IMA Journal of Numerical Analysis (2000) **20**, 499–519

Numerical analysis of a quadratic matrix equation

NICHOLAS J. HIGHAM AND HYUN-MIN KIM
Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK

[Received 5 August 1999 and in revised form 13 December 1999]

The quadratic matrix equation $AX^2 + BX + C = 0$ in $n \times n$ matrices arises in applications and is of intrinsic interest as one of the simplest nonlinear matrix equations. We give a complete characterization of solutions in terms of the generalized Schur decomposition and describe and compare various numerical solution techniques. In particular, we give a thorough treatment of functional iteration methods based on Bernoulli's method. Other methods considered include Newton's method with exact line searches, symbolic solution and continued fractions. We show that functional iteration applied to the quadratic matrix equation can provide an efficient way to solve the associated quadratic eigenvalue problem $(\lambda^2 A + \lambda B + C)x = 0$.

Keywords: quadratic matrix equation; solvent; generalized Schur decomposition; scaling; functional iteration; Bernoulli's method; Newton's method; exact line searches; continued fractions; quadratic eigenvalue problem.

1. Introduction

Matrix algebra was developed by Cayley in the 1850s. One hundred and fifty years later we can solve a wide variety of linear system and linear eigenvalue problems, exploiting all kinds of structure and computing architectures. Nonlinear matrix problems remain relatively unexplored, however. Our subject here is the simplest nonlinear matrix equation, the quadratic. A quadratic equation in matrices can be defined by

$$Q(X) = AX^2 + BX + C = 0, \qquad A, B, C \in \mathbb{C}^{n\times n}. \tag{1}$$

This is not the only possible definition: others include $X^2A + XB + C = 0$, for which an analogous treatment is possible, and the algebraic Riccati equation $XAX+BX+XC+D = 0$. Sylvester investigated the quadratic matrix equation (1) in several papers published in the 1880s; see, for example, Sylvester (1884).

It is natural to ask whether the usual formula for the roots of a scalar quadratic generalizes to (1). The answer is no, except in special cases such as when $A = I$, B and C commute, and $B^2 - 4C$ has a square root. In fact, the existence and characterization of solutions of (1) is not straightforward, and computing a solution poses interesting challenges.

In this work we investigate the theory and numerical solution of the quadratic matrix equation (1). We collect, unify and generalize earlier work and derive new or strengthened results, including a powerful result and algorithm based on the generalized Schur decomposition. We take care to point out applications and sometimes unexpected links, including links with continued fractions and the now little used Bernoulli method.

We will need three matrix norms on $\mathbb{C}^{n\times n}$: the Frobenius norm

$$\|A\|_F = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}$$

and two special cases of matrix norms subordinate to the vector p-norm:

$$\|A\|_2 = \sigma_{\max}(A), \qquad \|A\|_1 = \max_j \sum_i |a_{ij}|,$$

where $\sigma_{\max}$ denotes the largest singular value.

2. Theory of solvents

A solution to the quadratic matrix equation (1) is called a solvent. Existence and characterization of solvents can be treated via the quadratic eigenvalue problem or by directly attacking the matrix equation. We consider both approaches. First, we note that the quadratic matrix equation can have no solvents, a finite positive number, or infinitely many, as follows immediately from the theory of matrix square roots ($Q(X) = X^2 - A$) (Higham, 1987; Horn & Johnson, 1991, Section 6.4).

The quadratic matrix equation is intimately connected with the quadratic eigenvalue problem

$$Q(\lambda)x = (\lambda^2 A + \lambda B + C)x = 0. \tag{2}$$

It follows from the theory of λ-matrices that if S is a solvent of $Q(X)$ then $S - \lambda I$ is a right divisor of $Q(\lambda)$ (Gohberg *et al.*, 1982, Corollary 3.6; Lancaster, 1966, Theorem 3.3) and this fact is easily verified because the quotient has a simple form:

$$\lambda^2 A + \lambda B + C = -(B + AS + \lambda A)(S - \lambda I). \tag{3}$$

Hence any eigenpair of S is an eigenpair of $Q(\lambda)$. Indeed solvents can be constructed from eigenpairs of $Q(\lambda)$. If

$$Q(\lambda_i)v_i = 0, \quad i = 1\colon n$$

then

$$AV\Lambda^2 + BV\Lambda + CV = 0, \qquad V = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix}, \quad \Lambda = \operatorname{diag}(\lambda_i),$$

and provided that V is nonsingular we can postmultiply by V^{-1} to deduce that $S = V\Lambda V^{-1}$ is a solvent. It is the nonsingularity requirement on V that complicates the theory, because for the quadratic eigenvalue problem it is not necessarily the case that eigenvectors corresponding to distinct eigenvalues are linearly independent. Consider the example (Dennis *et al.*, 1976)

$$Q(X) = X^2 + \begin{bmatrix} -1 & -6 \\ 2 & -9 \end{bmatrix} X + \begin{bmatrix} 0 & 12 \\ -2 & 14 \end{bmatrix}. \tag{4}$$

The eigenpairs (λ_i, v_i) of $Q(\lambda)$ are given by

i	1	2	3	4
λ_i	1	2	3	4
v_i	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$

The construction described above yields solvents with all distinct pairs of eigenvalues except 3 and 4:

$$\begin{bmatrix}1 & 0\\0 & 2\end{bmatrix}, \quad \begin{bmatrix}1 & 2\\0 & 3\end{bmatrix}, \quad \begin{bmatrix}3 & 0\\1 & 2\end{bmatrix}, \quad \begin{bmatrix}1 & 3\\0 & 4\end{bmatrix}, \quad \begin{bmatrix}4 & 0\\2 & 2\end{bmatrix}. \tag{5}$$

This is the complete set of solvents.

This method of constructing S produces only diagonalizable solvents and therefore can fail to identify all solvents and can even produce no solvents when solvents exist. The method can, nevertheless, be made the basis of results stating the existence of a certain number of solvents (Higham & Kim, 1999, Theorem 2.1). Indeed, since $Q(\lambda)$ has $2n$ eigenvalues when A is nonsingular, we can expect at most $\binom{2n}{n}$ solvents, generically. (This number of solvents was identified by Sylvester, 1885, for the case $A = I$.)

It is desirable to have sufficient conditions for the existence of a solvent that do not involve knowledge of the eigensystem of $Q(\lambda)$. Eisenfeld (1973, Theorem 5.4) uses the contraction mapping principle to show that if A and B are nonsingular and

$$4\|B^{-1}A\|\|B^{-1}C\| < 1 \tag{6}$$

for a subordinate matrix norm, then at least two solvents exist (note the relation to the condition that the discriminant of a scalar quadratic be positive). A similar but more restrictive condition is derived by McFarland (1958). Lancaster & Rokne (1977) use the Kantorovich theorem on the convergence of Newton's method to derive several sets of sufficient conditions for the existence of a solvent, including (6) without the restriction that A is nonsingular.

It would be expected that quadratic matrix equations arising in applications would have properties that guarantee the existence of a solvent. An interesting class of problems is those related to quadratic eigenvalue problems arising in vibration problems with strong damping. The following definition was introduced by Duffin (1955).

DEFINITION 1 The quadratic eigenvalue problem (1) is overdamped if A and B are symmetric positive definite, C is symmetric positive semidefinite and

$$(x^TBx)^2 > 4(x^TAx)(x^TCx) \qquad \text{for all } x \neq 0. \tag{7}$$

Lancaster (1966, Section 7.6) proves several results for an overdamped problem: the eigenvalues are real and nonpositive and there is a gap between the n largest (the primary eigenvalues) and the n smallest (the secondary eigenvalues); there are n linearly independent eigenvectors associated with the primary eigenvalues and likewise for the secondary eigenvalues; and $Q(X)$ has at least two real solvents, having as their eigenvalues the primary eigenvalues and the secondary eigenvalues, respectively. The same results are shown by Krein & Langer (1978) in the case $A = I$.

Condition (7) is not easy to check, in general, but it is certainly satisfied if $\lambda_{\min}(B)^2 > 4\lambda_{\max}(A)\lambda_{\max}(C)$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalue, respectively.

An alternative way to approach the theory of solvents is via eigensystems of an associated generalized eigenvalue problem. The following development was inspired by analysis for the algebraic Riccati equation $XAX + BX + XC + D = 0$, which characterizes solutions in terms of the eigensystem of the matrix

$$\begin{bmatrix} -C & -A \\ D & B \end{bmatrix}.$$

This analysis originates with Anderson (1966) and Potter (1966), with more recent treatments that we have found useful given by Laub (1979), Lancaster & Rodman (1995, Theorem 7.1.2) and Van Dooren (1981).

We begin with a lemma that characterizes solvents. We define

$$F = \begin{bmatrix} 0 & I \\ -C & -B \end{bmatrix}, \qquad G = \begin{bmatrix} I & 0 \\ 0 & A \end{bmatrix}. \tag{8}$$

LEMMA 2 X is a solvent of $Q(X)$ in (1) if and only if

$$F \begin{bmatrix} I \\ X \end{bmatrix} = G \begin{bmatrix} I \\ X \end{bmatrix} X. \tag{9}$$

Proof. A direct computation. □

The lemma can be interpreted as saying that X is a solvent if and only if the columns of $\left[\begin{smallmatrix} I \\ X \end{smallmatrix}\right]$ span a deflating subspace (Stewart & Sun, 1990, Section 6.2.4) for the pair (F, G).

Lemma 2 can be developed into a more concrete characterization of solvents by representing the required deflating subspace in terms of the eigensystem of the pair (F, G). Instead of the Kronecker canonical form we use the more computationally satisfactory generalized Schur decomposition (Golub & Van Loan, 1996, Theorem 7.7.1). The following result does not appear to have been stated in the literature before, though it is closely related to existing results for algebraic Riccati equations.

THEOREM 3 All solvents of $Q(X)$ are of the form $X = Z_{21}Z_{11}^{-1} = Q_{11}T_{11}S_{11}^{-1}Q_{11}^{-1}$, where

$$Q^*FZ = T, \quad Q^*GZ = S \tag{10}$$

is a generalized Schur decomposition with Q and Z unitary and T and S upper triangular, and where all matrices are partitioned as block 2×2 matrices with $n \times n$ blocks.

Proof. First we show that $X = Z_{21}Z_{11}^{-1}$ (with nonsingular Z_{11}) is a solvent. The first block columns of $FZ = QT$ and $GZ = QS$ give

$$Z_{21} = Q_{11}T_{11}, \tag{11}$$

$$-CZ_{11} - BZ_{21} = Q_{21}T_{11}, \tag{12}$$

$$Z_{11} = Q_{11}S_{11}, \tag{13}$$

$$AZ_{21} = Q_{21}S_{11}. \tag{14}$$

Postmultiplying (12) by Z_{11}^{-1} and using (11) and then (13) and (14) gives

$$\begin{aligned} -C - BZ_{21}Z_{11}^{-1} &= Q_{21}T_{11}Z_{11}^{-1} \\ &= Q_{21} \cdot Q_{11}^{-1}Z_{21} \cdot Z_{11}^{-1} \\ &= AZ_{21}Z_{11}^{-1} \cdot Z_{21}Z_{11}^{-1}, \end{aligned}$$

which shows that $X = Z_{21}Z_{11}^{-1}$ is a solvent.

For the converse, let X be a solvent. Then by Lemma 2 it satisfies (9). Let

$$\begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = Z \begin{bmatrix} R \\ 0 \end{bmatrix} \tag{15}$$

be a QR factorization. It is easy to see that R and Z_{11} are nonsingular and $X = Z_{21}R = Z_{21}Z_{11}^{-1}$. Let

$$\begin{bmatrix} I & 0 \\ 0 & A \end{bmatrix} Z = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix}$$

be a QR factorization. On premultiplying (9) by Q^* and using (15) we obtain

$$\begin{aligned} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} &:= Q^* \begin{bmatrix} 0 & I \\ -C & -B \end{bmatrix} Z \begin{bmatrix} R \\ 0 \end{bmatrix} \\ &= Q^* \begin{bmatrix} I & 0 \\ 0 & A \end{bmatrix} Z \begin{bmatrix} R \\ 0 \end{bmatrix} X \\ &= \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} X. \end{aligned}$$

Equating (2, 1) blocks gives $T_{21}R = 0$, which implies $T_{21} = 0$. It is easy to see that the equations $T = Q^*FZ$ and $S = Q^*GZ$ can be completed to a generalized Schur decomposition while retaining $X = Z_{21}Z_{11}^{-1}$. Finally, we note that $Z_{21}Z_{11}^{-1} = Q_{11}T_{11}S_{11}^{-1}Q_{11}^{-1}$ follows from (11) and (13). □

A subtle aspect of Theorem 3 is worth noting. If A is singular then (14) shows that in the generalized Schur decomposition (10) Q_{21} or S_{11} (and hence Z_{11}) is singular. However, the 'converse' part of the proof shows that if a solvent exists then it is possible to choose a generalized Schur decomposition in which S_{11} and Z_{11} are nonsingular. For example, if $A = C = 0$ and $B = I$, so that the zero matrix is the only solvent, we can take $Q = Z = I$ and $S_{11} = Z_{11} = I$.

To see a connection between Theorem 3 and the construction of a solvent via the quadratic eigenvalue problem, note that (2) can be rewritten as the generalized eigenvalue problem

$$Fy = \begin{bmatrix} 0 & I \\ -C & -B \end{bmatrix} y = \lambda \begin{bmatrix} I & 0 \\ 0 & A \end{bmatrix} y = \lambda Gy, \qquad y = \begin{bmatrix} x \\ \lambda x \end{bmatrix}. \tag{16}$$

If A is nonsingular, as for an overdamped problem, then we can obtain an analogue of Theorem 3 based on a Schur decomposition of the matrix

$$\begin{bmatrix} 0 & I \\ -A^{-1}C & -A^{-1}B \end{bmatrix}.$$

3. Applications

We discuss briefly two applications of the quadratic matrix equation (1). Numerical examples of both of these applications are given in Section 8.

Quasi-birth–death processes are two-dimensional Markov chains with a block tridiagonal transition probability matrix. They are widely used as stochastic models in telecommunications, computer performance and inventory control. Analysis using the matrix-geometric method leads to three quadratic matrix equations whose elementwise minimal nonnegative solutions can be used to characterize most of the features of the Markov chain. An excellent reference is the recent book by Latouche & Ramaswami (1999).

A second application is the solution of the quadratic eigenvalue problem (2), which arises in the analysis of damped structural systems and vibration problems (Datta, 1995, Chapter 9; Lancaster, 1966). A standard approach is to convert the quadratic eigenvalue problem to a generalized eigenvalue problem of twice the dimension, $2n$: (16) is one of several possible reductions (Tisseur, 2000). The 'linearized' problem can be further converted to a standard eigenvalue problem of dimension $2n$ if A or C is nonsingular. However, as (3) shows, if we can find a solvent S of the associated quadratic matrix equation then the problem is reduced to solving two $n \times n$ eigenproblems: that of S and the generalized eigenproblem $(B + AS)x = -\lambda Ax$. If S can be found by working only with $n \times n$ matrices then this approach offers a potential saving of work and storage. We return to this possibility in Section 8.

4. Symbolic solution

Before attempting numerical solution we consider solving the quadratic matrix equation using a symbolic algebra package. We have experimented with MATLAB's Symbolic Math Toolbox (Moler & Costa, 1998), version 2.1 (Release 11), which makes use of Maple V Release 5.

MATLAB's syntax does not allow the quadratic matrix equation to be specified with a single command—the command `solve('A*X^2+B*X+C=0')` is syntactically valid, but it assumes that the variables are scalars, even if they have been defined as matrices. It is therefore necessary to provide the `solve` command with the n^2 individual equations. For $n = 2$ this is done by

```
syms w x y z
X = [w x; y z];
R = A*X^2 + B*X + C;
s = solve(R(1,1), R(2,1), R(1,2), R(2,2), w, x, y, z)
```

Our experience is that the `solve` command is remarkably effective for $n = 2$, finding all isolated solvents and even parametrized families of solvents in a few seconds on a 333 MHz Pentium. For $n \geqslant 3$ we have been able to find solvents with this approach only for very special A, B and C.

An algorithm for symbolic solution for $n = 2$ has been suggested by our colleague R. M. Wood. Suppose $A = I$ and C is nonsingular. Then X is nonsingular; write

$$X = \begin{bmatrix} w & x \\ y & z \end{bmatrix}, \qquad X^{-1} = u^{-1} \begin{bmatrix} z & -x \\ -y & w \end{bmatrix},$$

where $u = \det(X) = wz - yx$. The quadratic matrix equation can be written $X + B + CX^{-1} = 0$, which can be expressed as a linear system $Mp = g$, where $p = [w\ x\ y\ z]^T$ and the 4×4 matrix M depends on u. The system can be solved symbolically, yielding p as a rational function of u. Equating the numerators in $u = wz - yx$ leads to a sextic equation in u. Once the roots are found they can be substituted back into p to determine solvents X. We have found this approach less successful than using `solve` as described above. For example, for the problem (4) the sextic degenerates to a quartic and the method produces only the last four of the solvents in (5); for $u = 2$ (which is the determinant of the missing solvent) the matrix M is singular. This approach does not generalize to larger n.

Our conclusion is that symbolic solution appears to be of limited use except for $n = 2$, though we cannot rule out the possibility that more successful lines of symbolic attack can be developed. In the rest of this paper we consider numerical techniques.

5. Eigenvalue techniques

The eigenvalue-based constructions of Section 2 provide two ways to solve the quadratic matrix equation.

5.1 *Quadratic eigenvalue problem approach*

Suppose we solve the quadratic eigenvalue problem (2) and obtain eigenpairs $(\lambda_i, v_i)_{i=1}^{p}$, with $n \leqslant p \leqslant 2n$. To construct a solvent from these eigenpairs we need to identify a linearly independent set of n vectors v_i. This can be attempted by computing the QR factorization with column pivoting (Golub & Van Loan, 1996, Section 5.4.1)

$$[v_{i_1}\ \ldots\ v_{i_p}] := [v_1\ \ldots\ v_p]\Pi = Q[R_{11}\ R_{12}], \qquad Q, R_{11} \in \mathbb{C}^{n\times n},$$

where Π is a permutation matrix, Q is unitary and R_{11} is upper triangular. Provided that R_{11} is nonsingular the matrix $W = [v_{i_1}\ \ldots\ v_{i_n}]$ is nonsingular and $S = W\,\mathrm{diag}(\lambda_{i_1}, \ldots, \lambda_{i_n})W^{-1}$ is a solvent. This method fails when no linearly independent set of n eigenvectors v_i exists.

5.2 *Schur method*

In the second method, which we call the Schur method, we compute a generalized Schur decomposition $Q^*FZ = T$, $Q^*GZ = S$ of F and G in (8). Theorem 3 shows that if Z_{11} is nonsingular then $X = Z_{21}Z_{11}^{-1} = Q_{11}T_{11}S_{11}^{-1}Q_{11}^{-1}$ is a solvent. To obtain a nonsingular Z_{11} we may have to reorder the generalized Schur decomposition, which can be done using the methods of Van Dooren (1982) and Kågström & Poromaa (1996) (LAPACK 3.0 (Anderson *et al.*, 1999) has this capability). For an overdamped system we know from the properties stated in Section 2 that if the Schur form is ordered so that the eigenvalues t_{ii}/s_{ii} are arranged in increasing order then Z_{11} will be nonsingular.

The generalized Schur decomposition can be computed by the QZ algorithm (Golub & Van Loan, 1996, Section 7.7). This approach is numerically stable in that the computed Q and Z are nearly unitary and the computed T and S are the exact ones for a small normwise perturbation of F and G. These perturbations will not, in general, respect the structure of F

and G in (8), so may not correspond to a small perturbation of the quadratic matrix equation. However, a mixed forward–backward stability result for the quadratic matrix equation can be deduced in the special case where $\|A\|_2 = \|B\|_2 = \|C\|_2 = 1$, as a consequence of a result of Tisseur (2000, Theorem 7) concerning solution of the quadratic eigenvalue problem via a generalized eigenvalue problem.

The step of the Schur method that is numerically the most dangerous is the inversion of Z_{11}. However, the possibility of an ill-conditioned Z_{11} can be avoided (provided that Z_{11} is nonsingular) by scaling the quadratic matrix equation, adapting ideas from Kenney *et al.* (1989) for the Riccati equation. We need the following lemma, in which the condition number $\kappa(A) = \|A\|\|A^{-1}\|$.

LEMMA 4 Let $X = Z_{21}Z_{11}^{-1}$ be a solvent written in the form of Theorem 3 and let Z be the unitary matrix from the generalized Schur decomposition (10) of F and G. Then

$$\kappa_2(Z_{11}) \leqslant 1 + \|X\|_2^2.$$

Proof. The proof is very similar to that of Kenney *et al.* (1989, Lemma 1). Since Z is unitary, $Z_{11}^*Z_{11} + Z_{21}^*Z_{21} = I$, which implies $\|Z_{11}\|_2 \leqslant 1$. Using the nonsingularity of Z_{11},

$$\begin{aligned} I &= Z_{11}^*Z_{11} + Z_{11}^*(Z_{21}Z_{11}^{-1})^*Z_{21}Z_{11}^{-1}Z_{11} \\ &= Z_{11}^*Z_{11} + Z_{11}^*X^*XZ_{11} \\ &= Z_{11}^*(I + X^*X)Z_{11}. \end{aligned}$$

Hence $Z_{11}^{-1} = Z_{11}^*(I+X^*X)$, and the result follows on taking norms and using $\|Z_{11}\|_2 \leqslant 1$. □

The lemma shows that Z_{11} will be well conditioned if $\|X\|_2 \approx 1$. This can be achieved by the scaling

$$AX^2 + BX + C \longrightarrow \widetilde{A}(X/\rho)^2 + \widetilde{B}(X/\rho) + \widetilde{C}, \qquad \widetilde{A} = \rho^2 A, \quad \widetilde{B} = \rho B, \quad \widetilde{C} = C, \tag{17}$$

with $\rho = \|X\|_2$. From the relation

$$\begin{aligned} F - \lambda G &= \begin{bmatrix} 0 & I \\ -C & -B \end{bmatrix} - \lambda \begin{bmatrix} I & 0 \\ 0 & A \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \rho^{-1} \end{bmatrix} \left(\begin{bmatrix} 0 & I \\ -C & -\rho B \end{bmatrix} - \lambda\rho^{-1} \begin{bmatrix} I & 0 \\ 0 & \rho^2 A \end{bmatrix} \right) \begin{bmatrix} \rho & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

we see that the eigenvalues of the scaled pencil corresponding to $\widetilde{A}$, $\widetilde{B}$ and $\widetilde{C}$ are ρ^{-1} times those of the original pencil. Therefore, the same ordering of the Schur form can be used for the scaled pencil as for the original pencil.

Unfortunately, there is no generally applicable way to estimate any norm of X in advance. One possibility is to compute X and, if $\|X\|_F$ is found to be large, to repeat the computation on the scaled problem.

It is important to consider whether scaling worsens the conditioning or backward stability of the quadratic matrix equation. If we measure perturbations to A, B and C by

$$\left\| \left[\frac{\Delta A}{\|A\|_F} \quad \frac{\Delta B}{\|B\|_F} \quad \frac{\Delta C}{\|C\|_F} \right] \right\|_F,$$

then the condition number can be shown to be (Higham & Kim, 1999)

$$\Psi(X) = \|P^{-1}\left[\|A\|_F (X^2)^T \otimes I_n, \quad \|B\|_F X^T \otimes I_n, \quad \|C\|_F I_{n^2}\right]\|_2 / \|X\|_F,$$

where

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

It is easy to see that $\Psi(X)$ is invariant under the scaling (17). The effect of the scaling on the backward error of the computed solvent is unclear and merits further study. (A definition of backward error of an approximate solvent, and a computable expression for it, are given by Higham & Kim, 1999.)

6. Newton's method

A natural contender for solving the quadratic matrix equation is Newton's method, which has been investigated for this problem by Davis (1981, 1983) and Higham & Kim (1999). (Newton's method has also been considered for general degree matrix polynomials by Kratz & Stickel, 1987 and Shieh *et al.*, 1981.) The method can be derived by equating to zero the first-order part of the expansion

$$Q(X + E) = Q(X) + \big(AEX + (AX + B)E\big) + AE^2,$$

yielding the iteration

$$\left.\begin{array}{c} \text{Solve } AE_kX_k + (AX_k + B)E_k = -Q(X_k) \\ \text{Update } X_{k+1} = X_k + E_k \end{array}\right\} k = 1, 2, \ldots$$

Forming and solving the generalized Sylvester equation defining E_k takes at least $56n^3$ flops (Higham & Kim, 1999). The global convergence properties of Newton's method can be improved by incorporating line searches, so that E_k is regarded as a search direction and X_{k+1} is defined as $X_{k+1} = X_k + tE_k$, where t minimizes $\|Q(X_k + tE_k)\|_F$. Higham & Kim (1999) show that by exploiting the fact that Q is quadratic the optimal t can be computed exactly in $5n^3$ flops, yielding an exact line search method. They show that exact line searches improve the global convergence properties in theory and in practice. Newton's method with exact line searches has also been used by Benner & Byers (1998) and Benner *et al.* (1998) for solving algebraic Riccati equations.

While Newton's method is very attractive for solving the quadratic matrix equation, it has some weaknesses. Each iteration is rather expensive and many iterations can be required before quadratic convergence sets in; in the absence of a sufficiently good starting matrix, convergence cannot be guaranteed; and it is difficult to know in advance to which solvent the method will converge. In the next section we describe a class of methods for which the cost per iteration is much less and for which convergence to a particular solvent can be guaranteed for specified starting matrices, under certain conditions on the problem.

7. Functional iteration

A natural way to attempt to solve the quadratic matrix equation is to rewrite it in the form $X = F(X)$ and then define an iteration $X_{k+1} = F(X_k)$. This can be done in many ways, giving iterations such as

$$\begin{aligned} X_{k+1} &= (-A^{-1}(BX_k + C))^{1/2}, \\ X_{k+1} &= -B^{-1}(AX_k^2 + C), \\ X_{k+1} &= -A^{-1}(B + CX_k^{-1}). \end{aligned} \tag{18}$$

These iterations can not in general be transformed to a simpler form (such as diagonal form) and so it is difficult to obtain convergence results of practical applicability. However, it turns out that useful results for iteration (18) and some variants can be obtained from the analysis of an associated Bernoulli iteration.

To motivate the analysis we first consider briefly the scalar quadratic $ax^2 + bx + c = 0$, with $a \neq 0$. Iteration (18) can be written as

$$x_{k+1} = g(x_k), \qquad g(x) = -\frac{(c/x + b)}{a}. \tag{19}$$

From the standard theory of functional iteration (Stewart, 1996, Lec. 3) we know that the condition for convergence for starting values sufficiently close to a root α is $|g'(\alpha)| < 1$. Denote the two roots by α_1 and α_2 and note that $\alpha_1\alpha_2 = c/a$. Now

$$g'(\alpha_1) = \frac{c}{a\alpha_1^2} = \frac{\alpha_2}{\alpha_1}.$$

Thus we have local convergence to the larger root provided that the roots are of distinct modulus, which requires, in particular, that the discriminant be positive. Iteration (19) can be written $(ax_{k+1} + b)x_k + c = 0$ and a similar analysis shows that the iteration defined by $(ax_k + b)x_{k+1} + c = 0$ converges to the root of smaller modulus. It is clear that the matrix analogues of these iterations can converge only for problems satisfying some suitable generalization of the 'roots of distinct modulus' condition. In the next subsection we identify two solvents that are used to define the appropriate condition.

We assume throughout the rest of this section that A is nonsingular.

7.1 *Dominant and minimal solvents of $Q(X)$*

Since A is nonsingular, $Q(\lambda)$ in (2) has $2n$ eigenvalues, all finite, which we order by absolute value:

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_{2n}|. \tag{20}$$

Let $\lambda(A)$ denote the set of eigenvalues of A.

The solvents in the following definition play an important role in what follows.

DEFINITION 5 A solvent S_1 of $Q(X)$ is a dominant solvent if $\lambda(S_1) = \{\lambda_1, \ldots, \lambda_n\}$ and $|\lambda_n| > |\lambda_{n+1}|$, where the eigenvalues λ_i of $Q(\lambda)$ are ordered according to (20). A solvent S_2 of $Q(X)$ is a minimal solvent if $\lambda(S_2) = \{\lambda_{n+1}, \ldots, \lambda_{2n}\}$ and $|\lambda_n| > |\lambda_{n+1}|$.

A more elegant but less transparent definition is that a dominant solvent S of $Q(X)$ is one for which every eigenvalue of S exceeds in modulus every eigenvalue of the quotient $Q(\lambda)(\lambda I - S)^{-1}$ (Gohberg *et al.*, 1982, p 126), and similarly for a minimal solvent. Note that if S_1 is a dominant solvent and S_2 is a minimal solvent then

$$\min\{\,|\lambda| : \lambda \in \lambda(S_1)\,\} > \max\{\,|\lambda| : \lambda \in \lambda(S_2)\,\}.$$

We caution the reader that in referring to results on the convergence of Bernoulli's method some authors give imprecise or incorrect definitions of the dominant solvent: for example the definition that a dominant solvent is 'a solvent matrix whose eigenvalues strictly dominate the eigenvalues of all other solvents' (Kratz & Stickel, 1987) is far too restrictive.

It follows from the theory of λ-matrices (Gohberg *et al.*, 1982, Theorem 4.1) that a dominant solvent, if one exists, is unique, and likewise for a minimal solvent (recall that A is assumed nonsingular).

As we mentioned in Section 2, Lancaster has shown that the overdamping condition (Definition 1) is sufficient to ensure the existence of dominant and minimal solvents. The next theorem gives a sufficient condition of more general applicability.

THEOREM 6 Assume that the eigenvalues of $Q(\lambda)$, ordered as in (20), satisfy $|\lambda_n| > |\lambda_{n+1}|$ and that corresponding to $\{\lambda_i\}_{i=1}^{n}$ and $\{\lambda_i\}_{i=n+1}^{2n}$ there are two sets of linearly independent eigenvectors $\{v_1, \ldots, v_n\}$ and $\{v_{n+1}, \ldots, v_{2n}\}$. Then there exists a dominant solvent and a minimal solvent of $Q(X)$.

Proof. Let $W_1 = [v_1, \ldots, v_n]$ and $W_2 = [v_{n+1}, \ldots, v_{2n}]$. Since W_1 and W_2 are nonsingular, we can define

$$S_1 = W_1 \operatorname{diag}(\lambda_1, \ldots, \lambda_n) W_1^{-1}, \quad S_2 = W_2 \operatorname{diag}(\lambda_{n+1}, \ldots, \lambda_{2n}) W_2^{-1}.$$

It is easily seen that S_1 is a dominant solvent of $Q(X)$ and S_2 a minimal solvent. □

7.2 *Bernoulli iteration*

Bernoulli's method is an old method for finding the dominant root (if there is one) of a scalar polynomial (Henrici, 1964, Chapter 7; Young & Gregory, 1973, Section 5.7). It is essentially the power method applied to the companion matrix associated with the polynomial. In this subsection we consider a generalization of Bernoulli's method to the quadratic matrix polynomial. This generalization is not new, having been analysed previously for general degree polynomials by Busby & Fair (1975), Dennis *et al.* (1978) and Gohberg *et al.* (1982, Section 4.2). Our treatment of convergence differs from those in the references just cited, in particular by not assuming a monic quadratic. We follow the approach of Dennis *et al.* (1978) but correct some ambiguities and deficiencies in that analysis.

Consider the matrix recurrence

$$AY_{i+1} + BY_i + CY_{i-1} = 0, \qquad Y_0, Y_1 \text{ given.} \tag{21}$$

Define

$$V(S_1, S_2) = \begin{bmatrix} I & I \\ S_1 & S_2 \end{bmatrix}.$$

THEOREM 7 If S_1 and S_2 are solvents of $Q(X)$ for which $V(S_1, S_2)$ is nonsingular then

$$Y_i = S_1^i \Omega_1 + S_2^i \Omega_2 \tag{22}$$

is the general solution to *(21)*, where Ω_1 and Ω_2 are determined by the initial conditions.

Proof. For Y_i defined by (22) we have

$$\begin{aligned} AY_{i+1} + BY_i + CY_{i-1} &= (AS_1^2 + BS_1 + C)S_1^{i-1}\Omega_1 + (AS_2^2 + BS_2 + C)S_2^{i-1}\Omega_2 \\ &= 0, \end{aligned}$$

so Y_i is a solution to (21). For $i = 0$ and $i = 1$, (22) gives

$$\begin{bmatrix} I & I \\ S_1 & S_2 \end{bmatrix} \begin{bmatrix} \Omega_1 \\ \Omega_2 \end{bmatrix} = \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}$$

and since the coefficient matrix $V(S_1, S_2)$ is nonsingular, Ω_1 and Ω_2 are uniquely determined by Y_0 and Y_1. Finally, suppose that Z_i is a solution of (21) and let Y_i be the solution (22) with $Y_0 = Z_0$ and $Y_1 = Z_1$. Then

$$AY_2 = -BY_1 - CY_0 = -BZ_1 - CZ_0 = AZ_2.$$

Since A is nonsingular, $Z_2 = Y_2$. By induction, $Z_i = Y_i$ for all i. □

Note that $V(S_1, S_2)$ is a special case of a block Vandermonde matrix. By analogy with the scalar Vandermonde, it might be thought that for general X_1 and X_2 the condition $\lambda(X_1) \cap \lambda(X_2) = \emptyset$ would ensure the nonsingularity of $V(X_1, X_2)$. That it does not is shown by the example (Dennis *et al.*, 1976)

$$X_1 = \begin{bmatrix} 2 & 0 \\ -2 & 1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 4 & 2 \\ 0 & 3 \end{bmatrix},$$

for which $\det(V(X_1, X_2)) = \det(X_2 - X_1) = \det\left(\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\right) = 0$. The next result (which has weaker hypotheses than Dennis *et al.* 1976, Theorem 6.1) shows that the eigenvalue condition *is* sufficient to ensure nonsingularity when X_1 and X_2 are solvents of a quadratic matrix equation.

THEOREM 8 If S_1 and S_2 are solvents of $Q(X)$ with $\lambda(S_1) \cap \lambda(S_2) = \emptyset$ then $V(S_1, S_2)$ is nonsingular.

Proof. Suppose that $V(S_1, S_2)$ is singular. Let $v \in N := \text{null}(V(S_1, S_2))$. Then $[S_1\ S_2]v = 0$ and the identity

$$\begin{bmatrix} A^{-1}C & A^{-1}B \end{bmatrix} \begin{bmatrix} I & I \\ S_1 & S_2 \end{bmatrix} = -\begin{bmatrix} S_1^2 & S_2^2 \end{bmatrix}$$

implies $[S_1^2\ S_2^2]v = 0$, and so

$$0 = \begin{bmatrix} S_1 & S_2 \\ S_1^2 & S_2^2 \end{bmatrix} v = \begin{bmatrix} I & I \\ S_1 & S_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} v. \tag{23}$$

Thus $\mathrm{diag}(S_1, S_2)v \in N$, which means that N is an invariant subspace for $\mathrm{diag}(S_1, S_2)$. Hence N contains an eigenvector w of $\mathrm{diag}(S_1, S_2)$. Since the eigenvalues of S_1 are distinct from those of S_2, w has the form $w = [w_1^T\ 0]^T$ or $[0\ w_2^T]^T$. In either case

$$0 = V(S_1, S_2)w = \begin{bmatrix} I & I \\ S_1 & S_2 \end{bmatrix} w$$

implies $w = 0$, which is a contradiction. Thus $V(S_1, S_2)$ is nonsingular. □

To obtain a convergence result for Bernoulli's method, we need the following lemma.

LEMMA 9 Let Z_1 and Z_2 be square matrices such that

$$\min\{\,|\lambda| : \lambda \in \lambda(Z_1)\,\} > \max\{\,|\lambda| : \lambda \in \lambda(Z_2)\,\}.$$

Then Z_1 is nonsingular and, for any matrix norm,

$$\lim_{i\to\infty} \|Z_2^i\|\,\|Z_1^{-i}\| = 0.$$

Proof. See Gohberg *et al.* (1982, Lemma 4.9). □

THEOREM 10 Suppose that $Q(X)$ has a dominant solvent S_1 and a minimal solvent S_2 and let Y_i be the solution to the recurrence (21) with $Y_0 = 0$ and $Y_1 = I$. Then Y_i is nonsingular for sufficiently large i and

$$\lim_{i\to\infty} Y_i Y_{i-1}^{-1} = S_1.$$

Proof. Theorem 8 implies that $V(S_1, S_2)$ is nonsingular and so (22) is the general solution to (21). The initial conditions give

$$\Omega_1 + \Omega_2 = 0, \qquad S_1\Omega_1 + S_2\Omega_2 = I,$$

which imply

$$(S_1 - S_2)\Omega_1 = I.$$

Hence Ω_1 is nonsingular. First, we show that Y_i is nonsingular for sufficiently large i. From (22), since S_1 is nonsingular,

$$Y_i = S_1^i(\Omega_1 + S_1^{-i}S_2^i\Omega_2).$$

Since $S_1^{-i}S_2^i \to 0$ by Lemma 9 it follows that Y_i is nonsingular for sufficiently large i. Using (22) again we have

$$\begin{aligned} Y_iY_{i-1}^{-1} &= (S_1^i\Omega_1 + S_2^i\Omega_2)(S_1^{i-1}\Omega_1 + S_2^{i-1}\Omega_2)^{-1} \\ &= (S_1 + S_2^i\Omega_2\Omega_1^{-1}S_1^{-(i-1)})(S_1^{i-1}\Omega_1)\big((I + S_2^{i-1}\Omega_2\Omega_1^{-1}S_1^{-(i-1)})(S_1^{i-1}\Omega_1)\big)^{-1} \\ &= (S_1 + S_2^i\Omega_2\Omega_1^{-1}S_1^{-(i-1)})(I + S_2^{i-1}\Omega_2\Omega_1^{-1}S_1^{-(i-1)})^{-1}. \end{aligned}$$

From Lemma 9 we deduce

$$\lim_{i\to\infty} S_2^i\Omega_2\Omega_1^{-1}S_1^{-(i-1)} = 0, \qquad \lim_{i\to\infty} S_2^{i-1}\Omega_2\Omega_1^{-1}S_1^{-(i-1)} = 0,$$

which implies the result. □

We give an example from Gohberg *et al.* (1982, Example 4.4) to illustrate how Bernoulli iteration can fail to produce a solvent. Consider

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

The eigenvalues of $Q(\lambda)$ are $-1, 0, 0, 1$ and a dominant solvent is $\left[\begin{smallmatrix} 1 & -1 \\ 0 & -1 \end{smallmatrix}\right]$. However, there is no solvent with eigenvalues 0, 0, hence no minimal solvent. With $Y_0 = 0$ and $Y_1 = I$, the recurrence (21) gives

$$Y_2 = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}, \quad Y_3 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$

and then $Y_{2j} = Y_2$, $Y_{2j+1} = Y_3$ for $j \geqslant 1$. Hence all the Y_i are singular.

7.3 *Iterations for computing a solvent*

For practical computation it is desirable to rewrite (21) in the form of an iteration that directly computes a solvent. Postmultiplying (21) by Y_{i-1}^{-1}, we have

$$AY_{i+1}Y_{i-1}^{-1} + BY_iY_{i-1}^{-1} + C = 0,$$

which can be written as

$$(AY_{i+1}Y_i^{-1} + B)Y_iY_{i-1}^{-1} + C = 0.$$

Defining $X_i = Y_{i+1}Y_i^{-1}$ and setting $Y_0 = 0$ and $Y_1 = I$, we have the iteration

$$(AX_i + B)X_{i-1} + C = 0, \qquad X_1 = -A^{-1}B. \tag{24}$$

Similarly, postmultiplying (21) by Y_{i+1}^{-1} and letting $W_i = Y_iY_{i+1}^{-1}$, we obtain

$$A + (B + CW_{i-1})W_i = 0, \qquad W_0 = 0. \tag{25}$$

Note that this iteration is attempting to solve $CX^2 + BX + A = 0$, whose nonsingular solvents are the inverses of solvents of $Q(X) = 0$.

We deduce from Theorem 10 that if $Q(X)$ has a dominant solvent S_1 and a minimal solvent S_2 and the sequences X_i and W_i are defined then

$$\lim_{i\to\infty} X_i = S_1, \qquad \lim_{i\to\infty} W_i = S_1^{-1}.$$

Note from the proof of Theorem 10 that the role of Y_0 and Y_1 is simply to ensure that Ω_1 is nonsingular. Most choices of Y_0 and Y_1 will have the same effect and so we can try different starting matrices in (24) and (25), as may be necessary if an iteration breaks down.

Iterations (24) and (25) are fixed point iterations obtainable in an obvious way from $AX^2 + BX + C = 0$. Equally plausible variants are

$$(AX_{i-1} + B)X_i + C = 0, \qquad X_0 = 0 \tag{26}$$

and

$$A + (B + CW_i)W_{i-1} = 0, \qquad W_1 = -C^{-1}B. \tag{27}$$

These are related to the Bernoulli recurrence

$$AZ_{i-1} + BZ_i + CZ_{i+1} = 0, \qquad Z_0 = 0, \quad Z_1 = I. \tag{28}$$

The analysis above can be adapted to show that if $Q(X)$ has a dominant solvent S_1 and a nonsingular minimal solvent S_2 then the solution to (28) is $Z_i = S_1^{-i}\Omega_1 + S_2^{-i}\Omega_2$, with Ω_1 and Ω_2 determined by the initial conditions, and then that

$$\lim_{i\to\infty} Z_{i+1}Z_i^{-1} = S_2^{-1}.$$

Hence, as long as the sequences W_i and X_i are defined,

$$\lim_{i\to\infty} X_i = S_2, \qquad \lim_{i\to\infty} W_i = S_2^{-1}.$$

The convergence of all these iterations is linear, with error decreasing as $\|S_2^i\|\,\|S_1^{-i}\|$, and therefore with asymptotic error constant of order

$$e = \frac{|\lambda_n|}{|\lambda_{n+1}|} \tag{29}$$

(given the ordering (20)).

We explain a connection with continued fractions (Busby & Fair, 1975, 1994; Fair, 1971). Consider the quadratic $X^2 - BX - C = 0$. Assuming that X is nonsingular we have $X = B + CX^{-1}$ and then

$$X = B + C(B + CX^{-1})^{-1}.$$

On recurring this process we obtain a matrix continued fraction with approximants $X_k = R_kS_k^{-1}$ generated by

$$\begin{aligned} R_{k+1} &= BR_k + CR_{k-1}, \qquad R_0 = B, \quad R_1 = B^2 + C, \\ S_{k+1} &= BS_k + CS_{k-1}, \qquad S_0 = I, \quad S_1 = B. \end{aligned}$$

Clearly, $R_k = S_{k+1}$ and for Y_k defined in Theorem 10 we have $S_k = Y_{k+1}$, so the approximant $X_k = Y_{k+2}Y_{k+1}^{-1}$. Thus the continued fraction generates the same sequence as the iteration (24).

Finally, we give some numerical examples. Consider the equation

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

The eigenvalues of $Q(\lambda)$ are $1, 1, -2$ and -2 and there is only one eigenvector corresponding to each eigenvalue, the vector $[1\ 0]^T$ in both cases. It can be deduced that there are precisely two solvents,

$$\begin{bmatrix} -2 & -1/3 \\ 0 & -2 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1/3 \\ 0 & 1 \end{bmatrix},$$

the first of which is dominant, the second minimal. We applied the four iterations (24)–(27) in MATLAB using convergence test

$$\frac{\|X_i - X_{i-1}\|_1}{\|X_i\|_1} \leqslant u, \tag{30}$$

where $u = 2^{-53} \approx 1{\cdot}1 \times 10^{-16}$ is the unit roundoff. Each iteration converged in about 57 iterations to the expected solvent; this is the expected number of iterations since the error constant $e = 1/2$ in (29).

For the next example we take the equation

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

This is (4) with the coefficient matrices in reverse order, so its solvents are the inverses of the solvents of (4). The solvent $S_1 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1/2 \end{smallmatrix}\right]$ is dominant, but there is no minimal solvent. Theorem 10 is therefore not applicable. Applying iteration (24) we found that after 49 iterations $\|X_i - S_2\|_1 \approx 5 \times 10^{-8}$, where $S_2 = \left[\begin{smallmatrix} 1 & -2/3 \\ 0 & 1/3 \end{smallmatrix}\right]$. Thereafter the relative changes $\|X_{i+1} - X_i\|_1 / \|X_{i+1}\|_1$ increased until on the 91st iteration they started decreasing again. After 177 iterations the iteration had converged to S_1. Suspicious that this behaviour was caused by the effects of rounding error, we ran the iteration in exact arithmetic using MATLAB's Symbolic Math Toolbox. This time the iteration converged to S_2 (with monotonic decrease of the relative changes). Thus the condition that both dominant and minimal solvents exist (Theorem 10) is sufficient but not necessary for convergence of iteration (24) and when a dominant solvent but not a minimal one exists convergence can be to a non-dominant solvent.

8. Numerical experiments

We illustrate our methods on two practical problems.

We consider first a quadratic eigenvalue problem arising from a damped mass–spring system in which each mass is connected to its neighbour by a spring and a damper and also to the ground by a spring and a damper; see Tisseur (2000) for the details. We have chosen the masses and the spring and damper constants to give an $n \times n$ problem with $A = I$ and

$$B = \begin{bmatrix} 20 & -10 & & & & \\ -10 & 30 & -10 & & & \\ & -10 & 30 & -10 & & \\ & & -10 & \ddots & \ddots & \\ & & & \ddots & 30 & -10 \\ & & & & -10 & 20 \end{bmatrix}, \quad C = \begin{bmatrix} 15 & -5 & & & & \\ -5 & 15 & -5 & & & \\ & -5 & \ddots & \ddots & & \\ & & \ddots & \ddots & -5 & \\ & & & -5 & 15 & \end{bmatrix}.$$

We took $n = 100$. Since $\lambda_{\min}(B)^2 - 4\|A\|_2\|C\|_2 = 1{\cdot}9 \times 10^{-2} > 0$, the quadratic eigenvalue problem is overdamped. Hence all eigenvalues of the quadratic eigenvalue problem are real and nonpositive, the quadratic matrix equation has a dominant and a

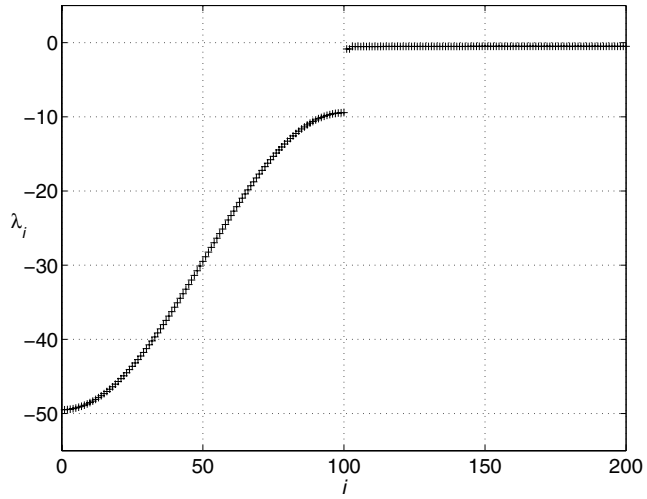

FIG. 1. Eigenvalues λ_i of the quadratic eigenvalue problem for a damped mass–spring system.

minimal solvent, and the Bernoulli iterations (24)–(27) all converge, provided that the iterates are defined. Figure 1 displays the eigenvalue distribution.

We applied the Bernoulli iterations and Newton's method to the quadratic matrix equation. The starting matrix for Newton's method was

$$X_0 = \left(\frac{\|B\|_F + \sqrt{\|B\|_F^2 + 4\|A\|_F\|C\|_F}}{2\|A\|_F} \right) I, \tag{31}$$

as in Higham & Kim (1999), and the iteration was terminated when

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

The Bernoulli iterations used the default starting matrices and the same stopping test as in (30) but with tolerance nu.

The Bernoulli iterations converged to the expected solvents in 13–15 iterations. This relatively quick convergence is consistent with the value 9×10^{-2} of the asymptotic error constant (29). Newton's method required 7 iterations without line searches and 6 iterations with exact line searches, converging to the minimal solvent in both cases.

A comparison of execution times on a 333 MHz Pentium II is instructive. The Bernoulli iterations converged after 3 seconds, while Newton's method took 45 seconds without line searches and 40 seconds with exact line searches. Computing the eigenvalues of the quadratic eigenvalue problem using MATLAB's `polyeig` (which solves a generalized eigenproblem of twice the dimension) took 17 seconds, which can be reduced to 8 seconds if we modify the M-file to avoid computing eigenvectors. Computing the eigenvalues of a solvent took 0·11 seconds. Thus computing the dominant solvent and the minimal solvent

by Bernoulli iteration and then finding their eigenvalues was faster than using `polyeig` to compute the eigenvalues directly.

This example shows that Bernoulli iteration can be significantly faster than Newton's method and that solving the quadratic eigenvalue problem via the associated quadratic matrix equation can be a viable approach.

Next, we consider a model from Bean *et al.* (1997) for the population of the bilby, an endangered Australian marsupial. Define the 5×5 matrix

$$Q(g,x) = \begin{bmatrix} gx_1 & (1-g)x_1 & 0 & 0 & 0 \\ gx_2 & & (1-g)x_2 & 0 & 0 \\ gx_3 & & 0 & (1-g)x_3 & 0 \\ gx_4 & & 0 & & (1-g)x_4 \\ gx_5 & & 0 & & (1-g)x_5 \end{bmatrix}.$$

The model is a quasi-birth–death process some of whose key properties are captured by the elementwise minimal solution $R_{\min}$ of the equation

$$R = \beta(A_0 + RA_1 + R^2A_2), \qquad A_0 = Q(g,b), \quad A_1 = Q(g, e-b-d), \quad A_2 = Q(g,d),$$

where b and d are vectors of probabilities and e is the vector of ones. We chose

$$g = 0{\cdot}2, \quad b = [1,\ 0{\cdot}4,\ 0{\cdot}25,\ 0{\cdot}1,\ 0], \quad d = [0,\ 0{\cdot}5,\ 0{\cdot}55,\ 0{\cdot}8,\ 1],$$

as in Bean *et al.* (1997), and $\beta = 0{\cdot}5$. We rewrite the quadratic equation as

$$AX^2 + BX + C = 0, \qquad A = \beta A_2^T, \quad B = \beta A_1^T - I, \quad C = \beta A_0^T.$$

Note that we make no attempt here to exploit the origin of this equation; doing so is an interesting problem to which we refer to Latouche & Ramaswami (1999).

Applying the Schur method (Section 5.2), with the generalized Schur decomposition ordered to maximize the absolute values of the diagonal elements of S_{11}, we obtain the desired minimal $R_{\min}$. The quadratic eigenvalue problem method (Section 5.1), using QR factorization with column pivoting, also produced $R_{\min}$. Newton's method converged with and without line searches with starting matrix (31), in 10 and 8 iterations, respectively; it converged to the same matrix in each case, but not to $R_{\min}$.

None of the Bernoulli iterations (24)–(27) is applicable because both A and C are singular, having a column of zeros. However, if we shift coordinates by defining $Y = X - I$ we have the equation $AY^2 + (2A+B)Y + A + B + C = 0$, and $A + B + C$ is nonsingular. Iteration (26) is now applicable and it converges to $R_{\min} - I$ in 111 iterations.

9. Concluding remarks

Over one hundred years after it was first investigated by Sylvester, the quadratic matrix equation (1) still poses interesting challenges. As we have seen, the theory and numerical solution bring together many different aspects of numerical linear algebra.

Our main contributions are twofold. We have given a new characterization of solvents in terms of the generalized Schur decomposition (Theorem 3) and a corresponding numerical method, with scaling to improve the accuracy. We have also given a thorough treatment of

functional iteration methods based on Bernoulli's method, identifying four iterations that converge to a dominant solvent, a minimal solvent, and the inverses of a dominant and minimal solvent.

The possibility of solving the quadratic eigenvalue problem via the quadratic matrix equation instead of solving a generalized eigenvalue problem of twice the dimension has been mentioned before in the literature, for example by Davis (1981). Our experiment in Section 8 appears to be the first demonstration that the quadratic matrix equation approach can be the faster. Whenever there is a large gap between the n smallest and n largest eigenvalues (in absolute value), solving the quadratic eigenvalue problem via Bernoulli iteration on the quadratic matrix equation is worth considering.

The quadratic matrix equation is just one of many nonlinear matrix equations that deserve further study. We hope that this work stimulates other researchers to attack this interesting and challenging class of problems.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council grant GR/L76532.

REFERENCES

ANDERSON, B. D. O. 1966 Solution of quadratic matrix equations. *Electron. Lett.* **2**, 371–372.

ANDERSON, E., BAI, Z., BISCHOF, C. H., BLACKFORD, S., DEMMEL, J. W., DONGARRA, J. J., DU CROZ, J. J., GREENBAUM, A., HAMMARLING, S. J., MCKENNEY, A., & SORENSEN, D. C. 1999 *LAPACK Users' Guide*. 3rd edn., Philadelphia, PA: Society for Industrial and Applied Mathematics.

BEAN, N. G., BRIGHT, L., LATOUCHE, G., PEARCE, C. E. M., POLLETT, P. K., & TAYLOR, P. G. 1997 The quasi-stationary behavior of quasi-birth-and-death processes. *Ann. Appl. Prob.* **7**, 134–155.

BENNER, P. & BYERS, R. 1998 An exact line search method for solving generalized continuous-time algebraic Riccati equations. *IEEE Trans. Automat. Control* **43**, 101–107.

BENNER, P., BYERS, R., QUINTANA-ORTÍ, E. S., & QUINTANA-ORTÍ, G. 1998 Solving algebraic Riccati equations on parallel computers using Newton's method with exact line search. *Report 98-05*, Zentrum fürTechnomathematik, Universität Bremen. Bremen, Germany, To appear in *Parallel Computing*.

BUSBY, R. C. & FAIR, W. 1975 Iterative solution of spectral operator polynomial equations and a related continued fraction. *J. Math. Anal. Appl.* **50**, 113–134.

BUSBY, R. C. & FAIR, W. 1994 Quadratic operator equations and periodic operator continued fractions. *J. Comput. Appl. Math.* **54**, 377–387.

DATTA, B. N. 1995 *Numerical Linear Algebra and Applications*. Pacific Grove, CA: Brooks/Cole Publishing Company.

DAVIS, G. J. 1981 Numerical solution of a quadratic matrix equation. *SIAM J. Sci. Stat. Comput.* **2**, 164–175.

DAVIS, G. J. 1983 Algorithm 598: an algorithm to compute solvents of the matrix equation $AX^2 + BX + C = 0$. *ACM Trans. Math. Software* **9**, 246–254.

DENNIS, J. E. JR., TRAUB, J. F., & WEBER, R. P. 1976 The algebraic theory of matrix polynomials. *SIAM J. Numer. Anal.* **13**, 831–845.

DENNIS, J. E. JR., TRAUB, J. F., & WEBER, R. P. 1978 Algorithms for solvents of matrix polynomials. *SIAM J. Numer. Anal.* **15**, 523–533.

DUFFIN, R. J. 1955 A minimax theory for overdamped networks. *J. Rat. Mech. Anal.* **4**, 221–233.

EISENFELD, J. 1973 Operator equations and nonlinear eigenparameter problems. *J. Funct. Anal.* **12**, 475–490.

FAIR, W. 1971 Noncommutative continued fractions. *SIAM J. Math. Anal.* **2**, 226–232.

GOHBERG, I., LANCASTER, P., & RODMAN, L. 1982 *Matrix Polynomials*. New York: Academic.

GOLUB, G. H. & VAN LOAN, C. F. 1996 *Matrix Computations*. 3rd edn., Baltimore, MD: Johns Hopkins University Press.

HENRICI, P. 1964 *Elements of Numerical Analysis*. New York: Wiley.

HIGHAM, N. J. 1987 Computing real square roots of a real matrix. *Linear Algebra Appl.* **88/89**, 405–430.

HIGHAM, N. J. & KIM, H.-M. 1999 Solving a quadratic matrix equation by Newton's method with exact line searches. *Numerical Analysis Report No. 339*, Manchester Centre for Computational Mathematics, Manchester, England.

HORN, R. A. & JOHNSON, C. R. 1991 *Topics in Matrix Analysis*. Cambridge: Cambridge University Press.

KÅGSTRÖM, B. & POROMAA, P. 1996 Computing eigenspaces with specified eigenvalues of a regular matrix pair (A, B) and condition estimation: theory, algorithms and software. *Numer. Algor.* **12**, 369–407.

KENNEY, C., LAUB, A. J., & WETTE, M. 1989 A stability-enhancing scaling procedure for Schur–Riccati solvers. *Math. Control Signals Systems* **12**, 241–250.

KRATZ, W. & STICKEL, E. 1987 Numerical solution of matrix polynomial equations by Newton's method. *IMA J. Numer. Anal.* **7**, 355–369.

KREIN, M. G. & LANGER, H. 1978 On some mathematical principles in the linear theory of damped oscillations of continua II. *Integral Equations Operator Theory* **1**, 539–566.

LANCASTER, P. 1966 *Lambda-Matrices and Vibrating Systems*. Oxford: Pergamon.

LANCASTER, P. & RODMAN, L. 1995 *Algebraic Riccati Equations*. Oxford: Oxford University Press.

LANCASTER, P. & ROKNE, J. G. 1977 Solutions of nonlinear operator equations. *SIAM J. Math. Anal.* **8**, 448–457.

LATOUCHE, G. & RAMASWAMI, V. 1999 *Introduction to Matrix Analytic Methods in Stochastic Modelling*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

LAUB, A. J. 1979 A Schur method for solving algebraic Riccati equations. *IEEE Trans. Automat. Control* AC-**24**, 913–921.

MCFARLAND, J. E. 1958 An iterative solution of the quadratic equation in Banach space. *Proc. Am. Math. Soc.* **9**, 824–830.

MOLER, C. & COSTA, P. J. 1998 *Symbolic Math Toolbox Version 2: User's Guide*. Natick, MA, USA: The MathWorks, Inc.

POTTER, J. E. 1966 Matrix quadratic solutions. *SIAM J. Appl. Math.* **14**, 496–501.

SHIEH, L. S., TSAY, Y. T., & COLEMAN, N. P. 1981 Algorithms for solvents and spectral factors of matrix polynomials. *Int. J. Systems Sci.* **12**, 1303–1316.

STEWART, G. W. 1996 *Afternotes on Numerical Analysis*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

STEWART, G. W. & SUN, J. 1990 *Matrix Perturbation Theory*. London: Academic.

SYLVESTER, J. J. 1973 *The Collected Mathematical Papers of James Joseph Sylvester*. vol. 4 (1882–1897), New York: Chelsea. Corrected reprint, published in four volumes, of work published

in four volumes by Cambridge University Press 1904–1912.

SYLVESTER, J. J. 1884 On Hamilton's quadratic equation and the general unilateral equation in matrices. *Phil. Mag.* **18**, 454–458. Reprinted in Sylvester (1973) pp. 231–235.

SYLVESTER, J. J. 1885 On the trinomial unilateral quadratic equation in matrices of the second order. *Quart. J. Math.* **20**, 305–312. Reprinted in Sylvester (1973) pp. 272–277.

TISSEUR, F. 2000 Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.* **309**, 339–361.

VAN DOOREN, P. M. 1981 A generalized eigenvalue approach for solving Riccati equations. *SIAM J. Sci. Stat. Comput.* **2**, 121–135.

VAN DOOREN, P. M. 1982 Algorithm 590: DSUBSP and EXCHQZ: FORTRAN subroutines for computing deflating subspaces with specified spectrum. *ACM Trans. Math. Software* **8**, 376–382.

YOUNG, D. M. & GREGORY, R. T. 1973 *A Survey of Numerical Mathematics*. vol. 2, Reading, MA: Addison-Wesley, Reprinted by Dover, New York, 1988.