The Numerical Solution of Algebraic Equations

R. Wait

The Numerical Solution of Algebraic Equations

R. Wait

Department of Computational and Statistical Science
University of Liverpool

A Wiley–Interscience Publication

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto

Library of Congress Cataloging in Publication Data:

Wait, R.
The numerical solution of algebraic equations.

'A Wiley–Interscience publication.'
Includes index.
1. Equations–Numerical solutions. I. Title.
QA218.W34 519.4 78-21869
ISBN 0 471 99755 2

Typeset by Preface Ltd., Salisbury, Wilts.
and printed in Great Britain by
Unwin Brothers Ltd., Woking, Surrey.

Preface

This book is intended as a half-year course for undergraduate students of mathematical science and is based on a one semester course given to second-year students at the University of Liverpool. The object of such a course is to provide guidance to those using or writing computer programs or intending to do so. At the same time it is an attempt to provide an insight into the mathematical analysis for those students with no more than a passing interest in the computational aspects of the methods. The individual sections are roughly equal in length and content, and they could each be reasonably covered in a single lecture. The sections can, in most cases, be classified as either theoretical or practical and there are approximately equal numbers of each.

The book is in two parts and, with one or two exceptions, there is no specific cross referencing, so the two parts can be read in either order. There are a few sections towards the end of Part II which, in view of their somewhat limited appeal, can be safely omitted.

In Part II special methods for polynomial equations are omitted as they have been dealt with effectively elsewhere and it would be difficult to improve on the presentation in, for example, Householder (1970). In addition, the usefulness of such methods in practice is highly questionable, since a polynomial equation can be formulated as a matrix eigenproblem and then solved by the QR algorithm. In the solution of arbitrary non-linear problems, there is no attempt to emulate the breadth and rigour of advanced texts such as Traub (1964), Ostrowski (1966), Ortega and Rheinboldt (1970) or Rheinboldt (1974). There is, however, an attempt to provide a thorough introduction to the theory of iterative methods such as cannot be found in general introductory texts on numerical analysis, which usually contain a very sketchy treatment of this area. At the same time numerous exercises and worked examples illustrate the directions in which the analysis proceeds in more advanced situations. It is intended herein to go beyond the ubiquitous Newton's method and give practical methods that work in practical problems, together with some indication of the relevance of current theory to current practice.

A knowledge of the calculus as far as Taylor series expansions is assumed, as is a degree of familiarity with the concepts of vectors and matrix norms. In order to avoid the book becoming a mere catalogue of all the available methods it has been necessary in certain instances to omit any discussion of some of the alternatives in order to concentrate on one of the best. Thus there is no explanation of Householder transformations because there is a careful assessment of Givens transformations: in

particular there is one of the first accounts at this level, of the square-root-free Givens transformations. No attempt has been made to discuss the associated subjects of matrix eigenproblems, linear programming and optimization, although their connexions with the solution of algebraic equations is indicated where appropriate.

This is one of the first introductory text books (as opposed to a specialist monograph or the proceedings of a conference) to include a discussion of many of the practical difficulties to be encountered when solving systems of equations and to describe some of the possible methods of overcoming them. In particular, there is a discussion of sparse-matrix techniques for linear systems (the cases of an irregular pattern and of a regular structure are both covered), Brown's Newton–Gauss–Seidel type of method for systems, quasi-Newton methods for non-linear systems and orthogonal factorization of matrices, together with a description of their practical application in the numerical solution of equations. Robust methods for single equations are covered; in particular the Pegasus method appears in a textbook for the first time. All these methods are used by practical people to solve practical problems and students should therefore be aware of them at an early stage. Such methods should be included in any book that claims to show how to solve problems; up to the present these methods have in general been restricted to research papers in the scientific journals and an elementary exposition is long overdue.

It is to be expected that practising scientists and engineers will find this book of value as an introduction to the modern methods that are available. It would be possible for them to proceed to a more detailed account by following up the references in the extensive bibliography provided. It should also prove to be a suitable textbook for many of the short courses in numerical methods that now form part of a typical course for engineering students. On the other hand, students with a more mathematical bias should find sufficient material to stimulate their particular interests in what is, without doubt, one of the key foundations of numerical mathematics, on which other areas – numerical solution of differential equations for example – can build.

I would like to acknowledge the assistance that I received from colleagues and students, past and present, in preparing the manuscript of this book. In particular J. M. Watt and N. G. Brown deserve a special mention. In addition I would like to thank the secretaries of the department of Computational and Statistical Science at the University of Liverpool for their painstaking typing of large parts of the manuscript.

Liverpool R. Wait
July 1978

Contents

PART II NON-LINEAR EQUATIONS

*Can be omitted on first reading.

1

Introduction

1.1 Linear Equations: An Introduction

One of the most fundamental problems of numerical analysis is the solution of a system of linear algebraic equations. Another cornerstone of the subject – the solution of non-linear equations – is also dealt with.

The numerical solution of linear algebraic systems is the object of a considerable amount of ongoing study, primarily in the solution of large sparse systems (see for example Rose and Willoughby, 1972; Bunch and Rose, 1976) and in the production of computer library packages, of which the most widely used are the I.M.S.L. Library, the FUNPACK, EISPACK, and LINPACK routines of the Argonne Laboratory, and the N.A.G. Library. This book will however be concerned only with a few of what might reasonably be termed *basic* methods and as much (or as little) theory as is considered appropriate. There is no attempt to provide an exhaustive catalogue of methods – merely enough to suggest the breadth of the subject. Whilst it is intended, as far as possible, to provide efficient and up-to-date methods, it should be remarked that it is on the computer implementation that such judgements ought to be made, and so without lengthy computer programs any assessments may of necessity be extremely vague and have limited validity. The important distinction between the mathematical statement of a method and the computer implementation should never be overlooked, and it is no coincidence that it is discussed again.

Classification

There are three largely distinct groups of methods:

(1) direct methods;
(2) iterative methods;
(3) other types.

The third group consists of such methods as *conjugate gradients* – see for

example Engeli, Ginsburg, Rutishauser and Stiefel (1959) for a rather elegant account, or Reid (1971b) for a more up-to-date approach. These methods have their uses in various applications but unfortunately cannot find a place in the present slim volume. The sections on direct methods are not only important as a summary of some of the basic methods for solving linear equations, but also provide an introduction to the concepts of *matrix factorization*, which has many other applications. One very important application is the solution of the *algebraic eigenvalue problem*, viz. for any matrix A determine non-trivial x and λ such that

$$Ax = \lambda x.$$

Several excellent texts are available that consider this problem – for example Wilkinson (1965a) and Gourlay and Watson (1973). The section on iterative methods for linear systems is delayed until a more appropriate place in Section 5.5 alongside descriptions of iterative methods in general, but it can be read out of sequence after Chapter 2.

1.2 Iterative Methods: An Introduction

Any iterative method consists of three parts:

(1) an initial estimate (or set of estimates) of the solution;
(2) a formula for updating the approximate solution;
(3) a 'fail-safe' procedure for stopping the updating process.

The different components and their relative importance are emphasized when the iterative process is represented by a flow chart such as Figure 1.

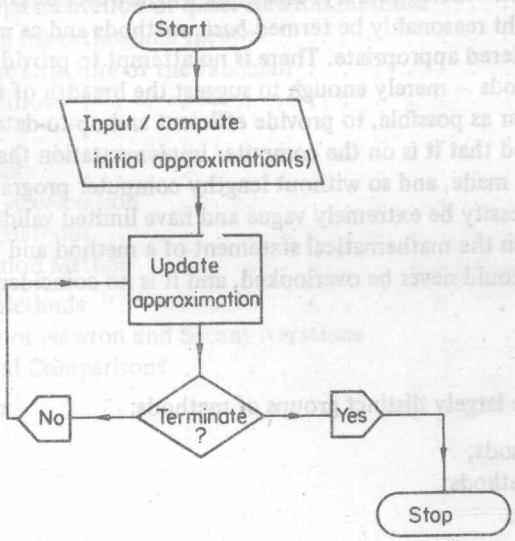

Figure 1 The iterative process

It is important to distinguish between the *iterative algorithm* – the total process from *start* to *stop* – and a single part of that process namely, the iteration formula. To be precise, mathematical (or non-mathematical) statements of (1), (2), (3) do not constitute an algorithm; it is only a *computer implementation* of the method that can justifiably be termed an algorithm. There are usually many different ways of expressing a particular mathematical formula, but they are rarely equally suitable as a basis for numerical computation.

A classical example of the difference between 'suitable' and 'unsuitable' mathematical formulae arises in computing the roots of the quadratic equation

$$ax^2 - 2bx + c = 0.$$

The roots can be expressed as either

$$\frac{b \pm \sqrt{(b^2 - ac)}}{a} \quad \text{or} \quad \frac{c}{b \mp \sqrt{(b^2 - ac)}}.$$

When $|ac| \ll b^2$, b and $\sqrt{(b^2 - ac)}$ are approximately equal in magnitude. Since subtracting nearly equal numbers invariably results in a loss of accuracy in the difference, the roots should be computed as

$$\frac{b + \sqrt{(b^2 - ac)}}{a}, \qquad \frac{c}{b + \sqrt{(b^2 - ac)}} \geqslant 0,$$

if $b < 0$, the minus sign is taken. This particular device for reducing *round-off errors* appears again in Sections 5.7 and 6.2.

The terminating condition is always of vital importance and it must anticipate all possible outcomes of the updating process. The method may not work every time and so the algorithm should be able to judge when the method has failed. Since iteration provides a sequence of approximate solutions, the algorithm will reach a successful conclusion when it adjudges the most recent approximation to be sufficiently accurate. An additional test in the algorithm should decide when the iteration has spent enough time/money even though the desired accuracy has not been attained. It should be noted that the first stage often needs to include more than a straightforward input of data, it may incorporate a test to ensure that the initial data are compatible with the update formula; for example, the iteration may require two numbers x and y such that $x \leqslant y$, in which case it is essential to check the ordering of the input data.

In general in this book, it is the updating section of the algorithm that is discussed. All the update formulae considered in Chapter 5 require one approximate solution, invariably denoted by $x^{(n)}$ ($\mathbf{x}^{(n)}$ for a system of equations), and generate a further approximation $x^{(n+1)}$ ($\mathbf{x}^{(n+1)}$). It is to be hoped that $x^{(n+1)}$ is a better approximation than $x^{(n)}$.

Examples of iteration formulae

(1) For finding $a^{1/2}$ for any positive number a; that is for solving the equation

$x^2 - a = 0$:

$$x^{(n+1)} = \frac{x^{(n)^2} + a}{2x^{(n)}}$$

or

$$x^{(n+1)} = \frac{x^{(n)}(3a - x^{(n)^2})}{2a}.$$

Note that both formulae have the solution $x^{(n)} = x^{(n+1)} = a^{1/2}$.

(2) For solving an arbitrary non-linear equation in the form $f(x) = 0$:

$$x^{(n+1)} = x^{(n)} - \left\{ \frac{df(x^{(n)})}{dx} \right\}^{-1} f(x^{(n)}). \tag{1.1}$$

(3) For solving a system of linear equations $A\mathbf{x} = \mathbf{b}$, where $A = \{a_{ij}\}$,

$$\mathbf{x} = [x_1, \ldots, x_N]^T \quad \text{and} \quad \mathbf{b} = [b_1, \ldots, b_N]^T,$$

that is the system of equations

$$\sum_{j=1}^{N} a_{ij}x_j = b_i, \qquad a_{ii} \neq 0, \qquad i = 1, \ldots, N:$$

$$\mathbf{x}^{(n+1)} = [x_1^{(n+1)}, \ldots, x_N^{(n+1)}]^T$$

is given in terms of

$$\mathbf{x}^{(n)} = [x_1^{(n)}, \ldots, x_N^{(n)}]^T$$

as

$$x_i^{(n+1)} = \frac{-1}{a_{ii}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij}x_j^{(n)} - b_i \right), \qquad i = 1, \ldots, N.$$

Exercise

1.1 Write out the formula (1.1) in the particular cases

(a) $f(x) = x^2 - a, \qquad a > 0$

(b) $f(x) = \frac{1}{x} - a.$

1.3 Algebraic Problems: An Introduction

The types of equations illustrated in the previous section arise naturally in the formulation of mathematical models. A few such models are given in this section to emphasize the importance of this area of numerical analysis.

(a) Chemical equilibrium problems

In simple models of chemical reactions, if it is assumed that all chemicals are kept well mixed, then spatial differences can be ignored. If, in addition, it is assumed that the equilibrium state has been reached, then transient effects can be ignored. The resulting conditions on the equilibrium state reduce to a set of non-linear algebraic equations.

As an example, consider the problem of finding the acidity of a solution made from a little soda lime and a lot of rhubarb leaves, that is a solution of oxalic acid $(COOH)_2$, and sodium hydroxide, NaOH. If we denote by [A] the molecular concentration of substance A, the equilibrium equations are conservation of mass and balance of electric charge in the solution.

The oxalic acid partially dissociates, so that

$$[(COOH)_2]_{total} = [(COOH)_2] + [H(COO)_2^-] + [(COO)_2^{2-}] \tag{1.2}$$

for which the dissociation constants are known to be

$$K_1 \equiv \frac{[H^+]\,[H(COO)_2^-]}{[(COOH)_2]} \tag{1.3}$$

and

$$K_2 \equiv \frac{[H^+]\,[(COO)_2^{2-}]}{[H(COO)_2^-]}. \tag{1.4}$$

If it is a weak solution, then the sodium hydroxide dissociates completely, so that

$$[NaOH]_{total} = [Na^+].$$

Since the change is balanced, it follows that

$$[Na^+] + [H^+] = [OH^-] + [H(COO)_2^-] + 2[(COO)_2^{2-}]. \tag{1.5}$$

The dissociation constant for water is known to be

$$K_3 \equiv [H^+]\,[OH^-].$$

Using (1.2)–(1.4) to eliminate the unknowns $[(COOH)_2]$, $[H(COO)_2^-]$ and $[(COO)_2^{2-}]$ from (1.5) leads to a quartic equation for $x \equiv [H^+]$ in terms of the known values

$$K_1, K_2, K_3, K_4 \equiv [(COOH)_2]_{total} \quad \text{and} \quad K_5 \equiv [NaOH]_{total}$$

in the form

$$f(x) \equiv (x^2 + xK_5 - K_3)(x^2 + xK_1 + K_1K_2) - x(K_1x + 2K_1K_2)K_4 = 0.$$

It is known that $[H^+] \geqslant 0$ and $[H^+] \leqslant (K_3)^{1/2}$, so once the values of $K_1, \ldots, K_5$ are determined it is possible to use the computer program in Section 6.4 to solve this problem.

(b) Leontieff input–output analysis

An economic model reduces to a system of algebraic equations only when the economy is assumed to be in equilibrium. The Leontieff analysis determines the level of production in an economy with many sections, each buying goods from the others in order to produce their own end products. The model could be of the national economy or of a single large industry with many divisions buying and selling amongst themselves as well as to the general public.

Let x_{ij} be the volume of sales from industry i to industry j. Let x_i be the gross output of industry i and let b_i be the external consumer demand for the products of industry i. Thus it follows that if there are N industries in the model, then

$$\sum_{j=1}^{N} x_{ij} + b_i = x_i \qquad i = 1, \ldots, N.$$

Assume that the sales from industry i to industry j are a fixed multiple of the production of industry j; then

$$x_{ij} = c_{ij} x_j, \quad i, j = 1, \ldots, N.$$

This situation would occur if industry i provided components for each product of industry j. The model then leads to

$$\sum_{j=1}^{N} a_{ij} x_j = b_i, \qquad i = 1, \ldots, N,$$

where all the

$$a_{ij} = \begin{cases} -c_{ij}, & i \neq j \\ 1 - c_{ii}, & i = j \end{cases}$$

and b_i $(i = 1, \ldots, N)$ are known.

If it is a model of a particular sector of the economy with some industries producing components for other industries, the matrix $A = \{a_{ij}\}$ is upper triangular, that is

$$a_{ij} = 0, \qquad i > j.$$

Alternatively if the model is of a very large system such as the national economy, there will be a large number of zeros corresponding to sections that have no direct trade. In this case the matrix will be large and sparse and a suitable candidate for the methods of Section 4.2.

(c) General market equilibrium

Assume that the supply and demand for each of the products $y_1, \ldots, y_N$ are not fixed but are governed by the price of the product and of its competitors. Let S_i, D_i and P_i be respectively the supply, demand and price of product y_i. If the products

can be substituted one for another to a greater or lesser extent, then the demand for product y_i can be written as

$$D_i = F_i(P_1, \ldots, P_N), \qquad i = 1, \ldots, N,$$

where F_i is a non-linear function with known (or estimated) coefficients such as

$$F_1 = a_0 - a_1 P_1 - a_2 P_1^2 + a_3 P_2 + a_4 P_2^2,$$

where $a_0, \ldots, a_4 > 0$. Supply is also a function of the prices, again not necessarily a linear function, so that

$$S_i = G_i(P_1, \ldots, P_N), \qquad i = 1, \ldots, N.$$

If the market is balanced, then

$$S_i = D_i, \qquad i = 1, \ldots, N$$

and the model can be represented as a system of non-linear equations

$$\mathrm{f}(\mathrm{x}) = 0,$$

where

$$\mathrm{x} \equiv [P_1, \ldots, P_N]^{\mathrm{T}} \qquad \text{and} \qquad \mathrm{f} \equiv [F_1 - G_1, \ldots, F_N - G_N]^{\mathrm{T}}.$$

The solution of this non-linear system is the system of prices to be charged to achieve a market in equilibrium.

(d) Electric power networks

Large sparse linear equations arise in the study of power networks. A power network contains a number of power stations connected by a network of transmission lines to a number of customers and the object of the analysis is to investigate the flow, subject to given load conditions.

Assume that there are M branches and N nodes. Along the jth branch of the network

$$\Delta p_j = k_j f_j, \qquad j = 1, \ldots, M$$

where, in electrical networks, Δp_j is the voltage drop, k_j is the resistance (ohms) and f_j is the current (amperes). If, at the ith node, the input is F_i, where $F_i < 0$ indicates a flow out of the system, then

$$\sum_j c_{ij} f_j = F_i, \qquad i = 1, \ldots, N$$

in which

$$c_{ij} = \begin{cases} 1 & \text{flow along branch } j \text{ towards node } i, \\ -1 & \text{flow along branch } j \text{ away from node } i, \\ 0 & \text{node } i \text{ not on branch } j. \end{cases}$$

The matrix $C = \{c_{ij}\}$ is the *incidence matrix* of the network and this latter can be interpreted as a *directed graph* (see Section 4.2). It follows that the pressure drops Δp_j can be written in terms of the nodal voltages P_i as

$$\sum_i c_{ij} P_i = \Delta p_j, \qquad j = 1, \ldots, M.$$

These three sets of equations can be combined as

$$A\mathbf{x} = \mathbf{b},$$

where

$$A = C \operatorname{diag}(k_j^{-1}) C^{\mathrm{T}},$$
$$\mathbf{x} = [P_1, \ldots, P_N]^{\mathrm{T}},$$
$$\mathbf{b} = [F_1, \ldots, F_N]^{\mathrm{T}}.$$

The same sets of equations can be used to analyse the flow in a grid of gas pipelines or any other flow network.

(e) Differential equations

In each of the preceding examples an equilibrium state was reached, but in many problems the transient state is of interest. Consider the photolysis of nitrogen dioxide, which is the first step in the production of photochemical smog:

$$\begin{aligned}
NO_2 + \text{sunlight} &\rightarrow NO + O \qquad (\text{reaction rate } k_1),\\
O + O_2 + N_2 &\rightarrow O_3 + N_2 \qquad (\text{reaction rate } k_2),\\
O + 2O_2 &\rightarrow O_3 + O_2 \qquad (\text{reaction rate } k_3),\\
NO + O_3 &\rightarrow NO_2 + O_2 \qquad (\text{reaction rate } k_4).
\end{aligned}$$

If the concentration of O_3 is c_1, the concentration of O is c_2, the concentration of NO_2 is c_3 and the concentration of NO is c_4, then the four reactions lead to a system of coupled differential equations. If it is assumed that the supplies of sunlight, N_2 and O_2 are sufficiently large, this system can be written as

$$\begin{aligned}
\frac{dc_1}{dt} &= (k_2 + k_3)c_2 - k_4 c_1 c_4,\\
\frac{dc_2}{dt} &= k_1 c_3 - (k_2 + k_3)c_2,\\
\frac{dc_3}{dt} &= -k_1 c_3 + k_4 c_1 c_4,\\
\frac{dc_4}{dt} &= k_1 c_3 - k_4 c_1 c_4.
\end{aligned} \tag{1.6}$$

The initial concentrations $c_1(0)$, $c_2(0)$, $c_3(0)$ and $c_4(0)$ are all known and the system can be written as

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \mathbf{F}(\mathbf{c}), \qquad \text{given } \mathbf{c}(0).$$

For some value $\Delta t > 0$, let $t_n = n\,\Delta t$ and let $\mathbf{C}_n$ be the finite-difference approximation to $\mathbf{c}(t_n)$ such that

$$\frac{1}{\Delta t}\{\mathbf{C}_{n+1} - \mathbf{C}_n\} = \tfrac{1}{2}\{\mathbf{F}(\mathbf{C}_{n+1}) + \mathbf{F}(\mathbf{C}_n)\}. \tag{1.7}$$

Proceeding from time t_n to time t_{n+1}, knowing $\mathbf{C}_n$, it is necessary to determine the $\mathbf{C}_{n+1}$ that satisfies this non-linear algebraic system. Very often the method used is a *predictor–corrector* method, which uses the formula (1.7) together with a less accurate form to provide a value for $\mathbf{C}_{n+1}$ explicitly. Such a formula is

$$\frac{1}{\Delta t}\{\mathbf{C}_{n+1} - \mathbf{C}_n\} = \mathbf{F}(\mathbf{C}_n). \tag{1.8}$$

Given $\mathbf{C}_n$, a sequence $\{\mathbf{x}^{(k)}\}$ of approximations to $\mathbf{C}_{n+1}$ is produced, where $\mathbf{x}^{(0)}$ satisfies

$$\frac{1}{\Delta t}\{\mathbf{x}^{(0)} - \mathbf{C}_n\} = \mathbf{F}(\mathbf{C}_n)$$

and where

$$\frac{1}{\Delta t}\{\mathbf{x}^{(k+1)} - \mathbf{C}_n\} = \tfrac{1}{2}\{\mathbf{F}(\mathbf{x}^{(k)}) - \mathbf{F}(\mathbf{C}_n)\}, \qquad k = 0, 1, \ldots.$$

This second equation defines an iteration of the type investigated in Chapter 5. Frequently such a simple iteration does not converge and the non-linear equation (1.7) has to be solved by alternative methods as in Section 5.8.

(f) Partial differential equations

It is probable that the largest single area in which large sparse systems of algebraic equations arises is the solution of partial differential equations. A large number of such problems involve diffusion: diffusion of heat, diffusion of pollutants in the atmosphere or in rivers, and so on. The diffusion of heat is governed by Fourier's law of heat conduction

$$\frac{\partial\theta}{\partial t} = \frac{k}{\rho c}\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2} + \frac{\partial^2\theta}{\partial z^2}\right), \tag{1.9}$$

where

θ = temperature,

k = coefficient of thermal conductivity,

c = specific heat

and ρ = density.

The coefficient $k/\rho c$ is called the thermal diffusivity and is denoted by α. Equations similar to (1.9) can be derived for other forms of diffusion. If the derivatives in (1.9) are replaced by finite differences, then it is necessary to replace the continuous function θ by a sequence of approximate values defined at mesh points as in the preceding example. An alternative approach based on finite elements also involves a grid of nodal values. In this example each time level t_n corresponds to a three-dimensional network of points. In the resulting system of linear equations that have to be solved at each time level, the structure is similar to the power network problems. The mesh points are the nodes and the positions of the branches joining near neighbours are governed by the choice of difference formula. Unlike the power networks and the economic models, the finite-difference or finite-element model has a regular structure that can be exploited in the solution procedure. For this reason it is appropriate to devote the whole of Section 4.3 to the solution of such structured problems; in addition, Section 5.5 deals with methods frequently applied to such problems.

The construction of the difference schemes themselves is beyond the scope of this book and the reader should consult a detailed text such as Mitchell (1969) or Mitchell and Wait (1977).