Distribution Category:
Mathematics and Computers
(UC-32)

ANL-80-106

ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, Illinois 60439

NEWTON'S METHOD WITH A MODEL TRUST-REGION MODIFICATION

by

D. C. Sorensen

Applied Mathematics Division

September 1980

TABLE OF CONTENTS

NEWTON'S METHOD WITH A MODEL TRUST-REGION MODIFICATION

by

D. C. Sorensen

ABSTRACT

A modified Newton method for unconstrained minimization is presented and analyzed. The modification is based upon the model trust region approach. This report contains a thorough analysis of the locally constrained quadratic minimizations that arise as subproblems in the modified Newton iteration. Several promising alternatives are presented for solving these subproblems in ways that overcome certain theoretical difficulties exposed by this analysis. Very strong convergence results are presented concerning the minimization algorithm. In particular the explicit use of second order information is justified by demonstrating that the iterates converge to a point which satisfies the second order necessary conditions for minimization. With the exception of very pathological cases this occurs whenever the algorithm is applied to problems with continuous second partial derivatives.

1. Introduction

The problem of minimizing a real valued function f of several real variables is generally attacked by some variant of Newton's method for finding a zero of the gradient of f. The term variant here is meant to include any method based upon maintaining an approximation to the Hessian matrix of mixed second order partial derivatives of f. When this matrix is actually computable, then Newton's method is probably the method of choice for the minimization problem.

As we shall point out in Section 3 there are several things to consider when attempting to provide a practical implementation of Newton's method for general use. Not the least of these is the problem of forcing convergence of the method when good initial guess at the solution is not available. The main purpose of this report is to describe and analyze a technique for the solution of this problem. The approach we shall present is well known. It is appropriately called a model trust region approach in that the step to a new iterate is obtained by minimizing a local quadratic model to the objective

function over a restricted ellipsoidal region centered about the current iterate. The diameter of this region is expanded and contracted in a controlled way based upon how well the local model predicts behavior of the objective function. It is possible to control the iteration in this way so that convergence is forced from any starting value assuming reasonable conditions on the objective function. In fact, we shall prove some very strong convergence properties for this method in Section 4. There it is shown that one can expect (but not ensure) that the iteration will converge to a point which satisfies the second order necessary conditions for a minimum.

The origin of this method properly lies with the work of Levenberg [14] and Marquardt [15] for nonlinear least squares calculations. The method was first discussed in connection with general minimization by Goldfeld, Quandt, and Trotter [11]. Powell [23] applied the modification in a more general situation of a quasi-Newton iteration. Hebden [12] made some important computational observations. This paper is most heavily influenced by the work of Moré [18] for the nonlinear least squares case.

The current interest stems from several recent efforts to obtain a practical implementation of a Modified Newton method that takes full advantage of the second order information. Several of the more recent works have attempted to explicitly use directions of negative curvature to accomplish various tasks such as escape from saddle points [5,9,13], search along more general paths [10,16,19,20,28], obtain convergence to points that satisfy second order necessary conditions [10,16,19,20] etc. We observe along with Gay [7,8] that the method proposed here will accomplish these things in a very elegant and intuitively appealing way.

It is hoped that this report will present a succinct but thorough analysis of this method. In particular we feel it is important to clearly describe the theoretical nature of the locally constrained quadratic minimization in Section 2. The analysis given in Section 4 is made sufficiently general to apply to several possible implementations. These possibilities are described in Section 5 where particular attention is paid to overcoming a practical problem of implementation exposed by the theoretical discussion in Section 2. We make an effort to offer several alternatives to implementation but shall make no recommendations until there is numerical evidence to present.

2. Constrained Quadratic Minimization

An important portion of the unconstrained minimization procedure presented in Section 3 will be concerned with the solution of the following problem:

(2.1) Let $\psi(w) = f + g^T w + \frac{1}{2} w^T B w$. Find $p \in \mathbb{R}^n$ such that
$\psi(p) = \min\{\psi(w): \|w\| \leq \Delta\}$.

In (2.1) $B = B^T \in \mathbb{R}^{n\times n}$; $w,g \in \mathbb{R}^n$; $f,\Delta \in \mathbb{R}$ with $\Delta > 0$, and $\|\cdot\|$ throughout is the 2-norm. There are some important subtleties to this problem. The purpose of this section is to give a complete discussion of the theoretical aspects of problem (2.1) and to expose the nature of the computational difficulties that may be present.

Several authors have considered problem (2.1) or related problems. This problem appears implicitly as a subsidiary calculation in Levenberg-Marquardt type algorithms for nonlinear least squares [14,15]. The computational aspect of this calculation was fully discussed by Moré in [18]. A relatively early paper by Forsythe and Golub [6] considers a closely related problem concerning minimization of the form

(2.2) $\min\{(x-b)^T A(x-b): \|x\| = 1\}$.

While their work gives an extensive study of problem (2.2), it is not fully applicable to problem (2.1) since $g \in$ range(B) may not hold, and the interior is not considered. Problem (2.1) first appeared as a subsidiary calculation in unconstrained minimization in the work of Goldfeld, Quandt, and Trotter [11]. Helden [12] made an important contribution concerning the practical computation of a solution to (2.1). More recently the problem has been discussed by Gay [7].

If the method of Lagrange is applied to the equivalent problem

(2.3) $$\begin{aligned} &\min \psi(w) \\ &\text{s.t. } w^T w \leq \Delta^2 \end{aligned}$$,

it is a straightforward conclusion of the first order necessary conditions that p solves (2.3) and hence (2.1) only if p satisfies an equation of the

form $(B+\lambda I)p = -g$ with $\lambda \geq 0$ the Lagrange multiplier associated with the constraint $w^Tw \leq \Delta^2$.

Lemma (2.4): If p is a nonzero solution to (2.1) then p is a solution to an equation of the form

$$(2.5) \qquad (B+\lambda I)p = -g$$

with $\lambda \geq 0$ and $B+\lambda I$ positive semidefinite.

Proof: It has been noted that p must solve an equation of the form of (2.5). It remains to show that $B+\lambda I$ is positive semidefinite. Since p solves (2.1), it also solves $\min\{\psi(w): \|w\| = \|p\|\}$. It follows that $\psi(w) \geq \psi(p)$ for all w such that $\|w\| = \|p\|$. This inequality together with equation (2.5) gives

$$(2.6) \qquad f - p^T(B+\lambda I)w + \frac{1}{2} w^TBw \geq f - p^T(B+\lambda I)p + \frac{1}{2} p^TBp \ .$$

Rearranging terms in (2.6) gives

$$(2.7) \qquad \frac{1}{2} (w-p)^T(B+\lambda I)(w-p) \geq \frac{1}{2} (w^Tw-p^Tp) = 0$$

for all w such that $\|w\| = \|p\|$. Since $p \neq 0$, it follows readily from (2.7) that $B+\lambda I$ is positive semidefinite. □

Lemma (2.4) establishes necessary conditions concerning the pair λ,p when p solves (2.1). Our next result establishes sufficient conditions that will ensure p is a solution to (2.1). These results are essentially given in [11]. However, we wish to present a statement and proof of these results that is more complete and better suited to this presentation.

Lemma (2.8): Let $\lambda \in \mathbb{R}$, $p \in \mathbb{R}^n$ satisfy

$$(2.9) \qquad (B+\lambda I)p = -g \text{ with } B+\lambda I \text{ positive semidefinite.}$$

(i) If $\lambda = 0$ and $\|p\| \leq \Delta$ then p solves (2.1).

(ii) If $\|p\| = \Delta$ then p solves

$$\psi(p) = \min\{\psi(w): \|w\| = \Delta\} .$$

(iii) If $\lambda \geq 0$ and $\|p\| = \Delta$ then p solves (2.1).

If, in fact, $B+\lambda I$ is positive definite then p is unique in each of the cases (i), (ii), (iii).

Proof: If λ, p satisfy (2.9) then

$$(2.10) \qquad f + g^T w + \frac{1}{2} w^T(B+\lambda I)w \geq f + g^T p + \frac{1}{2} p^T(B+\lambda I)p$$

holds for any $w \in \mathbb{R}^n$. It follows that

$$(2.11) \qquad \psi(w) \geq \psi(p) + \frac{\lambda}{2}(p^T p - w^T w) .$$

Statements (i), (ii), (iii) are immediate consequences of (2.11). The uniqueness statement follows from (2.10) because the inequality is strict when $B+\lambda I$ is positive definite and $w \neq p$. □

The solution of problem (2.1) is closely related to solving the nonlinear equation.

$$(2.12) \qquad \phi(\alpha) = \Delta , \quad \text{where} \quad \phi(\alpha) \equiv \|(B+\alpha I)^{-1} g\| .$$

Using the eigensystem of the symmetric matrix B together with the invariance of $\|\cdot\|$ under orthogonal transformations it is easy to show if $g \neq 0$ that $\phi^2(\alpha)$ is a rational function with second order poles all belonging to a subset of the eigenvalues of $-B$. Since $\lim_{\alpha \to +\infty} \phi(\alpha) = 0$ it follows that (2.12) has a solution whenever $\Delta > 0$ and $g \neq 0$.

We can construct a solution to problem (2.1) using a particular solution of (2.12). Let λ_1 be the smallest eigenvalue of B; let $S_1 = \{q \in \mathbb{R}^n: Bq = \lambda_1 q\}$; let $\hat{\alpha}$ be the largest root of (2.12) when $g \neq 0$ and $\hat{\alpha} = 0$ when $g = 0$. If there is any $q \in S_1$ such that $g^T q \neq 0$ then $\hat{\alpha} > -\lambda_1$ must hold. If $g \in S_1^{\perp}$ then $-\lambda_1$ is not a pole of ϕ. Thus $\phi(-\lambda_1)$ is well defined when $g \in S_1^{\perp}$ and this is the only possibility for $\hat{\alpha} \leq -\lambda_1$ to occur. Put $\lambda = \max\{0, -\lambda_1, \hat{\alpha}\}$

and let

$$\theta^2 = \Delta^2 - \phi^2(\lambda) \quad \text{if} \quad \lambda = -\lambda_1 , \quad \theta = 0 \text{ otherwise.}$$

We now construct a solution p to problem (2.1) by the formula

$$(2.13) \qquad p = -(B+\lambda I)^{\dagger} g + \theta q$$

where $q \in S_1$, $\|q\| = 1$, and $(\dagger)$ denotes pseudo-inverse [25]. Note $B+\lambda I$ must be positive semidefinite with this choice of λ. Since $q^T(B+\lambda I)^{\dagger} = 0$ when $\lambda = -\lambda_1$, it is easily checked that p is a solution to (2.9) and satisfies either condition (i) or (iii) of Lemma (2.8). Thus p solves (2.1) and $\|p\| = \Delta$ whenever $\lambda_1 \leq 0$. The solution given by (2.13) shows that p is not unique whenever $g \in S_1^{\perp}$ and $\phi(-\lambda_1) < \Delta$ due to the arbitrary choice of sign in defining θ.

This discussion of the theoretical subtleties of solving (2.1) indicates numerical difficulties may arise when a solution to problem (2.1) is sought. The case $g \in S_1^{\perp}$, $\lambda = -\lambda_1$ in (2.13) will give rise to a very sensitive numerical problem. Any computational technique for solving (2.9) will introduce roundoff error. However, in this sensitive case small perturbations in the quantities B,g,λ can lead to large perturbations of the solution p due to the fact that $B+\lambda I$ will be nearly singular. Apparently the true nature of the difficulty here is the non-uniqueness of the solution p given by (2.14). We illustrate this point with a simple example. Let

$$g^T = (1,0) , \quad B = \begin{pmatrix} 1 & 0 \\ 0 & \eta \end{pmatrix} \quad \text{with} \quad \eta \leq 0 .$$

The solutions to (2.1) are of the form

$$p^T = -\left(\frac{1}{1-\eta}, \theta\right) \quad \text{with} \quad \frac{1}{(1-\eta)^2} + \theta^2 = \Delta^2$$

whenever $1/(1-\eta)^2 < \Delta^2$. The perturbation $g_\varepsilon^T = (1,\varepsilon)$ gives a solution

$$p_\varepsilon^T = -\left(\frac{1}{1+\lambda}, \frac{\varepsilon}{\eta+\lambda}\right) , \quad \text{with} \quad \frac{1}{(1+\lambda)^2} + \frac{\varepsilon^2}{(\eta+\lambda)^2} = \Delta^2 .$$

Clearly for any choice of sign for θ there is a perturbation ε such that

$\|p-p_\varepsilon\|/\|p\|$ is "large". In case $\eta < 0$ we must have $\|p\| = \Delta$ to solve (2.1) and we can be led to extremely different solutions as a result of error introduced by roundoff.

The convergence analysis to be given in Section 4 will depend heavily upon the following technical result concerning the amount of decrease in the local quadratic model. A geometric interpretation of the result is that, for a quadratic function, any solution p to (2.1) produces a decrease $f-\psi(p)$ that is at least as much as the decrease a search along the steepest descent direction $-g$ would provide.

Lemma (2.14): Let p be a solution to (2.1). Then

$$f-\psi(p) \geq \frac{1}{2}\|g\|\min\left(\|p\|, \frac{\|g\|}{\|B\|}\right) .$$

A proof of this result may be found in [23]. □

In fact, the inequality in Lemma (2.14) is obtained by Powell's "dog-leg" step [22]. This inequality is the main ingredient used to show the sequence of gradients tend to zero for the Modified Newton's method we are about to present. The reason for solving (2.1) rather than using the dog-leg step is that second order information is used to greater advantage. This will become evident as we present some very strong convergence results in Section 4.

A particular method for obtaining numerical solutions to (2.1) will be suggested in Section 5. For the moment we assume that a numerical solution p to problem (2.1) can be obtained which satisfies

(i) $(B+\lambda I)p = -g+\delta g$ with $B+\lambda I$ positive semidefinite ,

(ii) $$\|\delta g\| \leq \begin{cases} \varepsilon_1\|g\| & \text{if } \|g\| \neq 0 , \\ \varepsilon_1\Delta & \text{if } \|g\| = 0 , \end{cases}$$

and

(iii) $$\begin{cases} |\|p\|-\Delta| \leq \varepsilon_2\Delta & (\text{when } \lambda > 0) , \\ \|p\| \leq (1+\varepsilon_1)\Delta & (\text{when } \lambda = 0) , \end{cases}$$

for some fixed $0 < \varepsilon_1 < \varepsilon_2$ in (0,1) that are consistent with the finite precision arithmetic. The results of Lemma (2.8) imply that such a p solves the modified problem.

$$(2.15) \qquad \min\{f + \tilde{g}^T w + \frac{1}{2} w^T B w : \|w\| \leq \tilde{\Delta}\}$$

where $(1-\varepsilon_2)\Delta \leq \tilde{\Delta} \leq (1+\varepsilon_2)\Delta$ and $\tilde{g} = g+\delta g$ with $\|\delta g\| \leq \varepsilon_1 \|g\|$ and $\|p\| = \tilde{\Delta}$ when $g \neq 0$. In our analysis we shall assume $\varepsilon_1 = \varepsilon_2 = 0$. A trivial but tedious modification of the analysis would apply to a computed step p which satisfies the above criteria. This is primarily because the crutial inequality of Lemma (2.14) will become

$$(2.16) \qquad f-\psi(p) \geq \frac{1}{2} \|\tilde{g}\| \min\left(\tilde{\Delta}, \frac{\|\tilde{g}\|}{\|B\|}\right)$$

$$\geq \frac{1}{2} (1-\varepsilon_1)\|g\| \min\left\{(1-\varepsilon_2)\Delta, \frac{(1-\varepsilon_1)\|g\|}{\|B\|}\right\} .$$

It is straightforward to see that the inequality of (2.16) is sufficient for purposes of the ensuing analysis, but we wish to refrain from including such complicated expressions at each stage of the analysis.

3. A Modified Newton Iteration

A well known method for solving the unconstrained minimization problem is Newton's method applied to finding a zero of the gradient of the objective function. However, this iteration is clearly not suitable as a general algorithm without modification. The basic iteration is

$$(3.1) \qquad x_{k+1} = x_k - G_k^{-1} \nabla f(x_k) , \qquad k=0,1,2,...$$

where an initial iterate x_0 must be specified, $\nabla f(x_k)$ is the gradient of f, $G_k = \nabla^2 f(x_k)$ is the $n \times n$ (symmetric) Hessian matrix of mixed second partial derivatives of f. The algorithm we shall discuss will require that f is twice differentiable at any point x in the domain of f, and that these derivatives can be evaluated explicitly.

There are three fundamental reasons why this basic method must be modified. First, the initial iterate may have to be very "close" to a local minimizer in order to be assured that the iteration will converge. Second, even if the iteration converges to a stationary value x^* ($\nabla f(x^*)=0$) there is no guarantee that x^* will be a local minimizer. Third, the iterate x_{k+1} may not be well defined by (3.1) if the Hessian G_k is singular or it may not be a sensible move if G_k is indefinite. Our purpose here is to discuss certain theoretical properties of a modification of the basic iteration (3.1). Our approach is not a new one, however we feel that the theoretical and numerical properties of the proposed method should be fully treated and that is the main goal of this discussion. The method we shall consider is called the model trust region method. We have already mentioned the history of this approach. The main concern here is the implementation of this type of algorithm. Therefore, this discussion is intended to apply to several possible implementations. Specific implementations are presented in Section 5.

Before the iteration is defined let us set out some of the properties desired of a modified-Newton iteration:

(3.2) (a) For a sufficiently general class of functions the iteration should be well defined and convergent given any initial iterate x_0.

(b) When the iteration converges to a point x^*, this point should satisfy as many necessary conditions for a minimizer as possible.

(c) The modification should not detract from the local quadratic rate of convergence enjoyed by Newton's method.

(d) The method should be invariant under linear affine scalings of the variables. That is, if we replace $f(x)$ by $\hat{f}(w) = f(Jw+z)$ where $J \in \mathbb{R}^{n\times n}$ is nonsingular and $w,z \in \mathbb{R}^n$, then applying the iteration to $\hat{f}$ with initial guess w_0 satisfying $x_0 = Jw_0+z$ should produce a sequence $\{w_k\}$ related to the sequence $\{x_k\}$ by $x_k = Jw_k+z$, where $\{x_k\}$ is produced by applying the algorithm to f with initial guess x_0.

The algorithm we are about to define will be shown to meet criteria a,b,c for all practical purposes. The last criterion (d) will be discussed in Section 6. To begin we introduce a factorization of the Hessian matrix. For each k we let

$$B_k = J_k^T G_k J_k$$

be a factorization of G_k with $B_k = B_k^T \in \mathbb{R}^{n\times n}$ and J_k nonsingular. It follows that B_k has the same inertia (see [17, p. 377]) as G_k. For each k we put $\phi_k(w) = f(x_k+J_k w)$. This function $\phi_k(w)$ may be regarded as a locally scaled objective function. The first three terms of the Taylor series of ϕ_k about w=0 will define a local quadratic model

$$\phi_k(w) = f_k + g_k^T w + \frac{1}{2} w^T B_k w$$

where $g_k^T = \nabla f(x_k)^T J_k$ and $f_k = f(x_k) = \phi_k(0)$. Along with the local quadratic model we shall maintain a control parameter $\Delta_k > 0$ which defines a local region of trust $\{w: \|w\| \leq \Delta_k\}$ where the model is considered valid. This parameter Δ_k will be revised during the iteration according to specific rules which are designed to force convergence of the iterates $\{x_k\}$.

We are potentially considering any symmetric factorization of the matrix G_k, but certain requirements should be kept in mind. For example, $g_k^T = \nabla f(x_k)^T J_k$ should be easily computed either explicitly or by solving $\nabla f(x_k)^T = g_k^T J_k^{-1}$. Also, it will be an advantage if the eigensystem of B_k is relatively inexpensive to compute or if the smallest eigenvalue and corresponding eigenvector(s) are easy to obtain. The reason for this is that the solution to problem (2.1) will play an important role in this iteration and as we have seen the eigensystem information may be required. This is especially true at points x_k where G_k is indefinite or singular.

Now we are ready to define the iteration.

Algorithm (3.3):

1) Let $k=1$ and let $0 < \eta_1 < \eta_2 < 1$, $0 < \gamma_1 < 1 < \gamma_2$ be prespecified constants;

2) Let $x_1 \in \mathbb{R}^n$, $\Delta_1 > 0$ be given;

3) If "convergence" then STOP;

4) Evaluate $f_k := f(x_k)$; $G_k := \nabla^2(fx_k)$;
Factor $B_k := J_k^T G_k J_k$; Evaluate $g_k := J_k^T \nabla f(x_k)$;

5) Compute $w_k \in \operatorname{argmin}\{\psi_k(w): \|w\| \leq \Delta_k\}$
Comment: $\psi_k(w) = f_k + g_k^T w + \frac{1}{2} w^T B_k w$;

6) Put ared $:= \phi_k(0)-\phi_k(w_k)$; pred $:= \phi_k(0)-\psi_k(w_k)$;
Comment: $\phi_k(w) = f(x_k+J_k w)$;

7) If $\frac{\text{ared}}{\text{pred}} < \eta_1$ then begin $\Delta_k := \gamma_1 \Delta_k$; goto 5; end;

8) If $\eta_1 \leq \frac{\text{ared}}{\text{pred}}$ then

   1) $x_{k+1} := x_k + J_k w_k$;

   2) if $\frac{\text{ared}}{\text{pred}} > \eta_2$ then $\Delta_k := \gamma_2 \Delta_k$;

   3) $\Delta_{k+1} := \Delta_k$; $k := k+1$;

9) GO to 2;

end.

There are ways to update the value of Δ at step 7 and step 8.2 which make better use of the information available at the current iterate x_k. For example, the cubic polynomial that fits $\Phi(\alpha) = \phi_k(\alpha w_k)$ by interpolating $\Phi(0)$,

$\phi'(0)$, $\phi''(0)$ and $\phi(1)$ will have a minimum $\hat{\alpha}$ in $(0,1)$ when the test at step 7 is passed. The region is contracted by setting $\gamma_1 = \hat{\alpha}$ if $\hat{\alpha}$ is not "too close" to 0 or 1. Details of this type of idea appear in [4,8,12,18]. Similar ideas may be applied at (8.2) to obtain an expansion factor $\gamma_2 \geq 1$ that depends upon available information. Other variations involving step 7 include accepting the predicted minimizer if $0 < \eta_0 \leq \frac{ared}{pred} \leq \eta_1$ but reducing the trust region. The analysis we shall perform on Algorithm (3.3) can be adapted to cover these possibilities in a fairly straightforward way. However, the gain in generality will result in a substantial loss in clarity of exposition in the analysis so we shall analyze the simple choices set forth in Algorithm (3.3).

Finally it should be pointed out that this iteration is well defined because step 7 will produce a sufficiently small Δ_k to obtain $\frac{ared}{pred} > \eta_1$ after a finite number of steps since the quadratic function $\psi_k(w)$ is defined by the first three terms of the Taylor series for $\phi_k(w)$. Our statement of the strategy is slightly different than the usual description in that x_{k+1} is always different from x_k. By doing this we avoid having to distinguish between "successful" and "unsuccessful" iterates in the analysis. With this exception the statement of the algorithm and the ensuing analysis are in the spirit of the paper presented by Powell [23]. Numerical schemes for producing the constrained quadratic minimization at step 5 will be presented in Section 5.

4. Convergence of the Modified Newton Iteration

In this section we shall establish that some very strong convergence properties are possessed by Algorithm (3.3). The first result is a slight modification of Powell's result in [23]. Our proof is much simpler due to the fact that here second order information is explicitly available.

Theorem (4.1): Let $f: \mathbb{R}^n \to \mathbb{R}$ be bounded below and let $G(x) = \nabla^2 f(x)$ be continuous and satisfy $\|G(x)\| \leq \beta$ for all $x \in \mathcal{L}(x_0)$. Let $\{x_k\} \subset \mathbb{R}^n$ be the sequence produced by Algorithm (3.3) applied to f given starting value x_0. Assume $\|J_k\|, \|J_k^{-1}\| \leq \sigma$, $k = 0,1,2,\ldots$ for some $\sigma \geq 1$. Then there is no constant $\varepsilon > 0$ such that $\|\nabla f(x_k)\| \geq \varepsilon$ for all k.

Proof: Assume there is an $\varepsilon > 0$ such that $\|\nabla f(x_k)\| \geq \varepsilon$ for all k. Since $g_k = J_k^T \nabla f(x_k)$ we have

$$\|g_k\| \geq \|\nabla f(x_k)\| / \|J_k^{-1}\| \geq \varepsilon/\sigma \equiv \gamma > 0 .$$

From step (7) of Algorithm (3.3) and from Lemma (2.14) we have

$$(4.2) \qquad f_k - f_{k+1} \geq \eta_1 (f_k - \psi_k(w_k)) \geq \frac{\eta_1}{2} \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$$

where $\|B_k\| = \|J_k^T G_k J_k\| \leq \beta\sigma^2$. Since f is bounded below and $f_{k+1} < f_k$ $k=0,1,2,\ldots$ we have $f_k - f_{k+1} \to 0$. Since $\|g_k\|/\|B_k\| \geq \gamma/\beta\sigma^2$ it follows that $\Delta_k \to 0$ from (4.2). However,

$$(4.3) \qquad \Delta_k \geq \|g_k\| / (\beta\sigma^2 + \lambda^{(k)})$$

is obtained from the inequality

$$\|g_k\| = \|(B_k + \lambda^{(k)} I) w_k\| \leq \|w_k\| (\|B_k\| + \lambda^{(k)}) ,$$

where $\lambda^{(k)}$ is the multiplier associated with the solution to step 5 of Algorithm (3.3). Thus Inequality (4.3) shows $\lambda^{(k)} \to +\infty$. Now, from Taylor's theorem and Lemma (2.8) it readily follows (for k sufficiently large) that

$$(4.4) \qquad \left|\frac{\text{ared}(k)}{\text{pred}(k)} - 1\right| = 2 \frac{\left| w_k^T \int_0^1 (B_k(\theta) - B_k)(1-\theta) d\theta w_k \right|}{w_k^T (B_k + \lambda^{(k)} I) w_k + \lambda^{(k)} w_k^T w_k}$$

$$\leq \frac{1}{\lambda^{(k)}} \sup_{0 \leq \theta \leq 1} \|B_k(\theta) - B_k\| ,$$

where $B_k(\theta) = J_k^T [G(x_k + \theta s_k)] J_k$ with $s_k = x_{k+1} - x_k$. Since $\lambda^{(k)} \to +\infty$ we obtain $\frac{\text{ared}(k)}{\text{pred}(k)} \to 1$ and thus the test at step (7) of Algorithm (3.3) is passed the first time through for all k sufficiently large. This implies the existence of a $K > 0$ such that $\Delta_k \geq \Delta_K$ for all $k \geq K$. Therefore, the assumption $\|\nabla f(x_k)\| \geq \varepsilon$ for all k has led to a contradiction.

□

We remark that the continuity of G(x) is only used to obtain the numerator on the right hand side of (4.4) and that the Theorem can also be established without this assumption. See Powell [23] for example.

This result has shown that at least one subsequence of $\{x_k\}$ converges to a critical point of f. The next result which is due to Thomas [26] will establish the much stronger fact that every accumulation point of the sequence $\{x_k\}$ is a critical point of f.

Theorem (4.5): Let the hypotheses of Theorem (4.1) hold. Then $\lim_{k\to\infty}\|\nabla f(x_k)\| = 0$.

Proof: Suppose there is a subsequence $\{x_{kj}\} \subset \{x_k\}$ such that $\|\nabla f(x_{kj})\| \geq \varepsilon > 0$, for all j=1,2,... . As in Theorem (4.1) this implies $\|g_k\| \geq \gamma > 0$. Moreover, due to Theorem (4.1), we may select an integer ℓj corresponding to each j such that

$$\ell j = \min\{\ell \geq kj: \|g_\ell\| \geq \gamma/2\sigma^2\} \tag{4.6}$$

and without loss of generality $kj \leq \ell j < k_{j+1}$, j=1,2,... . From inequality (4.2) we obtain that

$$f_\ell - f_{\ell+1} \geq \frac{\eta_1\gamma}{4\sigma^2}\min\left(\Delta_\ell, \frac{\gamma}{2\beta\sigma^4}\right) \tag{4.7}$$

will hold for all $kj \leq \ell \leq \ell j$, j=1,2,... . From (4.7) it follows that

$$\begin{aligned} f_{kj} - f_{\ell j+1} &= \sum_{\ell=kj}^{\ell j} f_\ell - f_{\ell+1} \\ &\geq \frac{\eta_1\gamma}{4\sigma^2}\min\left(\sum_{\ell=kj}^{\ell j}\Delta_\ell, \frac{\gamma}{2\beta\sigma^4}\right) . \end{aligned} \tag{4.8}$$

From inequality (4.8) it follows that

$$\|x_{kj} - x_{\ell j+1}\| \to 0 \quad \text{as} \quad j \to \infty$$

because

$$\|x_{kj} - x_{\ell j+1}\| \leq \sum_{\ell=kj}^{\ell j}\|s_\ell\| \leq \sigma \sum_{\ell=kj}^{\ell j}\Delta_\ell , \tag{4.9}$$

and the right hand side of (4.9) is forced to zero due to (4.8). The uniform bound on $G(\cdot)$ implies the uniform continuity of $\nabla f(x)$ on $\mathcal{L}(x_0)$ and it follows that

$$\|\nabla f(x_{k_j})-\nabla f(x_{\ell_j+1})\| < \frac{\gamma}{4\sigma}$$

for all j sufficiently large. Therefore

$$\begin{aligned}\|g_{k_j}\| &\leq \sigma\|\nabla f(x_{k_j})\| \\ &\leq \sigma\left(\|\nabla f(x_{k_j})-\nabla f(x_{\ell_j+1})\| + \|\nabla f(x_{\ell_j+1})\|\right) \\ &\leq \sigma\left(\frac{\gamma}{4\sigma} + \sigma\|g_{\ell_j+1}\|\right) \\ &\leq \sigma\left(\frac{\gamma}{4\sigma} + \frac{\gamma}{2\sigma}\right) \leq \frac{3\gamma}{4} < \gamma ,\end{aligned}$$

for all j sifficiently large. The assumption that $\|\nabla f(x_{k_j})\| \geq \varepsilon > 0$ has led to a contradiction and we must conclude that $\lim_{k\to\infty}\|\nabla f(x_k)\| = 0$. □

This result has established that every limit point of the sequence $\{x_k\}$ satisfies the first order necessary conditions for a minimum. Now we shall establish results which give added justification to the use of second order information when it is available. Several authors [10,16,19,20] have proposed modified Newton methods which guarantee convergence to a critical point x* with the additional feature that the Hessian G(x*) be positive semi-definite. Thus second order necessary conditions for a minimum are satisfied by x*. The following series of results show that Algorithm (3.3) shares this property.

Lemma (4.10): Let the hypotheses of Theorem (4.1) be satisfied. If G(x) is uniformly continuous on (x_{k_0}) then there is no positive number $\lambda > 0$ such that $\lambda^{(k)} \geq \lambda$ for $k \geq k_0$.

Proof: If $\lambda^{(k)} > 0$ then $\|w_k\| = \Delta_k$ due to Lemma 2.8. We conclude from inequality (4.4) that

$$\left|\frac{\text{ared}(k)}{\text{pred}(k)} - 1\right| \quad \leq \frac{1}{\lambda} \sup_{0\leq\theta\leq 1} \|B_k(\theta)-B_k\|$$

where $B_k(\theta) = J_k^T(G_k(x_k+\theta s_k))J_k$. Since $\text{pred}(k) \geq \lambda^{(k)}\Delta_k^2 \geq \lambda\Delta_k^2$ it follows that $\Delta_k \not\to 0$ because $\text{pred}(k) \to 0$. Now the boundedness of $\|J_k\|$, $\|J_k^{-1}\|$ together with the uniform continuity of $G(x)$ on (x_{k_0}) gives

$$\frac{\text{ared}(k)}{\text{pred}(k)} \to 1 , \quad \text{as} \quad k \to +\infty .$$

We must conclude as in the proof of Theorem (4.1) that $\Delta_k \geq \Delta_K$ for some $K > 0$. This contradiction establishes the result. □

Since $-\lambda^{(k)} \leq \lambda_1^{(k)}$ which is the smallest eigenvalue of B_k, the next theorem follows easily from the boundedness of $\|J_k\|, \|J_k^{-1}\|$ together with Lemma (4.10).

Theorem (4.11): Let the hypotheses of Lemma (4.10) be satisfied. If the sequence $\{x_k\}$ is convergent to a limit x^* say, then $\nabla f(x^*) = 0$ and $G(x^*)$ is positive semidefinite. □

At this point we should remark that failure of this iteration to converge will require an extremely pathological situation. A moments reflection will convince the reader that every limit point of the sequence $\{x_k\}$ must be a critical point of f, and f must have the same value at each of these critical points. Moreover, at least one of these critical points has a positive semidefinite Hessian. The next result shows that if any one of the limit points of the sequence, x^* say, satisfies $G(x^*)$ is positive definite then the entire sequence must converge to x^*.

Lemma (4.12): Let f satisfy the hypothesis of Theorem (4.1). Let $\{x_{k_j}\} \subset \{x_k\}$ be a subsequence which converges to a critical point x^*. If $G(x^*)$ is positive definite and $G(x)$ is continuous in a neighborhood of x^* then the entire sequence must converge to x^*.

Proof: Due to the continuity of G we must have $\|G(x)^{-1}\| \leq \hat{\beta}$ for all x in some neighborhood $\mathcal{N}$ of x^*. Thus for any $\varepsilon > 0$ there is a $\delta > 0$ and a corresponding ball $\mathcal{B}_\delta = \{x: \|x-x^*\| < \delta\} \subset \mathcal{N}$ such that

$$\|G(x)^{-1}\nabla f(x)\| \leq \varepsilon \quad \text{for} \quad x \in \mathcal{B}_\delta .$$

This follows from the continuity of ∇f and G, as well as the fact that $\nabla f(x^*) = 0$. The assumption that $G(x^*)$ is positive definite implies that x^* is an isolated local minimum. Therefore, there is a $\delta_1 > 0$ such that $f(x^*) < f(x)$, $G(x)$ is positive definite, $\nabla f(x) \neq 0$ for every $x \in \mathcal{B}_1 \equiv \{z: \|z-x^*\| < \delta_1\}$ that is different from x^*. Let $0 < \delta_2 < \frac{1}{4}\delta_1$ be chosen so that $\|G(x)^{-1}\nabla f(x)\| \leq \frac{\delta_1}{2\sigma^2}$ for all $x \in \mathcal{B}_2 \equiv \{z: \|z-x^*\| < \delta_2\}$. We shall show that for all k sufficiently large, the iterates x_k lie in $\mathcal{B}_2$.

Since $x_{kj} \to x^*$ and $f(x^*) < f(x)$ for all $x \in \mathcal{B}_1$, $x \neq x^*$ there is a j_0 such that

$$f(x_{kj_0}) < \inf\{f(x): \delta_2 \leq \|x-x^*\| \leq \delta_1\}, \tag{4.13}$$

and $x_{kj} \in \mathcal{B}_2$ for all $j \geq j_0$. Let $j \geq j_0$ be fixed and suppose that

$$x_\ell \in \mathcal{B}_2 \quad \text{but} \quad x_{\ell+1} \notin \mathcal{B}_2 \quad \text{for} \quad kj \leq \ell < k_{j+1}.$$

Since $\ell+1 > kj_0$ we have $f(x_{\ell+1}) \leq f(xj_0)$ and thus $x_{\ell+1} \notin \mathcal{B}_1$ due to (4.13). It follows that

$$\|x_{\ell+1}-x_\ell\| \geq \|x_{\ell+1}-x^*\| - \|x_\ell-x^*\| \geq \delta_1-\delta_2 \geq \frac{3}{4}\delta_1.$$

This implies that $\Delta_\ell \geq \|w_\ell\| \geq \frac{1}{\sigma}\|x_{\ell+1}-x_\ell\| \geq \frac{3\delta_1}{4\sigma}$. However, this is a contradiction since $x_\ell \in \mathcal{B}_2$ implies

$$\|G(x_\ell)^{-1}\nabla f(x_\ell)\| \leq \frac{\delta_1}{2\sigma^2}$$

so the Newton iterate $z = x_\ell - G(x_\ell)^{-1}\nabla f(x_\ell)$ satisifes

$$\|J_\ell^{-1}(z-x_\ell)\| \leq \sigma\|z-x_\ell\| \leq \frac{\delta_1}{2\sigma} < \Delta_\ell.$$

This argument shows that $x_k \in \mathcal{B}_2$ for all $k \geq kj_0$. Since x^* is the only critical point of f in $\mathcal{B}_2$ and since every limit point of the sequence $\{x_k\}$ is a critical point of f we must conclude that the entire sequence converges to x^*.

□

It would be more desirable to obtain a result that would ensure convergence of the sequence $\{x_k\}$ without assuming a subsequence converges to a strong local minimum. However, just extending this argument to the case of an isolated local minimum with singular Hessian would be difficult since one can no longer rely on the Newton step. Our final result will show, in conjunction with Lemma (4.12), that if there is a subsequence which converges to a strong local minimum then the entire sequence converges and ultimately the rate of convergence is quadratic.

Theorem (4.14); Let the hypotheses of Theorem (4.1) be satisfied. Suppose further that $x_k \to x^*$ with $G(x^*)$ positive definite and

$$\|G(x)-G(x^*)\| \leq L\|x-x^*\| \tag{4.15}$$

for all x in some neighborhood of x^*. Then there is a constant $\mathcal{K} > 0$ such that

$$\|x_{k+1}-x^*\| \leq \mathcal{K}\|x_k-x^*\|^2$$

for all k sufficiently large.

Proof: Since $x_k \to x^*$ and $G(x^*)$ is positive definite it follows from continuity that there are positive constants $\beta_1 \leq \|G(x_k)^{-1}\| \leq \beta_2$ for k sufficiently large. Thus $|\text{pred}(k)| \geq s_k^T G(x_k) s_k \geq \|s_k\|^2/\|G(x_k)^{-1}\| \geq \frac{1}{\beta_2}\|s_k\|^2$, and

$$|\text{ared}(k)-\text{pred}(k)| \leq \|s_k\|^2 \int_0^1 \|G_k(\theta)-G_k\|(1-\theta)d\theta ,$$

where $G_k(\theta) = G(x_k+\theta s_k)$ with $s_k = x_{k+1}-x_k$. It follows easily from (4.15) that

$$\left|\frac{\text{ared}(k)}{\text{pred}(k)} - 1\right| \to 0 \quad \text{as} \quad k \to \infty$$

and we must conclude that there is some $K > 0$ such that $\Delta_k \geq \Delta_K$ for all $k \geq K$.

Again, since $x_k \to x^*$ with $\nabla f(x^*) = 0$ it follows that $\|G(x_k)^{-1}\nabla f(x_k)\| < \sigma\Delta_K$ so the Newton step is accepted for all k sufficiently large. Hence the tail

of the sequence $\{x_k\}$ is the unmodified Newton iteration which is quadratically convergent to x* since G(x*) is positive definite [21, p. 421]. □

While these results hold little computational meaning in the presence of roundoff error, it is satisfying to have established such strong results about the iteration. This is especially true since the method has such an intuitive appeal. Our aim in this section has been to establish these theoretical results in a framework that is general enough to encompass many possible implementations. We shall consider some of these implementations in the next section.

5. Implementation

Numerical performance of the algorithm described in Section 3 and analyzed in Section 4 is obviously going to depend upon a careful implementation of the locally constrained minimization of the quadratic model. In Section 2 we pointed out several theoretical facts that indicate great care should be exercised in this computation. In this section we shall put forth several possible implementations. Each of these will have certain advantages and disadvantages depending upon the nature of the optimization problem at hand. The convergence theory provided in Section 4 was purposely made sufficiently general to apply to all of the alternative implementations to be presented here.

Our main concern is to provide an efficient and stable method for the solution of problem (2.1). To this end we consider factorizations

$$J^TGJ = B$$

of the symmetric n×n matrix G. We are assuming that $\|J\|, \|J^{-1}\| \leq \sigma$ where $\sigma > 1$ is some fixed number that is independent of G. Recall that the matrix B is also symmetric and must have the same inertia as G. Some specific examples are: (a) J orthogonal and B diagonal; (b) J orthogonal and B tridiagonal; (c) $J^T = L^{-1}P$ where L is unit lower triangular, P is a permutation matrix, and B is either tridiagonal [1] or block diagonal with 1×1 or 2×2 diagonal blocks

[2]. We shall also consider the case when J is just a diagonal nonsingular matrix.

If the eigensystem of B is easily obtained (i.e. in case a or case c when B is block diagonal) then we are able to solve problem (2.1) directly by solving the nonlinear equation (2.12) for the largest root and then constructing a solution to (2.1) using formula (2.13). This method of solution has the particular advantage that the case when $g \in S_1^{\perp}$ is explicitly revealed.

A disadvantage of using factorization (a) is that it is relatively expensive to compute. One of the reasons for introducing generality into the model trust region calculation was to allow use of the Bunch-Parlett factorization [2]. This factorization is very efficient due to the fact that symmetry is exploited. The matrix B for this factorization has an eigensystem that is easily computed. Moreover, the matrices J satisfy the criteria $\|J\|, \|J^{-1}\| \leq \sigma$ so in theory all of the results of Section 4 apply. There may be some cause for concern regarding the effect of the transformation J on the descent direction, because the triangular coordinate system may be very skewed even though the matrix J is well conditioned.

Nevertheless, our main concern with either of these factorizations is the efficient and reliable solution to an equation of the form

$$(5.1) \qquad \left[\sum_{j=1}^{n} \frac{\gamma_j^2}{(\alpha+\lambda_j)^2}\right]^{1/2} = \Delta \ .$$

for the largest root λ. The left hand side of (5.1) is precisely the form of $\phi(\alpha) = \|(B+\alpha I)^{-1} g\|$ in (2.12) regardless of whether or not B is diagonal. Several authors [12,18,24] discuss the solution of equations that closely resemble (5.1). The key observation is that Newton's method which is based on a local linear approxir tion to $\phi(\alpha)$ is not likely to be the best method for solving (5.1) because the rational structure of $\phi^2(\alpha)$ is ignored. Instead, an iteration for solving (5.1) can be derived based upon a local rational approximation to ϕ. The iteration is obtained by requiring $\hat{\phi}(\alpha) = \frac{\gamma}{\alpha+\beta}$ to satisfy

$$\hat{\phi}(\alpha) = \phi(\alpha) \ , \quad \hat{\phi}'(\alpha) = \phi'(\alpha)$$

where we regard α as the current approximation to the root λ. This approximation is then improved by solving for an $\hat{\alpha}$ that satisfies $\hat{\phi}(\hat{\alpha}) = \Delta$. The resulting iteration is

$$\alpha_{k+1} = \alpha_k + \frac{\phi(\alpha_k)}{\phi'(\alpha_k)} \left[\frac{\Delta - \phi(\alpha_k)}{\Delta} \right] . \tag{5.2}$$

If the form of $\phi(\alpha)$ is known explicitly then it is straightforward to safeguard (5.1). The local rate of convergence of this iteration is quadratic but the most important feature of (5.1) is that usually the number of iterations required to produce an acceptable approximation to λ is very small because the iteration is based upon the rational structure of ϕ^2.

Iteration (5.2) can be implemented without explicit knowledge of the eigensystem of B. This important observation which is due to Hebden [12] makes it possible to implement (5.2) merely by solving linear systems with $B+\alpha I$ as the coefficient matrix. This is easy to see since $\phi(\alpha) = \|p_\alpha\|$, and $\phi'(\alpha) = -\frac{1}{\phi(\alpha)} p_\alpha^T (B+\alpha I)^{-1} p_\alpha$ where $(B+\alpha I)p = -g$. Hebden [12] suggests a way to obtain $\alpha > -\lambda_1$ during the process of attempting to compute the Cholesky factorizatin of $B+\alpha I$. This is discussed in more detail by Gay in [7] where the difficult case $g \in S_1^{\perp}$ is addressed. Within this context we could allow J to be taken as a nonsingular diagonal matrix for solving purposes. Moré has used this idea in his adaptation of Hebden's work to the nonlinear least squares problem [18]. The result of Moré's work is a very elegant robust algorithm for nonlinear least squares. In [18] careful attention is paid to safeguarding the step calculation. The safeguarding task is somewhat more difficult in the present setting due to the fact that B may have negative eigenvalues. The essential difficulty seems to stem from the fact that without explicit knowledge of the eigensystem it is difficult to detect the case $g \in S_1^{\perp}$. Moreover, it seems to be necessary to have an estimate of the smallest eigenvalue and a corresponding eigenvector in order to obtain a solution to (2.1) in case $g \in S_1^{\perp}$ (see formula 2.13). This was recognized by Hebden but he did not provide a suitable solution. Gay [7] suggests obtaining an eigenvector using inverse iteration if the case $g \in S_1^{\perp}$ is detected because a factorization of the (nearly) singular matrix $B+\lambda I$ will be available.

Here we suggest an alternative to the methods which have been proposed previously. In the following we are considering J to be a diagonal

nonsingular matrix. Let us return to the derivation of iteration (5.2). Another way to obtain this iteration is to apply Newton's method to the problem

$$\frac{1}{\Delta} - \frac{1}{\phi(\alpha)} = 0 \ . \tag{5.3}$$

From this observation we can see that iteration (5.2) is closely related to Newton's method applied to the problem

$$\begin{bmatrix} r(p,\alpha) \\ \frac{1}{\Delta} - \frac{1}{\phi(\alpha)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ , \tag{5.4}$$

where we use the notation $r(p,\alpha) = B_\alpha p+g$ with $B_\alpha = B+\alpha I$. There is a serious disadvantage to this iteration when $g \in S_1^\perp$ or nearly so. This is because the Jacobian of (5.4) is

$$\begin{bmatrix} B_\alpha & p \\ 0 & \frac{\phi'(\alpha)}{\phi^2(\alpha)} \end{bmatrix} \ , \tag{5.5}$$

and this matrix is singular at a solution λ,p_λ of (2.1) in the sensitive case $g \perp S_1^\perp$, $\|B_\lambda^\perp g\| < \Delta$, where $\lambda = -\lambda_1$.

Of course, this situation impairs the local rate of convergence. Moreover, as the iteration converges to such a solution the method requires solving linear systems which have increasingly ill-conditioned coefficient matrices.

As an alternative we suggest removing the explicit dependence of $\phi(\alpha)$ on the variable α in (5.4). Instead of (5.4) we shall apply Newton's method to solve

$$\begin{bmatrix} r(p,\alpha) \\ \frac{1}{\Delta} - \frac{1}{\|p\|} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ . \tag{5.6}$$

Due to Lemma (2.8) a solution $\alpha = \lambda$, $p = p_\lambda$ to (5.6) provides a solution to problem (2.1) whenever B_λ is positive (semi) definite and $\lambda \geq 0$. The Jacobian of (5.6) is

$$(5.7) \qquad \begin{bmatrix} B_\alpha & p \\ \frac{1}{\|p\|^3} p^T & 0 \end{bmatrix} ,$$

and this matrix is nonsingular at a solution to (2.1) in the cases that are most likely to occur. This is important since it follows that Newton's method applied to (5.6) will usually enjoy a quadratic rate of convergence. A precise statement of when (5.7) is nonsingular at a solution is given in the following lemma.

Lemma (5.7): Let $\alpha = \lambda \geq 0$, $p = p_\lambda \neq 0$ be a solution to problem (5.6) with B_λ positive semidefinite. If B_λ is positive definite or if $\|B_\lambda^+ g\| = \Delta$ and $\dim(S_1) = 1$ then the Jacobian matrix (5.7) is nonsingular.

Proof: Let $p = p_\lambda$, $\alpha = \lambda$. It is sufficient to show

$$(5.8) \qquad \begin{pmatrix} B_\lambda & p \\ p^T & 0 \end{pmatrix}$$

is nonsingular. Suppose that

$$(5.9) \qquad \begin{pmatrix} B_\lambda & p \\ p^T & 0 \end{pmatrix} \begin{pmatrix} z \\ \zeta \end{pmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} .$$

Then

$$(5.9) \qquad \text{(i)} \quad B_\lambda z + p\zeta = 0 \quad \text{and} \quad \text{(ii)} \quad p^T z = 0 .$$

These two equations imply $z^T B_\lambda z = 0$. Therefore either $z = 0$ or B_λ is singular and $z \in S_1$. Both of these possibilities imply $\zeta = 0$ since $p \neq 0$, and $B_\lambda z = 0$. Thus, when B_λ is positive definite the only solution to (5.9) is

$z = 0$, $\zeta = 0$ so (5.8) is nonsingular. If on the other hand B_λ is singular and $\|B_\lambda^\dagger g\| < \Delta$ then $p = -B_\lambda^\dagger g + \theta q$ with $q \in S_1$, $\|q\| = 1$, and $\theta \neq 0$. Since $\dim(S_1) = 1$ it follows that $z = \gamma q$ and thus $0 = z^T p = \theta\gamma$ which shows $\gamma = 0$. Again we conclude (5.9) only has the trivial solution so (5.8) is nonsingular. □

The basic iteration (without safeguards) for solving (5.6) will be given now. The details of various suggested implementations will follow.

Algorithm (5.10):

1) Obtain an initial guess p_0 and α_0 such that $B_0 = B\alpha_0$ is positive definite;

2) for $k = 0,1,2,\ldots$

1) $r_k = B_k p_k + g$; $\rho_k = \frac{p_k^T p_k}{\Delta}(\|p_k\| - \Delta)$;

2) Solve

$$\begin{bmatrix} B_k & p_k \\ p_k^T & 0 \end{bmatrix} \begin{bmatrix} \delta p \\ \delta\alpha \end{bmatrix} = -\begin{bmatrix} r_k \\ \rho_k \end{bmatrix};$$

3) $p_{k+1} = p_k + \delta p$; $B_{k+1} = B_k + \delta\alpha I$;

We must address several computational questions concerning this iteration. These include what initial guess should be used, how to solve the linear systems at step 2.2, how to safeguard the basic iteration, and finally how to stop the iteration.

First of all we shall discuss some methods for solving the linear system at step 2.2. For matrices B that are of moderate size and those which have no particular structure we recommend the following. Compute an orthogonal matrix Q through a product of Householder transformations such that

$$(5.11) \qquad \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} B_0 & p_0 \\ p_0^T & 0 \end{pmatrix} \begin{pmatrix} Q^T & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} T & \tau e_n \\ \tau e_n^T & 0 \end{pmatrix}$$

where T is tridiagonal and $e_n^T = (0,\ldots,0,1)$. Initially this factorization is more expensive than some alternatives (such as the Bunch-Kaufman [3] factorization). However, it presents several advantages as we shall see. First of all, since $T = QBQ^T$ has the same eigenvalues as B we can easily compute a Sturm-sequence for T to tell very reliably whether or not T is positive definite [27]. Moreover, since good upper and lower bounds for the smallest eigenvalues of T are available, a good safeguarding scheme can be obtained. After applying transformation (5.11) to the linear system at step 2.2 of (5.10) a solution can be obtained using ordinary Gaussian-elimination with partial pivoting. It is preferable to ignore symmetry in this case for the same reason it is preferable in the case of inverse iteration for the computation of an eigenvector. See Wilkinson [27] for more detail. A more important observation to make here is that iteration (5.10) is invariant under transformation (5.11). Once the correction $\delta p = Q^T\hat{\delta p}$ is obtained we have the updated matrix

$$(5.12)\qquad \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} B_{k+1} & p_{k+1} \\ p_{k+1}^T & 0 \end{pmatrix}\begin{pmatrix} Q^T & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} T_k+\delta\alpha I & \hat{\delta p}+\tau e_n \\ \hat{\delta p}^T+\tau e_n^T & 0 \end{pmatrix} .$$

The form of the matrix on the right hand side of (5.12) is

$$(5.13)\qquad \begin{bmatrix} x & x & & & & & x \\ x & x & x & & & & x \\ & x & x & x & & & x \\ & & x & x & x & & x \\ & & & x & x & x & x \\ & & & & x & x & x \\ x & x & x & x & x & x & 0 \end{bmatrix} ,$$

where the x's denote nonzero elements. Gaussian elimination with partial pivoting preserves this structure if the pivots are taken from the tridiagonal part until the very last elimination step. The result of this strategy applied to (5.13) will be of the form

$$(5.14)\qquad \begin{bmatrix} x & x & + & & & & x \\ m & x & x & + & & & x \\ & m & x & x & + & & x \\ & & m & x & x & + & x \\ & & & m & x & x & x \\ & & & & m & x & x \\ m & m & m & m & m & m & + \end{bmatrix}$$

where m's denote multipliers which have overwritten the original matrix elements, and the +'s denote possible fill-in due to pivoting.

With this scheme only one expensive factorization is required. The rest of the iteration is performed under the transformation (5.11) and only after convergence to a vector $\hat{p}$ is obtained do we transform back to get $p = Q^T\hat{p}$ as a solution to problem (2.1).

Since the factorization given in equation (5.11) is roughly four times as expensive as a Cholesky factorization we might wish to consider the following alternate scheme. The system at step 2.2 of Algorithm (5.10) is equivalent (via symmetric permutation) to one of the form

$$\text{(5.15)} \qquad \begin{pmatrix} 0 & p^T \\ p & B \end{pmatrix}\begin{pmatrix} \delta\alpha \\ \delta p \end{pmatrix} = \begin{pmatrix} \rho \\ r \end{pmatrix} .$$

Use a single Householder transformation Q_1 to obtain

$$\text{(5.16)} \qquad \begin{pmatrix} 1 & 0 \\ 0 & Q_1 \end{pmatrix}\begin{pmatrix} 0 & p^T \\ p & B \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & Q_1^T \end{pmatrix} = \left[\begin{array}{cc:c} 0 & \tau & 0^T \\ \tau & \theta & v^T \\ \hdashline 0 & v & B \end{array}\right] ,$$

where

$$\begin{pmatrix} 0 & v^T \\ v & B \end{pmatrix} = Q_1 B Q_1^T \quad \text{and} \quad \tau = \|p\| ,$$

with $v \in \mathbb{R}^{n-1}$, $\theta \in \mathbb{R}$, $\hat{B} \in \mathbb{R}^{(n-1)\times(n-1)}$. The matrix on the right of (5.16) has the <u>exact</u> factorization

$$\text{(5.17)} \qquad \left[\begin{array}{cc:c} 1 & 0 & \\ 0 & 1 & \\ \hdashline \frac{1}{\tau}v & 0 & I_{n-1} \end{array}\right]\left[\begin{array}{cc:c} 0 & \tau & \\ \tau & \theta & \\ \hdashline & & \hat{B} \end{array}\right]\left[\begin{array}{cc:c} 1 & 0 & \frac{1}{\tau}v^T \\ 0 & 1 & 0^T \\ \hdashline & & I_{n-1} \end{array}\right] .$$

The eigenvalues of $\hat{B}$ separate the eigenvalues of B so $\hat{B}$ is positive definite when B is. Moreover, $\hat{B}$ is better conditioned than B whenever the separation is strict. (For a proof of separation see Wilkinson [27, pp. 95-104].) A solution to (5.15) is now possible using a Cholesky factorization of B together with factorization (5.17). The purpose of arranging the calculation this way is to avoid "pivoting" on the matrix B which is the essential result of factoring forward at step (2.2) of Algorithm (5.10).

This second scheme is much better suited to the problem of obtaining an initial guess α_0, p_0 at step 1 of Algorithm (5.10). If B is positive definite then we want to compute $p = -B^{-1}g$ and check to see if $\|p\| \leq \Delta$. If $\|p\| > \Delta$ then take $p_0 = \frac{\Delta}{\|p\|} p$. Thus it will be advantageous to attempt the computation of the Cholesky factorization of B. If B is not positive definite then we should compute α_0 so that B_{α_0} is positive definite and then take

$$p_0 = - \frac{\Delta}{\|B_{\alpha_0}^{-1} g\|} B_{\alpha_0}^{-1} b$$

as an initial guess. Various schemes for computing α_0 are possible. See Gay [7] for example.

Safeguarding this iteration is possible. At present several schemes are being considered but none of these are elegant. Therefore we shall postpone discussion of safeguarding at this time.

The decision to stop the iteration should be based upon the following tests:

Require α_{k+1}, p_{k+1} to satisfy

(a) $B_{k+1} := B + \alpha_{k+1} I$ positive semidefinite ,

(b) $|\delta\alpha| \|\delta p\| \leq \begin{cases} \varepsilon_1 \|g\| & \text{if } g \neq 0 \\ \varepsilon_1 \Delta & \text{if } g = 0 \end{cases}$,

where $\delta\alpha = \alpha_{k+1} - \alpha_k$, $\delta p = p_{k+1} - p_k$;

and (c) $|\|p_{k+1}\| - \Delta| \leq \varepsilon_2 \Delta$.

Note (from step 2.2 of Algorithm 5.10) that $(B+\alpha_{k+1}I)p_{k+1} = -g+\delta\alpha\delta p$. Therefore, if $g \neq 0$ then conditions (a) and (b) together with Lemma (2.8) imply that p_{k+1} solves

$$\min\{f + \tilde{g}^T w + \frac{1}{2} w^T B w: \quad \|w\| \leq \tilde{\Delta}\}$$

where $(1-\varepsilon_2)\Delta \leq \tilde{\Delta} \leq (1+\varepsilon_2)\Delta$ and $\tilde{g} = g+\delta g$ with $\|\delta g\| \leq \varepsilon_1\|g\|$ and $\|p_{k+1}\| = \tilde{\Delta}$. On the other hand, if $g = 0$ then p_{k+1} will be an approximate eigenvector for B which satisfies

$$\frac{\|Bp_{k+1}+\alpha_{k+1}p_{k+1}\|}{\|p_{k+1}\|} \leq \varepsilon_1$$

with α_{k+1} on approximation to $-\lambda_1$. Thus $\varepsilon_1 > 0$ should be taken quite small and $\varepsilon_2 > 0$ moderately small.

When these stopping rules are in effect the remarks at the end of Section 2 will apply. Therefore, the analysis of Section 4 will apply to the modified Newton iteration when the step is computed in the way described here.

6. Conclusions

The main purpose of this work has been to discuss the theory of the model trust region modification of Newton's method with an aim towards understanding the best way to implement it. Because of this goal we introduced sufficient generality into the analysis so that it would apply to many possible implementations based upon various factorizations of the Hessian matrix. Results similar to the second order properties given in Section 4 have been stated without proof by Gay in [8]. We feel that it is important to give proofs of these facts. This is of particular interest because no proof has been given that ensures convergence of the entire sequence (unless we make the assumption of a nonsingular Hessian at any critical point). This is despite the fact that the situation would have to be extremely pathological even in theory for convergence not to occur.

The basic ideas for possible implementations we have set forth in Section 5 are new alternatives which have been directed towards overcoming the theoretical difficulties of the locally constrained quadratic minimization

discussed in Section 2. In particular we considered using the Bunch-Parlett factorization and we also considered basing our method of solution on a more properly posed problem. It will be interesting to examine the behavior of these implementations in practice.

Finally, we have not overcome the problem of invariance under linear affine scalings of the variables. There is sufficient generality in the method to introduce uniformly bounded diagonal scalings of the variables. Ways to choose these scalings has been discussed by Fletcher [4], Gay [8], and Moré [18]. it is most appropriate to note here that the reason is that our method of proof of convergence is essentially based upon not doing worse than steepest descent at any step and this introduces a term that makes calculation of the step scale dependent. Nevertheless, we expect good performance on practical problems especially in the case that the variables can be well scaled.

7. Acknowledgement

Much of this work was done originally in 1977 at the Applied Mathematics Division of Argonne National Laboratory. Since that time it has undergone several revisions due to the encouragement of Jorge Moré. It is a pleasure to take this opportunity to thank him for the numerous discussions we have had over this work and to acknowledge what a valuable contribution his advice has been. I would also like to thank Roger Fletcher for several penetrating discussions concerning the material in Section 5. In particular, the factorization (5.17) came directly from one of these discussions.

8. References

[1] Aasen, J. O., On the reduction of a symmetric matrix to tridiagonal form, BIT 11, 233-242, 1971 .

[2] Bunch, J. R. and B. N. Parlett, Direct methods for solving symmetric indefinite systems of linear equations, SIAM J. Numer. Anal. 8, 639-655 1971.

[3] Bunch, J. R. and L. Kaufman, Some stable methods for calculating inertia and solving symmetric linear equations, Math. Comp. 31, 163-179, 1977.

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

[5] Fletcher, R. and T. L. Freeman, A modified Newton method for minimization, J.O.T.A. 23 (3), November 1977.

[6] Forsythe, G. E. and G. H. Golub, On the stationary values of a second-degree polynomial on the unit sphere, J. Soc. Indust. Appl. Math. 13 (4), December 1965.

[7] Gay, D. M., Computing Optimal Locally Constrained Steps, U. of Wisconsin MRC Rept. #2013, October 1979.

[8] Gay, D. M., On solving robust and generalized linear regressions problems, U. of Wisconsin MRC Rept. #2000, September 1979.

[9] Gill, P. E. and W. Murray, Newton type methods for unconstrained and linearly constrained optimization, Math. Prog. 7, 311-350, 1974.

[10] Goldfarb, D., Curvilinear path steplength algorithms for minimization which use directions of negative curvature, Report CCNY-CS-77-101, Dept. of Computer Science, City College of City University of New York, 1977.

[11] Goldfeld, S. M., R. E. Quandt, and H. F. Trotter, Maximization by quadratic hill-climbing, Econometrica 34, 541-551, 1966.

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

[13] Kaniel and Dax, A modified Newton's method for unconstrained minimization, SIAM J. Num. Anal., 1979.

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

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

[16] McCormick, G., A modification of Armijo's step-size rule for negative curvature, Math. Prog. 13, 111-115, 1977.

[17] Mirskey, L., An introduction to linear algebra, Oxford University Press (Clarendon), London and New York, 1955.

[18] Moré, J. J., The Levenberg-Marquardt algorithm: implementation and theory, pp. 105-116 of Lecture Notes in Mathematics 630, G. A. Watson, ed., Springer-Verlag, Berlin, Heidelberg and New York, 1978.

[19] Moré, J. J. and D. C. Sorensen, On the use of directions of negative curvature in a modified Newton method, Math. Programming 16, 1-20, 1979.

[20] Mukai, H. and E. Polak, A second order method for unconstrained optimization, J.O.T.A. 26, 501-513, 1978.

[21] Ortega, J. M. and W. C. Rheinboldt, Iterative solution of nonlinear equations in several variables. Academic Press, 1970.

[22] Powell, M. J. D., A hybrid method for nonlinear equations, pp. 87-114 of Numerical method for nonlinear algebraic equations, P. Rabinowitz, ed., Gordon and Breach, London, 1970.

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

[24] Reinsch, C. H., Smoothing by spline functions. II, Number. Math. 16, 451-454, 1971.

[25] Stewart, G. W., Introduction to matrix computations, Academic Press, New York, 1973.

[26] Thomas, S. W., Sequertial estimation techniques for quasi-Newton algorithms, Ph.D. thesis, Department of Computer Science, Cornell University, Report TR75-227, January 1975.

[27] Wilkinson, J. H., The algebraic eigenvalue problem, Oxford University Press (Clarendon), London and New York, 1965.

[28] Vial, J.-P. and I. Zang, Unconstrained optimization by approximation of the gradient path, CORE discussion paper, June 1975.