SIAM J. MATRIX ANAL. APPL.
Vol. 23, No. 2, pp. 303–316

SOLVING A QUADRATIC MATRIX EQUATION BY NEWTON'S METHOD WITH EXACT LINE SEARCHES*

NICHOLAS J. HIGHAM† AND HYUN-MIN KIM†

Abstract. We show how to incorporate exact line searches into Newton's method for solving the quadratic matrix equation $AX^2 + BX + C = 0$, where A, B and C are square matrices. The line searches are relatively inexpensive and improve the global convergence properties of Newton's method in theory and in practice. We also derive a condition number for the problem and show how to compute the backward error of an approximate solution.

Key words. quadratic matrix equation, solvent, Newton's method, generalized Sylvester equation, exact line searches, quadratic eigenvalue problem, condition number, backward error

AMS subject classifications. 65F30, 65H10

PII. S0895479899350976

1. Introduction.

Nonlinear matrix equations occur in a variety of applications. An important class of examples, arising in control theory, is algebraic Riccati equations, such as $XBX + XA + A^*X + C = 0$, where A, B, and C are given coefficient matrices. Theory of Riccati equations and numerical methods for their solution are well developed [1], [4], [31]. Our interest here is in the quadratic matrix equation

$$Q(X) = AX^2 + BX + C = 0, \qquad A, B, C \in \mathbb{C}^{n\times n}. \tag{1.1}$$

Although some Riccati equations are quadratic matrix equations, and vice versa, the two classes of equations require different techniques for analysis and solution in general.

Motivation for studying the quadratic matrix equation comes from the quadratic eigenvalue problem

$$Q(\lambda)x = \lambda^2 Ax + \lambda Bx + Cx = 0, \qquad A, B, C \in \mathbb{C}^{n\times n}, \tag{1.2}$$

which arises in the analysis of structural systems and vibration problems [30], [36], [37]. The standard approach is to reduce (1.2) to a generalized eigenproblem (GEP) $Gx = \lambda Hx$ of twice the dimension, $2n$. However, as is well known [7], [10], [30], if we can find a solution X of the associated quadratic matrix equation (1.1) then we can write

$$\lambda^2 A + \lambda B + C = -(B + AX + \lambda A)(X - \lambda I) \tag{1.3}$$

and so the eigenvalues of (1.2) are those of X together with those of the GEP $(B + AX)x = -\lambda Ax$, both of which are $n \times n$ problems. Bridges and Morris [5] employ this approach in the solution of differential eigenproblems.

A solution X of (1.1) is called a solvent [10]. More precisely, X is called a right solvent to distinguish it from a left solvent, which is a solution of $X^2A + XB + C = 0$.

*Received by the editors February 2, 1999; accepted for publication (in revised form) by A. Bunse-Gerstner June 9, 2000; published electronically August 8, 2001.
http://www.siam.org/journals/simax/23-2/35097.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/~higham/, kim@ma.man.ac.uk, http://www.ma.man.ac.uk/~kim/). The work of the first author was supported by Engineering and Physical Sciences Research Council grant GR/L76532.

Transposing the latter equation yields one of the form (1.1), so we concentrate on (1.1) here.

A dominant (minimal) solvent X is one for which every eigenvalue is greater (less than) in modulus than all the eigenvalues of the quotient $B + AX + \lambda A$ in (1.3). In earlier work, Dennis, Traub, and Weber gave two linearly convergent algorithms for computing a dominant solvent of an arbitrary degree matrix polynomial [11]. One of these is a generalization of Bernoulli's method for scalar polynomials and is also described by Gohberg, Lancaster, and Rodman [20, sec. 4.2]. These algorithms have the drawbacks that it is difficult to check in advance whether a dominant solvent exists and the convergence can be extremely slow (see [27] for more details). Davis [7], [8] applied Newton's method to the quadratic matrix equation, giving supporting theory and implementation details. Kratz and Stickel [29] investigated Newton's method for the general matrix polynomial.

This work has two main contributions. First, following an idea of Benner and Byers [2] (and, much earlier, of Man [33]) in the context of the algebraic Riccati equation, we incorporate exact line searches into Newton's method for the quadratic matrix equation in order to improve the global convergence properties. We show experimentally that exact line searches improve the reliability of Newton's method, leading to more frequent convergence and, often, faster convergence. Our second contribution is to derive the true condition number for the quadratic matrix equation, thus obtaining a sharper perturbation bound than Davis [7], and to obtain the backward error of an approximate solution.

Solving even the scalar quadratic equation reliably in floating point arithmetic is a difficult problem, as pointed out by Forsythe [15], principally due to the difficulty of handling underflow and overflow. We do not consider here the effects of underflow and overflow, but rather concentrate on the difficulties present with exact computation.

2. Theory. Before considering numerical solution of the quadratic matrix equation we examine the existence and enumeration of solvents. The fundamental theorem of algebra does not hold for matrix polynomials, as is shown by the special case of the matrix square root problem $X^2 = A$, which does not always have a solution when A is singular [28, sec. 6.4].

The quadratic matrix equation can be solved explicitly when $A = I$, B commutes with C, and $B^2 - 4C$ has a square root. We can complete the square in the usual way to obtain the solution

$$X = -\tfrac{1}{2}B + \tfrac{1}{2}(B^2 - 4C)^{1/2},$$

where $A^{1/2}$ denotes any square root that is a polynomial in A. This case pertains, for example, when A and B are scalar multiples of the identity and $B^2 - 4AC$ is nonsingular, after scaling though by A^{-1}. However, no generalization of the formula for the solution of a scalar quadratic is available for general A, B, and C.

Various sufficient conditions for the existence of a solvent are given by Eisenfeld [12] and Lancaster and Rokne [32]. In the former paper the results are obtained using the contraction mapping principle and in the latter paper using the Newton–Kantorovich theorem. Roughly speaking, all these results require that B or B^{-1} be small in norm compared with A and C, so they are of limited practical applicability.

The existence of dominant and minimal solvents is guaranteed for problems coming from overdamped quadratic eigenvalue problems (1.2): those for which A, B, and C are all symmetric positive definite and $(x^TBx)^2 > 4(x^TAx)(x^TCx)$ for all nonzero x; see Lancaster [30, sec. 7.6].

General information about existence of solvents comes from the connection between the quadratic matrix equation and the quadratic eigenvalue problem (1.2). Note, first, that if A is nonsingular then $\det(Q(\lambda)) = \det(A)\det(\lambda^2 I + A^{-1}B\lambda + A^{-1}C)$, so $\det(Q(\lambda))$ has degree exactly $2n$ and hence $Q(\lambda)$ has $2n$ eigenvalues, all of which are finite. If A is singular then $\det(Q(\lambda))$ has degree less than $2n$ and hence $Q(\lambda)$ has either less than $2n$ finite eigenvalues or infinitely many if $\det(Q(\lambda)) \equiv 0$.

The next result gives information on the number of solvents of $Q(X)$; it generalizes [10, Cor. 4.1].

THEOREM 2.1. *Suppose $Q(\lambda)$ has p distinct eigenvalues $\{\lambda_i\}_{i=1}^p$, with $n \le p \le 2n$, and that the corresponding set of p eigenvectors $\{v_i\}_{i=1}^p$ satisfies the Haar condition (that is, every subset of n of them is linearly independent). Then there are at least $\binom{p}{n}$ different solvents of $Q(X)$, and exactly this many if $p = 2n$, which are given by*

$$X = W\,\mathrm{diag}(\mu_i)W^{-1}, \qquad W = [\,w_1, \quad \ldots \quad , w_n\,], \tag{2.1}$$

where the eigenpairs $(\mu_i, w_i)_{i=1}^n$ are chosen from among the eigenpairs $(\lambda_i, v_i)_{i=1}^p$ of Q.

Proof. There are clearly $\binom{p}{n}$ choices of X in (2.1). Since $\mu_i^2 A w_i + \mu_i B w_i + C w_i = 0$, we have $AW\,\mathrm{diag}(\mu_i)^2 + BW\,\mathrm{diag}(\mu_i) + CW = 0$ and thence, on postmultiplying by W^{-1}, $Q(X) = 0$. That the $\binom{p}{n}$ solvents are different follows from the fact that no two have the same eigenvalues. Now suppose that $p = 2n$. From (1.3), every eigenpair of X is also an eigenpair of Q, and it follows that X is diagonalizable and of the form (2.1). ∎

When $p = n$ in Theorem 2.1 the distinctness of the eigenvalues is not needed in the proof, and we obtain a sufficient condition for the existence of a solvent.

COROLLARY 2.2. *If $Q(\lambda)$ has n linearly independent eigenvectors $v_1, \ldots, v_n$ then $Q(X)$ has a solvent.*

An example helps to clarify the theory. Consider the quadratic [10]

$$Q(X) = X^2 + \begin{bmatrix} -1 & -6 \\ 2 & -9 \end{bmatrix} X + \begin{bmatrix} 0 & 12 \\ -2 & 14 \end{bmatrix}.$$

$Q(\lambda)$ has four distinct eigenvalues, with eigenpairs (λ_i, v_i) given by

i	1	2	3	4
λ_i	1	2	3	4
v_i	$\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$	$\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$	$\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$	$\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$

To apply Theorem 2.1 we can take p no bigger than 3, in view of the Haar condition. If we take eigenvalues 1, 2, 3, then the theorem gives three solvents, having eigenvalues 1 and 2, 1 and 3, and 2 and 3. But the eigenvectors corresponding to eigenvalues 1, 2, 4 also satisfy the Haar condition and this gives us another two solvents, having eigenvalues 1 and 4, and 2 and 4. Note that there is no dominant solvent, which would have to have eigenvalues 3 and 4. The complete set of solvents is

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix}, \quad \begin{bmatrix} 1 & 3 \\ 0 & 4 \end{bmatrix}, \quad \begin{bmatrix} 4 & 0 \\ 2 & 2 \end{bmatrix}.$$

We were able to find all these solvents using the `solve` command of MATLAB's Symbolic Math Toolbox [34], but symbolic solution is clearly impractical for large n.

For a characterization of solvents via the generalized Schur decomposition of an associated matrix pencil, see [27].

3. Newton's method. Newton's method for solving the quadratic matrix equation (1.1) is readily obtained from the expansion

$$\begin{aligned} Q(X+E) &= Q(X) + \big(AEX + (AX+B)E\big) + AE^2 \\ &= Q(X) + D_X(E) + AE^2, \end{aligned} \tag{3.1}$$

where $D_X(E) : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is the Fréchet derivative of Q at X in the direction E. Newton's method drops the second order term, defines E as the solution of $Q(X) + D_X(E) = 0$, and replaces X by $X+E$. Each step of Newton's method involves finding the solution E of

$$AEX + (AX+B)E = -Q(X), \tag{3.2}$$

which is a special case of the generalized Sylvester equation "$AXB + CXD = E$."

We would like to know when the Fréchet derivative D_X is nonsingular, both at a solvent and at an iterate X, so that (3.2) has a solution. From a result of Chu [6] on the generalized Sylvester equation it follows that D_X is nonsingular if and only if the pair $(-A, AX+B)$ is regular (that is, $\det(-A - \lambda(AX+B))$ is not identically zero in λ) and the eigenvalues of the pair are distinct from the eigenvalues of X. If A is nonsingular, the regularity condition holds. When X is a solvent, we see from (1.3) that the second condition is equivalent to the eigenvalues of X being distinct from the remaining n eigenvalues of $Q(\lambda)$. We can therefore identify some sufficient conditions for nonsingularity of D_X at a solvent.

LEMMA 3.1. *If A is nonsingular then D_X is nonsingular at*

1. *a dominant or minimal solvent X,*
2. *all solvents X if the eigenvalues of $Q(\lambda)$ are distinct.*

For efficiency, $Q(X)$ should be calculated by nested multiplication as $(AX+B)X+C$, which requires two matrix multiplications instead of the three if X^2 is explicitly formed and provides the coefficient matrix $AX+B$ in (3.2) as a byproduct.

To solve (3.2) we can adapt methods for solving the generalized Sylvester equation described by Golub, Nash, and Van Loan [21] and Epton [13] (see also Chu [6] and Gardiner et al. [17], [18]). First we consider a Schur algorithm.

Compute the generalized Schur decomposition of A and $AX+B$ [22, Thm. 7.7.1],

$$W^*AZ = T, \qquad W^*(AX+B)Z = S, \tag{3.3}$$

where W and Z are unitary and T and S are upper triangular, and the Schur decomposition of X, $U^*XU = R$, where U is unitary and R is upper triangular. Then, pre- and postmultiplying (3.2) by W^* and U, respectively, transforms the system to

$$TYR + SY = F, \qquad F = -W^*Q(X)U, \quad Y = Z^*EU. \tag{3.4}$$

Equating kth columns and rearranging leads to

$$(S + r_{kk}T)y_k = f_k - \sum_{i=1}^{k-1} r_{ik}Ty_i, \qquad Y = [y_1, y_2, \ldots, y_n]. \tag{3.5}$$

By solving these upper triangular systems in the order $k = 1\colon n$, Y can be computed a column at a time. The cost of this algorithm is as follows, where a flop denotes a floating point operation. The generalized Schur decomposition requires $66n^3$ flops [22, sec. 7.7.6] and the Schur decomposition $25n^3$ flops [22, sec. 7.5.6]. Forming F

and transforming from Y to E in (3.4) costs $8n^3$ flops, and solving (3.5) requires $3n^3$ flops. The total is therefore $102n^3$ flops.

The Schur algorithm is used by Davis [7]. However, as noted by Golub, Nash, and Van Loan [21], Epton [13], and Gardiner et al. [17], one of the Schur decompositions can be replaced by a Hessenberg-triangular decomposition with a potentially substantial computational saving. Suppose we replace (3.3) by the Hessenberg-triangular decomposition [22, sec. 7.7.4]

$$W^*AZ = T, \qquad W^*(AX + B)Z = H,$$

where the only difference from (3.3) is that H is upper Hessenberg (this decomposition is a preliminary step to computing (3.3) by the QZ algorithm). The analogue of (3.5) is

$$(H + r_{kk}T)y_k = f_k - \sum_{i=1}^{k-1} r_{ik}Ty_i, \tag{3.6}$$

which is an upper Hessenberg system. The Hessenberg-triangular decomposition requires $15n^3$ flops [22, sec. 7.7.6] and the systems (3.6) can be solved in $4n^3$ flops. Hence the total cost of the Hessenberg–Schur algorithm is $52n^3$ flops, which is a 51 percent saving compared with the Schur algorithm.

Versions of the Schur and Hessenberg–Schur algorithms that employ real Schur decompositions and so use only real arithmetic can be developed; see [17] for details.

Standard convergence results for Newton's method apply [9, Thm. 5.2.1], as detailed in [29, Thm. 1]. In particular, if Newton's method is started sufficiently close to a solvent for which the Fréchet derivative is nonsingular, the iteration converges and at a quadratic rate. The Kantorovich theorem can also be applied to provide sufficient conditions for existence of a solvent and convergence of Newton's method to that solvent [9, Thm. 5.3.1].

4. Incorporating line searches. In the solution of unconstrained optimization problems by Newton or quasi-Newton methods it is common to use the Newton direction as a search direction and to define the next iterate by (approximately or exactly) minimizing the objective function along this direction [35, Chap. 3]; the minimization is called a line search. Line searches can also be used on nonlinear equation problems, given a suitable function for the line search to minimize. Benner and Byers [2] (see also [3]) investigate the use of exact line searches in Newton's method for solving the algebraic Riccati equation. (Man [33] had earlier used exact line searches in a quasi-Newton method for the same problem, but did not give any details.) Here, we apply exact line searches with Newton's method for the quadratic matrix equation.

The motivation for line searches is that, far from a solution, the linear model of $Q(X)$ on which Newton's method is based may be inaccurate, and so the Newton step E may not be a good one. Line searches are expected to give better global convergence (that is, convergence from arbitrary starting points). An example adapted from [2] illustrates the point. Consider the quadratic matrix equation

$$X^2 - \begin{bmatrix} 1 & 0 \\ 0 & \delta^{1/2} \end{bmatrix} = 0, \qquad 0 < \delta \ll 1,$$

which has solutions $X = \mathrm{diag}(\pm 1, \pm\delta^{1/4})$. With $X_0 = \mathrm{diag}(1, \delta)$, Newton's method gives $E = \mathrm{diag}(0, (\delta^{-1/2} - \delta)/2)$, so that $X_1 = X_0 + E$ is a much worse approximate

solvent than X_0. However, it is clear that $X_0 + tE$ is a solvent for suitable choice of the scalar t.

In our Newton method with line searches we take a multiple of the Newton step that minimizes the merit function

$$p(t) = \|Q(X + tE)\|_F^2, \tag{4.1}$$

where the Frobenius norm $\|A\|_F = (\mathrm{trace}(A^*A))^{1/2}$. Other choices of merit function could be tried (for example, based on other norms of Q), but this one is convenient to work with and has some theoretical backing, as explained below. Recalling that Newton's method defines E by $Q(X) + D_X(E) = 0$, from (3.1) we have, for this E,

$$\begin{aligned} Q(X + tE) &= Q(X) + tD_X(E) + t^2AE^2 \\ &= (1 - t)Q(X) + t^2AE^2. \end{aligned} \tag{4.2}$$

Thus

$$\begin{aligned} p(t) &= (1 - t)^2\|Q(X)\|_F^2 + t^4\|AE^2\|_F^2 \\ &\quad + (1 - t)t^2\,\mathrm{trace}\big(Q(X)^*AE^2(AE^2)^*Q(X)\big) \\ &\equiv \alpha(1 - t)^2 + \gamma t^4 + \beta(1 - t)t^2 \\ &= \gamma t^4 - \beta t^3 + (\alpha + \beta)t^2 - 2\alpha t + \alpha. \end{aligned} \tag{4.3}$$

If $\gamma = \|AE^2\|_F = 0$ then $p(t) = \alpha(1 - t)^2$, which attains its global minimum at $t = 1$, yielding the standard Newton step. If $\alpha = 0$ then X is a solvent. We can therefore assume that $\gamma > 0$ and $\alpha > 0$.

We have a quartic polynomial p of which we wish to find the global minimum. A quartic has at most two minima, of which one is the global minimum. We have

$$p'(t) = 2\alpha(t - 1) + \beta(2t - 3t^2) + 4\gamma t^3.$$

Hence

$$p'(0) = -2\alpha < 0, \tag{4.4}$$

and

$$\begin{aligned} p'(2) &= 2(\alpha - 4\beta + 16\gamma) \\ &= 2\,\mathrm{trace}\big(Q(X)^*Q(X) - 4(Q(X)^*AE^2 + (AE^2)^*Q(X)) + 16(AE^2)^*AE^2\big) \\ &= 2\,\mathrm{trace}\big((Q(X) - 4AE^2)^*(Q(X) - 4AE^2)\big) \\ &\geq 0. \end{aligned}$$

Since $p'(0) < 0$ and $p'(2) \geq 0$, p' has a real zero in the interval $(0, 2]$, and this zero corresponds to a minimum or a point of inflection of p. Since $t = 1$ corresponds to a pure Newton step, it is therefore reasonable to restrict our attention to the interval $[0, 2]$, although there is no guarantee that there is a minimum of p in this interval when we are far from a solution. Thus we define t by

$$p(t) = \min_{x \in [0,2]} p(x). \tag{4.5}$$

There are two cases to consider.

(1) If p' has one real zero and a (nonreal) complex conjugate pair of roots then the real zero, which must lie in $(0, 2]$, is the desired global minimum.

(2) If p' has three real zeros then at most two are minima of p. If the global minimum lies outside $(0, 2]$ then $t = 2$ needs to be checked, as it may yield a smaller value of p than the zero of p' in $(0, 2]$.

Knowing these cases, it is easy to implement the choice of t in (4.5), since the zeros of the cubic p' and the values of p at these zeros are easily computed.

The question arises of whether the exact line searches interfere with the quadratic convergence of Newton's method, necessitating the explicit setting of $t = 1$ once convergence is approached. The answer is no, under a mild assumption, as we now show. Assume that X_j is within a region where quadratic convergence to X occurs, and let $X_{j+1} = X_j + E_j$ and $\widetilde{X}_{j+1} = X_j + tE_j$ be the standard Newton update and the update with exact line search, respectively. Defining $\Delta_j = X - X_j$, we have

$$\|\Delta_{j+1}\| = O\big(\|\Delta_j\|^2\big).$$

The definition of t ensures that, using (4.2),

$$\begin{aligned} \|(1-t)Q(X_j) + t^2 A E_j^2\| &= \|Q(X_j + tE_j)\| \le \|Q(X_j + E_j)\| \\ &= \|Q(X_{j+1})\| = \|Q(X - \Delta_{j+1})\| \\ &= \|Q(X)\| + O(\|\Delta_{j+1}\|) \\ &= O\big(\|\Delta_j\|^2\big). \end{aligned} \tag{4.6}$$

Now $E_j = -\Delta_{j+1} + \Delta_j$, so $\|E_j\| = O(\|\Delta_j\|)$ and, by (3.1),

$$Q(X_j) = Q(X - \Delta_j) = -D_X(\Delta_j) + O(\|\Delta_j\|^2).$$

Hence, as long as the Fréchet derivative is nonsingular at X, (4.6) implies that $|1-t| = O(\|\Delta_j\|)$. Thus

$$X - \widetilde{X}_{j+1} = X - X_{j+1} + X_{j+1} - \widetilde{X}_{j+1} = O\big(\|\Delta_j\|^2\big) + (1-t)E_j = O\big(\|\Delta_j\|^2\big),$$

as required.

The global convergence properties of Newton's method with exact line searches can be obtained from standard theory. We are effectively solving a nonlinear system $f(x) = 0$ by Newton's method, where $f : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$, doing line searches on the function $F(x) = f(x)^T f(x)$, as advocated by Dennis and Schnabel [9, sec. 6.5] and Fletcher [14, sec. 6.2]. The global convergence results of [9, sec. 6.3], [14, sec. 2.5] apply provided that certain restrictions known as the Armijo–Goldstein conditions are imposed on the line search. In our notation these conditions may be written as

$$p(t) \le p(0) + c_1 t p'(0), \tag{4.7a}$$

$$p'(t) \ge c_2 p'(0), \tag{4.7b}$$

where c_1 and c_2 are parameters with $0 < c_1 < c_2 < 1$. The first condition ensures that the reduction in p is at least as big as that predicted by a first order model, while the second ensures that the step is not too small, by requiring that the derivative at t be at least some fraction of the derivative at 0. It is easy to see using (4.4) that (4.7a) is equivalent to

$$\|Q(X + tE)\|_F^2 \le (1 - 2c_1 t)\|Q(X)\|_F^2,$$

which requires a sufficient decrease in the merit function. The use of exact line searches does not necessarily imply that the conditions (4.7) are satisfied. However, (4.7b) certainly holds in the usual case when the optimal t is a zero of $p'(t)$, since $p'(0) < 0$. Both conditions have been checked and found to be satisfied in all our numerical tests (with $c_1 = 1/4$, $c_2 = 1/2$), so we have not considered any modifications to the exact line search.

The line search requires three matrix multiplications to compute the coefficients of p in (4.3) ($Q(X)$ is already available), the remaining computations being scalar ones. The total cost of the line search is $5n^3$ flops, which is negligible compared with the cost of computing the Newton direction E (at least $56n^3$ flops).

5. Conditioning. We now derive a condition number for a solvent of the quadratic matrix equation (1.1). The analyses in this section and the next have close connections with analyses for Sylvester and algebraic Riccati equations in [19], [23], [25], [26].

Consider the perturbed equation

$$(A + \Delta A)(X + \Delta X)^2 + (B + \Delta B)(X + \Delta X) + C + \Delta C = 0. \tag{5.1}$$

We will measure the perturbations normwise by

$$\epsilon = \| \, [\, \alpha^{-1}\Delta A, \quad \beta^{-1}\Delta B, \quad \gamma^{-1}\Delta C \,] \, \|_F,$$

where α, β, and γ are nonnegative parameters. A zero value of α, say, simply forces the corresponding perturbation ΔA to be zero. Expanding (5.1) we obtain

$$AX\Delta X + A\Delta X X + B\Delta X = -\Delta A X^2 - \Delta B X - \Delta C + O(\epsilon^2). \tag{5.2}$$

We now use the vec operator, which stacks the columns of a matrix into one long vector, and the Kronecker product $A \otimes B = (a_{ij}B)$, and we use the property $\operatorname{vec}(AXB) = (B^T \otimes A)\operatorname{vec}(X)$ [28, Chap. 4]. Applying the vec operator to (5.2) we obtain

$$\begin{aligned} P \operatorname{vec}(\Delta X) &= -((X^2)^T \otimes I_n)\operatorname{vec}(\Delta A) - (X^T \otimes I_n)\operatorname{vec}(\Delta B) - \operatorname{vec}(\Delta C) + O(\epsilon^2) \\ &= -\, [\, \alpha (X^2)^T \otimes I_n, \quad \beta X^T \otimes I_n, \quad \gamma I_{n^2} \,] \begin{bmatrix} \operatorname{vec}(\Delta A)/\alpha \\ \operatorname{vec}(\Delta B)/\beta \\ \operatorname{vec}(\Delta C)/\gamma \end{bmatrix} + O(\epsilon^2), \end{aligned}$$

where

$$P = I_n \otimes AX + X^T \otimes A + I_n \otimes B.$$

Multiplying by P^{-1}, taking 2-norms, and using $\| \operatorname{vec}(X) \|_2 = \|X\|_F$, we obtain the bound

$$\frac{\|\Delta X\|_F}{\|X\|_F} \le \Psi(X)\epsilon + O(\epsilon^2), \tag{5.3}$$

where

$$\Psi(X) = \|P^{-1} \, [\, \alpha (X^2)^T \otimes I_n, \quad \beta X^T \otimes I_n, \quad \gamma I_{n^2} \,] \, \|_2 / \|X\|_F.$$

This is a sharp bound, to first order in ϵ, so $\Psi(X)$ is the condition number of X. Note that P is nonsingular, and hence $\Psi(X)$ finite, precisely when the Fréchet derivative D_X in (3.1) is nonsingular.

An upper bound for $\Psi(X)$ involving $\|P^{-1}\|_2$ can of course be obtained by bounding the norm of the product by the product of the norms, but this bound can be arbitrarily weaker than (5.3). A perturbation bound for (1.1) that contains a factor $\|D_X^{-1}\|_F$ is derived by Davis [7], and it is easy to show that $\|P^{-1}\|_2 = \|D_X^{-1}\|_F$.

For the special case of the matrix square root we have $A = I$, $B = 0$, $\alpha = \beta = 0$, and the condition number Ψ simplifies to

$$\Psi(X) = \frac{\|P^{-1}\|_2 \gamma}{\|X\|_F}, \qquad P = I_n \otimes X + X^T \otimes I_n,$$

which is the matrix square root condition number identified in [25].

We give an illustrative example from [20, Ex. 4.4], with

$$A = I_2, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} -1 & 0 \\ -1 & 0 \end{bmatrix}.$$

The eigenvalues of $Q(\lambda)$ are -1, 0, 0, 1 and there are three solvents:

$$X_1 = \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad X_3 = \begin{bmatrix} -1 & 0 \\ -2 & 0 \end{bmatrix}.$$

The solvent X_1 is dominant and so Theorem 3.1 implies it has a finite condition number; in fact $\Psi(X_1) = 3.64$. The other two solvents are both easily seen to have singular P and hence infinite condition numbers.

6. Backward error. We define the backward error of an approximate solution Y to (1.1) by

$$\begin{aligned} \eta(Y) = \min\big\{ \epsilon : (A + \Delta A)Y^2 + (B + \Delta B)Y + C + \Delta C = 0, \\ \|\,[\,\alpha^{-1}\Delta A, \quad \beta^{-1}\Delta B, \quad \gamma^{-1}\Delta C\,]\,\|_F \le \epsilon \big\}. \end{aligned} \tag{6.1}$$

Defining

$$R = AY^2 + BY + C,$$

the constraint equation in (6.1) can be written as

$$\begin{aligned} -R &= \Delta A Y^2 + \Delta B Y + \Delta C \\ &= [\,\alpha^{-1}\Delta A, \quad \beta^{-1}\Delta B, \quad \gamma^{-1}\Delta C\,] \begin{bmatrix} \alpha Y^2 \\ \beta Y \\ \gamma I_n \end{bmatrix}. \end{aligned} \tag{6.2}$$

Taking Frobenius norms leads to the lower bound for the backward error

$$\eta(Y) \ge \frac{\|R\|_F}{(\alpha^2\|Y^2\|_F^2 + \beta^2\|Y\|_F^2 + n\gamma^2)^{1/2}}.$$

Applying the vec operator to (6.2) gives

$$[\,\alpha(Y^2)^T \otimes I_n, \quad \beta Y^T \otimes I_n, \quad \gamma I_{n^2}\,] \begin{bmatrix} \operatorname{vec}(\Delta A)/\alpha, \\ \operatorname{vec}(\Delta B)/\beta, \\ \operatorname{vec}(\Delta C)/\gamma \end{bmatrix} = -\operatorname{vec}(R), \tag{6.3}$$

which we write as

$$Hz = r, \qquad H \in \mathbb{R}^{n^2 \times 3n^2}.$$

We assume that H is of full rank, which guarantees that (6.3) has a solution, that is, that the backward error is finite. The backward error is the minimum 2-norm solution to this underdetermined system:

$$\eta(Y) = \|H^+ r\|_2,$$

where a superscript "+" denotes the pseudoinverse. To obtain an upper bound for $\eta(Y)$ we use

$$\eta(Y) \le \|H^+\|_2 \|r\|_2 = \frac{\|r\|_2}{\sigma_{\min}(H)},$$

where $\sigma_{\min}$ denotes the smallest singular value, which is nonzero by assumption. Now

$$\begin{aligned}
\sigma_{\min}(H)^2 &= \lambda_{\min}(HH^*) \\
&= \lambda_{\min}(\alpha^2 (Y^2)^T \overline{Y^2} \otimes I_n + \beta^2 Y^T \overline{Y} \otimes I_n + \gamma^2 I_{n^2}) \\
&= \lambda_{\min}(\alpha^2 (Y^2)^* Y^2 \otimes I_n + \beta^2 Y^* Y \otimes I_n + \gamma^2 I_{n^2}) \\
&\ge \alpha^2 \sigma_{\min}(Y^2)^2 + \beta^2 \sigma_{\min}(Y)^2 + \gamma^2.
\end{aligned}$$

Thus

$$\eta(Y) \le \frac{\|R\|_F}{(\alpha^2 \sigma_{\min}(Y^2)^2 + \beta^2 \sigma_{\min}(Y)^2 + \gamma^2)^{1/2}}.$$

We conclude from this analysis that a small relative residual does not necessarily imply a small backward error for the quadratic matrix equation. The same is true for the Sylvester equation [26] and, more generally, the algebraic Riccati equation [19].

7. Numerical experiments. Davis [7], [8] demonstrated the usefulness of Newton's method for solving the quadratic matrix equation. Our purpose in this section is to show experimentally the benefits of exact line searches in Newton's method. Our experiments were done in MATLAB, which has unit roundoff $u = 2^{-53} \approx 1.1 \times 10^{-16}$.

First, we give a few details about our Newton implementation. The default starting matrix is, as in [7],

$$X_0 = \left(\frac{\|B\|_F + \sqrt{\|B\|_F^2 + 4\|A\|_F \|C\|_F}}{2\|A\|_F} \right) I,$$

which is designed to have norm roughly of the same order of magnitude as a solvent. We terminate the iteration when the residual $Q(X_k)$ is of the same order of magnitude as the rounding error in computing it, namely, when the relative residual $\rho(X_k)$ satisfies

$$\rho(X_k) = \frac{\|fl(Q(X_k))\|_F}{\|A\|_F \|X_k\|_F^2 + \|B\|_F \|X_k\|_F + \|C\|_F} \le nu. \tag{7.1}$$

Our MATLAB code has an option to choose whether to use line searches. When line searches are being used, they are turned off (t is set to 1) once $\rho(X_k) \le 10^{-7}$; this is not necessary in theory (see section 4), but is done to save work and as a precaution to

avoid rounding errors destroying the quadratic convergence. In evaluating backward errors and condition numbers we took $\alpha = \|A\|_F$, $\beta = \|B\|_F$, $\gamma = \|C\|_F$.

The potential benefits of exact line searches are easily demonstrated. Consider the quadratic matrix equation with

$$A = I_2, \qquad B = \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{7.2}$$

As noted in [8] there are real solvents I_2 and $\left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$ and an infinite number of complex solvents. Applying Newton's method with and without line searches for the default X_0 and $X_0 = 10^j I$, $j = 1, 5, 10$, gave the results in Table 7.1, which show the substantial reduction in iterations that exact line searches can bring. In each case the computed solvent $\widehat{X}$ was within roundoff of I_2, with condition number $\Psi(\widehat{X}) = 1.4$ and backward error $\eta(\widehat{X}) \approx u$.

Our next example is the quadratic matrix equation with

$$A = B = I_2, \qquad C = \begin{bmatrix} -8 & -12 \\ -18 & -26 \end{bmatrix}, \tag{7.3}$$

again from [8], which has four solvents, all real and well conditioned. With the default starting matrix, convergence was obtained in 6 iterations with line searches and 10 without line searches, to the same matrix. We chose starting matrices

$$X_0 = \begin{bmatrix} 1 & x \\ y & 1 \end{bmatrix}, \qquad -1000 \le x,\ y \le 1000,$$

with an equally spaced grid of 100 points (x, y). Table 7.2 shows how many times a solvent was produced within 30, 50, and 100 iterations, respectively. Convergence was obtained to all four solvents, depending on the starting matrix, and a different solvent was sometimes obtained with exact line searches than without. Exact line searches result in more frequent convergence, though in 10 of the cases convergence was obtained without line searches but not with them, and in another 22 cases where both gave convergence faster convergence was obtained without line searches. Thus exact line searches do not lead to uniformly better convergence than when no line searches are used. An interesting phenomenon is that in 48 cases when line searches were not used the test (7.1) was satisfied within 100 iterations, but with $\|X\|_F \gg u^{-1}$, so that X was far from a solvent (these cases were counted as failure to converge for the statistics). This behavior did not happen with line searches: the line searches force $\|Q(X_k)\|_F$ to be a decreasing sequence, which tends to keep X_k from becoming large if there is no large solvent.

Our final example is based on a quadratic eigenvalue problem (1.2) from [16, sec. 10.11], with numerical values modified as in [30, sec. 5.3], modelling oscillations in an airplane wing:

$$A = \begin{bmatrix} 17.6 & 1.28 & 2.89 \\ 1.28 & 0.824 & 0.413 \\ 2.89 & 0.413 & 0.725 \end{bmatrix}, \qquad B = \begin{bmatrix} 7.66 & 2.45 & 2.1 \\ 0.23 & 1.04 & 0.223 \\ 0.6 & 0.756 & 0.658 \end{bmatrix},$$
$$C = \begin{bmatrix} 121 & 18.9 & 15.9 \\ 0 & 2.7 & 0.145 \\ 11.9 & 3.64 & 15.5 \end{bmatrix}. \tag{7.4}$$

TABLE 7.1
Number of iterations for convergence for problem (7.2).

X_0	Without line searches	With exact line searches
Default	6	5
$10I$	9	6
$10^5 I$	22	6
$10^{10} I$	39	7

TABLE 7.2
Number of times convergence obtained for problem (7.3) *with* 100 *different starting matrices.*

No. iterations allowed	Without line searches	With exact line searches
30	46	54
50	52	73
100	53	88

The 6 eigenvalues are distinct and come in 3 complex conjugate pairs; since any solvent must have 3 eigenvalues chosen from the 6, it follows that there are no real solvents. Starting Newton's method with $X_0 = iI$ we obtained the results displayed in Figure 7.1. Convergence was obtained to the same solvent with and without line searches, with condition number $\Psi(\widehat{X}) = 50$ and backward error $\eta(\widehat{X}) \approx u$. The eigenvalues of the computed solvent are

$$\begin{aligned} &\texttt{-8.8483e-001} + \texttt{8.4415e+000i}, \\ &\texttt{9.4722e-002} + \texttt{2.5229e+000i}, \\ &\texttt{-9.1800e-001} + \texttt{1.7606e+000i}, \end{aligned}$$

and these and their conjugates are the eigenvalues of the quadratic eigenvalue problem.

Finally, we note that in all our tests the global minimum of the merit function p in (4.1) was in $(0, 2]$ and never to the right of 2.

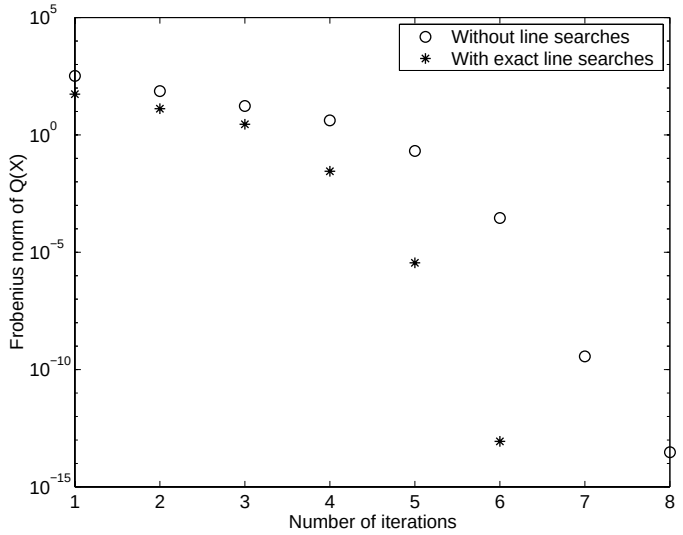

FIG. 7.1. *Convergence for problem* (7.4).

8. Concluding remarks. Newton's method is a useful tool in our stock of methods for solving quadratic matrix equations. In its favor is its applicability to the whole class of problems and its quadratic convergence, the latter making it a useful way to refine approximate solvents obtained with other methods. On the other hand each iteration is relatively expensive. The exact line searches introduced here frequently reduce the number of iterations and make standard global convergence results from optimization applicable.

A number of open problems remain, including guaranteeing convergence for particular starting matrices, determining to which solvent Newton's method will converge, and improving the convergence to solvents at which the Fréchet derivative is singular. These questions have been answered for certain types of Riccati equations, by exploiting their structure [2], [24], but the lack of structure in the quadratic matrix equation has so far precluded any useful results.

Acknowledgments. We thank Françoise Tisseur and the referees for their helpful suggestions.

REFERENCES

[1] P. Benner, *Computational methods for linear-quadratic optimization*, Rend. Circ. Mat. Palermo (2) Suppl., 58 (1999), pp. 21–56. Extended version available as Berichte aus der Technomathematik, Report 98–04, Universität Bremen, August 1998, from http://www.math.uni-bremen.de/zetem/berichte.html.

[2] P. Benner and R. Byers, *An exact line search method for solving generalized continuous-time algebraic Riccati equations*, IEEE Trans. Automat. Control, 43 (1998), pp. 101–107.

[3] P. Benner, R. Byers, E. S. Quintana-Ortí, and G. Quintana-Ortí, *Solving algebraic Riccati equations on parallel computers using Newton's method with exact line search*, Parallel Comput., 26 (2000), pp. 1345–1368.

[4] S. Bittanti, A. J. Laub, and J. C. Willems, eds., *The Riccati Equation*, Springer-Verlag, Berlin, 1991.

[5] T. J. Bridges and P. J. Morris, *Differential eigenvalue problems in which the parameter appears nonlinearly*, J. Comput. Phys., 55 (1984), pp. 437–460.

[6] K.-w. E. Chu, *The solution of the matrix equations* $AXB - CXD = E$ *and* $(YA - DZ, YC - BZ) = (E, F)$, Linear Algebra Appl., 93 (1987), pp. 93–105.

[7] G. J. Davis, *Numerical solution of a quadratic matrix equation*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 164–175.

[8] G. J. Davis, *Algorithm* 598: *An algorithm to compute solvents of the matrix equation* $AX^2 + BX + C = 0$, ACM Trans. Math. Software, 9 (1983), pp. 246–254.

[9] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[10] J. E. Dennis, Jr., J. F. Traub, and R. P. Weber, *The algebraic theory of matrix polynomials*, SIAM J. Numer. Anal., 13 (1976), pp. 831–845.

[11] J. E. Dennis, Jr., J. F. Traub, and R. P. Weber, *Algorithms for solvents of matrix polynomials*, SIAM J. Numer. Anal., 15 (1978), pp. 523–533.

[12] J. Eisenfeld, *Operator equations and nonlinear eigenparameter problems*, J. Funct. Anal., 12 (1973), pp. 475–490.

[13] M. A. Epton, *Methods for the solution of* $AXD - BXC = E$ *and its application in the numerical solution of implicit ordinary differential equations*, BIT, 20 (1980), pp. 341–345.

[14] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., Wiley, Chichester, UK, 1987.

[15] G. E. Forsythe, *What is a satisfactory quadratic equation solver?*, in Constructive Aspects of the Fundamental Theorem of Algebra, B. Dejon and P. Henrici, eds., Wiley-Interscience, London, 1969, pp. 53–61.

[16] R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices and Some Applications to Dynamics and Differential Equations*, 10th ed., Cambridge University Press, New York, 1963. Reprint of 1938 edition.

[17] J. D. Gardiner, A. J. Laub, J. J. Amato, and C. B. Moler, *Solution of the Sylvester matrix equation* $AXB^T + CXD^T = E$, ACM Trans. Math. Software, 18 (1992), pp. 223–231.

[18] J. D. Gardiner, M. R. Wette, A. J. Laub, J. J. Amato, and C. B. Moler, *Algorithm* 705: *A FORTRAN-77 software package for solving the Sylvester matrix equation* $AXB^T + CXD^T = E$, ACM Trans. Math. Software, 18 (1992), pp. 232–238.
[19] A. R. Ghavimi and A. J. Laub, *Backward error, sensitivity, and refinement of computed solutions of algebraic Riccati equations*, Numer. Linear Algebra Appl., 2 (1995), pp. 29–49.
[20] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[21] G. H. Golub, S. Nash, and C. F. Van Loan, *A Hessenberg–Schur method for the problem* $AX + XB = C$, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–913.
[22] G. H. Golub and C. F. Van Loan, *Matrix Computations*, third ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[23] T. Gudmundsson, C. S. Kenney, and A. J. Laub, *Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method*, IEEE Trans. Automat. Control, 37 (1992), pp. 513–518.
[24] C.-H. Guo and P. Lancaster, *Analysis and modification of Newton's method for algebraic Riccati equations*, Math. Comp., 67 (1998), pp. 1089–1105.
[25] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.
[26] N. J. Higham, *Perturbation theory and backward error for* $AX - XB = C$, BIT, 33 (1993), pp. 124–136.
[27] N. J. Higham and H.-M. Kim, *Numerical analysis of a quadratic matrix equation*, IMA J. Numer. Anal., 20 (2000), pp. 499–519.
[28] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.
[29] W. Kratz and E. Stickel, *Numerical solution of matrix polynomial equations by Newton's method*, IMA J. Numer. Anal., 7 (1987), pp. 355–369.
[30] P. Lancaster, *Lambda-Matrices and Vibrating Systems*, Pergamon Press, Oxford, UK, 1966.
[31] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, The Clarendon Press, Oxford University Press, New York, 1995.
[32] P. Lancaster and J. G. Rokne, *Solutions of nonlinear operator equations*, SIAM J. Math. Anal., 8 (1977), pp. 448–457.
[33] F. T. Man, *The Davidon method of solution of the algebraic matrix Riccati equation*, Internat. J. Control, 10 (1969), pp. 713–719.
[34] C. Moler and P. J. Costa, *Symbolic Math Toolbox Version* 2.0: *User's Guide*, The MathWorks, Inc., Natick, MA, 1997.
[35] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York, 1999.
[36] H. A. Smith, R. K. Singh, and D. C. Sorensen, *Formulation and solution of the non-linear, damped eigenvalue problem for skeletal systems*, Internat. J. Numer. Methods Engrg., 38 (1995), pp. 3071–3085.
[37] Z. C. Zheng, G. X. Ren, and W. J. Wang, *A reduction method for large scale unsymmetric eigenvalue problems in structural dynamics*, J. Sound Vibration, 199 (1997), pp. 253–268.