Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

630

Numerical Analysis

Proceedings of the Biennial Conference
Held at Dundee, June 28 – July 1, 1977

Edited by G. A. Watson

Springer-Verlag
Berlin Heidelberg New York 1978

Editor
G. A. Watson
University of Dundee
Department of Mathematics
Dundee, DD1 4HN/Scotland

AMS Subject Classifications (1970): 65-02, 65D10, 65F05, 65F20, 65K05, 65L05, 65L10, 65M99, 65N30, 65R05

ISBN 3-540-08538-6 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-08538-6 Springer-Verlag New York Heidelberg Berlin

Printed in Germany

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2141/3140-543210

Preface

For the 4 days June 28 - July 1, 1977, over 220 people attended the 7th Dundee Biennial Conference on Numerical Analysis at the University of Dundee, Scotland. The technical program consisted of 16 invited papers, and 63 short submitted papers, the contributed talks being given in 3 parallel sessions. This volume contains, in complete form, the papers given by the invited speakers, and a list of all other papers presented.

I would like to take this opportunity of thanking the speakers, including the after dinner speaker at the conference dinner, Professor D S Jones, all chairmen and participants for their contributions. I would also like to thank the many people in the Mathematics Department of this University who assisted in various ways with the preparation for, and running of, this conference. In particular, the considerable task of typing the various documents associated with the conference, and some of the typing in this volume has been done by Miss R Dudgeon; this work is gratefully acknowledged.

G A Watson

Dundee, September 1977.

CONTENTS

C T H BAKER: Runge-Kutta methods for Volterra integral equations of the second kind 1

I BARRODALE: Best approximation of complex-valued data 14

D BOLEY and G H GOLUB: Inverse eigenvalue problems for band matrices .. 23

J S R CHISHOLM: Multivariate approximants with branch points........... 32

L COLLATZ: The numerical treatment of some singular boundary value problems.. 41

M G COX: The incorporation of boundary conditions in spline approximation problems.. 51

J DOUGLAS JR, T DUPONT and P PERCELL: A time-stepping method for Galerkin approximations for nonlinear parabolic equations........... 64

A GEORGE: An automatic one-way dissection algorithm for irregular finite element problems 76

A R MITCHELL and D F GRIFFITHS: Generalised Galerkin methods for second order equations with significant first derivative terms 90

J J MORÉ: The Levenberg-Marquardt algorithm: Implementation and theory 105

M R OSBORNE and G A WATSON: Nonlinear approximation problems in vector norms 117

V PEREYRA: Finite difference solution of two-point boundary value problems and symbolic manipulation 133

M J D POWELL: A fast algorithm for nonlinearly constrained optimization calculations 144

R W H SARGENT: The decomposition of systems of procedures and algebraic equations.. 158

H J STETTER: Global error estimation in ODE-solvers 179

E L WACHSPRESS: Isojacobic crosswind differencing 190

INVITED SPEAKERS

C T H Baker
Department of Mathematics, University of Manchester,
Oxford Road, Manchester M13 9PL, England.

I Barrodale
Department of Mathematics, University of Victoria,
P.O. Box 1700, Victoria, B.C., Canada.

J S R Chisholm
Mathematical Institute, The University, Canterbury,
Kent CT2 7NF, England.

L Collatz
Institut für Angewandte Mathematik, Universitat Hamburg,
2 Hamburg 13, Bundesstr 55, W Germany.

M G Cox
Division of Numerical Analysis and Computing,
National Physical Laboratory, Teddington, Middlesex
TW11 OLW, England.

J Douglas, Jnr
Department of Mathematics, The University of Chicago,
5734 University Avenue, Chicago, Illinois 60637, USA.

J A George
Department of Computer Science, University of Waterloo,
Ontario, Canada.

G H Golub
Computer Science Department, Stanford University,
Stanford, California 94305, USA.

D S Jones
Department of Mathematics, University of Dundee,
Dundee DD1 4HN, Scotland.

A R Mitchell
Department of Mathematics, University of Dundee,
Dundee DD1 4HN, Scotland.

J J Moré
Applied Mathematics Division, Argonne National Laboratory,
9700 South Cass Avenue, Argonne, Illinois 60439, USA.

M R Osborne
Computer Centre, Australian National University,
Box 4 P.O., Canberra, A.C.T. 2600, Australia.

V Pereyra
Applied Mathematics 101-50, California Institute of
Technology, Pasadena, California 91125, USA.

M J D Powell
Department of Applied Mathematics and Theoretical Physics,
University of Cambridge, Silver Street, Cambridge CB3 9EW,
England.

R W H Sargent
Department of Chemical Engineering and Chemical
Technology, Imperial College, London SW7, England.

H J Stetter
Institut für Numerische Mathematik, Technische Hochschule
Wien, A-1040 Wien, Gusshausstr, 27-29 Austria.

E L Wachspress
General Electric Company, P.O. Box 1072, Schenectady,
New York 12301, USA.

SUBMITTED PAPERS

Z Aktas: Computer Science Dept, Middle East Technical University, Turkey.
An accuracy improvement for the method of lines.

P Alfeld: Mathematics Dept, University of Dundee, Scotland.
CDS - A new technique for certain stiff systems of ordinary differential equations.

K Balla: Computer and Automation Institute, Hungarian Academy of Science.
On error estimates of the substitution of the boundedness condition on solutions of systems of linear ordinary differential equations with regular singularity.

K E Barrett: Mathematics Dept, Lanchester Polytechnic, England.
The finite integral method for partial differential equations.

D G Bettis: Institute for Mathematics, Technical University of Munich, Germany.
An efficient embedded Runge-Kutta method.

Jean Beuneu: University of Lille I, France.
The rebalancing method for solving linear systems and eigenproblems.

Ake Björck: Mathematics Dept, Linköping University, Sweden.
Iterative solution of under- and overdetermined linear systems.

Klaus W A Böhmer: Mathematics Institute, University of Karlsruhe, Germany.
Defect corrections via neighbouring problems.

Claude Brezinski: University of Lille I, France.
Rational approximants to power series.

Hermann Brunner: Mathematics Department, Dalhousie University, Canada.
Volterra integral equations and their discretizations.

J P Coleman: Mathematics Dept, University of Durham, England.
Evaluation of the Bessel Functions J_0 and J_1 of complex argument.

I D Coope: Mathematics Dept, University of Dundee, Scotland.
Global convergence results for augmented Lagrangian methods.

G J Cooper: School of Math. and Physical Science, University of Sussex, England.
The order of convergence of linear methods for ordinary differential equations.

L J Cromme: Mathematics Dept, University of Bonn, Germany.
Numerical methods for nonlinear maximum norm approximations.

L M Delves: Dept of Comp and Statistical Science, University of Liverpool, England.
A global element method for the solution of elliptic partial differential equations.

P M Dew: Centre for Computer Studies, University of Leeds, England.
Numerical solution of quasi-linear heat problems with error estimates.

I S Duff: A.E.R.E. Harwell, England.
MA28 - a set of subroutines for solving sparse unsymmetric linear equations.

S Ellacott: Mathematics Dept, Brighton Polytechnic, England.
Practical complex best approximation: The state of the art.

C M Elliott: Computing Laboratory, Oxford University, England.
On the numerical solution of an electrochemical machining problem via a variational inequality formulation.

G Elliott: Mathematics Dept, Portsmouth Polytechnic, England.
The construction of Chebyshev approximations in the complex plane.

R England and J P Hennart: Universidad Nacional de Mexico.
Fractional steps finite element techniques for strongly anisotropic diffusion problems.

R Fletcher: Mathematics Dept, University of Dundee, Scotland.
The reduced Hessian in variable metric methods.

T L Freeman: Mathematics Dept., University of Manchester, England.
A method for computing the zeros of a polynomial with real coefficients.

Nima Geffen and Sara Yaniv: Tel-Aviv University, Israel.
Isoparametric characteristic elements for the Tricomi equation.

B Germain-Bonne, University of Lille I, France.
Shape and variation diminishing properties of spline curves.

Michael Ghil and Remesh Balgovind: Courant Institute of Mathematical Sciences, New York University, USA.
A fast Cauchy-Riemann solver with nonlinear applications.

Ian Gladwell: Mathematics Dept, University of Manchester, England.
The NAG library chapter for the solution of ordinary differential equations.

Moshe Goldberg: Mathematics Dept, University of California, USA.
Dissipative schemes for hyperbolic problems and boundary extrapolation.

R Gorenflo: Mathematics Dept, Freie Universität Berlin, Germany.
Conservative difference schemes for diffusion problems.

Myron S Henry: Mathematics Dept, Montana State University, USA.
Numerical comparisons of algorithms for polynomial and rational multivariate approximations.

J N Holt: Mathematics Dept, University of Queensland, Australia.
Free-knot cubic spline inversion of a Fredholm integral equation.

M K Horn: Institute for Mathematics, Technical University of Munich, Germany.
Developments in high-order Runge-Kutta-Nyström methods.

W D Hoskins, D S Meek, D J Walton: Dept of Computer Science, University of Manitoba, Canada.
An alternative method for the solution of Poisson-type equations on Rectangular Regions in two or three space dimensions.

K Jittorntrum, M R Osborne: Computer Centre, Australian National University.
Trajectory analysis and extrapolation in barrier function methods.

D C Joyce: Mathematics Dept, Massey University, New Zealand.
Extrapolation to the limit - algorithms and applications.

Bo Kagström: Dept of Information Processing and Numerical Analysis, University of Umea, Sweden.
On the numerical computation of matrix functions.

Malcolm S Keech: Mathematics Dept, University of Manchester, England.
Semi-explicit methods in the numerical solution of first kind Volterra integral equations.

R Kress: Universität Göttingen and University of Strathclyde, Scotland.
On improving the rate of convergence of successive approximation for integral equations of potential theory.

D P Laurie: National Research Institute for Mathematical Sciences, South Africa.
Exponentially fitted multipoint methods for two-point boundary value problems.

J D Lawson and J Ll Morris: Computer Science Dept, University of Waterloo, Canada.
The extrapolation of first order methods for parabolic partial differential equations.

A V Levy*and A Montalvo+: *Universidad Nacional Autónoma de México, +Universidad Iberoamericana, México.
The tunneling algorithm for the global minimization of functions.

I M Longman: Dept of Geophysics and Planetary Sciences, Tel-Aviv University, Israel.
A method of Laplace transform inversion by exponential series.

Jens Lorenz: Institute for Numerical Mathematics, University of Munster, Germany.
Stability inequalities for discrete boundary value problems.

J T Marti: Mathematics Dept, Swiss Federal Institute of Technology.
An algorithm for the computation of Fourier coefficients of non-analytic functions using B-splines of arbitrary order.

J C Mason: Mathematics Branch, Royal Military College of Science, Shrivenham, England.
A one-dimensional spline approximation method for the numerical solution of heat conduction problems.

S McKee: University of Oxford, England.
Multistep methods for solving linear Volterra integro-differential equations.

G Moore and A Spence: School of Mathematics, University of Bath, England.
Newton's method near a bifurcation point.

P Moore: Mathematics Dept, University of Aston in Birmingham, England.
Finite element multistep multiderivative schemes for linear parabolic equations.

Gerhard Opfer: Mathematics Dept, University of Hamburg, Germany.
Numerical solution of certain nonstandard approximation problems.

I Riddell: Dept of Computational and Statistical Science, University of Liverpool, England.
On comparing integral equation routines.

A Robinson and A Prothero: Shell Research Limited, Chester, England.
Global error estimates for solutions to stiff systems of ordinary differential equations.

J Barkley Rosser: Mathematics Research Center, University of Wisconsin-Madison, USA.
Harmonic functions on regions with reentrant corners.

A Sayfy: School of Maths. and Physical Sciences, University of Sussex, England.
Additive numerical methods for ordinary differential equations.

J Sinclair: Mathematics Dept, University of Dundee, Scotland.
A variable metric method generating orthogonal directions.

H J J te Riele: Mathematical Centre, Amsterdam, Holland.
Computation of zeros of partial sums of the Riemann ζ-function with real part > 1.

Per Grove Thomsen and Zahari Zlatev: Institute for Numerical Analysis, Technical University of Denmark.
The use of Backward Differentiation methods in the solution of non-stationary heat conduction problems.

Ph L Toint: F.N.D.P. Belguim.
On sparse and symmetric matrix updating subject to a linear equation.

J M Watt: Dept of Computational Science, University of Liverpool, England.
The convergence of deferred and defect corrections.

Richard Weiss: Technische Universität, Wien, Austria.
On the eigenvalue problem for singular systems of ordinary differential equations.

B Werner: Mathematics Dept., University of Hamburg, Germany.
About a connection between complementary and nonconforming finite elements.

Ragnar Winther: Institute of Informatics, University of Oslo, Norway.
A Galerkin method for a parabolic control problem.

G Woodford: Mathematics Dept, University of Dundee, Scotland.
Isoparametric cubic triangles in the finite element method.

K Wright: Computing Laboratory, University of Newcastle upon Tyne, England.
Asymptotic properties of quadrature weights based on zeros of orthogonal polynomials over partial and full ranges.

THE LEVENBERG-MARQUARDT ALGORITHM: IMPLEMENTATION AND THEORY[*]

Jorge J. Moré

1. Introduction

Let F: $R^n \to R^m$ be continuously differentiable, and consider the nonlinear least squares problem of finding a local minimizer of

$$\Phi(x) = \frac{1}{2}\sum_{i=1}^{m} f_i^2(x) = \frac{1}{2}\|F(x)\|^2 . \tag{1.1}$$

Levenberg [1944] and Marquardt [1963] proposed a very elegant algorithm for the numerical solution of (1.1). However, most implementations are either not robust, or do not have a solid theoretical justification. In this work we discuss a robust and efficient implementation of a version of the Levenberg-Marquardt algorithm, and show that it has strong convergence properties. In addition to robustness, the main features of this implementation are the proper use of implicitly scaled variables, and the choice of the Levenberg-Marquardt parameter via a scheme due to Hebden [1973]. Numerical results illustrating the behavior of this implementation are also presented.

Notation. In all cases $\|\cdot\|$ refers to the ℓ_2 vector norm or to the induced operator norm. The Jacobian matrix of F evaluated at x is denoted by $F'(x)$, but if we have a sequence of vectors $\{x_k\}$, then J_k and f_k are used instead of $F'(x_k)$ and $F(x_k)$, respectively.

2. Derivation

The easiest way to derive the Levenberg-Marquardt algorithm is by a linearization argument. If, given $x \in R^n$, we could minimize

$$\Psi(p) = \|F(x+p)\|$$

as a function of p, then x+p would be the desired solution. Since Ψ is usually a nonlinear function of p, we linearize F(x+p) and obtain the linear least squares problem

$$\psi(p) = \|F(x) + F'(x)p\| .$$

Of course, this linearization is not valid for all values of p, and thus we consider the constrained linear least squares problem

[*] Work performed under the auspices of the U.S. Energy Research and Development Administration

$$\min\{\psi(p): \|Dp\| \le \Delta\} . \tag{2.1}$$

In theory D is any given nonsingular matrix, but in our implementation D is a diagonal matrix which takes into account the scaling of the problem. In either case, p lies in the hyperellipsoid

$$E = \{p: \|Dp\| \le \Delta\} , \tag{2.2}$$

but if D is diagonal, then E has axes along the coordinate directions and the length of the ith semi-axis is Δ/d_i.

We now consider the solution of (2.1) in some generality, and thus the problem

$$\min\{\|f+Jp\|: \|Dp\| \le \Delta\} \tag{2.3}$$

where $f \in R^m$ and J is any m by n matrix. The basis for the Levenberg-Marquardt method is the result that if p^* is a solution to (2.3), then $p^* = p(\lambda)$ for some $\lambda \ge 0$ where

$$p(\lambda) = -(J^TJ + \lambda D^TD)^{-1}J^Tf . \tag{2.4}$$

If J is rank deficient and $\lambda = 0$, then (2.4) is defined by the limiting process

$$Dp(0) \equiv \lim_{\lambda \to 0^+} Dp(\lambda) = -(JD^{-1})^{\dagger}f .$$

There are two possibilities: Either $\lambda = 0$ and $\|Dp(0)\| \le \Delta$, in which case p(0) is the solution to (2.3) for which $\|Dp\|$ is least, or $\lambda > 0$ and $\|Dp(\lambda)\| = \Delta$, and then $p(\lambda)$ is the unique solution to (2.3).

The above results suggest the following iteration.

(2.5) Algorithm

(a) Given $\Delta_k > 0$, find $\lambda_k \ge 0$ such that if

$$(J_k^TJ_k + \lambda_k D_k^TD_k)p_k = -J_k^Tf_k ,$$

then either $\lambda_k = 0$ and $\|D_kp_k\| \le \Delta_k$, or $\lambda_k > 0$ and $\|D_kp_k\| = \Delta_k$.

(b) If $\|F(x_k+p_k)\| < \|F(x_k)\|$ set $x_{k+1} = x_k+p_k$ and evaluate J_{k+1}; otherwise set $x_{k+1} = x_k$ and $J_{k+1} = J_k$.

(c) Choose Δ_{k+1} and D_{k+1}.

In the next four sections we elaborate on how (2.5) leads to a very robust and efficient implementation of the Levenberg-Marquardt algorithm.

3. Solution of a Structured Linear Least Squares Problem

The simplest way to obtain the correction p is to use Cholesky decomposition on the linear system

$$(3.1) \qquad (J^TJ + \lambda D^TD)p = -J^Tf \ .$$

Another method is to recognize that (3.1) are the normal equations for the least squares problem

$$(3.2) \qquad \begin{pmatrix} J \\ \lambda^{\frac{1}{2}}D \end{pmatrix} p \cong -\begin{pmatrix} f \\ 0 \end{pmatrix} ,$$

and to solve this structured least squares problem using QR decomposition with column pivoting.

The main advantage of the normal equations is speed; it is possible to solve (3.1) twice as fast as (3.2). On the other hand, the normal equations are particularly unreliable when $\lambda = 0$ and J is nearly rank deficient. Moreover, the formation of J^TJ or D^TD can lead to unnecessary underflows and overflows, while this is not the case with (3.2). We feel that the loss in speed is more than made up by the gain in reliability and robustness.

The least squares solution of (3.2) proceeds in two stages. These stages are the same as those suggested by Golub (Osborne [1972]), but modified to take into account the pivoting.

In the first stage, compute the QR decomposition of J with column pivoting. This produces an orthogonal matrix Q and a permutation π of the columns of J such that

$$(3.3) \qquad QJ\pi = \begin{pmatrix} T & S \\ 0 & 0 \end{pmatrix}$$

where T is a nonsingular upper triangular matrix of rank (J) order. If $\lambda = 0$, then a solution of (3.2) is

$$p = \pi \begin{pmatrix} T^{-1} & 0 \\ 0 & 0 \end{pmatrix} Qf \equiv J^{-}f$$

where J^{-} refers to a particular symmetric generalized inverse of J in the sense that JJ^{-} is symmetric and $JJ^{-}J = J$. To solve (3.2) when $\lambda > 0$ first note that (3.3) implies that

$$(3.4) \qquad \begin{pmatrix} Q & 0 \\ 0 & \pi^T \end{pmatrix} \begin{pmatrix} J \\ \lambda^{\frac{1}{2}}D \end{pmatrix} \pi = \begin{pmatrix} R \\ 0 \\ D_\lambda \end{pmatrix}$$

where $D_\lambda = \lambda^{\frac{1}{2}}\pi^TD\pi$ is still a diagonal matrix and R is a (possibly singular) upper triangular matrix of order n.

In the second stage, compute the QR decomposition of the matrix on the right of (3.4). This can be done with a sequence of n(n+1)/2 Givens rotations. The result is an orthogonal matrix W such that

$$(3.5) \qquad W\begin{bmatrix} R \\ 0 \\ D_\lambda \end{bmatrix} = \begin{bmatrix} R_\lambda \\ 0 \end{bmatrix}$$

where R_λ is a nonsingular upper triangular matrix of order n. The solution to (3.2) is then

$$p = -\pi R_\lambda^{-1} u$$

where $u \in R^n$ is determined from

$$W\begin{bmatrix} Qf \\ 0 \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} .$$

It is important to note that if λ is changed, then only the second stage must be redone.

4. Updating the Step Bound

The choice of Δ depends on the ratio between the actual reduction and the predicted reduction obtained by the correction. In our case, this ratio is given by

$$(4.1) \qquad \rho(p) = \frac{\|F(x)\|^2 - \|F(x+p)\|^2}{\|F(x)\|^2 - \|F(x)+F'(x)p\|^2} .$$

Thus (4.1) measures the agreement between the linear model and the (nonlinear) function. For example, if F is linear then $\rho(p) = 1$ for all p, and if $F'(x)^T F(x) \neq 0$, then $\rho(p) \to 1$ as $\|p\| \to 0$. Moreover, if $\|F(x+p)\| \geq \|F(x)\|$ then $\rho(p) \leq 0$.

The scheme for updating Δ has the objective of keeping the value of (4.1) at a reasonable level. Thus, if $\rho(p)$ is close to unity (i.e. $\rho(p) \geq 3/4$), we may want to increase Δ, but if $\rho(p)$ is not close to unity (i.e. $\rho(p) \leq 1/4$), then Δ must be decreased. Before giving more specific rules for updating Δ, we discuss the computation of (4.1). For this, write

$$(4.2) \qquad \rho = \frac{\|f\|^2 - \|f_+\|^2}{\|f\|^2 - \|f+Jp\|^2}$$

with an obvious change in notation. Since p satisfies (3.1),

$$(4.3) \qquad \|f\|^2 - \|f+Jp\|^2 = \|Jp\|^2 + 2\lambda\|Dp\|^2 ,$$

and hence we can rewrite (4.2) as

(4.4)
$$\rho = \frac{1 - \left(\frac{\|f_+\|}{\|f\|}\right)^2}{\left(\frac{\|Jp\|}{\|f\|}\right)^2 + 2\left(\frac{\lambda^{\frac{1}{2}}\|Dp\|}{\|f\|}\right)^2} .$$

Since (4.3) implies that

$$\|Jp\| \le \|f\|, \quad \lambda^{\frac{1}{2}}\|Dp\| \le \|f\| ,$$

the computation of the denominator will not generate any overflows, and moreover, the denominator will be non-negative regardless of roundoff errors. Note that this is not the case with (4.2). The numerator of (4.4) may generate overflows if $\|f_+\|$ is much larger than $\|f\|$, but since we are only interested in positive values of ρ, if $\|f_+\| > \|f\|$ we can just set $\rho = 0$ and avoid (4.4).

We now discuss how to update Δ. To increase Δ we simply multiply Δ by a constant factor not less than one. To decrease Δ we follow Fletcher [1971] and fit a quadratic to $\delta(0)$, $\delta'(0)$ and $\delta(1)$ where

$$\delta(\theta) = \frac{1}{2}\|F(x+\theta p)\|^2 .$$

If μ is the minimizer of the resulting quadratic, we decrease Δ by multiplying Δ by μ, but if $\mu \notin \left[\frac{1}{10}, \frac{1}{2}\right]$, we replace μ by the closest endpoint. To compute μ safely, first note that (3.1) implies that

$$\gamma \equiv \frac{p^TJ^Tf}{\|f\|^2} = -\left[\left(\frac{\|Jp\|}{\|f\|}\right)^2 + \left(\lambda^{\frac{1}{2}}\frac{\|Dp\|}{\|f\|}\right)^2\right] ,$$

and that $\gamma \in [-1,0]$. It is now easy to verify that

(4.5)
$$\mu = \frac{\frac{1}{2}\gamma}{\gamma + \frac{1}{2}\left[1 - \left(\frac{\|f_+\|}{\|f\|}\right)^2\right]} .$$

If $\|f_+\| \le \|f\|$ we set $\mu = 1/2$. Also note that we only compute μ by (4.5) if say, $\|f_+\| \le 10\|f\|$, for otherwise, $\mu \le 1/10$.

5. The Levenberg-Marquardt Parameter

In our implementation $\alpha > 0$ is accepted as the Levenberg-Marquardt parameter if

(5.1)
$$|\phi(\alpha)| \le \sigma\Delta ,$$

where

(5.2)
$$\phi(\alpha) = \|D(J^TJ+\alpha D^TD)^{-1}J^Tf\| - \Delta ,$$

and $\sigma \in (0,1)$ specifies the desired relative error in $\|Dp(\alpha)\|$. Of course, if

$\phi(0) \le 0$ then $\alpha = 0$ is the required parameter, so in the remainder of this section we assume that $\phi(0) > 0$. Then ϕ is a continuous, strictly decreasing function on $[0,+\infty)$ and $\phi(\alpha)$ approaches $-\Delta$ at infinity. It follows that there is a unique $\alpha^* > 0$ such that $\phi(\alpha^*) = 0$. To determine the Levenberg-Marquardt parameter we assume that an initial estimate $\alpha_0 > 0$ is available, and generate a sequence $\{\alpha_k\}$ which converges to α^*.

Since ϕ is a convex function, it is very tempting to use Newton's method to generate $\{\alpha_k\}$, but this turns out to be very inefficient -- the particular structure of this problem allows us to derive a much more efficient iteration due to Hebden [1973]. To do this, note that

$$(5.3) \qquad \phi(\alpha) = \| (\tilde{J}^T\tilde{J}+\alpha I)^{-1}\tilde{J}^T f \| - \Delta, \quad \tilde{J} = JD^{-1} ,$$

and let $\tilde{J} = U\Sigma V^T$ be the singular value decomposition of $\tilde{J}$. Then

$$\phi(\alpha) = \left[\sum_{i=1}^{n} \frac{\sigma_i^2 z_i^2}{(\sigma_i^2+\alpha)^2} \right]^{\frac{1}{2}} - \Delta$$

where $z = U^T f$ and $\sigma_1,\ldots,\sigma_n$ are the singular values of $\tilde{J}$. Hence, it is very natural to assume that

$$\phi(\alpha) \doteq \frac{a}{b+\alpha} - \Delta \equiv \tilde{\phi}(\alpha) ,$$

and to choose a and b so that $\tilde{\phi}(\alpha_k) = \phi(\alpha_k)$ and $\tilde{\phi}'(\alpha_k) = \phi'(\alpha_k)$. Then $\tilde{\phi}(\alpha_{k+1}) = 0$ if

$$(5.4) \qquad \alpha_{k+1} = \alpha_k - \left[\frac{\phi(\alpha_k) + \Delta}{\Delta}\right] \left[\frac{\phi(\alpha_k)}{\phi'(\alpha_k)}\right] .$$

This iterative scheme must be safeguarded if it is to converge. Hebden [1973] proposed using upper and lower bounds u_k and ℓ_k, and that (5.4) be applied with the restriction that no iterate may be within $(u_k-\ell_k)/10$ of either endpoint. It turns out that this restriction is very detrimental to the progress of the iteration since in a lot of cases u_k is much larger than ℓ_k. A much more efficient algorithm can be obtained if (5.4) is only modified when α_{k+1} is outside of (ℓ_{k+1},u_{k+1}). To specify this algorithm we first follow Hebden [1973] and note that (5.3) implies that

$$u_0 = \frac{\| (JD^{-1})^T f \|}{\Delta}$$

is a suitable upper bound. If J is not rank deficient, then $\phi'(0)$ is defined and the convexity of ϕ implies that

$$\ell_0 = - \frac{\phi(0)}{\phi'(0)}$$

is a lower bound; otherwise let $\ell_0 = 0$.

(5.5) Algorithm

(a) If $\alpha_k \notin (\ell_k, u_k)$ let $\alpha_k = \max\{0.001\, u_k,\ (\ell_k u_k)^{\frac{1}{2}}\}$.

(b) Evaluate $\phi(\alpha_k)$ and $\phi'(\alpha_k)$. Update u_k by letting $u_{k+1} = \alpha_k$ if $\phi(\alpha_k) < 0$ and $u_{k+1} = u_k$ otherwise. Update ℓ_k by

$$\ell_{k+1} = \max\left\{\ell_k,\ \alpha_k - \frac{\phi(\alpha_k)}{\phi'(\alpha_k)}\right\} .$$

(c) Obtain α_{k+1} from (5.4).

The role of (5.5)(a) is to replace α_k by a point in (ℓ_k, u_k) which is biased towards ℓ_k; the factor 0.001 u_k was added to guard against exceedingly small values of ℓ_k, and in particular, $\ell_k = 0$. In (5.5)(b), the convexity of ϕ guarantees that the Newton iterate can be used to update ℓ_k.

It is not too difficult to show that algorithm (5.5) always generates a sequence which converges quadratically to α^*. In practice, less than two iterations (on the average) are required to satisfy (5.1) when $\sigma = 0.1$.

To complete the discussion of the Hebden algorithm, we show how to evaluate $\phi'(\alpha)$. From (5.2) it follows that

$$\phi'(\alpha) = -\frac{(D^Tq(\alpha))^T(J^TJ+\alpha D^TD)^{-1}(D^Tq(\alpha))}{\|q(\alpha)\|}$$

where $q(\alpha) = Dp(\alpha)$ and $p(\cdot)$ is defined by (2.4). From (3.4) and (3.5) we have

$$\pi^T(J^TJ+\alpha D^TD)\pi = R_\alpha^T R_\alpha ,$$

and hence,

$$\phi'(\alpha) = -\|q(\alpha)\| \left\| R_\alpha^{-T}\left(\frac{\pi^TD^Tq(\alpha)}{\|q(\alpha)\|}\right)\right\|^2 .$$

6. Scaling

Since the purpose of the matrix D_k in the Levenberg-Marquardt algorithm is to take into account the scaling of the problem, some authors (e.g. Fletcher [1971]) choose

(6.1) $$D_k = \mathrm{diag}(d_1^{(k)},\ldots,d_n^{(k)})$$

where

(6.2) $$d_i^{(k)} = \|\partial_i F(x_0)\|, \quad k \geq 0 .$$

This choice is usually adequate as long as $\|\partial_i F(x_k)\|$ does not increase with k. However, if $\|\partial_i F(x_k)\|$ increases, this requires a decrease in the length (= Δ/d_i) of the i^{th} semi-axis of the hyperellipsoid (2.2), since F is now changing faster along the

i^{th} variable, and therefore, steps which have a large i^{th} component tend to be unreliable. This argument leads to the choice

$$d_i^{(0)} = \| \partial_i F(x_0) \| \tag{6.3}$$

$$d_i^{(k)} = \max \left\{ d_i^{(k-1)}, \| \partial_i F(x_k) \| \right\} , \quad k \geq 1 .$$

Note that a decrease in $\| \partial_i F(x_k) \|$ only implies that F is not changing as fast along the i^{th} variable, and hence does not require a decrease in d_i. In fact, the choice

$$d_i^{(k)} = \| \partial_i F(x_k) \| , \quad k \geq 0 , \tag{6.4}$$

is computationally inferior to both (6.2) and (6.3). Moreover, our theoretical results support choice (6.3) over (6.4), and to a lesser extent, (6.2).

It is interesting to note that (6.2), (6.3), and (6.4) make the Levenberg-Marquardt algorithm scale invariant. In other words, for all of the above choices, if D is a diagonal matrix with positive diagonal elements, then algorithm (2.5) generates the same iterates if either it is applied to F and started at x_0, or if it is applied to $\tilde{F}(x) = F(D^{-1}x)$ and started at $\tilde{x}_0 = Dx_0$. For this result it is assumed that the decision to change Δ is only based on (4.1), and thus is also scale invariant.

7. Theoretical Results

It will be sufficient to present a convergence result for the following version of the Levenberg-Marquardt algorithm.

(7.1) Algorithm

(a) Let $\sigma \in (0,1)$. If $\| D_k J_k^- f_k \| \leq (1+\sigma)\Delta_k$, set $\lambda_k = 0$ and $p_k = -J_k^- f_k$. Otherwise determine $\lambda_k > 0$ such that if

$$\begin{bmatrix} J_k \\ \lambda_k^{\frac{1}{2}} D_k \end{bmatrix} p_k \cong - \begin{bmatrix} f_k \\ 0 \end{bmatrix}$$

then

$$(1-\sigma)\Delta_k \leq \| D_k p_k \| \leq (1+\sigma)\Delta_k .$$

(b) Compute the ratio ρ_k of actual to predicted reduction.

(c) If $\rho_k \leq 0.0001$, set $x_{k+1} = x_k$ and $J_{k+1} = J_k$.
If $\rho_k > 0.0001$, set $x_{k+1} = x_k + p_k$ and compute J_{k+1}.

(d) If $\rho_k \leq 1/4$, set $\Delta_{k+1} \in \left[\frac{1}{10} \Delta_k, \frac{1}{2} \Delta_k \right]$.
If $\rho_k \in \left(\frac{1}{4}, \frac{3}{4} \right)$ and $\lambda_k = 0$, or if $\rho_k \geq 3/4$, set $\Delta_{k+1} = 2\| D_k p_k \|$.

(e) Update D_{k+1} by (6.1) and (6.3).

The proof of our convergence result is somewhat long and will therefore be presented elsewhere.

Theorem. Let $F: R^n \to R^m$ be continuously differentiable on R^n, and let $\{x_k\}$ be the sequence generated by algorithm (7.1). Then

(7.2) $$\liminf_{k \to +\infty} \|(J_k D_k^{-1})^T f_k\| = 0 .$$

This result guarantees that eventually a scaled gradient will be small enough. Of course, if $\{J_k\}$ is bounded then (7.2) implies the more standard result that

(7.3) $$\liminf_{k \to +\infty} \|J_k^T f_k\| = 0 .$$

Furthermore, we can also show that if F' is uniformly continuous then

(7.4) $$\lim_{k \to +\infty} \|J_k^T f_k\| = 0 .$$

Powell [1975] and Osborne [1975] have also obtained global convergence results for their versions of the Levenberg-Marquardt algorithm. Powell presented a general algorithm for unconstrained minimization which as a special case contains (7.1) with $\sigma = 0$ and $\{D_k\}$ constant. For this case Powell obtains (7.3) under the assumption that $\{J_k\}$ is bounded. Osborne's algorithm directly controls $\{\lambda_k\}$ instead of $\{\Delta_k\}$, and allows $\{D_k\}$ to be chosen by (6.1) and (6.3). For this case he proves (7.4) under the assumptions that $\{J_k\}$ and $\{\lambda_k\}$ are bounded.

8. Numerical Results

In our numerical results we would like to illustrate the behavior of our algorithm with the three choices of scaling mentioned in Section 6. For this purpose, we have chosen four functions.

1) Fletcher and Powell [1963] n=3, m=3

$$f_1(x) = 10[x_3 - 10\theta(x_1,x_2)]$$
$$f_2(x) = 10[(x_1^2+x_2^2)^{\frac{1}{2}} - 1]$$
$$f_3(x) = x_3$$

where

$$\theta(x_1,x_2) = \begin{cases} \frac{1}{2\pi} \arctan (x_2/x_1), & x_1 > 0 \\ \frac{1}{2\pi} \arctan (x_2/x_1) + 0.5, & x_1 < 0 \end{cases}$$

$$x_0 = (-1,0,0)^T$$

2. Kowalik and Osborne [1968] n=4, m=11

$$f_i(x) = y_i - \frac{x_1[u_i^2 + x_2u_i]}{(u_i^2 + x_3u_i + x_4)}$$

where u_i and y_i are specified in the original paper.

$$x_0 = (0.25, 0.39, 0.415, 0.39)^T$$

3. Bard [1970] n=3, m=15

$$f_i(x) = y_i - \left[x_1 + \frac{u_i}{x_2v_i + x_3w_i}\right]$$

where $u_i = i$, $v_i = 16-i$, $w_i = \min\{u_i,v_i\}$, and y_i is specified in the original paper.

$$x_0 = (1,1,1)^T$$

4. Brown and Dennis [1971] n=4, m=20

$$f_i(x) = [x_1 + x_2t_i - \exp(t_i)]^2 + [x_3 + x_4\sin(t_i) - \cos(t_i)]^2$$

where $t_i = (0.2)i$.

$$x_0 = (25, 5, -5, 1)^T$$

These problems have very interesting features. Problem 1 is a helix with a zero residual at $x^* = (1,0,0)$ and a discontinuity along the plane $x_1 = 0$; note that the algorithm must cross this plane to reach the solution. Problems 2 and 3 are data fitting problems with small residuals, while Problem 4 has a large residual. The residuals are given below.

1. $\|F(x^*)\| = 0.0$
2. $\|F(x^*)\| = 0.0175358$
3. $\|F(x^*)\| = 0.0906359$
4. $\|F(x^*)\| = 292.9542$

Problems 2 and 3 have other solutions. To see this, note that for Kowalik and Osborne's function,

$$\lim_{\alpha\to\infty} f_i(\alpha,x_2,\alpha,\alpha) = y_i - \left(\frac{u_i}{u_i+1}\right)(x_2+u_i) \ , \tag{8.1}$$

while for Bard's function,

$$\lim_{\alpha\to\infty} f_i(x_1,\alpha,\alpha) = y_i - x_1 \ . \tag{8.2}$$

These are now linear least squares problems, and as such, the parameter x_2 in (8.1) and x_1 in (8.2) are completely determined. However, the remaining parameters only need to be sufficiently large.

In presenting numerical results one must be very careful about the convergence criteria used. This is particularly true of the Levenberg-Marquardt method since, unless $F(x^*) = 0$, the algorithm converges linearly. In our implementation, an approximation x to x^* is acceptable if either x is close to x^* or $\|F(x)\|$ is close to $\|F(x^*)\|$. We attempt to satisfy these criteria by the convergence tests

$$\Delta \leq \text{XTOL}\ \|Dx\|\ , \tag{8.3}$$

and

$$\left(\frac{\|Jp\|}{\|f\|}\right)^2 + 2\left(\lambda^{\frac{1}{2}}\ \frac{\|Dp\|}{\|f\|}\right)^2 \leq \text{FTOL}\ . \tag{8.4}$$

An important aspect of these tests is that they are scale invariant in the sense of Section 6. Also note that the work of Section 4 shows that (8.4) is just the relative error between $\|f+Jp\|^2$ and $\|f\|^2$.

The problems were run on the IBM 370/195 of Argonne National Laboratory in double precision (14 hexadecimal digits) and under the FORTRAN H (opt=2) compiler. The tolerances in (8.3) and (8.4) were set at FTOL = 10^{-8} and XTOL = 10^{-8}. Each problem is run with three starting vectors. We have already given the starting vector x_0 which is closest to the solution; the other two points are $10x_0$ and $100x_0$. For each starting vector, we have tried our algorithm with the three choices of $\{D_k\}$. In the table below, choices (6.2), (6.3) and (6.4) are referred to as initial, adaptive, and continuous scaling, respectively. Moreover, NF and NJ stands for the number of function and Jacobian evaluations required for convergence.

		x_0		$10x_0$		$100x_0$	
PROBLEM	SCALING	NF	NJ	NF	NJ	NF	NJ
1	Initial Adaptive Continuous	12 11 12	9 8 9	34 20 14	29 15 12	FC 19 176	FC 16 141
2	Initial Adaptive Continuous	19 18 18	17 16 16	81 79 63	71 71 54	365 348 FC	315 307 FC
3	Initial Adaptive Continuous	8 8 8	7 7 7	37 37 FC	36 36 FC	14 14 FC	13 13 FC
4	Initial Adaptive Continuous	268 268 FC	242 242 FC	423 57 FC	400 47 FC	FC 229 FC	FC 207 FC

Interestingly enough, convergence to the minimizer indicated by (8.1) only occurred for starting vector $10x_0$ of Problem 2, while for Problem 3 starting vectors $10x_0$ and $100x_0$ led to (8.2). Otherwise, either the global minimizer was obtained, or the algorithm failed to converge to a solution; the latter is indicated by FC in the table.

It is clear from the table that the adaptive strategy is best in these four examples. We have run other problems, but in all other cases the difference is not as dramatic as in these cases. However, we believe that the above examples adequately justify our choice of scaling matrix.

Acknowledgments. This work benefited from interaction with several people. Beverly Arnoldy provided the numerical results for several versions of the Levenberg-Marquardt algorithm, Brian Smith showed how to use pivoting in the two-stage process of Section 3, and Danny Sorensen made many valuable comments on an earlier draft of this paper. Finally, I would like to thank Judy Beumer for her swift and beautiful typing of the paper.

References

1. Bard, Y. [1970]. Comparison of gradient methods for the solution of nonlinear parameter estimation problem, *SIAM J. Numer. Anal. 7*, 157-186.

2. Brown, K. M. and Dennis, J. E. [1971]. New computational algorithms for minimizing a sum of squares of nonlinear functions, Department of Computer Science report 71-6, Yale University, New Haven, Connecticut.

3. Fletcher, R. [1971]. A modified Marquardt subroutine for nonlinear least squares, Atomic Energy Research Establishment report R6799, Harwell, England.

4. Fletcher, R. and Powell, M.J.D. [1963]. A rapidly convergent descent method for minimization, *Comput. J. 6*, 163-168.

5. Hebden, M. D. [1973]. An algorithm for minimization using exact second derivatives, Atomic Energy Research Establishment report TP515, Harwell, England.

6. Kowalik, J. and Osborne, M. R. [1968]. *Methods for Unconstrained Optimization Problems*, American Elsevier.

7. Levenberg, K. [1944]. A method for the solution of certain nonlinear problems in least squares, *Quart. Appl. Math. 2*, 164-168.

8. Marquardt, D. W. [1963]. An algorithm for least squares estimation of nonlinear parameters, *SIAM J. Appl. Math. 11*, 431-441.

9. Osborne, M. R. [1972]. Some aspects of nonlinear least squares calculations, in *Numerical Methods for Nonlinear Optimization*, F. A. Lootsma, ed., Academic Press.

10. Osborne, M. R. [1975]. Nonlinear least squares - the Levenberg algorithm revisited, to appear in Series B of the *Journal of the Australian Mathematical Society*.

11. Powell, M. J. D. [1975]. Convergence properties of a class of minimization algorithms, in *Nonlinear Programming 2*, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press.