Electronic Journal of Linear Algebra

Volume 34 *Volume 34 (2018)* Article 10

2018

Fast verified computation for the solvent of the quadratic matrix equation

Shinya Miyajima
Iwate University, miyajima@iwate-u.ac.jp

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

FAST VERIFIED COMPUTATION FOR THE SOLVENT OF THE QUADRATIC MATRIX EQUATION*

SHINYA MIYAJIMA†

Abstract. Two fast algorithms for numerically computing an interval matrix containing the solvent of the quadratic matrix equation $AX^2+BX+C=0$ with square matrices A, B, C and X are proposed. These algorithms require only cubic complexity, verify the uniqueness of the contained solvent, and do not involve iterative process. Let $\tilde{X}$ be a numerical approximation to the solvent. The first and second algorithms are applicable when A and $A\tilde{X}+B$ are nonsingular and numerically computed eigenvector matrices of $\tilde{X}^T$ and $\tilde{X}+A^{-1}B$, and $\tilde{X}^T$ and $(A\tilde{X}+B)^{-1}A$ are not ill-conditioned, respectively. The first algorithm moreover verifies the dominance and minimality of the contained solvent. Numerical results show efficiency of the algorithms.

Key words. Quadratic matrix equation, Verified numerical computation, Dominant solvent, Minimal solvent.

AMS subject classifications. 15A24, 65G20.

1. Introduction.

Consider the following quadratic matrix equation:

$$Q(X) := AX^2 + BX + C = 0, \tag{1.1}$$

where $A, B, C \in \mathbb{C}^{n\times n}$ are given and $X \in \mathbb{C}^{n\times n}$ is to be solved. A solution of (1.1) is called a solvent of $Q(X)$. The equation (1.1) arises in many applications such as multivariate rational expectations models [3], noisy Wiener-Hopf problems for Markov chains [5] and quasi-birth death process [8]. The other application is the solution of the quadratic eigenvalue problem

$$Q(\lambda)x := (\lambda^2 A + \lambda B + C)x = 0, \tag{1.2}$$

where $\lambda \in \mathbb{C}$ is an eigenvalue and $x \in \mathbb{C}^n \setminus \{0\}$ is an eigenvector corresponding to λ. The problem (1.2) arises in the analysis of damped structural systems and vibration problems [8, 9]. If X_* is a solvent of $Q(X)$, it then holds that

$$Q(\lambda) = -(AX_* + B + \lambda A)(X_* - \lambda I_n), \tag{1.3}$$

where I_n is the $n \times n$ identity matrix. Therefore, the eigenvalues of (1.2) are those of X_* together with those of the generalized eigenvalue problem

$$-(AX_* + B)x = \lambda Ax. \tag{1.4}$$

The problem (1.4) has n eigenvalues if and only if A is nonsingular (see [4, Section 7.7.1], e.g.). Hence, (1.2) has $2n$ eigenvalues if and only if A is nonsingular. Suppose A is nonsingular. Then, (1.2) has $2n$ eigenvalues, all finite and can be ordered by their absolute values as

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{2n}|. \tag{1.5}$$

*Received by the editors on October 5, 2017. Accepted for publication on February 20, 2018. Handling Editor: James G. Nagy.

†Faculty of Science and Engineering, Iwate University, Ueda, Morioka, 020-8551, Japan (miyajima@iwate-u.ac.jp). Supported by JSPS KAKENHI Grant Number JP16K05270.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

Let $\lambda(M)$ denote the spectrum of $M \in \mathbb{C}^{n\times n}$. A solvent X_D of $Q(X)$ is called a *dominant solvent* if $\lambda(X_D) = \{\lambda_1, \ldots, \lambda_n\}$ and $|\lambda_n| > |\lambda_{n+1}|$. A solvent X_M of $Q(X)$ is called a *minimal solvent* if $\lambda(X_M) = \{\lambda_{n+1}, \ldots, \lambda_{2n}\}$ and $|\lambda_n| > |\lambda_{n+1}|$. Numerical algorithms for computing the solvents are extensively studied (see [5, 8, 9, 11], e.g.).

The work presented in this paper addresses the problem of computing a verified solvent of $Q(X)$, specifically, numerically computing an interval matrix which is guaranteed to contain the solvent. The equation (1.1) may be written as nonlinear systems in $\mathbb{C}^{n^2}$, so that the verified computation of the solvent seems to be possible by executing a known verification algorithm for nonlinear systems (e.g. [15, 17]). On the other hand, this approach involves $\mathcal{O}(n^6)$ operations, which is prohibitively large for large n. In order to reduce the computational cost, Hashemi and Dehghan [6] have proposed two fast iterative verification algorithms. They skillfully exploit the special structure of $Q(X)$. These algorithms require only $\mathcal{O}(n^3)$ operations per iteration. The first and second algorithms are applicable when A and B are nonsingular, respectively. The first algorithm moreover verifies the uniqueness of the contained solvent.

The purpose of this paper is to propose two algorithms for computing the verified solvent. These algorithms also require only $\mathcal{O}(n^3)$ operations, verify the uniqueness, and do not involve iterative process. Let $\tilde{X}$ be a numerical result for the solvent. The first and second algorithms are applicable when A and $A\tilde{X}+B$ are nonsingular and numerically computed eigenvector matrices of $\tilde{X}^T$ and $\tilde{X}+A^{-1}B$, and $\tilde{X}^T$ and $(A\tilde{X}+B)^{-1}A$ are not ill-conditioned, and do not assume but prove the nonsingularities of A and $A\tilde{X}+B$, respectively. The first algorithm moreover verifies the dominance and minimality of the contained solvent.

This paper is organized as follows: Section 2 introduces notations and theories used in this paper. Sections 3 and 4 propose the first and second algorithms, respectively. Section 5 reports numerical results. Section 6 finally summarizes the result in this paper and highlights possible extension.

2. Preliminaries. For $M \in \mathbb{C}^{n\times n}$, let M_{ij}, $M_{:j}$ and $\lambda(M)$ be the (i,j) element, j-th column and spectrum of M, respectively, $|M| := (|M_{ij}|)$, $M^T := (M_{ji})$, $\|M\|_\infty := \max_i \sum_j |M_{ij}|$ and $\|M\|_{\max} := \max_{i,j} |M_{ij}|$. If M is nonsingular in particular, let $M^{-T} := (M^{-1})^T$. For $M, N \in \mathbb{R}^{m\times n}$, $M \le N$ means $M_{ij} \le N_{ij}$, $\forall i,j$. For $v \in \mathbb{C}^n$, v_i and $\mathrm{diag}(v)$ denote the i-th component of v and $n\times n$ diagonal matrix whose (i,i) element is v_i for $i = 1, \ldots, n$, respectively. For $v, w \in \mathbb{C}^n$ with $\|w\|_\infty < 1$ and $M, N \in \mathbb{C}^{n\times n}$ with $\|N\|_{\max} < 1$, define $\|v\|_w := \max_i(|v_i|/(1-|w_i|))$ and $\|M\|_N := \max_{i,j}(|M_{ij}|/(1-|N_{ij}|))$, respectively. Let `eps`, `realmin`, I_n and $\otimes$ be machine epsilon, the smallest positive normalized floating point number (especially `eps` $= 2^{-52}$ and `realmin` $= 2^{-1022}$ in IEEE 754 double precision), the $n\times n$ identity matrix and the Kronecker product (see e.g., [10]), respectively, and $e^{(n)} := [1, \ldots, 1]^T \in \mathbb{R}^n$. Let $\circ$ and ./ be the pointwise multiplication and division, respectively. For $C \in \mathbb{C}^{n\times n}$ and $R \in \mathbb{R}^{n\times n}$ with $R \ge 0$, $\langle C, R\rangle$ denotes the interval matrix whose midpoint and radius are C and R, respectively. For a Fréchet differentiable matrix function $F(X)$ where $X \in \mathbb{C}^{n\times n}$, denote the Fréchet derivative (see e.g., [7]) of F at X applied to the matrix H by $F'_X(H)$. The notations $\mathrm{fl}(\cdot)$ and $\mathrm{fl}_\triangle(\cdot)$ denote the results of floating point computations, where all operations insides the parentheses are executed by ordinary floating point arithmetic in rounding to nearest and towards $+\infty$ modes, respectively. The notations $\overline{\mathrm{fl}}(\cdot)$ and $\underline{\mathrm{fl}}(\cdot)$ denote rigorous upper and lower bounds for the insides of the parentheses obtained by rounding mode controlled floating point computations, respectively. Let $\mathbb{F}$ be the set of all floating point real numbers. For $M \in \mathbb{C}^{n\times n}$ and $v \in \mathbb{C}^{n^2}$, define

$$\mathrm{vec}(M) := \begin{bmatrix} M_{:1} \\ \vdots \\ M_{:n} \end{bmatrix} \quad \text{and} \quad \mathrm{mat}(v) := \begin{bmatrix} v_1 & v_{n+1} & \cdots & v_{n(n-1)+1} \\ \vdots & \vdots & \ddots & \vdots \\ v_n & v_{2n} & \cdots & v_{n^2} \end{bmatrix},$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

respectively. We then have $\mathrm{mat}(\mathrm{vec}(M)) = M$, $\mathrm{vec}(\mathrm{mat}(v)) = v$ and $\|M\|_N = \|\mathrm{vec}(M)\|_{\mathrm{vec}(N)}$ for $N \in \mathbb{C}^{n\times n}$ with $\|N\|_{\max} < 1$. If $v_i \neq 0$, $\forall i$, then

$$\mathrm{diag}(v)^{-1}\mathrm{vec}(M) = \mathrm{vec}(M./\mathrm{mat}(v)). \tag{2.6}$$

We cite Lemmas 2.1 to 2.3 and Corollary 2.5, which will be used hereafter.

LEMMA 2.1 (e.g., Golub and Van Loan [4]). *If $\|S\|_\infty < 1$ for $S \in \mathbb{C}^{n\times n}$, then $I_n - S$ is nonsingular.*

LEMMA 2.2 (e.g., Horn and Johnson [10]). *For any complex matrices K, L, M and N with compatible sizes, it holds that $(K \otimes L)(M \otimes N) = (KM \otimes LN)$ and $\mathrm{vec}(LMN) = (N^T \otimes L)\mathrm{vec}(M)$.*

LEMMA 2.3 (Minamihata [12]). *Let $S \in \mathbb{C}^{n\times n}$, $f \in \mathbb{C}^n$ and $s := |S|e^{(n)}$. If $\|s\|_\infty < 1$, then $I_n - S$ is nonsingular and $|(I_n - S)^{-1}||f| \leq |f| + \|f\|_s s$.*

REMARK 2.4. The proof of Lemma 2.3 can be found in [14, Section 2].

COROLLARY 2.5 (Miyajima [14]). *Let S and s be as in Lemma 2.3 and $F \in \mathbb{C}^{n\times n}$. Assume $\|s\|_\infty < 1$ and define $w := [\|F_{:1}\|_s, \ldots, \|F_{:n}\|_s]^T$. Then, $I_n - S$ is nonsingular and $|(I_n - S)^{-1}||F| \leq |F| + sw^T$.*

3. Verification algorithm when A is nonsingular. Let $Q(X)$ be as in (1.1), and $\tilde{X}$ be a numerically computed approximation to a solvent of $Q(X)$. Assume as the results of numerical generalized eigendecomposition, eigendecomposition and inversion, we have Λ_A,Λ_X,V_A,V_X,W_A,$W_X \in \mathbb{C}^{n\times n}$ with Λ_A and Λ_X being diagonal such that $(A\tilde{X} + B)V_A \approx AV_A\Lambda_A$, $\tilde{X}^T V_X \approx V_X\Lambda_X$, $W_A A V_A \approx I_n$ and $W_X V_X \approx I_n$. Define $S_A := I_n - W_A A V_A$, $S_X := I_n - W_X V_X$, $T_A := W_A(AV_A\Lambda_A - (A\tilde{X}+B)V_A)$ and $T_X := W_X(V_X\Lambda_X - \tilde{X}^T V_X)$. Provided that AV_A and V_X are not ill-conditioned, we can expect $S_A \approx 0$, $S_X \approx 0$, $T_A \approx 0$ and $T_X \approx 0$. If $\|S_A\|_\infty < 1$ and $\|S_X\|_\infty < 1$, then Lemma 2.1 gives $I_n - S_A$ and $I_n - S_X$ are nonsingular, respectively, which implies the nonsingularities of A, V_A, W_A, V_X and W_X. Then, define $U_A := (I_n - S_A)^{-1}T_A$, $U_X := (I_n - S_X)^{-1}T_X$, $Y := V_A^{-1}XV_X^{-T}$ and $\tilde{Y} := V_A^{-1}\tilde{X}V_X^{-T}$.

We first consider computing an interval matrix containing a solvent X_* of $Q(X)$. The verification of the uniqueness, dominance and minimality will be discussed later. We have $Q(X) = 0 \Leftrightarrow V_A^{-1}A^{-1}Q(X)V_X^{-T} = 0$. From $X = V_A Y V_X^T$, it holds that

$$V_A^{-1}A^{-1}Q(X)V_X^{-T} = YV_X^T V_A Y + V_A^{-1}A^{-1}BV_A Y + V_A^{-1}A^{-1}CV_X^{-T}.$$

Hence, (1.1) is equivalent to $R(Y) = 0$, where

$$R(Y) := YV_X^T V_A Y + V_A^{-1}A^{-1}BV_A Y + V_A^{-1}A^{-1}CV_X^{-T}.$$

Although $R(Y)$ seems to be more complicated than $Q(X)$, we can find its advantage when we consider its Fréchet derivative. In fact, we have

$$R(Y + H) = R(Y) + V_A^{-1}(V_A Y V_X^T + A^{-1}B)V_A H + HV_X^T V_A Y + HV_X^T V_A H, \tag{3.7}$$

which shows

$$R'_Y(H) = V_A^{-1}(V_A Y V_X^T + A^{-1}B)V_A H + HV_X^T V_A Y. \tag{3.8}$$

This and $\tilde{Y} = V_A^{-1}\tilde{X}V_X^{-T}$ give $R'_{\tilde{Y}}(H) = V_A^{-1}(\tilde{X} + A^{-1}B)V_A H + H(V_X^{-1}\tilde{X}^T V_X)^T$. From this and

$$V_A^{-1}(\tilde{X} + A^{-1}B)V_A = \Lambda_A - (\Lambda_A - V_A^{-1}(\tilde{X} + A^{-1}B)V_A) = \Lambda_A - V_A^{-1}A^{-1}W_A^{-1}T_A$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

$$= \Lambda_A - (W_A A V_A)^{-1} T_A = \Lambda_A - U_A, \tag{3.9}$$

$$V_X^{-1}\tilde{X}^T V_X = \Lambda_X - (\Lambda_X - V_X^{-1}\tilde{X}^T V_X) = \Lambda_X - V_X^{-1} W_X^{-1} T_X$$

$$= \Lambda_X - (W_X V_X)^{-1} T_X = \Lambda_X - U_X, \tag{3.10}$$

we obtain

$$R'_{\tilde{Y}}(H) = (\Lambda_A - U_A)H + H(\Lambda_X - U_X)^T,$$

whereas $Q'_{\tilde{X}}(H) = (A\tilde{X} + B)H + AH\tilde{X}$. We can expect that the coefficient matrices of $R'_{\tilde{Y}}(H)$ are not too far from diagonal, against that the coefficients $A\tilde{X}+B$, A and $\tilde{X}$ of $Q'_{\tilde{X}}(H)$ are dense in general. To exploit the special structure of the coefficient matrices of $R'_{\tilde{Y}}(H)$, we treat $R(Y) = 0$ instead of (1.1). Specifically, we compute an interval matrix $\boldsymbol{Y}$ containing $Y_* \in \mathbb{C}^{n\times n}$ such that $R(Y_*) = 0$, and enclose X_* by computing a superset of $\{V_A Y V_X^T : Y \in \boldsymbol{Y}\}$. Since $X_* = V_A Y_* V_X^T$, the superset contains X_*.

We now discuss the way for obtaining $\boldsymbol{Y}$. If $R'_{\tilde{Y}}(H)$ is invertible, we can define the Newton operator $N(Y) := Y - (R'_{\tilde{Y}})^{-1}(R(Y))$, and $N(Y) = Y$ is a fixed point equation for Y. For computing $\boldsymbol{Y}$, we thus verify the invertibility of $R'_{\tilde{Y}}(H)$ and inclusion $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\} \subseteq \langle \tilde{Y}, K\rangle$ for given $K \in \mathbb{R}^{n\times n}$ with $K > 0$. If these are true, Brouwer's fixed point theorem implies $Y_* \in \langle \tilde{Y}, K\rangle$, which gives $Y_* = N(Y_*) \in \{N(Y) : Y \in \langle \tilde{Y}, K\rangle\}$. Hence, an interval matrix including $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\}$ can be regarded as $\boldsymbol{Y}$.

We verify the invertibility of $R'_{\tilde{Y}}(H)$ by the following idea: From Lemma 2.2, $R'_{\tilde{Y}}(H)$ can be represented in terms of a matrix vector product as

$$\mathrm{vec}(R'_{\tilde{Y}}(H)) = \mathsf{P}\mathrm{vec}(H), \quad \mathsf{P} := I_n \otimes (\Lambda_A - U_A) + (\Lambda_X - U_X) \otimes I_n.$$

Therefore, if P is nonsingular, $R'_{\tilde{Y}}(H)$ is invertible. The nonsingularity of P can be verified with only $\mathcal{O}(n^3)$ operations by exploiting its special structure.

LEMMA 3.1. *Let $\nu, \mu \in \mathbb{C}^n$ and $\tilde{X}, V_A, V_X, W_A, W_X \in \mathbb{C}^{n\times n}$ be given, $\Lambda_A := \mathrm{diag}(\nu)$, $\Lambda_X := \mathrm{diag}(\mu)$, S_A, S_X, T_A and T_X be as the above, $s_A := |S_A|e^{(n)}$, $s_X := |S_X|e^{(n)}$, $t_A := |T_A|e^{(n)}$, $t_X := |T_X|e^{(n)}$ and $D := \nu e^{(n)^T} + e^{(n)}\mu^T$. Suppose $\|s_A\|_\infty < 1$, $\|s_X\|_\infty < 1$ and $|D| > 0$, and define $u_A := t_A + \|t_A\|_{s_A} s_A$, $u_X := t_X + \|t_X\|_{s_X} s_X$ and $E := (u_A e^{(n)^T} + e^{(n)} u_X^T)./|D|$. Then, A, V_A, W_A, V_X and W_X are nonsingular. If $\|E\|_{\max} < 1$, additionally, $R'_{\tilde{Y}}(H)$ is invertible for the above $R'_Y(H)$ and $\tilde{Y}$.*

Proof. Let P be as the above. We prove the nonsingularities of A, V_A, W_A, V_X, W_X and P. From $\|S_A\|_\infty = \|s_A\|_\infty < 1$, $\|S_X\|_\infty = \|s_X\|_\infty < 1$ and Lemma 2.1, $I_n - S_A$ and $I_n - S_X$ are nonsingular, which implies the nonsingularities of A, V_A, W_A, V_X and W_X. Let U_A and U_X be as the above, $\Delta := I_n \otimes \Lambda_A + \Lambda_X \otimes I_n$ and $\Omega := I_n \otimes U_A + U_X \otimes I_n$. Since Λ_A and Λ_X are diagonal, so is Δ. The elements $\Delta_{11}, \ldots, \Delta_{n^2n^2}$ can be written as $\nu_1 + \mu_1, \ldots, \nu_n + \mu_1, \ldots, \nu_1 + \mu_n, \ldots, \nu_n + \mu_n$, respectively. From this and $D_{ij} = \nu_i + \mu_j$, $i, j = 1, \ldots, n$, we have $\mathrm{mat}([\Delta_{11}, \ldots, \Delta_{n^2n^2}]^T) = D$. This and $|D| > 0$ give $\Delta_{kk} \neq 0$, $\forall k$. Thus, Δ is nonsingular, which shows

$$\mathsf{P} = \Delta - \Omega = \Delta(I_{n^2} - \Delta^{-1}\Omega). \tag{3.11}$$

Therefore, if $\|\Delta^{-1}\Omega\|_\infty < 1$, Lemma 2.1 yields the nonsingularity of P. We hence prove $\|\Delta^{-1}\Omega\|_\infty < 1$. From $\|s_A\|_\infty < 1$, $\|s_X\|_\infty < 1$ and Lemma 2.3, we have $|U_A|e^{(n)} \leq |(I_n - S_A)^{-1}|t_A \leq u_A$ and $|U_X|e^{(n)} \leq u_X$. It holds from these inequalities, $|\Delta^{-1}\Omega| = |\Delta^{-1}||\Omega|$, $|\Lambda_A|e^{(n)} = |\nu|$, $|\Lambda_X|e^{(n)} = |\mu|$, $\mathrm{mat}([\Delta_{11}, \ldots, \Delta_{n^2n^2}]^T) = D$, (2.6) and Lemma 2.2 that

$$|\Delta^{-1}\Omega|e^{(n^2)} = |\Delta^{-1}||\Omega|\mathrm{vec}(e^{(n)}e^{(n)^T}) \leq |\Delta^{-1}|(I_n \otimes |U_A| + |U_X| \otimes I_n)\mathrm{vec}(e^{(n)}e^{(n)^T})$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

$$\begin{aligned} &= |\Delta^{-1}|\mathrm{vec}(|U_A|e^{(n)}e^{(n)^T} + e^{(n)}(|U_X|e^{(n)})^T) \\ &\le |\Delta^{-1}|\mathrm{vec}(u_A e^{(n)^T} + e^{(n)}u_X^T) = \mathrm{vec}(E), \end{aligned} \tag{3.12}$$

which gives $\|\Delta^{-1}\Omega\|_\infty = \||\Delta^{-1}\Omega|e^{(n^2)}\|_\infty \le \|\mathrm{vec}(E)\|_\infty = \|E\|_{\max}$. This and $\|E\|_{\max} < 1$ show $\|\Delta^{-1}\Omega\|_\infty < 1$. ∎

We verify the inclusion $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\} \subseteq \langle \tilde{Y}, K\rangle$ by computing a superset of $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\}$. The superset can be computed by the following idea: The equality $N(Y) = Y - (R'_{\tilde{Y}})^{-1}(R(Y))$ is equivalent to $R'_{\tilde{Y}}(N(Y)) = R'_{\tilde{Y}}(Y) - R(Y)$. Therefore, $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\}$ is the set of all solutions to the parameterized Sylvester equation

$$(\Lambda_A - U_A)N_Y + N_Y(\Lambda_X - U_X)^T = R'_{\tilde{Y}}(Y) - R(Y), \tag{3.13}$$

where $N_Y \in \mathbb{C}^{n\times n}$ is unknown and $Y \in \langle \tilde{Y}, K\rangle$ is the parameter, which can be represented as $\mathsf{P}\mathrm{vec}(N_Y) = \mathrm{vec}(R'_{\tilde{Y}}(Y) - R(Y))$. Hence, the superset can be obtained by enclosing the solution set. The solution set can be enclosed with only $\mathcal{O}(n^3)$ operations by exploiting (3.11) and (3.12).

Lemma 3.2. *Under the conditions in Lemma 3.1, let $K \in \mathbb{R}^{n\times n}$ with $K > 0$ be given, $Q(X)$ be as in (1.1), $\tilde{X}$, V_A, V_X, W_A, W_X, s_A, s_X, D and E be as in Lemma 3.1, $\tilde{Y}$ and $N(Y)$ be as the above, $v_A := e^{(n)}./(e^{(n)} - s_A)$, $v_X := e^{(n)}./(e^{(n)} - s_X)$, $J \ge (I_n + s_A v_A^T)|W_A Q(\tilde{X})W_X^T|(I_n + v_X s_X^T)$, $L := (J + K|V_X^T V_A|K)./|D|$ and $M := L + \|L\|_E E$. Then, $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\} \subseteq \langle \tilde{Y}, M\rangle$.*

Proof. Let S_A, S_X, $R(Y)$, $R'_Y(H)$, P and N_Y be as the above, and Δ and Ω be as in the proof of Lemma 3.1. We prove $\langle \tilde{Y}, M\rangle$ includes the solution set of (3.13), i.e., $|\tilde{Y} - N_Y| \le M$ holds for any $Y \in \langle \tilde{Y}, K\rangle$. From Lemma 3.1 or its proof, $I_n - S_A$, $I_n - S_X$, A, V_A, W_A, V_X, W_X, Δ, $I_{n^2} - \Delta^{-1}\Omega$ and P are nonsingular. Any $Y \in \langle \tilde{Y}, K\rangle$ can be represented as $Y = \tilde{Y} + Y_P$, where $Y_P \in \mathbb{C}^{n\times n}$ satisfies $|Y_P| \le K$. This, $\mathrm{vec}(R'_{\tilde{Y}}(H)) = \mathsf{P}\mathrm{vec}(H)$, $\mathsf{P}\mathrm{vec}(N_Y) = \mathrm{vec}(R'_{\tilde{Y}}(Y) - R(Y))$ and (3.11) yield

$$\begin{aligned} \mathrm{vec}(\tilde{Y} - N_Y) &= \mathrm{vec}(\tilde{Y}) - \mathrm{vec}(N_Y) = \mathrm{vec}(\tilde{Y}) - \mathsf{P}^{-1}\mathrm{vec}(R'_{\tilde{Y}}(Y) - R(Y)) \\ &= \mathsf{P}^{-1}(\mathsf{P}\mathrm{vec}(\tilde{Y}) - \mathrm{vec}(R'_{\tilde{Y}}(\tilde{Y} + Y_P) - R(Y))) \\ &= (\Delta(I_{n^2} - \Delta^{-1}\Omega))^{-1}\mathrm{vec}(R'_{\tilde{Y}}(\tilde{Y}) - R'_{\tilde{Y}}(\tilde{Y}) - R'_{\tilde{Y}}(R_P) + R(Y)) \\ &= (I_{n^2} - \Delta^{-1}\Omega)^{-1}\Delta^{-1}\mathrm{vec}(R(\tilde{Y} + Y_P) - R'_{\tilde{Y}}(Y_P)). \end{aligned} \tag{3.14}$$

From (3.7) and (3.8) applied to $Y := \tilde{Y}$ and $H := Y_P$, $R(\tilde{Y} + Y_P) = R(\tilde{Y}) + R'_{\tilde{Y}}(Y_P) + Y_P V_X^T V_A Y_P$ holds. This, $\mathrm{mat}([\Delta_{11}, \ldots, \Delta_{n^2n^2}]^T) = D$, (2.6) and (3.14) give

$$\mathrm{vec}(\tilde{Y} - N_Y) = (I_{n^2} - \Delta^{-1}\Omega)^{-1}\mathrm{vec}((R(\tilde{Y}) + Y_P V_X^T V_A Y_P)./D). \tag{3.15}$$

It follows from $\tilde{Y} = V_A^{-1}\tilde{X}V_X^{-T}$ that

$$\begin{aligned} R(\tilde{Y}) &= V_A^{-1}\tilde{X}^2 V_X^{-T} + V_A^{-1}A^{-1}B\tilde{X}V_X^{-T} + V_A^{-1}A^{-1}CV_X^{-T} \\ &= V_A^{-1}A^{-1}Q(\tilde{X})V_X^{-T} = (W_A A V_A)^{-1}W_A Q(\tilde{X})W_X^T((W_X V_X)^{-1})^T \\ &= (I_n - S_A)^{-1}W_A Q(\tilde{X})W_X^T((I_n - S_X)^{-1})^T. \end{aligned} \tag{3.16}$$

From $[\|(I_n)_{:1}\|_{s_A}, \ldots, \|(I_n)_{:n}\|_{s_A}]^T = v_A$, $[\|(I_n)_{:1}\|_{s_X}, \ldots, \|(I_n)_{:n}\|_{s_X}]^T = v_X$, $\|s_A\|_\infty < 1$, $\|s_X\|_\infty < 1$ and Corollary 2.5, we have $|(I_n - S_A)^{-1}| = |(I_n - S_A)^{-1}|I_n \le I_n + s_A v_A^T$ and $|(I_n - S_X)^{-1}| \le I_n + s_X v_X^T$. These inequalities and (3.16) show

$$|R(\tilde{Y})| \le |(I_n - S_A)^{-1}||W_A Q(\tilde{X})W_X^T||(I_n - S_X)^{-1}|^T$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

$$\leq (I_n + s_A v_A^T)|W_A Q(\tilde{X}) W_X^T|(I_n + s_X v_X^T)^T \leq J. \tag{3.17}$$

From this, $|D| > 0$, $|Y_P| \leq K$, $\|E\|_{\max} < 1$, (3.12), (3.15) and Lemma 2.3, we obtain

$$\begin{aligned} \mathrm{vec}(|\tilde{Y} - N_Y|) &\leq |(I_{n^2} - \Delta^{-1}\Omega)^{-1}|\mathrm{vec}((|R(\tilde{Y})| + |Y_P||V_X^T V_A||Y_P|)./|D|) \\ &\leq |(I_{n^2} - \Delta^{-1}\Omega)^{-1}|\mathrm{vec}(L) \leq \mathrm{vec}(L) + \|\mathrm{vec}(L)\|_{|\Delta^{-1}\Omega|e^{(n^2)}}|\Delta^{-1}\Omega|e^{(n^2)} \\ &\leq \mathrm{vec}(L) + \|\mathrm{vec}(L)\|_{\mathrm{vec}(E)}\mathrm{vec}(E) = \mathrm{vec}(M), \end{aligned} \tag{3.18}$$

which proves $|\tilde{Y} - N_Y| \leq M$ for any Y_P, i.e., for any Y. ∎

Lemmas 3.1 and 3.2 yield a theory for computing an interval matrix containing a solvent X_* of $Q(X)$.

THEOREM 3.3. *Let $\tilde{X}$, V_A and V_X be as in Lemma 3.1, K and M be as in Lemma 3.2, and $M_S, G \in \mathbb{R}^{n\times n}$ be given. With all the assumptions in Lemma 3.1, suppose $M \leq M_S \leq K$ and $G \geq |V_A|M_S|V_X|^T$. Then, $\langle \tilde{X}, G\rangle$ contains the solvent X_*.*

REMARK 3.4. In practical executions of the algorithm, $G = \overline{\mathrm{fl}}(|V_A|M_S|V_X|^T)$.

Proof of Theorem 3.3. Let $\tilde{Y}$, $R(Y)$, $\boldsymbol{Y}$ and $N(Y)$ be as the above. From Lemmas 3.1 and 3.2, and $M \leq M_S \leq K$, we have $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\} \subseteq \langle \tilde{Y}, M\rangle \subseteq \langle \tilde{Y}, M_S\rangle \subseteq \langle \tilde{Y}, K\rangle$. Hence, the Brouwer theorem implies $\langle \tilde{Y}, K\rangle$ contains a solution Y_* to $N(Y) = Y$, i.e., $R(Y) = 0$, which shows $Y_* = N(Y_*) \in \{N(Y) : Y \in \langle \tilde{Y}, K\rangle\} \subseteq \langle \tilde{Y}, M_S\rangle$. We can thus put $\boldsymbol{Y} = \langle \tilde{Y}, M_S\rangle$. Thus, $\tilde{X} = V_A \tilde{Y} V_X^T$, $G \geq |V_A|M_S|V_X|^T$ and a center-radius interval arithmetic evaluation (e.g., [1]) yield

$$X_* = V_A Y_* V_X^T \in \{V_A Y V_X^T : Y \in \boldsymbol{Y}\} \subseteq \langle V_A \tilde{Y} V_X^T, |V_A|M_S|V_X|^T\rangle \subseteq \langle \tilde{X}, G\rangle. \quad ∎$$

In the practical executions, we need to determine K and M_S such that $K > 0$ and $M \leq M_S \leq K$. These matrices can be determined by:

LEMMA 3.5. *Under the conditions of Lemma 3.1, let V_A, V_X, D and E be as in Lemma 3.1, K, J and M be as in Lemma 3.2, M_S be as in Theorem 3.3, $M_0 \in \mathbb{F}^{n\times n}$ and $\sigma, \eta \in \mathbb{F}$ be given, and $L_0 := J./|D|$. If $J > 0$, $\mathrm{fl}_\triangle(a \bullet b) = (1+\delta)(a \bullet b)$ for $a, b \in \mathbb{F}$, where $\bullet \in \{+, *\}$ and $|\delta| \leq \mathtt{eps}$, $M_0 \geq L_0 + \|L_0\|_E E$, $\|(M_0|V_X^T V_A|M_0)./J\|_{\max} \leq \sigma \leq 1/(4(1+\mathtt{eps})^6)$,*

$$\frac{2(1+\mathtt{eps})^2}{1+\sqrt{1-4\sigma(1+\mathtt{eps})^6}} \leq \eta \leq \frac{1+\sqrt{1-4\sigma(1+\mathtt{eps})^6}}{2\sigma(1+\mathtt{eps})^4}, \tag{3.19}$$

$K = \eta M_0$ and $M_S = \mathrm{fl}_\triangle((1+\sigma\eta^2)M_0)$, then $K > 0$ and $M \leq M_S \leq K$.

Proof. We first prove $M \leq M_S$. Let L be as in Lemma 3.2. From $J > 0$, $K = \eta M_0$ and $\sigma \geq \|(M_0|V_X^T V_A|M_0)./J\|_{\max}$, we have

$$\begin{aligned} J + K|V_X^T V_A|K &= J + \eta^2 M_0|V_X^T V_A|M_0 = J + \eta^2(M_0|V_X^T V_A|M_0)./J \circ J \\ &\leq J + \eta^2\|(M_0|V_X^T V_A|M_0)./J\|_{\max} J \leq (1+\sigma\eta^2)J, \end{aligned}$$

so $L \leq (1+\sigma\eta^2)L_0$. Thus, $M_0 \geq L_0 + \|L_0\|_E E$ and $M_S = \mathrm{fl}_\triangle((1+\sigma\eta^2)M_0)$ give

$$M \leq (1+\sigma\eta^2)(L_0 + \|L_0\|_E E) \leq (1+\sigma\eta^2)M_0 \leq \mathrm{fl}_\triangle((1+\sigma\eta^2)M_0) = M_S.$$

We next prove $K > 0$ and $M_S \leq K$. The assumption regarding to $\mathrm{fl}_\triangle(\cdot)$ shows

$$M_S = (1+\delta_4)(1+\delta_3)(1+\sigma\eta^2(1+\delta_1)(1+\delta_2))M_0, \text{ where } |\delta_i| \leq \mathtt{eps},\ i = 1, \ldots, 4,$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

which yields $M_S \leq (1+\mathtt{eps})^2(1+\sigma\eta^2(1+\mathtt{eps})^2)M_0$. From $\sigma \leq 1/(4(1+\mathtt{eps})^6)$ and (3.19), we have $(1+\mathtt{eps})^2(1+\sigma\eta^2(1+\mathtt{eps})^2) \leq \eta$. The inequality $J > 0$ implies $M_0 > 0$. These discussions, $\eta > 1$ and $K = \eta M_0$ give $K > 0$ and $M_S \leq (1+\mathtt{eps})^2(1+\sigma\eta^2(1+\mathtt{eps})^2)M_0 \leq \eta M_0 = K$. ∎

Remark 3.6. In the proposed algorithm, K and M_S are "determined" based on Lemma 3.5, but M and K are not "numerically computed", and only M_S is computed. Note that $K > 0$ and $M \leq M_S \leq K$ are still valid even in this case, and computing M_S is sufficient for enclosing the solvent based on Theorem 3.3. The algorithm in Section 4 is designed similarly.

The uniqueness of the contained solvent can be verified by the following idea: Let $\langle \tilde{X}, G\rangle$ contain the solvent, and X_* and X_{**}, and X_1 and X_2 be the solvents and arbitrarily matrices contained in $\langle \tilde{X}, G\rangle$, respectively. We prove the invertibility of $Q'_{\tilde{X}}(H)$ and use $\mathcal{N}(X) := X - (Q'_{\tilde{X}})^{-1}(Q(X))$. Observe $\mathcal{N}(X_*) = X_*$ and $\mathcal{N}(X_{**}) = X_{**}$. We derive a function $\mathsf{S}(X_1, X_2)$ satisfying $\mathrm{vec}(\mathcal{N}(X_1) - \mathcal{N}(X_2)) = \mathsf{S}(X_1, X_2)\mathrm{vec}(X_1 - X_2)$, and prove $\|\mathsf{S}(X_1, X_2)\|_\infty < 1$, $\forall X_1, X_2 \in \langle \tilde{X}, G\rangle$, which gives $\|\mathsf{S}(X_*, X_{**})\|_\infty < 1$. The uniqueness can be shown from $\|\mathsf{S}(X_*, X_{**})\|_\infty < 1$, since

$$\begin{aligned}\|\mathrm{vec}(X_* - X_{**})\|_\infty &= \|\mathrm{vec}(\mathcal{N}(X_*) - \mathcal{N}(X_{**}))\|_\infty = \|\mathsf{S}(X_*, X_{**})\mathrm{vec}(X_* - X_{**})\|_\infty \\ &\leq \|\mathsf{S}(X_*, X_{**})\|_\infty \|\mathrm{vec}(X_* - X_{**})\|_\infty,\end{aligned}$$

so that $(1 - \|\mathsf{S}(X_*, X_{**})\|_\infty)\|\mathrm{vec}(X_* - X_{**})\|_\infty \leq 0$, which implies $X_* = X_{**}$. We establish Theorem 3.7 for verifying the uniqueness based on this idea.

Theorem 3.7. *Under the conditions in Lemma 3.1, let $\tilde{X}$, V_A, V_X, W_A, W_X, s_A, s_X, D and E be as in Lemma 3.1, and $G \in \mathbb{R}^{n\times n}$ with $G \geq 0$ be given. Define*

$$w^{(1)} := |W_A A|Ge^{(n)} + \||W_A A|Ge^{(n)}\|_{s_A} s_A, \quad w^{(2)} := |W_X|e^{(n)} + \||W_X|e^{(n)}\|_{s_X} s_X,$$

$$w^{(3)} := |W_A A|e^{(n)} + \||W_A A|e^{(n)}\|_{s_A} s_A, \quad w^{(4)} := |W_X|G^T e^{(n)} + \||W_X|G^T e^{(n)}\|_{s_X} s_X,$$

$$F := (w^{(1)}{w^{(2)}}^T + w^{(3)}{w^{(4)}}^T)./|D|, \quad \textit{and} \quad Z := |V_A|(F + \|F\|_E E)|V_X|^T.$$

If $\langle \tilde{X}, G\rangle$ contains the solvent and $\|Z\|_{\max} < 1$, then the contained solvent is unique.

Proof. Let Λ_A and Λ_X be as in Lemma 3.1, Δ and Ω be as in the proof of Lemma 3.1, and U_A, U_X, P, X_1, X_2, $\mathcal{N}(X)$ and $\mathsf{S}(X_1, X_2)$ be as the above. We prove the invertibility of $Q'_{\tilde{X}}(H)$, derive $\mathsf{S}(X_1, X_2)$, and show $\|\mathsf{S}(X_1, X_2)\|_\infty < 1$, $\forall X_1, X_2 \in \langle \tilde{X}, G\rangle$. We have $Q'_{\tilde{X}}(H) = (A\tilde{X} + B)H + AH\tilde{X}$, which can be written as

$$\mathrm{vec}(Q'_{\tilde{X}}(H)) = \mathsf{Q}\mathrm{vec}(H), \quad \mathsf{Q} := I_n \otimes (A\tilde{X} + B) + \tilde{X}^T \otimes A.$$

From Lemma 3.1 or its proof, $I_n - S_A$, $I_n - S_X$, A, V_A, V_X, Δ, $I_{n^2} - \Delta^{-1}\Omega$ and P are nonsingular. The equalities (3.9) and (3.10), and Lemma 2.2 yield

$$\begin{aligned}\mathsf{Q} &= (V_X \otimes AV_A)(I_n \otimes V_A^{-1}(\tilde{X} + A^{-1}B)V_A + V_X^{-1}\tilde{X}^T V_X \otimes I_n)(V_X^{-1} \otimes V_A^{-1}) \\ &= (V_X \otimes AV_A)(I_n \otimes (\Lambda_A - U_A) + (\Lambda_X - U_X) \otimes I_n)(V_X^{-1} \otimes V_A^{-1}) \\ &= (V_X \otimes AV_A)\mathsf{P}(V_X^{-1} \otimes V_A^{-1}). \qquad (3.20)\end{aligned}$$

From this and the nonsingularity of P, Q is nonsingular, so that $Q'_{\tilde{X}}(H)$ is invertible.

The equality $\mathcal{N}(X) = X - (Q'_{\tilde{X}})^{-1}(Q(X))$ gives $Q'_{\tilde{X}}(\mathcal{N}(X)) = Q'_{\tilde{X}}(X) - Q(X)$. From this, the equality $\mathrm{vec}(Q'_{\tilde{X}}(\mathcal{N}(X))) = \mathsf{Q}\mathrm{vec}(\mathcal{N}(X))$ and the nonsingularity of Q, we obtain

$$\mathrm{vec}(\mathcal{N}(X)) = \mathsf{Q}^{-1}\mathrm{vec}(Q'_{\tilde{X}}(X) - Q(X)). \qquad (3.21)$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

As mentioned in [6, Proof of Theorem 3.1], it follows that

$$Q(X_1) - Q(X_2) = \left(\frac{1}{2}A(X_1 + X_2) + B\right)(X_1 - X_2) + \frac{1}{2}A(X_1 - X_2)(X_1 + X_2),$$

which gives

$$\mathrm{vec}(Q(X_1) - Q(X_2)) = \left(I_n \otimes \left(\frac{1}{2}A(X_1 + X_2) + B\right) + \frac{1}{2}(X_1 + X_2)^T \otimes A\right)\mathrm{vec}(X_1 - X_2). \tag{3.22}$$

From (3.21), (3.22), $\mathsf{Q} = I_n \otimes (A\tilde{X} + B) + \tilde{X}^T \otimes A$ and $\mathrm{vec}(Q'_{\tilde{X}}(X_1) - Q'_{\tilde{X}}(X_2)) = \mathsf{Q}\mathrm{vec}(X_1 - X_2)$, it holds that

$$\begin{aligned}
&\mathrm{vec}(\mathcal{N}(X_1) - \mathcal{N}(X_2)) \\
&= \mathsf{Q}^{-1}\mathrm{vec}(Q'_{\tilde{X}}(X_1) - Q'_{\tilde{X}}(X_2) - (Q(X_1) - Q(X_2))) \\
&= \mathsf{Q}^{-1}\left(\mathsf{Q} - \left(I_n \otimes \left(\frac{1}{2}A(X_1 + X_2) + B\right) + \frac{1}{2}(X_1 + X_2)^T \otimes A\right)\right)\mathrm{vec}(X_1 - X_2) \\
&= \mathsf{Q}^{-1}\left(I_n \otimes A\left(\tilde{X} - \frac{1}{2}(X_1 + X_2)\right) + \left(\tilde{X} - \frac{1}{2}(X_1 + X_2)\right)^T \otimes A\right)\mathrm{vec}(X_1 - X_2).
\end{aligned}$$

Thus, $\mathsf{S}(X_1, X_2)$ is derived such that

$$\mathsf{S}(X_1, X_2) = \mathsf{Q}^{-1}\left(I_n \otimes A\left(\tilde{X} - \frac{1}{2}(X_1 + X_2)\right) + \left(\tilde{X} - \frac{1}{2}(X_1 + X_2)\right)^T \otimes A\right).$$

Since $X_1, X_2 \in \langle \tilde{X}, G\rangle$, X_1 and X_2 can be represented as $X_1 = \tilde{X} + \Gamma_1$ and $X_2 = \tilde{X} + \Gamma_2$, respectively, where $\Gamma_1, \Gamma_2 \in \mathbb{C}^{n\times n}$ satisfy $|\Gamma_1| \le G$ and $|\Gamma_2| \le G$. This representation, (3.20) and Lemma 2.2 yield

$$\begin{aligned}
\mathsf{S}(X_1, X_2) &= -(V_X \otimes V_A)\mathsf{P}^{-1}(V_X^{-1} \otimes (AV_A)^{-1})\left(I_n \otimes \frac{1}{2}A(\Gamma_1 + \Gamma_2)\frac{1}{2}(\Gamma_1 + \Gamma_2)^T \otimes A\right) \\
&= -(V_X \otimes V_A)\mathsf{P}^{-1}\left(V_X^{-1} \otimes \frac{1}{2}(AV_A)^{-1}A(\Gamma_1 + \Gamma_2) + \frac{1}{2}V_X^{-1}(\Gamma_1 + \Gamma_2)^T \otimes (AV_A)^{-1}A\right), \qquad (3.23)
\end{aligned}$$

so that

$$\begin{aligned}
|\mathsf{S}(X_1, X_2)| &\le (|V_X| \otimes |V_A|)|\mathsf{P}^{-1}|\left(|V_X^{-1}| \otimes \frac{1}{2}|(AV_A)^{-1}A|(|\Gamma_1| + |\Gamma_2|)\right. \\
&\qquad \left. + \frac{1}{2}|V_X^{-1}|(|\Gamma_1| + |\Gamma_2|)^T \otimes |(AV_A)^{-1}A|\right) \\
&\le (|V_X| \otimes |V_A|)|\mathsf{P}^{-1}|(|V_X^{-1}| \otimes |(AV_A)^{-1}A|G + |V_X^{-1}|G^T \otimes |(AV_A)^{-1}A|). \qquad (3.24)
\end{aligned}$$

From $\|s_A\|_\infty < 1$, $\|s_X\|_\infty < 1$ and Lemma 2.3, we have

$$|(AV_A)^{-1}A|Ge^{(n)} = |(I_n - S_A)^{-1}W_AA|Ge^{(n)} \le |(I_n - S_A)^{-1}||W_AA|Ge^{(n)} \le w^{(1)},$$

$$|V_X^{-1}|e^{(n)} = |(I_n - S_X)^{-1}W_X|e^{(n)} \le |(I_n - S_X)^{-1}||W_X|e^{(n)} \le w^{(2)},$$

$$|(AV_A)^{-1}A|e^{(n)} = |(I_n - S_A)^{-1}W_AA|e^{(n)} \le |(I_n - S_A)^{-1}||W_AA|e^{(n)} \le w^{(3)},$$

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

$$|V_X^{-1}|G^T e^{(n)} = |(I_n - S_X)^{-1} W_X| G^T e^{(n)} \le |(I_n - S_X)^{-1}||W_X| G^T e^{(n)} \le w^{(4)}.$$

These inequalities, (3.11), (3.12), (3.24), $\|E\|_{\max} < 1$ and Lemma 2.3 show

$$\begin{aligned}
|\mathsf{S}(X_1, X_2)| e^{(n^2)} &\le (|V_X| \otimes |V_A|)|(I_n - \Delta^{-1}\Omega)^{-1}\Delta^{-1}|(|V_X^{-1}| \otimes |(AV_A)^{-1}A|G \\
&\quad + |V_X^{-1}|G^T \otimes |(AV_A)^{-1}A|)\mathrm{vec}(e^{(n)} e^{(n)^T}) \\
&= (|V_X| \otimes |V_A|)|(I_n - \Delta^{-1}\Omega)^{-1}||\Delta^{-1}|\mathrm{vec}(|(AV_A)^{-1}A|Ge^{(n)}(|V_X^{-1}|e^{(n)})^T \\
&\quad + |(AV_A)^{-1}A|e^{(n)}(|V_X^{-1}|G^T e^{(n)})^T) \\
&\le (|V_X| \otimes |V_A|)|(I_n - \Delta^{-1}\Omega)^{-1}||\Delta^{-1}|\mathrm{vec}(w^{(1)} w^{(2)^T} + w^{(3)} w^{(4)^T}) \\
&= (|V_X| \otimes |V_A|)|(I_n - \Delta^{-1}\Omega)^{-1}|\mathrm{vec}(F) \\
&\le (|V_X| \otimes |V_A|)(\mathrm{vec}(F) + \|\mathrm{vec}(F)\|_{|\Delta^{-1}\Omega| e^{(n^2)}} |\Delta^{-1}\Omega| e^{(n^2)}) \\
&\le (|V_X| \otimes |V_A|)(\mathrm{vec}(F) + \|\mathrm{vec}(F)\|_{\mathrm{vec}(E)} \mathrm{vec}(E)) \\
&= (|V_X| \otimes |V_A|)\mathrm{vec}(F + \|F\|_E E) = \mathrm{vec}(Z), \qquad (3.25)
\end{aligned}$$

which gives $\|\mathsf{S}(X_1, X_2)\|_\infty = \||\mathsf{S}(X_1, X_2)| e^{(n^2)}\|_\infty \le \|\mathrm{vec}(Z)\|_\infty = \|Z\|_{\max}$. This and $\|Z\|_{\max} < 1$ show $\|\mathsf{S}(X_1, X_2)\|_\infty < 1$, $\forall X_1, X_2 \in \langle \tilde{X}, G \rangle$. ∎

We finally discuss verifying the dominance of the contained solvent. The verification of the minimality can be achieved completely analogously. Let $\lambda_1, \dots, \lambda_{2n}$ be as in (1.5), $\langle \tilde{X}, G \rangle$ contain the solvent X_*, $\{\nu_1^*, \dots, \nu_n^*\} := \lambda(-(X_* + A^{-1}B))$ and $\{\mu_1^*, \dots, \mu_n^*\} := \lambda(X_*)$. From (1.3), $\{\lambda_1, \dots, \lambda_{2n}\} = \{\nu_1^*, \dots, \nu_n^*\} \cup \{\mu_1^*, \dots, \mu_n^*\}$ holds. Therefore, if $\min_i |\mu_i^*| > \max_i |\nu_i^*|$, then $\lambda(X_*) = \{\lambda_1, \dots, \lambda_n\}$ and $|\lambda_n| > |\lambda_{n+1}|$, i.e., X_* is the dominant solvent. We hence check $\min_i |\mu_i^*| > \max_i |\nu_i^*|$, which is possible with only $\mathcal{O}(n^2)$ operations by reusing previously obtained matrices.

THEOREM 3.8. *Under the conditions in Lemma 3.1, let ν, μ, $\tilde{X}$, V_A, V_X, W_A, W_X, s_A, s_X, u_A and u_X be as in Lemma 3.1, and G be as in Theorem 3.7. Define $r_X := u_X + |W_X| G^T |V_X| e^{(n)} + \||W_X| G^T |V_X| e^{(n)}\|_{s_X} s_X$ and $r_A := u_A + |W_A| G |V_A| e^{(n)} + \||W_A| G |V_A| e^{(n)}\|_{s_A} s_A$. If $\langle \tilde{X}, G \rangle$ contains the solvent X_* and $\min_i(|\mu_i| - (r_X)_i) > \max_i(|\nu_i| + (r_A)_i)$, then X_* is the dominant solvent.*

Proof. Let $\lambda_1, \dots, \lambda_{2n}$ be as in (1.5), Λ_A and Λ_X be as in Lemma 3.1, and S_X, U_A, U_X, $\{\nu_1^*, \dots, \nu_n^*\}$ and $\{\mu_1^*, \dots, \mu_n^*\}$ be as the above. We prove $\min_i |\mu_i^*| > \max_i |\nu_i^*|$ by estimating lower and upper bounds for $\min_i |\mu_i^*|$ and $\max_i |\nu_i^*|$, respectively. Lemma 3.1 or its proof show $I_n - S_X$, A, V_A and V_X are nonsingular, and $|U_X| e^{(n)} \le u_X$. Since $X_* \in \langle \tilde{X}, G \rangle$, X_* can be written as $X_* = \tilde{X} + \Gamma$, where $\Gamma \in \mathbb{C}^{n \times n}$ satisfies $|\Gamma| \le G$.

We first show $\min_i |\mu_i^*| \ge \min_i(|\mu_i| - (r_X)_i)$. Since $\lambda(V_X^{-1} X_*^T V_X) = \{\mu_1^*, \dots, \mu_n^*\}$, we consider $\lambda(V_X^{-1} X_*^T V_X)$ instead. From $X_* = \tilde{X} + \Gamma$ and (3.10), we have

$$V_X^{-1} X_*^T V_X = V_X^{-1} (\tilde{X} + \Gamma)^T V_X = \Lambda_X + \Gamma_X,$$

where $\Gamma_X := V_X^{-1} \Gamma^T V_X - U_X$. This and the Gershgorin circle theorem give $\lambda(V_X^{-1} X_*^T V_X) \subseteq \bigcup_{i=1}^n \langle \mu_i + (\Gamma_X)_{ii}, (|\Gamma_X| e^{(n)})_i - |(\Gamma_X)_{ii}| \rangle$, whose superset is $\bigcup_{i=1}^n \langle \mu_i, (|\Gamma_X| e^{(n)})_i \rangle$. From $|U_X| e^{(n)} \le u_X$, $|\Gamma| \le G$, $\|s_X\|_\infty < 1$ and Lemma 2.3, we have

$$\begin{aligned}
|\Gamma_X| e^{(n)} &\le |(I_n - S_X)^{-1} W_X| |\Gamma|^T |V_X| e^{(n)} + |U_X| e^{(n)} \\
&\le |(I_n - S_X)^{-1}| |W_X| G^T |V_X| e^{(n)} + u_X \le r_X,
\end{aligned}$$

so that $\bigcup_{i=1}^n \langle \mu_i, (r_X)_i \rangle$ also contains $\lambda(V_X^{-1} X_*^T V_X) = \{\mu_1^*, \dots, \mu_n^*\}$. Hence, $\min_i |\mu_i^*| \ge \min_i(|\mu_i| - (r_X)_i)$ follows.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

We next prove $\max_i |\nu_i^*| \leq \max_i(|\nu_i| + (r_A)_i)$. Since $\lambda(V_A^{-1}(X_* + A^{-1}B)V_A) = \{-\nu_1^*, \ldots, -\nu_n^*\}$, we consider $\lambda(V_A^{-1}(X_*+A^{-1}B)V_A)$ instead. Similarly to the previous paragraph, we have $V_A^{-1}(X_*+A^{-1}B)V_A = \Lambda_A + \Gamma_A$, where $\Gamma_A := V_A^{-1}\Gamma V_A - U_A$, and $|\Gamma_A|e^{(n)} \leq r_A$, so that $\bigcup_{i=1}^n \langle \nu_i, (r_A)_i\rangle$ contains $\lambda(V_A^{-1}(X_* + A^{-1}B)V_A) = \{-\nu_1^*, \ldots, -\nu_n^*\}$. Therefore, $\max_i |\nu_i^*| \leq \max_i(|\nu_i|+(r_A)_i)$ is true. The proved inequalities and $\min_i(|\mu_i| - (r_X)_i) > \max_i(|\nu_i| + (r_A)_i)$ give $\min_i |\mu_i^*| > \max_i |\nu_i^*|$. ∎

Corollary 3.9. *Under the conditions in Lemma 3.1, let ν, μ, $\tilde{X}$, G, X_*, r_A and r_X be defined similarly to Theorem 3.8. If $\langle \tilde{X}, G\rangle \ni X_*$ and $\max_i(|\mu_i|+(r_X)_i) < \min_i(|\nu_i| - (r_A)_i)$, then X_* is the minimal solvent.*

Proof. From the proof of Theorem 3.8, we have $\max_i |\mu_i^*| \leq \max_i(|\mu_i|+(r_X)_i)$ and $\min_i |\nu_i^*| \geq \min_i(|\nu_i|-(r_A)_i)$ for $\{\nu_1^*, \ldots, \nu_n^*\}$ and $\{\mu_1^*, \ldots, \mu_n^*\}$ defined above. These inequalities and $\max_i(|\mu_i|+(r_X)_i) < \min_i(|\nu_i|-(r_A)_i)$ prove $\max_i |\mu_i^*| < \min_i |\nu_i^*|$. ∎

Based on the established theories, we propose:

Algorithm 1. This algorithm computes $\tilde{X}$ and G such that $\langle \tilde{X}, G\rangle \ni X_*$. The nonsingularity of A, uniqueness and dominance (or minimality) are moreover proved if successful.

Step 1. Compute $\tilde{X}$ via a known algorithm. Calculate Λ_A and V_A by numerical generalized eigendecomposition $(A\tilde{X} + B)V_A \approx AV_A\Lambda_A$. Compute Λ_X and V_X by numerical eigendecomposition $\tilde{X}^T V_X \approx V_X\Lambda_X$. Calculate W_A and W_X such that $W_A = \mathrm{fl}((AV_A)^{-1})$ and $W_X = \mathrm{fl}(V_X^{-1})$, respectively.

Step 2. Compute $\overline{\mathrm{fl}}(s_A)$. If $\overline{\mathrm{fl}}(\|s_A\|_\infty) \geq 1$, terminate with failure.

Step 3. Compute $\overline{\mathrm{fl}}(s_X)$. If $\overline{\mathrm{fl}}(\|s_X\|_\infty) \geq 1$, terminate with failure.

Step 4. Compute $\underline{\mathrm{fl}}(|D|)$. If $\min_{i,j} \underline{\mathrm{fl}}(|D_{ij}|) = 0$, terminate with failure.

Step 5. Compute $\overline{\mathrm{fl}}(E)$. If $\overline{\mathrm{fl}}(\|E\|_{\max}) \geq 1$, terminate with failure.

Step 6. Compute J such that $J = \overline{\mathrm{fl}}((I_n + s_A v_A^T)|W_A Q(\tilde{X})W_X^T|(I_n + v_X s_X^T))$. If $J_{ij} < \sqrt{\texttt{realmin}}$ for $i, j \in \{1, \ldots, n\}$, update J such that $J_{ij} = \sqrt{\texttt{realmin}}$.

Step 7. Compute M_0 such that $M_0 = \overline{\mathrm{fl}}(L_0 + \|L_0\|_E E)$. If $(M_0)_{ij} < \sqrt{\texttt{realmin}}$, update M_0 such that $(M_0)_{ij} = \sqrt{\texttt{realmin}}$.

Step 8. Compute σ such that $\sigma = \overline{\mathrm{fl}}(\|(M_0|V_X^T V_A|M_0)./J\|_{\max})$. If $\sigma < \sqrt{\texttt{realmin}}$, update σ such that $\sigma = \sqrt{\texttt{realmin}}$. If $\overline{\mathrm{fl}}(\sigma(1 + \texttt{eps})^6) > 1/4$, which means $\sigma \leq 1/(4(1 + \texttt{eps})^6)$ cannot be verified, terminate with failure.

Step 9. Compute η such that $\eta = \overline{\mathrm{fl}}(2(1 + \texttt{eps})^2/(1 + \sqrt{1 - 4\sigma(1 + \texttt{eps})^6}))$. If $\eta > \underline{\mathrm{fl}}((1 + \sqrt{1 - 4\sigma(1 + \texttt{eps})^6})/(2\sigma(1 + \texttt{eps})^4))$, terminate with failure.

Step 10. Compute $M_S = \mathrm{fl}_\triangle((1 + \sigma\eta^2)M_0)$. Then, $Y_* \in \langle \tilde{Y}, M_S\rangle$ holds.

Step 11. Compute G such that $G = \overline{\mathrm{fl}}(|V_A|M_S|V_X|^T)$. Then, $X_* \in \langle \tilde{X}, G\rangle$ follows.

Step 12. Compute $\overline{\mathrm{fl}}(Z)$. If $\overline{\mathrm{fl}}(\|Z\|_{\max}) < 1$, then X_* is unique in $\langle \tilde{X}, G\rangle$.

Step 13. Compute $\underline{\mathrm{fl}}(\min_i(|\mu_i| - (r_X)_i))$ and $\overline{\mathrm{fl}}(\max_i(|\nu_i| + (r_A)_i))$. If $\underline{\mathrm{fl}}(\min_i(|\mu_i| - (r_X)_i)) > \overline{\mathrm{fl}}(\max_i(|\nu_i| + (r_A)_i))$, then X_* is the dominant solvent. Terminate.

Step 14. Compute $\overline{\mathrm{fl}}(\max_i(|\mu_i| + (r_X)_i))$ and $\underline{\mathrm{fl}}(\min_i(|\nu_i| - (r_A)_i))$. If $\overline{\mathrm{fl}}(\max_i(|\mu_i| + (r_X)_i)) < \underline{\mathrm{fl}}(\min_i(|\nu_i| - (r_A)_i))$, then X_* is the minimal solvent. Terminate.

Remark 3.10. The inequalities $\min_{i,j}(M_0)_{ij} \geq \sqrt{\texttt{realmin}}$, $\sigma \geq \sqrt{\texttt{realmin}}$ and $\eta > 1$ show underflow does not occur during the computation of M_S, so that $\mathrm{fl}_\triangle(\cdot)$ satisfies the condition in Lemma 3.5 during this computation.

The following proposition clarifies the complexity of Algorithm 1:

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

PROPOSITION 3.11. *Algorithm* 1 *has a cost of* $\mathcal{O}(n^3)$ *operations.*

Proof. The costs for the generalized eigendecomposition, eigendecomposition and inversion are $\mathcal{O}(n^3)$. All the other matrix-matrix operations (multiplications, additions or pointwise divisions) involve $n \times n$ matrices, so their cost is again $\mathcal{O}(n^3)$. The cost for all the remaining operations is negligible, which finishes the proof. ∎

4. Verification algorithm when $A\tilde{X} + B$ is nonsingular. When A is singular or nearly singular, the condition $\|s_A\|_\infty < 1$ in Lemma 3.1 does not follow, so that Algorithm 1 fails. Therefore, we need an alternative approach in this case. The algorithm proposed in this section is applicable even when A is singular, but requires the nonsingularity of $A\tilde{X} + B$, where $\tilde{X}$ is defined as in Section 3.

Let Λ_X, V_X, W_X, S_X and T_X be as in Section 3, $AV_A \approx (A\tilde{X} + B)V_A\Lambda_A$ and $W_A \approx ((A\tilde{X} + B)V_A)^{-1}$ be numerical generalized eigendecomposition and inversion, respectively. Define $S_A := I_n - W_A(A\tilde{X} + B)V_A$ and $T_A := W_A((A\tilde{X} + B)V_A\Lambda_A - AV_A)$. If $\|S_A\|_\infty < 1$ and $\|S_X\|_\infty < 1$, then $I_n - S_A$, $I_n - S_X$, $A\tilde{X} + B$, V_A, W_A, V_X and W_X are nonsingular. Then, define U_A, U_X, Y and $\tilde{Y}$ similarly to Section 3.

We first consider computing an interval matrix containing the solvent X_*, and then discuss the verification of the uniqueness. We have $Q(X) = 0 \Leftrightarrow V_A^{-1}(A\tilde{X} + B)^{-1}Q(X)V_X^{-T} = 0$. Therefore, (1.1) is equivalent to $R(Y) = 0$, where

$$R(Y) := V_A^{-1}(A\tilde{X} + B)^{-1}AV_AYV_X^TV_AY + V_A^{-1}(A\tilde{X} + B)^{-1}BV_AY + V_A^{-1}(A\tilde{X} + B)^{-1}CV_X^{-T}.$$

From

$$\begin{aligned} R(Y + H) = R(Y) &+ V_A^{-1}(A\tilde{X} + B)^{-1}(AV_AYV_X^T + B)V_AH \\ &+V_A^{-1}(A\tilde{X} + B)^{-1}AV_AHV_X^TV_AY + V_A^{-1}(A\tilde{X} + B)^{-1}AV_AHV_X^TV_AH, \end{aligned} \tag{4.26}$$

we have

$$R'_Y(H) = V_A^{-1}(A\tilde{X} + B)^{-1}(AV_AYV_X^T + B)V_AH + V_A^{-1}(A\tilde{X} + B)^{-1}AV_AHV_X^TV_AY,$$

so that $R'_{\tilde{Y}}(H) = H + ((A\tilde{X} + B)V_A)^{-1}AV_AH(V_X^{-1}\tilde{X}^TV_X)^T$. This, $((A\tilde{X} + B)V_A)^{-1}AV_A = \Lambda_A - U_A$ and (3.10) show $R'_{\tilde{Y}}(H) = H + (\Lambda_A - U_A)H(\Lambda_X - U_X)^T$.

LEMMA 4.1. *Let* ν, μ, $\tilde{X}$, V_A, V_X, W_A, W_X, Λ_A *and* Λ_X *be defined similarly to Lemma 3.1, and* S_A *and* T_A *be as the above. Define* S_X, T_X, s_A, s_X, t_A *and* t_X *similarly to Lemma 3.1, and* $D := e^{(n)}e^{(n)^T} + \nu\mu^T$. *Suppose* $\|s_A\|_\infty < 1$, $\|s_X\|_\infty < 1$ *and* $|D| > 0$, *and define* u_A *and* u_X *similarly to Lemma 3.1, and* $E := (u_A|\mu|^T + (|\nu| + u_A)u_X^T)./|D|$. *Then,* $A\tilde{X} + B$, V_A, W_A, V_X *and* W_X *are nonsingular. If* $\|E\|_{\max} < 1$, *additionally,* $R'_{\tilde{Y}}(H)$ *is invertible for* $R'_Y(H)$ *and* $\tilde{Y}$ *defined above.*

Proof. The inequalities $\|s_A\|_\infty < 1$, $\|s_X\|_\infty < 1$ and $|D| > 0$ show the nonsingularities of $A\tilde{X} + B$, V_A, W_A, V_X, W_X and $\Delta := I_n \otimes I_n + \Lambda_X \otimes \Lambda_A$. Since $\mathrm{vec}(R'_{\tilde{Y}}(H)) = \mathsf{P}\mathrm{vec}(H)$, where $\mathsf{P} := I_n \otimes I_n + (\Lambda_X - U_X) \otimes (\Lambda_A - U_A)$, we prove the nonsingularity of P. We have $\mathsf{P} = \Delta(I_{n^2} - \Delta^{-1}\Omega)$, where $\Omega := \Lambda_X \otimes U_A + U_X \otimes (\Lambda_A - U_A)$. The estimation analogous to (3.12) yields

$$\begin{aligned} |\Delta^{-1}\Omega|e^{(n^2)} &\le |\Delta^{-1}|(|\Lambda_X| \otimes |U_A| + |U_X| \otimes (|\Lambda_A| + |U_A|))\mathrm{vec}(e^{(n)}e^{(n)^T}) \\ &= |\Delta^{-1}|\mathrm{vec}(|U_A|e^{(n)}(|\Lambda_X|e^{(n)})^T + (|\Lambda_A|e^{(n)} + |U_A|e^{(n)})(|U_X|e^{(n)})^T) \\ &\le |\Delta^{-1}|\mathrm{vec}(u_A|\mu|^T + (|\nu| + u_A)u_X^T) = \mathrm{vec}(E). \end{aligned}$$

This and $\|E\|_{\max} < 1$ prove the nonsingularity of P. ∎

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

LEMMA 4.2. *Under the conditions in Lemma 4.1, let V_A, V_X, W_A, s_A, D, $\tilde{Y}$ and E be as in Lemma 4.1, $N(Y)$, K, v_A and J be similar to those in Lemma 3.2, $L := (J + (I_n + s_A v_A^T)|W_A A V_A|K|V_X^T V_A|K)./|D|$ and $M := L + \|L\|_E E$. Then, $\{N(Y) : Y \in \langle \tilde{Y}, K\rangle\} \subseteq \langle \tilde{Y}, M\rangle$.*

Proof. Let S_A, U_A, U_X, $R(Y)$ and $R'_Y(H)$ be as the above, $\tilde{X}$, Λ_A and Λ_X be as in Lemma 4.1, Δ and Ω be as in the proof of Lemma 4.1, Y_P be as in the proof of Lemma 3.2, and N_Y be the solution of the parameterized Stein equation

$$N_Y + (\Lambda_A - U_A)N_Y(\Lambda_X - U_X)^T = R'_{\tilde{Y}}(Y) - R(Y),$$

where $Y \in \langle \tilde{Y}, K\rangle$ is the parameter. We prove $|\tilde{Y} - N_Y| \leq M$, $\forall Y \in \langle \tilde{Y}, K\rangle$. From Lemma 4.1 or its proof, $I_n - S_A$, $A\tilde{X} + B$, V_A, Δ, and $I_{n^2} - \Delta^{-1}\Omega$ are nonsingular. Similarly to (3.14), we have

$$\mathrm{vec}(\tilde{Y} - N_Y) = (I_{n^2} - \Delta^{-1}\Omega)^{-1}\Delta^{-1}\mathrm{vec}(R(\tilde{Y} + Y_P) - R'_{\tilde{Y}}(Y_P)). \tag{4.27}$$

From (4.26) and $V_A^{-1}(A\tilde{X} + B)^{-1} = (I_n - S_A)^{-1}W_A$, it holds that

$$R(\tilde{Y} + Y_P) = R(\tilde{Y}) + R'_{\tilde{Y}}(Y_P) + (I_n - S_A)^{-1}W_A A V_A Y_P V_X^T V_A Y_P. \tag{4.28}$$

Analogously to (3.17), $|(I_n - S_A)^{-1}| \leq I_n + s_A v_A^T$ and $|R(\tilde{Y})| \leq J$ follow. These inequalities, (4.27), (4.28) and the estimation analogous to (3.18) show $\mathrm{vec}(|\tilde{Y} - N_Y|) \leq \mathrm{vec}(M)$, $\forall Y \in \langle \tilde{Y}, K\rangle$, i.e., $|\tilde{Y} - N_Y| \leq M$, $\forall Y \in \langle \tilde{Y}, K\rangle$. ∎

THEOREM 4.3. *Let $\tilde{X}$ be as in Lemma 4.1, K and M be as in Lemma 4.2, and M_S and G be similar to those in Theorem 3.3. With all the assumptions in Lemma 4.1, suppose $M \leq M_S \leq K$ and $G \geq |V_A|M_S|V_X|^T$. Then, $\langle \tilde{X}, G\rangle$ contains the solvent X_*.*

Proof. The discussion similar to the proof of Theorem 3.3 shows the result. ∎

LEMMA 4.4. *Under the conditions in Lemma 4.1, let V_A, V_X, W_A and s_A be as in Lemma 4.1, K, v_A, J and M be as in Lemma 4.2, M_S be as in Theorem 4.3, and M_0, σ and η be defined similarly to Lemma 3.5. If $J > 0$, $\|((I_n + s_A v_A^T)|W_A A V_A|M_0|V_X^T V_A|M_0)./J\|_{\max} \leq \sigma \leq 1/(4(1 + \mathtt{eps})^6)$, and $\mathrm{fl}_\triangle(\cdot)$, M_0, η, K, and M_S satisfy the conditions in Lemma 3.5, then $K > 0$ and $M \leq M_S \leq K$.*

Proof. The discussion analogous to the proof of Lemma 3.5 proves the result. ∎

THEOREM 4.5. *Under the conditions in Lemma 4.1, let $\tilde{X}$ be as in Lemma 4.1, and G and Z be defined similarly to Theorem 3.7. If $\langle \tilde{X}, G\rangle$ contains the solvent and $\|Z\|_{\max} < 1$, then the contained solvent is unique.*

Proof. Let X_1, X_2 and $\mathsf{S}(X_1, X_2)$ be as in Section 3, Q, Γ_1 and Γ_2 be as in the proof of Theorem 3.7, V_A and V_X be as in Lemma 4.1, and P be as in the proof of Lemma 4.1. We prove the nonsingularity of Q and $\|\mathsf{S}(X_1, X_2)\|_\infty \leq \|Z\|_{\max}$, $\forall X_1, X_2 \in \langle \tilde{X}, G\rangle$. From Lemma 4.1 or its proof, $A\tilde{X} + B$, V_A, V_X and P are nonsingular. Analogously to (3.20), we have $\mathsf{Q} = (V_X \otimes (A\tilde{X} + B)V_A)\mathsf{P}(V_X^{-1} \otimes V_A^{-1})$, which shows the nonsingularity of Q and $\mathsf{Q}^{-1} = (V_X \otimes V_A)\mathsf{P}^{-1}(V_X^{-1} \otimes ((A\tilde{X} + B)V_A)^{-1})$. This and the derivation analogous to (3.23) give

$$\begin{aligned}\mathsf{S}(X_1, X_2) = -(V_X \otimes V_A)\mathsf{P}^{-1}\Bigg(&V_X^{-1} \otimes \frac{1}{2}((A\tilde{X} + B)V_A)^{-1}A(\Gamma_1 + \Gamma_2)\\ &+ \frac{1}{2}V_X^{-1}(\Gamma_1 + \Gamma_2)^T \otimes ((A\tilde{X} + B)V_A)^{-1}A\Bigg).\end{aligned}$$

This and the estimations analogous to (3.24) and (3.25) prove $\|\mathsf{S}(X_1, X_2)\|_\infty \leq \|Z\|_{\max}$, $\forall X_1, X_2 \in \langle \tilde{X}, G\rangle$. ∎

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

REMARK 4.6. As mentioned in Section 1, (1.2) has $2n$ eigenvalues if and only if A is nonsingular. Therefore, we cannot discuss the dominant and minimal solvents when A is singular. Since this section takes the case when A is singular into account, we do not mention the verification of the dominance and minimality.

From the established theories, we propose:

ALGORITHM 2. This algorithm computes $\tilde{X}$ and G such that $\langle \tilde{X}, G\rangle \ni X_*$. The nonsingularity of $A\tilde{X}+B$ and uniqueness are moreover proved if successful.

Step 1. Similar to that in Algorithm 1 except the following: Compute Λ_A and V_A by numerical generalized eigendecomposition $AV_A \approx (A\tilde{X}+B)V_A\Lambda_A$. Calculate W_A such that $W_A = \mathrm{fl}(((A\tilde{X}+B)V_A)^{-1})$.

Steps 2 to 7. Similar to those in Algorithm 1.

Step 8. Similar to that in Algorithm 1 except the following: Compute σ such that $\sigma = \overline{\mathrm{fl}}(\|((I_n + s_A v_A^T)|W_A A V_A|M_0|V_X^T V_A|M_0)./J\|_{\max})$.

Steps 9 to 12. Similar to those in Algorithm 1. Terminate.

PROPOSITION 4.7. *Algorithm* 2 *has a cost of* $\mathcal{O}(n^3)$ *operations.*

Proof. The discussion similar to the proof of Proposition 3.11 shows the result. ∎

5. Numerical results. We used a computer with Intel Core 2.60GHz CPU, 8.00GB RAM and MATLAB R2012a with Intel Math Kernel Library and IEEE 754 double precision. In this environment, $\mathrm{fl}_\triangle(\cdot)$ satisfies the condition in Lemma 3.5 except the cases of underflow and overflow. We denote compared algorithms as follows:

`HD1`: Algorithm 4 in [6], where the nonsingularity of A and uniqueness are verified,
`HD2`: the iteration (5.1) in [6], where the nonsingularity of B is verified,
`M1`: Algorithm 1, where the nonsingularity of A, uniqueness, dominance and minimality are verified, and
`M2`: Algorithm 2, where the nonsingularity of $A\tilde{X}+B$ and uniqueness are verified.

In all the algorithms, the numerical eigendecompositon and generalized eigendecompositon, and inversion were executed by the MATLAB commands `eig` and `\`, respectively. In `HD1` and `HD2`, interval matrices containing inverse matrices were computed by the INTLAB [16] routine `verifylss`, and the maximum numbers of the iterations were set to 30. In `HD1`, `M1` and `M2`, we computed $\tilde{X}$ via the iteration (26) in [8] with stopping criterion (30) in [8]. The maximum number of the iteration was set to 30. Although Newton method with exact line search [8, 9] is also applicable, the iteration (26) was faster in the sense of actual CPU times, and gave $\tilde{X}$ with $Q(\tilde{X})$ having approximately equal ∞-norm. See http://web.cc.iwate-u.ac.jp/~miyajima/QME.zip for details of the implementations, where the MATLAB codes of the iteration (26) and compared algorithms (denoted by `B26.m`, `HD1.m`, `HD2.m`, `M1.m` and `M2.m`) are uploaded. To the author's best knowledge, the implementations of `HD1` and `HD2` by the authors of [6] are not available. Therefore, we implemented them by ourselves.

Let $\langle \tilde{X}, G\rangle \ni X_*$. To assess the qualities of the enclosures, define the maximum radius as $\max_{i,j} G_{ij}$. The algorithm `HD1`, `M1` and `M2` proved the uniqueness for the problems in which they succeeded. In Example 1, `M1` moreover proved the minimality for all the problems. The compared algorithms failed for some problems. The reasons for the failure of `HD1` in Examples 1 and 2 are it did not succeed after 30 iterations, and `NaN` and `Inf` were included in $\mathrm{fl}(\tilde{X} + A^{-1}B)$, respectively. That of `HD2` is it did not succeed after 30 iterations.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

That of M1 is NaN and Inf were included in W_A.

Example 1. In this example, we observe the maximum radii and CPU times for various n. Consider (1.1), where $A = I_n$,

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

This problem arises in a damped mass-spring system [8] and is treated also in [6, Section 6]. Table 1 reports the obtained radii and CPU times (sec) for various n. The actual iteration numbers of HD1 were one when $n = 500, 600, 700$, two when $n = 800$, and four when $n = 900$. We see from Table 1 that M1 and M2 were faster than HD1.

n	HD1	HD2	M1	M2	HD1	HD2	M1	M2
500	4.3e–12	failed	4.3e–12	4.3e–12	1.1e+1	failed	2.8e+0	2.8e+0
600	4.9e–12	failed	6.4e–12	5.2e–12	2.0e+1	failed	5.7e+0	5.6e+0
700	6.1e–12	failed	5.7e–12	5.7e–12	3.1e+1	failed	1.0e+1	1.0e+1
800	7.9e–12	failed	6.8e–12	7.0e–12	1.2e+2	failed	1.6e+1	1.6e+1
900	1.1e–11	failed	7.4e–12	7.4e–12	3.6e+2	failed	2.5e+1	2.5e+1
1000	failed	failed	8.6e–12	8.6e–12	failed	failed	3.7e+1	3.8e+1

TABLE 1
Obtained radii (left part) and CPU times (sec) (right part) in Example 1.

REMARK 5.1. When we compared the algorithms through [11, Example 4.1], which also treats the case where A is nonsingular, we observed the tendency analogous to Example 1. More specifically, M1 and M2 were faster than HD1, and the radii were comparable.

Example 2. In this example, we observe behavior of the algorithms when A is singular. Consider (1.1), where

$$A = \begin{bmatrix} 0 & 0.05 & 0.055 & 0.08 & 0.1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0.22 & 0 & 0 \\ 0 & 0 & 0 & 0.32 & 0.4 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0.01 & 0.02 & 0.01 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0.04 & -1 & 0 & 0 \\ 0 & 0 & 0.08 & -1 & 0 \\ 0 & 0 & 0 & 0.04 & -1 \end{bmatrix},$$

$$C = \begin{bmatrix} 0.1 & 0.04 & 0.025 & 0.01 & 0 \\ 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0.16 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.04 & 0 \end{bmatrix}.$$

This problem arises in a quasi-birth death process [8] and is treated also in [6, Section 6]. Table 2 displays the similar quantities to Table 1. The actual iteration number of HD2 was 21. Predictably, HD2 and M2 succeeded, whereas HD1 and M1 failed.

Electronic Journal of Linear Algebra, ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 34, pp. 137-151, March 2018.
http://repository.uwyo.edu/ela

HD1	HD2	M1	M2	HD1	HD2	M1	M2
failed	2.2e–16	failed	1.1e–13	failed	1.9e–2	failed	1.5e–2

TABLE 2
Obtained radii (left part) and CPU times (sec) (right part) in Example 2.

6. Conclusion. In this paper, we proposed Algorithms 1 and 2, and reported the numerical results. By exploiting the theory in [13, Section 2.2], modification of these algorithms adopting block diagonalization [2] instead of the generalized eigendecomposition and/or eigendecomposition will be possible. This modification will be effective when V_A and/or V_X are singular or ill-conditioned.

Acknowledgment. The author acknowledges the referees for valuable comments.

REFERENCES

[1] H. Arndt. On the interval systems $[x] = [A][x] + [b]$ and the powers of interval matrices in complex interval arithmetics. *Reliab. Comput.*, 13:245–259, 2007.
[2] A. Bavely and G. Stewart. An algorithm for computing reducing subspaces by block diagonalization. *SIAM J. Numer. Anal.*, 16:359–367, 1979.
[3] M. Binder and M. Hashem Pesaran. Multivariate rational expectations models and macroeconometric modelling: A review and some new results. In: M. Hashem Pesaran and M. Wickens. (editors), *Handbook of Applied Econometrics: Macroeconomics*, Basil Blackwell, 139–187, 1999.
[4] G.H. Golub and C.F. Van Loan. *Matrix Computations*, fourth edition. The Johns Hopkins University Press, Baltimore, 2013.
[5] C.-H. Guo. On a quadratic matrix equation associated with an M-matrix. *IMA J. Numer. Anal.*, 23:11–27, 2003.
[6] B. Hashemi and M. Dehghan. Efficient computation of enclosures for the exact solvents of a quadratic matrix equation. *Electron. J. Linear Algebra*, 20:519–536, 2010.
[7] N.J. Higham. *Functions of Matrices: Theory and Computation*. SIAM Publications, Philadelphia, 2008.
[8] N.J. Higham and H.-M. Kim. Numerical analysis of a quadratic matrix equation. *IMA J. Numer. Anal.*, 20:499–519, 2000.
[9] N.J. Higham and H.-M. Kim. Solving a quadratic matrix equation by Newton's method with exact line searches. *SIAM J. Matrix Anal. Appl.*, 23:303–316, 2001.
[10] R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1994.
[11] J.-H. Long, X.-Y. Hu, and L. Zhang. Improved Newton's method with exact line searches to solve quadratic matrix equation. *J. Comp. Appl. Math.*, 222:645–654, 2008.
[12] A. Minamihata. Private communication, 2013.
[13] S. Miyajima. Fast enclosure for solutions of Sylvester equations. *Linear Algebra Appl.*, 439:856–878, 2013.
[14] S. Miyajima. Fast enclosure for all eigenvalues and invariant subspaces in generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 35:1205–1225, 2014.
[15] S.M. Rump. Verification methods for dense and sparse systems of equations. In: J. Herzberger (editor), *Topics in Validated Computations - Studies in Computational Mathematics*, Elsevier, Amsterdam, 63–136, 1994.
[16] S.M. Rump. INTLAB - INTerval LABoratory. In: T. Csendes (editor), *Developments in Reliable Computing*, Kluwer Academic Publishers, Dordrecht, 77–107, 1999.
[17] S.M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numer.*, 19:287–449, 2010.