SIAM REVIEW
Vol. 22, No. 1, January 1980

0036-1445/80/2201-0001$01.00/0

SOME RECENT APPROACHES TO SOLVING LARGE RESIDUAL NONLINEAR LEAST SQUARES PROBLEMS*

LARRY NAZARETH†

Abstract. Gauss–Newton based algorithms are widely used for solving nonlinear least squares problems. However, for reasons we shall discuss, they can be expected to perform poorly in certain circumstances. This leads to two other classes of algorithms corresponding to methods which implicitly take account of the second term in the Hessian of the function and methods which explicitly take account of the second term. We discuss the more promising approaches in these two areas and illustrate our discussion with a set of test results.

1. Introduction. The nonlinear least squares problem is

$$\underset{x\in\mathbb{R}}{\text{Minimize}} \quad F(x)=\frac{1}{2}\sum_{k=1}^{m}[f_k(x)]^2 \tag{1}$$

where each $f_k(x)$ is a nonlinear functional defined over $\mathbb{R}^n$.

The gradient and Hessian of $F(x)$ have a special form, often exploited by algorithms. They are given by

$$\nabla F(x)=J^T f, \tag{2}$$

where J is the $m\times n$ Jacobian matrix whose (i,j)th element is $J_{ij}=\partial f_i/\partial x_j$ and $f^T=(f_1,\cdots,f_m)$, both J and f being evaluated at the point x. The Hessian of F is

$$\nabla^2 F(x)=J^TJ+\sum_{k=1}^{m} f_kH_k, \tag{3}$$

where H_k is the $n\times n$ Hessian matrix of $f_k(x)$, whose (i,j)th element is $\partial^2 f_k/\partial x_i\partial x_j$.

Most nonlinear least squares methods are variants of Newton's method. Since the cost of providing the complete Hessian is often exorbitant, methods have been derived which can make do with first derivative information or just function values. Methods of this sort fall into three categories.

(a) Methods based on the Gauss–Newton algorithm, which do not take account of the second term $\sum_{k=1}^{m} f_kH_k$ of the Hessian (3). These are discussed in § 2, along with some results which indicate when they are likely to perform poorly.

(b) Methods which implicitly take the second term into account, discussed in § 3.

(c) Methods which, in some manner explicitly take the second term into account, discussed in § 4.

Finally, the results of some computer runs on a set of test functions are given in § 5.

2. Gauss–Newton methods and factors which influence their efficiency.

2.1. Gauss–Newton based methods. Gauss–Newton methods use the approximation $H_{gn}^{(k)}=J^{(k)T}J^{(k)}$ for the Hessian of the quadratic approximating or model function at the kth iteration. The corresponding search direction and subsequent iterate are

* Received by the editors November 14, 1977, and in revised form January 23, 1979. This work was performed under the auspices of the U.S. Energy Research and Development Administration.

† Applied Mathematics Division, Argonne National Laboratory, Argonne, Illinois 60439.

obtained from

$$J^{(k)T}J^{(k)}d^{(k)} = -J^{(k)T}f^{(k)},$$
$$x^{(k+1)} = x^{(k)} + \alpha^{(k)}d^{(k)} \tag{4}$$

where $\alpha^{(k)}$ is the step length.

The algorithm may be interpreted as solving a sequence of linear least squares problems, for which numerous stable techniques are available [1]. Given an approximation to the minimum, the primary goal of an iteration is to develop a search direction along which a sufficient decrease in function value can be obtained. Therefore, it is not imperative that the model function be determined by the exact, or a close approximation to the Jacobian. This gives an added dimension to algorithm development in that strategies of great cleverness can be devised (i) for reducing the amount of fresh information about the function required during an iteration of the algorithm, and (ii) for developing effective search directions. An example is the widely used VA05A method of Powell [3]. This uses Broyden's unsymmetric rank-1 updating of the Jacobian, and the "dog-leg" strategy for choosing a step in the subspace spanned by the Levenberg–Marquardt vector (discussed below) and the negative gradient vector, this step being restricted to lie within an adaptively updated "region of trust" of the model function.

In the Levenberg–Marquardt method, a term of the form $\mu^{(k)}D^{(k)}$ is added to the Gauss–Newton approximation $H_{gn}^{(k)}$, where $D^{(k)}$ is a diagonal matrix with positive diagonal entries, and $\mu^{(k)}$ is a positive scalar. Given a current estimate $x^{(k)}$ to the solution, the algorithm determines a search direction $d^{(k)}$ and subsequent iterate $x^{(k+1)}$ by

$$(J^{(k)T}J^{(k)} + \mu^{(k)}D^{(k)})d^{(k)} = -J^{(k)T}f^{(k)},$$
$$x^{(k+1)} = x^{(k)} + d^{(k)}. \tag{5}$$

$\mu^{(k)}$ is adaptively chosen, see e.g. [4], and four related reasons may be identified for using it:

(a) To maintain nonsingularity of the Hessian matrix of the model function.

(b) To conduct a search. Thus if $F(x^{(k+1)}) > F(x^{(k)})$, $\mu^{(k)}$ can be increased and the system (5) resolved.

(c) To permit adaptive choices between a Guass–Newton model function, and a "Cauchy" model function; i.e. as $\mu^{(k)}$ increases, the contours of the model function increasingly approach hyperspheres for suitable scaling of variables.

(d) To establish a "region of trust" of the model function and to step to an optimal point of the model function within this region of trust. To see this consider the following problem. Given a model function with Hessian $G(x^{(k)})$ of the form

$$M(x) = f(x^{(k)}) + g(x^{(k)})^T(x - x^{(k)}) + \tfrac{1}{2}(x - x^{(k)})G^{(k)}(x - x^{(k)})$$

we wish to find the point $z^{(k+1)}$ such that the reduction in function value $f(x^{(k)}) - M(x^{(k)})$ is maximized, subject to the constraint $(x - x^{(k)})D^{(k)}(x - x^{(k)}) \leqq \delta^{(k)}$. Thus we seek the solution to the problem

$$\begin{aligned} &\text{Minimize} \quad g(x^{(k)})(x - x^{(k)}) + \tfrac{1}{2}(x - x^{(k)})G^{(k)}(x - x^{(k)}) \\ &\text{subject to} \quad (x - x^{(k)})D^{(k)}(x - x^{(k)}) \leqq \delta^{(k)}. \end{aligned} \tag{6}$$

By forming the Lagrangian and finding its stationary point, we find the solution x^* to (6) as

$$x^* = x^{(k)} - (G^{(k)} + \mu^* D^{(k)})^{-1} g(x^{(k)}),$$

where μ^* is the corresponding Lagrange multiplier. Furthermore, if some $\mu^{(k)}$ other than μ^* is chosen, then the point

$$x^{(k+1)} = x^{(k)} - (G^{(k)} + \mu^{(k)} D^{(k)}) g(x^{(k)}) \tag{7}$$

solves the problem

$$\begin{aligned} &\text{Minimize} \quad g(x^{(k)})^T (x - x^{(k)}) + \tfrac{1}{2}(x - x^{(k)}) G^{(k)} (x - x^{(k)}) \\ &\text{subject to} \quad (x - x^{(k)}) D^{(k)} (x - x^{(k)}) \leqq (x^{(k+1)} - x^{(k)}) D^{(k)} (x^{(k+1)} - x^{(k)}), \end{aligned}$$

i.e. (7) is the solution within an altered region of trust.

In practice, because of the potential of conflict between these goals, it is best to introduce additional parameters (e.g. to permit search along $d^{(k)}$ or to control the region of trust), and to employ matrix factorizations of $J^{(k)}$. We shall not deal further with the various possible implementations.

We conclude this section by reviewing briefly two reasons why Gauss–Newton and Levenberg–Marquardt methods might perform poorly.

2.2. Rate of convergence analysis of Gauss–Newton and Levenberg–Marquardt methods. The following result, Meyer [5] points to one area of weakness of GN methods. Assuming $J^{*T} J^*$ nonsingular, define

$$r^* = \rho\left[(J^{*T} J^*)^{-1} \left(\sum_{i=1}^{m} f_i^* H_i^* \right) \right] \tag{8}$$

where $\rho[A]$ is the maximum absolute eigenvalue of A, and the quantities in (8) are evaluated at x^*, the optimum point. If $r^* < 1$, $\exists$ a neighborhood of x^* within which the full step GN algorithm (i.e. in (4), $\alpha^{(k)} = 1 \forall k$) is convergent at a linear rate, given by

$$\limsup [\|x^{(k+1)} - x^*\| / \|x^{(k)} - x^*\|] \leqq r^*. \tag{9}$$

Furthermore, McKeown [8] shows that for arbitrary r^*, linear convergence is retained by employing a suitable restriction on $\alpha^{(k)}$. (Similar results holds for the Levenberg–Marquardt method, given suitable restrictions on $\mu^{(k)}$ and $D^{(k)}$.)

The quantity r^* being large is thus one way of identifying a large residual problem. Note that scaling each of the functions $f_k(x)$ by the same multiplicative constant does not change the value of r^*. The value of r^* can be expected to be small and, in consequence, Gauss–Newton and Levenberg–Marquardt methods can be expected to perform well when any of the following conditions hold.

(i) residual f_i are small;
(ii) functions are close to linear, i.e., $\|H_i\|$ are small;
(iii) Oscillating signs between residuals result, on average, in cancellation between terms resulting in $\sum_{i=1}^{m} f_i H_i$ being small, even though individual terms are not small, see [9].

When $r^* = 0$, the GN method has at least superlinear convergence and when $f_i^* = 0$ the GN method has quadratic convergence.

2.3. Examinations of subspaces within which directions must lie [10]. We now look at the second reason why GN and LM methods may perform poorly. Let the

columns of $V^{(k)}$ be an orthonormal basis for the null space of $J^{(k)T}J^{(k)}$, of dimension $(n-t)$, i.e., $J^{(k)}V^{(k)}=0$.

Let the columns of $W^{(k)}$ be an orthonormal basis for the range of $J^{(k)T}J^{(k)}$, i.e., $W^{(k)T}V^{(k)}=0$. $V^{(k)}$ and $W^{(k)}$ can be obtained from the complete orthogonal factorization of

$$J^{(k)}=P^{(k)}\left[\begin{array}{c|c} R^{(k)} & 0 \\ \hline 0 & 0 \end{array}\right]Q^{(k)}$$

where $P^{(k)T}P^{(k)}=I$, $Q^{(k)T}Q^{(k)}=I$, and $R^{(k)}$ is a nonsingular upper triangular matrix.

$$Q^{(k)}=\left[\begin{array}{c} W^{(k)T} \\ \hline V^{(k)T} \end{array}\right].$$

Thus any vector $d^{(k)}$ can be written as $d^{(k)}=d_1^{(k)}+d_2^{(k)}$ where $d_1^{(k)}=W^{(k)}w^{(k)}$ and $d_2^{(k)}=V^{(k)}v^{(k)}$.

Substituting into (5), and assuming $\mu^{(k)}\neq 0$ gives

$$J^{(k)T}J^{(k)}W^{(k)}w^{(k)}+\mu^{(k)}W^{(k)}w^{(k)}+\mu^{(k)}V^{(k)}v^{(k)}=-J^{(k)T}f^{(k)}. \tag{10}$$

Multiplying on the left by $V^{(k)T}$ gives

$$\mu^{(k)}V^{(k)T}V^{(k)}v^{(k)}=0.$$

Thus $v^{(k)}=0$, and $d^{(k)}$ lies in the range of $J^{(k)T}J^{(k)}$.

On the other hand for the Newton step $\bar{d}^{(k)}$,

$$\left(J^{(k)T}J^{(k)}+\sum_{i=1}^{m}f_i^{(k)}H_i^{(k)}\right)\bar{d}^{(k)}=-J^{(k)T}f^{(k)}. \tag{11}$$

Writing $\bar{d}^{(k)}=\bar{d}_1^{(k)}+\bar{d}_2^{(k)}$, where $\bar{d}_1^{(k)}=W^{(k)}\bar{w}^{(k)}$ and $\bar{d}_2^{(k)}=V^{(k)}\bar{v}^{(k)}$, substituting to (11) and multiplying on the left by $V^{(k)T}$ gives

$$[V^{(k)T}(\Sigma f_i^{(k)}H_i^{(k)})V^{(k)}]\bar{v}^{(k)}=-V^{(k)T}(\Sigma f_i^{(k)}H_i^{(k)})W^{(k)}\bar{w}^{(k)}. \tag{12}$$

Thus in general $\bar{v}^{(k)}\neq 0$ and $\bar{d}^{(k)}$ is not in the range of $J^{(k)T}J^{(k)}$. Gill and Murray [10] argue that when $J^{(k)}$ is not of full column rank, information about the Hessian in the null space of $J^{(k)}$ should be used in determining the search direction. See also § 4.5.

3. Methods which implicitly take account of the second term in the Hessian.

3.1. Quasi-Newton methods. One alternative to the use of the specialized methods of § 2 is to use a quasi-Newton method, e.g., [11], [12], [13], [23]. Many are known to possess superlinear convergence [14], and McKeown [15] has shown a variety of examples for which the quality r^* in (8) is large and for which performance of QN methods is superior to that of the specialized methods. Search directions are developed from

$$B^{(k)}d^{(k)}=-J^{(k)T}f^{(k)} \tag{13}$$

where $B^{(k)}$ is an approximation to the Hessian, chosen to satisfy

$$B^{(k)}(x^{(k)}-x^{(k-1)})=(J^{(k)T}f^{(k)}-J^{(k-1)T}f^{(k-1)}) \tag{14}$$

and developed from $B^{(k-1)}$ using a rank 1 or rank 2 updating rule.

3.2. Nazareth's method [16]. Since two different model functions have been considered, it is natural to use a model function that combines the two through

adaptively chosen parameters. The search directions are thus developed from

$$[\phi^{(k)}J^{(k)T}J^{(k)}+(1-\phi^{(k)})B^{(k)}]d^{(k)}=-J^{(k)T}f^{(k)} \tag{15}$$

$0\leqq\phi^{(k)}\leqq 1$, and $B^{(k)}$ is maintained in factored form $\chi^{(k)T}\chi^{(k)}$.

The approximation implicitly employed for the second term is

$$S^{(k)}=(1-\phi^{(k)})(\chi^{(k)T}\chi^{(k)}-J^{(k)T}J^{(k)}).$$

The action of the Hessian of the model function on $\Delta x^{(k-1)}=x^{(k)}-x^{(k-1)}$ is given by

$$\begin{aligned}(\phi^{(k)}J^{(k)T}J^{(k)}+(1-\phi^{(k)})\chi^{(k)T}\chi^{(k)})\,\Delta x^{(k-1)}\\ =\Delta g^{(k-1)}+\phi^{(k)}(J^{(k)T}J^{(k)}\,\Delta x^{(k-1)}-\Delta g^{(k-1)})\end{aligned} \tag{16}$$

where

$$\Delta g^{(k-1)}=J^{(k)T}f^{(k)}-J^{(k-1)T}f^{(k-1)}.$$

Thus one has explicit control over the extent to which the Hessian of the model function disobeys the quasi-Newton relation, this being governed by examining at each step, the extent to which the Gauss–Newton model can be trusted, e.g., by comparing the actual reduction in function value against the predicted reduction a Gauss–Newton model would give.

There are several alternative implementations of this method. The implementation described in [16] introduces a parameter $\mu^{(k)}$ into the model as in the Levenberg–Marquardt method discussed in § 2 and for the reasons outlined there. Thus we solve

$$\begin{aligned}&((\phi^{(k)})J^{(k)T}J^{(k)}+(1-\phi^{(k)})\chi^{(k)T}\chi^{(k)}+\mu^{(k)}I)d^{(k)}=-J^{(k)T}f^{(k)},\\ &x^{(k+1)}=x^{(k)}+d^{(k)}\end{aligned} \tag{17}$$

and the main features of the algorithm may briefly be summarized as follows.

(i) A strategy for choosing $\phi^{(k)}$ and $\mu^{(k)}$ is based upon a comparison of predicted reduction in function value versus actual reduction, see e.g. [4].

(ii) The quasi-Newton method used is Davidon's optimally conditioned algorithm [13].

(iii) Each term in the model Hessian is positive semi-definite and the system of equations (17) may be solved in a numerically stable manner as the solution of a linear least squares problem.

(iv) The algorithm can be easily specialized to the quasi-Newton method ($\phi^{(k)}=0\ \forall k$) and to the Levenberg–Marquardt method ($\phi^{(k)}=1\ \forall\ k$).

(v) If $\phi^{(k)}\to 0$ as $k\to\infty$, we can expect superlinear convergence of the algorithm. Also if the initial quasi-Newton approximation is $J^{(1)T}J^{(1)}$ and $\mu^{(1)}=0$, then the algorithm will converge for linear problems in one step.

4. Methods which explicitly take account of second term in Hessian. We outline each basic method and subsequently discuss some features of the approach.

4.1. Brown–Dennis method. Approximate each $H_i^{(k)}$ by a quasi-Newton approximation $B_i^{(k)}$, i.e.,

$$B_i^{(k)}\,\Delta^{(k-1)}=\Delta g_i^{(k-1)}\quad\text{for each } i \tag{18}$$

where $\Delta g_i^{(k-1)}=\nabla f_i^{(k)}--\nabla f_i^{(k-1)}$ and $\nabla f_i^{(k)}$ is the ith row of $J^{(k)}$. $B_i^{(k)}$ is derived from

$B_i^{(k-1)}$ using a quasi-Newton updating rule. Then the search direction is given by solving

$$(J^{(k)^T}J^{(k)} + \sum_{i=1}^{m} f_i^{(k)} B_i^{(k)}) d^{(k)} = -J^{(k)^T} f^{(k)}.$$

The method has good convergence properties [18] but is mainly of theoretical interest since it requires the storage of m matrices each of $n(n+1)/2$ elements. The following method thus seeks to utilize only a single such matrix.

4.2. Broyden–Dennis method and Betts' method. Consider the step and the associated information about the function illustrated in Fig. 1.

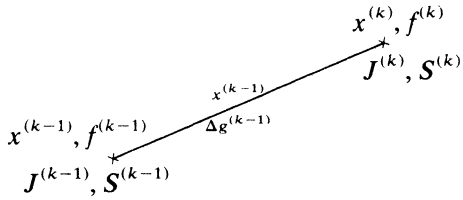

FIGURE 1

By Taylor's formulae

$$\nabla F(x^{(k)}) - \nabla F(x^{(k-1)}) = \bar{H}^{(k)}(x^{(k)} - x^{(k-1)}) \tag{19}$$

where $\bar{H}^{(k)} = \int_0^1 \nabla^2 F(x^{(k-1)} + t(x^{(k)} - x^{(k-1)}))\, dt$. Equation (19) corresponds to the QN relation. Since

$$H(x) = J(x)^T J(x) + R(x)$$

where

$$R(x) = \sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x)$$

we may write (19) as

$$\bar{R}^{(k)}\, \Delta x^{(k-1)} = (\Delta g^{(k-1)} - \bar{G}^{(k)}\, \Delta x^{(k-1)}) \equiv \Delta \tilde{g}^{(k-1)} \tag{20}$$

where

$$\bar{R}^{(k)} = \sum_{i=1}^{m} \int_0^1 f_i(x^{(k-1)} + t\, \Delta x^{(k-1)}) \nabla^2 f_i(x^{(k-1)} + t\, \Delta x^{(k-1)})\, dt$$

and

$$\bar{G}^{(k)} = \int^1 J^T(x^{(k-1)} + t\, \Delta x^{(k-1)}) J(x^{(k-1)} + t\, \Delta x^{(k-1)})\, dt.$$

Given an approximation $S^{(k-1)}$ to the second term, the next approximation $S^{(k)}$ is developed from $S^{(k-1)}$ using a QN updating rule such that $S^{(k)}$ satisfies (20) with $\bar{R}^{(k)}$ replaced by $S^{(k)}$. The model function for the next iteration has for its Hessian $(J^{(k)^T}J^{(k)} + S^{(k)})$, and the next search direction is given by solving

$$(J^{(k)^T}J^{(k)} + S^{(k)}) d^{(k)} = -J^{(k)^T} f^{(k)}. \tag{21}$$

Now, since $\Delta \tilde{g}^{(k-1)}$ is unknown, it must be approximated in some manner, and different ways of doing this give rise to different methods.

In the Broyden–Dennis method [18] we have

$$S_{\mathrm{BD}}^{(k)}\,\Delta x^{(k-1)} = \Delta \tilde{g}_{\mathrm{BD}}^{(k-1)} \equiv (\Delta g^{(k-1)} - J^{(k)^T} J^{(k)}\,\Delta x^{(k-1)}) \tag{22}$$

and $S_{\mathrm{BD}}^{(k)}$ is developed from $S_{\mathrm{BD}}^{(k-1)}$ using Powell's symmetric rank 2 update [20]. Note from (22) that the model function Hessian for the next iteration $(J^{(k)^T}J^{(k)} + S_{\mathrm{BD}}^{(k)})$ satisfies the usual QN relation.

In Bett's method [19] we have

$$S_{\mathrm{B}}^{(k)}\,\Delta x^{(k-1)} = \Delta \tilde{g}_{\mathrm{B}}^{(k-1)} \equiv (\Delta g^{(k-1)} - J^{(k-1)^T} J^{(k-1)}\,\Delta x^{(k-1)}) \tag{23}$$

and $S_{\mathrm{B}}^{(k)}$ is developed from $S_{\mathrm{B}}^{(k-1)}$ using Davidon's Symmetric Rank 1 update [21].

Finally we may approximate $\bar{G}^{(k)}$ in (20) by

$$\bar{G}_{\mathrm{D}}^{(k)} = \int_0^1 J^{(k-1)^T} J(x^{(k-1)} + t\,\Delta x^{(k-1)})\,dt = J^{(k-1)^T} \bar{J}^{(k-1)} \tag{24}$$

where

$$\bar{J}^{(k-1)} = \int_0^1 J(x^{(k-1)} + t\,\Delta x^{(k-1)})\,dt. \tag{25}$$

We shall now see that this gives the DGW method discussed in the next section.

Note also that other approximations to $\bar{G}^{(k)}$ are feasible, e.g., employing convex combinations of $J^{(k-1)}$ and $J^{(k)}$.

4.3. Dennis, Gay and Welsch (DGW) method. From (18) we have

$$f_i^{(k)} B_i^{(k)}\,\Delta x^{(k-1)} = f_i^{(k)}\,\Delta g_i^{(k-1)} \quad \forall\, i. \tag{26}$$

Writing

$$S_{\mathrm{D}}^{(k)} = \Sigma f_i^{(k)} B_i^{(k)}$$

we have

$$\begin{aligned} S_{\mathrm{D}}^{(k)}\,\Delta x^{(k-1)} &= \Sigma f_i^{(k)}\,\Delta g_i^{(k-1)} \\ &= J^{(k)^T} f^{(k)} - J^{(k-1)^T} f^{(k)} \\ &= \Delta g^{(k-1)} - J^{(k-1)^T} \bar{J}^{(k-1)}\,\Delta x^{(k-1)} \end{aligned} \tag{27}$$

where $f^{(k)} = f^{(k-1)} + \bar{J}^{(k-1)}\,\Delta x^{(k-1)}$ by Taylor's formulae and $\bar{J}^{(k-1)}$ is defined by (25).

Comparing (27) with (20) and (24) we see the equivalence of the two formulations of the DGW method.

4.4. Some observations on the above approaches. Note that $\sum_{i=1}^m f_i^{(k)} H_i^{(k)}$, is, in general, an indefinite matrix. If we use positive definite approximations $S^{(k)}$ to it, then (21) can be posed as a linear least squares problem and solved without explicitly forming the normal equations. However, to make such positive definite updates one must ensure that the somewhat artificial condition $\Delta \tilde{g}^{(k-1)^T}\,\Delta x^{(k-1)} > 0$ holds at each step. On the other hand, if we allow indefinite approximations to $\sum_{i=1}^m f_i^{(k)} H_i^{(k)}$, e.g., using Powell's symmetric rank 2 update, then (21) cannot always be posed as a linear

least squares problem; moreover $d^{(k)}$ may not be a descent direction, entailing further modification to the basic algorithm.

Such modifications are discussed in [22]. In particular an adaptation of Oren–Luenberger scaling of $S_D^{(k)}$ (termed sizing) has substantially improved performance.

Note also that the DGW Hessian approximation could be used in place of $B^{(k)}$ in the method discussed in § 3.2. In this case the Hessian of the model function is

$$J^{(k)T}J^{(k)}+(1-\phi^{(k)})S_D^{(k)},$$

and a sizing of $S_D^{(k)}$ has again been introduced in another way. Indeed the algorithms of §§ 3.2 and 4 may mutually benefit from a cross fertilization of the ideas underlying them.

4.5. Gill and Murray's method. We use the same notation as in § 2.3. In this algorithm $J^{(k)T}J^{(k)}$ is regarded as a suitable approximation to the Hessian in the space spanned by the columns of $W^{(k)}$. If $d_1^{(k)}$ is the Gauss–Newton vector and we write $\sum_i f_i^{(k)}H_i^{(k)}=\varepsilon B$, $\|B\|=1$, then

$$d_1^{(k)}=\bar{d}_1^{(k)}+O(\varepsilon).$$

However, for reasons similar to those discussed in § 2.3, $J^{(k)T}J^{(k)}$ is not a good approximation to the Hessian in the space spanned by the columns of $V^{(k)}$. Thus to the Gauss–Newton vector, $d_1^{(k)}$, is added an estimate of the component of the Newton vector in the space spanned by $V^{(k)}$, say $d_2^{(k)}$. The latter is obtained from the solution $v^{(k)}$ of the following approximation to (12)

$$\varepsilon V^{(k)T}BV^{(k)}v^{(k)}=-\varepsilon V^{(k)}BW^{(k)}w^{(k)} \tag{28}$$

where

$$w^{(k)}=W^{(k)T}d_1^{(k)}$$

and then

$$d_2^{(k)}=V^{(k)}v^{(k)}=\bar{d}_2^{(k)}+O(\varepsilon). \tag{29}$$

For details on how to cope with $(J^{(k)T}J^{(k)}+\varepsilon B)$ not positive definite and on the considerable savings in computation that can be achieved in approximating $\varepsilon BV^{(k)}$ see [10]. However, it is entirely possible that $J^{(k)T}J^{(k)}$ is not close to being singular while $\sum_{i=1}^m f_iH_i$ is significant, i.e., ε is not small. In this case no correction $d_2^{(k)}$ is made, whereas it may be advantageous to take the second term into account as some of the earlier algorithms do.

5. Numerical results and conclusions. We describe here the results of some simple numerical experiments obtained by extending the implementation of the hybrid method (N) [16] (which was flexibly designed specifically for this sort of experimentation) to include the methods of Broyden–Dennis (BD), Betts (B) and Dennis–Gay–Welsch (DGW) discussed in § 4. The model function in each case takes the form

$$(\alpha^{(k)}J^{(k)T}J^{(k)}+\beta^{(k)}S^{(k)}+\mu^{(k)}I)d^{(k)}=-J^{(k)T}f^{(k)} \tag{30}$$

where for the hybrid method $\alpha^{(k)}=\phi^{(k)}$, $\beta^{(k)}=(1-\phi^{(k)})$ and $S^{(k)}=\chi^{(k)T}\chi^{(k)}$ as described in § 3.

For the Broyden–Dennis, Betts, and Dennis–Gay-Welsch methods $\alpha^{(k)}\equiv 1$, $\beta^{(k)}\equiv 1\ \forall\ k$, and $S^{(k)}$ was developed from $S^{(k-1)}$ and chosen to satisfy the QN relations (22),

(23), and (27), respectively. In order to avoid further modifications to the algorithms (see § 4.4), we chose to maintain positive definite updates using Davidon's optimally conditioned technique, employing the same subroutine implementing this as was used in the hybrid method.

The hybrid method is specialized to Levenberg–Marquardt (LM) by setting $\alpha^{(k)} = 1$, $\beta^{(k)} = 0 \; \forall \, k$, and to the quasi-Newton method (DQN) by setting $\alpha^{(k)} = 0$, $\beta^{(k)} = 1 \; \forall \, k$.

Since the implementations differed only in the choice of the model functions, other portions, e.g., technique for choosing $\mu^{(k)}$, convergence criteria, etc., remaining the same, a fairly uniform comparison was possible.

The algorithms were run on a set of 11 test functions as listed below. We should emphasize that tests results on a few test functions should be viewed with extreme caution and that their primary uses are for discerning certain broad trends in performance, and in confirming what might be expected from the theoretical soundness of the algorithms.

TABLE 1

Problem	Dimension No. of f_i	H	LM	DQN	B	BD	DGW
1	2	35	28	48	34	31	27
	(2)	21	16	39	29	30	22
2	4	37	97	47	207	301	301
	(7)	21	71	34	204	301	301
3	4	14	9	28	14	18	15
	(4)	11	9	24	14	18	15
4	2	322	278	249	205	149	135
	(2)	160	143	207	178	144	101
5	3	21	15	38	24	20	31
	(3)	15	12	31	22	18	30
6	9	18	5	44	200	134	17
	(31)	11	5	39	200	134	17
7	3	24	26	24	37	35	27
	(10)	13	13	22	33	31	24
8	4	32	41	50	27	32	25
	(6)	28	24	42	22	32	24
9	2	22	18	2*	17	19	12
	(10)	16	13	2	13	17	11
10	2	11	25	17	11	8	8
	(2)	11	23	13	11	8	8
11	4	21	36	27	14	18	12
	(20)	16	28	19	13	14	12

* Found a different local minimum.

Test functions.

(a) *Zero residual.*

1. Rosenbrock's parabolic valley. See Brent [24, p. 139].
2. Wood's quartic. See Brent [24, p. 141].
3. Powell's quartic. See Brent [24, p. 141].
4. Powell's badly scaled function. See Powell [25, p. 146].
5. Fletcher's helical valley. See Brent [24, p. 140].
6. Watson's function, $n = 9$. See Brent [24, p. 142].

(b) *Large residual.*

7. Modified Box's exponential obtained by adding 20 to the 2nd function f_2 and 10 to f_4. This gives a residual of $\approx 0.308\text{E}+3$ at the solution. See Brent [24, p. 141].
8. Brown–Dennis. See Brown [26, p. 12].
9. Jennrich and Sampson problem. See Gill and Murray [27, p. 25].
10. Freudenstein and Roth problem. See Gill and Murray [27, p. 23].
11. Davidon problem. See Davidon [28] or Gill and Murray [27, p. 24].

In Table 1 above, the upper number in each entry is the number of function calls and the lower the number of Jacobian calls.

Convergence was declared when $\|\text{gradient}\| \leqq 10^{-4}$.

Further details of the implementation are given in Nazareth [29].

On small or zero residual problems the Levenberg–Marquardt algorithm performs very well, but our results for large residual problems show quite clearly that the algorithms of §§3 and 4 are superior. In particular the hybrid method (H) performs overall better than its specializations (LM and DQN) and the DGW method comes out a clear winner (even with positive definite approximations to the second term of the Hessian).

It is clear that much interesting work remains to be done in the area of nonlinear least squares. Here we have sought to provide an overall framework for some recent algorithms and reported computational experience. The way in which the algorithms of § 4.2 are presented is new, and we have also discussed some new, and perhaps interesting, variants on these algorithms in the remarks at end of §§ 4.2 and 4.4.

Acknowledgment. The author would like to thank Jorge Moré for many useful discussions on this subject, and Judy Beumer for her typing.

REFERENCES

[1] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[2] M. J. D. POWELL, *A method for minimizing a sum of squares of non-linear functions without calculating derivatives*, Comput. J., 7 (1965), pp. 303–307.

[3] ———, AERE Harwell Library Subroutine VA05A, Harwell, Oxfordshire, England, 1970.

[4] R. FLETCHER, *A modified Marquardt subroutine for non-linear least squares*, UKAEA Research Group Report, AERE R6799, Harwell, Oxfordshire, England, 1971.

[5] R. R. MEYER, *Theoretical and computational aspects of non-linear regression*, Non-linear Programming, Rosen, Mangasarian and Ritter, eds., Academic Press, New York–London, 1970.

[6] M. R. OSBORNE, *Some aspects of non-linear least squares calculations*, Numerical Methods for Non-linear Optimization, Lootsma, ed., Academic Press, New York–London, 1972.

[7] H. RAMSIN AND P. A. WEDIN, *A comparison of some algorithms for the non-linear least squares problem*, Report LU/CS-75:06, Lund University, Dept. of Computer Sciences, Solvegatan, Sweden, 1975.

[8] J. J. MCKEOWN, *On algorithms for sums of squares problems*, Toward Global Optimization, Dixon, ed., American Elsevier, New York, 1975.

[9] D. H. GLOECKNER, M. H. MACFARLANE AND S. C. PIEPER, *The use of first and second derivatives in optical model parameter searches*, Physics Division, Argonne National Laboratory, 1976. (Under review by Computer Physics Communications.)

[10] P. E. GILL AND W. MURRAY, *Non-linear least squares and non-linearly constrained optimization*, Numerical Analysis, Lecture Notes in Mathematics No. 506, Dundee, Springer-Verlag, Berlin–Heidelberg–New York, 1975.

[11] W. C. DAVIDON, *Variable metric method for minimization*, AEC Research and Development Report, ANL-5990, (Rev.), Argonne National Laboratory, Argonne, Il., 1959.

[12] C. G. BROYDEN, *Quasi-Newton methods and their application to function minimization*, Math. Comput., 21 (1967), pp. 368–381.

[13] W. C. DAVIDON, *Optimally conditioned optimization algorithms without line searches*, Math. Prog., 9, (1975), pp. 1–30.

[14] J. E. DENNIS AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 519–560.

[15] J. J. MCKEOWN, *Specialized versus general purpose algorithms for minimizing functions that are sums of squared terms*, Math. Prog., 9 (1975), pp. 57–68.

[16] J. L. NAZARETH, *A hybrid least squares method*, AMD TM-254, Appl. Math. Div., Argonne National Laboratory, 1975. (Rev.)

[17] K. M. BROWN AND J. E. DENNIS, *New computational algorithms for minimizing a sum of squares of non-linear functions*, Rept. 71–6, Computer Science Dept., Yale University, New Haven, CT, 1971.

[18] J. E. DENNIS, *Some computational techniques for the non-linear least squares problems*, Numerical Solution of Systems of Non-linear Algebraic Equations, Byrne and Hall, eds., Academic Press, New York–London, 1973.

[19] J. T. BETTS, *Solving the non-linear least squares problem: Application of a general method*, The Aerospace Corp., El Segundo, CA, 1975.

[20] M. J. D. POWELL, *A new algorithm for unconstrained optimization*, Nonlinear Programming, Mangasarian and Ritter, eds., Academic Press, New York.

[21] W. C. DAVIDON, *Variance algorithms for minimization*, Comput. J., 10 (1963), pp. 106–410.

[22] J. E. DENNIS, D. M. GAY AND R. E. WELSCH, *An adaptive non-linear least squares algorithm*, MIT Oper. Res. Center Report 142, 1977.

[23] J. E. DENNIS AND J. J. MORÉ, *Quasi-Newton methods—motivation and theory*, this Review, 19 (1977), pp. 46–89.

[24] R. P. BRENT, *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[25] M. J. D. POWELL, *A Hybrid method for non-linear equations*, Numerical Methods for Nonlinear Equations, Gordon & Breach, 1970, pp. 87–114.

[26] K. M. BROWN, *Computer oriented methods for fitting data in the linear and nonlinear least squares sense*, Proc. FJCC, Anaheim, CA, 1972.

[27] P. E. GILL AND W. MURRAY, *Algorithms for the solution of the non-linear least squares problem*, NPL Report NAC 71, National Physical Laboratory, England, 1976.

[28] W. C. DAVIDON, *New least squares algorithms*, JOTA, 18 (1976), pp. 187–198.

[29] J. L. NAZARETH, *Minkit—an optimization*, ANL-AMD Tech. Memo 305, Applied Math. Division, Argonne National Laboratory, 1977.