2

This is an informal report intended primarily for internal or limited external distribution. (The opinions and conclusions stated are those of the author and may or may not be those of the laboratory.) This report is not to be given additional external distribution or cited in external documents without the consent of the author or LLL Technical Information Department.

22,985

UCID - 16331

LAWRENCE LIVERMORE LABORATORY

University of California/Livermore, California

LECTURE SERIES ON NONLINEAR LEAST SQUARES

George D. Byrne

July 1973

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

LECTURE SERIES ON NONLINEAR LEAST SQUARES

George D. Byrne[†]

Lawrence Livermore Laboratory, University of California

Livermore, California

July 1973

ABSTRACT

This series of tutorial lectures on nonlinear least squares was presented July 16, 18, and 20, 1973. The series describes the least squares problem, some old methods for solving it and a new one.

[†]Currently with the Departments of Mathematics and Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, PA 15260.

1. INTRODUCTION

In this series of lectures, we state the nonlinear least squares problem, indicate its origin, outline some currently used methods for solving the nonlinear least squares problem and note a few connections among these methods. After pointing out some shortcomings of these methods, we then motivate and briefly describe a new method and give some numerical results.

2. THE PROBLEM (See [3], [15], [16].)

The nonlinear least squares problem can be simply described as

$$\text{Minimize } F : E^n \rightarrow E \quad (2.1)$$

where

$$F(\theta) = \sum_{i=1}^N w_i \left[y_i - f_i(\theta) \right]^2, \quad \theta \in E^n \quad (2.2)$$

and

$$N \geq n.$$

There are three common problem types.

TYPE I.

Here $f_i(\theta) = g(x_i, \theta)$ with $x_i \in E^p$, which corresponds to the curve fitting problem. For example, suppose $n = 4$ and $N = 25$. If the y_i is interpreted as a datum corresponding to x_i and $w_i > 0$ is taken to be a weight assigned to that point, then w_i is an assessment of the worth of that datum. That is if w_i is near zero the worth is less than if w_i were near unity. In this particular example

$$\left[y_i - \left(\theta_4 x_i^3 + \theta_3 x_i^2 + \theta_2 x_i + \theta_1 \right) \right]$$

denotes the distance from a cubic polynomial at x_i to y_i . The weighted sum of squares of twenty-five such distances is to be minimized. Such curve

fitting occurs in the study of trajectories, surfaces, equipment performance, etc.

TYPE II.

In this type, the solution of the nonlinear system

$$f_1(\theta) = y_1$$

$$f_2(\theta) = y_2$$

$$\vdots$$

$$f_N(\theta) = y_N$$

is sought. Usually for such a system $w_i = 1$, for $i = 1, 2, \dots, n$. The minimization of (2.2) is sought.

TYPE III.

This type is just a combination or hybrid comprised of the previous two types.

$$f_1(\theta) = g_1(x_1, \theta)$$

$$f_2(\theta) = g_1(x_2, \theta)$$

$$\vdots$$

$$f_{N_1}(\theta) = g_2(x_{N_1}, \theta)$$

$$f_{N_1+1}(\theta) = g_2(x_{N_1+1}, \theta)$$

$$\vdots$$

$$f_{N_k}(\theta) = g_k(x_{N_k}, \theta)$$

$$f_{N_k+1}(\theta) = y_{N_k+1}$$

$$\vdots$$

$$f_N(\theta) = y_N$$

There are k sets of experimental data subject to linear constraints based on auxiliary information and a nonlinear system, say. Such problems are fairly common. Here again $w_i = 1$, $N_k < i \leq N$, $0 \leq w_i \leq 1$ for $1 \leq i \leq N$.

In the least squares problem $W = \text{diag} [w_1, \dots, w_N]$ can be taken as a covariance matrix. It is also worth mentioning that instead of minimizing $\phi : E^n \rightarrow E$ we could solve the system of TYPE II with

$$f_i(\theta) = \phi_{,i}(\theta) = \frac{\partial \phi(\theta)}{\partial \theta_i} \quad \text{and} \quad y_i = 0.$$

3. SOME METHODS OF SOLUTION

Here we outline a few methods for solving the nonlinear least squares problem (2.1) - (2.2) (see [3], [15], and [16]).

3.1. The Gradient Method or Method of Steepest Descent

This technique was described in [2], where the basic idea is attributed to Cauchy [1]. The fundamental notion is simply this. Think of F as describing a hill and always attempt to move downhill. This can be done, theoretically at least, by examining the negative of the gradient of F , namely

$$-\nabla F(\theta) \equiv \delta_g \quad (3.1.1)$$

where

$$\delta_{gj} = 2 \sum_{i=1}^N w_i \left\{ y_i - f(\theta) \right\} f_{i,j}(\theta). \quad (3.1.2)$$

is the j th component of δ_g . The $(k+1)$ -st approximation to θ^* , the minimizing point, can then be given by

$$\theta^{(k+1)} = \theta^{(k)} + \gamma^{(k)} \delta_g^{(k)} \quad (3.1.3)$$

where the superscript k is used to indicate dependence on $\theta^{(k)}$. The scalar

$\gamma^{(k)}$ is chosen to insure that

$$F(\theta^{(k)} + \gamma^{(k)} \delta_g^{(k)}) < F(\theta^{(k)}). \quad (3.1.4)$$

Under fairly mild conditions, it can be shown that this method is convergent. However, the gradient direction is only downhill locally. As a consequence, for many problems $\|\gamma^{(k)} \delta_g^{(k)}\|$ must be chosen so small that the convergence rate is painfully slow (see [4] for choices of $\gamma^{(k)}$).

3.2 Newton's Method

Newton's method is not directly applicable to the minimization of a function. So, we impose the necessary condition that θ^* be a minimum of F and seek the zero of ∇F . The method can then be described by

$$\begin{aligned} \theta^{(k+1)} &= \theta^{(k)} - \gamma^{(k)} \left\{ \square F(\theta^{(k)}) \right\}^{-1} \nabla F(\theta_k) \\ &= \theta^{(k)} + \gamma^{(k)} \left\{ \square F(\theta^{(k)}) \right\}^{-1} \delta_g^{(k)} \end{aligned} \quad (3.2.1)$$

where $\square F$ denotes the Hessian of F and

$$(\square F(\theta))_{ij} = \frac{\partial^2 F(\theta)}{\partial \theta_i \partial \theta_j} = F_{,ij}(\theta), \quad i, j = 1, 2, \dots, n, \quad (3.2.2)$$

denotes the element occupying the i th row and j th column of the Hessian of F .

It is easily seen that

$$(\square F(\theta))_{ij} = 2 \sum_{\ell=1}^N w_{\ell} \left\{ f_{\ell,i}(\theta) f_{\ell,j}(\theta) + [f_{\ell}(\theta) - y_{\ell}] f_{\ell,ij}(\theta) \right\} \quad (3.2.3)$$

where f_{ℓ} originally appeared on the right hand side of (2.2). We could recast (3.2.3) as

$$\square F(\theta) = 2 \left\{ J^T(\theta) W J(\theta) + \sum_{\ell=1}^N [f_{\ell}(\theta) - y_{\ell}(\theta)] w_{\ell} \square f_{\ell}(\theta) \right\} \quad (3.2.4)$$

where

$$(J(\theta))_{ij} = f_{i,j}(\theta) = \frac{\partial f_i(\theta)}{\partial \theta_j} \quad (3.2.5)$$

denotes the component of the Jacobian of $\underline{f} = [f_1, f_2, \dots, f_N]^T$ which occupies the i th row and j th column and which is evaluated at θ . Again $\gamma^{(k)}$ is a scalar so chosen as to force condition (3.1.4). This method is locally convergent, independent of $F(\theta^*)$, is ultimately of second order and can be used independent of the rank of J . On the other hand, $F(\theta)$ must be nonsingular, $J(\theta)$ is required, and so is $\square F(\theta)$ (in particular N matrices in $E^{n \times n}$ must be evaluated and, hence, one Newton step is expensive).

3.3 The Gauss-Newton Method

There are two ways for thinking of this method. The simpler is this. Drop the f_{ℓ} terms from (3.2.4) and apply (3.2.1). The other approach is this. Look at

$$\tilde{F}(\theta^{(k)} + \delta_t^{(k)}) = \sum_{i=1}^N w_i \left\{ y_i - f_i(\theta^{(k)}) - \sum_{m=1}^n f_{i,m}(\theta^{(k)}) \delta_{tm}^{(k)} \right\}^2 \quad (3.3.1)$$

where $\tilde{F}(\theta^{(k)} + \delta_t^{(k)})$ is an approximant to $F(\theta^{(k)} + \delta_t^{(k)})$ (see (2.2)) and is obtained by using the first two terms of the Taylor series for the squared terms in (2.2). If we minimize (3.3.1) with respect to $\delta_t^{(k)}$, then there are n necessary conditions,

$$\frac{\partial \tilde{F}(\theta^{(k)} + \delta_t^{(k)})}{\partial \delta_{tj}^{(k)}} = 0, \quad j = 1, 2, \dots, n,$$

which lead to the linear system

$$A^{(k)} \delta_t^{(k)} = \delta_g^{(k)} \quad (3.3.2)$$

where

$$A^{(k)} = 2J^T(\theta^{(k)}) WJ(\theta^{(k)}). \quad (3.3.3)$$

The formula for computing the kth step is then

$$\theta^{(k+1)} = \theta^{(k)} + \gamma^{(k)} \delta_t^{(k)}. \quad (3.3.4)$$

The Gauss-Newton method is locally convergent if[#] $\underline{\lambda}(J^T(\theta^*) WJ(\theta^*)) > K'(F(\theta^*))^{\frac{1}{2}}$, the convergence rate is ultimately one if the preceding condition holds and two if the condition holds and $F(\theta^*) = 0$. The Jacobian, J , is required and must be of full rank (A can be ill-conditioned, though) and there is usually some difficulty if the residuals are large. Unfortunately this is often the case in real life problems [3], [15], [16].

There are many nuances to the Gauss-Newton method. Hartley [6] suggested that one use $\gamma = 1, 1/2$, and $1/4$, use quadratic interpolation and, if $F(\theta^{(k+1)}) < F(\theta^{(k)})$ then to move to the new point. Otherwise, use $\gamma = 1/4, 1/8, 1/16$, and repeat the process.

Levenberg [8] suggested a damped Taylor series or Gauss-Newton method. Essentially, his idea was to replace $\tilde{F}$ by

$$R(\theta^{(k)} + \delta^{(k)}) = V\tilde{F}(\theta^{(k)} + \delta^{(k)}) + \sum_{k=1}^n C_k \delta_k \quad (3.3.5)$$

where V and C_k are positive scalars. The idea was to stabilize the Gauss-Newton method. Indeed, the Levenberg variant was the precursor of the interpolatory methods to be described now and later.

[#] $\underline{\lambda}(A) = \min \{\lambda_i(A) : i = 1, 2, \dots, n\}$, where the $\lambda_i(A)$ are the eigenvalues of A , and $K'(F(\theta^*))^{\frac{1}{2}}$ is a strict lower bound for the spectrum of A^* .

3.4 The Marquardt or Levenberg-Marquardt Method [9]

Marquardt made some interesting observations: The gradient method converges, but is too slow. The Gauss-Newton method allows big steps, but can lead to iterations in the wrong region or to divergence. The gradient vector and Gauss-Newton vector usually lie between 89° and 90° of one another, i.e., if[#]

$$\cos \alpha = \frac{\langle \delta_t, \delta_g \rangle}{\|\delta_t\| \|\delta_g\|}, \quad \text{then} \quad 89^\circ \leq \alpha \leq 90^\circ.$$

Marquardt employed some of Levenberg's ideas to construct a method of the following type. Consider

$$(A^{(k)} + \mu^{(k)} I) \delta_M^{(k)} = \delta_g^{(k)}, \quad (3.4.1)$$

where $\mu^{(k)}$ is a scalar, and

$$\theta^{(k+1)} = \theta^{(k)} + \delta_M^{(k)} \quad (3.4.2)$$

Note that for $\mu^{(k)} = 0$, $\delta_M^{(k)} = \delta_t^{(k)}$ as in the Gauss-Newton method and $\|\delta_M^{(k)}\| \rightarrow 0$ as $\mu^{(k)} \rightarrow \infty$. Further it can be shown that

$$\frac{\delta_M^{(k)}, \delta_g^{(k)}}{\|\delta_M^{(k)}\| \|\delta_g^{(k)}\|} \rightarrow 1 \quad \text{as} \quad \mu^{(k)} \rightarrow \infty.$$

This method is fairly stable; for each k and each $\mu^{(k)}$ (3.4.1) must be solved; no real analysis of correct incremental vector length was given; and the parameters must be carefully chosen for large residual problems. The Jacobian J is required; Gauss-Newton convergence properties are obtained if $\mu^{(k)} = O(\|[f_1(\theta^{(k)}), f_2(\theta^{(k)}), \dots, f_N(\theta^{(k)})]\|)$; and the linear system can usually be solved easily.

This method is really an interpolatory method, with a step somewhere between δ_g and δ_t . The incremental length in the δ_g direction is zero.

[#]We shall use $\langle a, b \rangle \equiv a^T b \equiv \sum_{i=1}^n a_i b_i$ when $a = [a_1, \dots, a_n]$ and $b = [b_1, \dots, b_n]$.

3.5 SPIRAL [7]

The paper which describes SPIRAL is not very rigorous mathematically. SPIRAL is an interpolatory method and, at each iteration, the algorithm attempts a full step in the δ_t direction. If F is not reduced, step size is reduced and the direction is rotated toward that of δ_g . The process is continued until F is reduced. For certain test problems, SPIRAL appears to perform better than Marquardt's algorithm and it does allow a step in the δ_g direction of positive length.

3.6 Powell's 1970 Methods [10],[11],[13]

Powell has two methods which are called his 1970 methods. We shall designate by 1970a the method for solving systems of nonlinear equations [10], [11] and 1970b the method for the unconstrained minimization of a functional [13]. The basic idea for both methods is this. Write

$$\phi(\theta^{(k)} + \delta^{(k)}) = F(\theta^{(k)}) - \langle \delta_g^{(k)}, \delta^{(k)} \rangle + \frac{1}{2} \langle \delta^{(k)}, B^{(k)} \delta^{(k)} \rangle, \quad (3.6.1)$$

where

$$\delta^{(k)} = \beta \delta_g^{(k)} + (1 - \beta) H^{(k)} \delta_g \quad (3.6.2)$$

and

$$H^{(k)} = [B^{(k)}]^{-1}. \quad (3.6.3)$$

If the direction of $\delta^{(k)}$ is sufficiently different from $\delta_t^{(k)} = B^{(k)} \delta_g^{(k)}$, then β is so chosen as to minimize 3.6.1). Otherwise, the δ_t direction is used.

Ad hoc rules are used to determine the length of $\gamma^{(k)} \delta^{(k)}$, where

$$\theta^{(k+1)} = \theta^{(k)} + \gamma^{(k)} \delta^{(k)}.$$

Ordinarily $\gamma^{(k)} = 1/2, 1$, or 2 .

The matrices $B^{(k)}$ and $H^{(k)}$ are symmetric matrices which are updated at the end of each successful step. The update consists of a complicated looking but easily calculated update. $B^{(k)}$ is an approximation to $\square F(\theta^{(k)})$. In 1970a, $B^{(1)}$ is the finite difference approximation to $J^T(\theta^{(1)}) W J(\theta^{(1)})$ and $H^{(1)} = [B^{(1)}]^{-1}$. In 1970b, $B^{(1)} = [H^{(1)}]^{-1}$ and $B^{(1)}$ is a scalar matrix. Powell gives some numerical results, which indicate the methods are certainly worth consideration [10], [11], [13].

3.7 Summary of Methods Descriptions

We can summarize the above by noting that each method, in its own way, makes an attempt at minimizing

$$\tilde{F}(\theta^{(k)} + \delta^{(k)}) = F(\theta^{(k)}) - \langle \delta_g^{(k)}, \delta^{(k)} \rangle + \frac{1}{2} \langle \delta^{(k)}, A^{(k)} \delta^{(k)} \rangle. \quad (3.7.1)$$

We simply note that the following table is appropriate.

Table 3.7.1

Method	Choice of Matrix $A^{(k)}$
GRADIENT	I
NEWTON	$2J^T(\theta^{(k)}) W J(\theta^{(k)}) + 2 \sum_{\ell=1}^N [f_{\ell}(\theta^{(k)}) - y_{\ell}] w_{\ell} \square f_{\ell}(\theta^{(k)})$
NEWTON-GAUSS	$2J^T(\theta^{(k)}) W J(\theta^{(k)})$
MARQUARDT	$\mu^{(k)} I + 2J^T(\theta^{(k)}) W J(\theta^{(k)})$

The Marquardt method is one of four interpolatory methods. SPIRAL, 1970a, 1970b, and the Marquardt algorithm all seek to minimize F by moving in a direction between δ_t and δ_g . Each seeks to combine good features of the Gauss-Newton and gradient methods and each succeeds, to a certain extent.

In passing we remark that there is substantial computational evidence to justify the use of finite difference approximations in lieu of exact analytic derivatives in Jacobians and Hessians.

Discussions of methods for least squares can be found in [3], [15], and [16]. Indeed much of what is given in Section 3 is a mixture from these references.

4. A NEW METHOD (See [15], [16], [17].)

Before embarking on a description of a new method, we simply observe that some basic questions for an interpolatory algorithm are:

- 1) How long is its increment?
- 2) What direction does the increment take?
- 3) Does $\tilde{F}$ really approximate F ?
- 4) Does it work?
- 5) It is practical?

We feel that the new algorithm addresses itself to the questions and answers 3) - 5) affirmatively.

4.1 Some Basic Theory

The following two definitions are probably familiar.

DEFINITION 4.1.1 Let X be a set in E^n . Then X is *convex* if and only if for every u in X , for every v in X and every $0 < a < 1$, $Z(a) = au + (1-a)v \in X$.

DEFINITION 4.1.2 Let $F : X \rightarrow E$. F is *convex* on $X \subset E^n$ if and only if for every $u \in X$ and for every $v \in X$ and for every $0 \leq a < 1$,

$$F(au + (1 - a)v) \leq aF(u) + (1 - a)F(v).$$

We note that for some Z on the abstract line connecting $\theta^{(k)}$ and $\theta^{(k+1)} = \theta^{(k)} + \delta^{(k)}$ ($F \in C^2(X)$, of course)

$$F(\theta^{(k+1)}) = S(\theta^{(k+1)}) \equiv F(\theta^{(k)}) - \langle \delta_g^{(k)}, \delta^{(k)} \rangle + \frac{1}{2} \langle \delta^{(k)}, \square F(Z) \delta^{(k)} \rangle \quad (4.1.1)$$

and $\square F(\theta)$ is positive semi-definite on X if and only if F is convex on X . Also, it does not make much sense to speak of a relative minimum of F on a set X unless F is convex there. Further since we are speaking of interpolatory methods, we will want X to be convex. The following is not standard.

DEFINITION 4.1.3 $F : X \rightarrow E$ is *bounded-convex* on the convex set $X \in E^n$ if and only if $F \in C^2(X)$ and $\square F$ has bounded positive spectrum on X .

If X is compact, then convexity is equivalent to bounded convexity. The example $F(\theta) = \theta^2$, $\theta \in \mathbb{R}$ serves to illustrate that F can satisfy Definition 4.1.3, and still remain unbounded.

The following theorem can be established in several different ways [4], [15], [17] and serves to illustrate why there has been interest in approximations of type (3.7.1) and methods which use the iteration

$$\theta^{(k+1)} = \theta^{(k)} + \gamma^{(k)} \delta^{(k)} \quad (4.1.2)$$

where

$$A^{(k)} \delta^{(k)} = \delta_g^{(k)}. \quad (4.1.3)$$

THEOREM 4.1.1 [16]

- (H1) $F : E^n \rightarrow E$ is bounded convex on the closed bounded set X in E^n .
 (H2) $A : X \rightarrow E^{n \times n}$.
 (H3) $0 < \lambda_i(A(\theta)) \leq M < \infty$ for every $\theta \in X$, some $M \in R$, and $i = 1, 2, \dots, n$.
 (H4) $\delta^{(k)}$ satisfies (4.1.3).
 (H5) If $\theta^{(k)} \in X$, then $\theta^{(k)} + \delta^{(k)} \in X$.

Then for every $\theta^{(0)} \in X$, there are sequences $\{\theta^{(k)}\} \in X$ and $\{\gamma^{(k)}\} \in [R]$ such that

$$(C1) \quad \theta^{(k+1)} = \theta^{(k)} + \gamma^{(k)} \delta^{(k)}, \quad \gamma^{(k)} > 0, \quad k = 0, 1, \dots;$$

$$(C2) \quad F(\theta^{(k+1)}) \leq F(\theta^{(k)}), \quad \text{for all } k, \text{ and} \\ F(\theta^{(k+1)}) = F(\theta^{(k)}), \quad \text{if and only if } \delta_g^{(k)} = 0;$$

and

$$(C3) \quad \theta^{(k)} \rightarrow \theta^* \in X, \quad \text{where } F(\theta^*) < F(\theta) \quad \text{for } \theta \neq \theta^* \text{ and } \theta \in X.$$

The theorem which describes the mechanism for the new method utilizes the following unorthodox definition.

DEFINITION 4.1.4 [15],[16],[17] The $n \times n$ matrix A is of *type H1* if and only if it is real, symmetric, and has distinct eigenvalues greater than two. [#]

Again, we use (3.7.1), but with $\delta^{(k)}$ defined by

$$\begin{aligned} \delta^{(k)} &= \gamma^{(k)} \delta_p^{(k)} = \gamma^{(k)} [p\delta_g + (1-p)\delta_t] \\ &= \gamma^{(k)} \left\{ pI + (1-p)[A^{(k)}]^{-1} \right\} \delta_g. \end{aligned} \quad (4.1.4)$$

[#]The subset P of positive definite matrices with distinct eigenvalues in $E^{n \times n}$ is dense in the set of positive semi-definite matrices in $E^{n \times n}$. Finally matrices of type H1 can be obtained by scaling the members of P .

THEOREM 4.1.2 [15],[16],[17]

(H1) $A^{(k)}$ is of type H1.

(H2) $A^{(k)}\delta_g^{(k)} \neq \lambda\delta_g$.

Then there is an r_0 , $0 < r_0 < 1$, such that for any r , $r_0 \leq r < 1$, there exists an ordered pair (γ, p) with the following properties

(C1) $\|r\delta_t^{(k)}\| = \|\gamma^{(k)}\delta_p^{(k)}\|$ where $A^{(k)}\delta_t^{(k)} = \delta_g^{(k)}$ and if $r = 1$, $(\gamma, p) = (1, 0)$;

(C2) $\frac{\partial \tilde{F}(\theta^{(k)} + \delta_p^{(k)})}{\partial p} = 0$;

(C3) as r decreases, $\delta_p^{(k)}$ moves away from $\delta_t^{(k)}$ and toward $\delta_g^{(k)}$, i.e., $p \geq 0$ and p increases as r decreases;

(C4) $\delta_p^{(k)}$ is always within 90° of both $\delta_t^{(k)}$ and $\delta_g^{(k)}$; and

(C5) r_0 may be so chosen that $r = r_0$, $\delta_p = \delta_g$.

Proofs of Theorems 4.1.1 and 4.1.2 are not given here since they are fairly technical. However, they may be found in [15] and [17]. We note that since we take $r = 1$ to be the maximum allowed value of r for least squares work, C1 connects the length of a (perhaps shortened) Gauss-Newton step and the step of the algorithm. Next C2 indicates that $\tilde{F}$ is really minimized in some sense and the remaining hypotheses simply describe the nature of the interpolant. In summary, the method picks both increment direction and length (subject to some constraint) so as to minimize $\tilde{F}$, the approximant to F , in some sense. The connection between the new method and Theorem 4.1.1 is simple. The matrix $\{pI + (1-p)[A^{(k)}]^{-1}\}$ in (4.1.4) plays the role of A in (4.1.3) and Theorem 4.1.1. As a consequence, convergence for the new method is established. Note that A need not be a good approximation to $\square F(\theta)$ in either case!

The equations describing the required entities and the mechanism are as follows.

To obtain p , γ , and r_0 , define

$$\sigma \equiv \delta_g - \delta_t, \quad (4.1.5)$$

$$\eta \equiv \langle \sigma, A\sigma \rangle / \langle \delta_g, \sigma \rangle, \quad (4.1.6)$$

$$b_1 \equiv \langle \delta_t, \sigma \rangle / \|\delta_t\|^2, \quad (4.1.7)$$

$$b_2 \equiv \|\sigma\|^2 / \|\delta_t\|^2, \quad (4.1.8)$$

$$r_1^2 \equiv (b_2 - b_1)^2 / (b_2 - 2b_1\eta + \eta^2), \quad (4.1.9)$$

$$r_2^2 \equiv b_2 / \eta^2, \quad (4.1.10)$$

$$r_3^2 \equiv (1 + 2b_1 + b_2) / (1 + \eta^2), \quad (4.1.11)$$

and

$$r_0 = \max \{r_1, r_2, r_3\}. \quad (4.1.12)$$

Notice that at $r = r_0$, $\delta_p = \delta_g$ and that dependence on $\theta^{(k)}$ has been suppressed for brevity.

In actual execution (see Fig. 4.2.1) two possibilities arise.

Case I. $r \geq r_0$. Define

$$a_1(r) \equiv b_1 - \eta r^2, \quad (4.1.13)$$

$$a_2(r) \equiv b_2 - \eta^2 r^2, \quad (4.1.14)$$

$$v(r) = \left\{ (b_1^2 - b_2) + (b_2 - 2b_1\eta + \eta^2)r^2 \right\}, \quad (4.1.15)$$

$$p \equiv p(r) = -[a_1(r) + \{v(r)\}^{1/2}] / a_2(r), \quad (4.1.16)$$

$$\gamma = \gamma(r) = (1 + \eta p)^{-1}. \quad (4.1.17)$$

Case II. $0 \leq r \leq r_0$. We simply use (4.1.13) - (4.1.17) to define

$$p = p_0 = p(r_0) \quad (4.1.18)$$

and

$$\gamma = (r/r_0)[1 + \eta p_0]^{-1}. \quad (4.1.19)$$

That is, if $r \leq r_0$, set $r = r_0$ so that Theorem 4.1.2 applies.

4.2 The Algorithm

The accompanying flow chart (Figure 4.2.1) indicates that certain program constants are required [16]. Some choices for these are given in Table 4.2.1. We also take

$$r = \min \left\{ d / \|\delta_t\|^*, r_{\max} \right\} \quad (4.2.1)$$

where $r_{\max} = 1$ for a least squares problem and $\|\cdot\|^*$ denotes a pseudo-norm.

Thus far, we have successfully used

$$\|\delta_t\|^* = \begin{cases} \max \left\{ |\delta_{tj}| / (|\theta_j| + \epsilon_j) : j = j_1, j_2, \dots, j_i \right\} \\ \max \left\{ |\delta_{tj}| : j = j_1, j_2, \dots, j_i \right\} \\ \|\delta_t\|_{\infty} \end{cases} \quad (4.2.2)$$

Table 4.2.1

Condition	Reasonable Choice
$\epsilon' > 0$	$\epsilon' = 1/2$
$c_0 > 2$	$c_0 = 2.5$
$c_1 > 1$	$c_1 = 3.0$
$v < 1$	$v = 1/2$

Here, $j_1, j_2, \dots, j_i$ denotes some set of components in which we have special interest. We could take $j_1 = 1, \dots, j_i = n$ or we could leave linear functions f_i out of consideration. Next, $d > 0$, can be specified by the user or, if not, it is determined by the code. This parameter is a bound on $\|\delta_t\|^*$.

Figure 4.2.1. Minimizing nonlinear functionals.

Notice that we only require A to be of type (H1). We have usually taken $A = 2\tilde{J}^T W \tilde{J}$ where $\tilde{J}$ denotes a finite difference approximation to J . In particular, we used

$$(\tilde{J}(\theta))_{ij} = \frac{f_i(\theta + e_j(\Delta\theta_j)) - f(\theta)}{\Delta\theta_j} \quad (4.2.3)$$

with e_j the j th Euclidean basis element and

$$\Delta\theta_j = \begin{cases} 1.0E - 10, & \text{if } |\theta_j| < 1.0E - 10 \\ |\theta_j| \cdot 0.00001, & \text{otherwise} \end{cases} \quad (4.2.4)$$

For general minimization problems several approximations to the Hessian $\square F$ have been used, including the main diagonal of the central difference approximation to the Hessian [15].

4.3 Numerical Results

This method has been used to analyze fission neutron spectra. For such problems, the new method outperformed the Bettis Atomic Power Laboratory variant of Marquardt's method. For a description of the problem, see [16]. Next we sketch some classical test problems. These problems have pathologies which are usually built-in to cause someone's algorithm to fail.

In the last line of Table 4.3.1, $r(\theta_1, \theta_2) = [\theta_1^2 + \theta_2^2]^{\frac{1}{2}}$

and

$$2\pi\phi(\theta_1, \theta_2) = \begin{cases} \tan^{-1}(\theta_2/\theta_1), & \theta_1 \geq 0 \\ \pi + \tan^{-1}(\theta_2/\theta_1), & \theta_1 < 0 \end{cases}$$

In determining the effectiveness of methods, we use equivalent function evaluations (E.F.E.). We count one evaluation whenever the function is evaluated by itself and two when the Jacobian is approximated by finite differences. For the

Table 4.3.1

Problem Name	$F(\theta)$	Starting Point	Solution
Rosenbrock's Parabolic Valley I [14]	$100(\theta_2 - \theta_1^2)^2 + (1 - \theta_1)^2$	$(-1.2, 1.0)$	$(1, 1)$
Rosenbrock's Parabolic Valley II [7]	$100(\theta_2 - \theta_1^2)^2 + (1 - \theta_1)^2$	$(-0.86, 1.14)$	$(1, 1)$
Zangwill's Problem I [18]	$\{(\theta_1 + \theta_2 - \theta_3)^2 + (\theta_1 + \theta_3 - \theta_2)^2 + (\theta_2 + \theta_3 - \theta_1)^2\}$	$(1/2, 1, 1/2)$	$(0, 0, 0)$
Zangwill's Problem II [7]	$\{(\theta_1 + \theta_2 - \theta_3)^2 + (\theta_1 + \theta_3 - \theta_2)^2 + (\theta_2 + \theta_3 - \theta_1)^2\}$	$(100, -1, 2.5)$	$(0, 0, 0)$
Powell's Quartic Function I [12]	$\{(\theta_1 + 10\theta_2)^2 + 5(\theta_3 - \theta_4)^2 + (\theta_2 - 2\theta_3)^4 + (\theta_1 - \theta_4)^4\}$	$(3, -1, 0, 1)$	$(0, 0, 0, 0)$
Powell's Quartic Function II [7]	$\{(\theta_1 + 10\theta_2)^2 + 5(\theta_3 - \theta_4)^2 + (\theta_2 - 2\theta_3)^4 + (\theta_1 - \theta_4)^4\}$	$(3, 2, 1, 3)$	$(0, 0, 0, 0)$
Fletcher & Powell Helical Valley [5]	$\{100[(\theta_3 - 10\phi(\theta_1, \theta_2))^2 + [r(\theta_1, \theta_2) - 1]^2] + \theta_3^2\}$	$(-1, 0, 0)$	$(1, 0, 0)$

simple problems above, the total number of equivalent function evaluations to solve the whole set of problems in Table 4.3.1 is given by Table 4.3.2 (see [16] for a more detailed breakdown). We note that SPIRAL performed better than the new algorithm in the solution of the two Rosenbrock problems. Otherwise the new algorithm uniformly performed better than the others.

Table 4.3.2

Algorithm	Marquardt (Bettis A.P.L.)	Marquardt [7]	SPIRAL [7]	This Work [16]
Total E.F.E. [#]	503	496	247	172

Since our algorithm has been compared to 1970a we now turn to some examples given by Powell.

Powell's badly scaled example is precisely what the name implies [10, pp. 146-147].

$$\begin{cases} f_1(\theta) = 1 - 10,000 \theta_1 \theta_2 \\ f_2(\theta) = 1.0001 - [\exp(-\theta_1) + \exp(-\theta_2)] \\ y_1 = y_2 = 0 \\ \theta(0) = (0,1) \end{cases} \quad (4.3.1)$$

The true solution is given as (1.098E-05, 9.106). In 223 function evaluations Powell obtained

$$(\theta_1, \theta_2) = (1.1064 \text{ E-}05, 9.0380), F(\theta_1, \theta_2) = 7.3881 \text{ E-}11.$$

The new method yielded

$$(\theta_1, \theta_2) = (1.098199 \text{ E-}05, 9.105729), F(\theta_1, \theta_2) = 8.7025667 \text{ E-}11$$

in 172 E.F.E. and 32 iterations. The convergence test was satisfied on the next iteration with 176 E.F.E.

Next Powell's cautionary example [11, pp. 90-93, 113] causes Newton's method to fail in the solution of a system of equations.

[#]In all calculations the convergence criterion we used was $\max \{ |\delta_{tk}^{(n)}| / (|\theta_k^{(m)}| + \tau_k) : k = 1, 2, \dots, n \} < \epsilon$ with $\tau_k = 0.001$ and $\epsilon = 0.0001$. Jones [7] used the same criterion.

$$\left\{ \begin{array}{l} f_1(\theta) = \theta_1 \\ f_2(\theta) = [10\theta_1/(\theta_1 + 0.1)] + 2\theta_2^2 \\ y_1 = y_2 = 0 \\ \theta^{(0)} = (3,1) \\ \theta^* = (0,0) \end{array} \right. \quad (4.3.2)$$

For this problem, 1970a took 60 E.F.E. to obtain $(\theta_1, \theta_2) = (-0.000072, 0.060228)$, where $F(\theta) = 1.45 \text{ E-}08$ and performance of the new algorithm is given in Table 4.3.3. Powell notes that 1970a took an erratic path to the listed solution.

Table 4.3.3

E.F.E.	Iteration Number	θ_1	θ_2	$F(\theta_1, \theta_2)$
11	2	4.698117 E-10	9.206140 E-06	2.223407 E-15
32	6	0.	5.001051 E-11	2.502103 E-41
36	7	Convergence Test Satisfied		

4.4 Some Comments

If equivalent function evaluations are the measure of performance of a least squares method, then it seems that the new method works well. So far, we have run it at Bettis Atomic Power Laboratory on a CDC 6600, at the University of Pittsburgh on a PDP-10, and at Lawrence Livermore Laboratory on a CDC 6600. Surprisingly there doesn't seem to be much difference between PDP-10 results and CDC 6600 results.

We currently plan to modify our existing code so that the linear system is solved more efficiently and $\min \{\lambda_i(A) : i = 1, \dots, n\}$ is estimated more

efficiently. Currently, we actually invert A and use the maximum absolute row sums of A^{-1} for these estimates.

Someday, we hope to try some of the update ideas suggested by Dennis [3]. We also hope to leave a clean code with documentation at the end of the summer. If this is not possible we will make a code with documentation available to Lawrence Livermore Laboratory as soon as we can.

5. ACKNOWLEDGMENTS

The author is indebted to Dr. Niel K. Madsen who made this summer visit at Lawrence Livermore Laboratory both possible and enjoyable and to Dr. Fred N. Fritsch who encouraged some of the work and the lectures.

Many computations supporting this work were done by Dr. Norman M. Steen at Bettis Atomic Power Laboratory in connection with his thesis [15]. Several examples (in particular the last two) were worked by the author at Lawrence Livermore Laboratory on a CDC 6600 and at the University of Pittsburgh on a PDP-10. In all cases a variation of FORTRAN was employed.

Finally the author wishes to thank Dr. Fritsch for his careful, constructive criticism of these notes.

REFERENCES

- [1] Cauchy, A. L. Methode generale pour la resolution des systems d'equations simultanees, Compte Rendus 25(1847) pp. 536-538.
- [2] Curry, H. B. The method of steepest descent for non-linear problems, Quart. Appl. Math., 2 (1944) pp. 258-261.
- [3] Dennis, Jr. J. E. Some computational techniques for the nonlinear least squares problem, in Numerical Solution of Systems of Nonlinear Algebraic Equations, G. D. Byrne and C. A. Hall, eds., Academic Press, N.Y., 1973, pp. 157-183.
- [4] Elkin, R. M. Convergence theorems for Gauss-Seidel and other minimization algorithms, Ph.D. thesis, University of Maryland, College Park, MD, 1968.
- [5] Fletcher, R. and Powell, M. J. D. A rapidly convergent descent method for minimization, Computer J. 6 (1963) pp. 163-168.
- [6] Hartley, H. O. The modified Gauss-Newton method for the fitting of nonlinear regression functions by least squares, Technometrics 3 (1961) pp. 269-280.
- [7] Jones, A. SPIRAL — A new algorithm for nonlinear parameters estimation using least squares, Computer J., 13 (1970) pp. 301-308.
- [8] Levenberg, K. A method for the solution of certain non-linear problems in least squares, Quart. Appl. Math. 2 (1944) pp. 164-168.
- [9] Marquardt, D. W. An algorithm for least squares estimation of nonlinear parameters, J. Soc. Indust. Appl. Math., 11 (1963) pp. 431-441.
- [10] Powell, M. J. D. A FORTRAN subroutine for solving systems of nonlinear algebraic equations, in Numerical Methods for Nonlinear Algebraic Equations, Rabinowitz, ed., Gordon and Breach, London, 1970, pp. 115-161.

- [11] Powell, M. J. D. A hybrid method for nonlinear equations, in Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz, ed., Gordon and Breach, London, 1970, pp. 87-114.
- [12] Powell, M. J. D. An iterative method for finding stationary values of a function of several variables without calculating derivatives, Computer J. 5 (1962) pp. 147-151.
- [13] Powell, M. J. D. A new algorithm for unconstrained optimization, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, 1970, pp. 31-65.
- [14] Rosenbrock, H. H. An automatic method for finding the greatest or least value of a function, Computer J., 3 (1960) pp. 175-184.
- [15] Steen, N. M. An algorithm for nonlinear parameter estimation, Ph. D. thesis, University of Pittsburgh, 1972.
- [16] Steen, N. M. and Byrne, G. D. The problem of minimizing nonlinear functionals, I. Least squares, in Numerical Solution of Systems of Nonlinear Algebraic Equations, G. D. Byrne and C. A. Hall, eds., Academic Press, 1973, pp. 185-239.
- [17] Steen, N. M. and Byrne, G. D. An algorithm for minimizing nonlinear functionals, in progress.
- [18] Zangwill, W. I. Minimizing a function without calculating derivatives, Computer J., 10 (1967) pp. 293-296.

DISTRIBUTION

LLL Internal

Technical Information Division (15 cys) L-9

J. S. Chang	L-71	L. J. Leonard	L-310
Richard Cooper	L-437	N. K. Madsen	L-310
R. P. Dickinson	L-310	D. H. Moore	L-533
S. Fernbach	L-61	W. J. O'Connell	L-313
F. N. Fritsch (20 cys)	L-310	R. L. Pexton	L-310
C. D. Gardner	L-310	R. A. Riviello	L-326
P. V. Grantham	L-24	M. E. Rundquist (2 cys)	L-310
M. A. Harrison	L-61	B. L. Shroyer	L-402
R. F. Hausman	L-310	D. R. Slaughter	L-518
Ted Harvey	L-313	N. G. Smiriga	L-310
N. W. Hetherington	L-421	J. R. Stedinger	L-523
A. C. Hindmarsh	L-518	J. B. Trenholme	L-216
R. H. Howell	L-518	R. E. von Holdt	L-61
J. I. Karush	L-307	H. H. Whiting	L-402
R. P. Koopman	L-313	Y. M. Wong	L-519

External

Library
Lawrence Berkeley Laboratory
University of California
Bldg 50, Room 134
Berkeley, CA 94720

Loren P. Meissner
Mathematics & Computing Group
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

Ron J. Detry
Applied Mathematics Division 2642
Sandia Laboratories
Sandia Base
Albuquerque, NM 87115

Larry F. Shampine
Numerical Analysis Division 5121
Sandia Laboratories
Sandia Base
Albuquerque, NM 87115

Robert J. Thompson
Applied Mathematics Division 5122
Sandia Laboratories
Sandia Base
Albuquerque, NM 87115

Fred W. Dorr
Division C-4
Los Alamos Scientific Laboratory
P. O. Box 1663
Los Alamos, NM 87544

Blair K. Swartz
Division C-6
Los Alamos Scientific Laboratory
P. O. Box 1663
Los Alamos, NM 87544

Robert E. Huddleston
Tom Jefferson
Jim Lathrop
Roy Lee
James N. Rogers
Numerical Applications Division 8441
Sandia Laboratories
P. O. Box 969
Livermore, CA 94550

W. J. Cody, Jr.
Wayne R. Cowell
Applied Mathematics Division
Argonne National Laboratory
9700 South Cass Avenue
Argonne, IL 60439

Software Services Section
Attn: Gene Anderson
Code 5137
Naval Weapons Center
China Lake, CA 93555

C. W. Gear
Computer Science Department
University of Illinois
Urbana, IL 61801

Richard F. Sincovec
Computer Science Department
Kansas State University
Manhattan, KS 66506

Henry C. Thacher, Jr.
Department of Computer Science
University of Kentucky
Lexington, KY 40506

Cleve B. Moler
Department of Mathematics
University of New Mexico
Albuquerque, NM 87106

Dr. R. J. Lambert
Department of Mathematics
400 Carver Hall
Iowa State University
Ames, IA 50010

Professor George D. Byrne (25 cys)
Department of Mathematics
804 Schenley Hall
University of Pittsburgh
Pittsburgh, PA 15260

George H. Miley, Room 214
Nuclear Engineering
University of Illinois
Urbana, IL 61801

George D. Byrne
Mathematics Department
University of Pittsburgh
Pittsburgh, PA 15213

Joseph F. Traub
Computer Science Department
Carnegie-Mellon University
Pittsburgh, PA 15213

Gene H. Golub
Computer Science Department
Stanford University
Stanford, CA 94305

Brian Ford
NAG Central Office
13 Banbury Road
Oxford
ENGLAND

Dr. Norman M. Steen
Bettis Atomic Power Laboratory (2 cys)
Naval Reactors Facility
Box 2068
Idaho Falls, ID 83407

Dr. Leonard Tornheim
Chevron Research Company
576 Standard Avenue
Richmond, CA 94801

Professor Davis M. Himmelblau
Chemical Engineering Department
University of Texas at Austin
Austin, TX 78712

Lee W. Lucas
Code 6075
Naval Weapons Center
China Lake, CA 93555

NOTICE

"This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately-owned rights."