On updating the inverse of a KKT matrix[1]

M.J.D. Powell

Abstract: A KKT matrix, W say, is symmetric and nonsingular, with a leading $\hat{n}\times\hat{n}$ block that has a conditional positive definite property and a trailing $\hat{m}\times\hat{m}$ block that is identically zero, the dimensions of W being $(\hat{n}+\hat{m})\times(\hat{n}+\hat{m})$. The author requires the inverse matrix $H=W^{-1}$ explicitly in an iterative algorithm for unconstrained minimization without derivatives, and only one of the first $\hat{n}$ rows and columns of W is altered on each iteration. The corresponding change to H can be calculated in $\mathcal{O}(\hat{n}^2)$ operations. We study the accuracy and stability of some methods for this updating problem, finding that huge errors can occur in the application to optimization, which tend to be corrected on later iterations. Let Ω be the leading $\hat{n}\times\hat{n}$ submatrix of H. We give particular attention to the remark that the rank of Ω is only $\hat{n}-\hat{m}$, due to the zero block of W. Thus Ω can be expressed as the sum of $\hat{n}-\hat{m}$ matrices of rank one, and this factorization can also be updated in $\mathcal{O}(\hat{n}^2)$ operations. We find, in theory and in practice, that the use of the factored form of Ω reduces the damage from rounding errors and improves the stability of the updating procedure. These conclusions are illustrated by numerical results from the algorithm for unconstrained minimization.

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

January, 2004.

[1]Presented at the International Conference on Numerical Linear Algebra and Optimization (October, 2003), Guilin, China.

1. Introduction

The KKT conditions at the solution of a constrained optimization problem give a system of equations that is satisfied by the variables and Lagrange multipliers of the calculation, provided that the relevant functions are differentiable, and the gradients of the active constraints are linearly independent (see Fletcher, 1987, for instance). In particular, for a quadratic programming problem in $\hat{n}$ variables, with $\hat{m}$ equality and no inequality constraints, the system of equations is linear, and its matrix has the form

$$W \;=\; \left(\begin{array}{c|c} A & X^T \\ \hline X & 0 \end{array}\right) \quad \begin{array}{l} \updownarrow \hat{n} \\ \updownarrow \hat{m}\,. \end{array} \tag{1.1}$$

Here A is the symmetric second derivative matrix of the objective function, and the rows of the $\hat{m}\times\hat{n}$ matrix X are the gradients of the constraints. The second order conditions for optimality state that the quadratic programming problem has a unique solution if the rank of X is $\hat{m}$, and if $\underline{v}^T A\underline{v}$ is strictly positive for every nonzero vector $\underline{v}\in\mathcal{R}^{\hat{n}}$ that satisfies $X\underline{v}=0$. In this case the matrix (1.1) is called a KKT matrix and is nonsingular. We write its inverse in the form

$$H \;=\; W^{-1} \;=\; \left(\begin{array}{c|c} \Omega & \Xi^T \\ \hline \Xi & \Upsilon \end{array}\right) \quad \begin{array}{l} \updownarrow \hat{n} \\ \updownarrow \hat{m}\,. \end{array} \tag{1.2}$$

Systems of equations with the matrix (1.1) are also important to the iterations of sequential quadratic programming algorithms for constrained optimization calculations. Then A is an estimate of a second derivative matrix of the Lagrange function, and the rows of X are gradients of constraint functions for a particular choice of the vector of variables. Therefore, when this vector is revised and the constraints are nonlinear, the change to X may be of full rank. In this case it is usually not advantageous to employ updating techniques for the solutions of the systems of equations of consecutive iterations.

On the other hand, the author is constructing an iterative algorithm for unconstrained minimization without derivatives, where a KKT matrix occurs on every iteration, and where the change to the KKT matrix from one iteration to the next is confined to a single row and column. Further, the inverse matrix (1.2) is stored instead of W, because the elements of H provide some coefficients of Lagrange functions that are required. Therefore in this paper we address the updating of H when only the t-th row and column of W are altered, preserving symmetry, where t is any integer from $[1,\hat{n}]$. As far as the author knows, this situation does not arise in algorithms for general constrained optimization calculations. Nevertheless, because of the importance of the KKT matrix, our results may be of interest to many researchers in mathematical programming.

The new algorithm for unconstrained minimization is still under development, a report on a provisional version being available (Powell, 2003), which presents

an outline of the method with a few numerical results. The author has also written a paper already on the updating calculation of the previous paragraph, which gives not only a useful formula for the new H that can be applied in $\mathcal{O}(\{\hat{m}+\hat{n}\}^2)$ operations, but also some analysis of the stability of the formula and some favourable numerical results. Indeed, most of the errors that are introduced by the use of inverse matrices are corrected automatically as the iterations proceed. Therefore the author had not expected to write another paper on this subject. The work on the new algorithm has been frustrated, however, by occasional huge losses in efficiency due to the effects of computer rounding errors, although the stability that has been mentioned gives excellent accuracy in the final value of the objective function eventually. A technique that reduces much or all of this damage was found by the author recently. It is based on the remark that, because equations (1.1) and (1.2) imply $X\Omega=0$, the rank of Ω is at most $\hat{n}-\hat{m}$. We are going to store and update Ω in a way that guarantees this property, even in the presence of computer rounding errors. Thus we will find that the performance of the new algorithm is improved greatly in some troublesome cases.

The relevance of KKT matrices to the new algorithm for minimization without derivatives is explained by Powell (2002b, 2003). That question is also addressed in the remainder of this section, because of its importance to our work. Specifically, each iteration employs a quadratic model $Q(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, of the objective function $F(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, that is required to satisfy interpolation conditions of the form

$$Q(\underline{x}_i) \;=\; F(\underline{x}_i), \qquad i=1,2,\ldots,m, \tag{1.3}$$

where m is a prescribed fixed integer from the interval $[n+2,\frac{1}{2}(n+1)(n+2)]$, the value $m=2n+1$ being typical. The quadratic models are updated as the calculation proceeds, but, after allowing for the constraints (1.3), there are $\frac{1}{2}(n+1)(n+2)-m$ degrees of freedom in each new quadratic model. The main idea of the new algorithm is to take up this freedom on each iteration by minimizing the Frobenius norm of the change to the second derivative matrix of the model. This technique is analogous to applying the symmetric Broyden formula when the gradient $\underline{\nabla}F(\underline{x})$, $\underline{x}\in\mathcal{R}^n$, is available (see Section 9.1 of Dennis and Schnabel, 1983, for instance).

We write the change to the quadratic model in the form

$$Q(\underline{x})-Q_{\rm old}(\underline{x}) \;=\; c+\underline{g}^T\underline{x}+\tfrac{1}{2}\,\underline{x}^T\Delta\,\underline{x}, \qquad \underline{x}\in\mathcal{R}^n, \tag{1.4}$$

where Q and $Q_{\rm old}$ are the new and old models, respectively. and where $c\in\mathcal{R}$, $\underline{g}\in\mathcal{R}^n$ and $\Delta\in\mathcal{R}^{n\times n}$ have to be calculated. The values of these parameters are defined by minimizing the squared Frobenius norm

$$\|\Delta\|_F^2 \;=\; \textstyle\sum_{i=1}^n\,\sum_{j=1}^n\,\Delta_{ij}^2, \tag{1.5}$$

subject to the interpolation equations

$$c+\underline{g}^T\underline{x}_i+\tfrac{1}{2}\,\underline{x}_i^T\Delta\,\underline{x}_i \;=\; F(\underline{x}_i)-Q_{\rm old}(\underline{x}_i), \qquad i=1,2,\ldots,m. \tag{1.6}$$

We ignore the symmetry condition $\Delta^T=\Delta$, because it is achieved automatically. Indeed, it follows from the KKT conditions of this subproblem that there exist multipliers λ_j, $j=1,2,\ldots,m$, satisfying the constraints

$$\sum_{j=1}^m \lambda_j \;=\; 0 \quad \text{and} \quad \sum_{j=1}^m \lambda_j\,\underline{x}_j \;=\; 0, \tag{1.7}$$

such that Δ is the matrix

$$\Delta \;=\; \sum_{j=1}^m \lambda_j\,\underline{x}_j\,\underline{x}_j^T. \tag{1.8}$$

Thus the updating of the quadratic model is reduced to the calculation of $c\in\mathcal{R}$, $\underline{g}\in\mathcal{R}^n$ and $\underline{\lambda}\in\mathcal{R}^m$.

Expressions (1.8) and (1.6) provide the identities

$$c+\underline{g}^T\underline{x}_i+\tfrac{1}{2}\sum_{j=1}^m \lambda_j\,(\underline{x}_i^T\underline{x}_j)^2 \;=\; F(\underline{x}_i)-Q_{\rm old}(\underline{x}_i), \qquad i=1,2,\ldots,m. \tag{1.9}$$

It follows from the constraints (1.7) that the required parameters satisfy the square system of linear equations

$$W\left(\begin{array}{c}\underline{\lambda}\\ \hline c\\ \underline{g}\end{array}\right) \;=\; \left(\begin{array}{c|c}A & X^T\\ \hline X & 0\end{array}\right)\left(\begin{array}{c}\underline{\lambda}\\ \hline c\\ \underline{g}\end{array}\right) \;=\; \left(\begin{array}{c}\underline{r}\\ \hline 0\end{array}\right)\begin{array}{l}\updownarrow\ m\\ \updownarrow\ n+1\,,\end{array} \tag{1.10}$$

where A has the elements

$$A_{ij} \;=\; \tfrac{1}{2}\,(\underline{x}_i^T\underline{x}_j)^2, \qquad 1\le i,j\le m, \tag{1.11}$$

where X is the matrix

$$X \;=\; \left(\begin{array}{cccc}1 & 1 & \cdots & 1\\ \hline \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_m\end{array}\right)\begin{array}{l}\updownarrow 1\\ \updownarrow n\,,\end{array} \tag{1.12}$$

and where the components of $\underline{r}\in\mathcal{R}^m$ are the right hand sides of the interpolation conditions (1.9). We see that the matrix of the system (1.10) has the form (1.1), the values of $\hat{n}$ and $\hat{m}$ being m and $n+1$, respectively. It is highly useful that the number of unknowns is only $m+n+1$, because the quadratic polynomial (1.4) has $\frac{1}{2}(n+1)(n+2)$ independent parameters.

One can deduce from the elements (1.11) that A has no negative eigenvalues (see equation (2.10) of Powell, 2002b), and the positions of the points $\underline{x}_i$, $i=1,2,\ldots,m$, are chosen so that the system (1.10) is nonsingular, which will receive attention in the next section. Therefore W is a KKT matrix. Moreover, each iteration of the optimization algorithm changes at most one of the interpolation points $\underline{x}_i$, $i=1,2,\ldots,m$. Let $\underline{x}_t$ be replaced by $\underline{x}^+$, where t is an integer from $[1,m]$. It follows from expressions (1.10)–(1.12) that all changes to W are confined to the t-th row and column, which gives the updating calculation of the third paragraph of this section. The current inverse matrix (1.2) is stored, as mentioned already. Thus the parameters c, $\underline{g}$ and $\underline{\lambda}$ of each change to the quadratic model

are calculated in only $\mathcal{O}(\{m+n\}^2)$ operations, which is also the amount of work for updating H.

Section 2 addresses the revision of H when the t-th row and column of W are altered. A simple updating formula is presented, which is equivalent to the first one that is derived by Powell (2002b). Some lemmas show the change in the determinant of W, and the dependence of the new matrix H on the t-th diagonal element of the new W. The stability of the updating method is considered in Section 3. An example exposes the huge loss of accuracy that can occur in the application to unconstrained optimization. Therefore the stability is crucial to the precision that can be achieved in the final vector of variables. In Section 4 we study the new idea of expressing the partition Ω of the matrix (1.2) as the sum of only $\hat{n}-\hat{m}=m-n-1$ rank one matrices, as suggested by the property $X\Omega=0$ that has been mentioned. It is proved that this technique gives a major improvement to the stability of the updating formula. We find too that some of the immediate damage from computer rounding errors is reduced. Ways of updating the new expression for Ω are the subject of Section 5. They allow the factorization of Ω to be included in the new algorithm for unconstrained minimization. Thus the efficiency of the algorithm is improved greatly, as shown by a few numerical experiments in Section 6. That final section includes also some remarks on the development of the new algorithm.

2. A simple version of the updating formula

The $\hat{m}\times\hat{m}$ zero partition of the matrix (1.1) is irrelevant throughout this section. Therefore we let W be any $d\times d$ nonsingular symmetric matrix, d being an integer that is at least two, and we consider the updating of $H=W^{-1}$, when only the t-th row and column of W are altered, their new values being $\underline{v}^T$ and $\underline{v}$, respectively. In other words, W is replaced by the matrix W^+ that has the elements

$$\left.\begin{array}{rcll} W^+_{kt} &=& W^+_{tk} = v_k, & k=1,2,\ldots,d, \\ W^+_{ij} &=& W_{ij}, & i,j\in\{1,2,\ldots,d\}\backslash\{t\}. \end{array}\right\} \tag{2.1}$$

Therefore, if W^+ is nonsingular, its inverse matrix $H^+=(W^+)^{-1}$ is characterized by the conditions

$$\left.\begin{array}{rcll} (H^+)^{-1}_{kt} &=& (H^+)^{-1}_{tk} = v_k, & k=1,2,\ldots,d, \\ (H^+)^{-1}_{ij} &=& H^{-1}_{ij}, & i,j\in\{1,2,\ldots,d\}\backslash\{t\}. \end{array}\right\} \tag{2.2}$$

The following theorem suggests a way of calculating H^+ from H and $\underline{v}\in\mathcal{R}^d$ in $\mathcal{O}(d^2)$ operations, when W is not available, provided that the diagonal element H_{tt} is nonzero.

Theorem: Let H be a nonsingular $d\times d$ real symmetric matrix and let t be any integer from $[1,d]$ such that H_{tt} is nonzero. Further, let $\underline{v}$ be any vector in $\mathcal{R}^d$

with the property

$$(\underline{e}_t - \tilde{H}\underline{v})^T \underline{v} \neq 0, \tag{2.3}$$

where $\underline{e}_t$ is the t-th coordinate vector in $\mathcal{R}^d$, and where $\tilde{H}$ is the singular matrix

$$\tilde{H} = H - \frac{H\,\underline{e}_t\,\underline{e}_t^T H}{\underline{e}_t^T H\,\underline{e}_t}. \tag{2.4}$$

Then the matrix

$$H^+ = \tilde{H} + \frac{(\underline{e}_t - \tilde{H}\underline{v})\,(\underline{e}_t - \tilde{H}\underline{v})^T}{(\underline{e}_t - \tilde{H}\underline{v})^T \underline{v}} \tag{2.5}$$

is nonsingular, and it satisfies the conditions (2.2).

Proof: The nonsingularity of H and the definition (2.4) imply that $\tilde{H}$ has a null space of dimension one, spanned by $\underline{e}_t$. Thus the matrix (2.5) has the property

$$H^+ \underline{e}_t = (\underline{e}_t - \tilde{H}\underline{v}) \,/\, (\underline{e}_t - \tilde{H}\underline{v})^T \underline{v}, \tag{2.6}$$

this vector being nonzero because of assumption (2.3). We address the nonsingularity of H^+ by letting $\underline{z} \in \mathcal{R}^d$ satisfy $H^+ \underline{z} = 0$. Then the symmetry of H^+ and equation (2.6) give the relation

$$0 = \underline{e}_t^T H^+ \underline{z} = (H^+ \underline{e}_t)^T \underline{z} = (\underline{e}_t - \tilde{H}\underline{v})^T \underline{z} \,/\, (\underline{e}_t - \tilde{H}\underline{v})^T \underline{v}. \tag{2.7}$$

Therefore the definition (2.5) provides $\tilde{H}\underline{z} = H^+ \underline{z} = 0$. It follows from the first statement of this proof that $\underline{z} = \theta\,\underline{e}_t$ holds for some $\theta \in \mathcal{R}$, and θ is zero due to $H^+ \underline{z} = 0$, because the vector (2.6) is nonzero. Hence $H^+ \underline{z} = 0$ implies $\underline{z} = 0$, which establishes that H^+ is nonsingular as required.

The matrix (2.5) also has the property $H^+ \underline{v} = \underline{e}_t$, which we write in the form $\underline{v} = (H^+)^{-1} \underline{e}_t$. Thus, remembering the symmetry of H^+, we deduce the first part of expression (2.2).

Turning to the second part, we let i and j be any integers from $[1, d]$ that are different from t. Because the definition (2.4) implies $\tilde{H}(H^{-1}\underline{e}_j) = \underline{e}_j$, formulae (2.5) and (2.6) give the condition

$$\begin{aligned} H^+(H^{-1}\underline{e}_j) &= \underline{e}_j + (\underline{e}_t - \tilde{H}\underline{v})\,\{(\underline{e}_t - \tilde{H}\underline{v})^T (H^{-1}\underline{e}_j)\} \,/\, (\underline{e}_t - \tilde{H}\underline{v})^T \underline{v} \\ &= \underline{e}_j + \{(\underline{e}_t - \tilde{H}\underline{v})^T (H^{-1}\underline{e}_j)\}\, H^+ \underline{e}_t. \end{aligned} \tag{2.8}$$

By pre-multiplying this equation by $\underline{e}_i^T (H^+)^{-1}$, and by employing $\underline{e}_i^T \underline{e}_t = 0$, we find the identity $\underline{e}_i^T H^{-1} \underline{e}_j = \underline{e}_i^T (H^+)^{-1} \underline{e}_j$, which is equivalent to the second part of expression (2.2). The proof is complete. □

The theorem presents a simple method for updating H, assuming that divisions by zero do not occur. Specifically, $\tilde{H}$ and H^+ are rank one modifications of H and $\tilde{H}$, respectively, that can be applied in only $\mathcal{O}(d^2)$ operations. On the other hand,

we find below that the division by $\underline{e}_t^T H \underline{e}_t$ is avoided if the intermediate matrix $\tilde{H}$ is eliminated analytically from the updating procedure.

We begin the elimination of $\tilde{H}$ by deducing from expression (2.4) the equation

$$\underline{e}_t - \tilde{H}\underline{v} \;=\; (\underline{e}_t - H\underline{v}) + \frac{\underline{e}_t^T H \underline{v}}{\underline{e}_t^T H \underline{e}_t} \, H \underline{e}_t, \tag{2.9}$$

which provides the denominator

$$\begin{aligned} (\underline{e}_t - \tilde{H}\underline{v})^T \underline{v} \;&=\; \left\{ (\underline{e}_t^T \underline{v} - \underline{v}^T H \underline{v}) \, \underline{e}_t^T H \underline{e}_t + (\underline{e}_t^T H \underline{v})^2 \right\} \Big/ \, \underline{e}_t^T H \underline{e}_t \\ &=\; \sigma \Big/ \, \underline{e}_t^T H \underline{e}_t, \end{aligned} \tag{2.10}$$

where the last line defines the real parameter σ. Therefore the updating procedure of the theorem gives the formula

$$H^+ \;=\; H - \frac{H \underline{e}_t \, \underline{e}_t^T H}{\underline{e}_t^T H \underline{e}_t} \;+\; \frac{\underline{e}_t^T H \underline{e}_t}{\sigma} \, (\underline{e}_t - \tilde{H}\underline{v}) \, (\underline{e}_t - \tilde{H}\underline{v})^T. \tag{2.11}$$

It follows from expression (2.9) that H^+ is the matrix

$$\begin{aligned} H^+ \;=\; H + \sigma^{-1} \Big[& \alpha \, (\underline{e}_t - H\underline{v}) \, (\underline{e}_t - H\underline{v})^T - \beta \, H \underline{e}_t \, \underline{e}_t^T H \\ & + \tau \left\{ H \underline{e}_t \, (\underline{e}_t - H\underline{v})^T + (\underline{e}_t - H\underline{v}) \, \underline{e}_t^T H \right\} \Big], \end{aligned} \tag{2.12}$$

where α, τ and β are the real parameters

$$\left. \begin{aligned} \alpha \;&=\; \underline{e}_t^T H \underline{e}_t, \qquad \tau \;=\; \underline{e}_t^T H \underline{v}, \qquad \text{and} \\ \beta \;&=\; \frac{\sigma}{\underline{e}_t^T H \underline{e}_t} \;-\; \frac{(\underline{e}_t^T H \underline{v})^2}{\underline{e}_t^T H \underline{e}_t} \;=\; \underline{e}_t^T \underline{v} - \underline{v}^T H \underline{v}. \end{aligned} \right\} \tag{2.13}$$

We see that analytic cancellation removes the division by $\underline{e}_t^T H \underline{e}_t$ in the last line. Therefore expression (2.12) can be used to calculate H^+ whenever $\sigma = \alpha\beta + \tau^2$ is nonzero. This expression is the first updating formula of Section 4 of Powell (2002b).

The equation $H^+ = (W^+)^{-1}$, stated in the first paragraph of this section, suggests that $\sigma = 0$ occurs in the updating formula (2.12) if and only if W^+ is singular. We investigate this question, retaining the assumption that H is any nonsingular $d \times d$ real symmetric matrix. Because H^{-1} is calculated from H and $\underline{v}$, we define $W = H^{-1}$, and we let W^+ have the elements (2.1). The relation between σ and the singularity of W^+ is as follows.

Lemma 1: The parameter σ of formula (2.12) has the value

$$\sigma \;=\; \det W^+ \,/\, \det W, \tag{2.14}$$

for any $\underline{v}\in\mathcal{R}^d$, the matrices W and W^+ being defined above.

Proof: The definition of W^+ and expression (2.13) with $\sigma=\alpha\beta+\tau^2$ imply that both sides of equation (2.14) depend continuously on $\underline{v}\in\mathcal{R}^d$. Therefore we may assume without loss of generality that $\tau=\underline{e}_t^T H\underline{v}$ is nonzero.

We show that W^+ can be expressed in the form

$$W^+ \;=\; (I-\underline{e}_t\,\underline{u}^T)\,W(I-\underline{u}\,\underline{e}_t^T)+\beta\,\underline{e}_t\,\underline{e}_t^T, \tag{2.15}$$

where $\underline{u}=\underline{e}_t-H\underline{v}$. Clearly this matrix is symmetric and satisfies the second part of the conditions (2.1), so it is sufficient to establish $W^+\underline{e}_t=\underline{v}$. The choice of $\underline{u}$ and the identity $W=H^{-1}$ provide the equation

$$W(I-\underline{u}\,\underline{e}_t^T)\,\underline{e}_t \;=\; W(\underline{e}_t-\underline{u}) \;=\; WH\underline{v} \;=\; \underline{v}. \tag{2.16}$$

Hence the matrix (2.15) has the required property

$$W^+\underline{e}_t \;=\; (I-\underline{e}_t\,\underline{u}^T)\,\underline{v}+\beta\,\underline{e}_t \;=\; \underline{v}-\underline{e}_t\,(\underline{e}_t-H\underline{v})^T\underline{v}+\beta\,\underline{e}_t \;=\; \underline{v}, \tag{2.17}$$

the last equation being due to the definition (2.13) of β.

We are going to combine expression (2.15) with the product rule for determinants. The matrix $(I-\underline{e}_t\underline{u}^T)$ has the determinant $1-\underline{u}^T\underline{e}_t=\tau$, which is nonzero by assumption, so equation (2.15) is equivalent to the relation

$$(I-\underline{e}_t\,\underline{u}^T)^{-1}\,W^+(I-\underline{u}\,\underline{e}_t^T)^{-1} \;=\; W+\beta\,\underline{z}\,\underline{z}^T, \tag{2.18}$$

where $\underline{z}$ is the vector

$$\underline{z} \;=\; (I-\underline{e}_t\,\underline{u}^T)^{-1}\underline{e}_t \;=\; \Big(I+\frac{\underline{e}_t\,\underline{u}^T}{1-\underline{u}^T\underline{e}_t}\Big)\,\underline{e}_t \;=\; \frac{\underline{e}_t}{1-\underline{u}^T\underline{e}_t} \;=\; \tau^{-1}\underline{e}_t. \tag{2.19}$$

Thus $WH=I$ implies the identity

$$(I-\underline{e}_t\,\underline{u}^T)^{-1}\,W^+(I-\underline{u}\,\underline{e}_t^T)^{-1} \;=\; W\left(I+\beta\tau^{-2}\,H\,\underline{e}_t\,\underline{e}_t^T\right). \tag{2.20}$$

Now $\det(I+\beta\tau^{-2}H\,\underline{e}_t\,\underline{e}_t^T)$ has the value $1+\beta\tau^{-2}\underline{e}_t^T H\,\underline{e}_t=1+\alpha\beta\tau^{-2}$. Hence, by taking determinants of the matrix equation (2.20), we obtain the condition

$$\tau^{-2}\det W^+ \;=\; (1+\alpha\,\beta\tau^{-2})\,\det W \;=\; \sigma\,\tau^{-2}\det W. \tag{2.21}$$

Therefore the lemma is true. □

The author has found that, when the updating formula (2.12) fails to provide good accuracy in practice, the trouble arises usually from severe cancellation in the calculation of β. Strong support for this claim is given in the next section. Therefore he investigated the errors that occur in the conditions (2.2) if β is incorrect in expression (2.12), the formulae

$$\alpha=\underline{e}_t^T H\,\underline{e}_t, \quad \tau=\underline{e}_t^T H\underline{v} \quad\text{and}\quad \sigma=\alpha\beta+\tau^2 \tag{2.22}$$

being retained. He had in mind the technique of modifying β, in order to avoid denominators σ that are very close to zero, in the application to the new algorithm for unconstrained optimization, which is mentioned in Section 1. Work on that technique has now been abandoned, however, because of the success of the factorization that we will study in Sections 4–6, but it did produce the following interesting result.

Lemma 2: Let $H^+(\beta)$ be the matrix (2.12) when t, H and $\underline{v}$ are as before, when α, τ and σ have the values (2.22), and when β is any real number such that $H^+(\beta)$ is finite and nonsingular. Then all the elements of the inverse matrix $\{H^+(\beta)\}^{-1}$ are independent of β, except for the t-th diagonal element.

Proof: Let $H^+(\hat{\beta})$ and $H^+(\check{\beta})$ be finite and nonsingular, where $\hat{\beta}\neq\check{\beta}$. It follows from equation (2.12) that the difference $H^+(\check{\beta})-H^+(\hat{\beta})$ is the matrix

$$(\hat{\sigma}\,\check{\sigma})^{-1}\Big[\alpha\,(\hat{\sigma}-\check{\sigma})\,(\underline{e}_t-H\underline{v})\,(\underline{e}_t-H\underline{v})^T-(\check{\beta}\,\hat{\sigma}-\hat{\beta}\,\check{\sigma})\,H\,\underline{e}_t\,\underline{e}_t^T H \\ +\tau\,(\hat{\sigma}-\check{\sigma})\Big\{H\,\underline{e}_t\,(\underline{e}_t-H\underline{v})^T+(\underline{e}_t-H\underline{v})\,\underline{e}_t^T H\Big\}\Big]\,, \tag{2.23}$$

where $\hat{\sigma}=\alpha\hat{\beta}+\tau^2$ and $\check{\sigma}=\alpha\check{\beta}+\tau^2$. The multipliers $\alpha\,(\hat{\sigma}-\check{\sigma})$, $-(\check{\beta}\hat{\sigma}-\hat{\beta}\check{\sigma})$ and $\tau(\hat{\sigma}-\check{\sigma})$ are the numbers $\alpha^2(\hat{\beta}-\check{\beta})$, $\tau^2(\hat{\beta}-\check{\beta})$ and $\alpha\tau(\hat{\beta}-\check{\beta})$, respectively. Therefore expression (2.23) is the symmetric outer product

$$(\hat{\beta}-\check{\beta})\,(\hat{\sigma}\,\check{\sigma})^{-1}\Big\{\alpha\,(\underline{e}_t-H\underline{v})+\tau\,H\,\underline{e}_t\Big\}\Big\{\alpha\,(\underline{e}_t-H\underline{v})+\tau\,H\,\underline{e}_t\Big\}^T\,. \tag{2.24}$$

Moreover, equations (2.12) and (2.22) in the case $\beta=\hat{\beta}$ give the identity

$$\begin{aligned} H^+(\hat{\beta})\,\underline{e}_t &= H\,\underline{e}_t+\hat{\sigma}^{-1}\left[(\underline{e}_t-H\underline{v})\,\{\alpha-\alpha\tau+\tau\alpha\}+H\,\underline{e}_t\,\{-\hat{\beta}\alpha+\tau-\tau^2\}\right] \\ &= \hat{\sigma}^{-1}\,\{\alpha\,(\underline{e}_t-H\underline{v})+\tau\,H\,\underline{e}_t\}\,. \end{aligned} \tag{2.25}$$

These remarks provide the relation

$$H^+(\check{\beta}) \;=\; H^+(\hat{\beta})+(\hat{\beta}-\check{\beta})\,(\hat{\sigma}/\check{\sigma})\,H^+(\hat{\beta})\,\underline{e}_t\,\underline{e}_t^T H^+(\hat{\beta}). \tag{2.26}$$

Hence $H^+(\check{\beta})$ has the inverse

$$\{H^+(\check{\beta})\}^{-1} \;=\; \{H^+(\hat{\beta})\}^{-1}-\;\frac{(\hat{\beta}-\check{\beta})\,\hat{\sigma}}{\check{\sigma}+(\hat{\beta}-\check{\beta})\,\hat{\sigma}\,\underline{e}_t^T H^+(\hat{\beta})\,\underline{e}_t}\;\underline{e}_t\,\underline{e}_t^T, \tag{2.27}$$

the denominator being nonzero because $\underline{e}_t$ is not in the null space of the matrix (2.26). Equation (2.27) shows that the lemma is true. $\square$

3. On the stability of the updating method

Both the updating method of the theorem and formula (2.12) have a highly useful stability property, because H^+ is calculated from t, H and $\underline{v}$. The property is a consequence of the conditions (2.2), when substantial errors are present in H, due to the effects of the computer rounding errors of the previous iterations. We consider the errors that are inherited by H^+, if the calculation of H^+ from t, H and $\underline{v}$ is exact, taking the view that H^{-1} and $(H^+)^{-1}$ should be the matrices W and W^+, respectively, where W^+ is defined by expression (2.1). Equations (2.1) and (2.2) imply the values

$$\left.\begin{aligned}\left((H^+)^{-1}-W^+\right)_{kt} &= \left((H^+)^{-1}-W^+\right)_{tk} \\ &= v_k-v_k \;=\; 0, \qquad\qquad k=1,2,\ldots,d, \\ \left((H^+)^{-1}-W^+\right)_{ij} &= (H^{-1}-W)_{ij}, \qquad i,j\in\{1,2,\ldots,d\}\backslash\{t\}.\end{aligned}\right\} \tag{3.1}$$

Therefore the updating procedure reduces to zero all the errors in the t-th row and column of $H^{-1}-W$, while all other elements of this matrix are unchanged, except for the new errors that occur within the current iteration.

It follows that, if H is set to an arbitrary nonsingular symmetric matrix initially, if the updating formula is applied many times sequentially to H and W without a division by zero, and if the final H is nonsingular, then, in exact arithmetic, each element of the final matrix $H^{-1}-W$ is the same as the corresponding element of the initial matrix $H^{-1}-W$ or is zero. Further, $(H^{-1}-W)_{ij}$ becomes zero if and only if i and/or j is in the set $\mathcal{T}$, where $\mathcal{T}$ contains the indices t of all the applications of the updating method. This highly advantageous property is the subject of Test 5 in Section 7 of Powell (2002b), which runs for 10^5 iterations in computer arithmetic. Each matrix W is taken from the system (1.10), the values of m and n being 101 and 50, respectively. The initial H is such that the elements of the first error matrix $H^{-1}-W$ are of magnitude 10^{-3}. Then on each iteration the updating of W and H arises from a change to the position of the interpolation point $\underline{x}_t$, as mentioned in Section 1. Hence every integer t is from the interval $[1,m]$. Thus the stability property of the previous paragraph tends to correct the errors in the first m rows and columns of $H^{-1}-W$, but it does not reduce the errors in the bottom right $(n+1)\times(n+1)$ submatrix of $H^{-1}-W$. Those errors retain their magnitude of about 10^{-3} throughout the experiment. On the other hand, when the t-th interpolation point is shifted for the first time, t being an integer from $[1,m]$, then the errors in the t-th row and column of $H^{-1}-W$ become of magnitude 10^{-15}, and they remain at this level throughout the subsequent iterations of the calculation.

In this experiment, no very large new errors are introduced by an application of the updating procedure, because the current interpolation points $\underline{x}_i$, $i=1,2,\ldots,m$, are clustered round the origin on every iteration. In unconstrained

minimization calculations without derivatives, however, it is usual for the initial interpolation points to be clustered about a given initial vector of variables, while the final interpolation points are clustered about the optimal vector of variables, which may be far from the initial vector. Further, because it is inefficient to keep points in clusters when they are moved a long way, some points may be far apart and others may be close together on a typical intermediate iteration. Therefore several situations that occur in practice are not tested by the experiment of the previous paragraph. They can introduce huge errors into H^{-1}–W, an example being shown below. Recovery from such damage has to be achieved by the stability properties of the updating formula.

We consider the effects of computer rounding errors on the calculation of H^+ in a case with $n=2$ and $m=5$. Let W be the partitioned matrix of expression (1.10), when the interpolation points are the vectors

$$\underline{x}_1=\begin{pmatrix}\xi\\0\end{pmatrix},\ \ \underline{x}_2=\begin{pmatrix}\xi+\eta\\0\end{pmatrix},\ \ \underline{x}_3=\begin{pmatrix}\xi-\eta\\0\end{pmatrix},\ \ \underline{x}_4=\begin{pmatrix}\xi\\\eta\end{pmatrix}\ \ \text{and}\ \ \underline{x}_5=\begin{pmatrix}\xi\\-\eta\end{pmatrix} \tag{3.2}$$

in the definitions (1.11) and (1.12), where ξ and η are real numbers that satisfy $0<\eta<\xi$. We let H be exactly W^{-1}, we pick $t=4$, and we let $\underline{x}^+=(\xi+\eta,\eta)^T$ be the new position of $\underline{x}_t$. It follows from some algebra that H is the matrix

$$H\ =\ \left(\begin{array}{ccccc|ccc}
\frac{4}{\eta^4} & -\frac{1}{\eta^4} & -\frac{1}{\eta^4} & -\frac{1}{\eta^4} & -\frac{1}{\eta^4} & \frac{\eta^2-\xi^2}{\eta^2} & \frac{2\xi}{\eta^2} & 0\\
-\frac{1}{\eta^4} & \frac{1}{2\eta^4} & \frac{1}{2\eta^4} & 0 & 0 & \frac{\xi(\xi-\eta)}{\eta^2} & \frac{\eta-2\xi}{\eta^2} & 0\\
-\frac{1}{\eta^4} & \frac{1}{2\eta^4} & \frac{1}{2\eta^4} & 0 & 0 & \frac{\xi(\xi+\eta)}{\eta^2} & \frac{-\eta-2\xi}{\eta^2} & 0\\
-\frac{1}{\eta^4} & 0 & 0 & \frac{1}{2\eta^4} & \frac{1}{2\eta^4} & 0 & 0 & \frac{1}{2\eta}\\
-\frac{1}{\eta^4} & 0 & 0 & \frac{1}{2\eta^4} & \frac{1}{2\eta^4} & 0 & 0 & -\frac{1}{2\eta}\\
\hline
\frac{\eta^2-\xi^2}{\eta^2} & \frac{\xi(\xi-\eta)}{2\eta^2} & \frac{\xi(\xi+\eta)}{2\eta^2} & 0 & 0 & 0 & 0 & 0\\
\frac{2\xi}{\eta^2} & \frac{\eta-2\xi}{2\eta^2} & \frac{-\eta-2\xi}{2\eta^2} & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & \frac{1}{2\eta} & -\frac{1}{2\eta} & 0 & 0 & -\xi^2
\end{array}\right). \tag{3.3}$$

. Further, because $\underline{v}\in\mathcal{R}^8$ is the t-th column of W^+, we find the vectors

$$\underline{v}\ =\ \left(\begin{array}{c}
\frac{1}{2}\,(\underline{x}_1^T\underline{x}^+)^2\\
\frac{1}{2}\,(\underline{x}_2^T\underline{x}^+)^2\\
\frac{1}{2}\,(\underline{x}_3^T\underline{x}^+)^2\\
\frac{1}{2}\,\|\underline{x}^+\|^4\\
\frac{1}{2}\,(\underline{x}_5^T\underline{x}^+)^2\\
\hline
1\\
\xi+\eta\\
\eta
\end{array}\right)\ =\ \left(\begin{array}{c}
\frac{1}{2}\,(\underline{x}_1^T\underline{x}^+)^2\\
\frac{1}{2}\,(\underline{x}_2^T\underline{x}^+)^2\\
\frac{1}{2}\,(\underline{x}_3^T\underline{x}^+)^2\\
\frac{1}{2}\,(\underline{x}_4^T\underline{x}^+)^2+\gamma\\
\frac{1}{2}\,(\underline{x}_5^T\underline{x}^+)^2\\
\hline
1\\
\xi+\eta\\
\eta
\end{array}\right)\ \ \text{and}\ \ H\underline{v}\ =\ \left(\begin{array}{c}
-1-\eta^{-4}\gamma\\
1\\
0\\
1+\frac{1}{2}\eta^{-4}\gamma\\
\frac{1}{2}\eta^{-4}\gamma\\
\hline
0\\
0\\
\xi\,\eta^2+\frac{1}{2}\eta^{-1}\gamma
\end{array}\right), \tag{3.4}$$

where $\gamma=\xi^3\eta+\frac{5}{2}\xi^2\eta^2+3\xi\eta^3+\frac{3}{2}\eta^4$. Thus the parameters of the updating formula take the values

$$\left.\begin{aligned}&\alpha=\underline{e}_t^T H\,\underline{e}_t=\tfrac{1}{2}\,\eta^{-4},\qquad \beta=\underline{e}_t^T\underline{v}-\underline{v}^T H\,\underline{v}=\eta^4-2\gamma-\tfrac{1}{2}\,\eta^{-4}\gamma^2,\\&\tau=\underline{e}_t^T H\,\underline{v}=1+\tfrac{1}{2}\,\eta^{-4}\gamma\qquad\text{and}\qquad \sigma=\alpha\,\beta+\tau^2=\tfrac{3}{2}.\end{aligned}\right\}\tag{3.5}$$

We see that both the leading 5×5 submatrix of H and σ are independent of ξ. These properties are addressed in Section 5 of Powell (2002b), since a change to ξ in the example can be regarded as a shift of the origin in $\mathcal{R}^2$.

Let ε be the relative precision of the computer arithmetic. Because the values $H_{11}=4\eta^{-4}$ and $v_1=\frac{1}{2}(\xi^2+\xi\eta)^2>\frac{1}{2}\xi^4$ occur in the previous paragraph, the first component of $H\underline{v}$ includes a contribution from rounding errors of magnitude $(\xi/\eta)^4\varepsilon$. It follows from the $-v_1H_{11}v_1$ term of $\beta=\underline{e}_t^T\underline{v}-\underline{v}^TH\underline{v}$, that some errors in β are of size $(\xi^8/\eta^4)\,\varepsilon$, which causes errors in $\sigma=\alpha\beta+\tau^2$ of magnitude $(\xi/\eta)^8\varepsilon$, due to $\alpha=\frac{1}{2}\eta^{-4}$. The accuracy $\varepsilon=10^{-16}$ is typical in double precision computation. Thus we conclude that all accuracy may be lost from the denominator of the updating formula (2.12) if we pick $\eta\leq0.01\xi$ in the example.

Let $\underline{x}_*$ be the best of the current interpolation points, which means that it satisfies the condition

$$F(\underline{x}_*)\;=\;\min\{F(\underline{x}_i):i=1,2,\ldots,m\}.\tag{3.6}$$

The example suggests that intolerable loss of accuracy may occur if the length of the next change to the variables, $\underline{d}$ say, is less than $0.01\|\underline{x}_*-\underline{x}_0\|$, where $\underline{x}_0$ is the origin. Therefore $\underline{x}_0$ is altered occasionally, but each origin shift requires a revision of the matrix H that takes $\mathcal{O}(m^2n)$ operations, which, for large n, is much more onerous than all the other work of an iteration. The compromise of the algorithm is to change $\underline{x}_0$ in the case $\|\underline{d}\|\leq10^{-3/2}\|\underline{x}_*-\underline{x}_0\|$, the new position of $\underline{x}_0$ being $\underline{x}_*$. Thus most iterations are without shifts.

Each change to $\underline{x}_0$ does not reduce the damage that has been caused to H by the rounding errors of previous iterations, but it should provide less bad accuracy in the updating of H by the current and future iterations. It has been mentioned already that most of the present damage will be eliminated later by the stability property (3.1), the main exception being that all the errors in the bottom right $(n+1)\times(n+1)$ submatrix of $H^{-1}-W$ are inherited by $(H^+)^{-1}-W^+$ on every iteration. Therefore we would like these errors to be zero, which is possible in practice, by applying the factorization technique that is studied in Section 4.

Usually the new interpolation point $\underline{x}^+$ is generated by the algorithm for unconstrained optimization before picking the index t of the point that will be deleted to make room for $\underline{x}^+$. An advantage of this freedom in t is that it can provide a relatively large value of $|\sigma|$ in the updating formula (2.11), which is important to the nonsingularity of the system (1.10), as shown in Lemma 1. Therefore the algorithm calculates the value of σ that would occur for each choice of t, using

a method that is suggested by expression (3.4). Specifically, we let $\underline{w}\in\mathcal{R}^{m+n+1}$ have the components $\frac{1}{2}(\underline{x}_i^T\underline{x}^+)^2$, $i=1,2,\ldots,m$, followed by 1 and the components of $\underline{x}^+$, so the t-th column of W^+ is the vector

$$\underline{v} \;=\; \underline{w}+\gamma\,\underline{e}_t, \tag{3.7}$$

where $\gamma=\frac{1}{2}\|\underline{x}^+\|^4-\frac{1}{2}(\underline{x}_t^T\underline{x}^+)^2$. By making the substitution (3.7) in equation (2.12), Powell (2002b) derives the formula

$$\begin{aligned} H^+ \;=\; & H+\hat{\sigma}^{-1}\Big[\hat{\alpha}\,(\underline{e}_t-H\underline{w})\,(\underline{e}_t-H\underline{w})^T-\hat{\beta}\,H\,\underline{e}_t\,\underline{e}_t^T\,H \\ & +\hat{\tau}\left\{H\,\underline{e}_t\,(\underline{e}_t-H\underline{w})^T+(\underline{e}_t-H\underline{w})\,\underline{e}_t^T\,H\right\}\Big], \end{aligned} \tag{3.8}$$

with the parameters

$$\left.\begin{aligned} &\hat{\alpha}=\alpha=\underline{e}_t^T\,H\,\underline{e}_t, \qquad \hat{\beta}=\tfrac{1}{2}\,\|\underline{x}^+\|^4-\underline{w}^T\,H\,\underline{w}, \\ &\hat{\tau}=\underline{e}_t^T\,H\underline{w} \quad\text{and}\quad \hat{\sigma}=\sigma=\alpha\beta+\tau^2=\hat{\alpha}\hat{\beta}+\hat{\tau}^2. \end{aligned}\right\} \tag{3.9}$$

We see that the value of $\hat{\sigma}=\sigma$ for every t can be found in only $\mathcal{O}(m^2)$ operations, because $\underline{w}$ and $\hat{\beta}$ are independent of t. Thus the algorithm makes a choice of t that assists the nonsingularity of the system (1.10).

Another strong advantage of formula (3.8) over formula (2.12) is that the parameters $\hat{\tau}$ and $\hat{\beta}$ are usually smaller than τ and β, which avoids cancellation in the denominator $\hat{\sigma}=\sigma$. In particular, equation (3.5) shows that, in our example, the terms $\alpha\beta$ and τ^2 are both of magnitude $\gamma^2\eta^{-8}\approx(\xi/\eta)^6$, which can be huge for $0<\eta<\xi$, although $\sigma=\alpha\beta+\tau^2=\frac{3}{2}$ is independent of ξ/η in exact arithmetic. On the other hand, if formula (3.8) is applied to the example, then, because $H\underline{w}$ is the rightmost vector of expression (3.4) in the case $\gamma=0$, we deduce $\hat{\beta}=\eta^4$ and $\hat{\tau}=1$, so $\hat{\sigma}=\sigma=\hat{\alpha}\hat{\beta}+\hat{\tau}^2$ is now the sum of two positive numbers. Powell (2002b) proves that $\hat{\alpha}$ and $\hat{\beta}$ are nonnegative for general positions of the points $\underline{x}_i$, $i=1,2,\ldots,m$, and $\underline{x}^+$. Therefore expression (3.8) is preferable to formula (2.12) for updating H in the given application to unconstrained optimization.

The switch from expression (2.12) to (3.8) can also reduce the unwelcome contributions from computer rounding errors to the denominator of the updating formula, provided that one includes the factorization method of the next section. Otherwise, the following argument applies. Equation (3.7) with $t=4$ gives $v_1H_{11}v_1=w_1H_{11}w_1$ in our example, so an error of magnitude

$$v_1H_{11}v_1\,\varepsilon=4\,\eta^{-4}\,\{\tfrac{1}{2}(\underline{x}_1^T\underline{x}^+)^2\}^2\,\varepsilon=\xi^4\,(\xi+\eta)^4\,\eta^{-4}\,\varepsilon, \tag{3.10}$$

mentioned earlier, occurs not only in β but also in $\hat{\beta}$. It follows from $\alpha=\hat{\alpha}=\frac{1}{2}\eta^{-4}$ that the resultant damage to both $\sigma=\alpha\beta+\tau^2$ and $\hat{\sigma}=\hat{\alpha}\hat{\beta}+\hat{\tau}^2$ is of magnitude $(\xi/\eta)^8\varepsilon$.

4. A partial factorization of H

Equations (1.1) and (1.2) imply $X\Omega = 0$, as mentioned in Section 1. Moreover, the rows of X are linearly independent, because the matrix (1.1) is nonsingular. Therefore the rank of the $\hat{n} \times \hat{n}$ symmetric matrix Ω is at most $\hat{n} - \hat{m}$. It follows that Ω can be expressed in the form

$$\Omega \;=\; \sum_{j=1}^{\hat{n}-\hat{m}} s_j \, \underline{z}_j \, \underline{z}_j^T \;=\; Z \, S \, Z^T, \tag{4.1}$$

where each s_j is -1 or $+1$, where each $\underline{z}_j$ is in $\mathcal{R}^{\hat{n}}$, where Z is the $\hat{n} \times (\hat{n} - \hat{m})$ matrix with the columns $\underline{z}_j$, and where S is the diagonal matrix with the diagonal elements s_j, $j = 1, 2, \ldots, \hat{n} - \hat{m}$. We are going to study the factorization (4.1) of Ω. We find that in practice it is highly advantageous to store the factors instead of Ω itself, the main reason being given in the following lemma, which is valid even if much damage has been done to Z by computer rounding errors.

Lemma 3: Let H be a nonsingular symmetric matrix that has the form

$$H \;=\; \left(\begin{array}{c|c} Z \, S \, Z^T & \Xi^T \\ \hline \Xi & \Upsilon \end{array} \right) \begin{array}{l} \updownarrow \hat{n} \\ \updownarrow \hat{m} \, , \end{array} \tag{4.2}$$

where Z has only $\hat{n} - \hat{m}$ columns. Then all the elements of the bottom right $\hat{m} \times \hat{m}$ submatrix of H^{-1} are zero.

Proof: Let i and j be any integers from the interval $[\hat{n}+1, \hat{n}+\hat{m}]$, and let the matrix $C^{(ij)}$ be formed by deleting the j-th row and i-th column of H. Because of the elementary identity

$$H_{ij}^{-1} \;=\; (-1)^{i+j} \det(C^{(ij)}) \, / \det(H), \tag{4.3}$$

we ask whether $C^{(ij)}$ is singular. The first $\hat{n}$ rows of $C^{(ij)}$ give the submatrix

$$\left(\;\; Z \, S \, Z^T \;\; \middle| \;\; (\hat{m}-1) \text{ columns of } \Xi^T \;\; \right) \quad \updownarrow \hat{n} \, , \tag{4.4}$$

whose rank is bounded above by the sum

$$\text{rank} \, (Z \, S \, Z^T) + (\hat{m} - 1) \;\leq\; \hat{n} - 1. \tag{4.5}$$

It follows that the first $\hat{n}$ rows of $C^{(ij)}$ are linearly dependent. Therefore the element (4.3) of H^{-1} is zero, which completes the proof. □

The numerical results of Section 6 will show that, by using the factorization (4.1) in the new algorithm for unconstrained optimization, huge gains in efficiency are obtained in some calculations. Substantial gains are expected, because Lemma 3 implies a major improvement to the stability properties of the procedure for

updating H. Indeed, because W has the structure (1.1), no errors occur in the bottom right $\hat{m}\times\hat{m}$ submatrix of H^{-1}, as stated in the paragraph after equation (3.6), where $\hat{n}$ and $\hat{m}$ take the values m and $n{+}1$, respectively. Another advantage of the factorization is that it may reduce the damage from the new rounding errors of each application of the updating procedure. We recall the example of Section 3, in order to illustrate this remark.

We give most of our attention to the calculation of $\hat{\sigma}$ in expression (3.9), remembering that errors in $\hat{\sigma}$ of magnitude $(\xi/\eta)^8\varepsilon$ are found at the end of Section 3. We separate $\hat{\beta}=\frac{1}{2}\|\underline{x}^+\|^4-\underline{w}^TH\underline{w}$ into four parts, namely $\frac{1}{2}\|\underline{x}^+\|^4$ and the contributions to $-\underline{w}^TH\underline{w}$ from the off diagonal, bottom right and top left partitions of the matrix (3.3). We see that the damage to the first three parts from the computer arithmetic is $\mathcal{O}(\xi^4\varepsilon)$, $\mathcal{O}(\xi^6\eta^{-2}\varepsilon)$ and $\mathcal{O}(\xi^2\eta^2\varepsilon)$, respectively. The last part is $-\underline{\hat{w}}^T\Omega\,\underline{\hat{w}}$, where $\underline{\hat{w}}$ has the components $\frac{1}{2}(\underline{x}_i^T\underline{x}^+)$, $i=1,2,\ldots,m$, and the damage to this part in Section 3 is $\mathcal{O}(\xi^8\eta^{-4}\varepsilon)$. Now, however, the factorization (4.1) provides the identity

$$-\underline{\hat{w}}^T\Omega\,\underline{\hat{w}} \;=\; -\sum_{j=1}^{m-n-1} s_j\,(\underline{z}_j^T\underline{\hat{w}})^2. \tag{4.6}$$

We deduce below that, if one calculates the right hand side of this equation instead of the left hand side, then the contributions from rounding errors are smaller than before.

The product ZSZ^T in the example has to be the top left submatrix of expression (3.3). It follows that S is the 2×2 unit matrix, but Z is not unique, because ZSZ^T remains the same if Z is post-multiplied by any 2×2 orthogonal rotation. We make the choice

$$Z^T \;=\; \sqrt{2\,\eta^{-4}}\left(\begin{array}{ccccc} 1 & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} \end{array}\right), \tag{4.7}$$

which agrees with the leading part of expression (3.3). Thus computer arithmetic gives the vector

$$Z^T\underline{\hat{w}} \;=\; \sqrt{2\,\eta^{-4}}\left(\begin{array}{c} -\frac{1}{2}\eta^2(\xi{+}\eta)^2+\mathcal{O}(\xi^4\varepsilon) \\ -\frac{1}{2}\eta^4+\mathcal{O}(\xi^4\varepsilon) \end{array}\right) \;=\; \left(\begin{array}{c} \underline{z}_1^T\underline{\hat{w}}+\mathcal{O}(\xi^4\eta^{-2}\varepsilon) \\ \underline{z}_2^T\underline{\hat{w}}+\mathcal{O}(\xi^4\eta^{-2}\varepsilon) \end{array}\right). \tag{4.8}$$

Hence, by squaring the components on the right hand side, we find that the dominant part of the damage from rounding errors to the term (4.6) has the form

$$\begin{aligned} 2\sum_{j=1}^{m-n-1}|\underline{z}_j^T\underline{\hat{w}}|\;\mathcal{O}(\xi^4\eta^{-2}\varepsilon) \;&=\; \sqrt{2\,\eta^{-4}}\;\left\{\eta^2(\xi{+}\eta)^2+\eta^4\right\}\;\mathcal{O}(\xi^4\eta^{-2}\varepsilon) \\ &=\; \mathcal{O}(\xi^6\eta^{-2}\varepsilon). \end{aligned} \tag{4.9}$$

Therefore the resultant contributions to $\hat{\sigma}=\hat{\alpha}\hat{\beta}+\hat{\tau}^2$ are now of magnitude $(\xi/\eta)^6\varepsilon$. Moreover, $\hat{\tau}=\underline{e}_4^TH\underline{w}$ and the error in $\hat{\tau}$ are 1 and $\mathcal{O}(\xi^4\eta^{-4}\varepsilon)$, respectively, so the damage to $\hat{\sigma}$ from errors in $\hat{\tau}$ is negligible.

This gain in accuracy is achieved from condition (4.9), because, for each j, the scalar product $\underline{z}_j^T\underline{\hat{w}}$ on the left hand side is at most $\mathcal{O}(\xi^2)$, although it is a linear combination of terms of magnitude $\xi^4\eta^{-2}$. Thus the gain occurs also in the following generalization of the example. Let the points $\underline{x}_i$, $i=1,2,\ldots,m$, and $\underline{x}^+$ be in a cluster of radius η, and let ξ be the distance of the centre of the cluster from the origin, where $0<\eta\ll\xi$. Then the elements of the submatrix Ω of expression (1.2) are independent of ξ, and, assuming no unusual tendencies towards singularity in W, they are $\mathcal{O}(\eta^{-4})$, so the elements of Z are $\mathcal{O}(\eta^{-2})$. We write the components of $\underline{\hat{w}}$ in the form

$$\hat{w}_i=\tfrac{1}{2}(\underline{x}_i^T\underline{x}^+)^2=\tfrac{1}{2}(\underline{x}_i^T\underline{x}^+-\bar{\underline{x}}^T\underline{x}^+)^2+(\underline{x}_i^T\underline{x}^+)\,(\bar{\underline{x}}^T\underline{x}^+)-\tfrac{1}{2}(\bar{\underline{x}}^T\underline{x}^+)^2, \tag{4.10}$$

$i=1,2,\ldots,m$, where $\bar{\underline{x}}$ is the centre of the cluster. Now, according to the first paragraph of this section, Z has the property $XZ=0$, so the definition (1.12) of X gives the equations

$$\sum_{i=1}^m Z_{ij}=0\quad\text{and}\quad\sum_{i=1}^m Z_{ij}\,\underline{x}_i^T=0,\qquad i=1,2,\ldots,m-n-1. \tag{4.11}$$

Therefore expression (4.10) with $\|\underline{x}_i-\bar{\underline{x}}\|=\mathcal{O}(\eta)$ and $\|\underline{x}^+\|=\mathcal{O}(\xi)$ imply the bound

$$\begin{aligned}|\underline{z}_j^T\underline{\hat{w}}| &= \Big|\sum_{i=1}^m Z_{ij}\left\{\tfrac{1}{2}(\underline{x}_i^T\underline{x}^+-\bar{\underline{x}}^T\underline{x}^+)^2+(\underline{x}_i^T\underline{x}^+)\,(\bar{\underline{x}}^T\underline{x}^+)-\tfrac{1}{2}(\bar{\underline{x}}^T\underline{x}^+)^2\right\}\\ &= \Big|\tfrac{1}{2}\sum_{i=1}^m Z_{ij}\left((\underline{x}_i-\bar{\underline{x}})^T\underline{x}^+\right)^2\Big| \;=\; \mathcal{O}(\xi^2),\qquad j=1,2,\ldots,m-n-1,\end{aligned} \tag{4.12}$$

which shows the cancellation that is vital to the argument of the previous paragraph. Thus the improvement there to the accuracy in $\hat{\sigma}$ is achieved in the present generalization.

Furthermore, if we apply the updating formula (2.12), which is not recommended, then a small reduction in the dominant error of β can be gained by employing the right hand side instead of the left hand side of the identity

$$-\underline{\hat{v}}^T\Omega\,\underline{\hat{v}}\;=\;-\sum_{j=1}^{m-n-1}s_j\,(\underline{z}_j^T\underline{\hat{v}})^2. \tag{4.13}$$

A simple calculation gives the value $\underline{z}_2^T\underline{\hat{v}}\approx-2^{-1/2}\xi^3\eta^{-1}$ in the example of Section 3, which is unfavourable in comparison to the bound (4.12) on $\underline{z}_j^T\underline{\hat{w}}$. Therefore the damage to $\sigma=\hat{\sigma}$ from computer rounding errors is now of magnitude $(\xi/\eta)^7\varepsilon$.

5. Updating the factorization of Ω

Let the symmetric matrix $H \approx W^{-1}$ have the form (4.2), where S, Z, Ξ and Υ are available, and let W^+ be constructed as before, the t-th row and column of W being replaced by $\underline{v}^T$ and $\underline{v}$. We recall that $H^+ \approx (W^+)^{-1}$ is defined by the conditions (2.2), assuming that H and H^+ are nonsingular. Therefore H^+ can be generated by one of the updating methods that have been described already. In this section, however, we address the problem of expressing the leading $\hat{n} \times \hat{n}$ submatrix of H^+ as the product

$$\Omega^+ \;=\; \sum_{j=1}^{\hat{n}-\hat{m}} s_j^+ \, \underline{z}_j^+ \, \underline{z}_j^{+T} \;=\; Z^+ S^+ Z^{+T}, \tag{5.1}$$

in order that the bottom right $\hat{m} \times \hat{m}$ submatrix of $(H^+)^{-1}$ is identically zero, as proved in Lemma 3. Therefore we seek a convenient procedure that generates $s_j^+ = \pm 1$ and $\underline{z}_j^+ \in \mathcal{R}^{\hat{n}}$, $j = 1, 2, \ldots, \hat{n} - \hat{m}$, explicitly from S, Z, Ξ, Υ, t and $\underline{v}$.

We pick the submatrices Ξ^+ and Υ^+ of the partition

$$H^+ \;=\; \left(\begin{array}{c|c} Z^+ S^+ Z^{+T} & \Xi^{+T} \\ \hline \Xi^+ & \Upsilon^+ \end{array} \right) \begin{array}{l} \updownarrow \hat{n} \\ \updownarrow \hat{m} \end{array} \tag{5.2}$$

from equation (2.12) or (3.8), the latter formula being used by the new algorithm for unconstrained optimization. Therefore we require $H\underline{v}$ or $H\underline{w}$, respectively. These vectors are generated in the obvious way from the form (4.2) of H, taking advantage of the availability of $Z^T \underline{\hat{v}}$ or $Z^T \underline{\hat{w}}$ from the calculation of $\sigma = \hat{\sigma}$, as described in Section 4. The vector

$$H \, \underline{e}_t \;=\; \left(\begin{array}{c} \sum_{j=1}^{\hat{n}-\hat{m}} s_j \, (\underline{e}_t^T \, \underline{z}_j) \, \underline{z}_j \\ \hline \Xi \, \underline{e}_t \end{array} \right) \begin{array}{l} \updownarrow \hat{n} \\ \updownarrow \hat{m} \end{array} \tag{5.3}$$

is also required after t has been chosen from the interval $[1, \hat{n}]$, where $\underline{e}_t$ on the left and right hand sides is in $\mathcal{R}^{\hat{n}+\hat{m}}$ and $\mathcal{R}^{\hat{n}}$, respectively. We find below that we may have applied a rotation to Z that forces $\underline{e}_t^T \, \underline{z}_j$ to be zero, and then in practice we drop the j-th term from the sum of expression (5.3).

Immediately after the selection of t, we expect the scalar products $\underline{e}_t^T \, \underline{z}_j$, $j = 1, 2, \ldots, \hat{n} - \hat{m}$, to be nonzero. However, if $\underline{e}_t^T \, \underline{z}_j$ is nonzero, where $1 \le j < \hat{n} - \hat{m}$, then, by using the identity

$$\begin{aligned} \underline{a} \, \underline{a}^T + \, \underline{b} \, \underline{b}^T \;&=\; (\cos\theta \, \underline{a} + \sin\theta \, \underline{b}) \, (\cos\theta \, \underline{a} + \sin\theta \, \underline{b})^T \\ &\quad + (\cos\theta \, \underline{b} - \sin\theta \, \underline{a}) \, (\cos\theta \, \underline{b} - \sin\theta \, \underline{a})^T, \quad \theta \in \mathcal{R}, \;\; \underline{a}, \underline{b} \in \mathcal{R}^{\hat{n}}, \end{aligned} \tag{5.4}$$

or

$$\begin{aligned} \underline{a} \, \underline{a}^T - \, \underline{b} \, \underline{b}^T \;&=\; (\cosh\theta \, \underline{a} + \sinh\theta \, \underline{b}) \, (\cosh\theta \, \underline{a} + \sinh\theta \, \underline{b})^T \\ &\quad - (\cosh\theta \, \underline{b} + \sinh\theta \, \underline{a}) \, (\cosh\theta \, \underline{b} + \sinh\theta \, \underline{a})^T, \quad \theta \in \mathcal{R}, \;\; \underline{a}, \underline{b} \in \mathcal{R}^{\hat{n}}, \end{aligned} \tag{5.5}$$

in the case $s_j = s_{\hat{n}-\hat{m}}$ or $s_j = -s_{\hat{n}-\hat{m}}$, respectively, we can overwrite $\underline{z}_{\hat{n}-\hat{m}}$ and $\underline{z}_j$ by vectors that preserve the matrix $\Omega = ZSZ^T$, and usually we can choose θ so that $\underline{e}_t^T \underline{z}_j$ becomes zero. Thus we may satisfy the conditions

$$\underline{e}_t^T \underline{z}_j = 0, \qquad j = 1, 2, \ldots, \hat{n}-\hat{m}-1. \tag{5.6}$$

The reason for doing so is not to simplify expression (5.3), but because the following lemma suggests a highly convenient procedure for generating the required factorization (5.1).

Lemma 4: Let the conditions of the theorem at the beginning of Section 2 be satisfied, let H have the form (4.2), where Z has only $\hat{n}-\hat{m}$ columns, and where S is a diagonal matrix as in equation (4.1), and let t be from the interval $[1, \hat{n}]$. Then the leading $\hat{n} \times \hat{n}$ submatrix of H^+ has a factorization of the form (5.1). Further, if Z has the property (5.6), then the first $\hat{n}-\hat{m}-1$ columns of Z^+ and diagonal elements of S^+ can be given the values

$$\underline{z}_j^+ = \underline{z}_j \qquad \text{and} \qquad s_j^+ = s_j, \qquad j = 1, 2, \ldots, \hat{n}-\hat{m}-1, \tag{5.7}$$

respectively. In this case, the final column of Z^+ is the vector

$$\underline{z}_{\hat{n}-\hat{m}}^+ \;=\; \pm \left| (\underline{e}_t - \tilde{H}\,\underline{v})^T \underline{v} \right|^{-1/2} \text{chop}\,(\underline{e}_t - \tilde{H}\,\underline{v}), \tag{5.8}$$

where the choice of sign can be made later, where $\tilde{H}$ is the matrix (2.4), and where the components of $\text{chop}\,(\underline{e}_t - \tilde{H}\,\underline{v}) \in \mathcal{R}^{\hat{n}}$ are the first $\hat{n}$ components of $\underline{e}_t - \tilde{H}\underline{v}$. The corresponding diagonal element of S^+ is $s_{\hat{n}-\hat{m}}^+ = \pm 1$, its sign being the same as the sign of the denominator $(\underline{e}_t - \tilde{H}\underline{v})^T \underline{v}$ of formula (2.5).

Proof: The conditions of the theorem include the nonsingularity of H. It follows from Lemma 3 that the bottom right $\hat{m} \times \hat{m}$ submatrix of H^{-1} is zero. The second part of expression (2.2) shows that $(H^+)^{-1}$ has this property too, the value of t being from the interval $[1, \hat{n}]$. Therefore the existence of the factorization (5.1) is established by the argument in the first paragraph of Section 4.

Let H^+ be generated by the method of the theorem of Section 2, which is possible because of the assumptions (2.3) and $H_{tt} \neq 0$, and let $\tilde{\Omega}$ be the leading $\hat{n} \times \hat{n}$ submatrix of $\tilde{H}$. Then equations (2.4), (4.1), (5.6) and (5.7) give the formula

$$\begin{aligned} \tilde{\Omega} \;=\; \Omega - \frac{\Omega\,\underline{e}_t\,\underline{e}_t^T\,\Omega}{\underline{e}_t^T\,\Omega\,\underline{e}_t} \;&=\; \sum_{j=1}^{\hat{n}-\hat{m}} s_j\,\underline{z}_j\,\underline{z}_j^T - \frac{s_{\hat{n}-\hat{m}}^2\,(\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})^2\,\underline{z}_{\hat{n}-\hat{m}}\,\underline{z}_{\hat{n}-\hat{m}}^T}{s_{\hat{n}-\hat{m}}\,(\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})^2} \\ &=\; \sum_{j=1}^{\hat{n}-\hat{m}-1} s_j\,\underline{z}_j\,\underline{z}_j^T \;=\; \sum_{j=1}^{\hat{n}-\hat{m}-1} s_j^+\,\underline{z}_j^+\,\underline{z}_j^{+\,T}. \end{aligned} \tag{5.9}$$

We see that Ω^+ is the sum of the matrix (5.9) and the leading $\hat{n} \times \hat{n}$ submatrix of the rank one term of expression (2.5). In other words, because of the choices of

$\underline{z}^+_{\hat{n}-\hat{m}}$ and $s^+_{\hat{n}-\hat{m}}$ in and just after equation (5.8), Ω^+ is the sum in the middle of expression (5.1), as required. $\square$

We derive a convenient way of updating Z in the case (5.6) by eliminating $\tilde{H}$ from the construction in Lemma 4. Equations (4.1) and (5.6) provide the formulae

$$\left.\begin{aligned} \Omega\,\underline{e}_t &= s_{\hat{n}-\hat{m}}\,(\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})\,\underline{z}_{\hat{n}-\hat{m}} \\ \alpha = H_{tt} &= \Omega_{tt} = s_{\hat{n}-\hat{m}}\,(\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})^2 \end{aligned}\right\}, \tag{5.10}$$

as in the first line of expression (5.9). Therefore the definition (2.4) of $\tilde{H}$ gives the vector

$$\begin{aligned} \mathrm{chop}\,(\underline{e}_t-\tilde{H}\,\underline{v}) &= \mathrm{chop}\,(\underline{e}_t-H\,\underline{v}) + (\underline{e}_t^T H\,\underline{v})\,\alpha^{-1}\,\Omega\,\underline{e}_t \\ &= (\underline{e}_t^T H\,\underline{v})\,(\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})^{-1}\underline{z}_{\hat{n}-\hat{m}} + \mathrm{chop}\,(\underline{e}_t-H\,\underline{v}) \end{aligned} \tag{5.11}$$

and the scalar product

$$(\underline{e}_t-\tilde{H}\,\underline{v})^T\underline{v} = (\underline{e}_t-H\,\underline{v})^T\underline{v} + \alpha^{-1}\,(\underline{e}_t^T H\,\underline{v})^2 = \alpha^{-1}\sigma, \tag{5.12}$$

where the last equation depends on the values (2.13) and $\sigma=\alpha\beta+\tau^2$. It follows from the definition (5.8) and conditions (5.10)–(5.12) that $\underline{z}^+_{\hat{n}-\hat{m}}$ can be written in the form

$$\begin{aligned} \underline{z}^+_{\hat{n}-\hat{m}} &= \pm|\alpha|^{1/2}\,|\sigma|^{-1/2}\,\left\{(\underline{e}_t^T H\,\underline{v})\,(\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})^{-1}\underline{z}_{\hat{n}-\hat{m}} + \mathrm{chop}\,(\underline{e}_t-H\,\underline{v})\right\} \\ &= |\sigma|^{-1/2}\left\{\tau\,\underline{z}_{\hat{n}-\hat{m}} + (\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})\,\mathrm{chop}\,(\underline{e}_t-H\,\underline{v})\right\}, \end{aligned} \tag{5.13}$$

by making the choice $\pm|\alpha|^{1/2}=\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}}$. Further, because the signs of $s^+_{\hat{n}-\hat{m}}$ and $(\underline{e}_t-\tilde{H}\,\underline{v})^T\underline{v}=\alpha^{-1}\sigma$ have to be the same, it follows from the last part of expression (5.10) that the sign of $s^+_{\hat{n}-\hat{m}}$ is opposite to the sign of $s_{\hat{n}-\hat{m}}$ if and only if σ is negative. The use of equations (5.7) and (5.13) for updating Z is recommended, one reason being that again the division by $\alpha=\underline{e}_t^T H\,\underline{e}_t$, which occurs in the theorem, has been removed.

When the partitions Ξ^+ and Υ^+ of H^+ are obtained from formula (3.8), instead of from formula (2.12), we prefer to put $\underline{v}=\underline{w}+\gamma\,\underline{e}_t$ into expression (5.13). The conditions (5.10) show that the term $(\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})\,\mathrm{chop}\,(-\gamma H\,\underline{e}_t)$ is just $-\alpha\,\gamma\,\underline{z}_{\hat{n}-\hat{m}}$, and equations (2.13), (3.7) and (3.9) imply that $\tau-\alpha\,\gamma$ has the value $\underline{e}_t^T H(\underline{v}-\gamma\underline{e}_t)=\underline{e}_t^T H\,\underline{w}=\hat{\tau}$. Thus, recalling $\hat{\sigma}=\sigma$, we can write the vector (5.13) in the form

$$\underline{z}^+_{\hat{n}-\hat{m}} = |\hat{\sigma}|^{-1/2}\left\{\hat{\tau}\,\underline{z}_{\hat{n}-\hat{m}} + (\underline{e}_t^T\,\underline{z}_{\hat{n}-\hat{m}})\,\mathrm{chop}\,(\underline{e}_t-H\,\underline{w})\right\}. \tag{5.14}$$

The method of the last two paragraphs has another reassuring property in the situation when $\alpha=\hat{\alpha}$ is zero. This possibility is unlikely in practice with the conditions (5.6), because then all of the terms $\underline{e}_t^T\,\underline{z}_j$, $j=1,2,\ldots,\hat{n}-\hat{m}$, are zero, so the t-th row and column of Ω are identically zero. In this case, $\sigma=\hat{\sigma}$

has the positive value $\alpha\beta+\tau^2=\tau^2$ or $\hat{\alpha}\hat{\beta}+\hat{\tau}^2=\hat{\tau}^2$ in equation (5.13) or (5.14), respectively. Thus the method gives $\underline{z}_{\hat{n}-\hat{m}}^+ = \pm\underline{z}_{\hat{n}-\hat{m}}$ and $s_{\hat{n}-\hat{m}}^+ = s_{\hat{n}-\hat{m}}$, which implies $\Omega^+=\Omega$. Now the conditions (5.6) are preserved if the columns of Z are permuted. Therefore it is reassuring that $\Omega^+=\Omega$ is independent of the particular column of Z that is chosen to be last. In the present case, α, $\hat{\alpha}$ and the first $\hat{n}$ components of $H\underline{e}_t$ are all zero in formulae (2.12) and (3.8), so these formulae confirm that $\Omega^+=\Omega$ should occur.

We have to give further attention to the method for updating Z, because $\alpha=\hat{\alpha}=\Omega_{tt}$ may be zero with nonzero elements elsewhere in the t-th row and column of Ω. Then the conditions (5.6) cannot be obtained by rotations of the form (5.4) and (5.5), the trouble in exact arithmetic being the choice of θ in equation (5.5) in the case $|\underline{e}_t^T\underline{a}| = |\underline{e}_t^T\underline{b}| \neq 0$. Indeed, neither $\cosh\theta\,\underline{a}+\sinh\theta\,\underline{b}$ nor $\cosh\theta\,\underline{b}+\sinh\theta\,\underline{a}$ can have a zero t-th component if θ is finite, and one of these vectors is intended to overwrite $\underline{z}_j$. Another difficulty in practice is that the proposed use of formula (5.5) tends to magnify any errors in $\underline{z}_j$ and $\underline{z}_{\hat{n}-\hat{m}}$. Therefore the following way of updating Z without this formula is recommended. If $\underline{e}_t^T\underline{z}_j$ is nonzero, and if $s_k=s_j$ holds for $k\neq j$, then, by applying the Givens rotation (5.4) to $\underline{z}_j$ and $\underline{z}_k$, we can preserve ZSZ^T and achieve $\underline{e}_t^T\underline{z}_j=0$. Thus all but one or two of the elements $\underline{e}_t^T\underline{z}_j$, $j=1,2,\ldots,\hat{n}-\hat{m}$, are made zero in a stable way. The number is one if all the signs s_j, $j=1,2,\ldots,\hat{n}-\hat{m}$, are the same, and then the updating of Z is completed by the procedure that has been described already. The alternative case is addressed in the remainder of this section. For ease of notation, we assume that the rotations (5.4) have provided the conditions

$$\underline{e}_t^T\underline{z}_j \;=\; 0, \qquad j=3,4,\ldots,\hat{n}-\hat{m}, \tag{5.15}$$

and that the values of s_1 and s_2 are $+1$ and -1, respectively. This situation can occur in the application to unconstrained optimization, but it would be due to computer rounding errors, because, when Ω has rank $\hat{n}-\hat{m}$ and no negative eigenvalues, then all the signs s_j in the factorization (4.1) are positive.

Let the conditions (5.15) hold with $s_1=+1$ and $s_2=-1$. If $|\underline{e}_t^T\underline{z}_1| \neq |\underline{e}_t^T\underline{z}_2|$ occurs, then, by combining a rotation of the form (5.5) with the procedure given earlier, we deduce that both $\underline{z}_1^+$ and $\underline{z}_2^+$ are in the linear space spanned by $\underline{z}_1$, $\underline{z}_2$ and $\mathrm{chop}\,(\underline{e}_t-H\underline{v})$. By continuity, we also expect this property in the case $|\underline{e}_t^T\underline{z}_1|=|\underline{e}_t^T\underline{z}_2|$. Therefore we pick the values

$$\underline{z}_j^+ = \underline{z}_j \qquad \text{and} \qquad s_j^+ = s_j, \qquad j=3,4,\ldots,\hat{n}-\hat{m}, \tag{5.16}$$

for the factorization (5.1). The corresponding values of $\underline{z}_1^+$, $\underline{z}_2^+$, s_1^+ and s_2^+ are addressed in the lemma below, two alternatives being presented, in order to avoid cancellation when the parameter β of formula (2.12) is positive or negative. The identity (5.5) implies that the number of suitable choices of $\underline{z}_1^+$ and $\underline{z}_2^+$ is infinite.

Lemma 5: Let the conditions of the first sentence of Lemma 4 be satisfied with the equations (5.15), let s_1 and s_2 be $+1$ and -1, respectively, and let H^+ be

defined by formula (2.12), where $\alpha = \underline{e}_t^T H \underline{e}_t$ and $\sigma = \alpha\beta + \tau^2$, the value of σ being nonzero. Then the factorization (5.1) of the leading $\hat{n} \times \hat{n}$ submatrix of H^+ allows the values (5.16). The choices

$$\left.\begin{aligned} \underline{z}_1^+ &= |\tau^2 + \beta\xi_1^2|^{-1/2} \left\{ \tau \underline{z}_1 + \xi_1 \underline{u} \right\} \\ \underline{z}_2^+ &= |\sigma(\tau^2 + \beta\xi_1^2)|^{-1/2} \left\{ -\beta\xi_1\xi_2\underline{z}_1 + (\tau^2 + \beta\xi_1^2)\underline{z}_2 + \tau\xi_2\underline{u} \right\} \end{aligned}\right\} \tag{5.17}$$

or the choices

$$\left.\begin{aligned} \underline{z}_1^+ &= |\sigma(\tau^2 - \beta\xi_2^2)|^{-1/2} \left\{ (\tau^2 - \beta\xi_2^2)\underline{z}_1 + \beta\xi_1\xi_2\underline{z}_2 + \tau\xi_1\underline{u} \right\} \\ \underline{z}_2^+ &= |\tau^2 - \beta\xi_2^2|^{-1/2} \left\{ \tau \underline{z}_2 + \xi_2 \underline{u} \right\} \end{aligned}\right\} \tag{5.18}$$

can also be made, where $\xi_1 = \underline{e}_t^T \underline{z}_1$, $\xi_2 = \underline{e}_t^T \underline{z}_2$ and $\underline{u} = \mathrm{chop}\,(\underline{e}_t - H\underline{v})$. If expression (5.17) is preferred in the case $\beta \geq 0$, then s_1^+ and s_2^+ take the values $+1$ and $-\mathrm{sign}\,\sigma$, respectively. Their values in the alternative case (5.18) with $\beta \leq 0$ are $s_1^+ = \mathrm{sign}\,\sigma$ and $s_2^+ = -1$.

Proof: Equations (2.12), (4.1), (5.1) and (5.16) imply that it is sufficient to establish the identity

$$\begin{aligned} s_1^+ \underline{z}_1^+ \underline{z}_1^{+T} + s_2^+ \underline{z}_2^+ \underline{z}_2^{+T} &= \underline{z}_1 \underline{z}_1^T - \underline{z}_2 \underline{z}_2^T \\ &\quad + \sigma^{-1} \Big\{ \alpha\, \underline{u}\, \underline{u}^T - \beta\Omega\, \underline{e}_t\, \underline{e}_t^T \Omega + \tau(\Omega\, \underline{e}_t\, \underline{u}^T + \underline{u}\, \underline{e}_t^T \Omega) \Big\}. \end{aligned} \tag{5.19}$$

The conditions (5.15), with $s_1 = +1$ and $s_2 = -1$, provide the values

$$\Omega\, \underline{e}_t = \xi_1\, \underline{z}_1 - \xi_2\, \underline{z}_2 \qquad \text{and} \qquad \alpha = \underline{e}_t^T \Omega\, \underline{e}_t = \xi_1^2 - \xi_2^2. \tag{5.20}$$

It follows from $\sigma = \tau^2 + (\xi_1^2 - \xi_2^2)\beta$ that the right hand side of expression (5.19) is σ^{-1} times the matrix

$$\begin{aligned} &(\xi_1^2 - \xi_2^2)\, \underline{u}\, \underline{u}^T + \tau\xi_1 (\underline{z}_1\, \underline{u}^T + \underline{u}\, \underline{z}_1^T) - \tau\xi_2 (\underline{z}_2\, \underline{u}^T + \underline{u}\, \underline{z}_2^T) \\ &\qquad + (\tau^2 - \beta\xi_2^2)\, \underline{z}_1\, \underline{z}_1^T + \beta\xi_1\xi_2 (\underline{z}_1\, \underline{z}_2^T + \underline{z}_2\, \underline{z}_1^T) - (\tau^2 + \beta\xi_1^2)\, \underline{z}_2\, \underline{z}_2^T. \end{aligned} \tag{5.21}$$

Moreover, in the case (5.17) with the signs $s_1^+ = +1$, $s_2^+ = -\mathrm{sign}\,\sigma$ and $\tau^2 + \beta\xi_1^2 > 0$, the left hand side of expression (5.19) is σ^{-1} times the matrix

$$\begin{aligned} &(\tau^2 + \beta\xi_1^2)^{-1} \Big[(\tau^2 + \beta\xi_1^2 - \beta\xi_2^2) \Big\{ \tau \underline{z}_1 + \xi_1 \underline{u} \Big\} \Big\{ \tau \underline{z}_1 + \xi_1 \underline{u} \Big\}^T - \Big\{ -\beta\xi_1\xi_2 \underline{z}_1 \\ &\qquad + (\tau^2 + \beta\xi_1^2)\, \underline{z}_2 + \tau\xi_2 \underline{u} \Big\} \Big\{ -\beta\xi_1\xi_2 \underline{z}_1 + (\tau^2 + \beta\xi_1^2)\, \underline{z}_2 + \tau\xi_2 \underline{u} \Big\}^T \Big], \end{aligned} \tag{5.22}$$

and in the case (5.18) with the signs $s_1^+ = \mathrm{sign}\,\sigma$, $s_2^+ = -1$ and $\tau^2 - \beta\xi_2^2 > 0$, the left hand side is σ^{-1} times the matrix

$$\begin{aligned} &(\tau^2 - \beta\xi_2^2)^{-1} \Big[-(\tau^2 + \beta\xi_1^2 - \beta\xi_2^2) \Big\{ \tau \underline{z}_2 + \xi_2 \underline{u} \Big\} \Big\{ \tau \underline{z}_2 + \xi_2 \underline{u} \Big\}^T + \Big\{ (\tau^2 - \beta\xi_2^2)\, \underline{z}_1 \\ &\qquad + \beta\xi_1\xi_2\, \underline{z}_2 + \tau\xi_1 \underline{u} \Big\} \Big\{ (\tau^2 - \beta\xi_2^2)\, \underline{z}_1 + \beta\xi_1\xi_2\, \underline{z}_2 + \tau\xi_1 \underline{u} \Big\}^T \Big]. \end{aligned} \tag{5.23}$$

By comparing the multipliers of each of the terms $\underline{u}\,\underline{u}^T$, $\underline{z}_1\underline{u}^T+\underline{u}\,\underline{z}_1^T$, $\underline{z}_2\underline{u}^T+\underline{u}\,\underline{z}_2^T$, $\underline{z}_1\underline{z}_1^T$, $\underline{z}_1\underline{z}_2^T+\underline{z}_2\underline{z}_1^T$ and $\underline{z}_2\underline{z}_2^T$, we find that the matrices (5.21), (5.22) and (5.23) are the same. Therefore the lemma is true. $\Box$

We can replace α, β, τ, σ and $\underline{v}$ by $\hat{\alpha}$, $\hat{\beta}$, $\hat{\tau}$, $\hat{\sigma}$ and $\underline{w}$ throughout the statement and proof of Lemma 5, the definition of H^+ being changed from formula (2.12) to formula (3.8). Thus the lemma is useful to the new algorithm for unconstrained minimization, if a practical failure occurs in the theoretical property $s_i = +1$, $i=1,2,\ldots,\hat{n}$.

6. Numerical results and discussion

Only one table of numerical results is presented. It illustrates three of the main stages of the research of the author in the last five years on algorithms for minimization without derivatives. The first stage provided the UOBYQA software (Powell, 2002a), which defines all the parameters of every quadratic model by interpolation to values of the objective function. Therefore the number of interpolation conditions is $\frac{1}{2}(n+1)(n+2)$, so the amount of routine work of each iteration is $\mathcal{O}(n^4)$. Thus UOBYQA is unsuitable for large numbers of variables. Then the second stage began with the challenge of deriving suitable updates of quadratic models from far fewer interpolation conditions. The least Frobenius norm method, described in Section 1, was tried, and it performed brilliantly. There were not only huge reductions in the amount of work for $n \geq 20$, but also an unconstrained problem with $n = 160$ was solved to high accuracy using fewer than 10000 values of the objective function, although UOBYQA would require $\frac{1}{2}(n+1)(n+2) = 13041$ values to construct its first quadratic model. Whenever the author has discovered techniques of this importance to practical algorithms on previous occasions, he has developed Fortran software that makes the discoveries available for general use. His efforts to do so again were frustrated for 18 months by loss of efficiency due to computer rounding errors, as shown in the example of Section 3. That loss is addressed in our discussion of Table 1 below, the relevant results being obtained by the version of the new algorithm that is the subject of Powell (2003). The factorization method of Section 4 was introduced in June, 2003. It provided the jump from stage two to stage three, by making tolerable the damage from rounding errors in difficult situations, which is also illustrated in Table 1. The latest Fortran software, namely NEWUOA, is available free of charge from the author at the e-mail address mjdp@cam.ac.uk.

We apply the software that is mentioned above to the objective function

$$F(\underline{x}) \;=\; \sum_{i=1}^{2n} \Big\{ b_i - \sum_{j=1}^{n} \Big(C_{ij} \cos(\theta_j x_j) + S_{ij} \sin(\theta_j x_j) \Big) \Big\}^2, \qquad \underline{x} \in \mathcal{R}^n, \qquad (6.1)$$

in the case $n = 40$, where each C_{ij} and S_{ij} is a random integer from $[-100, 100]$,

	UOBYQA	Powell (2003)		NEWUOA	
	$m=861$	$m=81$	$m=861$	$m=81$	$m=861$
Problem 1	2181	2345	11335	1886	2216
Problem 2	2539	2295	3945	1601	2120
Problem 3	3035	2298	3205	1595	2199
Problem 4	2305	2469	6098	2062	2002
Problem 5	2028	2062	4412	1838	1996
$\max F(\underline{x}_*)$	4.6×10^{-9}	3.0×10^{-8}	9.0×10^{-10}	1.5×10^{-7}	1.1×10^{-9}
$\max \|\underline{x}_*-\underline{x}^*\|_\infty$	7.6×10^{-7}	2.0×10^{-6}	3.1×10^{-7}	5.4×10^{-6}	3.6×10^{-7}
Average time	814 secs	62 secs	3423 secs	37 secs	1162 secs

Table 1: Numbers of function values etc in the case (6.1) with $n=40$

each θ_j is chosen randomly from the logarithmic distribution on $[0.1, 1]$, and the parameters b_i, $i=1,2,\ldots,2n$, are defined by $F(\underline{x}^*)=0$, where the component x_j^* of $\underline{x}^*$ is picked randomly from the uniform distribution on $[-\pi/\theta_j, \pi/\theta_j]$ for $j=1,2,\ldots,n$. An initial vector of variables, $\underline{x}^0$ say, is required by all versions of the software, and its j-th component is taken at random from the uniform distribution on $[x_j^*-0.1\pi/\theta_j,\ x_j^*+0.1\pi/\theta_j]$ for $j=1,2,\ldots,n$. The software also requires initial and final values of a trust region radius, which are set to $\rho_{\rm beg}=10^{-1}$ and $\rho_{\rm end}=10^{-6}$, respectively. The number of interpolation conditions, namely the integer m of equation (1.3), is $m=\frac{1}{2}(n+1)(n+2)=861$ in UOBYQA, but otherwise it has to be prescribed. We pick both $m=81$ and $m=861$ for the software of stages two and three. Thus five different methods provide the results of the numerical experiments. Only five different sets of random numbers were tried, each set giving a particular objective function (6.1) and a particular starting point $\underline{x}^0$, which are used throughout the relevant row of Table 1. The calculations were run in double precision arithmetic on a Sun Ultra 10 workstation.

The numbers of function evaluations that occur when each of the five test problems is solved by the five different methods are shown separately in the main part of Table 1. We see that the NEWUOA software compares favourably with the other implementations, and that fewer function values are usual when m is decreased from 861 to 81. Other experiments by the author have supported these findings for $n\geq 40$, especially when n is large, but they are not reported here, because the most important feature of the table to our work is the severe inefficiency of the Powell (2003) algorithm in the case $m=861$. The crucial point is that in theory the least Frobenius norm updating technique is redundant if m has the value $\frac{1}{2}(n+1)(n+2)$, because there is no freedom in a quadratic model

that satisfies the current interpolation conditions. Thus, apart from the effects of computer rounding errors, the performance of the Powell (2003) algorithm with $n = 40$ and $m = 861$ should be similar to that of UOBYQA. Therefore one of the main aims during the development of the more recent software was to find techniques for auxiliary tasks that would provide the efficiency of UOBYQA in the case $m = \frac{1}{2}(n+1)(n+2)$. Particular attention was given to the frequency of shifts of the origin $\underline{x}_0$, mentioned in Section 3. Those attempts failed miserably during the second stage of the work, however, so the last column of Table 1 provides excellent motivation for the use of the factorization (4.1).

The last three rows of Table 1 show the accuracy and running times of the calculations. Here $\underline{x}_*$ is the final vector of variables that is returned by the software, and we recall that $\underline{x}^*$ is the optimal vector of variables, the parameters b_i, $i = 1, 2, \ldots, 2n$, being defined by $F(\underline{x}^*) = 0$. The maximum and average values in these rows are taken over the five test problems. The magnitudes of $F(\underline{x}_*)$ and $\|\underline{x}_* - \underline{x}^*\|_\infty$ are very acceptable, because $\rho_{\rm end} = 10^{-6}$ is a lower bound on the radii of the trust regions that are chosen automatically. The good accuracy in the $m = 861$ column of Powell (2003) is particularly welcome, because it demonstrates the success of the stability properties of Section 3 in recovering from large errors in the approximation $H \approx W^{-1}$. Those errors were investigated by monitoring the signs of the diagonal elements of Ω, which should all be positive. In Problem 1, where 11335 function values are required, the updating formula (3.8) was applied 10473 times, 3 negative signs were introduced, and they survived for 8, 58 and 3 consecutive iterations. The times in the last row confirm that huge savings can be achieved by the new software, provided that n is sufficiently large and the calculation of the objective function is relatively cheap. Here the advantage of NEWUOA over Powell (2003), in the $m = 81$ columns of the table, is due mainly to the inclusion of a truncated conjugate gradient method, instead of an $\mathcal{O}(n^3)$ procedure, for solving the trust region subproblems that occur.

The positive semi-definiteness of Ω is important in theory, but some violations of the conditions $\Omega_{ii} \geq 0$, $i = 1, 2, \ldots, m$, are mentioned in the previous paragraph. Therefore we ask briefly whether it may be advantageous to alter computed numbers if such violations become obvious in practice. A technique of this kind is proposed in formula (7.5) of Powell (2003), namely the replacement of $\hat{\sigma}$ in our expression (3.9) by the value

$$\hat{\sigma} \;=\; \max[\hat{\alpha}, 0]\, \max[\hat{\beta}, 0] + \hat{\tau}^2, \tag{6.2}$$

because $\hat{\alpha}$ and $\hat{\beta}$ should be nonnegative, but the calculated values of $\hat{\alpha}$ and $\hat{\beta}$ are retained in the numerator of equation (3.8). Experiments on the device (6.2) during stage two of the development of the new software were highly unfavourable. Moreover, by changing $\hat{\beta}$ and then defining $\hat{\sigma} = \hat{\alpha}\hat{\beta} + \hat{\tau}^2$, as suggested in Lemma 2, the author was unable to avoid the inefficiency that is shown by the large numbers in the middle column of Table 1. With the benefit of hindsight, we take

the view now that, if a need for the modification of parameters is detectable, then substantial errors must have occurred already that require attention. We rely on the stability property, described in the second paragraph of Section 3, to remove old errors automatically. Therefore we should preserve the useful feature that the t-th rows and columns of $(H^+)^{-1}$ and W^+ are the same, although H^{-1} may be very different from W. It follows that changes to $\hat{\beta}$ and other parameters are not recommended, even if it is clear from the factorization (5.1) that Ω^+ is going to have a negative eigenvalue.

The difficulties addressed in this paper are due to working with sequences of inverse matrices. If W is nearly singular, then $H = W^{-1}$ is often the sum of a matrix of low rank with huge elements and another matrix with much smaller elements that is important. Then the rounding errors of the first matrix damage the elements of the second matrix severely. Therefore the standard advice of many numerical analysts is to employ factorizations of matrices instead of storing and updating their inverses. The new algorithms, however, do require coefficients of Lagrange functions, and those coefficients are particular elements of $H = W^{-1}$. The author has not investigated whether it would be less painful to derive them from a factorization of W. He does not intend to do so, because, due to the good performance of the NEWUOA software in practice, his current research on least Frobenius norm updating in unconstrained minimization without derivatives may be nearly complete.

References

J.E. Dennis and R.B. Schnabel (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall (Englewood Cliffs).

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

M.J.D. Powell (2002a), "UOBYQA: unconstrained optimization by quadratic approximation", *Math. Programming*, Vol. 92, pp. 555–582.

M.J.D. Powell (2002b), "Least Frobenius norm updating of quadratic models that satisfy interpolation conditions", Report No. DAMTP 2002/NA08, University of Cambridge (to be published in *Math. Programming*).

M.J.D. Powell (2003), "On the use of quadratic models in unconstrained minimization without derivatives", Report No. DAMTP 2003/NA03, University of Cambridge (to be published in *Optimization Methods and Software*).