically (5.3.15) implies that $R(z^*,\mu^*)$ is singular if and only if the solution branch at (x^*,λ^*) is tangential to the surface S on which $A(x,\lambda)$ is singular. This follows because $\left[\partial_x\phi(x^*,\lambda^*)^T \quad \partial_\lambda\phi(x^*,\lambda^*)\right]$ is normal to S at (x^*,λ^*).

5.3.4. Choices for $\phi(x,\lambda)$

Now we consider some specific choices for $\phi(x,\lambda)$. From (5.3.3), the obvious choice is

$$\phi_1(x,\lambda) = \mathrm{Det}\big(A(x,\lambda)\big) .$$

This choice proved acceptable except in two cases. When $A(x,\lambda)$ is large and sparse, the evaluation of $\phi_1(x,\lambda)$ may be inconvenient since it requires the factorisation of $A(x,\lambda)$ into matrices which are not necessarily sparse. Secondly, if $\mathrm{Det}\big(A(x,\lambda)\big)$ is very small compared with $\|H(x,\lambda)\|$, then loss of significance occurs in the evaluation of $\Psi_i(z,\mu)$, and therefore of Δ_i, in (5.3.8), which adversely affects the convergence rate of the method. Also, in severe problems, underflow may occur. Despite these difficulties, this choice proved successful for several small problems, but we discuss two further choices which do not suffer the same disadvantages.

Define $\phi_2(x,\lambda)$ by

$$\phi_2(x,\lambda) = e_r^T\big(B(x,\lambda)P_r\big)^{-1}a_r(x,\lambda) ,$$

where r is the index described earlier. Since $B(x,\lambda)P_r$ is nonsingular in a neighbourhood of (x^*,λ^*), it follows from Theorem 5.3.1 that $\phi_2(x,\lambda) = 0$ if and only if $A(x,\lambda)$ is singular. Thus $\phi_2(x,\lambda)$ is a suitable choice. Its evaluation requires the solution of a system of linear

equations and so is suitable in the case when $B(x,\lambda)$ is sparse. Finally, it is straightforward to show that $q(z,\mu)$ and $\psi(z,\mu)$, where $\psi(z,\mu) = \phi_2(x,\lambda)$, satisfy the continuity conditions required in Theorem 5.3.2 if $H(x,\lambda)$ is twice Frechet differentiable on D and its second derivative satisfies a Lipschitz condition in a neighbourhood of (x^*,λ^*).

Our final choice for $\phi(x,\lambda)$ is given by defining $\phi_3(x,\lambda)$ by the relation

$$A(x,\lambda)v(x,\lambda) = \phi_3(x,\lambda)w \tag{5.3.16a}$$

and

$$c^T v(x,\lambda) = 1 , \tag{5.3.16b}$$

for some fixed c and w such that $\|c\| = \|w\| = 1$. This choice is an extension of the method of Osborne and Michaelson [55], [57] for the nonlinear eigenvalue problem in one variable. We describe the details of the method as they affect our problem and refer the reader to [55] and [57] for further details.

Firstly we show that $\phi_3(x,\lambda)$ is well defined for certain choices of w and c.

THEOREM 5.3.3. *Suppose $A(x,\lambda)$ is continuous in an open neighbourhood of (x^*,λ^*) and $\phi_3(x,\lambda)$ is defined by* (5.3.16). *Then $\phi_3(x,\lambda)$ is well defined and continuous in an open neighbourhood of (x^*,λ^*) if*

$$\operatorname{Det}\begin{bmatrix} A(x^*,\lambda^*) & -w \\ c^T & 0 \end{bmatrix} \neq 0 . \tag{5.3.17}$$

Proof. $v(x,\lambda)$ and $\phi_3(x,\lambda)$ are defined by

$$\begin{bmatrix} A(x,\lambda) & -w \\ c^T & 0 \end{bmatrix}\begin{bmatrix} v(x,\lambda) \\ \phi_3(x,\lambda) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} . \tag{5.3.18}$$

From (5.3.17) and the continuity of $A(x,\lambda)$, (5.3.18) has a unique continuous solution in an open neighbourhood of (x^*,λ^*). □

So we must choose w and c to guarantee that (5.3.17) is true. Since $\text{rank}[A(x^*,\lambda^*)] = n - 1$ there exists a nonsingular matrix V and an $n \times n$ matrix Λ of the form

$$\Lambda = \begin{bmatrix} 0 & 0 \\ 0 & \Omega \end{bmatrix},$$

where Ω is $(n\text{-}1) \times (n\text{-}1)$ and nonsingular, such that

$$A(x^*,\lambda^*) = V\Lambda V^{-1} .$$

It is simple to show that

$$\left| \text{Det} \begin{bmatrix} A(x^*,\lambda^*) & -w \\ c^T & 0 \end{bmatrix} \right| = \left| e_1^T V^{-1} w \quad c^T V e_1 \quad \text{Det}(\Omega) \right|$$

and that $e_1^T V^{-1}$ and Ve_1 are left and right null vectors of $A(x^*,\lambda^*)$. Let $u_r = Ve_1$ and $u_l^T = e_1^T V^{-1}$, then a sufficient condition for (5.3.17) to be true is that $u_l^T w \neq 0$ and $c^T u_r \neq 0$. It follows that, in some sense, the best choice of w and c is, with suitable scaling,

$$w = u_l , \quad c = u_r ,$$

for this choice maximises the determinant in (5.3.17). However, to generate an approximation to u_l requires extra work and so we choose w as an approximation to u_r. Such an approximation is readily available in the course of computation. We note that, if $A(x^*,\lambda^*)$ is symmetric, $u_r = u_l$ and the choice of w is best in this case. In the general case, $u_r^T u_l = 1$ and the choice of w ensures that $u_l^T w \neq 0$, at least if w is a good approximation to u_r. It is convenient to choose $c = e_k$, for some k chosen so that $e_k^T u_r \neq 0$. In practice, if w is a reasonable approximation to u_r, the choice of k which maximises $\left| e_j^T w \right|$, $j = 1,\ldots,n$ is

suitable. Also, it is most efficient to change w, and therefore c, at each iteration, always using the best estimate of u_r for w. $A(x_0,\lambda_0)^{-1}d(x_0,\lambda_0)$, which is the first n components of $\dot{y}$ at (x_0,λ_0) and so has already been calculated, is a good initial choice for w when suitably scaled. Finally, we note that the differentiability requirements on $q(z,\mu)$ and $\psi(z,\mu)$ in Theorem 5.3.2 follow if $H(x,\lambda)$ is twice Frechet differentiable on D and its second derivative satisfies a Lipschitz condition in a neighbourhood of (x^*,λ^*).

To complete our description of the method for this choice of $\phi(x,\lambda)$, we define the subspace L_i, as in section 5.3.2, and the matrix $M_i(\mu)$ by

$$M_i(\mu) = A\left(\hat{z}_{i+1}-Q_i^{-1}u_i(\mu-\mu_i),\mu\right)$$

where, for brevity, we are considering A to be a function of z and μ. Then if w_i is our current estimate of u_r and e_k is our current choice of c, then $\Psi_i(\mu_i)$ is given by

$$M_i(\mu_i)v_{i+1} = \Psi_i(\mu_i)w_i$$

and

$$e_k^T v_{i+1} = 1 . \tag{5.3.19}$$

v_{i+1} is found by solving

$$M_i(\mu_i)v = w_i \tag{5.3.20}$$

and scaling the solution to satisfy (5.3.19) and also

$$\Psi_i(\mu_i) = 1/e_k^T v .$$

This represents one step of inverse iteration and so v_{i+1} will be richer in u_r then w_i. Thus v_{i+1} is a better choice for w_{i+1} than w_i. It is shown by Osborne [56] that another efficient choice of w_{i+1} is v'_{i+1}, given by

$$v'_{i+1} = \frac{dM_i}{d\mu}(\mu_i)v_{i+1} .$$

We do not have the derivative of $M_i(\mu)$, but, in the estimation of $\frac{d\Psi_i}{d\mu}(\mu)$ on L_i, we also calculate $M_i(\mu_i+\delta_i)$ and so we can improve v_{i+1} by forming

$$v'_{i+1} = [M_i(\mu_i)-M_i(\mu_i+\delta_i)]v_{i+1} . \tag{5.3.21}$$

Then we set w_{i+1} to be v_{i+1} or v'_{i+1} and scale it suitably. It is important to note that, as the process converges, $\{M_i(\mu_i)\}$ approaches $A(x^*,\lambda^*)$, however, in the same way as inverse iteration, no difficulties arise when solving (5.3.20) due to $M_i(\mu_i)$ being nearly singular. All that is necessary is that care be taken in solving (5.3.20) so that the solution remains within machine bounds.

We conclude this section with two remarks.

REMARK 5.3.1. For each choice of ϕ, an iteration requires the solution of four linear systems and gives second order convergence to the turning point. This compares favourably with the method described by Simpson [69]. Also we note that with ϕ_2 and ϕ_3 the work in solving these systems can be reduced as follows. If a direct method is to be employed for solving the linear equations then, when calculating $Q_i^{-1}u_i$ and $Q_i^{-1}q_i$ in (5.3.6) and (5.3.7), it is only necessary to decompose Q_i into its appropriate factors once. This saving cannot be made if an iterative method is being used to solve the linear systems. In this case, however, the calculation of $\Psi_i(\mu_i+\delta_i)$ and $\Psi_i(\mu_i)$ each require the solution of a linear system. Moreover, the solution of the first will provide an excellent estimate of the solution of the second. The result is that few iterations will be required for the second system.

REMARK 5.3.2. The method of Osborne and Michaelson is just one of a class of methods for the nonlinear eigenvalue problem which could be applied to this problem. Some of these are discussed in [67].

5.4. The Determination of Certain Simple Bifurcation Points

We point out, in this section, that the method of section 5.3 can sometimes be applied to finding simple bifurcation points. To find a point (x_B,λ_B) defined in (5.1.4) we can solve

(5.4.1a) $$H(x,\lambda) = 0 ,$$

and

(5.4.1b) $$\phi(x,\lambda) = 0 ,$$

with $\phi(x,\lambda)$ given by ϕ_1 or ϕ_3 from section 5.3. In this case, however, the resulting Jacobian is singular at the solution and so the method converges only linearly. However, it is often the case that, on a primary branch, we have independent information about the solution curve $x(\lambda)$. For example, in the problems discussed in section 5.5, noting the symmetry gives the required information. If x, on the solution branch, also satisfies

$$g(x,\lambda) = 0 ,$$

$g : D \subset R^n \times R \to R^m$, $m < n$, then it may be possible to replace certain components of H by components of g in such a way that the resulting system has full rank at (x_B,λ_B). In the case when $A(x,\lambda)$ is factorised, we can first apply the method to (5.4.1) and then convergence to (x_B,λ_B) is linear. In solving systems of the form

$$A(x_i,\lambda_i)v = b ,$$

where $A(x_i,\lambda_i)$ replaces Q_i in section 5.3.2, we factorise $A(x_i,\lambda_i)$ into

$$PA(x_i,\lambda_i) = LU$$

where P is a permutation matrix and U is upper triangular and L is unit lower triangular. We extend the decomposition to form

$$\begin{bmatrix} PA(x_i,\lambda_i) \\ G(x_i,\lambda_i) \end{bmatrix} = \begin{bmatrix} L \\ W \end{bmatrix} [U] ,$$

where $G(x,\lambda) = \partial_x g(x,\lambda)$. If, at some stage, the best choice of pivot in the decomposition of A , from the kth row say, becomes small compared with the elements of A , we replace that row of A by a row of $G(x,\lambda)$, the jth say, which maximises the pivot. We then continue with a new system, in which $H_k(x,\lambda)$ in (5.4.1a) is replaced by $g_j(x,\lambda)$. This new system satisfies the conditions of Theorem 5.3.2 and so we can attain rapid convergence to (x_B,λ_B) .

It is particularly convenient to use $\phi(x,\lambda) = \phi_3(x,\lambda)$ from section 5.3.4 since, on converging to (x_B,λ_B) , the final value of w_i gives a good approximation to the zero eigenvector of $A(x_B,\lambda_B)$ which is useful when looking for a point on the secondary branch. (See [37], [64] for further details of methods for this problem.)

5.5. Numerical Results

We have applied the methods of sections 5.2, 5.3 and 5.4 to several problems with success and we describe two which have appeared in the literature. The trussed dome problem [33], which was also considered in [63], is a physical example of stability loss. The dome of Figure 5.3, if subjected to vertical forces at nodes $1,2,\ldots,7$, deforms until it loses stability at a turning point. The equations defining the equilibrium positions of the structure are of the form

$$W(x)x = \lambda w ,$$

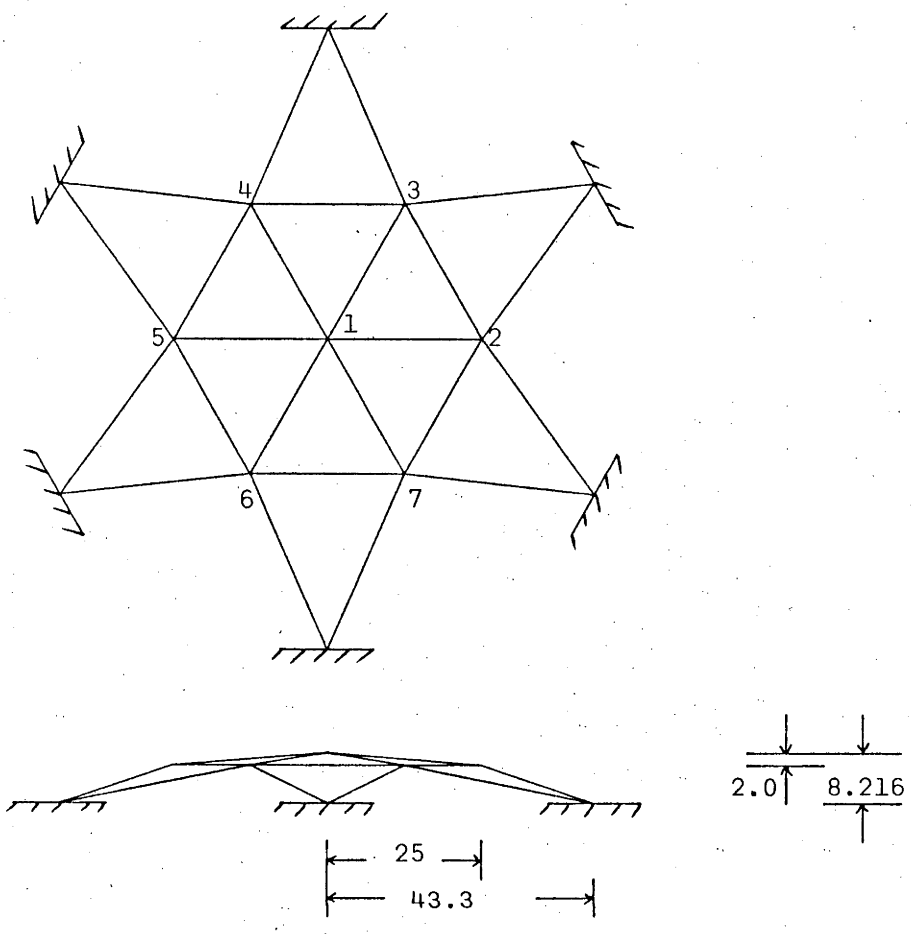

FIGURE 5.3: Geometry of Trussed Dome (from [33]).

where $W(x)$ is a matrix and w is a constant vector, when the force at node i is $\lambda\beta_i$ for fixed β_i, $i = 1,\ldots,7$. The vector x defines the position of the seven nodes and so the dimension of the problem is 21. The details and the derivation of these equations, together with a Fortran subroutine for the relevant calculations were provided by Professor W.C. Rheinboldt [62]. For the case where $\beta_1 = 10^{-4}$ and $\beta_j = 2 \times 10^{-4}$, $j = 2,\ldots,7$, Figure 5.4 shows the displacement, ξ, of the central node for varying λ and the turning point was found to be at

$$\lambda^* = 9.074147\ldots ,$$

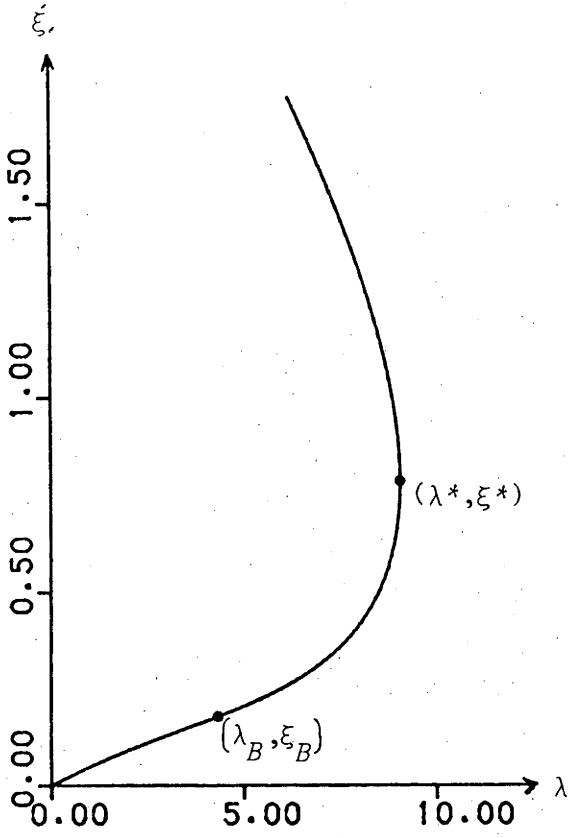

FIGURE 5.4: Vertical displacement of central node (ξ) vs. λ .

when for example, $\xi^* = 0.7865549\ldots$. With the choices of $\phi = \phi_2$ and ϕ_3 the algorithm displayed second order convergence to (x^*,λ^*) . The choice of $\phi(x,\lambda) = \mathrm{Det}\big(A(x,\lambda)\big)$ suffered from the loss of significance described in section 5.3.4. Typical values of the relevant functions in the region of (x^*,λ^*) were

$$\|H(x,\lambda)\| = 10^{-5}\ , \quad |\phi_1(x,\lambda)| = 10^{-37}\ , \quad |\phi_2(x,\lambda)| = 10^{-1}\ , \quad |\phi_3(x,\lambda)| = 10^{-4}$$

and so the choice of $\phi_1(x,\lambda)$ was less effective than the other choices.

The second problem was described by Simpson [69] and is the solution of the boundary value problem

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = \lambda e^u \ , \qquad (x,y) \in D \ ,$$

$$u(x,y) = 0 \ , \qquad (x,y) \in \partial D \ ,$$

where D is the unit square. The problem was discretised using the 9-point box form of the Laplacian (see Fox [26]) on a uniform mesh of size h. The resulting system is of the form (5.1.1) where λ appears non-linearly. If $m = 1/h$, the problem is of dimension m^2 and is sparse, so we used the iterative method of Paige and Saunders [59] to solve the linear systems. We used the choices $\phi_2(x,\lambda)$ and $\phi_3(x,\lambda)$ and both were successful. Figure 5.5 shows how $u(0.5,0.5)$ varies with λ (calculated with $h = 1/12$). We calculated the turning point on mesh sizes $h = 1/16$

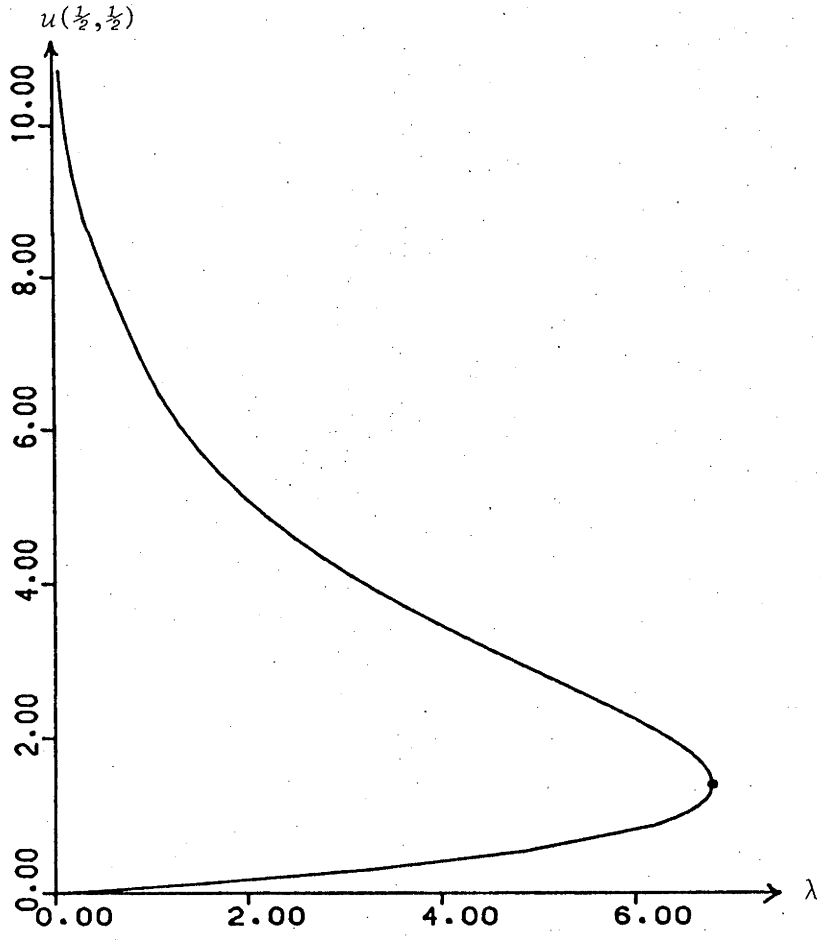

FIGURE 5.5: $u(\frac{1}{2},\frac{1}{2})$ vs. λ .

and $h = 1/24$ and derived the results, for $h = 1/16$

$$\lambda^* = 6.8080865\ldots\ , \qquad u(0.5,0.5) = 1.3916567\ldots$$

and for $h = 1/24$

$$\lambda^* = 6.80811698\ldots\ , \quad u(0.5,0.5) = 1.3916603\ldots\ ,$$

with convergence, in each case, being attained to more than the figures shown. These results for λ^* should be more accurate than those given by Simpson.

Typically, the number of iterations were the same for $\phi_2(x,\lambda)$ and $\phi_3(x,\lambda)$ with the correction (5.3.21). Without this correction, on average, using $\phi_3(x,\lambda)$ cost about one extra iteration. But in all cases the second order convergence to the turning point was apparent.

The method of section 5.4 was applied to finding the simple bifurcation point which occurs in the trussed dome problem. The value of $\mathrm{Det}\big(A(x,\lambda)\big)$ was monitored along $\big(x(\lambda),\lambda\big)$ to bracket $\big(x_B,\lambda_B\big)$ and then the method of section 5.4 was applied with $\phi(x,\lambda)$ given by $\phi_1(x,\lambda)$ and $\phi_3(x,\lambda)$. The extra information, which is satisfied only on the primary branch, was provided by several of the obvious symmetry relations satisfied by the dome. The methods were again successful and, on replacing a component of $H(x,t)$ by an appropriate symmetry relation, the convergence to $\big(x_B,\lambda_B\big)$ was second order. The bifurcation point was found to be at

$$\lambda_B = 4.341092788\ldots$$

where, for example, $\xi_B = 0.1796179807\ldots$. Note that when using $\phi_3(x,\lambda)$, the initial choice of $w_0 = A\big(x_0,\lambda_0\big)^{-1}d\big(x_0,\lambda_0\big)$ is not suitable since, as in this example, w_0 may have a very small component in the direction of the appropriate eigenvector. For the bifurcation point problem we have found choosing $w_0^T = (1,1,\ldots,1)$ is acceptable.

APPENDIX TO CHAPTER 5

We now prove Theorem 5.3.2. We use the same notation as described in section 5.3 and so refrain from restating the theorem.

Proof of Theorem 5.3.2. Throughout this proof we define the norm on $R^n \times R$ in terms of a norm on R^n as

$$\left\|\begin{matrix} a \\ \alpha \end{matrix}\right\| = \|a\| + |\alpha| \tag{A5.1}$$

for any $(a,\alpha) \in R^n \times R$. For any $\delta > 0$ we define the set $S(\gamma)$ as

$$S(\gamma) = \left\{(z,\mu) \;\middle|\; \left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\| < \gamma\right\} .$$

Also we define the functions $w(a,\alpha)$ and $\zeta(a,\alpha;b,\beta)$, for any $(a,\alpha) \in S$ and $(b,\beta) \in S$, by

$$0 = q(z^*,\mu^*) = q(a,\alpha) + Q(a,\alpha)(z^*-a) + u(a,\alpha)(\mu^*-\alpha) + w(a,\alpha) \tag{A5.2}$$

and

$$\psi(b,\beta) = \psi(a,\alpha) + \partial_z\psi(a,\alpha)^T(b-a) + \partial_\mu\psi(a,\alpha)(\beta-\alpha) + \zeta(a,\alpha;b,\beta) . \tag{A5.3}$$

It follows from [53, Theorem 3.2.5] and the Lipschitz continuity of the derivatives of q and ψ that, if $S(\varepsilon) \subset S$, there are constants K_1 and K_2 such that

$$\|w(a,\alpha)\| \le K_1\left\|\begin{matrix} a-z^* \\ \alpha-\mu^* \end{matrix}\right\|^2 \tag{A5.4}$$

and

$$|\zeta(a,\alpha;b,\beta)| \le K_2\left\|\begin{matrix} a-b \\ \alpha-\beta \end{matrix}\right\|^2 \tag{A5.5}$$

for all $(a,\alpha),(b,\beta) \in S(\varepsilon)$.

Throughout the following we will frequently omit the arguments (z,μ) on functions of z and μ. For example we will write Q for $Q(z,\mu)$ and u for $u(z,\mu)$, etc. From the assumptions, there exist constants $B_1,\ldots,B_4$ such that

(A5.6) $\|Q^{-1}\| \le B_1$; $\left\|\begin{matrix} -Q^{-1}u \\ 1 \end{matrix}\right\| \le B_2$; $\|\partial_z \psi\| \le B_3$; $\|R\| \le B_4$,

for all $(z,\mu) \in S$. Finally, throughout the proof, for any $(z,\mu) \in S$ we will define $\hat{z}(z,\mu)$ by

(A5.7) $$\hat{z}(z,\mu) = z - Q^{-1}q \ .$$

Much of the proof is in the derivation of intermediate results which we present as three lemmas.

LEMMA A5.1. *Let* $\varepsilon > 0$ *be such that* $S(\varepsilon) \subset S$. *Then, if there exists a constant* $C > 0$ *such that, for all* $(z,\mu) \in S$, δ *satisfies*

(A5.8) $$0 < |\delta| < C\left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\| ,$$

it follows that, for all $(z,\mu) \in S(\varepsilon)$, $(\hat{z},\mu)$ *and* $\left(\hat{z}-Q^{-1}u\delta,\mu+\delta\right)$ *are in* $S(\gamma)$, *where*

(A5.9) $$\gamma = B_2(1+C)\varepsilon + B_1K_1\varepsilon^2 \ .$$

Proof. Let $(z,\mu) \in S(\varepsilon)$. Then substituting for $q(z,\mu)$, from (A5.2) into (A5.7), we have

$$\hat{z} = z - Q^{-1}[Q(z-z^*)+u(\mu-\mu^*)-w] ,$$

from which we have

(A5.10) $$\hat{z} - z^* = -Q^{-1}u(\mu-\mu^*) + Q^{-1}w \ .$$

Thus

(A5.11) $$\left\|\begin{matrix} \hat{z}-z^* \\ \mu-\mu^* \end{matrix}\right\| \le \left\|\begin{matrix} -Q^{-1}u \\ 1 \end{matrix}\right\| |\mu-\mu^*| + \|Q^{-1}\|\|w\| \ .$$

Now $|\mu-\mu^*| \le \varepsilon$ and, since $(z,\mu) \in S(\varepsilon)$, it follows from (A5.4) and (A5.6) that

(A5.12) $$\left\|\begin{matrix} \hat{z}-z^* \\ \mu-\mu^* \end{matrix}\right\| \le B_2\varepsilon + B_1K_1\varepsilon^2 = \sigma$$

and $\sigma < \gamma$. Thus $(z,\mu) \in S(\gamma)$.

Next, we note that

$$\left\| \begin{matrix} \hat{z}-Q^{-1}u\delta-z^* \\ \mu+\delta-\mu^* \end{matrix} \right\| \le \left\| \begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix} \right\| + \left\| \begin{matrix} -Q^{-1}u \\ 1 \end{matrix} \right\| |\delta|$$

$$\le \sigma + B_2 C\varepsilon = \gamma \ .$$

Thus $\left(\hat{z}-Q^{-1}u\delta,\mu+\delta\right) \in S(\gamma)$. □

LEMMA A5.2. *Define the function* $\Delta(z,\mu,\delta)$, *for* $\delta \neq 0$, *by*

(A5.13) $$\Delta(z,\mu,\delta) = \left(\psi\left(\hat{z}-Q^{-1}u\delta,\mu+\delta\right)-\psi(\hat{z},\mu)\right)/\delta \ .$$

If, for each (z,μ) , δ *satisfies* (A5.8) *for some constant* C , *then there exists an* $\varepsilon > 0$ *and a* $\rho > 0$ *such that, for all* $(z,\mu) \in S(\varepsilon)$,

(A5.14) $$|\Delta(z,\mu,\delta)| \ge \rho \ .$$

Proof. First we suppose that ε is sufficiently small so that $S(\gamma) \subset S$, where γ is given in (A5.9). Suppose also that $(z,\mu) \in S(\varepsilon)$, then it follows from Lemma A5.1 that $(\hat{z},\mu)$ and $\left(\hat{z}-Q^{-1}u\delta,\mu+\delta\right)$ are in $S(\gamma)$. Now substituting for $\psi\left(\hat{z}-Q^{-1}u\delta,\mu+\delta\right)$ in (A5.13), from (A5.3) with $(b,\beta) = \left(\hat{z}-Q^{-1}u\delta,\mu+\delta\right)$ and $(a,\alpha) = (\hat{z},\mu)$, we have

(A5.15) $$\Delta(z,\mu,\delta) = \left[-\partial_z\psi(\hat{z},\mu)^T Q^{-1}u\delta+\partial_\mu\psi(\hat{z},\mu)\delta+\zeta\left(\hat{z},\mu;\hat{z}-Q^{-1}u\delta,\mu+\delta\right)\right]/\delta \ .$$

It follows from (A5.5) that

$$|\zeta\left(\hat{z},\mu;\hat{z}-Q^{-1}u\delta,\mu+\delta\right)| \le K_2 \left\| \begin{matrix} Q^{-1}u \\ -1 \end{matrix} \right\|^2 \delta^2$$

and so, from (A5.6) and (A5.8), we have

(A5.16) $$|\zeta\left(\hat{z},\mu;\hat{z}-Q^{-1}u\delta,\mu+\delta\right)|/\delta \le K_2 B_2^2 C \left\| \begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix} \right\| \ .$$

Thus, for some constant K , it follows from (A5.15) that

(A5.17) $$|\Delta(z,\mu,\delta)| \ge \left|-\partial_z\psi(\hat{z},\mu)^T Q^{-1}u+\partial_\mu\psi(\hat{z},\mu)\right| - K\left\| \begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix} \right\| \ .$$

Next, define $\hat{R}(z,\mu)$ by,

$$\hat{R} = \begin{bmatrix} Q & u \\ \partial_z\psi(\hat{z},\mu)^T & \partial_\mu\psi(\hat{z},\mu) \end{bmatrix} \ .$$

Then, from (5.3.11), $\hat{R} = R + \tilde{e}_{n+1}p^T$, where $p(z,\mu)$ is given by

(A5.18) $$p(z,\mu) = \begin{bmatrix} \partial_z\psi(\hat{z},\mu)-\partial_z\psi(z,\mu) \\ \partial_\mu\psi(\hat{z},\mu)-\partial_\mu\psi(z,\mu) \end{bmatrix} .$$

By assumption, R is nonsingular for all $(z,\mu) \in S$, and so, from (5.3.2),

$$\mathrm{Det}(\hat{R}) = \mathrm{Det}(R)\left[1+p^T R^{-1}\tilde{e}_{n+1}\right] .$$

Thus, for all $(z,\mu) \in S(\varepsilon)$,

$$|\mathrm{Det}(\hat{R})| \geq |\mathrm{Det}(R)|\left[1-\left|p^T R^{-1}\tilde{e}_{n+1}\right|\right] .$$

Now $\left|p^T R^{-1}\tilde{e}_{n+1}\right| \leq \|p\|\|R^{-1}\|$. Moreover, it follows from (A5.18) and the Lipschitz continuity of $\partial_z\psi$ and $\partial_\mu\psi$ that there is a constant B_5, independent of (z,μ), such that, for all $(z,\mu) \in S$,

$$\|p\| \leq B_5\|\hat{z}-z\| \leq B_5(\|\hat{z}-z^*\|+\|z-z^*\|) .$$

Then, from (A5.12),

$$\|p\| \leq B_5(\sigma+\varepsilon) .$$

Thus, if ε is chosen small enough so that

$$B_5(\sigma+\varepsilon)B_4 \leq \kappa$$

for some $\kappa < 1$, where B_4 is given in (A5.6), it follows that

(A5.19) $$|\mathrm{Det}(\hat{R})| \geq |\mathrm{Det}(R)|(1-\kappa) .$$

Since $R(z,\mu)$ is nonsingular with bounded inverse on S, $|\mathrm{Det}(R)|$ is bounded away from zero on S. Thus, from (A5.19), there exists a constant $\nu > 0$, independent of (z,μ), such that

$$|\mathrm{Det}(\hat{R})| \geq \nu$$

for all $(z,\mu) \in S(\varepsilon)$.

Now, from (5.2.10),

$$\mathrm{Det}(\hat{R}) = \mathrm{Det}(Q)\left\{\partial_\mu\psi(\hat{z},\mu)-\partial_z\psi(\hat{z},\mu)^T Q^{-1}u\right\}$$

and, since Q is bounded on S, there is a constant $L > 0$ such that

$|\mathrm{Det}(Q)| < L$ for all $(z,\mu) \in S$. Thus

$$\left|\partial_\mu \psi(\hat{z},\mu) - \partial_z \psi(\hat{z},\mu)^T Q^{-1} u\right| \geq \nu/L$$

for all $(z,\mu) \in S(\varepsilon)$.

Finally, if ε is chosen so that $\rho = (\nu/L) - K\varepsilon > 0$, it follows from (A5.17) that $\Delta(z,\mu,\delta) \geq \rho$ for all $(z,\mu) \in S(\varepsilon)$. □

LEMMA A5.3. *For any* $(z,\mu) \in S$, *define* $\tilde{\mu}(z,\mu,\delta)$, *for* $\delta \neq 0$, *by*

$$\tilde{\mu}(z,\mu,\delta) = \mu - \frac{\psi(\hat{z},\mu)}{\Delta(z,\mu,\delta)}. \tag{A5.20}$$

If, for each $(z,\mu) \in S$, δ *satisfies* (A5.8) *for some constant* C, *then there exists an* $\varepsilon > 0$ *and an* $M > 0$ *such that*

$$|\tilde{\mu}(z,\mu,\delta) - \mu^*| \leq M \left\| \begin{matrix} z - z^* \\ \mu - \mu^* \end{matrix} \right\|^2$$

for all $(z,\mu) \in S(\varepsilon)$.

Proof. Suppose that ε is sufficiently small so that $S(\gamma) \subset S$, where γ is given in (A5.9). Suppose also that $(z,\mu) \in S(\varepsilon)$, then it follows from Lemma A5.1 that $(\hat{z},\mu)$ and $\left(\hat{z} - Q^{-1}u\delta, \mu+\delta\right)$ are in $S(\gamma)$. From (A5.20) we have

$$\tilde{\mu}(z,\mu,\delta) - \mu^* = \left(\Delta(z,\mu,\delta)(\mu-\mu^*) - \psi(\hat{z},\mu)\right)/\Delta(z,\mu,\delta)$$

and, from (A5.15),

$$\tilde{\mu}(z,\mu,\delta) - \mu^* = \Big[-\partial_z \psi(\hat{z},\mu)^T Q^{-1} u(\mu-\mu^*) + \partial_\mu \psi(\hat{z},\mu)(\mu-\mu^*) - \psi(\hat{z},\mu) + \frac{\hat{\zeta}}{\delta}(\mu-\mu^*)\Big]/\Delta(z,\mu,\delta), \tag{A5.21}$$

where we have written $\hat{\zeta}$ for $\zeta\left(\hat{z},\mu;\hat{z} - Q^{-1}u\delta, \mu+\delta\right)$. We now replace $\psi(\hat{z},\mu)$ in (A5.21) using (A5.3) with $(b,\beta) = (z^*,\mu^*)$ and $(a,\alpha) = (\hat{z},\mu)$ and derive

$$\tilde{\mu}(z,\mu,\delta) - \mu^* = \left\{-\partial_z \psi(\hat{z},\mu)^T\left[\hat{z} - z^* + Q^{-1}u(\mu-\mu^*)\right] + \zeta^* + \frac{\hat{\zeta}}{\delta}(\mu-\mu^*)\right\}/\Delta(z,\mu,\delta),$$

where we have written ζ^* for $\zeta(\hat{z},\mu;z^*,\mu^*)$. It follows immediately from (A5.10) that

(A5.22) $\tilde{\mu}(z,\mu,\delta) - \mu^* = \left\{-\partial_z \psi(\hat{z},\mu)^T Q^{-1} w + \zeta^* + \frac{\hat{\zeta}}{\delta}(\mu-\mu^*)\right\}/\Delta(z,\mu,\delta)$.

From (A5.5) and (A5.8), with $K_3 = K_2 B_2^2 C$,

(A5.23) $$\left|\frac{\hat{\zeta}}{\delta}\right| \leq K_3 \left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\|$$

and by the definition (A5.1),

(A5.24) $$|\mu-\mu^*| \leq \left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\| .$$

Also, from (A5.5),

$$|\zeta^*| \leq K_2 \left\|\begin{matrix} \hat{z}-z^* \\ \mu-\mu^* \end{matrix}\right\|^2 ,$$

which from (A5.11), (A5.6), (A5.4) and (A5.24) gives

(A5.25) $$|\zeta^*| \leq K_4 \left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\|^2$$

for some constant, K_4 , independent of (z,μ) . Finally, using (A5.22)-(A5.25), (A5.6), (A5.4) and Lemma A5.2, it follows that

$$|\tilde{\mu}(z,\mu,\delta)-\mu^*| \leq M \left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\|^2 ,$$

where $M = \left(B_1 B_3 K_1 + K_3 + K_4\right)/\rho$. □

Let $\varepsilon > 0$ be such that the conditions of Lemmas A5.1-A5.3 be satisfied. Also, let $(z_i,\mu_i) \in S(\varepsilon)$. It follows from [53, Theorem 3.2.3] that, for any $(z,\mu) \in S$, $\psi(z,\mu)$ satisfies

$$|\psi(z,\mu)| \leq C \left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\|$$

for some constant C , independent of (z,μ) . Thus, for any $(z,\mu) \in S(\varepsilon)$, the choice of δ given by

$$\delta = |\psi(z,\mu)|/\left(1+\|Q^{-1}u\|\right)$$

satisfies

$$|\delta| \leq C \left\|\begin{matrix} z-z^* \\ \mu-\mu^* \end{matrix}\right\| .$$

Also if, whenever $\psi(z,\mu) = 0$, δ is chosen sufficiently small, the choice

of δ satisfies (A5.8).

Now, from (5.3.6) and (A5.7), $\hat{z}_{i+1} = \hat{z}(z_i,\mu_i)$ and so, from (A5.10),

$$\hat{z}_{i+1} - z^* = -Q_i^{-1}u_i(\mu_i-\mu^*) + Q_i^{-1}w_i ,$$

where $Q_i = Q(z_i,\mu_i)$ etc. Substituting into (5.3.10) gives

$$z_{i+1} - z^* = -Q_i^{-1}u_i(\mu_{i+1}-\mu^*) + Q_i^{-1}w_i$$

and hence

$$\left\|\begin{matrix} z_{i+1}-z^* \\ \mu_{i+1}-\mu^* \end{matrix}\right\| \leq \left\|\begin{matrix} -Q_i^{-1}u_i \\ 1 \end{matrix}\right\| |\mu_{i+1}-\mu^*| + \left\|Q_i^{-1}\right\| \|w_i\| .$$

But it follows from (5.3.7), (5.3.9) and (A5.20) that $\mu_{i+1} = \tilde{\mu}(z_i,\mu_i,\delta_i)$ and so from (A5.4), (A5.6) and Lemma A5.3,

$$\left\|\begin{matrix} z_{i+1}-z^* \\ \mu_{i+1}-\mu^* \end{matrix}\right\| \leq A\left\|\begin{matrix} z_i-z^* \\ \mu_i-\mu^* \end{matrix}\right\|^2 \tag{A5.26}$$

where $A = B_2M + B_1K_1$. It follows that, if ε is such that $A\varepsilon < 1$, then $(z_{i+1},\mu_{i+1}) \in S(\varepsilon)$ and, by induction, that the sequence $\{(z_i,\mu_i)\}$ converges to (z^*,μ^*) . Finally, it follows trivially from (A5.26) that the sequence converges with R-order ≥ 2 . □

CHAPTER 6

FINDING SEVERAL SOLUTIONS OF NONLINEAR EQUATIONS

6.1. Introduction

In this chapter we consider the problem of finding *several* solutions of the nonlinear system of equations

$$f(x) = 0 \ , \tag{6.1.1}$$

$f : D \subset R^n \to R^n$ and, unless stated otherwise, we shall assume that f is twice differentiable on D. This problem is often of interest although it has received little attention in the literature. The approach frequently adopted is to use an iterative scheme, often based on Newton's method, with a variety of starting guesses. However, Brown and Gearhardt [14] have noted that this approach can fail on quite simple problems, when the method continually finds the same root or, of course, the method may continually diverge and fail to find the desired roots. Recently two approaches have been suggested for overcoming these difficulties and in this chapter we consider the two methods and draw some conclusions about their computational efficiency.

In section 6.2 we consider the approach suggested by Branin [11]. The basis of his method was presented in section 4.4 as a method with wide convergence for finding a single solution of (6.1.1). In his paper, Branin proposed an extension of the method as a means of finding several solutions. He suggested following the solution trajectory of

$$\dot{x}(t) = -J(x)^{-1}f(x) \ , \qquad x(0) = x_0 \ , \tag{6.1.2}$$

where, as in Chapters 1-4, $J(x)$ is the Jacobian of $f(x)$ and x_0 is an estimate of a solution of (6.1.1). Under the conditions of Theorem 2.2.1 the solution trajectory of (6.1.2) converges to a solution of (6.1.1). To

find a second solution, Branin suggested reversing the sign in (6.1.2) and following the solution of the new differential equation in a direction away from the first root and away from x_0 . On crossing a region where $J(x)$ is singular he reverted to following the solution of (6.1.2), hopefully giving convergence to a new root. In this chapter we shall refer to Branin's method not as the means of following the trajectory, as described in section 4.4, but as the principle of following the whole solution trajectory of (6.1.1). In fact, in our numerical tests we used the method NEW/2 of Chapter 4 to follow the trajectory.

In section 6.3 we describe the approach due to Brown and Gearhardt [14] who extended the idea of deflation, usually associated with finding roots of a polynomial, to dimensions greater than one. On finding a root, r , of $f(x)$ they suggested finding a zero of the deflated function $g(x) = f(x)/\|x-r\|$, where $\|\cdot\|$ is some norm on R^n . If r is a simple root of f then r is not a root of g . It is shown in section 6.3 that, if Newton's method is used to solve the deflated equation, then the resulting method is similar to Branin's method and can be considered as differing only in the way in which it chooses the sign in (6.1.2) and in the accuracy with which it follows the resulting trajectory.

Branin's method is more successful than the deflation method for finding several solutions of (6.1.1), however it is more costly in terms of the amount of computation per zero found. In section 6.4 we present numerical results which demonstrate this and, for completeness, describe a modification of the methods which is more efficient than Branin's method and is more successful in finding zeros than Brown and Gearhardt's method.

We note that Chao, Liu and Pan [16] used a modification of Branin's method, however, in their paper they gave little detail about the computational efficiency of the resulting method.

6.2. Branin's Method

The method can be described as one which follows the solution trajectory $x(\lambda)$ of

$$f(x(\lambda)) - (1-\lambda)f(x_0) = 0 , \qquad x(0) = x_0 . \tag{6.2.1}$$

It is shown in Chapter 1 that the solutions of (6.2.1) are essentially the same as those of (6.1.2). The method attempts to follow $x(\lambda)$ from $x(0) = x_0$ to $x(1) = r$, which is a root of f. Then the method continues to follow $x(\lambda)$ for $\lambda > 1$ until either the trajectory passes through a turning point or the trajectory diverges to infinity (or, in practice, goes beyond some prescribed bound). For the former case the method continues to follow the trajectory, with λ decreasing now, possibly on to another solution. Each point on the solution branch $(x(\lambda),\lambda)$, at which $\lambda = 1$, represents a solution of (6.1.1). Note that this process is exactly that described in Chapter 5 except that here we are not interested in the accurate determination of the whole solution branch.

When the solution trajectory diverges, the method returns to $x(0)$, follows the trajectory with λ decreasing and repeats the whole process until divergence occurs again. Thus the method follows the trajectory through x_0 in both directions.

Branin actually suggested integrating (6.1.2) rather than (6.2.1) and the corresponding version of (6.2.1) is derived by applying the transformation $1 - \lambda = e^{-t}$. Since we wish to follow the solution trajectory in both directions we need to modify the resulting equation and to follow the solution of

$$f(x(t)) - e^{-\delta t}f(x_0) = 0 , \qquad x(0) = x_0 , \tag{6.2.2}$$

where $\delta = 1$ if we are approaching a zero or $\delta = -1$ if we are leaving a zero in search of a turning point. For this formulation there is an added

complication, due to integrating over the infinite interval. On finding a solution of (6.1.1) and before following the correct trajectory away from the root, it is necessary to "step over" this root onto the solution of (6.2.2). In accordance with our comments of Chapters 1-4, we prefer to use (6.2.2), for then we can follow the solution trajectory using a method designed to take advantage of the Liapunov stability of (6.2.2) when $\delta = 1$.

In his paper, Branin gave several examples and diagrams which are important as they give an insight into the behaviour of the solution trajectories. In particular he notes that the method is not always successful since $x(t)$ may not pass through all the roots or indeed may not pass through any. Actually, the method is less reliable than Branin hoped since he offered a conjecture that the method finds all the solutions of (6.1.1) if the problem has no *extraneous singularities* (defined below). This conjecture is not true, as shown in the following example. A singularity of the differential equation $\dot{x}(t) = q(x)$ is a point x such that $q(x) = 0$. Thus, any solution of (6.1.1) is a singularity of (6.1.2) and, as described in Chapter 1, is a stable node in the Liapunov sense. Branin modifies (6.1.2) and considers the differential equation

$$\dot{x}(t) = \mathrm{Adj}\big(J(x)\big)f(x)/\mathrm{Det}\big(J(x)\big) \ ,$$

where $\mathrm{Adj}(\cdot)$ denotes the adjoint. Apart from a sign, this is equivalent to (6.1.2). He then defines an extraneous singularity as a point x such that $f(x) \neq 0$ and $\mathrm{Adj}\big(J(x)\big)f(x) = 0$. Such points often give rise to a region of non-convergence, i.e. a region S such that, for any $x_0 \in S$, the solution trajectory of (6.2.1) does not pass through all the zeros of f . The following problem is shown to possess no extraneous singularities but does have a region of non-convergence and so disproves Branin's conjecture. Consider (6.1.1) with $f(x)$ given by

$$f(x) = \begin{bmatrix} x_1^2 - x_2 \\ x_2^2 - 1 \end{bmatrix},$$

where we have written $x^T = (x_1, x_2)$. Then

$$\text{Adj}\big(J(x)\big) f(x) = \begin{bmatrix} 2x_1^2 x_2 - x_2^2 - 1 \\ 2x_1\left(x_2^2 - 1\right) \end{bmatrix} \tag{6.2.3}$$

and it is easy to show that $\text{Adj}(J)f = 0$ if and only if $f = 0$. Thus the problem has no extraneous singularities. The problem has two solutions at $(1,1)$ and $(-1,1)$, however solution trajectories of (6.1.2) passing through points (x_1, x_2) such that $x_2 < -1$ do not converge to a solution. A full analysis of the trajectories shows this, however we briefly note that, when $x_2 = -1$, the unit vector in the direction of a solution trajectory of (6.1.2) is $\dot{x}(t)/\|\dot{x}(t)\|$ and from (6.2.3) is given by

$$\frac{\dot{x}(t)}{\|\dot{x}(t)\|} = \xi \begin{bmatrix} 2\left(x_1^2 + 1\right) \\ 0 \end{bmatrix},$$

where ξ is a scaling factor. Thus, the trajectory is parallel to the x_1 axis. This indicates that no trajectories cross the line $x_2 = -1$ which, on further analysis, proves to be the case. So the region $S = \{x \mid x_2 < -1\}$ is a region of non-convergence.

In general then, following a trajectory does not guarantee finding all or even any solutions. Despite this failing, Branin's method appears to be the most reliable of the methods currently available. Unfortunately, to be sure of finding all the solutions on a trajectory requires a large amount of computation and so some balance must be found between efficiency and guaranteed success in finding all zeros on a trajectory. We discuss this further in section 6.4.

6.3. The Deflation Technique

6.3.1. A New Formulation

In this section we consider a method of deflation similar to that applied to polynomial equations. Having found a root, ξ, of a polynomial $\rho(\alpha)$, other roots can be found by solving the equation

$$\rho(\alpha)/(\alpha-\xi) = 0 .$$

The process is described in detail by Wilkinson [73]. Brown and Gearhardt [14] extended this idea to solving (6.1.1).

Let $r \in R^n$ and $M(x,r)$ be a matrix on R^n which is defined for all $x \in U_r$, where U_r is open in $D \subset R^n$ and r belongs to the closure of U_r. Then Brown and Gearhardt define M to be a *deflation matrix* if, for any differentiable function $f : D \subset R^n \to R^n$ such that $f(r) = 0$ and $J(r)$ is nonsingular, we have

$$\liminf_{i\to\infty} \|M(x_i,r)f(x_i)\| > 0$$

for any sequence $\{x_i\}$ such that

$$\lim_{i\to\infty} x_i = r$$

and $x_i \in U_r$. Thus any iterative method which converges to a solution of

(6.3.1) $$M(x,r)f(x) = 0$$

will not converge to r. The process suggested by Brown and Gearhardt is to find a root r, by some method, then with some deflation matrix $M(x,r)$, to solve (6.3.1). If, in addition, $M(x,r)$ is chosen to be nonsingular for all $x \in R^n\backslash\{r\}$, then any solution of (6.3.1) will also be a solution of (6.1.1) and, by choice of $M(x,r)$, will be different from r. The process can be repeated to deflate out a number of roots $r_1,r_2,\ldots,r_k$ by solving

$$M(x,r_k)\ldots M(x,r_2)M(x,r_1)f(x) = 0$$

and we consider this further in section 6.3.3.

The most obvious choice of $M(x,r)$ is $I/\|x-r\|$, for some norm $\|\cdot\|$ on R^n, and it is this form of deflation that we shall consider. Define $\eta : R^n \to R$ by

$$\eta(x) = \|x-r\| ,$$

then the deflated function $g : D\backslash\{r\} \subset R^n \to R^n$ is defined by

$$g(x) = \frac{f(x)}{\eta(x)} . \tag{6.3.2}$$

Brown and Gearhardt suggested taking differences of $g(x)$ to form an estimate of $G(x)$, the Jacobian of $g(x)$, and using a discrete version of Newton's method to solve $g(x) = 0$. We prefer to form $G(x)$ explicitly in terms of $J(x)$ and to use Newton's method to solve the deflated equation. From (6.3.2) we have

$$G(x) = \frac{1}{\eta(x)} \left[J(x) - f(x) \frac{\eta'(x)^T}{\eta(x)} \right] , \tag{6.3.3}$$

where we have written

$$\frac{d\eta}{dx}(x) = \eta'(x) .$$

We note that $G(x)$ is defined only where $\eta(x) \neq 0$ and $\eta'(x)$ is defined. For example, if $\eta(x) = \|x-r\|_2 = \left[(x-r)^T(x-r)\right]^{\frac{1}{2}}$, then $G(x)$ is defined on $D\backslash\{r\}$. For $\eta(x) = \|x-r\|_p$, $p = 1,\infty$, $G(x)$ is defined on $D\backslash S_p$, where

$$S_1 = \{x \mid x_i - r_i = 0 \text{ for some } i\}$$

and

$$S_\infty = \{x \mid |x_i-r_i| = |x_j-r_j| = \|x-r\|_\infty \text{, for some } i,j \text{, } i \neq j\} .$$

These restrictions do not present any difficulties since, in practice, we extend the definition of $\eta'(x)$ so that it is defined for all $x \neq r$. We demonstrate this extension for $\eta(x) = \|x-r\|_p$, $p = 1,\infty$ for, together with $p = 2$, these are the norms which are most convenient to use in practice.

Writing $\eta^{(1)}(x) = \|x-r\|_1$, then formally $\frac{d\eta^{(1)}}{dx}(x)$ is defined only on $R^n \backslash S_1$. However, $\eta^{(1)}(x)$ can be written

$$\eta^{(1)}(x) = \sum_{i=1}^{n} \gamma_i (x_i - r_i) , \tag{6.3.4}$$

where

$$\gamma_i = \begin{cases} 1 & \text{if } x_i \geq r_i , \\ -1 & \text{if } x_i < r_i . \end{cases} \tag{6.3.5}$$

We can define $\frac{d\eta^{(1)}}{dx}(x)$ arbitrarily on S_1 without affecting the results, so we define it in the natural way by

$$\frac{\partial \eta^{(1)}}{\partial x_i}(x) = \gamma_i , \tag{6.3.6}$$

$i = 1,2,\ldots,n$, and then $\frac{d\eta^{(1)}}{dx}(x)$ is defined for all $x \neq r$. Similarly, writing $\eta^{(\infty)}(x) = \|x-r\|_\infty$, then $\eta^{(\infty)}(x)$ can be written

$$\eta^{(\infty)}(x) = \sum_{i=1}^{n} \delta_i (x_i - r_i) , \tag{6.3.7}$$

where

$$\delta_i = \begin{cases} 1 & \text{if } i = i_0 \text{ and } x_{i_0} \geq r_{i_0} , \\ -1 & \text{if } i = i_0 \text{ and } x_{i_0} < r_{i_0} , \\ 0 & \text{if } i \neq i_0 \end{cases} \tag{6.3.8}$$

and i_0 is the smallest i such that

$$|x_i - r_i| \geq |x_j - r_j|$$

for $j = 1,2,\ldots,n$. Normally $\frac{\partial \eta^{(\infty)}}{\partial x_i}(x)$ is undefined if $|x_i - r_i| = |x_{i_0} - r_{i_0}|$, however we make the formal definition

(6.3.9) $$\frac{\partial \eta}{\partial x_i}^{(\infty)}(x) = \delta_i ,$$

$i = 1,2,\ldots,n$, and then $\frac{d\eta}{dx}^{(\infty)}(x)$ is defined for all $x \neq r$.

If A is a nonsingular matrix and x and y are vectors, then $A + xy^T$ is nonsingular if and only if $1 + y^T A^{-1} x \neq 0$ and the Sherman-Morrison formula states that

(6.3.10) $$\left(A+xy^T\right)^{-1} = A^{-1} - \frac{A^{-1}xy^T A^{-1}}{1+y^T A^{-1}x} ,$$

(see Householder [35]). Thus, from (6.3.3),

(6.3.11) $$G(x)^{-1} = \eta(x)\left[J(x)^{-1} + \frac{J(x)^{-1}f(x)\ \frac{\eta'(x)^T}{\eta(x)}\ J(x)^{-1}}{1 - \frac{\eta'(x)^T}{\eta(x)}\ J(x)^{-1}f(x)}\right] .$$

Writing $q(x) = -J(x)^{-1}f(x)$ and

(6.3.12) $$\sigma(x) = \frac{1}{1+q(x)^T\ \frac{\eta'(x)}{\eta(x)}} ,$$

some simple algebra using (6.3.2) and (6.3.11) shows that

$$-G(x)^{-1}g(x) = q(x)\sigma(x) .$$

Therefore the Newton iteration

(6.3.13) $$x_{i+1} = x_i - G(x_i)^{-1}g(x_i)$$

for solving $g(x) = 0$ can be written

(6.3.14) $$x_{i+1} = x_i + q(x_i)\sigma(x_i) .$$

This represents an improved formulation of the deflation technique for we see that Newton's method applied to the equations $g(x) = 0$ can be implemented without the need to evaluate derivatives of $g(x)$ directly. Moreover, for some i , $G(x_i)$ may be nearly singular and whilst this may be because $J(x_i)$ is nearly singular we see from (6.3.11) that it may also

be because $\sigma(x_i)$ is large. If this is the case then, in applying (6.3.13), we have none of the difficulties involved in calculating $G(x_i)^{-1}g(x_i)$. If $\sigma(x_i)$ is large then suitable damping can easily be applied to the step in (6.3.14).

6.3.2. The Relation with Branin's Method

Our view of deflation follows from (6.3.14) which shows that the resulting method is essentially Euler's method for integrating (6.1.2) with a special choice of step-size and as such is similar to Branin's method. We note that Euler's method is only first order for (6.1.2) and so will be less successful in following the solution trajectory than the higher order methods previously discussed. This is borne out in practice, as we show in section 6.4. Also we note that $\sigma(x_i)$ is often an unsuitable choice of step-size in (6.3.14) and we demonstrate this below. First we prove two lemmas.

LEMMA 6.3.1. *Let* $\eta : R^n \to R$ *be defined by* $\eta(x) = \|x-r\|_p$, $p = 1,2$ *or* ∞, *for some* r. *Assume that, for* $p = 1$ *and* ∞, $\eta'(x)$ *is defined by* (6.3.6) *and* (6.3.9) *respectively. Then, for any* $x \neq r$,

$$\eta'(x)^T(x-r) = \eta(x) . \tag{6.3.15}$$

Proof. For $p = 1$, $\frac{\partial \eta}{\partial x_i}(x) = \gamma_i$, where γ_i is given in (6.3.5). Thus

$$\eta'(x)^T(x-r) = \sum_{i=1}^{n} \gamma_i(x_i-r_i)$$

and from (6.3.4), the result follows.

For $p = 2$, $\eta'(x) = (x-r)/\|x-r\|$ and so (6.3.15) follows immediately.

Finally, for $p = \infty$, $\frac{\partial \eta}{\partial x_i}(x) = \delta_i$, where δ_i is given in (6.3.8).

So

$$\eta'(x)^T(x-r) = \sum_{i=1}^{n} \delta_i(x_i-r_i)$$

and the result follows from (6.3.7). □

We note that (6.3.15) is true for all norms of the form

$$\|x\| = (x^TAx)^{\frac{1}{2}} ,$$

for any positive definite matrix A or of the form

$$\|x\| = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} ,$$

$p \geq 1$, assuming that, in each case, $\eta'(x)$ is defined so that it exists for all $x \neq r$. We have restricted our attention to the cases $p = 1,2$ and ∞ since these are the practical choices.

LEMMA 6.3.2. *Let* $f : D \subset R^n \to R^n$ *and* $f(r) = f(\tilde{r}) = 0$ *where* $r \in \text{Int}(D)$, $\tilde{r} \in \text{Int}(D)$ *and* $r \neq \tilde{r}$. *Let* $N_r \subset D$ *and* $N_{\tilde{r}} \subset D$ *be convex neighbourhoods of* r *and* $\tilde{r}$ *respectively such that* $N_r \cap N_{\tilde{r}}$ *is empty. Suppose also that* $J(x)$ *is Lipschitz continuous on* N_r *and on* $N_{\tilde{r}}$ *and that* $J(x)^{-1}$ *exists and is bounded on both* N_r *and* $N_{\tilde{r}}$. *Finally, let* $\eta(x) = \|x-r\|_p$ *for* $p = 1,2$ *or* ∞ *and suppose, for the cases* $p = 1,\infty$, *that* $\eta'(x)$ *is defined by* (6.3.6) *and* (6.3.9) *respectively. Then with* $\sigma(x)$ *defined in* (6.3.12), *it follows that*

(i) if $\{x_i\}$ *is a sequence such that* $\lim_{i\to\infty} x_i = \tilde{r}$, *then*

$$\lim \sigma(x_i) = 1$$

and

(ii) there exists a constant K *such that for all* $x \in N_r\backslash\{r\}$,

$$|\sigma(x)| \geq K/\|x-r\| .$$

Proof. We assume that the norm used in the proof is the same norm that

defines $\eta(x)$. From our assumptions, $\eta'(x)$ exists on $N_{\tilde{r}}$. Also $\eta'(x)$ is bounded on $D\setminus\{r\}$, in particular, for each $x \neq r$, $\|\eta'(x)\| = n$ if $p = 1$ and $\|\eta'(x)\| = 1$ if $p = 2$ or ∞ . In addition $\eta(x)$ is bounded away from zero on $N_{\tilde{r}}$ and $J(x)^{-1}$ is bounded on $N_{\tilde{r}}$. Writing

$$\alpha(x) = \frac{\eta'(x)^T J(x)^{-1} f(x)}{\eta(x)} ,$$

it follows that there is a constant L_1 such that

$$|\alpha(x)| \leq L_1 \|f(x)\|$$

for all $x \in N_{\tilde{r}}$. The result now follows because $\sigma(x) = 1/\big(1+\alpha(x)\big)$ and $f(\tilde{r}) = 0$.

To prove (ii) we define the function $u(x)$ by

$$u(x) = f(x) + J(x)(x-r) .$$

Then, from [53, Theorem 3.2.5] and the Lipschitz continuity of $J(x)$ on N_r , there is a constant $L_2 > 0$ such that, for all $x \in N_r$,

$$\|u(x)\| \leq L_2 \|x-r\|^2 .$$

Also, from the assumptions, there is a constant $K_2 > 0$ such that

$$\|J(x)^{-1}\| \leq K_2$$

for all $x \in N_r$. Now $\alpha(x)$ can be written

$$\alpha(x) = \frac{\eta'(x)^T}{\eta(x)} J(x)^{-1}[u(x)-J(x)(x-r)]$$

$$= -\frac{\eta'(x)^T(x-r)}{\eta(x)} + \frac{\eta'(x)^T J(x)^{-1} u(x)}{\eta(x)} .$$

From Lemma 6.3.1 we have

$$\frac{\eta'(x)^T(x-r)}{\eta(x)} = 1$$

and so

$$1 + \alpha(x) = \frac{\eta'(x)^T J(x)^{-1} u(x)}{\eta(x)} .$$

Because $\eta(x) = \|x-r\|$ and the way in which $\|\eta'(x)\|$ is bounded, it follows that, for each $x \in N_r \backslash \{r\}$,

$$|1+\alpha(x)| \leq nK_2L_2\|x-r\| .$$

Since $\sigma(x) = \big(1+\alpha(x)\big)^{-1}$ the result follows with $K = \big(nK_2L_2\big)^{-1}$. □

Part (i) of Lemma 6.3.2 shows that, in the region of a solution other than r , the method behaves like Newton's method and tends to Newton's method as iterates converge to the new zero. Part (ii) shows that, in the region of r , the stepsize $\sigma(x)$ is large and, under the conditions of the theorem, acts to force iterates away from r . Unfortunately, far from $\tilde{r}$, the signs of $\sigma(x)$ and $q(x)\sigma(x)$ and, close to r , the magnitude of $q(x)\sigma(x)$ are all unpredictable. This is in contrast to the scalar case where, under extra differentiability conditions, $q(x)\sigma(x)$ tends to a finite limit as $x \to r$. For the vector case this unpredictability means that the behaviour of the method is also unpredictable. This is borne out in practice and the following example shows that the method can fail in simple cases. Consider (6.1.1) with $f(x)$ given by

$$f(x) = \begin{bmatrix} 4x_1^3 - 3x_1 - x_2 \\ x_1^2 - x_2 \end{bmatrix}$$

where we have written $x = [x_1, x_2]^T$. This problem is taken from Brown and Gearhardt's paper and has three zeros at $(1,1)$, $(0,0)$ and $(-0.75, 0.5625)$. Suppose the first zero found is at $r = [1,1]^T$ and we perform deflation with $\eta(x)$ given by $\eta(x) = \|x-r\|_\infty$. We define the function $s(x)$ by

(6.3.16) $$s(x) = q(x)\sigma(x)$$

and the set B by

$$B = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \;\middle|\; x_1 > 1,\ x_2 < 2 - x_1 \right\} .$$

With $e^T = (1,1)$ and $e_1^T = (1,0)$ an equivalent definition of B is

$$B = \left\{ x \mid e_1^T(x-r) > 0,\ e^T(x-r) < 0 \right\} . \tag{6.3.17}$$

It is simple to show that, for all $x \in B$, $\eta(x) = 1 - x_2$ and so $\eta'(x)^T = [0,-1]$ for $x \in B$. Some straightforward algebra gives

$$q(x) = -\frac{1}{\Delta(x)} \begin{bmatrix} \left(3+x_1-4x_1^2\right)x_1 \\ 4x_1^4+3x_1^2+\Delta(x)x_2 \end{bmatrix} \tag{6.3.18}$$

where $\Delta(x) = -12x_1^2 + 3 + 2x_1$. Then it follows that, for all $x \in B$,

$$\sigma(x) = \frac{\Delta(x)\left(1-x_2\right)}{4\left(x_1-1\right)^2\left(x_1+1/2\right)\left(x_1+3/2\right)} .$$

Now $\Delta(x) < 0$ for all $x \in B$ and it follows that $\sigma(x) < 0$ for all $x \in B$. Therefore, from (6.3.16) and (6.3.18), for all $x \in B$,

$$e_1^T s(x) = -\frac{\sigma(x)}{\Delta(x)}\left(3+x_1-4x_1^2\right)x_1 = \frac{\sigma(x)}{\Delta(x)}\left(x_1-1\right)\left(4x_1+3\right)x_1$$

and so, for all $x \in B$,

$$e_1^T s(x) > 0 . \tag{6.3.19}$$

Similarly, for $x \in B$,

$$e^T s(x) = -\frac{\sigma(x)}{\Delta(x)}\left(4x_1^4-4x_1^3+4x_1^2+3x_1+\Delta(x)x_2\right) .$$

But, if $x \in B$, $x_2 < 2 - x_1$, so, for $x \in B$,

$$e^T s(x) < -\frac{\sigma(x)}{\Delta(x)}\left(4x_1^4-4x_1^3+4x_1^2+3x_1+\Delta(x)\left(2-x_1\right)\right) .$$

From the definition of $\Delta(x)$ we can now derive

$$e^T s(x) < -\frac{\sigma(x)}{\Delta(x)}\left(x-1\right)^2\left(4x_1^2+16x_1+6\right) < 0 \tag{6.3.20}$$

for all $x \in B$. Thus, if $x \in B$, it follows from the definition in

(6.3.17) and from (6.3.19) and (6.3.20) that $x + s(x) \in B$. Therefore, if $x_0 \in B$, the iterates defined by (6.3.14) must all be in B and so can never converge to another zero. This is in contrast to Branin's method which is globally convergent for this problem, in that the solution trajectory of (6.2.1) passes through all three zeros for any x_0 .

6.3.3. Multiple Deflation

All we have said can be applied to deflation with respect to several zeros. Suppose we have found zeros $r_1, r_2, \ldots, r_k$ of $f(x)$. Then, letting $\eta_k(x) = \|x-r_k\|$, the deflated function $g_k(x)$ is given by

$$g_k(x) = \frac{f(x)}{\eta_1(x)\ldots\eta_k(x)} ,$$

where $g_k : D\backslash\{r_1,\ldots,r_k\} \subset R^n \to R^n$. The Jacobian, $G_k(x)$, of $g_k(x)$ is given by

$$G_k(x) = \frac{1}{\eta_1(x)\ldots\eta_k(x)} \left[J(x) - f(x) \sum_{j=1}^{k} \frac{\eta_j'(x)^T}{\eta_j(x)} \right]$$

which is the generalisation of (6.3.3). Now using formula (6.3.10), simple algebra gives

$$-G_k(x)^{-1} g_k(x) = q(x)\sigma_k(x)$$

where

$$\sigma_k(x) = \frac{1}{1+q(x)^T \sum_{j=1}^{k} \frac{\eta_j'(x)}{\eta_j(x)}} .$$

Thus, in the general case, (6.3.14) becomes

$$x_{i+1} = x_i + q(x_i)\sigma_k(x_i) .$$

Again with this formulation, it is often easy to overcome the problem of $G_k(x)$ being almost singular. This is detected if $\sigma_k(x)$ is large and

again, suitable damping can easily be applied.

6.4. Numerical Results

In this section we describe some numerical tests performed on the methods of sections 6.2 and 6.3. The implementation of Branin's method (Method 1) is essentially an extension of the method NEW/2 described in Chapter 4. Changes were included, first to step over a solution and follow the trajectory away from the root. On finding a root, to step over and find a new starting point, the equation

$$f(x) = (-1)^k f(x_0) \cdot 10^{-5}$$

was solved, where k is the number of roots found so far. Secondly, the sign of $\mathrm{Det}(J(x))$ was monitored so that the method was aware of passing a turning point. The deflation technique (Method 2) was implemented as described in section 6.3 with the 1, 2 and ∞ norms. Since there was no appreciable difference in the results, the ∞ norm, being the simplest to evaluate, was used in the experiments. Given x_0, the method based upon that of Brown and Gearhardt can be written as

(i) find a root r_1 by Newton's method,

(ii) deflate with respect to r_1 and begin again, *using* x_0 *as starting guess*, to find a further root r_2,

(iii) deflate by r_2 and repeat until a termination criterion is satisfied.

Also, for both methods, iterations were continued until

(a) x_i became too large and a test of the form $\|x_i - c\| < \Delta$ was violated, where c and Δ were function dependent, e.g., $c^T = (0,0)$ and $\Delta = 10$ for function 3 below,

(b) the required number of zeros were found, (for each function the number of zeros required was preset - the values are given below as zmax), or

(c) the maximum number of iterations was exceeded, (if a method took more than 35 function evaluations to find one zero then iterations were terminated).

The methods were tested on the following eight functions, chosen because they were known to have more than one zero. In each case the methods were initiated with ten starting guesses which were chosen at random from a region surrounding the zeros of interest. The first four functions were described in Brown and Gearhardt [14].

1.
$$f_1 = 4x_1^3 - 3x_1 - x_2 ,$$
$$f_2 = x_1^2 - x_2 .$$

This system has three zeros and zmax = 3 .

2.
$$f_1 = \left(x_1 - x_2^2\right)\left(x_1 - \sin x_2\right) ,$$
$$f_2 = \left(\cos x_2 - x_1\right)\left(x_2 - \cos x_1\right) .$$

This function has four zeros in the unit square with others elsewhere. zmax was set equal to 4 .

3.
$$f_1 = x_1 x_2 - 1 ,$$
$$f_2 = x_1^2 + x_2^2 - 4 ,$$

which has 4 zeros, zmax = 4 .

4.
$$f_1 = x_1^2 + 2x_2^2 - 4 ,$$
$$f_2 = x_1^2 + x_2^2 + x_3 - 8 ,$$
$$f_3 = \left(x_1 - 1\right)^2 + \left(2x_2 - \sqrt{2}\right)^2 + \left(x_3 - 5\right)^2 - 4 .$$

This function has two roots and zmax = 2 .

5. Problem 1 of section 2.5, which has several zeros, $zmax = 2$.

6. Problem 3 of section 2.5, which has two close zeros in the positive quadrant and others elsewhere, $zmax = 2$.

7. A function from Chao, Liu and Pan [16],

$$f_1 = x_1 + x_2 + x_3 + x_4 - 1 ,$$

$$f_2 = x_1 + x_2 - x_3 + x_4 - 3 ,$$

$$f_3 = x_1^2 + x_2^2 + x_3^2 + x_4^2 - 4 ,$$

$$f_4 = (x_1-1)^2 + x_2^2 + x_3^2 + x_4^2 - 4 ,$$

which has two zeros so $zmax = 2$.

8.
$$f_1 = x_1^2 - x_2 + x_4 + (x_3-x_5)^2 ,$$

$$f_2 = x_2 - x_4 - 1 ,$$

$$f_3 = x_3 - x_5 - 2x_1^2 + 2 ,$$

$$f_4 = x_4 - x_1^2 - x_3 + x_5 ,$$

$$f_5 = x_5 - x_1 + (x_2-x_4)^2 ,$$

and this function has four zeros, $zmax = 4$.

The results of the numerical tests are given in Table 6.1 where, for each method, the first line gives the number of zeros found in the ten runs and the second line gives the number of equivalent function evaluations *per zero*, where one Jacobian evaluation is considered as n equivalent function evaluations. This measure of the amount of work done was used since Method 1 attempts to improve efficiency by evaluating $J(x)$ only when necessary and not necessarily each time $f(x)$ is evaluated. The criterion that iterations be terminated at a zero was that $\|f(x_i)\|_\infty < 10^{-6}$.

TABLE 6.1

METHOD	FUNCTION							
	1	2	3	4	5	6	7	8
1	23	31	29	17	20	17	20	24
	49	52	50	64	37	53	60	129
2	10	28	17	13	12	12	11	11
	24	22	19	39	31	30	41	47
3	20	30	29	15	18	17	18	15
	22	31	28	52	23	40	55	82

The results show, as predicted, that Method 1 is more successful in finding zeros and, in fact, found 79% of the maximum possible whereas Method 2 found only 50%. However, Method 2 was considerably more efficient in terms of the amount of work expended per zero. This was largely due to the fact that Method 1 follows a trajectory in *both* directions and often requires several more iterations than Method 2 before terminating because of criterion (a) above.

In order to attempt a balance, a new method was written (Method 3) which followed the solutions of (6.1.2) like Method 1 but only as accurately as Method 2. The basic iteration is therefore

$$x_{i+1} = x_i - J(x_i)^{-1} f(x_i) \, |\sigma_k(x_i)| \, \delta \tag{6.4.1}$$

where δ is as described in equation (6.2.2). Notice that the presence of the term $|\sigma_k(x_i)|$ in (6.4.1) precludes the possibility of converging again to a known simple root. The results for this method are also given in Table 6.1 and show it to be a possible compromise between Methods 1 and 2. Method 3 found 70% of the possible zeros but was less efficient than Method 2, primarily because, like Method 1, it follows trajectories in both directions. Note that Method 3 represents a simple modification of the method of Brown and Gearhardt and gives a significant improvement to the

performance of that method. For practical problems, the actual choice of method would depend upon how one balances computation cost with the need to find as many zeros as possible.

Finally, we note that we also used the method of section 5.2 to follow the solution of (6.2.1) with $H(x,t)$ given by (6.2.2). (This was the motivation for the work of Menzel and Schwetlick [49].) The method was modified to give second order convergence to solutions of (6.1.1) however, since the method is designed to follow a solution trajectory with some accuracy and since this is not required in this application, the method did not give any improvement over our implementation of Branin's method, which did not demand very high accuracy in the region of a turning point.

REFERENCES

[1] J.P. Abbott, Methods for finding several solutions of simultaneous nonlinear equations, Proc. 7th Australian Computer Conference, Perth, 1976, pp. 1014-1022.

[2] J.P. Abbott and R.P. Brent, Fast local convergence with single and multistep methods for nonlinear equations, to appear, J. Austral. Math. Soc., Ser. B.

[3] P.M. Anselone and R.H. Moore, An extension of the Newton-Kantorovich method for solving nonlinear equations with application to elasticity, J. Math. Anal. Appl., 13 (1966), pp. 476-501.

[4] J.H. Avila, Continuation methods for nonlinear equations, Ph.D. Thesis, Tech. Rep. TR-142, Computer Science Center, University of Maryland, 1971.

[5] J.H. Avila, The feasibility of continuation methods for nonlinear equations, SIAM J. Numer. Anal., 11 (1974), pp. 102-122.

[6] L. Bauer, E.L. Reiss and H.B. Keller, Axisymmetric buckling of hollow spheres and hemispheres, Comm. Pure Appl. Math., 23 (1970), pp. 529-568.

[7] L. Bittner, Einige kontinuierliche Analogien von Iterationsverfahren, in Funktionalanalysis, Approximationstheorie Numerische Mathematik, ISNM7, pp. 114-135, Birkhauser-Verlag, Basel, 1967.

[8] P.T. Boggs, The solution of nonlinear operator equations by A-stable integration techniques, Ph.D. Thesis, Cornell University, 1970.

[9] P.T. Boggs, The solution of nonlinear systems of equations by A-stable integration techniques, SIAM J. Numer. Anal., 8 (1971), pp. 767-785.

[10] W.E. Bosarge, Iterative continuation and the solution of nonlinear two point boundary value problems, Numer. Math., 17 (1971), pp. 268-283.

[11] F.H. Branin Jr., Widely convergent method for finding multiple solutions of simultaneous nonlinear equations, IBM J. Res. Develop. 16 (1972), pp. 504-522.

[12] R.P. Brent, Some efficient algorithms for solving systems of nonlinear equations, SIAM J. Numer. Anal., 10 (1973), pp. 327-344.

[13] K.M. Brown, A quadratically convergent Newton-like method based on Gaussian elimination, SIAM J. Numer. Anal., 6 (1969), pp. 560-569.

[14] K.M. Brown and W.B. Gearhardt, Deflation techniques for the calculation of further solutions of a nonlinear system, Numer. Math., 16 (1971), pp. 334-342.

[15] C.G. Broyden, A new method of solving nonlinear simultaneous equations, Comput. J., 12 (1969), pp. 94-99.

[16] K.-S. Chao, D.-K. Liu, and C.-T. Pan, A systematic search method for obtaining multiple solutions of simultaneous nonlinear euqations, IEEE Transactions on Circuits and Systems, Vol. CAS-22, 9 (1975), pp. 748-753.

[17] M.G. Crandall and P.H. Rabinowitz, Bifurcation from simple eigenvalues, J. Functional Anal. 8 (1971), pp. 321-340.

[18] G.G. Dahlquist, A special stability problem for linear multistep methods, B.I.T., 3 (1963), pp. 27-43.

[19] D.F. Davidenko, On a new method of numerical solution of systems of nonlinear equations, Dokl. Akad. Nauk, SSSR (N.S.)., 88 (1953), pp. 601-604, (Russian).

[20] D.F. Davidenko, On the approximate solution of a system of nonlinear equations, Ukrain. Mat. Z., 5 (1953), pp. 196-206, (Russian).

[21] J. Davis, The solution of nonlinear operator equations with critical points, Tech. Rep. No. 25, Department of Mathematics, Oregon State University, Corvallis, Oregon, 1966.

[22] F.H. Deist and L. Sefor, Solution of systems of nonlinear equations by parameter variation, Comput. J., 10 (1967), pp. 78-82.

[23] P. Deuflhard, A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting, Numer. Math., 22 (1974), pp. 289-315.

[24] P. Deuflhard, H.-J. Pesch and P. Rentrop, A modified continuation method for the numerical solution of nonlinear two-point boundary value problems by shooting techniques, Numer. Math., 26 (1976), pp. 327-343.

[25] F. Ficken, The continuation method for functional equations, Comm. Pure Appl. Math., 4 (1951), pp. 435-456.

[26] L. Fox, Numerical Solution of Ordinary and Partial Differential Equations, Pergamon Press, 1962, p. 261.

[27] F. Freudenstein and B. Roth, Numerical solution of systems of nonlinear equations, J. Assoc. Comput. Mach., 10 (1963), pp. 550-556.

[28] M.K. Gavurin, Nonlinear functional equations and continuous analogs of iterative methods, Izv. Vyss. Ucebn. Zaved. Mathematika, 6 (1958), pp. 18-31; English Transl., Tech. Rep. 68-70, Computer Science Center, University of Maryland, 1968.

[29] C.W. Gear, Numerical Initial Value Problems in Ordinary Differential Equations, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[30] C.W. Gear and K.W. Tu, The effect of variable mesh size on the stability of multistep methods, SIAM J. Numer. Anal., 11 (1974), pp. 1025-1043.

[31] C.W. Gear and D.S. Watanabe, Stability and convergence of variable order multistep methods, SIAM J. Numer. Anal., 11 (1974), pp. 1044-1058.

[32] J.K. Hale, Ordinary Differential Equations, Wiley-Interscience, N.Y., 1969.

[33] Y. Hangai and S. Kawamata, Analysis of geometrically nonlinear and stability problems by static perturbation method, University of Tokyo, Report of the Inst. of Ind. Science, Vol. 22, 5, No. 143, 1973.

[34] P. Henrici, Discrete Variable Methods for Ordinary Differential Equations, John Wiley, N.Y., 1962.

[35] A. Householder, The Theory of Matrices in Numerical Analysis, Ginn (Blaisdell), Boston, Massachusetts, 1964.

[36] H.B. Keller, Numerical solution of bifurcation and nonlinear eigenvalue problems, in Applications of Bifurcation Theory, Academic Press, New York, 1977.

[37] H.B. Keller and W.F. Langford, Iterations, perturbations and multiplicities for nonlinear bifurcation problems, Arch. Rat. Mech. Anal., 48 (1972), pp. 83-108.

[38] H.B. Keller and A.W. Wolfe, On the nonunique equilibrium states and buckling mechanism of spherical shells, J. Soc. Ind. Appl. Math., 13 (1965), pp. 674-705.

[39] R.B. Kellogg, T.Y. Liu and J. Yorke, A constructive proof of the Brower fixed-point theorem and computational results, SIAM J. Numer. Anal., 13 (1976), pp. 473-483.

[40] W. Kizner, A numerical method for finding solutions of nonlinear equations, SIAM J. Appl. Math., 12 (1964), pp. 424-428.

[41] H. Kleinmichel, Stetige Analoga und Iterationsverfahren für nichtlineare Gleichungen in Banachräumen, Math. Nachr., 37 (1968), pp. 313-344.

[42] R.W. Klopfenstein and R.S. Millman, Numerical stability of a one-evaluation predictor-corrector algorithm for numerical solution of ordinary differential equations, Math. Comp., 22 (1968), pp. 557-564.

[43] E. Lahaye, Une méthode de résolution d'une catégorie d'équations transcendantes, C.R. Acad. Sci. Paris, 198 (1934), pp. 1840-1842.

[44] E. Lahaye, Sur la représentation des racines systèms d'équations transcendantes, Deuxièmne Congrès National des Sciences, 1 (1935), pp. 141-146.

[45] J.P. Lasalle and S. Lefshetz, Stability by Liapunov's Direct Method with Applications, Academic Press, N.Y., 1961.

[46] J.D. Lawson, An order five Runge-Kutta process with extended region of stability, SIAM J. Numer. Anal., 3 (1966), pp. 593-597.

[47] K. Levenberg, A method for the solution of certain nonlinear problems in least squares, Quart. Appl. Math., 2 (1944), pp. 164-168.

[48] D. Marquardt, An algorithm for least squares estimation of non-linear parameters, SIAM J. Appl. Math., 11 (1963), pp. 431-441.

[49] R. Menzel and H. Schwetlick, Zur Behandlung von Singularitäten bei Einbettungs algorithmen, manuscript, Mathematics Dept., Dresden Technical University, 1975.

[50] G.H. Meyer, On solving nonlinear equations with a one-parameter operator embedding, SIAM J. Numer. Anal., 4 (1968), pp. 739-752.

[51] J.J. Moré and M.Y. Cosnard, Numerical comparison of three nonlinear equation solvers, Rep. TM-286, Argonne Nat. Laboratory, 1976.

[52] A. Nordsieck, On the numerical integration of ordinary differential equations, Math. Comp. 16 (1962), pp. 22-49.

[53] J. Ortega and W. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, N.Y., 1970.

[54] J. Ortega and M. Rockoff, Nonlinear difference equations and Gauss-Seidel type iterative methods, SIAM J. Numer. Anal., 3 (1966), pp. 497-513.

[55] M.R. Osborne, A new method for the solution of eigenvalue problems, Computer J., 7 (1964), pp. 228-232.

[56] M.R. Osborne, Numerical methods for hydrodynamic stability problems, SIAM J. Appl. Math., 15 (1967), pp. 539-557.

[57] M.R. Osborne and S. Michaelson, The numerical solution of eigenvalue problems in which the eigenvalue appears nonlinearly, with an application to differential equations, Computer J., 7 (1964), pp. 66-71.

[58] A. Ostrowski, Solution of Equations and Systems of Equations, Academic Press, N.Y., (second edition), 1966.

[59] C.C. Paige and M.A. Saunders, Solution of sparse indefinite systems of equations and least squares problems, Tech. Rep. CS-73-399, Computer Science Dept., Stanford University, 1973.

[60] W.C. Rheinboldt, Local mapping relations and global implicit function theorems, Trans. Amer. Math. Soc., 138 (1969), pp. 183-198.

[61] W.C. Rheinboldt, An adaptive continuation process for solving systems of nonlinear equations, Tech. Rep. TR-393, Computer Science Center, University of Maryland, 1975.

[62] W.C. Rheinboldt, Large displacements in truss-structures, (Working Manuscript, Computer Science Center, University of Maryland, February, 1976).

[63] W.C. Rheinboldt, Numerical continuation methods for finite element applications, Tech. Rep. TR-454, Computer Science Center, University of Maryland, 1976.

[64] W.C. Rheinboldt, Numerical methods for a class of finite dimensional bifurcation problems, Tech. Rep. TR-490, Computer Science Center, University of Maryland, 1976.

[65] W.C. Rheinboldt, On parameter adaption, (Working manuscript, Computer Science Center, University of Maryland, November 1976).

[66] E. Riks, The application of Newton's method to the problem of elastic stability, J. Appl. Mech., Dec. 1972, pp. 1060-1065.

[67] A. Ruhe, Algorithms for the nonlinear eigenvalue problem, SIAM J. Numer. Anal., 10 (1973), pp. 674-689.

[68] V.E. Shamanskii, A modification of Newton's method, Ukrain. Mat. Zh., 19 (1967), pp. 133-138, (Russian).

[69] R.B. Simpson, A method for the numerical determination of bifurcation states of nonlinear systems of equations, SIAM J. Numer. Anal., 12 (1975), pp. 439-451.

[70] J.F. Traub, Iterative Methods for the Solution of Equations, Prentice-Hall, Englewood Cliffs, N.J., 1971, p. 224.

[71] R.G. Voigt, Rates of convergence for a class of iterative procedures, SIAM J. Numer. Anal., 8 (1971), pp. 127-134.

[72] M.N. Yakovlev, On some methods of solving nonlinear equations, Trudy Mat. Inst. Steklov., 84 (1965), pp. 8-40; English Transl. Tech. Rep. 68-75, Computer Science Center, University of Maryland, 1968.

[73] J.H. Wilkinson, Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, N.J., 1963.