DAMTP 2010/NA04

Beyond symmetric Broyden for updating quadratic models in minimization without derivatives

M.J.D. Powell

Abstract: Some highly successful algorithms for unconstrained minimization without derivatives construct changes to the variables by applying trust region methods to quadratic approximations to the objective function $F(\underline{x})$, $\underline{x} \in \mathcal{R}^n$. A quadratic model has $(n+1)(n+2)/2$ independent parameters, but each new model may interpolate only $2n+1$ values of F, for instance. The symmetric Broyden method takes up the remaining freedom by minimizing the Frobenius norm of the difference between the second derivative matrices of the old and new models, which usually works well in practice. We consider an extension of this technique that combines changes in first derivatives with changes in second derivatives. A simple example suggests that the extension does bring some advantages, but numerical experiments on three test problems with up to 320 variables are disappointing. On the other hand, rates of convergence are investigated numerically when F is a homogeneous quadratic function, which allows very high accuracy to be achieved in practice, the initial and final errors in the variables being about 10 and 10^{-5000}, respectively. It is clear in some of these experiments that the extension does reduce the number of iterations. The main difficulty in the work was finding a way of implementing the extension sufficiently accurately in only $\mathcal{O}(n^2)$ operations on each iteration. A version of the truncated conjugate gradient procedure is suitable, that is used in the numerical experiments, and that is described in detail in an appendix.

Keywords: Minimization without derivatives; Quadratic models; Symmetric Broyden; Truncated conjugate gradients.

Department of Applied Mathematics and Theoretical Physics,
Centre for Mathematical Sciences,
Wilberforce Road,
Cambridge CB3 0WA,
England.

March, 2010 (revised April, 2011).

1. Introduction

The symmetric Broyden method provides a very useful technique for updating second derivative matrices of quadratic models in iterative algorithms for unconstrained minimization when first derivatives of the objective function are available. At the beginning of the k-th iteration, the model has the form

$$Q_k(\underline{x}) \;=\; F(\underline{x}_k) + (\underline{x}-\underline{x}_k)^T \underline{\nabla} F(\underline{x}_k) + \tfrac{1}{2}\,(\underline{x}-\underline{x}_k)^T B_k\,(\underline{x}-\underline{x}_k), \quad \underline{x}\in\mathcal{R}^n, \qquad (1.1)$$

where $F(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, is the objective function, where $\underline{x}_k$ is the current best vector of variables, which means usually that $F(\underline{x}_k)$ is the least calculated value of $F(\underline{x})$ so far, and where B_k is an $n\times n$ symmetric matrix, chosen before the start of the k-th iteration. Termination may occur if $\|\underline{\nabla} F(\underline{x}_k)\|$ is sufficiently small, but otherwise the iteration generates a new vector of variables, $\underline{x}_k+\underline{d}_k$ say, the construction of the nonzero step $\underline{d}_k$ being guided by the approximation $Q_k(\underline{x}) \approx F(\underline{x})$, $\underline{x}\in\mathcal{R}^n$. If the strict reduction $F(\underline{x}_k+\underline{d}_k) < F(\underline{x}_k)$ is achieved, then $\underline{x}_{k+1}$ is set to $\underline{x}_k+\underline{d}_k$, and the quadratic model for the next iteration is expression (1.1) with k increased by one. The symmetric Broyden method includes a formula that defines the new matrix B_{k+1}.

That formula is derived from two considerations. Firstly, when F is twice differentiable, it has the property

$$\Big\{ \int_{\alpha=0}^{1} \nabla^2 F(\underline{x}_k+\alpha\,\underline{d}_k)\,d\alpha \Big\}\,\underline{d}_k \;=\; \underline{\nabla} F(\underline{x}_k+\underline{d}_k) - \underline{\nabla} F(\underline{x}_k), \qquad (1.2)$$

so the constraint

$$B_{k+1}\,\underline{d}_k \;=\; \underline{\nabla} F(\underline{x}_{k+1}) - \underline{\nabla} F(\underline{x}_k) \qquad (1.3)$$

supplies B_{k+1} with some true second derivative information from the objective function. Secondly, changes to models should not be unnecessarily large, in order to avoid both instabilities and the replacement of good models by less accurate ones. Specifically, B_{k+1} is set to a symmetric matrix that satisfies equation (1.3), and all the remaining freedom in B_{k+1} is taken up by minimizing the Frobenius norm $\|B_{k+1}-B_k\|_F$, the Frobenius norm of a real matrix being the square root of the sum of squares of its elements. It is well known that the difference $B_{k+1}-B_k$ is a matrix of rank two, and that B_{k+1} can be calculated from B_k in only $\mathcal{O}(n^2)$ operations, as shown in equation (3.6.5) of Fletcher (1987), for instance.

Let $\mathcal{Q}$ be the linear space of polynomials of degree at most two in n variables. We regard $\mathcal{Q}$ as an inner product space by equipping it with the semi-norm

$$\|Q\|_\theta \;=\; \Big\{ \|\nabla^2 Q\|_F^2 + 2\theta\,\|\underline{\nabla} Q(\underline{v})\|_2^2 \Big\}^{1/2}, \qquad Q\in\mathcal{Q}, \qquad (1.4)$$

where θ and $\underline{v}$ for the moment are a nonnegative constant and a fixed point in $\mathcal{R}^n$, respectively. The title of our work begins "Beyond symmetric Broyden", because we study the idea of letting θ be positive, which extends the usual approach of taking up the freedom in Q_{k+1} by minimizing $\|Q_{k+1}-Q_k\|_\theta$ with $\theta=0$, after satisfying the constraint (1.3) and the equations

$$Q_{k+1}(\underline{x}_{k+1}) \;=\; F(\underline{x}_{k+1}) \qquad \text{and} \qquad \underline{\nabla} Q_{k+1}(\underline{x}_{k+1}) \;=\; \underline{\nabla} F(\underline{x}_{k+1}). \qquad (1.5)$$

From now on, however, our attention is given to algorithms when first derivatives of F are not available. Then, instead of the constraints (1.3) and (1.5), we let the conditions on Q_{k+1} take the form

$$Q_{k+1}(\underline{y}_j^+) \;=\; F(\underline{y}_j^+), \qquad j=1,2,\ldots,m, \tag{1.6}$$

where the set of points $\{\underline{y}_j^+ \in \mathcal{R}^n : j=1,2,\ldots,m\}$ includes $\underline{x}_{k+1}$ and is revised automatically as the iterations proceed. The number m is a prescribed constant, the choice $m=2n+1$ being mentioned in the abstract.

We combine the situations with and without first derivatives by defining $\mathcal{A}$ to be the affine set in the linear space $\mathcal{Q}$, such that $Q\in\mathcal{Q}$ satisfies the constraints on Q_{k+1} if and only if Q is in $\mathcal{A}$. Thus, when calculating Q_{k+1} from Q_k, we require Q_{k+1} to be the quadratic $Q\in\mathcal{Q}$ that solves the subproblem

$$\text{Minimize} \quad \|Q-Q_k\|_\theta \quad \text{subject to} \quad Q\in\mathcal{A}. \tag{1.7}$$

In other words, Q_{k+1} is the closest point to Q_k in the affine set $\mathcal{A}$, distance being measured by the semi-norm $\|\cdot\|_\theta$. It follows from the basic theory of inner product spaces and least squares projection that Q_{k+1} has the property

$$\|Q_{k+1}-Q\|_\theta^2 \;=\; \|Q_k-Q\|_\theta^2 - \|Q_{k+1}-Q_k\|_\theta^2, \qquad Q\in\mathcal{A}. \tag{1.8}$$

The case when F is a quadratic function is of interest. Then F is an element of the affine set $\mathcal{A}$, because it must satisfy the conditions on Q_{k+1} that are taken from itself. Therefore the property (1.8) gives the equation

$$\|Q_{k+1}-F\|_\theta^2 \;=\; \|Q_k-F\|_\theta^2 - \|Q_{k+1}-Q_k\|_\theta^2, \qquad k=1,2,3,\ldots\,. \tag{1.9}$$

We see that the errors $\|Q_k-F\|_\theta$, $k=1,2,3,\ldots$, decrease monotonically and that $\|Q_{k+1}-Q_k\|_\theta$ tends to zero as $k\to\infty$. It follows from the definition (1.4) that, if $\theta>0$ and $\underline{v}\in\mathcal{R}^n$ are fixed, then the changes to the quadratic models achieve both of the conditions

$$\lim_{k\to\infty}\|\nabla^2 Q_{k+1}-\nabla^2 Q_k\|_F=0 \quad \text{and} \quad \lim_{k\to\infty}\|\underline{\nabla} Q_{k+1}(\underline{v})-\underline{\nabla} Q_k(\underline{v})\|_2=0. \tag{1.10}$$

If θ were zero, however, then equation (1.9) would yield only the first of these limits. Therefore positive values of θ may be very helpful to proofs of convergence.

The book by Conn, Scheinberg and Vicente (2009) includes much careful work on optimization without derivatives, with some convergence theory of algorithms that employ quadratic models. That analysis requires the derivatives of the models to be sufficiently accurate approximations to derivatives of the objective function, and, when necessary, there are some extra evaluations of F in order to achieve these conditions. The remarks in the last paragraph suggest, however, that a version of the symmetric Broyden method may provide some useful theory without any extra calculations of values of F. Further attention is given to this possibility at the beginning of Section 5.

An advantage of $\theta > 0$ over $\theta = 0$ is shown by a simple algebraic example in Section 2, the number of variables and the number of interpolation conditions (1.6) being only $n = 2$ and $m = 4$, respectively. Choices of θ and $\underline{v}$ for the norm (1.4) are the subject of Section 3, with an explanation that constant values of these parameters are not suitable in practice.

Section 4 compares $\theta > 0$ with $\theta = 0$ by presenting numerical results for three different forms of F that allow large numbers of variables, the values of n being 20, 40, 80, 160 and 320. Many features need careful attention in such calculations without derivatives, including the changes that are made to the variables by the iterative procedure, the positions of the interpolation points of expression (1.6), the adjustments of trust region radii, and the stability of updating procedures that provide economically some required factorizations and inverses of matrices. Answers to these needs have been included already in the development of the BOBYQA Fortran software (Powell, 2009), so our numerical results were obtained by an extended version of BOBYQA, where the extension allows θ to be positive in the subproblem (1.7) that provides the new quadratic model Q_{k+1}, the choices of $\theta > 0$ and $\underline{v}$ being taken from Section 3. The subproblem is not solved exactly, because in experiments with hundreds of variables it is very welcome if the amount of computation for each k is only $\mathcal{O}(m^2 + n^2)$. Instead a conjugate gradient procedure is truncated when enough attention seems to have been given to the contribution from θ to the norm (1.4), a full description of this construction of Q_{k+1} being supplied in Appendix A.

Section 5 provides numerically some insight into whether or not positive values of θ may help the theoretical convergence properties of algorithms for minimization without derivatives. The objective functions F of that section are homogeneous quadratics, which means that they have the property

$$F(\lambda\,\underline{x}) \;=\; \lambda^2\,F(\underline{x}), \qquad \lambda\in\mathcal{R}, \qquad \underline{x}\in\mathcal{R}^n, \tag{1.11}$$

and every second derivative matrix $\nabla^2 F$ is positive definite. Therefore the calculated vectors $\underline{x}_k$, $k = 1, 2, 3, \ldots$, should converge to the zero vector, and the speed of convergence can be observed even if the limited precision of computer arithmetic causes substantial relative errors in every $\underline{x}_k$. Furthermore, the property (1.11) implies that, if all calculated function values and all calculated vectors of variables are scaled by λ^2 and λ, respectively, at any time during the iterative procedure, where λ is any positive constant, then the resultant changes to the later iterations are by the same scaling factors. Thus the current problem is replaced by an equivalent one occasionally in a way that avoids computer underflows. In terms of the original scaling, this technique allows the initial and final values of $\|\underline{x}_k\|$ in the experiments of Section 5 to be of magnitude 10 and 10^{-5000}, respectively, which is sufficient to expose situations where positive values of θ are beneficial to the achievement of very high accuracy.

2. An algebraic example

Only one new value of the objective function F is calculated on each iteration of the BOBYQA algorithm. Let $F(\underline{x}_k+\underline{d}_k)=F(\underline{y}_t^+)$, say, be the new value of the k-th iteration, the reason for the notation $\underline{y}_t^+=\underline{x}_k+\underline{d}_k$ being that the new value is always included in the interpolation conditions (1.6) on Q_{k+1}. It follows that the other $m-1$ function values $F(\underline{y}_j^+)$, $j\neq t$, were available when Q_k was chosen, and BOBYQA provides the property

$$Q_k(\underline{y}_j^+)=F(\underline{y}_j^+), \qquad j\in\{1,2,\ldots,m\}\backslash\{t\}. \tag{2.1}$$

Hence the difference $Q_{k+1}(\underline{x})-Q_k(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, is a quadratic that vanishes at the points $\underline{x}=\underline{y}_j^+$, $j\neq t$, which allows Q_{k+1} to be written in the form

$$Q_{k+1}(\underline{x}) \;=\; Q_k(\underline{x})+\{F(\underline{y}_t^+)-Q_k(\underline{y}_t^+)\}\,\Lambda_t(\underline{x}), \qquad \underline{x}\in\mathcal{R}^n, \tag{2.2}$$

for some quadratic function Λ_t that satisfies $\Lambda_t(\underline{y}_j^+)=\delta_{jt}$, $j=1,2,\ldots,m$, where δ_{jt} is the Kronecker delta. We recall from Section 1 that Q_{k+1} is derived from the projection (1.7), which is equivalent to the construction of $\Lambda_t\in\mathcal{Q}$ from the subproblem

$$\text{Minimize} \quad \|\Lambda_t\|_\theta \quad \text{subject to} \quad \Lambda_t(\underline{y}_j^+)=\delta_{jt}, \qquad j=1,2,\ldots,m. \tag{2.3}$$

The example of this section addresses the dependence on θ of the solution Λ_t of subproblem (2.3) in a case with $n=2$, $m=4$ and $\underline{v}=0$ in the definition (1.4). We pick the interpolation points

$$\underline{y}_1^+=\begin{pmatrix}0\\0\end{pmatrix}, \quad \underline{y}_2^+=\begin{pmatrix}1\\0\end{pmatrix}, \quad \underline{y}_3^+=\begin{pmatrix}0\\M\end{pmatrix} \quad \text{and} \quad \underline{y}_4^+=\begin{pmatrix}1/2\\\eta/M\end{pmatrix}, \tag{2.4}$$

M being large and positive and η being of magnitude one, and we pick $t=4$. These values are possible if the initial interpolation points are about distance M apart, one of them being at the origin, and if the first three iterations of BOBYQA replace the far interpolation points by new ones that are within distance one from the origin. Then the situation (2.4) would occur during the second iteration, $\underline{y}_2^+$ being the new point of the first iteration, $\underline{y}_4^+=\underline{y}_t^+$ being the new point of the second iteration, and $\underline{y}_3^+$ being the initial interpolation point that is going to be replaced by the third iteration.

The positions (2.4) with $t=4$ imply that $\Lambda_t\in\mathcal{Q}$ satisfies $\Lambda_t(\underline{y}_j^+)=0$, $j\neq t$, if and only if it has the form

$$\Lambda_t(\underline{x}) \;=\; \Lambda_t(x,y) \;=\; p\,x\,(1-x)+q\,y\,(M-y)+r\,x\,y, \qquad \underline{x}\in\mathcal{R}^2, \tag{2.5}$$

for some real multipliers p, q and r. Therefore the calculation (2.3) is equivalent to finding the values of p, q and r that minimize $\|\Lambda_t\|_\theta$ subject to the constraint

$$\Lambda_t(\underline{y}_4^+) \;=\; p/4+\eta\,(1-\eta/M^2)\,q+\eta\,r/(2M) \;=\; 1. \tag{2.6}$$

In the case $\underline{v}=0$, equations (1.4) and (2.5) provide the expression

$$\tfrac{1}{4}\,\|\Lambda_t\|_\theta^2 \;=\; \tfrac{1}{2}\,(2+\theta)\,p^2+\tfrac{1}{2}\,(2+\theta M^2)\,q^2+\tfrac{1}{2}\,r^2. \tag{2.7}$$

It follows from first order conditions for optimality that we require p, q and r to be such that the gradient of expression (2.7) with respect to them is a multiple of the gradient of the constraint function (2.6). Thus the subproblem (2.3) has the solution

$$\Lambda_t(\underline{x}) = \chi^{-1}\left\{\frac{1}{8+4\,\theta}\,x\,(1-x)+\frac{\eta(1-\eta/M^2)}{2+\theta M^2}\,y\,(M-y)+\frac{\eta}{2M}\,x\,y\right\},\quad \underline{x}\in\mathcal{R}^2, \tag{2.8}$$

for some denominator $\chi\in\mathcal{R}$, which takes the value

$$\chi \;=\; \frac{1}{32+16\,\theta}+\frac{\eta^2(1-\eta/M^2)^2}{2+\theta M^2}+\frac{\eta^2}{4M^2}, \tag{2.9}$$

in order to satisfy the constraint $\Lambda_t(\underline{y}_t^+)=1$.

When θ is zero, which holds in the construction of Q_{k+1} by the unextended version of BOBYQA, equations (2.8) and (2.9) imply that Λ_t is approximately the function

$$\Lambda_t(\underline{x}) \approx \{\tfrac{1}{32}+\tfrac{1}{2}\eta^2\}^{-1}\,\{\,\tfrac{1}{8}\,x\,(1-x)+\tfrac{1}{2}\,\eta\,y\,(M-y)+\tfrac{1}{2}\,\eta\,M^{-1}\,x\,y\,\},\quad \underline{x}\in\mathcal{R}^2, \tag{2.10}$$

the approximation being the removal of the η/M^2 and η^2/M^2 terms because they are much less than one. We find that expression (2.10) gives the values

$$\|\nabla^2\Lambda_t\|_F^2 \approx 4\,(\eta^2+\tfrac{1}{16})^{-1}\quad\text{and}\quad \|\underline{\nabla}\Lambda_t(0)\|_2^2 \approx (\eta^2M^2+\tfrac{1}{16})\,(\eta^2+\tfrac{1}{16})^{-2}, \tag{2.11}$$

where again an η^2/M^2 term has been dropped. The fact that $\|\underline{\nabla}\Lambda_t(0)\|_2$ is of magnitude M is highly unwelcome, especially because the quadratic

$$\Lambda(\underline{x}) \;=\; 4\,x\,(1-x),\qquad \underline{x}\in\mathcal{R}^2, \tag{2.12}$$

also satisfies the constraints $\Lambda(\underline{y}_j^+)=\delta_{jt}$, $j=1,2,\ldots,m$, of subproblem (2.3), and it is not troubled by large values of M.

We wish to update the quadratic model in a way that avoids changes to first and second derivatives that are much larger than necessary near $\underline{x}_{k+1}$. In particular, in the example of this section where M is large, we would prefer to employ the function (2.12) instead of the function (2.8) in the case $\theta=0$. The use of a positive value of θ in the norm (1.4) is intended to provide such a replacement automatically, because then the construction of Λ_t gives attention to both first and second derivatives.

We compare Λ_t with Λ when the magnitude of θ is one, which is assumed to mean that both M/θ and $M\theta$ are of magnitude M. Then the denominator (2.9) has the property

$$\chi^{-1} \;=\; 4\,(8+4\,\theta)+\mathcal{O}(M^{-2}). \tag{2.13}$$

It follows from equations (2.8) and (2.12) that the multiple of $x(1-x)$ in $\Lambda_t(\underline{x})$ is an excellent approximation to $\Lambda(\underline{x})$, $\underline{x}\in\mathcal{R}^2$. Further, if $\|\underline{x}\|$ is $\mathcal{O}(1)$, then the other terms of $\Lambda_t(\underline{x})$ in equation (2.8) have magnitude M^{-1} or less. Therefore, in the present setting, the subproblem (2.3) supplies a quadratic Λ_t that is suitable for the updating formula (2.2), provided that θ is $\mathcal{O}(1)$, without any need to give careful attention to the actual value of θ. Similarly, estimates of magnitudes are employed in the technique of the next section, that defines θ and $\underline{v}$ for the subproblem (2.3) on every iteration of the extended version of BOBYQA.

The value $\theta=\infty$ is also of interest. Then, because of the definition (1.4), the subproblem (2.3) becomes the minimization of $\|\nabla^2\Lambda_t\|_F^2$ subject to the constraints

$$\Lambda_t(\underline{y}_j^+)=\delta_{jt}, \qquad j=1,2,\ldots,m, \qquad \text{and} \qquad \underline{\nabla}\Lambda_t(\underline{v})=0. \tag{2.14}$$

It is possible that these constraints cannot be satisfied by any quadratic Λ_t when the subproblem (2.3) has a solution for finite θ, this situation being usual in the case $m=(n+1)(n+2)/2$. In the example of this section, however, the choice $\theta=\infty$ in expressions (2.8) and (2.9) gives the function

$$\Lambda_t(\underline{x}) \;=\; (2M/\eta)\,x\,y, \qquad \underline{x}\in\mathcal{R}^2, \tag{2.15}$$

that has the properties (2.14), but $\|\nabla^2\Lambda_t\|_F^2$ is unacceptably large. Instead we want θ to be reasonably small.

3. The parameters of the semi-norm

It is suggested in Section 1 that the parameters θ and $\underline{v}$ of the semi-norm (1.4) be fixed with θ positive, in order to provide the theoretical limits (1.10) when F itself is quadratic, but this suggestion is unsuitable in practice. Our explanation of this assertion begins by asking whether a good choice of θ in expression (1.4) remains good if the variables $\underline{x}\in\mathcal{R}^n$, are scaled by the factor $\sigma>0$, say. Then the general quadratic $Q(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, becomes $Q^+(\underline{x}^+)=Q(\sigma^{-1}\underline{x}^+)$, $\underline{x}^+\in\mathcal{R}^n$, where $\underline{x}^+=\sigma\underline{x}$ is the new vector of variables. Further, because first and second derivatives of Q^+ at $\underline{x}^+=\sigma\underline{x}$ are the same as first and second derivatives of Q at $\underline{x}=\sigma^{-1}\underline{x}^+$ multiplied by σ^{-1} and σ^{-2}, respectively, the semi-norm (1.4) has the property

$$\Big\{\|\nabla^2 Q\|_F^2+2\theta\,\|\underline{\nabla}Q(\underline{v})\|_2^2\Big\}^{1/2} \;=\; \sigma^2\Big\{\|\nabla^2 Q^+\|_F^2+2\theta^+\,\|\underline{\nabla}Q^+(\sigma\underline{v})\|_2^2\Big\}^{1/2}, \tag{3.1}$$

where $\theta^+=\sigma^{-2}\theta$. Thus, if θ provides a good balance between the first and second derivative terms on the left hand side, then $\theta^+=\sigma^{-2}\theta$ provides a good balance on the right hand side.

The objective function F is irrelevant to the present argument, because the purpose of the semi-norm is to take up the freedom in the quadratic Λ_t that is constructed from the subproblem (2.3). In particular, the right hand sides δ_{jt} of the constraints on Λ_t do not depend on F. Further, the main difference between

the positions of the interpolation points $\underline{y}_j^+$, $j=1,2\ldots,m$, on the early and late iterations of BOBYQA is that, as the calculation proceeds, the points become much closer together, perhaps by the factor $\sigma=10^{-6}$ when six decimal places of accuracy are required in the final vector of variables. This factor would cause first and second derivatives of Λ_t to increase by factors of about 10^6 and 10^{12}, respectively, as mentioned already, and such factors can also be deduced directly from the conditions $\Lambda_t(\underline{y}_j^+)=\delta_{jt}$, $j=1,2\ldots,m$. In this setting a good balance in the semi-norm (1.4) requires θ to be increased by about the factor 10^{12}, as shown in equation (3.1), so we abandon the idea that θ be a constant.

Keeping $\underline{v}$ fixed would be even more disastrous in practice, unless its position is very close to the final vector of variables, but this condition is unacceptable in the usual practical situation when the solution to the optimization calculation is not known in advance. One can pick $\underline{v}$ within the initial cluster of interpolation points. When the points come closer together, however, it would be usual for $\underline{v}$ to become outside the convex hull of the cluster, and then eventually all the distances $\|\underline{v}-\underline{y}_j^+\|_2$, $j=1,2,\ldots,m$, may be bounded below by a positive constant. Moreover, if the algorithm were working as intended, then the magnitudes of the terms $\|\nabla^2\Lambda_t\|_F$, θ and $\|\underline{\nabla}\Lambda_t(\underline{v})\|_2$ of $\|\Lambda_t\|_\theta$ would be $\mathcal{O}(\Delta^{-2})$, $\mathcal{O}(\Delta^{-2})$ and $\mathcal{O}(\Delta^{-1})$, respectively, where Δ is about the distance between the current interpolation points. A positive lower bound on $\|\underline{v}-\underline{x}_{k+1}\|$ with the identity

$$\underline{\nabla}\Lambda_t(\underline{x}_{k+1}) \;=\; \underline{\nabla}\Lambda_t(\underline{v}) + \nabla^2\Lambda_t\,(\underline{x}_{k+1}-\underline{v}), \tag{3.2}$$

however, would imply that the magnitude of the term $\|\underline{\nabla}\Lambda_t(\underline{x}_{k+1})\|_2$ is likely to be $\mathcal{O}(\Delta^{-2})$ instead of $\mathcal{O}(\Delta^{-1})$ as required. Therefore we want the ratio $\|\underline{v}-\underline{x}_{k+1}\|_2/\Delta$ to be not much larger than one throughout the calculation. The extension to BOBYQA achieves this condition by setting $\underline{v}=\underline{x}_{k+1}$ on every iteration, the vector $\underline{x}_{k+1}$ being available before the updating formula (2.2) is applied, as stated in Section 1.

The choice of θ for the subproblem (2.3) on the k-th iteration of the extension to BOBYQA is $\theta=\eta_k/(2\delta_k)$, where η_k and δ_k are estimates of the magnitudes of $\|\nabla^2\Lambda_t\|_F^2$ and $\|\underline{\nabla}\Lambda_t(\underline{v})\|_2^2$, respectively, the actual values of η_k and δ_k being given below. Thus we try to balance the two terms of the semi-norm (1.4). Because distances between interpolation points on the first iteration are about $\rho_{\rm beg}$, where $\rho_{\rm beg}$ is the initial trust region radius supplied by the user of BOBYQA, and because Λ_t has to satisfy the constraints of subproblem (2.3), the values

$$\eta_1=\rho_{\rm beg}^{-4} \qquad \text{and} \qquad \delta_1=\rho_{\rm beg}^{-2} \tag{3.3}$$

are picked for the first iteration.

Both η_{k+1} and δ_{k+1} are set after the construction of Λ_k on the k-th iteration, partly because the actual value of $\|\underline{\nabla}\Lambda_t(\underline{x}_{k+1})\|_2^2=\|\underline{\nabla}\Lambda_t(\underline{v})\|_2^2$ gives an indication of its magnitude. On the other hand, if $m=4$ and $n=2$ occur in the subproblem (2.3), and if we pick the points

$$\underline{y}_1^+=\begin{pmatrix}0\\0\end{pmatrix},\quad \underline{y}_2^+=\begin{pmatrix}1\\-1\end{pmatrix},\quad \underline{y}_3^+=\begin{pmatrix}1\\1\end{pmatrix}\quad\text{and}\quad \underline{y}_4^+=\begin{pmatrix}2\\0\end{pmatrix}, \tag{3.4}$$

with $t=4$, $\theta=0$ and $\underline{v}=\underline{x}_{k+1}=\underline{x}_k=\underline{y}_1^+$, then straightforward calculation shows that the subproblem has the solution

$$\Lambda_t(\underline{x}) \;=\; \tfrac{1}{4}\,(x^2-y^2), \qquad \underline{x}=(x,y)\in\mathcal{R}^2. \tag{3.5}$$

The data (3.4) were chosen so that the function (3.5) has the property $\underline{\nabla}\Lambda_t(\underline{v})=0$. It follows that this function is also the solution of subproblem (2.3) for any $\theta>0$. Further, the obvious choice $\delta_{k+1}=\|\underline{\nabla}\Lambda_t(\underline{v})\|_2^2$ would cause $\theta=\eta_{k+1}/(2\delta_{k+1})$ to be infinite on the next iteration, which would introduce the disadvantages that are mentioned at the end of Section 2. We respond to this possibility by imposing the lower bound Δ_k^{-2} on δ_{k+1}, where Δ_k is the current trust region radius, this bound being a reasonable estimate of $\|\underline{\nabla}\Lambda_t(\underline{v})\|_2^2$ on the next iteration, due to the constraints $\Lambda_t(\underline{y}_j^+)=\delta_{jt}$, $j=1,2,\ldots,m$. We also try to take advantage of information from previous iterations. Specifically, the value of δ_{k+1} in the extended version of BOBYQA is given by the formula

$$\delta_{k+1} \;=\; \max\left[\,0.7\,\delta_k,\,\Delta_k^{-2},\,\|\underline{\nabla}\Lambda_t(\underline{x}_{k+1})\|_2^2\,\right], \qquad k=1,2,3,\ldots\;. \tag{3.6}$$

The term $0.7\delta_k$ occurs instead of δ_k on the right hand side, in order to admit the possibility $\delta_{k+1}<\delta_k$, which is often helpful if there are several increases in the trust region radius.

It may be adequate to pick $\eta_{k+1}=\|\nabla^2\Lambda_t\|_F^2$ after Λ_t is constructed on the k-th iteration, but, instead of being guided by Λ_t, we prefer to give attention to $\|\nabla^2\Lambda_t^{(0)}\|_F^2$, where $\Lambda_t^{(0)}(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, is defined to be the solution to subproblem (2.3) in the case $\theta=0$, which has the property $\|\nabla^2\Lambda_t^{(0)}\|_F^2\le\|\nabla^2\Lambda_t\|_F^2$. The main reason for this preference is that, if θ became unsuitably large, there would be a tendency for $\|\nabla^2\Lambda_t\|_F^2$ to be unsuitably large too, and then $\eta_{k+1}=\|\nabla^2\Lambda_t\|_F^2$ with $\theta=\eta_{k+1}/(2\delta_{k+1})$ would cause the large value of θ to be inherited by the next iteration. Instead the choice $\eta_{k+1}=\|\nabla^2\Lambda_t^{(0)}\|_F^2$ with formula (3.6) gives an upper bound on the next value of θ that is independent of the current value of θ, this bound being the product $\|\nabla^2\Lambda_t^{(0)}\|_F^2\,\Delta_k^2$. It is prudent to let θ be too small in cases of doubt, because of the good performance in practice of the unmodified version of BOBYQA, which corresponds to setting $\theta=0$ on every iteration. Furthermore, we also wish to take advantage of information from previous iterations, without forcing the sequence η_k, $k=1,2,3,\ldots$, to be monotonic. Therefore the extended version of BOBYQA employs the formula

$$\eta_{k+1} \;=\; \max\left[\,0.7\,\eta_k,\,\|\nabla^2\Lambda_t^{(0)}\|_F^2\,\right], \qquad k=1,2,3,\ldots\;. \tag{3.7}$$

The specification of the values $\theta=\eta_k/(2\delta_k)$, $k=1,2,3,\ldots$, is complete.

4. Numerical results

The data that are required to run the extended version of BOBYQA with $\theta>0$ are the same as the data for the unmodified BOBYQA Fortran software that has

been sent by e-mail to about 200 people. The objective function is specified by a subroutine, provided by the user, that calculates $F(\underline{x})$ for any $\underline{x}$ in $\mathcal{R}^n$ that satisfies the bounds $\underline{a} \le \underline{x} \le \underline{b}$, where $\underline{a} \in \mathcal{R}^n$ and $\underline{b} \in \mathcal{R}^n$ are also given by the user. The components of $\underline{a}$ and $\underline{b}$ are set to -10^{60} and 10^{60}, respectively, in all our numerical experiments, which makes the bounds irrelevant. The user also supplies the number of interpolation conditions m, introduced in equation (1.6), and we are going to compare $m = n+6$ with $m = 2n+1$. A starting point $\underline{x}_0 \in \mathcal{R}^n$, say, has to be given too, and the first quadratic model satisfies $Q_1(\underline{y}_j) = F(\underline{y}_j)$, $j = 1, 2, \ldots, m$, where $\underline{x}_0$ is one of the points $\underline{y}_j$, the other points being positioned nearby automatically (Powell, 2009). Finally, the initial and final trust region radii, namely $\rho_{\rm beg}$ and $\rho_{\rm end}$, are required, the choices $\rho_{\rm beg} = 0.5$ or 0.1 and $\rho_{\rm end} = 10^{-6}$ being made for the numerical results of this section.

We recall that Δ_k is the trust region radius of the k-th iteration. It satisfies $\Delta_k \ge \rho_k$, where the sequence ρ_k, $k = 1, 2, 3, \ldots$, decreases monotonically from $\rho_{\rm beg}$ to $\rho_{\rm end}$. The reduction $\rho_{k+1} < \rho_k$ is made only when it seems that the bound $\Delta_k \ge \rho_k$ is preventing further progress, a typical reduction being $\rho_{k+1} = 0.1\,\rho_k$. Therefore it is usual for ρ_k to remain fixed for many consecutive iterations, during which well-known techniques are employed for updating Δ_k subject to $\Delta_k \ge \rho_k$. Termination occurs when the criteria for reducing ρ_k are achieved but ρ_k has reached its final value $\rho_{\rm end}$. Therefore we expect the error in the final vector of variables to be of magnitude $\rho_{\rm end}$, and this parameter provides a crude control of accuracy.

Four versions of the extension to BOBYQA are compared for each test problem of this section, these versions being given by combining $m = n+6$ or $m = 2n+1$ with $\theta = 0$ or $\theta > 0$. Every $\theta > 0$ is taken from the previous section, the value of θ being changed automatically from iteration to iteration. When θ is zero, however, the unmodified BOBYQA software could have been used, but instead all updating of quadratic models is done by the procedure in Appendix A, so all our numerical results are new. The three different forms of F, mentioned in Section 1, are given below. For each of them, and for each n from $\{20, 40, 80, 160, 320\}$, we pick five test problems by making five different choices of random numbers. Thus there are 75 different test problems altogether for our experiments with the four versions of BOBYQA.

The first form of F has been used very much by the author, in particular in the development of the original BOBYQA software. The objective function is the sum of squares

$$F(\underline{x}) \;=\; \sum_{i=1}^{2n} \Big\{ c_i - \sum_{j=1}^{n} \left[S_{ij} \sin(x_j/\sigma_j) + C_{ij} \cos(x_j/\sigma_j) \right] \Big\}^2, \qquad \underline{x} \in \mathcal{R}^n, \qquad (4.1)$$

the parameters S_{ij} and C_{ij} being independent random integers from $[-100, 100]$, and the divisors σ_j being random constants from the continuous logarithmic distribution on $[1, 10]$. Then the components of $\underline{c} \in \mathcal{R}^{2n}$ are defined by $F(\underline{x}_*) = 0$, after choosing $\underline{x}_*$ randomly from the uniform distribution on $[-\pi, \pi]^n$. The starting

n	$m=n+6$ $\theta=0$	$m=n+6$ $\theta>0$	$m=2n+1$ $\theta=0$	$m=2n+1$ $\theta>0$
20	1419—1584	1354—2223	786—919	991—1230
40	3622—4445	4442—5440	1692—1990	2201—3535
80	8549—9528	9048—12081	3186—3510	5535—6746
160	20169—21901	22021—39455	5733—6483	11228—19128
320	46923—51964	50379—75588	11064—12034	21723—35208

Table 1: Ranges of $\#F$ when F is the sum of squares (4.1)

vector $\underline{x}_0$ is picked by letting the weighted differences $[\underline{x}_0-\underline{x}_*]_j/\sigma_j$, $j=1,2,\ldots,n$, be random numbers from $[-\pi/10,\pi/10]$, where $[\underline{x}_0-\underline{x}_*]_j$ is the j-th component of $\underline{x}_0-\underline{x}_*$. The values of $\rho_{\rm beg}$ and $\rho_{\rm end}$ are set to 0.1 and 10^{-6}, respectively.

Table 1 presents some of the numerical results when F has the form (4.1). There is a row in the table for each n, and the four main columns are provided by the four versions of the extension to BOBYQA. We recall there are five test problems for each row, generated by different choices of random numbers. Entries of the form a—b in the table show the least and greatest of the five values of $\#F$ that occurred, using the version of BOBYQA that belongs to the column, where $\#F$ is the total number of calculations of $F(\underline{x})$ when a version of BOBYQA is applied to a test problem. We see that, in most of the experiments of Table 1, it is hardly ever advantageous to prefer $\theta>0$ instead of $\theta=0$, which is disappointing.

Next we let the objective function have the form

$$F(\underline{x}) \;=\; \textstyle\sum_{j=1}^{n-1}\,\{(x_j^2+x_n^2)^2-4\,x_j+3\}, \qquad \underline{x}\in\mathcal{R}^n, \tag{4.2}$$

taken from the Appendix of Conn *et al* (1994), and having the name ARWHEAD because of the sparsity structure of $\nabla^2 F$. We employ the usual starting point of this example, namely $\underline{x}_0=\underline{e}\in\mathcal{R}^n$, $\underline{e}$ being the vector of ones, and we set $\rho_{\rm beg}=0.5$ and $\rho_{\rm end}=10^{-6}$. These data define the first test problem for every n, but there are no random numbers yet for supplying four more test problems. Therefore we investigate whether changes to the order of the variables make much difference in practice, by working with the five objective functions $F_i(\underline{x})=F(P_i\,\underline{x})$, $\underline{x}\in\mathcal{R}^n$, $i=1,2,3,4,5$, where the definition (4.2) is retained, where P_1 is the $n\times n$ unit matrix, and where P_i, $i=2,3,4,5$, are $n\times n$ permutation matrices, generated randomly for each n. The least value of $F_i(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, occurs at the point $\underline{x}_*$ that satisfies $P_i\,\underline{x}_*=\underline{e}-\underline{e}_n$, where $\underline{e}_n$ is the n-th coordinate vector in $\mathcal{R}^n$, but the reorderings of the variables do not disturb the starting point $\underline{x}_0=\underline{e}$.

Table 2 gives some results of the ARWHEAD calculations, all details of the format being the same as those of Table 1. We find again that most of the $\theta>0$ values of $\#F$ are greater than the $\theta=0$ values. It is interesting that several of the a—b entries in the table have the property $b>1.5a$, because now all the differences between the five test problems for each n are due only to the ordering

n	$m=n+6$ $\theta=0$	$m=n+6$ $\theta>0$	$m=2n+1$ $\theta=0$	$m=2n+1$ $\theta>0$
20	305—616	367—576	738—801	678—765
40	822—1401	929—1569	1758—2201	2504—3083
80	1665—1915	1623—1989	5795—7016	6702—7530
160	3543—4004	3658—9492	11898—12970	15963—20382
320	8295—8992	12462—13624	8327—30130	33491—44068

Table 2: Ranges of $\#F$ when F is the ARWHEAD function (4.2)

of the variables. In particular, the five values of $\#F$ that provide the $n = 40$ entry in the first column are 822, 845, 852, 870 and 1401, so only one of the calculations is much more inefficient than the rest, which suggests some kind of instability. The author had hoped that such inefficiencies would be removed by making θ positive, but the $n = 40$ entry in the second column of Table 2 comes from the values 929, 932, 985, 1008 and 1569 of $\#F$, while the $n = 160$ entry in this column is due to the values 3658, 3757, 4697, 4958 and 9492, so it seems that any instability has survived. Furthermore, the very wide a—b range at the bottom of the third column is given by the numbers 8327, 12103, 28403, 29195 and 30130. Fortunately, the accuracy of the calculations is not impaired by these anomalies, the greatest value of $\|\underline{x}_{\rm fin}-\underline{x}_*\|_\infty$ throughout the experiments of Table 2 being 1.7×10^{-5}, where $\underline{x}_{\rm fin}$ and $\underline{x}_*$ are the final and optimal vectors of variables, respectively. Another interesting feature of Table 2, unlike Table 1, is that many of the $m=n+6$ values of $\#F$ are much smaller than the corresponding $m=2n+1$ values.

The third and final form of F for the experiments of this section is the "chained Rosenbrock" function

$$F(\underline{x}) \;=\; \textstyle\sum_{j=1}^{n-1} \{4\,(x_j - x_{j+1}^2)^2 + (1-x_{j+1})^2\}, \qquad \underline{x}\in\mathcal{R}^n, \tag{4.3}$$

which can be found in the Buckley (1989) collection of test problems. Again the least value of F is zero, the optimal vector of variables being $\underline{e} \in \mathcal{R}^n$. The usual starting point for minimization calculations is $\underline{x}_0 = -\underline{e} \in \mathcal{R}^n$, but then convergence to a local minimum may occur. Instead we let the components of $\underline{x}_0$ be independent random numbers from the logarithmic distribution on $[0.5, 2]$, and we let this randomness provide five different problems for each n. We pick $\rho_{\rm beg} = 0.1$ and $\rho_{\rm end} = 10^{-6}$. Some results of these calculations are presented in Table 3, using the same format as before. The entries in the $m=n+6$ columns for $n=160$ are unusual, because they suggest clearly that $\#F$ becomes smaller when θ becomes positive. Therefore ten more random starting points $\underline{x}_0$ were tried for these choices of m and n. The new values of $\#F$ are in the intervals $[9350, 11412]$ and $[7675, 8986]$ for $\theta = 0$ and $\theta > 0$, respectively, which confirms that positive values of θ are helpful in this case. The author is puzzled by this finding.

n	$m=n+6$ $\theta=0$	$m=n+6$ $\theta>0$	$m=2n+1$ $\theta=0$	$m=2n+1$ $\theta>0$
20	609—880	697—1664	672—808	747—970
40	1661—2222	2272—5763	1732—2164	2010—2623
80	4015—4876	4045—18547	3849—4362	3694—5406
160	9556—11557	7048—7994	8388—9280	7621—15275
320	22684—28862	16954—51500	10559—33477	19925—40106

Table 3: Ranges of $\#F$ for the chained Rosenbrock function (4.3)

There are more numerical results in the next section, but they are not relevant to practical computation, because they investigate rates of convergence in a setting that requires the optimal vector of variables to be at the origin, in order that very high accuracy can be achieved. Therefore we ask now whether or not positive values of θ in the semi-norm (1.4) are likely to provide more efficient software for general optimization calculations without derivatives. Most of the comparisons of values of $\#F$ in Tables 1–3 are not promising. Furthermore, it is disadvantageous that the procedure in Appendix A for the subproblem (2.3) requires much more work than the updating of quadratic models in the unextended version of BOBYQA. Good support for $\theta>0$, however, is given by the example of Section 2. Another consideration is that Tables 1 and 2 show that efficiency can depend strongly on the choice of m, which provides encouragement for the development of new techniques that choose and adjust m automatically. Therefore more research is expected, and it is too early to advance the view that going beyond symmetric Broyden will not be useful in practice.

5. On the speed of convergence

Algorithms for unconstrained optimization that employ quadratic models are expected to achieve fast convergence eventually if the models become sufficiently accurate approximations to the objective function within a suitable region of $\mathcal{R}^n$. Furthermore, Broyden, Dennis and Moré (1973) identify conditions on quasi-Newton methods that are sufficient for superlinear convergence, this work being a major breakthrough, because the errors in the approximations $B_k=\nabla^2 Q_k\approx\nabla^2 F(\underline{x}_k)$, $k=1,2,3,\ldots$, may remain bounded away from zero as $k\to\infty$, where the notation is taken from equation (1.1). Similarly, BOBYQA often completes its iterations successfully without calculating enough values of $F(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, for the construction of an accurate quadratic model. For example, in all five experiments of Table 1 with $n=320$, $m=2n+1$ and $\theta=0$, the difference between the final vector of variables $\underline{x}_{\rm fin}$ and the optimal vector $\underline{x}_*$ satisfies $\|\underline{x}_{\rm fin}-\underline{x}_*\|_\infty\le 1.5\times10^{-5}$, but a good quadratic model cannot be constructed from only 12034 values of F,

n	$\#F$	$\|\underline{x}_{\rm fin}-\underline{x}_*\|_\infty$	$\|E_1\|_F$	$\|E_{\rm fin}\|_F$
20	967.2	1.7×10^{-6}	116.3	57.2
40	2069.4	2.6×10^{-6}	161.9	114.7
80	4176.8	2.9×10^{-6}	226.3	191.8
160	7633.0	3.0×10^{-6}	317.9	294.3
320	13751.6	6.4×10^{-6}	448.3	433.9

Table 4: Some results when F is quadratic, $m=2n+1$ and $\rho_{\rm end}=10^{-6}$

as a quadratic in 320 variables has 51681 degrees of freedom.

The author ran some numerical experiments in 2009 that investigated the monotonically decreasing sequence $\|\nabla^2Q_k-\nabla^2F\|_F$, $k=1,2,3,\ldots$, when homogeneous quadratics are minimized by the unextended version of BOBYQA. The eigenvectors of ∇^2F, namely $\underline{v}_j$, $j=1,2,\ldots,n$, were generated randomly, $\underline{v}_1$ being from the uniform distribution on the surface of the unit ball $\{\underline{x}:\|\underline{x}\|_2=1\}$, and $\underline{v}_j$, $j=2,3,\ldots,n$, being from the uniform distribution on the set $\{\underline{x}:\|\underline{x}\|_2=1;\ \underline{x}^T\underline{v}_i=0,\ i=1,2,\ldots,j-1\}$, but the eigenvalues were in a geometric progression from 1 to 100, which provided the objective function

$$F(\underline{x})\ =\ \tfrac{1}{2}\,\underline{x}^T\Big\{\sum_{j=1}^n\ 100^{(j-1)/(n-1)}\,\underline{v}_j\,\underline{v}_j^T\Big\}\,\underline{x},\qquad \underline{x}\in\mathcal{R}^n. \tag{5.1}$$

The initial vector of variables $\underline{x}_0$ was also chosen randomly subject to $\|\underline{x}_0\|_2=1$, five different values of m were tried, and BOBYQA was given $\rho_{\rm beg}=0.1$ and $\rho_{\rm end}=10^{-6}$. Further, five sets of random numbers supplied five different test problems for each n, as in Section 4. Some results of these calculations with $m=2n+1$ are shown in Table 4, all the entries being averages over the five test problems of each row of the table. In the column headings, $\underline{x}_{\rm fin}$ is still the vector of variables that is returned by BOBYQA, the matrix E_k is defined to be the difference $\nabla^2Q_k-\nabla^2F$, and $E_{\rm fin}$ is the error of the approximation $\nabla^2Q_k\approx\nabla^2F$ on the last iteration.

We see in the table that $\#F$ grows no faster than linearly as $n\to\infty$, so again $\#F$ becomes too small to allow the construction of good quadratic models when n is large. Nevertheless, the author had not expected the entries in the last column of the table to be so close to the entries in the previous column. When $n=320$ and $\#F=13751$, there are about 13110 terms in the monotonically decreasing sequence $\|E_k\|_F$, $k=\ 1,2,\ldots,$fin, and the final term is about 96.8% of the first term. Hence, on average, the improvement to the approximation $\nabla^2Q_k\approx\nabla^2F$ by each update of the model is less than 0.00025%. Therefore, although BOBYQA achieves good accuracy in $\underline{x}_{\rm fin}$, it may be the world's worst procedure for estimating second derivatives of objective functions.

The function (5.1) is useful for investigating whether positive values of θ in the semi-norm (1.4) may be helpful to the rate of convergence of the sequence $\underline{x}_k$, $k=1,2,3,\ldots$, generated by the extended version of BOBYQA. We recall from

n	$m=n+6$ $\theta=0$	$m=n+6$ $\theta>0$	$m=2n+1$ $\theta=0$	$m=2n+1$ $\theta>0$
20	33.6—34.6	34.1—34.8	30.1*—35.4*	31.5—34.4
40	106.7—110.6	103.1—107.0	102.8—105.9	102.8—105.4
80	436.9—447.1	399.8—405.5	364.8—371.9	356.3—366.5
160	1819.9—1839.6	1537.6—1553.8	1180.5—1194.9	1159.3—1170.2

Table 5: Ranges of $10^{-3}\,\#F$ when F is quadratic and $\rho_{\rm end}=10^{-5000}$

the last paragraph of Section 1 that the homogeneity of expression (5.1) gives a way of avoiding underflows in computer arithmetic when $\rho_{\rm end}$ is tiny. Specifically, the values $\rho_{\rm beg}=1$ and $\rho_{\rm end}=10^{-5000}$ are assumed in the experiments below, and every initial vector $\underline{x}_0$ is chosen randomly from the set $\{\underline{x}:\|\underline{x}\|_2=10\}$. Because the reductions in the lower bound ρ_k on Δ_k are by factors of ten, as mentioned in the second paragraph of Section 4, we employ the identity

$$100^\sigma\,F(\underline{x})\;=\;F(10^\sigma\underline{x}),\qquad\underline{x}\in\mathcal{R}^n,\tag{5.2}$$

which is equivalent to equation (1.11), and we let σ take integer values in the range $[0,5000]$, beginning with $\sigma=0$ and ending with $\sigma=5000$ on the final iteration. Also, σ is increased by one whenever the reduction $\rho_{k+1}=0.1\rho_k$ occurs. It follows that, if $\underline{x}$ is of magnitude ρ_k for any iteration number k, then expression (5.2) is of magnitude one, so its value can be calculated without overflows or underflows in floating point computer arithmetic. Further, instead of storing such an $\underline{x}$, the values of σ and $10^\sigma\underline{x}$ are retained, which avoids underflows too. In particular, the vectors $10^\sigma\underline{y}_j^+$, $j=1,2,\ldots,m$, are stored with σ instead of the interpolation points of the conditions (1.6), and these vectors are multiplied by 10 whenever σ is increased by one. Overflows do not occur, because the replacement of one interpolation point on each iteration of BOBYQA tends to avoid automatically the retention of points that are relatively far from $\underline{x}_k$.

The four versions of BOBYQA that gave the numerical results of Section 4 were extended further to include the device in the previous paragraph, for minimizing a homogeneous quadratic function without underflow when $\rho_{\rm end}$ is tiny. We apply the extended software to the functions F that occur in the calculations of Table 4, except that the functions with $n=320$ are omitted, because otherwise the tiny value of $\rho_{\rm end}$ would cause the total amount of computation to be inconvenient. The choices $\rho_{\rm beg}=1$, $\rho_{\rm end}=10^{-5000}$ and $\|\underline{x}_0\|_2=10$ are employed, as mentioned already. Some results of these calculations are shown in Table 5, with $\#F$ scaled by 10^{-3}, but otherwise the format is taken from the tables of Section 4. We find now that, in most cases with $m=n+6$ and in all of the tests with $n=160$, going "beyond symmetric Broyden" does reduce the number of iterations.

The $n=20$ entry in the third column of Table 5 includes asterisks that denote failure to achieve the required accuracy in one of the five test problems, but all

n	$\|E_{\rm fin}\|_F$	Average $\#F$ for each $\rho\in[10^{-1500},10^{-501}]$	Average $\#F$ for each $\rho\in[10^{-5000},10^{-4001}]$
20	0.000032 / 0.000027	6.67 / 6.76	3.91 / 3.90
40	0.002107 / 0.001448	23.56 / 22.89	9.18 / 8.96
80	0.304445 / 0.230877	119.05 / 103.92	32.86 / 32.21
160	34.62499 / 26.13208	436.25 / 352.69	200.73 / 173.07

Table 6: On the rate of convergence of the Table 5 tests with $m=n+6$

the other runs finished successfully. In the bad case, an error return occurs after 26370 calculations of F, the final accuracy being $\|\underline{x}_{\rm fin}-\underline{x}_*\|_\infty=0.84\times10^{-3720}$, so this value of $\#F$ is excluded from the table. In all other cases of Table 5, the norms $\|\underline{x}_{\rm fin}-\underline{x}_*\|_\infty$ are satisfactory, the greatest of them being 1.64×10^{-5000}. A major difficulty in maintaining adequate accuracy is that some of the distances $\|\underline{y}_j^+-\underline{x}_{k+1}\|$, $j=1,2,\ldots,m$, from the interpolation points to the trust region centre, become much larger than Δ_k. For instance, in the experiment of Table 5 with $n=20$, $m=2n+1$ and $\theta=0$ that gives $10^{-3}\#F=30.1$, there are on average about 6 calculations of the objective function for every reduction in Δ_k by a factor of 10, so $m=41$ implies that on average Δ_k is decreased about 7 times during the retention of each new F. Thus some of the ratios $\|\underline{y}_j^+\|/\Delta_k$, $j=1,2,\ldots,m$, exceed 10^7. At the error return mentioned above, the greatest ratio is 1.06×10^9, and there are even larger ratios in some of the other calculations.

One purpose of the tests of Table 5 is to investigate the sequence of errors $E_k=\nabla^2Q_k-\nabla^2F$, $k=1,2,3,\ldots$, as k becomes large. We recall the property of the symmetric Broyden method that $\|E_k\|_F$ decreases monotonically as k increases when F is quadratic and θ is zero. The accuracy of our experiments is sufficient to demonstrate this property numerically with $\rho_{\rm end}=10^{-5000}$ and $n\geq20$, but, for $n=10$, the norms $\|E_k\|_F$ become of the magnitude of computer rounding errors in the difference $\nabla^2Q_k-\nabla^2F$. The final values of these norms provide the $\|E_{\rm fin}\|_F$ column of Table 6 for the $m=n+6$ calculations of Table 5, each entry being of the form $a\,/\,b$, where a and b are taken from the $\theta=0$ and $\theta>0$ tests, respectively. Both a and b are averages of the final values of $\|E_k\|_F$ for the five objective functions that are generated randomly for each n. We see that smaller values of $\|E_{\rm fin}\|_F$ are obtained by going beyond symmetric Broyden, in spite of the loss of monotonicity in the sequence $\|E_k\|_F$, $k=1,2,\ldots,$fin. It seems that $\|E_k\|_F$ tends to zero in exact arithmetic as $k\to\infty$, although, when n is 160, more than 10^6 iterations are required to reduce $\|E_k\|_F$ from its initial value by a factor of ten.

If our extended version of BOBYQA had superlinear convergence, then, in exact arithmetic, the number of reductions in the lower bound ρ on Δ_k by a factor of ten would be infinite, and the average number of calculations of the objective function for each ρ would tend to zero. This situation would be unsustainable

ℓ	$m=n+6$ $\theta=0$	$m=n+6$ $\theta>0$	$m=2n+1$ $\theta=0$	$m=2n+1$ $\theta>0$
1	222.4—228.2	198.9—200.9	157.1—159.9	161.2—166.0
2	308.4—317.5	277.2—279.8	229.8—234.2	231.6—238.4
3	363.2—373.5	328.9—332.9	283.5—289.1	280.4—289.0
4	404.0—414.3	367.9—373.0	326.8—333.5	320.0—329.7
5	436.9—447.1	399.8—405.5	364.8—371.9	356.3—366.5

Table 7: Ranges of $10^{-3}\,\#F$ to achieve $F(\underline{x}_{k+1})\leq 10^{-2000\ell}$ twice when $n=80$

in practice, however, because, as mentioned at the end of the paragraph before last, some of the ratios $\|\underline{y}_j^+\|/\Delta_k$, $j=1,2,\ldots,m$, would become infinite. Therefore BOBYQA includes the feature that usually a reduction in ρ is not allowed until at least three new values of F have been generated for the current ρ, which excludes superlinear convergence. This limiting behaviour is nearly achieved in practice by the $n=20$ experiments of Table 5, the average number of calculations of F for each new ρ in the interval $[10^{-5000},10^{-4001}]$ being less than four. The last column of Table 6 shows these averages for $m=n+6$ and our usual choices of n, the notation $a\,/\,b$ being employed as before to separate the $\theta=0$ results from the $\theta>0$ results. The averages when ρ is in the interval $[10^{-1500},10^{-501}]$ are also recorded. We note that the entries in the last column of Table 6 are much less than the corresponding entries in the previous column. The growth in the entries of the final column as n increases may be due to the magnitudes of $\|\nabla^2Q_k-\nabla^2F\|_F$, which have been addressed briefly already. All the figures in the last three rows of Table 6 suggest that $\theta>0$ is more efficient than $\theta=0$, but the reader should look again at the last paragraph of Section 4.

Table 7 was added to the original version of this paper, in order to show faster convergence as k increases in a way proposed by a referee. It goes beyond the $n=80$ row of Table 5 by providing values of $10^{-3}\#F$ when the algorithm achieves $F(\underline{x}_{k+1})\leq 10^{-2000\ell}$ for the second time for $\ell\in\{1,2,3,4,5\}$, except that each $\ell=5$ entry is replaced by the final value of $10^{-3}\#F$. The use of second time instead of first time avoids distortion of the results by iterations that cause $F(\underline{x}_{k+1})$ to be much smaller than usual for the current ρ. The differences between the rows of Table 7 demonstrate that, for all four versions of BOBYQA, the speed of convergence becomes faster as the iterations proceed, the gains being stronger when θ is positive.

Acknowledgements

Much of the work of this paper was done at the Liu Bie Ju Centre for Mathematical Sciences of the City University of Hong Kong. The author is very grateful for the excellent support, facilities and welcome that he enjoyed there from January to March, 2010. He is also very grateful to two referees for several suggestions that have improved the presentation.

References

C.G. Broyden, J.E. Dennis and J.J. Moré (1973), "On the local and superlinear convergence of quasi-Newton methods", *J. Inst. Math. Appl.*, Vol. 12, pp. 223–245.

A.G. Buckley (1989), "Test functions for unconstrained minimization", Technical Report 1989 CS-3, Dalhousie University, Canada.

A.R. Conn, N.I.M. Gould, M. Lescrenier and Ph.L. Toint (1994), "Performance of a multifrontal scheme for partially separable optimization", in *Advances in Optimization and Numerical Analysis*, eds. Susana Gomez and Jean-Pierre Hennart, Kluwer Academic (Dordrecht), pp. 79–96.

A.R. Conn, K. Scheinberg and L.N. Vicente (2009), *Introduction to Derivative-Free Optimization*, SIAM Publications (Philadelphia).

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

M.J.D. Powell (2004a), "Least Frobenius norm updating of quadratic models that satisfy interpolation conditions", *Math. Programming B*, Vol. 100, pp. 183–215.

M.J.D. Powell (2004b), "On updating the inverse of a KKT matrix", in *Numerical Linear Algebra and Optimization*, editor Y. Yuan, Science Press (Beijing), pp. 56–78.

M.J.D. Powell (2006), "The NEWUOA software for unconstrained optimization without derivatives", in *Large-Scale Optimization*, editors G. Di Pillo and M. Roma, Springer (New York), pp. 255–297.

M.J.D. Powell (2009), "The BOBYQA algorithm for bound constrained optimization without derivatives", Report No. DAMTP 2009/NA06, CMS, University of Cambridge.

Appendix A. The calculation of Λ_t

We recall from Sections 1 and 2 that the quadratic model of our extension to BOBYQA is updated by the formula

$$Q_{k+1}(\underline{x}) \;=\; Q_k(\underline{x}) + \{F(\underline{y}_t^+) - Q_k(\underline{y}_t^+)\}\,\Lambda_t(\underline{x}), \qquad \underline{x}\in\mathcal{R}^n. \tag{A.1}$$

The vector $\underline{y}_t^+ = \underline{x}_k + \underline{d}_k$ is the new interpolation point of the k-th iteration, and the function Λ_t is a convenient estimate of the quadratic polynomial that minimizes the semi-norm

$$\|\Lambda_t\|_\theta \;=\; \left\{\|\nabla^2\Lambda_t\|_F^2 + 2\theta\,\|\underline{\nabla}\Lambda_t(\underline{v})\|_2^2\right\}^{1/2}, \tag{A.2}$$

subject to the Lagrange interpolation equations

$$\Lambda_t(\underline{y}_j^+) \;=\; \delta_{jt}, \qquad j=1,2,\ldots,m, \tag{A.3}$$

the choices of θ and $\underline{v}$ being specified in Section 3. The procedure that constructs Λ_t in all the calculations of Tables 1–3 and 5–7 is described in this appendix, after some relevant theory. Usually it requires only $\mathcal{O}(n^2)$ operations when m is of magnitude n, this small amount of work being achieved by applying a version of truncated conjugate gradients that provides an approximation to the optimal Λ_t. Then a refinement procedure satisfies the conditions (A.3) exactly except for contributions from computer rounding errors.

We write Λ_t in the form

$$\Lambda_t(\underline{x}) \;=\; c + (\underline{x}-\underline{v})^T\underline{g} + \tfrac{1}{2}\,(\underline{x}-\underline{v})^T G\,(\underline{x}-\underline{v}), \qquad \underline{x}\in\mathcal{R}^n, \tag{A.4}$$

and assume for the moment that the number $c\in\mathcal{R}$, the components of $\underline{g}\in\mathcal{R}^n$ and the elements of the $n\times n$ matrix G have the values that minimize the expression

$$\tfrac{1}{4}\,\|\Lambda_t\|_\theta^2 \;=\; \tfrac{1}{4}\,\|G\|_F^2 + \tfrac{1}{2}\,\theta\,\|\underline{g}\|_2^2, \tag{A.5}$$

subject to the constraints

$$c + (\underline{y}_j^+-\underline{v})^T\underline{g} + \tfrac{1}{2}\,(\underline{y}_j^+-\underline{v})^T G\,(\underline{y}_j^+-\underline{v}) \;=\; \delta_{jt}, \qquad j=1,2,\ldots,m, \tag{A.6}$$

which is a quadratic programming problem. The contributions to the first order KKT conditions of this problem from first derivatives with respect to the elements of G supply the identity

$$G \;=\; \textstyle\sum_{\ell=1}^m \mu_\ell\,(\underline{y}_\ell^+-\underline{v})\,(\underline{y}_\ell^+-\underline{v})^T, \tag{A.7}$$

the multipliers μ_ℓ, $\ell=1,2,\ldots,m$, being the Lagrange parameters of the constraints (A.6) in the KKT conditions. Furthermore, differentiation with respect to c and the components of $\underline{g}$ shows that these multipliers also have the properties

$$\textstyle\sum_{\ell=1}^m \mu_\ell = 0 \qquad \text{and} \qquad \sum_{\ell=1}^m \mu_\ell\,(\underline{y}_\ell^+-\underline{v}) = \theta\,\underline{g}. \tag{A.8}$$

By substituting the form (A.7) into expression (A.6), we deduce that these constraints with the properties (A.8) provide the $(m+n+1)\times(m+n+1)$ linear system of equations

$$\left(\begin{array}{c|c|c} A & \underline{e} & Y^T \\ \hline \underline{e}^T & 0 & 0 \\ \hline Y & 0 & 0 \end{array}\right) \left(\begin{array}{c} \underline{\mu} \\ \hline c \\ \hline \underline{g} \end{array}\right) = \left(\begin{array}{c} \underline{e}_t \\ \hline 0 \\ \hline \theta\,\underline{g} \end{array}\right), \tag{A.9}$$

where A is the $m\times m$ matrix with the elements

$$A_{j\ell} = \tfrac{1}{2}\left\{(\underline{y}_j^+-\underline{v})^T(\underline{y}_\ell^+-\underline{v})\right\}^2, \qquad j=1,2,\ldots,m, \quad \ell=1,2,\ldots,m, \tag{A.10}$$

where all the components of $\underline{e}\in\mathcal{R}^m$ are 1, where Y is the $n\times m$ matrix with the columns $\underline{y}_\ell^+-\underline{v}$, $\ell=1,2,\ldots,m$, and where $\underline{e}_t$ is the t-th coordinate vector in $\mathcal{R}^m$. Putting the $\theta\underline{g}$ term of the system (A.9) on the right hand side brings the advantage that the matrix on the left hand side is independent of θ.

The techniques of BOBYQA for updating the interpolation points $\underline{y}_j^+$, $j=1,2,\ldots,m$, ensure that the matrix on the left hand side of expression (A.9) is nonsingular. We write its inverse in the partitioned form

$$\left(\begin{array}{c|c|c} A & \underline{e} & Y^T \\ \hline \underline{e}^T & 0 & 0 \\ \hline Y & 0 & 0 \end{array}\right)^{-1} = \left(\begin{array}{c|c|c} \Omega_v & \underline{\varepsilon}_v & \Gamma_v^T \\ \hline \underline{\varepsilon}_v^T & \xi_v & \underline{\omega}_v^T \\ \hline \Gamma_v & \underline{\omega}_v & \Upsilon_v \end{array}\right), \tag{A.11}$$

the dependence on $\underline{v}$ that is indicated by the subscripts being important later. We also assume for the moment that the submatrices Ω_v, Γ_v and Υ_v are available, their dimensions being $m\times m$, $n\times m$ and $n\times n$, respectively. Let the system (A.9) be multiplied by the inverse matrix (A.11). Then the last n components of the product are the relation

$$\underline{g} = \Gamma_v\,\underline{e}_t + \theta\,\Upsilon_v\,\underline{g}. \tag{A.12}$$

The elements (A.10) make A positive definite or semi-definite, which is going to be demonstrated neatly when we address the calculation of the term $\|\nabla^2\Lambda_t^{(0)}\|_F^2$ in formula (3.7). It follows that Υ_v in expressions (A.11) and (A.12) has no positive eigenvalues, so the eigenvalues of the $n\times n$ matrix $I-\theta\Upsilon_v$ are bounded below by one for $\theta\geq 0$. We regard the solution of the system (A.12) as the unconstrained minimization of the strictly convex quadratic function

$$\Phi(\underline{g}) = -\underline{g}^T\Gamma_v\,\underline{e}_t + \tfrac{1}{2}\,\underline{g}^T(I-\theta\,\Upsilon_v)\,\underline{g}, \qquad \underline{g}\in\mathcal{R}^n. \tag{A.13}$$

Indeed, unless the choice $\underline{g}=\Gamma_v\underline{e}_t$ is made as mentioned below, we construct $\underline{g}$ by applying a few iterations of the conjugate gradient method with exact line searches to the function (A.13), starting at $\underline{g}=0=\underline{g}^{(0)}$, say, this method being described in Fletcher (1987), for instance. The ℓ-th conjugate gradient iteration obtains $\underline{g}^{(\ell)}$ by searching from $\underline{g}^{(\ell-1)}$ along a direction $\underline{d}^{(\ell)}$, the work for each ℓ being only $\mathcal{O}(n)$, except that $\underline{d}^{(\ell)}$ has to be multiplied by the matrix $(I-\theta\,\Upsilon_v)$.

The numbers $\Phi(\underline{g}^{(\ell)})$, $\ell \geq 0$, and the gradients $\underline{\nabla}\Phi(\underline{g}^{(\ell)})$ are calculated during the conjugate gradient iterations, $\underline{\nabla}\Phi(\underline{g}^{(\ell)})$ being important to the choice of $\underline{d}^{(\ell+1)}$. The iterations are terminated, $\underline{g}^{(\ell)}$ being accepted as a sufficiently accurate estimate of the minimizer of the function (A.13), $\underline{g}^{(\rm opt)}$ say, when one (or both) of the conditions

$$\left.\begin{array}{rcl} \Phi(\underline{g}^{(\ell-1)}) - \Phi(\underline{g}^{(\ell)}) & \leq & 0.01\left\{\Phi(\underline{g}^{(0)}) - \Phi(\underline{g}^{(\ell-1)})\right\} \\ \text{and} \quad \|\underline{\nabla}\Phi(\underline{g}^{(\ell)})\| & \leq & 0.01\,\|\underline{y}_s^+ - \underline{x}_{k+1}\|^{-1} \end{array}\right\} \tag{A.14}$$

is achieved, where $\underline{y}_s^+$ is the closest point to $\underline{x}_{k+1}$ in the set $\{\underline{y}_j^+ : j = 1, 2, \ldots, m\}$ excluding $\underline{x}_{k+1}$. The first test causes truncation if the reductions in Φ become relatively small. The other test is derived from the view that, because gradients of magnitude $\|\underline{y}_s^+ - \underline{x}_{k+1}\|^{-1}$ are unavoidable near $\underline{x}_{k+1}$ in some Lagrange functions even if θ is huge, errors of at most $0.01\,\|\underline{y}_s^+ - \underline{x}_{k+1}\|^{-1}$ in the estimate $\underline{g}^{(\ell)} \approx \underline{g}^{(\rm opt)}$ are tolerable. The negative definiteness of Υ_v in the definition (A.13) provides the bound

$$\begin{aligned} \|\,\underline{g}^{(\ell)} - \underline{g}^{(\rm opt)}\| &\leq \|\,(I - \theta\Upsilon_v)\,(\underline{g}^{(\ell)} - \underline{g}^{(\rm opt)})\| \\ &= \|\,\underline{\nabla}\Phi(\underline{g}^{(\ell)}) - \underline{\nabla}\Phi(\underline{g}^{(\rm opt)})\| \;=\; \|\,\underline{\nabla}\Phi(\underline{g}^{(\ell)})\|. \end{aligned} \tag{A.15}$$

Therefore the second test of expression (A.14) is suitable.

We return to the case when the estimate $\underline{g}^{(\ell)} \approx \underline{g}^{(\rm opt)}$ is picked with $\ell \geq 1$, after considering two other situations that are without conjugate gradient iterations. This happens when θ is zero, because then expressions (A.9) and (A.11) show that the parameters $\underline{g} = \Gamma_v \underline{e}_t$ and $\underline{\mu} = \Omega_v \underline{e}_t$ provide the required Lagrange function (A.4), G being the matrix (A.7). The value $c = \Lambda_t(\underline{v})$ is taken directly from the constraints (A.3), as $\underline{v} = \underline{x}_{k+1}$ is one of the points $\underline{y}_j^+$, $j = 1, 2, \ldots, m$. Furthermore, this Lagrange function is selected in all situations without conjugate gradients even if θ is positive.

In particular, the second of the inequalities (A.14) is tested for $\ell = 0$ before any conjugate gradient iterations, and if it holds we switch to the procedure of the previous paragraph. In this case $\underline{\nabla}\Phi(\underline{g}^{(\ell)})$ is the vector $\underline{\nabla}\Phi(\underline{g}^{(0)}) = \underline{\nabla}\Phi(0) = -\Gamma_v \underline{e}_t$, and $\underline{g}$ is set to $\Gamma_v \underline{e}_t$, which supplies the estimate $\Gamma_v \underline{e}_t \approx \underline{g}^{(\rm opt)}$ instead of the approximation $\underline{g}^{(\ell)} \approx \underline{g}^{(\rm opt)}$ that occurred before. Therefore we ask whether the error $\|\Gamma_v \underline{e}_t - \underline{g}^{(\rm opt)}\|$ is also bounded above by $0.01\,\|\underline{y}_s^+ - \underline{x}_{k+1}\|^{-1}$. The identity $\underline{g}^{(\rm opt)} = (I - \theta\Upsilon_v)^{-1}\Gamma_v \underline{e}_t$ provides the equation

$$\|\,\Gamma_v\,\underline{e}_t - \underline{g}^{(\rm opt)}\| \;=\; \|\,(I - \theta\,\Upsilon_v)^{-1}\,(-\theta\,\Upsilon_v)\,\Gamma_v\,\underline{e}_t\|, \tag{A.16}$$

and the negative definiteness of Υ_v with $\theta \geq 0$ implies that all the eigenvalues of the symmetric matrix $(I - \theta\Upsilon_v)^{-1}(-\theta\Upsilon_v)$ are in the interval $[0, 1)$. Thus expression (A.16) is at most $\|\Gamma_v \underline{e}_t\| = \|\underline{\nabla}\Phi(\underline{g}^{(0)})\|$, which gives the required result.

We now address the situation when the inverse matrix (A.11) is not available, which happens on most iterations, because the task of calculating the matrix

directly, or of constructing it by an updating procedure from the inverse matrix of a previous iteration when $\underline{v}$ is changed, requires $\mathcal{O}(n^3)$ operations, due mainly to the property that the elements (A.10) are quartic functions of $\underline{v}$. Therefore BOBYQA works with the inverse matrix on the right hand side of the equation

$$\left(\begin{array}{c|c|c} A_0 & \underline{e} & Y_0^T \\ \hline \underline{e}^T & 0 & 0 \\ \hline Y_0 & 0 & 0 \end{array}\right)^{-1} = \left(\begin{array}{c|c|c} \Omega_0 & \underline{\varepsilon}_0 & \Gamma_0^T \\ \hline \underline{\varepsilon}_0^T & \xi_0 & \underline{\omega}_0^T \\ \hline \Gamma_0 & \underline{\omega}_0 & \Upsilon_0 \end{array}\right), \tag{A.17}$$

where A_0 has the elements $\frac{1}{2}\{(\underline{y}_j^+ - \underline{x}_0)^T(\underline{y}_\ell^+ - \underline{x}_0)\}^2$, $1 \le j, \ell \le m$, where Y_0 has the columns $\underline{y}_\ell^+ - \underline{x}_0$, $\ell = 1, 2, \ldots, m$, and where $\underline{x}_0$ is a suitable point that usually remains fixed for several consecutive iterations. When the k-th iteration replaces the old point $\underline{y}_t$ by $\underline{y}_t^+ = \underline{x}_k + \underline{d}_k$, as mentioned in the first paragraph of Section 2, only the t-th row and column of the matrix on the left hand side of expression (A.17) are altered for fixed $\underline{x}_0$, which allows the right hand side to be updated in only $\mathcal{O}(n^2)$ operations, details being given in Powell (2004b). That procedure has excellent stability properties, it stores and revises only the submatrices Ω_0, Γ_0 and Υ_0, and Ω_0 is in a factored form that preserves its rank and its positive semi-definiteness. These features are retained in our extension to BOBYQA.

It is important to the accuracy of BOBYQA in practice to keep $\underline{x}_0$ fairly close to the interpolation points $\underline{y}_j^+$, $j = 1, 2, \ldots, m$, so $\underline{x}_0$ is shifted occasionally. We assume temporarily that the k-th iteration moves $\underline{x}_0$ to the new position $\underline{v} = \underline{x}_{k+1}$, knowing this does not happen on most iterations, because then the matrices (A.11) and (A.17) become the same. Thus the operations of the shift, described in Section 5 of Powell (2004a), include formulae that express Ω_v, Γ_v and Υ_v in terms of the available submatrices Ω_0, Γ_0 and Υ_0. Specifically, for fixed $\underline{x}_0$, the required submatrices are given by the product

$$\left(\begin{array}{c|c} \Omega_v & \Gamma_v^T \\ \hline \Gamma_v & \Upsilon_v \end{array}\right) = \left(\begin{array}{c|c} I & 0 \\ \hline Z & I \end{array}\right) \left(\begin{array}{c|c} \Omega_0 & \Gamma_0^T \\ \hline \Gamma_0 & \Upsilon_0 \end{array}\right) \left(\begin{array}{c|c} I & Z^T \\ \hline 0 & I \end{array}\right), \tag{A.18}$$

where Z is the $n \times m$ matrix that has the columns

$$\underline{z}_j = \left\{(\underline{y}_j^+ - \tfrac{1}{2}\,\underline{x}_0 - \tfrac{1}{2}\,\underline{v})^T(\underline{v} - \underline{x}_0)\right\}(\underline{y}_j^+ - \tfrac{1}{2}\,\underline{x}_0 - \tfrac{1}{2}\,\underline{v}), \qquad j = 1, 2, \ldots, m. \tag{A.19}$$

The reader may notice that this choice of $\underline{z}_j$ is without the term $\frac{1}{4}\|\underline{v} - \underline{x}_0\|^2(\underline{v} - \underline{x}_0)$ that occurs in the definition of $\underline{z}_j$ in Powell (2004a). These terms alter Z by the rank one matrix $\frac{1}{4}\|\underline{v} - \underline{x}_0\|^2(\underline{v} - \underline{x}_0)\underline{e}^T$, which does not disturb the product (A.18), because the definition (A.17) gives the conditions $\underline{e}^T\Omega_0 = 0$ and $\underline{e}^T\Gamma_0^T = 0$.

We now return to the calculation of $\underline{g}^{(\ell)}$ from $\underline{g}^{(\ell-1)}$ by the ℓ-th iteration of the truncated conjugate gradient procedure. Equation (A.18) provides our way of multiplying the direction $\underline{d}^{(\ell)}$ by $(I - \theta\Upsilon_v)$ in only $\mathcal{O}(n^2)$ operations. Indeed, it gives the formula

$$\begin{aligned} \Upsilon_v\,\underline{d}^{(\ell)} &= \left\{Z\Omega_0 Z^T + Z\Gamma_0^T + \Gamma_0 Z^T + \Upsilon_0\right\}\underline{d}^{(\ell)} \\ &= Z\left\{\Omega_0(Z^T\underline{d}^{(\ell)}) + \Gamma_0^T\underline{d}^{(\ell)}\right\} + \Gamma_0\,(Z^T\underline{d}^{(\ell)}) + \Upsilon_0\,\underline{d}^{(\ell)}. \end{aligned} \tag{A.20}$$

The vector $(Z^T\underline{d}^{(\ell)}) \in \mathcal{R}^m$ has the components $\underline{z}_j^T\underline{d}^{(\ell)}$, $j=1,2,\ldots,m$, and the product $Z\{\Omega_0(Z^T\underline{d}^{(\ell)})+\Gamma_0^T\underline{d}^{(\ell)}\}$ can be formed as a linear combination of the columns of Z. The other multiplications of vectors by matrices in expression (A.20) are straightforward, the matrices being Ω_0, Γ_0 and Υ_0, which are available. The description of the truncated conjugate gradient procedure for seeking the approximate minimum of the function (A.13) is now complete, except for the calculation of the initial search direction $\underline{d}^{(1)}=-\underline{\nabla}\Phi(\underline{g}^{(0)})=-\underline{\nabla}\Phi(0)=\Gamma_v\underline{e}_t$.

Equation (A.18) implies that $\Gamma_v\underline{e}_t$ is the vector $(\Gamma_0+Z\Omega_0)\underline{e}_t$, but we prefer an equivalent form that is without the dependence on Z, this form being the expression

$$\Gamma_v\underline{e}_t \;=\; \Gamma_0\,\underline{e}_t \;+\; \sum_{j=1}^m \mu_j\,(\underline{y}_j^+-\underline{v})\,(\underline{y}_j^+-\underline{v})^T(\underline{v}-\underline{x}_0), \tag{A.21}$$

where μ_j, $j=1,2,\ldots,m$, are now the components of $\Omega_0\,\underline{e}_t$. In order to show the equivalence, we note that, because $Z\Omega_0\,\underline{e}_t$ is the sum $\sum_{j=1}^m \mu_j\underline{z}_j$, a rearrangement of the definition (A.19) supplies the identity

$$Z\,\Omega_0\,\underline{e}_t \;=\; \sum_{j=1}^m \mu_j\,(\underline{y}_j^+-\tfrac{1}{2}\,\underline{x}_0-\tfrac{1}{2}\,\underline{v})\,(\underline{y}_j^+-\tfrac{1}{2}\,\underline{x}_0-\tfrac{1}{2}\,\underline{v})^T(\underline{v}-\underline{x}_0). \tag{A.22}$$

Moreover, equation (A.17) gives not only $\underline{e}^T\Omega_0=0$ but also $Y_0\Omega_0=0$, providing both $\underline{e}^T\Omega_0\,\underline{e}_t=0$ and $Y_0\Omega_0\,\underline{e}_t=0$, which are the conditions

$$\sum_{j=1}^m \mu_j \;=\; 0 \qquad \text{and} \qquad \sum_{j=1}^m \mu_j\,(\underline{y}_j^+-\underline{x}_0) \;=\; 0. \tag{A.23}$$

It follows that the sum of equation (A.21) is the vector $Z\Omega_0\,\underline{e}_t$ as required.

Formula (A.21) is also useful in the construction of $\Lambda_t^{(0)}(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, which is defined in the last paragraph of Section 3 to be the quadratic that minimizes expression (A.2) subject to the constraints (A.3) when θ is zero. The parameters of $\Lambda_t^{(0)}$ are given by the system (A.9) with $\theta=0$, so the definition (A.11) provides $\underline{\mu}=\Omega_v\underline{e}_t$ and $\underline{g}=\Gamma_v\underline{e}_t$. Therefore, because the product (A.18) implies $\Omega_v=\Omega_0$, the components of $\underline{\mu}$ are those that occur in the previous paragraph, and $\underline{g}$ is the vector (A.21). These remarks give the parameters of Λ_t when Λ_t is set to $\Lambda_t^{(0)}$, which happens if and only if there are no conjugate gradient iterations. Furthermore, the value of $\|\nabla^2\Lambda_t^{(0)}\|_F^2$ is required for equation (3.7), its calculation being addressed next.

The square of the Frobenius norm of a real matrix is the trace of the matrix times its transpose. Thus expression (A.7) provides the formula

$$\begin{aligned}\|G\|_F^2 \;&=\; \text{Trace}\Big\{\sum_{j=1}^m \mu_j\,(\underline{y}_j^+-\underline{v})\,(\underline{y}_j^+-\underline{v})^T\,\sum_{\ell=1}^m \mu_\ell\,(\underline{y}_\ell^+-\underline{v})\,(\underline{y}_\ell^+-\underline{v})^T\Big\}\\ &=\; \text{Trace}\Big\{\sum_{j=1}^m\sum_{\ell=1}^m \mu_j\,\mu_\ell\,(\underline{y}_j^+-\underline{v})\,(\underline{y}_j^+-\underline{v})^T\,(\underline{y}_\ell^+-\underline{v})\,(\underline{y}_\ell^+-\underline{v})^T\Big\}\\ &=\; \sum_{j=1}^m\sum_{\ell=1}^m \mu_j\,\mu_\ell\,\Big\{(\underline{y}_j^+-\underline{v})^T\,(\underline{y}_\ell^+-\underline{v})\Big\}^2 \;=\; 2\,\underline{\mu}^T A\,\underline{\mu},\end{aligned} \tag{A.24}$$

where A has the elements (A.10). This equation is valid for general $\underline{\mu}\in\mathcal{R}^m$, and the left hand side is always nonnegative, which establishes the result stated earlier that A is positive definite or semi-definite. The work of applying this formula is reduced to $\mathcal{O}(n^2)$ operations by employing the scalar products

$$\Xi_{j\ell} \;=\; (\underline{y}_j^+-\underline{v})^T\,(\underline{y}_\ell^+-\underline{v}) \;=\; (\underline{y}_j^+-\underline{x}_{k+1})^T\,(\underline{y}_\ell^+-\underline{x}_{k+1}),\quad j,\ell=1,2,\ldots,m. \tag{A.25}$$

They are calculated for $\underline{v}=\underline{x}_1$ before the first iteration, and they are updated on every iteration, the work of this updating for each k being only $\mathcal{O}(n^2)$, even if $\underline{x}_{k+1}$ is different from $\underline{x}_k$. We recall from the previous paragraph that $G=\nabla^2\Lambda_t^{(0)}$ is the matrix (A.7) when the multipliers μ_j, $j=1,2,\ldots,m$, are the components of $\Omega_0\,\underline{e}_t$. Then equations (A.24) and (A.25) supply the value

$$\|\nabla^2\Lambda_t^{(0)}\|_F^2 \;=\; \sum_{j=1}^m\,\sum_{\ell=1}^m\,\mu_j\,\mu_\ell\,\Xi_{j\ell}^2. \tag{A.26}$$

Thus the implementation of formula (3.7) by the extended version of BOBYQA takes only $\mathcal{O}(n^2)$ operations, due to our assumption $m=\mathcal{O}(n)$.

The remainder of this appendix completes the description of the choice of Λ_t after the truncated conjugate gradient method has constructed an estimate $\underline{g}^{(\ell)}$ with $\ell\geq 1$ of the vector $\underline{g}^{(\rm opt)}$ that minimizes the function (A.13). We recall that $\underline{g}^{(\rm opt)}$ is the $\underline{g}$ defined by the system (A.9), and we let $\underline{\mu}^{(\rm opt)}$ be the corresponding value of $\underline{\mu}$. By multiplying the system by the inverse matrix (A.11) again, we find that the first m components of the product are the expression

$$\underline{\mu}^{(\rm opt)} \;=\; \Omega_v\,\underline{e}_t+\theta\,\Gamma_v^T\underline{g}^{(\rm opt)} \;=\; \Omega_0\,\underline{e}_t+\theta\,(\Gamma_0^T+\Omega_0\,Z^T)\,\underline{g}^{(\rm opt)}, \tag{A.27}$$

where the last equation depends on the relation (A.18). The replacement of $\underline{g}^{(\rm opt)}$ by $\underline{g}^{(\ell)}=\underline{g}^{(\rm est)}$, say, provides the estimate

$$\underline{\mu}^{(\rm est)} \;=\; \Omega_0\,\underline{e}_t+\theta\,\Gamma_0^T\underline{g}^{(\rm est)}+\theta\,\Omega_0\,(Z^T\underline{g}^{(\rm est)}) \tag{A.28}$$

of $\underline{\mu}^{(\rm opt)}$, which is calculated in a way that depends on the properties $\Omega_0\,\underline{e}=0$ and $\Omega_0 Y_0^T=0$, mentioned already and implied by equation (A.17). Specifically, the definition (A.19) allows the components of $Z^T\underline{g}^{(\rm est)}$ to be replaced by the numbers $(\underline{y}_j^+-\underline{v})^T(\underline{v}-\underline{x}_0)\,(\underline{y}_j^+-\underline{v})^T\underline{g}^{(\rm est)}$, $j=1,2,\ldots,m$.

After $\underline{g}^{(\rm est)}=\underline{g}^{(\ell)}$ and $\underline{\mu}^{(\rm est)}$ are constructed, it is assumed there is no need to give further attention to the fact that θ is positive in the semi-norm (A.2). Indeed, our stopping conditions for the conjugate gradient iterations are intended to take sufficient account of $\theta>0$ in a way that keeps ℓ small. On the other hand, we require our choice of Λ_t to satisfy the Lagrange interpolation equations (A.3) to high accuracy, assuming that computer rounding errors are negligible.

Therefore the final choice of Λ_t is a perturbation of the estimate

$$\Lambda_t^{(\rm est)}(\underline{x}) \;=\; \Lambda_t(\underline{v})+(\underline{x}-\underline{v})^T\underline{g}^{(\rm est)}+\tfrac{1}{2}\,(\underline{x}-\underline{v})^TG^{(\rm est)}(\underline{x}-\underline{v}),\qquad \underline{x}\in\mathcal{R}^n, \tag{A.29}$$

where the value of $\Lambda_t(\underline{v})$ is taken from equation (A.3) with $\underline{v}=\underline{x}_{k+1}$, where $G^{(\rm est)}$ is the matrix

$$G^{(\rm est)} \;=\; \textstyle\sum_{\ell=1}^m \mu_\ell^{(\rm est)}\,(\underline{y}_\ell^+-\underline{v})\,(\underline{y}_\ell^+-\underline{v})^T, \tag{A.30}$$

and where the perturbation is defined in a way that achieves all of the conditions (A.3).

The values $\Lambda_t^{(\rm est)}(\underline{y}_j^+)$, $j=1,2,\ldots,m$, are calculated by applying the formula

$$\begin{aligned}\Lambda_t^{(\rm est)}(\underline{y}_j^+) \;&=\; \Lambda_t(\underline{v})+(\underline{y}_j^+-\underline{v})^T\underline{g}^{(\rm est)}+\tfrac{1}{2}\,(\underline{y}_j^+-\underline{v})^T G^{(\rm est)}(\underline{y}_j^+-\underline{v})\\ &=\; \Lambda_t(\underline{v})+(\underline{y}_j^+-\underline{v})^T\underline{g}^{(\rm est)}+\tfrac{1}{2}\sum_{\ell=1}^m \mu_\ell^{(\rm est)}\,\Xi_{j\ell}^2,\quad j=1,2,\ldots,m,\end{aligned} \tag{A.31}$$

which requires only $\mathcal{O}(n^2)$ operations altogether, because the scalar products (A.25) are available. Then Λ_t is the sum $\Lambda_t^{(\rm est)}(\underline{x})+\Pi(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, where the perturbation Π is a quadratic polynomial that satisfies the constraints

$$\Pi(\underline{y}_j^+) \;=\; \delta_{jt}-\Lambda_t^{(\rm est)}(\underline{y}_j^+),\qquad j=1,2,\ldots,m. \tag{A.32}$$

It is convenient to take up the freedom in Π by minimizing $\|\nabla^2\Pi\|_F$, which is the familiar symmetric Broyden method.

Indeed, by putting $\theta=0$ and $\underline{v}=\underline{x}_0$ in the variational calculation that gives the linear system (A.9), we find that the perturbation Π is the function

$$\Pi(\underline{x}) \;=\; \Pi(\underline{x}_0)+(\underline{x}-\underline{x}_0)^T\underline{g}^{(\pi)}+\tfrac{1}{2}\,(\underline{x}-\underline{x}_0)^T G^{(\pi)}(\underline{x}-\underline{x}_0),\qquad \underline{x}\in\mathcal{R}^n, \tag{A.33}$$

with the second derivative matrix

$$G^{(\pi)} \;=\; \textstyle\sum_{\ell=1}^m \mu_\ell^{(\pi)}\,(\underline{y}_\ell^+-\underline{x}_0)\,(\underline{y}_\ell^+-\underline{x}_0)^T, \tag{A.34}$$

where $\underline{g}^{(\pi)}\in\mathcal{R}^n$ and $\underline{\mu}^{(\pi)}\in\mathcal{R}^m$ are defined by the linear system of equations

$$\left(\begin{array}{c|c|c} A_0 & \underline{e} & Y_0^T\\ \hline \underline{e}^T & 0 & 0\\ \hline Y_0 & 0 & 0\end{array}\right)\left(\begin{array}{c}\underline{\mu}^{(\pi)}\\ \hline \Pi(\underline{x}_0)\\ \hline \underline{g}^{(\pi)}\end{array}\right) \;=\; \left(\begin{array}{c}\underline{r}\\ \hline 0\\ \hline 0\end{array}\right), \tag{A.35}$$

the components of $\underline{r}$ being the right hand sides of expression (A.32). Therefore the inverse matrix (A.17), which is available, supplies the values $\underline{g}^{(\pi)}=\Gamma_0\,\underline{r}$ and $\underline{\mu}^{(\pi)}=\Omega_0\,\underline{r}$. Further, because the properties $\underline{e}^T\Omega_0=0$ and $Y_0\,\Omega_0=0$ provide both $\underline{e}^T\underline{\mu}^{(\pi)}=0$ and $Y_0\,\underline{\mu}^{(\pi)}=0$, which are the constraints (A.23) with $\underline{\mu}=\underline{\mu}^{(\pi)}$, we may change $\underline{x}_0$ to $\underline{v}$ in expression (A.34). It follows that, writing $\Lambda_t=\Lambda_t^{(\rm est)}+\Pi$ in the form (A.4) with the second derivative matrix (A.7), its parameters $\underline{g}\in\mathcal{R}^n$ and $\underline{\mu}\in\mathcal{R}^m$ take the values

$$\begin{aligned}\underline{g} \;&=\; \underline{g}^{(\rm est)}+\underline{\nabla}\Pi(\underline{v}) \;=\; \underline{g}^{(\rm est)}+\underline{g}^{(\pi)}+G^{(\pi)}\,(\underline{v}-\underline{x}_0)\\ &=\; \underline{g}^{(\rm est)}+\Gamma_0\,\underline{r}+\textstyle\sum_{\ell=1}^m \mu_\ell^{(\pi)}\,(\underline{y}_\ell^+-\underline{v})\,(\underline{y}_\ell^+-\underline{v})^T(\underline{v}-\underline{x}_0)\end{aligned} \tag{A.36}$$

and $\underline{\mu}=\underline{\mu}^{(\rm est)}+\underline{\mu}^{(\pi)}$. The description of the construction of Λ_t by the extended version of BOBYQA is complete.