DAMTP 1997/NA18

Trust region calculations revisited[1]

M.J.D. Powell

Abstract: Trial changes to the variables in many trust region algorithms for optimization calculations are derived from the following problem. Minimize a quadratic function $Q(\underline{d})$, $\underline{d} \in \mathcal{R}^n$, subject to $\|\underline{d}\| \leq \Delta$, where Q and $\Delta \in \mathcal{R}$ are given, where the vector norm is Euclidean, and where n is the number of variables. Each trial change is an estimate of a solution of this problem, some estimates being better than others. We argue that the estimate of Moré and Sorensen (1983) should be employed more often, because it gives a value of $Q(\underline{d})$ that differs from the optimal value by at most a prescribed tolerance. On the other hand, some researchers believe that achieving this property requires an excessive amount of computation. Therefore the details of the method are reviewed. We note that the tolerance can usually be satisfied by very few cycles of an iterative procedure. Each cycle requires the solution of a system of equations of the form $(\nabla^2 Q + \lambda I)\, \underline{d} = -\underline{\nabla} Q(0)$, where λ is a nonnegative parameter that is adjusted by each cycle. Thus the amount of computation is of magnitude n^3 for general $\nabla^2 Q$. When a sequence of trust region problems is solved, however, and when each new $\nabla^2 Q$ differs from the previous one by a matrix of rank one or two, which happens in quasi-Newton algorithms for unconstrained optimization, then the work of each iteration can be reduced to $\mathcal{O}(n^{7/3})$ by applying orthogonal similarity transformations to $\nabla^2 Q$ (Powell, 1997). This technique is described. It should allow the Moré and Sorensen method to be used routinely for much larger values of n than before.

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

September, 1997.

[1]Presented at the 17th Biennial Conference on Numerical Analysis (Dundee).

1. Introduction

Trust region calculations are an important part of many optimization algorithms (see Fletcher, 1987, for instance). In particular, a trust region algorithm for minimizing an objective function $F(\underline{x})$, $\underline{x} \in \mathcal{R}^n$, when there are no constraints on the variables, may proceed as follows. An iterative method is applied until termination. We let $\underline{\widehat{x}} \in \mathcal{R}^n$ be the current vector of variables at the start of an iteration. A quadratic approximation to the objective function, namely Q, is also required, and we write it in the form

$$F(\underline{\widehat{x}}+\underline{d}) \approx Q(\underline{d}) = F(\underline{\widehat{x}}) + \underline{d}^T\underline{g} + \tfrac{1}{2}\underline{d}^T B\,\underline{d}, \qquad \underline{d}\in\mathcal{R}^n. \tag{1.1}$$

The vector $\underline{g}$ is often the gradient $\underline{\nabla}F(\underline{\widehat{x}})$, and B is an $n\times n$ symmetric matrix, which is sometimes the second derivative matrix $\nabla^2 F(\underline{\widehat{x}})$. Furthermore, a positive real number Δ is required, its purpose being that the length of the change in the variables by an iteration is not allowed to exceed Δ. Therefore we define $\underline{d}^* \in \mathcal{R}^n$ to be an "optimal trial step" if it is a solution of the subproblem

$$\text{Minimize } \; Q(\underline{d}), \quad \underline{d}\in\mathcal{R}^n, \quad \text{subject to } \; \|\underline{d}\| \le \Delta. \tag{1.2}$$

A typical iteration generates a vector $\underline{d}$ that is an approximation to $\underline{d}^*$ and that has the properties $Q(\underline{d}) < Q(0)$ and $\|\underline{d}\| \le \Delta$. Then the new function value $F(\underline{\widehat{x}}+\underline{d})$ is calculated. The iteration overwrites $\underline{\widehat{x}}$ by $\underline{\widehat{x}}+\underline{d}$ if the difference $F(\underline{\widehat{x}})-F(\underline{\widehat{x}}+\underline{d})$ is positive and acceptably large, but otherwise $\underline{\widehat{x}}$ remains unchanged. Different algorithms perform these tasks in different ways. We are going to consider the calculation of $\underline{d}$.

The name "trust region" was introduced because one can take the view that the approximation (1.1) is being trusted if $\underline{d}$ satisfies $\|\underline{d}\| \le \Delta$. We let the vector norm be Euclidean throughout this paper, which is a highly useful choice that loses generality. The trust region bound allows the symmetric matrix B to be indefinite, because in all cases the subproblem (1.2) has a solution. Furthermore, suitable choices of Δ can be helpful to the efficiency of algorithms and to theoretical convergence properties.

Our reasons for revisiting trust region algorithms are due to the works of Steihaug (1983), Moré and Sorensen (1983) and Powell (1997). The two earlier papers present different ways of generating $\underline{d}$. The Moré and Sorensen method provides a trial step $\underline{d}$ that has the property

$$F(\underline{\widehat{x}}) - Q(\underline{d}) \ge (1-\varepsilon)\,[F(\underline{\widehat{x}})-Q(\underline{d}^*)], \tag{1.3}$$

where ε is a prescribed positive constant and where $\underline{d}^*$ is the optimal trial step that has been defined already. The value $\varepsilon = 0.1$ is typical and $\varepsilon < 1$ is required. The "truncated conjugate gradient" method of Steihaug, however, can cause the ratio $[F(\underline{\widehat{x}})-Q(\underline{d})]\,/\,[F(\underline{\widehat{x}})-Q(\underline{d}^*)]$ to be much less than one when the matrix B

has one or more negative eigenvalues, as shown in an example that is given below. Nevertheless, in recent algorithms that require an approximation to the solution of the trust region subproblem (1.2), the truncated conjugate gradient method is usually preferred. It seems that this preference is becoming stronger, because many researchers believe that the amount of computation of the alternative technique of Moré and Sorensen is too expensive for algorithms for general use. We question this point of view, not only because of the advantage of achieving condition (1.3), but also because of the recent results of Powell (1997). They show that large reductions can be made in the amount of computation of the Moré and Sorensen method when B has no sparsity, and when B is updated on each iteration by a quasi-Newton formula that changes B by a matrix of low rank. These findings are presented briefly in Section 3, after consideration of the details of the Moré and Sorensen procedure, which is the subject of Section 2.

The truncated conjugate gradient method is the same as the algorithm of Fletcher and Reeves (1964), with a particular termination condition, when the objective function is $Q(\underline{d})$, $\underline{d} \in \mathcal{R}^n$, and when the initial vector of variables is $\underline{d} = 0$. Indeed, both methods increase $\|\underline{d}\|$ monotonically from zero by moving along the same piecewise linear path in $\mathcal{R}^n$ until the termination condition holds. Thus, if the optimal trial step has the property $\|\underline{d}^*\| < \Delta$, then $\underline{d} = \underline{d}^*$ is achieved, but otherwise the final $\underline{d}$ is the point on the piecewise linear path that satisfies $\|\underline{d}\| = \Delta$. In order to present the example that has been mentioned, we note also that the initial line segment of the path has the direction of $-\underline{g}$, and it goes all the way to the trust region boundary if Q has negative curvature along this direction.

We pick $\Delta = 1$, and we consider the case when Q is the quadratic function of two variables

$$Q(\underline{d}) = F(\underline{\widehat{x}}) + \underline{d}^T \underline{g} + \tfrac{1}{2} \underline{d}^T B \, \underline{d} = -10^{-6} d_1 - 10^{-6} d_1^2 - d_2^2, \qquad \underline{d} \in \mathcal{R}^2, \tag{1.4}$$

where d_1 and d_2 are the components of $\underline{d}$. Thus the initial line segment is along the first coordinate direction, so the truncated conjugate gradient method provides $\underline{d} = (1,0)^T$. On the other hand, the choice $\underline{d} = (0,1)^T$ would yield a much smaller value of $Q(\underline{d})$. Specifically, $F(\underline{\widehat{x}}) - Q(\underline{d}^*)$ is greater than 1.0, but the actual choice of $\underline{d}$ provides only $F(\underline{\widehat{x}}) - Q(\underline{d}) = 2 \times 10^{-6}$. This unwelcome situation occurs because B is ill-conditioned, and the truncated conjugate gradient method has become the method of steepest descents.

Some other methods for the trust region subproblem are proposed by Gould (1997), Rendl and Wolkowicz (1997) and Sorensen (1997). We are taking the view, however, that the procedure of Moré and Sorensen (1983) is a highly useful way of generating $\underline{d}$. Therefore the interesting properties that allow it to be implemented efficiently are reviewed in Sections 2 and 3. The value of this work to practical algorithms is discussed briefly in Section 4.

2. The method of Moré and Sorensen

Necessary and sufficient conditions for a vector $\underline{d}^*$ to be a solution of the trust region subproblem (1.2) are given below. They are of fundamental importance to several algorithms for trust region calculations, including the method of Moré and Sorensen (1983).

Characterization Theorem: Let $\underline{d}^* \in \mathcal{R}^n$ satisfy the trust region bound $\|\underline{d}^*\| \leq \Delta$. It is an "optimal trial step" if and only if there exists a nonnegative number, λ^* say, such that the following three conditions hold. (1) The matrix $B+\lambda^* I$ has no negative eigenvalues, where I is the $n \times n$ unit matrix. (2) $\underline{d}^*$ is a solution of the linear system of equations

$$(B+\lambda^* I)\, \underline{d}^* = -\underline{g}. \tag{2.1}$$

(3) If λ^* is positive, then $\underline{d}^*$ is on the trust region boundary, which is the complementarity property $\lambda^*(\Delta - \|\underline{d}^*\|) = 0$.

Outline of Proof: When seeking an optimal trial step, we can restrict attention to vectors $\underline{d}^*$ that satisfy $\|\underline{d}^*\| \leq \Delta$. Furthermore, $\underline{d}^*$ has to be a stationary point of the Lagrangian function of the subproblem (1.2), after writing the constraint in the form $\frac{1}{2}\|\underline{d}\|^2 \leq \frac{1}{2}\Delta^2$, in order to provide differentiability at $\underline{d}=0$. Therefore we can restrict attention to vectors $\underline{d}^*$ that are stationary points of the function

$$L(\underline{d}) = F(\hat{\underline{x}}) + \underline{d}^T \underline{g} + \tfrac{1}{2} \underline{d}^T B\, \underline{d} + \lambda^* \left(\tfrac{1}{2}\|\underline{d}\|^2 - \tfrac{1}{2}\Delta^2\right), \qquad \underline{d} \in \mathcal{R}^n, \tag{2.2}$$

for some $\lambda^* \geq 0$. It follows that $\underline{d}^*$ has to be a solution of the equations (2.1). Moreover, the third condition of the theorem is a well-known property of inequality constraints. Therefore conditions (2) and (3) are necessary if $\underline{d}^*$ is optimal, which defines λ^* uniquely for every $\underline{d}^*$ that can satisfy (2) and (3).

Let $\underline{d}^*$ be one of these vectors. If $\|\underline{d}^*\|$ is strictly less than Δ, then λ^* is zero, so equation (2.1) shows that $\underline{d}^*$ is a stationary point of $Q(\underline{d})$, $\underline{d} \in \mathcal{R}^n$. In this case any sufficiently small change to $\underline{d}^*$ is allowed by the trust region bound. Therefore $\underline{d}^*$ is an optimal trial step if and only if Q is convex, which is equivalent to the condition that $B+\lambda^* I$ has no negative eigenvalues. This remark allows $\|\underline{d}^*\| = \Delta$ to be assumed for the remainder of the proof.

Let $\underline{d}^*$ satisfy conditions (2) and (3) of the theorem and let it be on the trust region boundary, but let $B+\lambda^* I$ have a negative eigenvalue. Then we pick $\underline{s} \in \mathcal{R}^n$ such that $\underline{s}^T (B+\lambda^* I) \underline{s} < 0$. Further, by reversing the sign of $\underline{s}$ and by making a small perturbation to $\underline{s}$ if necessary, we may assume the strict inequality $\underline{s}^T \underline{d}^* < 0$. Then we consider points in $\mathcal{R}^n$ of the form $\underline{d}^* + \alpha\, \underline{s}$, $\alpha \geq 0$. The inequality $\underline{s}^T \underline{d}^* < 0$ implies that these points are inside the trust region for sufficiently small α, so we can let α^* be a positive number such that $\|\underline{d}^* + \alpha^* \underline{s}\| = \Delta$. Moreover, the quadratic function of one variable $\psi(\alpha) = L(\underline{d}^* + \alpha\, \underline{s})$, $\alpha \geq 0$, has the properties $\psi'(0) = 0$ and $\psi''(\alpha) < 0$, so it decreases strictly monotonically. Thus we deduce the relation

$$Q(\underline{d}^* + \alpha^* \underline{s}) = L(\underline{d}^* + \alpha^* \underline{s}) < L(\underline{d}^*) = Q(\underline{d}^*). \tag{2.3}$$

It follows from $\|\underline{d}^*+\alpha^*\underline{s}\|=\Delta$ that $\underline{d}^*$ is not a solution of the subproblem (1.2). Therefore conditions (1), (2) and (3) of the theorem are all necessary for the optimality of $\underline{d}^*$.

The sufficiency of these conditions for optimality in the case $\|\underline{d}^*\| = \Delta$ is elementary. Indeed, the conditions imply not only that $\underline{d}^*$ is a stationary point of the Lagrangian function (2.2) but also that L is convex, so we have $L(\underline{d}) \geq L(\underline{d}^*)$, $\underline{d} \in \mathcal{R}^n$. Moreover, the assumption $\|\underline{d}^*\|=\Delta$ provides the identity $L(\underline{d}^*)=Q(\underline{d}^*)$, while the definitions of Q and L provide the inequality $L(\underline{d}) \leq Q(\underline{d})$, $\|\underline{d}\| \leq \Delta$. Thus we deduce the bound $Q(\underline{d}) \geq Q(\underline{d}^*)$, $\|\underline{d}\| \leq \Delta$, which shows that $\underline{d}^*$ is an "optimal trial step". The proof is complete. $\square$

One can expose some of the information in the characterization theorem by assuming that B is a diagonal matrix. Many of the conclusions in this special case apply generally, because B can be made diagonal by an orthogonal similarity transformation, and such transformations preserve the Euclidean lengths of vectors, in particular the trust region bound. Therefore we assume that the diagonal elements of B are its eigenvalues λ_i, $i=1,2,\ldots,n$, and that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Hence $\underline{d}^*$ is optimal if and only if the equations

$$(\lambda_i+\lambda^*)\, d_i^* = -g_i, \quad i=1,2,\ldots,n, \quad \text{and} \quad \lambda^*(\Delta-\|\underline{d}^*\|) = 0 \tag{2.4}$$

hold for some λ^* in the set $\mathcal{L} = \{\lambda : \lambda \geq \max[0, -\lambda_1]\}$, where d_i^* and g_i, $i = 1,2,\ldots,n$, are the components of $\underline{d}^*$ and $\underline{g}$. We also take the view that Δ is a parameter. Therefore, if $\underline{d}^*$ satisfies the first part of expression (2.4), then it is an optimal trial step in the case $\Delta=\|\underline{d}^*\|$, provided that λ^* is in $\mathcal{L}$.

When λ is in the set $\mathcal{L}^- = \mathcal{L}\backslash\{-\lambda_1\}$, there is a unique vector $\underline{d}(\lambda)$ that has the components $-g_i/(\lambda_i+\lambda)$, $i=1,2,\ldots,n$. Similarly, if B is a general symmetric matrix and $\lambda \in \mathcal{L}^-$, then $B+\lambda I$ is positive definite and we define $\underline{d}(\lambda)$ by the system

$$(B+\lambda I)\, \underline{d}(\lambda) = -\underline{g}. \tag{2.5}$$

It follows that $\underline{d}(\lambda)$ is the required vector $\underline{d}^*$ if its length is Δ. Moreover, by considering the case when B is diagonal, we find that $\|\underline{d}(\lambda)\|$ is a continuous and monotonically decreasing function of $\lambda \in \mathcal{L}^-$. In theory, therefore, the characterization of $\underline{d}^*$ can be divided into three disjoint cases. Specifically, the "Newton–Raphson case" is when B is positive definite and $\underline{d}^*=\underline{d}(0)$, because the trust region radius satisfies $\Delta \geq \|\underline{d}(0)\|$. The "ordinary case" is when both λ^* and $\lambda_1+\lambda^*$ are positive, so $\underline{d}^*$ is the vector $\underline{d}(\lambda^*)$, where λ^* is the only element of $\mathcal{L}^-$ that solves the equation $\|\underline{d}(\lambda^*)\|=\Delta$. Thirdly, the "hard case" is when $\lambda_1+\lambda^*$ is zero, so $B+\lambda^* I$ is a singular matrix with nonnegative eigenvalues, and the equations (2.5) do not have a unique solution. This variety of possibilities may be one of the reasons for the apparent lack of popularity of the method of Moré and Sorensen (1983).

The method may try several nonnegative values of λ, and for each one it seeks the Cholesky factorization of the matrix $B+\lambda I$, in order to calculate $\underline{d}(\lambda)$ by solving the system (2.5) if $B+\lambda I$ is positive definite. If the factorization attempt fails, then $B+\lambda I$ has a negative eigenvalue, so λ is less than λ^*. In practice one may not be able to identify the case when the factorization exists and $B+\lambda I$ is singular, but, if it occurs, then one has encountered the "hard case" just when the equations (2.5) have a solution that satisfies $\|\underline{d}(\lambda)\| \leq \Delta$. Then the strict inequality $\|\underline{d}(\lambda)\| < \Delta$ holds for every larger value of λ, so λ^* is defined by the remark that $B+\lambda^* I$ is a singular matrix with no negative eigenvalues. Usually, however, $B+\lambda I$ is positive definite and $\underline{d}(\lambda)$ is well-defined. Then λ is too small if $\underline{d}(\lambda)$ violates the trust region bound, while a positive value of λ is too large if $\|\underline{d}(\lambda)\|$ is strictly less than Δ. Of course it is important in practice to have a tolerance condition to decide whether or not to try another value of λ. The Moré and Sorensen method includes a highly efficient solution to this question, which will be described later.

In the "ordinary case", the method seeks the solution of $\|\underline{d}(\lambda^*)\| = \Delta$ by applying an extension of the Newton–Raphson algorithm to the equation $\phi(\lambda)=0$, where ϕ is the function

$$\phi(\lambda) = \Delta^{-1} - \|\underline{d}(\lambda)\|^{-1}, \qquad \lambda \in \mathcal{L}^-. \tag{2.6}$$

Some highly useful properties of this function can be identified easily by letting B be the diagonal matrix of eigenvalues again. Indeed, assuming $\underline{g} \neq 0$, we find the formulae

$$\left.\begin{array}{rcl} \phi(\lambda) & = & \Delta^{-1} - \{\sum_{i=1}^n g_i^2/(\lambda_i+\lambda)^2\}^{-1/2} \\ \phi'(\lambda) & = & -\{\sum_{i=1}^n g_i^2/(\lambda_i+\lambda)^3\}\,\{\sum_{i=1}^n g_i^2/(\lambda_i+\lambda)^2\}^{-3/2} \\ \phi''(\lambda) & = & 3\,\beta(\lambda)\,\{\sum_{i=1}^n g_i^2/(\lambda_i+\lambda)^2\}^{-5/2} \end{array}\right\}, \quad \lambda \in \mathcal{L}^-, \tag{2.7}$$

where $\beta(\lambda)$ is the number

$$\beta(\lambda) = \{\textstyle\sum_{i=1}^n g_i^2/(\lambda_i+\lambda)^4\}\,\{\sum_{i=1}^n g_i^2/(\lambda_i+\lambda)^2\} - \{\sum_{i=1}^n g_i^2/(\lambda_i+\lambda)^3\}^2. \tag{2.8}$$

Hence $\phi(\lambda)$, $\lambda \in \mathcal{L}^-$, decreases strictly monotonically and is convex, because we deduce the property $\phi''(\lambda) \geq 0$, $\lambda \in \mathcal{L}^-$, by applying the Cauchy–Schwarz inequality to expression (2.8). Therefore the first derivative $\phi'(\lambda)$ is an increasing function of $\lambda \in \mathcal{L}^-$. Further, by taking limits as λ tends to $\max[0, -\lambda_1]$ from above and to $+\infty$ from below, we derive the bounds

$$-1\,/\,|g_\ell| \leq \phi'(\lambda) \leq -1\,/\,\|\underline{g}\|, \qquad \lambda \in \mathcal{L}^-, \tag{2.9}$$

where ℓ is the least integer in $[1, n]$ such that $g_\ell \neq 0$. Thus $|\phi'(\lambda)|$, $\lambda \in \mathcal{L}^-$, is finite and is bounded away from zero, which is helpful to the Newton–Raphson algorithm for solving $\phi(\lambda) = 0$. Therefore a strong advantage has been gained by using the inverse of $\|\underline{d}(\lambda)\|$ in expression (2.6). On the other hand, the first derivative of $\|\underline{d}(\lambda)\|$ or $\|\underline{d}(\lambda)\|^2$ with respect to λ usually becomes infinite if λ tends to $-\lambda_1$. These useful properties of ϕ are retained when B is a general symmetric matrix.

The "hard case" occurs only if ℓ is greater than one in expression (2.9). Then the lower limits of all the sums of formulae (2.7) and (2.8) can be increased from $i=1$ to $i=\ell$. Thus, if $\lambda_\ell > \lambda_1$, the values and derivatives of $\phi(\lambda)$, $\lambda \in \mathcal{L}^-$, provide no indication that the least eigenvalue of B is λ_1. Therefore an algorithm for calculating an estimate of $\underline{d}^*$ may require some information that is independent of ϕ. We will find that suitable information is present in the tolerance condition of Moré and Sorensen that has been mentioned already.

The method of this section employs upper and lower bounds on λ^*, namely λ_L and λ_U, but there is no need to set λ_U initially. The characterization theorem provides the first two terms of the initial choice

$$\lambda_L = \max\,[\,0,\, \max\{-B_{ii} : i=1,2,\ldots,n\},\, -\|B\|_1 + \|\underline{g}\|/\Delta\,], \tag{2.10}$$

while the last term is derived as follows. The system (2.5), the triangle inequality and the trust region bound imply $\|B\,\underline{d}^*\| + \lambda^*\Delta \geq \|\underline{g}\|$. Furthermore, $\|B\,\underline{d}^*\|$ is at most Δ times the largest modulus of an eigenvalue of B, and these moduli cannot exceed $\|B\|_1$. Hence the last part of expression (2.10) is a lower bound on λ^*, which is particularly useful when Δ is very small. Expression (2.10) can also be the first trial value of λ, but, when a trust region subproblem is solved by each iteration of an algorithm for unconstrained optimization, it may be better to let the initial λ depend on the final λ of the previous iteration (Moré and Sorensen, 1983).

Every trial λ satisfies $\lambda \geq \lambda_L$, and next we consider the case when the Cholesky factorization of $B+\lambda I$ fails because of a negative eigenvalue. Then the attempt to calculate the factorization provides a least integer k in $[1,n]$ such that the leading $k \times k$ submatrix of $B+\lambda I$ is not positive definite. Further, we let $\underline{u}$ be the vector in $\mathcal{R}^n$ whose last $n-k$ components are zero, whose k-th component is one, and, if $k \geq 2$, the first $k-1$ components of $\underline{u}$ are chosen so that $\underline{u}$ is orthogonal to the $k-1$ columns of the partial factorization that has been formed already. This construction provides the property $\underline{u}^T(B+\lambda I)\,\underline{u} \leq 0$, while the Rayleigh quotient theorem states that $\underline{u}^T(B+\lambda I)\,\underline{u}/\|\underline{u}\|^2$ is an upper bound on the least eigenvalue of $B+\lambda I$. Now the required λ^* is bounded below by minus any upper bound on the least eigenvalue of B. Therefore the construction of this paragraph allows λ_L to be replaced by the number

$$\lambda_L = \lambda - \underline{u}^T(B+\lambda I)\,\underline{u}\,/\,\|\underline{u}\|^2, \tag{2.11}$$

which is no less than the previous value, due to $\underline{u}^T(B+\lambda I)\,\underline{u} \leq 0$.

Next a new trial value of λ is selected, and we make a choice that differs from the one that is stated by Moré and Sorensen (1983). If the upper bound on λ^*, namely λ_U, has not been set, we aim for a slight over-estimate of λ^*. Specifically, because $\phi(\lambda)$ decreases from Δ^{-1} to 0 as λ increases from $-\lambda_1$ to λ^*, and because condition (2.9) includes the inequality $\phi'(\lambda) \le -1/\|\underline{g}\|$, we let the new value of λ be $\lambda_L + \|\underline{g}\|/\Delta$. Alternatively, if λ_U exists, then $\phi(\lambda_U)$ and $\phi'(\lambda_U)$ are available, the value of $\phi(\lambda_U)$ being negative. Further, the Newton–Raphson step from λ_U must have provided an estimate of λ^* that is too low for the stopping condition of our calculation. Therefore we do not assume that $\phi(\lambda)$, $\lambda \in \mathcal{L}^-$, is approximately linear. Instead we write our choice of λ in the form $\lambda_L + \theta(\lambda_U - \lambda_L)$, and we pick $\theta \le 0.1$, in order that the new λ is likely to be in the interval $[-\lambda_1, \lambda^*]$. The reason for allowing $\theta < 0.1$ occasionally is that, due to the convexity of ϕ that is mentioned after equation (2.8), the new λ should not exceed any estimate that is given by applying the rule of false position to a bracket on the zero of ϕ. We employ not only the bracket $\phi(-\lambda_1) = \Delta^{-1}$ and $\phi(\lambda_U) < 0$ but also the assumption that λ_L is very close to the unknown number $-\lambda_1$. These remarks give the new value

$$\lambda = \lambda_L + \min[\, 0.1,\ \{1 - \Delta\, \phi(\lambda_U)\}^{-1}\,]\, (\lambda_U - \lambda_L), \tag{2.12}$$

when λ_U exists. Then there is a branch back to the factorization at the beginning of the previous paragraph, so the procedure cycles until a Cholesky factorization of $B + \lambda I$ is found.

If special attention were given to the case when $B + \lambda I$ is singular and the factorization

$$B + \lambda I = L L^T \tag{2.13}$$

exists, where L is a lower triangular matrix, then the algorithm would have to perform two tasks that would not be needed otherwise. Firstly, there would have to be a test for the consistency of the system (2.5), which could be assisted by letting the nonzero columns of L be linearly independent. Then, if the test were satisfied, the shortest solution of the system (2.5) would be required, so $\underline{d}(\lambda)$ would be in the linear subspace of $\mathcal{R}^n$ that is spanned by the columns of L. It is highly unlikely in practice, however, that $B + \lambda I$ would be found to be exactly singular, and, if this case occurred, then probably the test for consistency would fail. Therefore, if $B + \lambda I$ is singular, we prefer to adjust λ by the method of the previous paragraph.

It remains to address the actions that occur when a trial value of λ causes $B + \lambda I$ to be positive definite. In this case the values of both $\phi(\lambda)$ and $\phi'(\lambda)$ are usually required. The calculation of $\phi(\lambda)$ is straightforward, because the triangular matrix L of the factorization (2.13) has positive diagonal elements. Therefore the system (2.5) is solved by a forward substitution that provides $\underline{y}$ such that $L\underline{y} = -\underline{g}$, followed by a back substitution that obtains $\underline{d}(\lambda)$ from $L^T \underline{d}(\lambda) = \underline{y}$. Then $\phi(\lambda)$ is set to the value (2.6). We also obtain the solution $\underline{\widehat{y}}$ of the triangular system $L\underline{\widehat{y}} = \underline{d}(\lambda)$, because it is important to the calculation of $\phi'(\lambda)$. Here we make use of the property

$$\phi'(\lambda) = \frac{d}{d\lambda} \left\{ -\sum_{i=1}^{n} d_i(\lambda)^2 \right\}^{-1/2} = \frac{\underline{d}(\lambda)^T \underline{d}'(\lambda)}{\|\underline{d}(\lambda)\|^3}, \qquad (2.14)$$

where $\underline{d}'(\lambda)$ is defined by the relation

$$\{B+\lambda I+\eta I\}\,\{\underline{d}(\lambda)+\eta\,\underline{d}'(\lambda)\} = -\underline{g} + \mathcal{O}(\eta^2) \qquad (2.15)$$

as $\eta \to 0$. Specifically, the first order terms in η provide the equation

$$(B+\lambda I)\,\underline{d}'(\lambda) = -\underline{d}(\lambda), \qquad (2.16)$$

so the factorization (2.13) continues to be useful. Indeed, we find the identity

$$\underline{d}(\lambda)^T \underline{d}'(\lambda) = \underline{d}(\lambda)^T \{-(LL^T)^{-1} \underline{d}(\lambda)\} = -\|L^{-1}\underline{d}(\lambda)\|^2 = -\|\underline{\widehat{y}}\|^2. \qquad (2.17)$$

Thus we deduce the formula $\phi'(\lambda) = -\|\underline{\widehat{y}}\|^2 / \|\underline{d}(\lambda)\|^3$, which is highly convenient in practice. It is a generalization of the middle line of expression (2.7).

When the current matrix $B+\lambda I$ is positive definite and another value of λ has to be tried, then $\phi(\lambda)$ and $\phi'(\lambda)$ are calculated by the method of the previous paragraph. They provide the Newton–Raphson choice

$$\lambda_{NR} = \lambda - \phi(\lambda) \,/\, \phi'(\lambda), \qquad (2.18)$$

which is a lower bound on λ^*, due to the convexity of ϕ. We also pay attention to the rule of false position, as it gives an upper bound on λ^*. Therefore, after setting λ_L or λ_U to the current λ when $\phi(\lambda)$ is positive or negative, respectively, we let λ_{FP} be the number

$$\lambda_{FP} = \{\phi(\lambda_L)\,\lambda_U - \phi(\lambda_U)\,\lambda_L\} \,/\, \{\phi(\lambda_L) - \phi(\lambda_U)\}, \qquad (2.19)$$

if both $\phi(\lambda_L)$ and $\phi(\lambda_U)$ are available. If λ_U has not been set, however, then we define $\lambda_{FP} = \lambda_L + \phi(\lambda_L)\,\|\underline{g}\|$, because it is an over-estimate of λ^*, due to the second of the inequalities (2.9). Moreover, if $\phi(\lambda_L)$ is not present, which occurs when the matrix $B+\lambda_L I$ is not positive definite, then we replace $\phi(\lambda_L)$ by Δ^{-1} in expression (2.19). We let the new λ be at least $\max[\lambda_L, \lambda_{NR}]$ and at most λ_{FP}, although Moré and Sorensen do not make use of λ_{FP}. Specifically, we propose the updating formula

$$\lambda_{\text{new}} = \begin{cases} 0.9\,\lambda_{NR} + 0.1\,\lambda_{FP}, & \phi(\lambda) > 0, \\ \max[\,\lambda_{NR},\, \min\{\lambda_{FP},\, 0.9\,\lambda_L + 0.1\,\lambda_U\}\,], & \phi(\lambda) < 0. \end{cases} \qquad (2.20)$$

It provides a bias towards the fast convergence of the Newton–Raphson method in a way that tends to replace λ_L and λ_U alternately, if several values of λ are tried. Further, if the old value of λ is λ_U, and if the new value has to be in the first tenth of the interval $[\lambda_L, \lambda_U]$, then the second line of the formula sets $\lambda = \lambda_{FP}$, because the circumstances may make this choice more suitable than λ_{NR}.

A brilliant feature of the method of Moré and Sorensen (1983) is that the adjustments to λ are terminated not when condition (1.3) and $\|\underline{d}\| \leq \Delta$ are satisfied by $\underline{d} = \underline{d}(\lambda)$, but when $\underline{d}(\lambda)$ allows the construction of a vector $\underline{d}$ that achieves these conditions. We begin our consideration of this technique by dismissing the case when $\|\underline{d}^*\|$ has to be less than Δ. Indeed, if this case occurs, then B is positive definite, and expression (2.10) gives $\lambda_L = 0$, so it is usual to calculate the Cholesky factorization of B initially. Hence $\underline{d}(0) = \underline{d}^*$ is found by solving the system (2.5), which allows the strict inequality $\|\underline{d}^*\| < \Delta$ to be identified automatically. Therefore we assume for the remainder of this section that the trust region bound is the equation $\|\underline{d}\| = \Delta$.

Then the test for termination is applied for all trial values of λ in the algorithm such that $B + \lambda I$ is positive definite. It depends on the Cholesky factorization (2.13), on the solution $\underline{d}(\lambda)$ of the system (2.5), and on a vector $\underline{z} \in \mathcal{R}^n$ that is chosen so that $\underline{d} = \underline{d}(\lambda) + \underline{z}$ satisfies the constraint $\|\underline{d}\| = \Delta$. Specific choices of $\underline{z}$ will be considered after a proof of a crucial discovery of Moré and Sorensen (1983). It states that the vector $\underline{d}$ achieves the required condition (1.3) if $\underline{z}$ has the property

$$\|L^T \underline{z}\|^2 \leq \varepsilon \left(\|L^T \underline{d}(\lambda)\|^2 + \lambda\, \Delta^2 \right). \tag{2.21}$$

The justification of this assertion depends on the remark that $\underline{d}(\lambda)$ is a stationary point of the convex quadratic function $Q(\underline{d}) + \frac{1}{2}\lambda\, \|\underline{d}\|^2$, $\underline{d} \in \mathcal{R}^n$. Further, LL^T is the Cholesky factorization of the second derivative matrix of this function. Thus we deduce the identity

$$Q(\underline{d}) + \tfrac{1}{2}\lambda\, \|\underline{d}\|^2 = Q(\underline{d}(\lambda)) + \tfrac{1}{2}\lambda\, \|\underline{d}(\lambda)\|^2 + \tfrac{1}{2}\, \|L^T \{\underline{d} - \underline{d}(\lambda)\}\|^2, \quad \underline{d} \in \mathcal{R}^n. \tag{2.22}$$

By subtracting this relation from its form when $\underline{d} = 0$, we find the equation

$$F(\underline{\widehat{x}}) - Q(\underline{d}) = \tfrac{1}{2}\lambda\, \|\underline{d}\|^2 + \tfrac{1}{2}\, \|L^T \underline{d}(\lambda)\|^2 - \tfrac{1}{2}\, \|L^T \{\underline{d} - \underline{d}(\lambda)\}\|^2, \quad \underline{d} \in \mathcal{R}^n. \tag{2.23}$$

Therefore the choice of $\underline{d}$ in the previous paragraph and inequality (2.21) provide the bound

$$\begin{aligned} F(\underline{\widehat{x}}) - Q(\underline{d}) &= \tfrac{1}{2}\lambda\, \Delta^2 + \tfrac{1}{2}\, \|L^T \underline{d}(\lambda)\|^2 - \tfrac{1}{2}\, \|L^T \underline{z}\|^2 \\ &= \left(1 - \frac{\|L^T \underline{z}\|^2}{\lambda\, \Delta^2 + \|L^T \underline{d}(\lambda)\|^2} \right) \left(\tfrac{1}{2}\lambda\, \Delta^2 + \tfrac{1}{2}\, \|L^T \underline{d}(\lambda)\|^2 \right) \\ &\geq (1 - \varepsilon) \left(\tfrac{1}{2}\lambda\, \Delta^2 + \tfrac{1}{2}\, \|L^T \underline{d}(\lambda)\|^2 \right). \end{aligned} \tag{2.24}$$

Moreover, by letting $\underline{d} = \underline{d}^*$ in equation (2.23), we obtain the condition

$$F(\underline{\widehat{x}}) - Q(\underline{d}^*) \leq \tfrac{1}{2}\lambda\, \Delta^2 + \tfrac{1}{2}\, \|L^T \underline{d}(\lambda)\|^2. \tag{2.25}$$

It follows from $\varepsilon < 1$ that the right hand side of expression (2.24) is bounded below by $(1-\varepsilon)\,[F(\underline{\widehat{x}})-Q(\underline{d}^*)]$. Therefore the assertion at the end of the previous paragraph is true.

We consider inequality (2.21) for the choice

$$\underline{d} = \{\Delta \,/\, \|\underline{d}(\lambda)\|\}\, \underline{d}(\lambda). \tag{2.26}$$

It is straightforward to express the inequality in the form

$$\left(1 - \frac{\Delta}{\|\underline{d}(\lambda)\|}\right)^2 \leq \varepsilon \left(1 + \frac{\lambda\,\Delta^2}{\|L^T\underline{d}(\lambda)\|^2}\right), \tag{2.27}$$

because $\underline{z}$ is the vector $\{-1+\Delta/\|\underline{d}(\lambda)\|\}\,\underline{d}(\lambda)$. Therefore, if $\underline{d}(\lambda)$ satisfies condition (2.27), then termination occurs, and the algorithm provides the vector (2.26) as the approximation to $\underline{d}^*$. It is highly advantageous that no more trial values of λ are required if $|1-\Delta/\|\underline{d}(\lambda)\||$ is less than $\sqrt{\varepsilon}$.

Another choice of $\underline{z}$ may also be used in order to respond to the difficulty of the "hard case". This is done for the current λ if inequality (2.27) fails, if $\phi(\lambda)<0$, which implies $\lambda>\lambda^*$, and if the matrix $B+\lambda_L I$ is not known to be positive definite, where λ_L is still the current lower bound on λ^*. The last condition always holds when $B+\lambda^* I$ is singular, and it may be impossible to satisfy inequality (2.27) in the hard case, because then it is usual for $1-\Delta/\|\underline{d}(\lambda)\|$, $\lambda>\lambda^*$, to be bounded away from zero. Indeed, as $\lambda\rightarrow\lambda^*$, the vector $\underline{d}(\lambda)$ tends to be in the column space of $B+\lambda^* I$, but equation (2.1) and $\|\underline{d}^*\|=\Delta$ may require part of $\underline{d}^*$ to be in the null space of $B+\lambda^* I$. It follows from the construction $\underline{d}=\underline{d}(\lambda)+\underline{z}$, introduced just before expression (2.21), that it is suitable to try to pick $\underline{z}$ from this null space. Therefore, because of the factorization $B+\lambda I=LL^T$, we seek $\underline{z}$ such that $\|L^T\underline{z}\|/\|\underline{z}\|$ is small. Inequality (2.21) confirms that this aim should assist a choice of $\underline{d}$ that satisfies the required condition (1.3). Further, we expect L^T to be nearly singular when an additional choice of $\underline{z}$ is necessary for good efficiency.

Moré and Sorensen (1983) suggest letting the $\underline{z}$ of the previous paragraph be a multiple of $(B+\lambda I)^{-1}\underline{\widehat{e}}=(LL^T)^{-1}\underline{\widehat{e}}$, where each component of $\underline{\widehat{e}}$ has modulus one. Thus $\underline{z}=\gamma\underline{\widehat{z}}$ is set for some $\gamma\in\mathcal{R}$, where $L\underline{w}=\underline{\widehat{e}}$ and $L^T\underline{\widehat{z}}=\underline{w}$. The signs of the components of $\underline{\widehat{e}}$ are chosen by setting $\widehat{e}_1=1$ initially. Then, during the forwards substitution calculation of $\underline{w}$ from $L\underline{w}=\underline{\widehat{e}}$, the index i runs through the values $2,3,\ldots,n$, and, for each i, the sign of $\widehat{e}_i$ is chosen to maximize $|w_i|$. Further, γ is the root of the quadratic equation $\|\underline{d}(\lambda)+\gamma\underline{\widehat{z}}\|^2=\Delta^2$ that has the smaller modulus, the roots being real because $\phi(\lambda)<0$ implies $\|\underline{d}(\lambda)\|<\Delta$. If this choice of $\underline{z}$ satisfies inequality (2.21), then the calculation is terminated, because condition (1.3) is achieved by $\underline{d}=\underline{d}(\lambda)+\gamma\underline{\widehat{z}}$. Otherwise, a new value of λ will be chosen in the way that has been described already.

Moré and Sorensen also make use of the remark that, when $\underline{\widehat{z}}$ is generated by the above method, which happens only when $B+\lambda_L I$ is not positive definite, then the Rayleigh quotient

$$\underline{\widehat{z}}^T (B+\lambda I)\, \underline{\widehat{z}} \,/\, \|\underline{\widehat{z}}\|^2 = \|L^T \underline{\widehat{z}}\|^2 / \,\|\underline{\widehat{z}}\|^2 \tag{2.28}$$

may be highly suitable for updating λ_L. Indeed, the construction of $\underline{\widehat{z}}$ should cause expression (2.28) to be very close to the least eigenvalue of $B+\lambda I$. Therefore, by analogy with equation (2.11), the value of λ_L is increased to $\lambda - \|L^T \underline{\widehat{z}}\|^2 / \|\underline{\widehat{z}}\|^2$, if this number is larger than the old value of λ_L. Thus it becomes likely that $B+\lambda I$ will be positive definite for all future trial values of λ.

The efficiency of the method can be measured by the number of trial values of λ that occur before termination. Moré and Sorensen (1983) report that on average this number is only 1.6 for the range of experiments that they tried.

3. A band matrix technique for quasi-Newton algorithms

The range of applications of the Moré and Sorensen (1983) method is restricted by the need to calculate the Cholesky factorization $B+\lambda I = LL^T$ for several values of λ. Indeed, if B is a general symmetric matrix without any sparsity, then each factorization requires $\mathcal{O}(n^3)$ operations. Fortunately, however, we find in this section that, in an important special case, the amount of work can be reduced substantially by making an orthogonal change of variables in $\mathcal{R}^n$. Specifically, the band matrix technique of Powell (1997) for quasi-Newton algorithms is described. Each iteration employs an orthogonal matrix Ω such that the product

$$M = \Omega^T B\, \Omega \tag{3.1}$$

is an $n \times n$ matrix of band-width $2s+1$, say, where s is much less than n for large n. We introduce Ω into the equations (2.5) by writing them in the form

$$\Omega^T (B+\lambda I)\, (\Omega\, \Omega^T)\, \underline{d}(\lambda) = -\Omega^T \underline{g}, \tag{3.2}$$

which is equivalent to the system

$$(M+\lambda I)\, \underline{\widehat{d}}(\lambda) = -\underline{\widehat{g}}, \tag{3.3}$$

where $\underline{\widehat{d}}(\lambda) = \Omega^T \underline{d}(\lambda)$ and $\underline{\widehat{g}} = \Omega^T \underline{g}$. Further, because the orthogonality of Ω gives $\|\underline{\widehat{d}}(\lambda)\| = \|\underline{d}(\lambda\|$, the function (2.6) can be defined by the equation

$$\phi(\lambda) = \Delta^{-1} - \|\underline{\widehat{d}}(\lambda)\|^{-1}, \qquad \lambda \in \mathcal{L}^-. \tag{3.4}$$

It follows that we may replace $\underline{g}$, B and $\underline{d}(\lambda)$ by $\underline{\widehat{g}}$, M and $\underline{\widehat{d}}(\lambda)$, respectively, throughout the trust region method of Section 2, and otherwise the algorithm is as before. The strong advantage of this reformulation is that, due to the band structure of M, the Cholesky factorization of $M+\lambda I$ for each λ takes only $\mathcal{O}(ns^2)$

operations. In particular, by choosing $s \sim n^{2/3}$, the complexity of the factorization becomes $\mathcal{O}(n^{7/3})$. This magnitude of s is mentioned, because it will be shown that it allows all the routine calculations of a quasi-Newton algorithm for unconstrained optimization to be done in $\mathcal{O}(n^{7/3})$ operations, including the updating of Ω and M when the matrix B is revised by a quasi-Newton formula.

The admissible quasi-Newton formulae are those that alter B by a matrix of rank one or two at the end of an iteration. For example, the BFGS formula overwrites B by the matrix

$$B^* = B - \frac{B\,\underline{\delta}\,\underline{\delta}^T B}{\underline{\delta}^T B\,\underline{\delta}} + \frac{\underline{\gamma}\,\underline{\gamma}^T}{\underline{\delta}^T \underline{\gamma}}, \tag{3.5}$$

in order to satisfy $B^*\underline{\delta} = \underline{\gamma}$, where $\underline{\delta}$ is the change to the vector of variables that is made by the current iteration, and where $\underline{\gamma}$ is the corresponding change in $\underline{\nabla} F$. We regard a rank two change to B as the sum of two rank one changes, however, so we address the updating formula

$$B^* = B + \phi\, \underline{r}\, \underline{r}^T, \tag{3.6}$$

where B is a general $n \times n$ symmetric matrix, and where $\phi \in \mathcal{R}$ and $\underline{r} \in \mathcal{R}^n$ are given. Assuming that Ω and M are available and have the properties that have been mentioned, we seek an orthogonal matrix Ω^* and an $n \times n$ symmetric matrix M^* of band-width $2s+1$ that satisfy the equation

$$M^* = \Omega^{*T} B^*\, \Omega^*. \tag{3.7}$$

The given algorithm that generates Ω^* and B^* does not require B explicitly.

Equations (3.1) and (3.6) provide the relation

$$B^* = \Omega\,(M + \phi\, \widehat{\underline{r}}\, \widehat{\underline{r}}^T)\, \Omega^T = \Omega\, W \Omega^T, \tag{3.8}$$

say, where $\widehat{\underline{r}}$ is the vector $\Omega^T \underline{r}$, which can be calculated easily. The algorithm overwrites Ω and W by $\Omega\, G_i^T$ and $G_i W G_i^T$, respectively, for $i = 1, 2, \ldots, m$, where each G_i is a Givens rotation with the property that its four elements that depend on the angle of rotation are adjacent. Further, we say that G_i is of type ℓ, where ℓ is the integer from $[1, n-1]$ such that these four elements have the indices (ℓ, ℓ), $(\ell, \ell+1)$, $(\ell+1, \ell)$ and $(\ell+1, \ell+1)$. The rotations are chosen so that the final W has band width $2s+1$. Thus W and Ω become M^* and Ω^* as required. The work of the overwriting for each i is $\mathcal{O}(n)$, because Ω has no sparsity in general. Therefore the main task of this section is to show that the number of rotations, namely m, is of magnitude $n^{4/3}$ when $s \sim n^{2/3}$.

The first half of the calculation will be introduced by considering the form

$$W = \begin{pmatrix}
\times & \times & \times & \cdot & \odot & \odot & \odot & \odot & \odot & \odot \\
\times & \times & \times & \times & \cdot & \odot & \odot & \odot & \odot & \odot \\
\times & \times & \times & \times & \times & \cdot & \odot & \odot & \odot & \odot \\
\cdot & \times & \times & \times & \times & \times & \cdot & \boxdot & \boxdot & \boxdot \\
\odot & \cdot & \times & \times & \times & \times & \times & \cdot & \boxdot & \boxdot \\
\odot & \odot & \cdot & \times & \times & \times & \times & \times & \cdot & \boxdot \\
\odot & \odot & \odot & \cdot & \times & \times & \times & \times & \times & \cdot \\
\odot & \odot & \odot & \boxdot & \cdot & \times & \times & \times & \times & \times \\
\odot & \odot & \odot & \boxdot & \boxdot & \cdot & \times & \times & \times & \times \\
\odot & \odot & \odot & \boxdot & \boxdot & \boxdot & \cdot & \times & \times & \times
\end{pmatrix}, \tag{3.9}$$

which is taken from Powell (1997). Here n and s have the values $n=10$ and $s=2$, so the '$\times$' elements of the expression denote the nonzero band of M. Therefore all the other elements belong to $\phi\,\widehat{\underline{r}}\,\widehat{\underline{r}}^T$ initially. In this case we let G_i be of type $(n{-}i)$ for $i=1,2,\ldots,6$, and we choose these rotations so that the last six elements of $G_6G_5\cdots G_1\widehat{\underline{r}}$ are zero. Thus all the elements in the '$\odot$' and '$\boxdot$' positions of the lower triangular part of expression (3.9) would be zero if this expression were the product $G_6G_5\cdots G_1(M{+}\phi\,\widehat{\underline{r}}\,\widehat{\underline{r}}^T)$. When the similarity transformation is completed, however, by post-multiplication by $G_1^TG_2^T\cdots G_6^T$, then usually the '$\boxdot$' elements in the lower triangular part of expression (3.9) become nonzero, but the zeros in the '$\odot$' positions are not disturbed. It follows from symmetry that, after the first six rotations, W has zeros in all the '$\odot$' positions that are shown.

The zeros that have been mentioned, including those that are made nonzero by the completion of the similarity transformation, are due to the fact that initially the '$\cdot$', '$\odot$' and '$\boxdot$' elements of W are the elements of a symmetric rank one matrix. We consider the more general case when these elements of W below the diagonal are those of the matrix $\phi\,\widehat{\underline{r}}\,\widehat{\underline{s}}^T$, while the corresponding elements of W above the diagonal are taken from $\phi\,\widehat{\underline{s}}\,\widehat{\underline{r}}^T$, where ϕ and $\widehat{\underline{r}}$ are as before, but $\widehat{\underline{s}}$ is any vector in $\mathcal{R}^n$. The symmetry of W is preserved, and again we pick the first six Givens rotations so that the last six elements of $G_6G_5\cdots G_1\widehat{\underline{r}}$ are zero. It follows from the remarks of the previous paragraph that these rotations cause W to acquire zeros in all the '$\odot$' positions of expression (3.9), although $\widehat{\underline{s}}$ is arbitrary. This generalization is useful because, after the first six rotations, the '$\cdot$' and '$\boxdot$' elements below the diagonal of the bottom right $7{\times}7$ submatrix of expression (3.9) are the elements of a rank one matrix, due to the way in which the '$\boxdot$' elements become nonzero when the similarity transformation is completed. Therefore we can introduce some zeros into this $7{\times}7$ submatrix, by applying a further sequence of Givens similarity transformations in a way that is analogous to the method of the previous paragraph. Three rotations occur in the new sequence, namely G_7, G_8 and G_9, and they are of types $n{-}1$, $n{-}2$ and $n{-}3$, respectively. They annihilate the '⊡' elements of expression (3.9), by mixing the last four rows and columns of W. Therefore the '⊙' elements of expression (3.9) remain at zero.

These remarks suggest the following procedure for generating most of the required zero elements of W. It assumes that W has the following two properties. Firstly, there exists a nonnegative integer k such that, when an element of W is in one of the first k rows or columns of W and is outside the diagonal band of width $2s{+}3$, then the element is zero. Secondly, the elements of the bottom right $(n{-}k)\times(n{-}k)$ submatrix of W, that are below the diagonal band of width $2s{+}1$, are the elements of a matrix of rank one, say $\phi\,\underline{\widehat{r}}\,\underline{\widehat{s}}^T$, which extends our use of the notation $\underline{\widehat{r}}$. Then, if $k<n-s-2$, we set $\widehat{m}=n-s-k-2$, and we apply $\widehat{m}$ Givens rotations to $\underline{\widehat{r}}$ so that the last $\widehat{m}$ components of $\underline{\widehat{r}}$ become zero, the i-th Givens rotation being of type $n{-}i$ for $i=1,2,\ldots,\widehat{m}$. Further, W is pre-multiplied by these rotations and is post-multiplied by their transposes, in order that the change to W is an orthogonal similarity transformation. It follows, as at the end of the previous paragraph, that the new W inherits the two properties that have been mentioned for a larger value of k, the new value being the sum of the old value and the number $\min[\widehat{m},s{+}1]$. For example, the integers $n=10$, $s=2$ and $\widehat{m}=6$ occur in the calculation of the paragraph that includes expression (3.9), and the value of k is increased from 0 to 3. Then the further rotations at the end of the previous paragraph are an application of the procedure with $\widehat{m}=3$ that increases k from 3 to 6. In the general case, the procedure is employed recursively, beginning with $k=0$, until the value $k=n-s-2$ is achieved. Thus W becomes a matrix of band-width $2s+3$. The number of applications of the procedure is of magnitude $n/(s{+}1)$, and on average an application employs about $\frac{1}{2}n$ Givens rotations. Hence the total number of rotations is $\mathcal{O}(n^2/(s{+}1))$, which gives the required $n^{4/3}$ order when $s\sim n^{2/3}$.

It is shown by Powell (1997) that the total number of rotations so far can be reduced by about a factor of two. Indeed, $\widehat{m}$ has the value $n-s-2$ when $k=0$, and we consider choosing the first $n-s-2$ rotations so that the vector

$$G_{n-s-2}\,G_{n-s-1}\cdots G_1\,\underline{\widehat{r}} \tag{3.10}$$

still has $n-s-2$ zero components by design, but we make the first p and last $n-s-p-2$ components zero, where p is the integer part of $\frac{1}{2}(n-s-2)$. This can be done by letting G_i be of type $n-i$ for $i\in[1,q]$, and of type $i-q$ for $i\in[q+1,q+p]$, where $q=n-s-p-2$. Thus, after the first $p+q$ similarity transformations, the nonzero elements of W are confined to the diagonal band of width $2s+3$, to the leading $(p{+}1)\times(p{+}1)$ submatrix of W, and to the bottom right $(q{+}1)\times(q{+}1)$ submatrix of W. The bottom right submatrix is transformed to band-width $2s+3$ by applying the procedure that has been described about $q/(s{+}1)$ times. Further, an analogous procedure is applied about $p/(s{+}1)$ times to the leading $(p{+}1)\times(p{+}1)$ submatrix. Thus the band-width of W becomes $2s+3$, using about the same number of applications of the procedure as before.

The savings occur because now the average number of Givens rotations of an application of the procedure is about $\frac{1}{4}n$, but before it was about $\frac{1}{2}n$.

The other half of the calculation of Ω^* and M^* has to reduce the band-width of W from $2s$+3 to $2s$+1, which is also achieved by a sequence of Givens rotations. Each Givens similarity transformation causes a particular subdiagonal element of W to become zero. Further, if the element is W_{ij}, where $i > j$, then the Givens rotation is of type $i-1$. The rotations have the property that, whenever W does not have band-width $2s$+3, then there is only one nonzero subdiagonal element of W outside this band. Therefore the choice of rotations can be defined as follows. If the band-width of W is greater than $2s+3$, then the next rotation annihilates the nonzero element that has just been mentioned. If the band-width of W is $2s$+1 as required, then the procedure is terminated. Otherwise, we let W_{ij} be the first nonzero element of W below the diagonal that is not in the band of width $2s+1$, and we pick the rotation that annihilates W_{ij}.

For example, we consider this construction when $n=10$ and $s=2$, and the rotations G_i, $i=1,2,\ldots,9$, have caused just the '$\odot$' and '$\boxdot$' elements of the matrix (3.9) to be zero. Then G_{10} is the rotation of type 3 that annihilates $W_{4,1}$. Thus, when W is overwritten by $G_{10} W G_{10}^T$, the third column of the new matrix is a linear combination of the third and fourth columns of $G_{10} W$. It follows from $W_{7,4} \neq 0$ that $W_{7,3}$ becomes nonzero too, and $W_{7,3}$ is outside the diagonal band of width $2s+3$. Therefore G_{11} is the rotation of type 6 that restores this element to zero, but then $W_{10,6}$ becomes nonzero. Hence G_{12} is of type 9, and it causes the band-width of W to be $2s+1$ again. Furthermore, the new W has the property $W_{4,1} = W_{1,4} = 0$. Therefore G_{13} is the rotation of type 4 that provides $W_{5,2} = 0$, but $W_{8,4}$ becomes nonzero, so G_{14} is the type 7 rotation that causes W to have bandwidth 7 again. Subsequently, in accordance with the previous paragraph, the calculation is completed by seven more Givens rotations, their types being 5, 8, 6, 9, 7, 8 and 9, respectively.

For general s and n, this procedure can be divided into stages, where, for $j = 1,2,\ldots,n-s-1$, the j-th stage provides $W_{j+s+1,j} = W_{j,j+s+1} = 0$, and it corrects any increases in the band-width of W. Furthermore, at the end of the j-th stage, the leading $(j+s+1)\times(j+s+1)$ submatrix of W has band-width $2s$+1, while the band-width of the remainder of W is $2s$+3. Therefore termination occurs at $j = n-s-1$. The first similarity transformation of the j-th stage annihilates $W_{j+s+1,j}$, but it causes $W_{j+2s+2,j+s}$ to become nonzero if $j+2s+2 \leq n$. Then, for $k \geq 2$, the k-th rotation of the j-th stage annihilates $W_{i,i-s-2}$, where $i = j+k(s+1)$, but the new $W_{i+s+1,i-1}$ is nonzero if $j+(k+1)(s+1) \leq n$. Thus the number of rotations of the j-th stage is the integer part of $(n-j)/(s+1)$. It follows that the total number of Givens rotations of this procedure is at most the sum

$$\sum_{j=1}^{n-s-1} \frac{n-j}{s+1} = \frac{(n+s)\,(n-s-1)}{2\,(s+1)} < \frac{n^2}{2\,(s+1)}. \tag{3.11}$$

n	s	band	full	ratio
50	20	0.07	0.10	1.43
100	30	0.35	0.74	2.11
200	40	1.51	5.90	3.91
400	60	7.60	48.42	6.37
800	80	36.62	397.48	10.85

Table 1: A comparison of the "band" and "full" calculations

We see that the magnitude of this expression is $n^{4/3}$ when $s \sim n^{2/3}$. Therefore it has been shown that the required calculation of Ω^* and M^* from Ω and M can be done in $\mathcal{O}(n^{7/3})$ computer operations.

It follows that, if each iteration of a quasi-Newton method revises B by a rank one or rank two formula, if B has no useful sparsity, and if each trial change to the variables is calculated by the Moré and Sorensen (1983) trust region procedure, then, by transforming each B to a matrix of band-width about $n^{2/3}$, the amount of computation can be reduced by a factor of about $c\,n^{2/3}$ for some constant c. The usefulness of this conclusion depends on c, so this constant was estimated by some numerical experiments. Specifically, the time of a "band" calculation was compared with the time of a "full" calculation for several values of n, all times being in seconds on a Sun 4 workstation. In the "full" case, the system (2.5) was solved for two values of λ such that $B+\lambda I$ is positive definite, assuming no sparsity in B, and formula (3.6) was applied once. On the other hand, the "band" calculation solves the system (3.3) twice, and it generates Ω^* and M^* from Ω and M by the method of this section, for a suitable choice of s. The values of n and s, the "band" times, the "full" times and the full/band ratios are given in Table 1, which is copied from Powell (1997). We see that the use of band matrices has reduced the amount of computation when $n \geq 40$, and that substantial savings occur for $n \geq 100$. Further, we find that c is approximately 0.12.

4. Discussion

The importance of the given work to practical applications is debatable, partly because current methods of theoretical analysis have many inadequacies. Indeed, nearly all of the theory of algorithms for unconstrained optimization is confined to the following three questions, where $\underline{x}_{k+1}$ is the vector of variables that is calculated by the k-th iteration of the algorithm. (1) If the number of iterations is infinite and if the sequence of points $\underline{x}_{k+1}$, $k=1,2,3,\ldots$, remains in a bounded region of $\mathcal{R}^n$, does the sequence have a limit point that is a stationary point of the objective function? (2) If the infinite sequence of calculated points converges, what is the rate of convergence? (3) Is the algorithm efficient when the objective function is quadratic? A disadvantage of question (1) is that the answer is unknown for several algorithms that perform well in practice, but the answer is affirmative for some very slow methods. The main criticism of question (2) is that, in a typical calculation by a good algorithm, one sees a fast rate of convergence eventually, but it is relevant to less than one tenth of the iterations. An objection to question (3) is that the techniques of unconstrained optimization are intended for objective functions that are not quadratic. Therefore we take the view that one should try to develop algorithms that provide favourable answers to the questions, but that the questions do not go far enough.

Actually, question (3) was vital to the development of variable metric algorithms, which provided major advances over steepest descents in the 1960s. One expects steady progress for smooth objective functions if quadratics are treated well, and if the changes to the variables are so small that third order terms of Taylor series expansions are negligible, but it is difficult to predict the magnitudes of the changes to the variables that will occur. From this point of view the results of numerical experiments were very encouraging. Therefore it is sensible to employ quadratic models, such as expression (1.1), in ways that are not damaged unnecessarily by third and higher order terms of F. Moreover, in trust region methods one should try to avoid small values of the radius Δ when the variables are far from a minimum, because the number of iterations has to be large if small steps are taken in this case.

Now, when the matrix B of the quadratic model (1.1) is the exact second derivative matrix $\nabla^2 F(\underline{\widehat{x}})$, the error $|Q(\underline{d}) - F(\underline{\widehat{x}}+\underline{d})|$ is typically of magnitude $\|\underline{d}\|^3$ when $\|\underline{d}\|$ is small enough for the model to be useful. We consider the possibility that the model is the function (1.4) and the error is just $\|\underline{d}\|^3$. Then the trial step $\underline{d}$ provides the required reduction $F(\underline{\widehat{x}}+\underline{d}) < F(\underline{\widehat{x}})$ if the inequality

$$Q(0) - Q(\underline{d}) > \|\underline{d}\|^3 \tag{4.1}$$

holds, but otherwise it is likely that the iteration will decrease the trust region radius instead of adjusting the current vector of variables. It follows that, if $\underline{d}$ is a multiple of the first coordinate direction, as in the truncated conjugate gradient method, then the new value of Δ may have to be less than the positive root of the equation $10^{-6}(\Delta+\Delta^2) = \Delta^3$, which implies $\Delta < 0.001$, approximately. On the other hand, condition (4.1) is always achieved if $\Delta < 1$ and if $\underline{d}$ is a multiple of the second coordinate direction. These remarks provide a good reason for preferring the method of Moré and Sorensen (1983) when second derivatives are available.

In the other situation, when the matrix B of the quadratic model is revised on each iteration by a quasi-Newton algorithm, then the case for the Moré and Sorensen method is different. It is certainly advantageous that the technique of Section 3 allows the routine work of each iteration to be done in only $\mathcal{O}(n^{7/3})$

computer operations, which compares favourably with the truncated conjugate gradient method if B is full and if that method take more than $n^{1/3}$ conjugate gradient iterations to generate the trial step $\underline{d}$. Moreover, the remarks of the previous paragraph are important if B can have some negative eigenvalues. In many quasi-Newton algorithms, however, every matrix B is positive definite, but most of those algorithms prevent increases in the value of F by applying line searches instead of trust regions. Then the quadratic model (1.1) provides a change to the variables that is a positive multiple of the direction $-B^{-1}\underline{g}$, and it is well-known that the routine work of an iteration can be restricted to $\mathcal{O}(n^2)$. On the other hand, when a trust region method is employed for convex quadratic models, then some recent results of Yuan (1997) suggest that the bad behaviour of example (1.4) cannot occur. Specifically, guided by several numerical experiments and by some analysis in a special case, he conjectures that the trial step $\underline{d}$ of the truncated conjugate gradient method satisfies condition (1.3) with $\varepsilon = \frac{1}{2}$ if B is positive definite.

The average amount of work per iteration of the procedure of Section 3 can be reduced from $\mathcal{O}(n^{7/3})$ to $\mathcal{O}(n^{11/5})$ by storing the orthogonal matrix Ω in product form (Powell, 1997). In order to explain this remark, we recall the notation (3.1), where M is an $n\times n$ symmetric matrix of band-width $2s+1$. At the beginning of each iteration, both B and the elements in the band of M are available as before, but now Ω is expressed as the product

$$\Omega = \Omega_0\,\Gamma_1^T\Gamma_2^T\cdots\Gamma_\mu^T, \tag{4.2}$$

for some integer μ, where Ω_0 is an orthogonal matrix that is retained for several iterations. Further, the matrices Γ_i, $i=1,2,\ldots,\mu$, are the Givens rotations, stored in a condensed form, that occur in the technique of Section 3, so μ is zero at the beginning of the first iteration. Expression (4.2) implies that the calculations of the vectors $\underline{\widehat{g}} = \Omega^T\underline{g}$ and $\underline{d}(\lambda) = \Omega\,\underline{\widehat{d}}(\lambda)$ from $\underline{g}$ and $\underline{\widehat{d}}(\lambda)$ now require $\mathcal{O}(n^2+\mu)$ operations, but the work of solving the system (3.3) for each λ is still $\mathcal{O}(ns^2)$.

When applying the updating formula (3.6), the new procedure also overwrites W by $G_iWG_i^T$ for $i=1,2,\ldots,m$, which provides M^* explicitly. The overwriting takes only $\mathcal{O}(n^2)$ operations, because m is of magnitude n^2/s, and because we take advantage of the strucure of W. Moreover, instead of replacing Ω by $\Omega\,G_i^T$, the new procedure adds G_i, $i=1,2,\ldots,m$, to the end of the list Γ_i, $i=1,2,\ldots,\mu$, so G_i becomes $\Gamma_{\mu+i}$ for each i, and then μ is increased by m. Another change is that B is actually updated by formula (3.6). Hence the total work of the iteration is $\mathcal{O}(n^2+\mu+ns^2)$, which includes solving the system (3.3) for a few values of λ. The reason for retaining B is that Ω_0 is revised and μ is reset to zero every t iterations, say, in order that μ does not become too large. Thus the average value of μ is $\mathcal{O}(n^2t/s)$. The matrix M is also revised when Ω_0 is changed, because the new matrices are defined by a direct calculation that gives the factorization $B=\Omega_0M\,\Omega_0^T$ for the current B, subject to the orthogonality of Ω_0 and the symmetric band-structure of M. This calculation requires $\mathcal{O}(n^3)$ operations every t iterations. Thus the average work per iteration is $\mathcal{O}(n^2+n^2t/s+ns^2+n^3/t)$. The best balance between the terms in brackets occurs for $s\sim n^{3/5}$ and $t\sim n^{4/5}$, and then the average work is only $\mathcal{O}(n^{11/5})$, as claimed at the beginning of the previous paragraph. Further details are given by Powell (1997).

We have found that the iterations of quasi-Newton trust region algorithms need not be much more expensive than the iterations of line search algorithms when each trial step is calculated by the method of Moré and Sorensen (1983). It is hoped that this conclusion will be valuable to the future development of software for nonlinear optimization calculations.

Acknowledgement

I am very grateful to Jorge Moré for several helpful suggestions on the first draft of Sections 1 and 2.

References

R. Fletcher and C.M. Reeves (1963), "Function minimization by conjugate gradients", *Comput J.*, Vol. 7, pp. 149–154.

R. Fletcher (1987), *Practical Methods of Optimization*, John Wiley & Sons (Chichester).

N.I.M. Gould (1997), "Solving the trust region subproblem using the Lanczos method", presented at the 17th Biennial Conference on Numerical Analysis (Dundee).

J.J. Moré and D.C. Sorensen (1983), "Computing a trust region step", *SIAM J. Sci. Stat. Comput.*, Vol. 4, pp. 553–572.

M.J.D. Powell (1997), "The use of band matrices for second derivative approximations in trust region algorithms", Report No. DAMTP 1997/NA12, University of Cambridge.

F. Rendl and H. Wolkowicz (1997), "A semidefinite framework for trust region subproblems with applications to large scale minimization", *Math. Programming*, Vol. 77, pp. 273–299.

D.C. Sorensen (1997), "Minimization of a large-scale quadratic function subject to a spherical constraint", *SIAM J. Optimization*, Vol. 7, pp. 141–161.

T. Steihaug (1983), "The conjugate gradient method and trust regions in large scale optimization", *SIAM J. Numer. Anal.*, Vol. 20, pp. 626–637.

Y. Yuan (1997), "Some properties of a trust region subproblem", presented at the 16th International Symposium on Mathematical Programming (Lausanne).