THE NUMERICAL SOLUTION OF NONLINEAR PROBLEMS

Edited by
Christopher T.H. Baker
and Chris Phillips

THE NUMERICAL SOLUTION OF NONLINEAR PROBLEMS

EDITED BY

CHRISTOPHER T. H. BAKER AND CHRIS PHILLIPS

CLARENDON PRESS · OXFORD

1981

Oxford University Press, Walton Street, Oxford OX2 6DP
London Glasgow New York Toronto
Delhi Bombay Calcutta Madras Karachi
Kuala Lumpur Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Wellington
and associate companies in
Beirut Berlin Ibadan Mexico City

Published in the United States by
Oxford University Press, New York

British Library Cataloguing in Publication Data

Baker, Christopher T. H.
The numerical solution to nonlinear problems.
1. Differential equations, Nonlinear—Numerical solutions 2. Boundary value problems—Numerical solutions
I. Title
515.3'55 QA371

ISBN 0-19-853354-3

Printed in Great Britain by Thomson Litho Ltd., East Kilbride, Scotland

PREFACE

The present volume is based upon the lectures presented at a Summer School held in Liverpool in July 1980, the fifth in a sequence of such schools organized jointly by the Department of Computational and Statistical Science at Liverpool University and the Department of Mathematics at the University of Manchester with the help and participation of a number of external speakers.

The proceedings of the previous four summer schools have been published by Oxford University Press and it is a pleasure for the editors to thank the Press for their continued interest and assistance. The present volume should be read in conjunction with its companion volumes from earlier summer schools. It has been our intention to combine the various contributions into a coherent whole, a delicate task which has been made easier by the willing co-operation of the individual contributors, who were:

I. Barrodale (Victoria), E. Bohl (Konstanz), D. Kershaw (Lancaster), A. Spence (Bath), M.R. Osborne (Canberra), C. Phillips (Hull); C.T.H. Baker, T.L. Freeman, I. Gladwell, G. Hall, Joan E. Walsh, J. Williams (Manchester); R. Cook, L.M. Delves, L.E. Scales, R. Wait, J.M. Watt (Liverpool).

The editors have endeavoured to adopt a fairly uniform house-style, but to relax it where slavish adherence would cause unnecessary complications, and they wish to thank the individual contributors for the thought they have given to accommodating the suggestions of the editors. As a safeguard, final (proof) versions have been referred to contributors.

Finally, the editors dedicate their efforts in producing this volume to their respective wives, whose support and ecouragement they value and appreciate.

Manchester
and Hull, 1981.

Christopher T.H. Baker
Chris Phillips

CONTENTS

PART ONE

ALGEBRAIC AND TRANSCENDENTAL EQUATIONS

INTRODUCTION TO NONLINEAR ALGEBRAIC EQUATIONS

Contributed by T.L. Freeman and J.E. Walsh

1.1. TYPES OF PROBLEM

Nonlinear algebraic equations arise in many different contexts in numerical analysis, the most important being the discretization of nonlinear operator equations, the algebraic eigenvalue problem, and curve fitting with nonlinear parameters. Let us consider the general system

$$f_i(\mathrm{x}) \equiv f_i(x_1, x_2, x_3, \ldots, x_n) = 0, \quad i = 1, 2, \ldots, m, \tag{1.1}$$

which may be written in vector notation as

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

If $m = n = 1$, we have a single equation in a single variable; important examples of this case are the eigenvalue problem and general polynomial equations in one variable, which will be treated in later chapters. The problem of curve fitting leads to a system with $m > n$, where it is not possible to satisfy the equations exactly, and we define the solution as the value of x which minimizes some function of the residuals. This case will also be discussed later. In the present chapter, we shall consider systems for which $m = n$ and $n > 1$, and we shall assume that the variables and functions take real values only.

When the nonlinear system arises from the discretization of an operator equation, the functions $f_i(\mathrm{x})$ generally have a special structure, which depends on the form of the operator and on the type of discretization. An example of a structure which arises frequently is the *band* system, which is an extension of the idea of a band matrix in linear equations. The system (1.2) is said to have a band form if the ith function $f_i(\mathrm{x})$ depends only on a limited number of variables lying within some band. This can be stated as

$$\frac{\partial f_i}{\partial x_j} \equiv 0, \quad \text{for} \quad |j - i| > k, \tag{1.3}$$

where $k < n$. The system is *full* if no such restriction applies.

Systems of band form arise from boundary-value problems in ordinary differential equations when we use local methods of discretization, such as finite differences or finite elements. When global discretization

is used the solution is approximated by a single expression over the whole region, and the resulting system will generally be full. For elliptic partial differential equations with local approximation of the solution, we obtain band systems which are also sparse, i.e. in addition to (1.3) we have $\frac{\partial f_i}{\partial x_j} \equiv 0$ for most of the variables x_j within the band. The systems which arise in solving elliptic problems are generally of high order, and it is important to take advantage of the sparsity and the band structure in devising efficient methods of solution.

Another type of structure which is significant in numerical work is the "almost linear" system. Suppose

$$f(x) \equiv Ax - p(x) = 0, \tag{1.4}$$

where A is a non-singular $n \times n$ matrix, and $p(x)$ is a nonlinear function. If $p(x)$ is small compared with Ax we can regard it as a correction term, and we derive a simple iterative scheme for solving (1.4) as follows

$$Ax^{[k+1]} = p(x^{[k]}), \quad k = 0,1,2,\ldots . \tag{1.5}$$

The iteration starts with a suitable estimate $x^{[0]}$, and the successive iterates $x^{[1]}$, $x^{[2]}$ ... are obtained by solving a sequence of linear problems. If $x^{[k]}$ converges to some limit, this clearly gives a solution of (1.4). Another example of an almost linear system is

$$f(x) \equiv [A(x)]x - b = 0, \tag{1.6}$$

where $A(x)$ is an $n \times n$ matrix whose elements have a 'small' dependence on x. This form suggests the iterative scheme

$$[A(x^{[k]})]x^{[k+1]} = b, \tag{1.7}$$

which is slightly less easy to apply than (1.5), because the matrix changes at each step.

To investigate the behaviour of a general iteration of this type, suppose the system $f(x) = 0$ can be written in the equivalent form

$$x = g(x), \tag{1.8}$$

which has the same roots as the original equations. The iterative method based on (1.8) is

$$x^{[k+1]} = g(x^{[k]}), \quad k = 0,1,2,\ldots, \tag{1.9}$$

which includes (1.5) and (1.7) as special cases. Clearly the function $g(x)$ must be chosen suitably, so that the calculation of each step is

reasonably simple and the iteration converges rapidly. We consider the general conditions for convergence in the next section, but first we need to define the basic vector and matrix norms which will be used.

The vector norm is a measure of the size of a vector, and it enables us to give a precise meaning to the rate of convergence of a vector iteration. The two most important examples of a vector norm are

$$\begin{aligned} &\text{infinity norm:} && \|x\|_\infty = \max_i |x_i|, \\ &\ell_2 \text{ norm:} && \|x\|_2 = \{\sum_i x_i^2\}^{\frac{1}{2}}. \end{aligned} \tag{1.10}$$

To obtain a corresponding matrix norm, we take

$$\|A\| = \max_{\|x\| \neq 0} \{\|Ax\| / \|x\|\}. \tag{1.11}$$

It is easily shown that

$$\|A\|_\infty = \max_i \{\sum_j |A_{ij}|\}, \qquad \|A\|_2 = \max_i |\lambda_i(A^T A)|^{\frac{1}{2}}, \tag{1.12}$$

where λ_i is an eigenvalue of $A^T A$. Other examples of norms, and the relations between them, are discussed by Stewart (1973).

1.2. CONTRACTION MAPPING THEOREM

For a general nonlinear system it is not usually possible to prove the existence of a solution *a priori*, and if a solution does exist, it may not be unique. The iterative method (1.9) is important for the theory as well as for practical applications, because in suitable cases it can be used to establish the existence and uniqueness of a solution in a specified region of the space $\mathbb{R}^n$.

For limited classes of problems we can establish the "global" convergence of certain iterative methods (Section 1.8), but in general we have to consider the iteration within a restricted region D_0 say, where the function $f(x)$ is defined and the successive iterates $g(x^{[k]})$ can be computed. For nonlinear iterations the main problem often lies in finding a suitable region D_0 in the neighbourhood of the solution, and this problem will recur frequently in later chapters. (For linear iterations, by contrast, the iterative methods are always globally convergent if they converge at all, and the choice of starting points is not important.)

Suppose the vector $\underset{\sim}{\alpha}$ is a solution of the equations $f(x) = 0$. Then it also satisfies the equivalent system (1.8), and from (1.9) we have

$$x^{[k+1]} - \underset{\sim}{\alpha} = g(x^{[k]}) - g(\underset{\sim}{\alpha}). \tag{1.13}$$

The left-hand side is the error after the $(k+1)$th step, $e^{[k+1]}$ say, and the iteration converges if $\|e^{[k+1]}\| \to 0$ as k increases. The basic theorems on convergence require that the function $g(x)$ has a *contraction* property, defined in (1.14) below. At this stage we do not need to assume that $g(x)$ is differentiable.

THEOREM 1.1. Contraction Mapping: Suppose that D_0 is a set of points in $\mathbb{R}^n$ and that $g(x)$ lies in D_0 for any x in D_0. If there exists a constant λ such that

$$\|g(x) - g(y)\| \leq \lambda\|x - y\|, \qquad \lambda < 1, \tag{1.14}$$

for all x, y in D_0, then the iteration (1.9) converges to a unique fixed point for any $x^{[0]}$ lying in D_0.

The proof is given by Ortega and Rheinboldt (1970, p.120).

THEOREM 1.2. Error of the Iterates: Under the conditions of Theorem 1.1, the following error estimate holds

$$\|e^{[k]}\| = \|x^{[k]} - \underset{\sim}{\alpha}\| \leq \frac{\lambda}{1-\lambda}\|x^{[k]} - x^{[k-1]}\|. \tag{1.15}$$

The proof is in Ortega and Rheinboldt (1970, p.385).

In practice it may not be easy to verify that $g(x)$ lies in D_0 for any x in D_0 and an alternative condition may be used. Given $x^{[0]}$ in D_0, let $x^{[1]} = g(x^{[0]})$ and consider the sphere

$$S = \{z: \|z - x^{[1]}\| \leq \frac{\lambda}{1-\lambda}\|x^{[1]} - x^{[0]}\|\}. \tag{1.16}$$

Then if S lies in D_0 and (1.14) holds, the results of Theorems 1.1 and 1.2 follow (Collatz, 1966, p.214).

From (1.15) it is easy to show that

$$\|e^{[k]}\| \leq \frac{\lambda^k}{1-\lambda}\|x^{[1]} - x^{[0]}\|, \tag{1.17}$$

so the rate of convergence of the iteration is at least geometric. However, this result does not give a good error bound unless we have a good estimate of λ. If $g(x)$ is strongly nonlinear, the bound in (1.17) is likely to be a considerable over-estimate of the error unless D_0 is very small.

When $g(x)$ is differentiable in D_0, we can express λ in terms of partial derivatives. The Jacobian of $g(x)$ is the $n \times n$ matrix

$$J_g(x) = \left[\frac{\partial g_i}{\partial x_j}\right]. \tag{1.18}$$

From (1.14) and the mean-value theorem, we can obtain a sufficient condition for convergence in the form

$$\lambda = \max_{x \in D_0} \|J_g(x)\| < 1. \tag{1.19}$$

In general, the iteration (1.9) will not be implemented exactly. This is because any numerical calculation is subject to rounding error, and also because the equations for $x^{[k+1]}$ may not be solved exactly. In the examples (1.5) and (1.7), the new iterate $x^{[k+1]}$ is obtained from a system of linear equations, and the latter system is often solved by another iteration, making an inner loop in the main calculation. To save computation time, we do not want to obtain the intermediate values $x^{[k]}$ to high accuracy, and so we need to estimate how much error can be allowed at each iterative step without affecting the overall convergence. A simple result on the effect of perturbations is given by the following theorem.

THEOREM 1.3. Perturbed Iterations: Suppose that $g(x)$ satisfies the conditions of Theorem 1.1, and that $\bar{g}(x)$ is a perturbation of $g(x)$ such that

$$\|\bar{g}(x) - g(x)\| \leq \varepsilon, \quad x \text{ in } D_0. \tag{1.20}$$

Let $\bar{x}^{[1]} = \bar{g}(x^{[0]})$, and let $\bar{S}$ be the sphere

$$\bar{S} = \{z: \|z - \bar{x}^{[1]}\| \leq \frac{\lambda}{1-\lambda}\|\bar{x}^{[1]} - x^{[0]}\| + \frac{2\varepsilon}{1-\lambda}\}. \tag{1.21}$$

Then if $\bar{S}$ lies in D_0 the iterates $\bar{x}^{[k+1]} = \bar{g}(\bar{x}^{[k]})$ remain in $\bar{S}$, and

$$\|\bar{x}^{[k]} - x^{[k]}\| \leq \frac{\varepsilon}{1-\lambda}. \tag{1.22}$$

The proof is given by Collatz (1966, p.218).

We note that $\bar{g}(x)$ is not necessarily a continuous perturbation of $g(x)$. Since the unperturbed iteration is convergent, we see that the perturbed iterates will converge to the true solution up to a certain accuracy, after which we must reduce ε to obtain any further improvement.

All the results above assume that the conditions hold in a region D_0 containing the solution $\underset{\sim}{\alpha}$. The set of all points for which the iteration (1.9) converges to $\underset{\sim}{\alpha}$ is called the *domain of attraction* of $\underset{\sim}{\alpha}$. For certain solutions there may be no domain of attraction (other than the point itself); in such cases the iteration is unstable in the neighbourhood of the solution, and either diverges or converges to some

more distant point.

1.3. NEWTON'S METHOD

We now consider the construction of suitable iteration functions $g(x)$ for systems which do not have any special features such as those of (1.4) and (1.6). We return to the original equation $f(x) = 0$, and suppose that $f(x)$ has continuous second partial derivatives in some convex region D_0. By Taylor expansion about the point $x^{[k]}$ we have

$$f(x) = f(x^{[k]}) + J(x^{[k]})\,(x - x^{[k]}) + O(\|x - x^{[k]}\|^2)\,, \tag{1.23}$$

where $J(x)$ is the Jacobian matrix of $f(x)$. (Note that $J(x)$, without subscript, denotes the Jacobian of the original function $f(x)$ throughout.) Suppose $x^{[k]}$ is close to $\underset{\sim}{\alpha}$, then putting $x = \underset{\sim}{\alpha}$ gives the approximate equation

$$J(x^{[k]})(\underset{\sim}{\alpha} - x^{[k]}) \simeq -f(x^{[k]}). \tag{1.24}$$

This leads us to define the basic Newton iteration for solving $f(x) = 0$,

$$x^{[k+1]} = x^{[k]} - [J(x^{[k]})]^{-1} f(x^{[k]}), \quad k = 0, 1, 2, \ldots . \tag{1.25}$$

To use the method in this form, $J(x^{[k]})$ must be non-singular for all $x^{[k]}$. If $x^{[0]}$ is in the neighbourhood of the root, it is fairly easy to prove convergence with some additional conditions, but the question of the behaviour of the iteration from a general starting point is much more difficult.

THEOREM 1.4. Convergence of Newton's Method: If the convex region D_0 contains a solution $\underset{\sim}{\alpha}$ of $f(x) = 0$, and if $[J(x)]^{-1}$ and the second derivatives of $f(x)$ exist and are bounded in D_0, the Newton iteration converges quadratically for $x^{[k]}$ sufficiently close to $\underset{\sim}{\alpha}$, that is,

$$\|x^{[k+1]} - \underset{\sim}{\alpha}\| = O(\|x^{[k]} - \underset{\sim}{\alpha}\|^2)\,. \tag{1.26}$$

The proof is given by Collatz (1966, p.292).

This result is about the local convergence of (1.25), assuming that the root is known to exist. If we want to prove the existence of a root in D_0, more conditions on $f(x)$ are required.

Sufficient conditions for the Newton iteration (1.25) to converge to a solution $\underset{\sim}{\alpha}$ of (1.2) are given by the Newton-Kantorovich theorem. We state the version of the theorem given in Ortega (1972). This is not as general as the version in Chapter 3, but will adequately illustrate

the conditions of the theorem.

THEOREM 1.5 (Ortega (1972), p.155). Assume that $f: \mathbb{R}^n \to \mathbb{R}^n$ is differentiable on a convex set D_0 and that

$$\|J(x) - J(y)\| \leq \gamma\|x-y\| \tag{1.27a}$$

for all $x, y \in D_0$. Suppose that there is an $x^{[0]} \in D_0$ such that

$$\|[J(x^{[0]})]^{-1}\| \leq \beta \tag{1.27b}$$

$$\|[J(x^{[0]})]^{-1} f(x^{[0]})\| \leq \eta \tag{1.27c}$$

and

$$\theta = \beta\gamma\eta < \tfrac{1}{2}.$$

Assume that

$$S \equiv \{x: \|x-x^{[0]}\| \leq t^*\} \subset D_0$$

$$t^* = \frac{1}{\beta\gamma}\,(1-(1-2\theta)^{\frac{1}{2}}).$$

Then the iterates $x^{[k+1]}$, $k = 0,1,\ldots,$ given by (1.25), are well defined and converge to a solution $\underset{\sim}{\alpha}$ of (1.2) in S.

Proof. We reproduce the proof of Ortega (1972), but omit some of the details by referring to this work. We have

$$\|J(x)-J(x^{[0]})\| \leq \gamma\|x-x^{[0]}\| \leq \gamma t^* < 1/\beta,$$

for any $x \in S$. Hence, by the Banach lemma (Ortega (1972), p.32), $J(x)$ is nonsingular, and

$$\|J(x)^{-1}\| \leq \frac{\beta}{1-\beta\gamma\,\|x-x^{[0]}\|}, \quad x \in S.$$

Consequently, the Newton function

$$g(x) = x - [J(x)]^{-1} f(x)$$

is well-defined on S and if $x, g(x) \in S$, then

$$\|g(g(x))-g(x)\| = \|[J(g(x))]^{-1} f(g(x))\| \leq \frac{\beta\|f(g(x))\|}{1-\beta\gamma\,\|x^{[0]}-g(x)\|}, \quad x, g(x) \in S.$$

But

$$\|f(g(x))\| = \|f(g(x))-f(x)-J(x)(g(x)-x)\| \leq \tfrac{1}{2}\gamma\,\|g(x)-x\|^2,$$

where the inequality follows from the Lipschitz continuity of $J(x)$ (see

Ortega (1972), section 8.1.5).

Hence,

$$\|g(g(x))-g(x)\| \leq \frac{\beta\gamma\|g(x)-x\|^2}{2(1-\beta\gamma\|x^{[0]}-g(x)\|)} . \tag{1.28}$$

We now define the scalar iteration

$$t^{[k+1]} = t^{[k]} - \frac{\frac{1}{2}\beta\gamma t^{[k]2} - t^{[k]} + \eta}{\beta\gamma t^{[k]} - 1}, \qquad k = 0,1,\ldots,$$

$$t^{[0]} = 0.$$

It can easily be shown that this scalar iteration is Newton's method for the quadratic $q(t) \equiv \frac{1}{2}\beta\gamma t^2 - t + \eta$, whose smaller root is t^*.

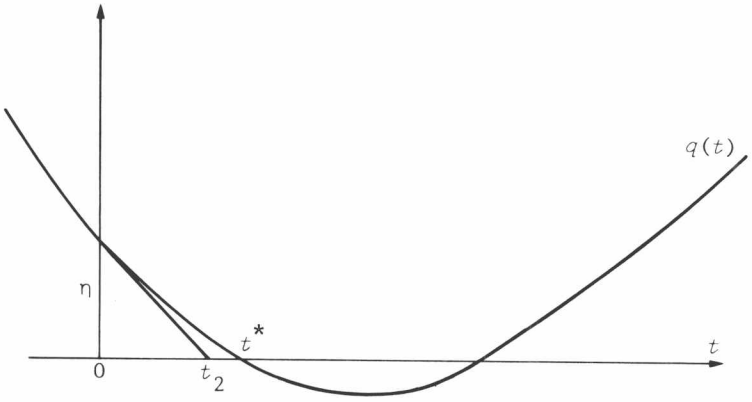

Figure 1

It follows from this observation that the sequence $\{t^{[k]}\}$ is well-defined, monotonically increasing, and converges to t^*.

The significance of the sequence $\{t^{[k]}\}$ is that it is a majorising sequence for the sequence $\{x^{[k]}\}$:

$$\|x^{[k+1]} - x^{[k]}\| \leq t^{[k+1]} - t^{[k]}, \qquad k = 0,1,\ldots .$$

This result can be proved inductively using inequality (1.28) (see Ortega (1972)); by a simultaneous induction it can be shown that $x^{[k+1]} \in S$, $k = 0,1,\ldots$. Hence, for any $k \geq 1$ and $p \geq 1$,

$$\|x^{[k+p]} - x^{[k]}\| \leq \sum_{i=1}^{p} (t^{[k+i]} - t^{[k+i-1]}) = t^{[k+p]} - t^{[k]},$$

and since $t^{[k]} \to t^*$ as $k \to \infty$, $\{t^{[k]}\}$ is a Cauchy sequence. Hence $\{x^{[k]}\}$ is a Cauchy sequence which, by the closure of S has a limit $\underset{\sim}{\alpha} \in S$.

But

$$\|f(x^{[k]})\| = \|J(x^{[k]})(x^{[k+1]}-x^{[k]})\| \leq \|J(x^{[k]})\| \, \|x^{[k+1]}-x^{[k]}\|,$$

and since $\|x^{[k+1]}-x^{[k]}\| \to 0$ as $k \to \infty$, and $J(x)$ is bounded for $x \in S$, it follows that

$$f(x^{[k]}) \to 0 \quad \text{as} \quad k \to \infty.$$

Finally the continuity of f implies that $f(\underset{\sim}{\alpha}) = 0$.

It is interesting to consider the significance of the conditions of the theorem. The Lipschitz continuity condition (1.27a) is simply a smoothness condition on the function $f(x)$ on the convex set D_0. This smoothness condition, together with the bound (1.27b), guarantees that the Jacobian is nonsingular on the set S and hence that the Newton iteration is well-defined on this set. Finally condition (1.27c), which can be rewritten as

$$\|x^{[1]} - x^{[0]}\| \leq \eta,$$

ensures that the second iterate $x^{[1]} \in S$, hence starting the inductive proof of the theorem.

1.4. DAMPED NEWTON METHOD

Theorem 1.5 provides sufficient, but not necessary, conditions for the convergence of Newton's method. Even when the conditions are not satisfied, which in practice is often the case, the method may nonetheless converge. The probability of convergence from starting points not satisfying the conditions of Theorem 1.5 is improved if a line search is incorporated in Newton's method. This line search ensures that each iteration of the resulting damped Newton method reduces some norm of the function $f(x)$. The method is given by

$$x^{[k+1]} = x^{[k]} + r^{[k]} p^{[k]}, \quad k = 0,1,\ldots, \tag{1.29}$$

where

$$J^{[k]} p^{[k]} = -f^{[k]}, \tag{1.30}$$

and $r^{[k]}$ is chosen so that, for example,