SIAM J. MATRIX ANAL. APPL.
Vol. 33, No. 1, pp. 235–262

ON A NONLINEAR MATRIX EQUATION ARISING IN NANO RESEARCH*

CHUN-HUA GUO†, YUEH-CHENG KUO‡, AND WEN-WEI LIN§

Abstract. The matrix equation $X + A^\top X^{-1}A = Q$ arises in Green's function calculations in nano research, where A is a real square matrix and Q is a real symmetric matrix dependent on a parameter and is usually indefinite. In practice one is mainly interested in those values of the parameter for which the matrix equation has no stabilizing solutions. The solution of interest in this case is a special weakly stabilizing complex symmetric solution X_*, which is the limit of the unique stabilizing solution X_η of the perturbed equation $X + A^\top X^{-1}A = Q + i\eta I$, as $\eta \to 0^+$. It has been shown that a doubling algorithm can be used to compute X_η efficiently even for very small values of η, thus providing good approximations to X_*. It has been observed by nano scientists that a modified fixed-point method can sometimes be quite useful, particularly for computing X_η for many different values of the parameter. We provide a rigorous analysis of this modified fixed-point method and its variant and of their generalizations. We also show that the imaginary part X_I of the matrix X_* is positive semidefinite and we determine the rank of X_I in terms of the number of unimodular eigenvalues of the quadratic pencil $\lambda^2 A^\top - \lambda Q + A$. Finally we present a new structure-preserving algorithm that is applied directly on the equation $X + A^\top X^{-1}A = Q$. In doing so, we work with real arithmetic most of the time.

Key words. nonlinear matrix equation, complex symmetric solution, weakly stabilizing solution, fixed-point iteration, structure-preserving algorithm, Green's function

AMS subject classifications. 15A24, 65F30

DOI. 10.1137/100814706

1. Introduction.

The nonlinear matrix equation $X + A^\top X^{-1}A = Q$, where A is real and Q is real symmetric positive definite, arises in several applications and has been studied in [3, 7, 11, 21, 22, 27], for example.

In this paper we further study the nonlinear matrix equation

$$X + A^\top X^{-1}A = Q + i\eta I,$$

where $A \in \mathbb{R}^{n\times n}$, $Q = Q^\top \in \mathbb{R}^{n\times n}$, and $\eta \geq 0$, but Q is usually not positive definite. The equation arises from the nonequilibrium Green's function approach for treating quantum transport in nanodevices, where the system Hamiltonian is a semi-infinite or bi-infinite real symmetric matrix with special structures [1, 5, 17, 18, 25]. A first systematic mathematical study of the equation has already been undertaken in [12].

*Received by the editors November 12, 2010; accepted for publication (in revised form) December 7, 2011; published electronically March 15, 2012.
http://www.siam.org/journals/simax/33-1/81470.html

†Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada (chguo@math.uregina.ca). The work of this author was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

‡Department of Applied Mathematics, National University of Kaohsiung, Kaohsiung 811, Taiwan (yckuo@nuk.edu.tw). The work of this author was partially supported by the National Science Council in Taiwan.

§Department of Mathematics, National Taiwan University, Taipei 106, Taiwan, and Center of Mathematical Modelling and Scientific Computing, National Chiao Tung University, Hsinchu 300, Taiwan (wwlin@math.nctu.edu.tw). The work of this author was partially supported by the National Science Council and the National Center for Theoretical Sciences in Taiwan.

For the bi-infinite case, the Green's function corresponding to the scattering region $G_S \in \mathbb{C}^{n_s \times n_s}$, in which the nano scientists are interested, satisfies the relation [4, 17]

$$G_S = \left((\mathcal{E} + i0^+)I - H_S - C_{L,S}^\top G_{L,S} C_{L,S} - D_{S,R} G_{S,R} D_{S,R}^\top\right)^{-1},$$

where $\mathcal{E}$ is energy, a real number that may be negative, $H_S \in \mathbb{R}^{n_s \times n_s}$ is the Hamiltonian for the scattering region, $C_{L,S} \in \mathbb{R}^{n_\ell \times n_s}$ and $D_{S,R} \in \mathbb{R}^{n_s \times n_r}$ represent the coupling with the scattering region for the left lead and the right lead, respectively, and $G_{L,S} \in \mathbb{C}^{n_\ell \times n_\ell}$ and $G_{S,R} \in \mathbb{C}^{n_r \times n_r}$ are special solutions of the matrix equations

$$G_{L,S} = \left((\mathcal{E} + i0^+)I - B_L - A_L^\top G_{L,S} A_L\right)^{-1} \tag{1.1}$$

and

$$G_{S,R} = \left((\mathcal{E} + i0^+)I - B_R - A_R G_{S,R} A_R^\top\right)^{-1} \tag{1.2}$$

with A_L, $B_L = B_L^\top \in \mathbb{R}^{n_\ell \times n_\ell}$, and A_R, $B_R = B_R^\top \in \mathbb{R}^{n_r \times n_r}$. Since (1.1) and (1.2) have the same type, we only need to study (1.1). We simplify the notation n_ℓ to n. In nano research, one is mainly interested in the values of $\mathcal{E}$ for which $G_{L,S}$ in (1.1) has a nonzero imaginary part [18].

For each fixed $\mathcal{E}$, we replace 0^+ in (1.1) by a sufficiently small positive number η and consider the matrix equation

$$X = \left((\mathcal{E} + i\eta)I - B_L - A_L^\top X A_L\right)^{-1}. \tag{1.3}$$

It is shown in [12] that the required special solution $G_{L,S}$ of (1.1) is given by $G_{L,S} = \lim_{\eta \to 0^+} G_{L,S}(\eta)$ with $X = G_{L,S}(\eta)$ being the unique complex symmetric solution of (1.3) such that $\rho\left(G_{L,S}(\eta) A_L\right) < 1$, where $\rho(\cdot)$ denotes the spectral radius. Thus $G_{L,S}$ is a special complex symmetric solution of $X = \left(\mathcal{E}I - B_L - A_L^\top X A_L\right)^{-1}$ with $\rho\left(G_{L,S} A_L\right) \le 1$.

The question as to when $G_{L,S}$ has a nonzero imaginary part is answered in the following result from [12], where $\mathbb{T}$ denotes the unit circle.

THEOREM 1.1. *For $\lambda \in \mathbb{T}$, let the eigenvalues of $\psi_L(\lambda) = B_L + \lambda A_L + \lambda^{-1} A_L^\top$ be $\mu_{L,1}(\lambda) \le \cdots \le \mu_{L,n}(\lambda)$. Let*

$$\Delta_{L,i} = \left[\min_{|\lambda|=1} \mu_{L,i}(\lambda),\ \max_{|\lambda|=1} \mu_{L,i}(\lambda)\right]$$

and $\Delta_L = \bigcup_{i=1}^n \Delta_{L,i}$. Then $G_{L,S}$ is a real symmetric matrix if $\mathcal{E} \notin \Delta_L$. When $\mathcal{E} \in \Delta_L$, the quadratic pencil $\lambda^2 A_L^\top - \lambda(\mathcal{E}I - B_L) + A_L$ has eigenvalues on $\mathbb{T}$. If all these eigenvalues on $\mathbb{T}$ are simple (they must then be nonreal, as explained in the proof of Theorem 3.3), then $G_{L,S}$ has a nonzero imaginary part.

By replacing X in (1.3) with X^{-1}, we get the equation

$$X + A^\top X^{-1} A = Q_\eta, \tag{1.4}$$

where $A = A_L$ and $Q_\eta = Q + i\eta I$ with $Q = \mathcal{E}I - B_L$. So Q is a real symmetric matrix dependent on the parameter $\mathcal{E}$ and is usually indefinite. For $\eta > 0$, we need the stabilizing solution X of (1.4), which is the solution with $\rho(X^{-1}A) < 1$, and then $G_{L,S}(\eta) = X^{-1}$. The existence of the stabilizing solution was proved in [12] using advanced tools; an elementary proof has been given recently in [10]. When $\eta = 0$ and

$\mathcal{E} \in \Delta_L$, it follows from Theorem 1.1 that the required solution $X = G_{L,S}^{-1}$ of (1.4) is only weakly stabilizing, in the sense that $\rho(X^{-1}A) = 1$.

The size of the matrices in (1.4) can be very small or very large, depending on how the system Hamiltonian is obtained. If the Hamitonian is obtained from layer-based models, as in [18] and [25], then the size of the matrices is just the number of principal layers. In [18] considerable attention is paid to single-layer models and the more realistic double-layer models, which correspond to $n = 1$ and $n = 2$ in (1.4). We can say that (1.4) with $n \leq 10$ is already of significant practical interest. On the other hand, if the Hamiltonian is obtained from the discretization of a differential operator, as in [1], then the size of the matrices in (1.4) can be very large if a fine mesh grid is used.

One way to approximate $G_{L,S}$ is to take a very small $\eta > 0$ and compute $G_{L,S}(\eta)$. It is proved in [12] that the sequence $\{X_k\}$ from the basic fixed-point iteration (FPI)

$$X_{k+1} = Q_\eta - A^\top X_k^{-1} A, \tag{1.5}$$

with $X_0 = Q_\eta$, converges to $G_{L,S}(\eta)^{-1}$. And it follows that the sequence $\{Y_k\}$ from the basic FPI

$$Y_{k+1} = (Q_\eta - A^\top Y_k A)^{-1}, \tag{1.6}$$

with $Y_0 = Q_\eta^{-1}$, converges to $G_{L,S}(\eta)$. However, the convergence is very slow for $\mathcal{E} \in \Delta_L$ since $\rho(G_{L,S}(\eta)A) \approx 1$ for η close to 0. It is also shown in [12] that a doubling algorithm (DA) can be used to compute the desired solution $X = G_{L,S}(\eta)^{-1}$ of the equation (1.4) efficiently for each fixed value of $\mathcal{E}$. However, in practice the desired solution needs to be computed for many different $\mathcal{E}$ values. Since the DA is not a correction method, it cannot use the solution obtained for one $\mathcal{E}$ value as an initial approximation for the exact solution at a nearby $\mathcal{E}$ value. To compute the solutions corresponding to many $\mathcal{E}$ values, it may be more efficient to use a modified FPI together with the DA. Indeed, it is suggested in [25] that the following modified FPI be used to approximate $G_{L,S}(\eta)$:

$$Y_{k+1} = \frac{1}{2} Y_k + \frac{1}{2}(Q_\eta - A^\top Y_k A)^{-1}. \tag{1.7}$$

A variant of this FPI is given in [12] to approximate $G_{L,S}(\eta)^{-1}$,

$$X_{k+1} = \frac{1}{2} X_k + \frac{1}{2}(Q_\eta - A^\top X_k^{-1} A), \tag{1.8}$$

which requires less computational work each iteration. However, the convergence analysis of these two modified FPIs has been an open problem even for the special initial matrices $Y_0 = Q_\eta^{-1}$ and $X_0 = Q_\eta$, respectively.

Our first contribution in this paper is a proof of convergence (to the desired solutions) of these two modified FPIs and their generalizations for many choices of initial matrices. These methods can be used as correction methods. In this process we will show that the unique stabilizing solution $X = G_{L,S}(\eta)^{-1}$ of (1.4) is also the unique solution of (1.4) with a positive definite imaginary part. It follows that the imaginary part X_I of the matrix $G_{L,S}^{-1}$ is positive semidefinite. Our second contribution in this paper is a determination of the rank of X_I in terms of the number of eigenvalues on $\mathbb{T}$ of the quadratic pencil $\lambda^2 A^\top - \lambda Q + A$. Our third contribution is a structure-preserving algorithm (SA) that is applied directly on (1.4) with $\eta = 0$. In doing so, we work with real arithmetic most of the time.

2. Convergence analysis of FPIs. In this section we perform convergence analysis for some FPIs, including the modified FPIs (1.7) and (1.8). The main tool we need is the following important result due to Earle and Hamilton [6]. The presentation here follows [14, Theorem 3.1] and its proof.

THEOREM 2.1 (Earle–Hamilton theorem). *Let $\mathcal{D}$ be a nonempty domain in a complex Banach space Z and let $h : \mathcal{D} \to \mathcal{D}$ be a bounded holomorphic function. If $h(\mathcal{D})$ lies strictly inside $\mathcal{D}$, then h has a unique fixed point in $\mathcal{D}$. Moreover, the sequence $\{z_k\}$ defined by the fixed point iteration $z_{k+1} = h(z_k)$ converges to this fixed point for any $z_0 \in \mathcal{D}$.*

Now let Z be the complex Banach space $\mathbb{C}^{n\times n}$ equipped with the spectral norm. For any $K \in \mathbb{C}^{n\times n}$, its imaginary part is the Hermitian matrix

$$\mathrm{Im}K = \frac{1}{2i}(K - K^*).$$

For any Hermitian matrices X and Y, $X > Y$ ($X \geq Y$) means that $X - Y$ is positive definite (semidefinite). Let $\mathcal{D}_+ = \{X \in \mathbb{C}^{n\times n} : \mathrm{Im}X > 0\}$, $\mathcal{D}_- = \{X \in \mathbb{C}^{n\times n} : \mathrm{Im}X < 0\}$.

We start with a proof of convergence for the basic FPI (1.5) for many different choices of X_0, not just for $X_0 = Q_\eta$.

THEOREM 2.2. *For any $X_0 \in \mathcal{D}_+$, the sequence $\{X_k\}$ produced by the FPI* (1.5) *converges to the unique fixed point X_η in $\mathcal{D}_+$.*

Proof. Let $\mathcal{D} = \{X \in \mathbb{C}^{n\times n} : \mathrm{Im}X > \frac{\eta}{2}I\}$. For each $X \in \mathcal{D}$, X is invertible by Bendixson's theorem (see [26], for example) and we also have $\|X^{-1}\| < \frac{2}{\eta}$ (see [2, Corollary 4] or [13, Lemma 3.1]). Now let

$$f(X) = Q_\eta - A^\top X^{-1} A.$$

Then for $X \in \mathcal{D}$

$$\begin{aligned}\mathrm{Im}f(X) &= \mathrm{Im}Q_\eta - \frac{1}{2i}\left(A^\top X^{-1}A - (A^\top X^{-1}A)^*\right)\\ &= \eta I + A^\top X^{-1}(\mathrm{Im}X)(A^\top X^{-1})^* \geq \eta I.\end{aligned}$$

It follows that $f : \mathcal{D} \to \mathcal{D}$ is a bounded holomorphic function and $f(\mathcal{D})$ lies strictly inside $\mathcal{D}$. By the Earle–Hamilton theorem, f has a unique fixed point X_η in $\mathcal{D}$ and X_k converges to X_η for any $X_0 \in \mathcal{D}$. The theorem is proved by noting that $X_1 \in \mathcal{D}$ for any $X_0 \in \mathcal{D}_+$ and that any fixed point X_* in $\mathcal{D}_+$ must be in $\mathcal{D}$ by $X_* = Q_\eta - A^\top X_*^{-1} A$. ∎

Remark 2.1. Since $\{X_k\}$ converges to $G_{L,S}(\eta)^{-1}$ for $X_0 = Q_\eta \in \mathcal{D}_+$, we know that $X_\eta = G_{L,S}(\eta)^{-1}$ in Theorem 2.2. Thus X_η is the unique solution of (1.4) such that $\rho(X_\eta^{-1}A) < 1$, and it is also the unique solution of (1.4) in $\mathcal{D}_+$. If we have obtained a particular solution of (1.4) by some method and would like to know whether it is the required solution, the latter condition is easier to check.

The matrix $G_{L,S}(\eta)$ can also be computed directly by using (1.6) for many different choices of Y_0. Note that the FPI (1.6) is $Y_{k+1} = \hat{f}(Y_k)$ with

$$\hat{f}(Y) = (Q_\eta - A^\top Y A)^{-1}.$$

COROLLARY 2.3. *For any $Y_0 \in \mathcal{D}_-$, the sequence $\{Y_k\}$ produced by the FPI* (1.6) *converges to the unique fixed point $Y_\eta = G_{L,S}(\eta)$ in $\mathcal{D}_-$.*

Proof. For any $Y_0 \in \mathcal{D}_-$, Y_0 is invertible by Bendixson's theorem. We now take $X_0 = Y_0^{-1}$ in (1.5). Then $X_0 \in \mathcal{D}_+$ since $\mathrm{Im}X_0 = -Y_0^{-1}(\mathrm{Im}Y_0)Y_0^{-*}$. It follows that the sequence $\{Y_k\}$ is well defined and related to the sequence $\{X_k\}$ from (1.5) by $Y_k = X_k^{-1}$. Thus $\{Y_k\}$ converges to $Y_\eta = X_\eta^{-1} \in \mathcal{D}_-$. Since X_η is the unique fixed point of f in $\mathcal{D}_+$, Y_η is the unique fixed point of $\hat{f}$ in $\mathcal{D}_-$. ∎

For faster convergence, we consider the modified FPI for (1.4)

$$X_{k+1} = (1-c)X_k + c(Q_\eta - A^\top X_k^{-1} A), \quad 0 < c < 1, \tag{2.1}$$

or $X_{k+1} = g(X_k)$ with the function g defined by

$$g(X) = (1-c)X + cf(X). \tag{2.2}$$

Note that $f(X) = X$ if and only if $g(X) = X$. So f and g have the same fixed points. Note also that the FPI (1.8) is a special case of the FPI (2.1) with $c = \frac{1}{2}$. We can now prove the following general result.

THEOREM 2.4. *For any $X_0 \in \mathcal{D}_+$, the FPI $X_{k+1} = g(X_k)$ converges to the unique fixed point X_η in $\mathcal{D}_+$.*

Proof. For any $X_0 \in \mathcal{D}_+$, X_1 is well defined and $\mathrm{Im}X_1 > c\eta I$. Let b be any number such that $b > \|X_1\|$ and $b > 2(\|Q_\eta\| + \frac{1}{c\eta}\|A\|^2)$. Let $\mathcal{D} = \{X \in \mathbb{C}^{n\times n} : \mathrm{Im}X > c\eta I,\ \|X\| < b\}$. Thus $X_1 \in \mathcal{D}$. For each $X \in \mathcal{D}$, X is invertible and $\|X^{-1}\| < \frac{1}{c\eta}$, as before. Then for $X \in \mathcal{D}$

$$\mathrm{Im}g(X) = (1-c)\mathrm{Im}X + c\mathrm{Im}f(X) > (1-c)c\eta I + c\eta I = (2-c)c\eta I$$

and

$$\|g(X)\| \le (1-c)\|X\| + c\left(\|Q_\eta\| + \frac{1}{c\eta}\|A\|^2\right) < (1-c)b + c\frac{b}{2} = \left(1 - \frac{c}{2}\right)b.$$

It follows that $g : \mathcal{D} \to \mathcal{D}$ is a bounded holomorphic function and $g(\mathcal{D})$ lies strictly inside $\mathcal{D}$. By the Earle–Hamilton theorem, X_k converges to the unique fixed point of g in $\mathcal{D}$, which must be X_η. ∎

Similarly, we consider the modified FPI

$$Y_{k+1} = (1-c)Y_k + c(Q_\eta - A^\top Y_k A)^{-1}, \quad 0 < c < 1, \tag{2.3}$$

or $Y_{k+1} = \hat{g}(Y_k)$ with the function $\hat{g}$ defined by

$$\hat{g}(Y) = (1-c)Y + c\hat{f}(Y). \tag{2.4}$$

The FPI (2.3) includes (1.7) as a special case. Note that $\hat{f}(Y) = Y$ if and only if $\hat{g}(Y) = Y$. So $\hat{f}$ and $\hat{g}$ have the same fixed points. However, there are no simple relations between X_k from (2.1) and Y_k from (2.3).

THEOREM 2.5. *For any $Y_0 \in \mathcal{D}_-$, the FPI $Y_{k+1} = \hat{g}(Y_k)$ converges to the unique fixed point Y_η in $\mathcal{D}_-$.*

Proof. Take $b > 2/\eta$, and let $\mathcal{D} = \{Y \in \mathbb{C}^{n\times n} : \mathrm{Im}Y < 0,\ \|Y\| < b\}$. For any $Y \in \mathcal{D}$, $\mathrm{Im}(Q_\eta - A^\top Y A) \ge \eta I$. So $\hat{g}(Y)$ is well defined and $\|(Q_\eta - A^\top Y A)^{-1}\| \le \frac{1}{\eta}$. Thus

$$\|\hat{g}(Y)\| < (1-c)b + c\frac{1}{\eta} < \left(1 - \frac{1}{2}c\right)b.$$

Moreover,

$$\begin{aligned}\mathrm{Im}(\hat{g}(Y)) &< c\mathrm{Im}(Q_\eta - A^\top Y A)^{-1}\\ &= -c(Q_\eta - A^\top Y A)^{-1}\mathrm{Im}(Q_\eta - A^\top Y A)(Q_\eta - A^\top Y A)^{-*}\\ &\le -c\eta\left((Q_\eta - A^\top Y A)^*(Q_\eta - A^\top Y A)\right)^{-1}\\ &\le -c\eta\|Q_\eta - A^\top Y A\|^{-2}I\\ &\le -\frac{c\eta}{(\|Q_\eta\| + b\|A\|^2)^2}I.\end{aligned}$$

It follows that $\hat{g} : \mathcal{D} \to \mathcal{D}$ is a bounded holomorphic function and $\hat{g}(\mathcal{D})$ lies strictly inside $\mathcal{D}$. By the Earle–Hamilton theorem, Y_k converges to Y_η for any $Y_0 \in \mathcal{D}$ and hence for any $Y_0 \in \mathcal{D}_-$ since we can take $b > \|Y_0\|$. ∎

We remark that the modified FPI (2.1) is slightly less expensive than the modified FPI (2.3) for each iteration. These two methods make improvements over the basic FPIs (1.5) and (1.6) in the same way, as explained below.

The rate of convergence of each FPI can be determined by computing the Fréchet derivative of the fixed-point mapping, as in [11]. For (1.5) and (1.6), we have

$$\limsup_{k\to\infty} \sqrt[k]{\|X_k - X_\eta\|} \le (\rho(X_\eta^{-1}A))^2, \quad \limsup_{k\to\infty} \sqrt[k]{\|Y_k - Y_\eta\|} \le (\rho(Y_\eta A))^2,$$

where equality typically holds. Recall that $Y_\eta = X_\eta^{-1}$. Note also that if $Y_0 = X_0^{-1}$ (with $X_0 \in \mathcal{D}_+$), then

$$\limsup_{k\to\infty} \sqrt[k]{\|Y_k - Y_\eta\|} = \limsup_{k\to\infty} \sqrt[k]{\|X_k - X_\eta\|}.$$

The Fréchet derivative at X_η of the function g in (2.2) is given by

$$g'(X_\eta)E = (1-c)E + cA^\top X_\eta^{-1} E X_\eta^{-1} A.$$

It follows that for FPI (2.1)

$$\limsup_{k\to\infty} \sqrt[k]{\|X_k - X_\eta\|} \le \rho\left((1-c)I + c(A^\top X_\eta^{-1}) \otimes (A^\top X_\eta^{-1})\right).$$

Similarly, the Fréchet derivative at Y_η of the function $\hat{g}$ in (2.4) is given by

$$\hat{g}'(Y_\eta)E = (1-c)E + cY_\eta A^\top E A Y_\eta.$$

It follows that for FPI (2.3)

$$\limsup_{k\to\infty} \sqrt[k]{\|Y_k - Y_\eta\|} \le \rho\left((1-c)I + c(Y_\eta A^\top) \otimes (Y_\eta A^\top)\right).$$

The rate of convergence of both modified FPIs is then determined by

$$r_\eta = \max_{i,j} \left|1 - c + c\lambda_i(X_\eta^{-1}A)\lambda_j(X_\eta^{-1}A)\right|.$$

The convergence of the modified FPIs is often much faster because r_η may be much smaller than 1 for a proper choice of c. An extreme example is the following.

Example 2.1. Consider the scalar equation (1.4) with $A = 1$ and $Q_\eta = \eta i$. It is easy to find that

$$X_\eta = \frac{1}{2}(\eta + \sqrt{4+\eta^2})i.$$

Thus $\rho(X_\eta^{-1}A) \to 1$ as $\eta \to 0^+$, while for $c = \frac{1}{2}$ we have $r_\eta \to 0$ as $\eta \to 0^+$.

Note that for $i, j = 1, \ldots, n$, the n^2 numbers $\lambda_i(X_\eta^{-1}A)\lambda_j(X_\eta^{-1}A)$ are inside $\mathbb{T}$ for each $\eta > 0$. In the limit $\eta \to 0^+$, at least one of them is on $\mathbb{T}$ if $\mathcal{E} \in \Delta_L$. So each of these numbers has the form $re^{i\theta}$ with $0 \le r \le 1$ and $0 \le \theta < 2\pi$. We may allow $c = 1$ in (2.1) and (2.3). In this case, the basic FPIs (1.5) and (1.6) are recovered. To get some insight, we first consider choosing $c \in (0, 1]$ such that for fixed (r, θ)

$$p(c) = \left|1 - c + cre^{i\theta}\right|$$

is minimized. If $(r, \theta) = (1, 0)$, then $p(c) = 1$ for all $c \in (0, 1]$. So assume $(r, \theta) \ne (1, 0)$. In this case,

$$p(c) = \left|1 - re^{i\theta}\right| \left|\frac{1}{1 - re^{i\theta}} - c\right|$$

is minimized on $(0, 1]$ when

$$c = \min\left\{1, \mathrm{Re}\frac{1}{1 - re^{i\theta}}\right\} = \min\left\{1, \frac{1 - r\cos\theta}{1 + r^2 - 2r\cos\theta}\right\} \ge \frac{1}{2},$$

where we have used $1 - r\cos\theta - \frac{1}{2}(1 + r^2 - 2r\cos\theta) = \frac{1}{2}(1 - r^2) \ge 0$. It follows that $c = 1$ is the best choice when $\frac{1-r\cos\theta}{1+r^2-2r\cos\theta} \ge 1$ or, in other words, when $z = re^{i\theta}$ is in the disk $\{z \in \mathbb{C} : |z - \frac{1}{2}| \le \frac{1}{2}\}$. Note that $p(1) = r$. It also follows that $c = \frac{1}{2}$ is the best choice when $r = 1$. Note that $p(\frac{1}{2}) = \frac{1}{2}\sqrt{1 + r^2 + 2r\cos\theta} = \frac{1}{2}\sqrt{2(1 + \cos\theta)}$ for $r = 1$.

We know from [12] that the eigenvalues of $X_\eta^{-1}A$ are precisely the n eigenvalues inside $\mathbb{T}$ of the quadratic pencil $\lambda^2 A^\top - \lambda Q_\eta + A$. We also know from Theorem 1.1 that the quadratic pencil $P(\lambda) = \lambda^2 A^\top - \lambda Q + A$ has some eigenvalues on $\mathbb{T}$ when $\mathcal{E} \in \Delta_L$. We can then make the following conclusions.

Remark 2.2. If $P(\lambda)$ has some eigenvalues near 1 or -1, the convergence of the FPI (2.1) is expected to be very slow for any choice of the parameter c. The DA is recommended for this case. If all eigenvalues of $P(\lambda)$ are clustered around $\pm i$, then the FPI (2.1) with $c = \frac{1}{2}$ is expected to be very efficient. In the general case, the optimal c is somewhere between $\frac{1}{2}$ and 1. If all eigenvalues of $X_\eta^{-1}A$ are available, we can determine the optimal c to minimize r_η using the bisection procedure in [9, section 6] with $[\frac{1}{2}, 1]$ as the initial interval. In practice we would not compute these eigenvalues for every $\mathcal{E}$ value. But we may use DA to compute X_η for one $\mathcal{E}$ value, determine the optimal c value for this $\mathcal{E}$, and use it as a suboptimal c for many nearby $\mathcal{E}$ values. If one does not want to compute any eigenvalues when using the FPI (2.1), then $c = \frac{1}{2}$ is recommended since this c value is best in handling the eigenvalues of $X_\eta^{-1}A$ that are extremely close to $\mathbb{T}$, which are the eigenvalues responsible for the extreme slow convergence of the basic FPIs (1.5) and (1.6).

We note that the approximate solution from the DA or the FPI for any $\mathcal{E}$ value is in $\mathcal{D}_+$ and can be used as an initial guess for the exact solution when using the FPI for other $\mathcal{E}$ values, with guaranteed convergence. However, even for small problems

the convergence of the FPI (2.1), with $c = \frac{1}{2}$, for example, can still be very slow when $P(\lambda)$ has some eigenvalues near 1 or -1. Moreover, the latter situation will happen for some energy values since $P(\lambda)$ is palindromic and thus, as $\mathcal{E}$ varies, eigenvalues of $P(\lambda)$ leave or enter the unit circle typically through the points ± 1. When n is large, there may be some eigenvalues of $P(\lambda)$ near ± 1 for almost all energy values of interest, and thus the convergence of the FPI may be very slow and other methods should be used.

3. Rank of $\mathbf{Im}(G_{L,S})$. The equation (1.4) has a unique stabilizing solution $X_\eta = G_{L,S}(\eta)^{-1}$ for any $\eta > 0$. Thus

$$X_\eta + A^\top X_\eta^{-1} A = Q_\eta \tag{3.1}$$

with $\rho(X_\eta^{-1}A) < 1$. We also know that X_η is complex symmetric. Write $X_\eta = X_{\eta,R} + iX_{\eta,I}$ with $X_{\eta,R}^\top = X_{\eta,R}$, $X_{\eta,I}^\top = X_{\eta,I} \in \mathbb{R}^{n\times n}$. We know from the previous section that $\mathrm{Im}(X_\eta) = X_{\eta,I} > 0$. Let

$$\varphi_\eta(\lambda) = \lambda A^\top + \lambda^{-1} A - Q_\eta. \tag{3.2}$$

By (3.1) the rational matrix-valued function $\varphi_\eta(\lambda)$ has the factorization

$$\varphi_\eta(\lambda) = \left(\lambda^{-1} I - S_\eta^\top\right) X_\eta \left(-\lambda I + S_\eta\right),$$

where $S_\eta = X_\eta^{-1} A$. Let $X = \lim_{\eta\to 0^+} X_\eta = G_{L,S}^{-1}$. Then

$$X + A^\top X^{-1} A = Q \tag{3.3}$$

with $\rho(X^{-1}A) \le 1$ and $\mathrm{Im}(X) \ge 0$. Note that $\varphi_0(\lambda) = \lambda A^\top + \lambda^{-1} A - Q$ has the factorization

$$\varphi_0(\lambda) = \left(\lambda^{-1} I - S^\top\right) X \left(-\lambda I + S\right), \tag{3.4}$$

where $S = X^{-1}A$. In particular, $\varphi_0(\lambda)$ is regular, i.e., its determinant is not indentically zero. In this section we will determine the rank of $\mathrm{Im}(X)$, which is the same as the rank of $\mathrm{Im}(G_{L,S})$ since $\mathrm{Im}(G_{L,S}) = \mathrm{Im}(X^{-1}) = -X^{-1}\mathrm{Im}(X)X^{-*}$.

Let

$$\mathcal{M} = \begin{bmatrix} A & 0 \\ Q & -I \end{bmatrix}, \quad \mathcal{L} = \begin{bmatrix} 0 & I \\ A^\top & 0 \end{bmatrix}. \tag{3.5}$$

Then the pencil $\mathcal{M} - \lambda\mathcal{L}$, also denoted by $(\mathcal{M}, \mathcal{L})$, is a linearization of the quadratic matrix polynomial

$$P(\lambda) = \lambda\varphi_0(\lambda) = \lambda^2 A^\top - \lambda Q + A. \tag{3.6}$$

It is easy to check that y and z are the right and left eigenvectors, respectively, corresponding to an eigenvalue λ of $P(\lambda)$ if and only if

$$\begin{bmatrix} y \\ Qy - \lambda A^\top y \end{bmatrix}, \quad \begin{bmatrix} z \\ -\overline{\lambda} z \end{bmatrix} \tag{3.7}$$

are the right and left eigenvectors of $(\mathcal{M}, \mathcal{L})$, respectively.

THEOREM 3.1. *Suppose that λ_0 is a semisimple eigenvalue of $\varphi_0(\lambda)$ on $\mathbb{T}$ with multiplicity m_0 and $Y \in \mathbb{C}^{n\times m_0}$ forms an orthonormal basis of right eigenvectors*

corresponding to λ_0. Then $iY^*(2\lambda_0 A^\top - Q)Y$ is a nonsingular Hermitian matrix. Let d_j, $j = 1, \ldots, \ell$, be the distinct eigenvalues of $iY^*(2\lambda_0 A^\top - Q)Y$ with multiplicities m_0^j, and let $\xi_j \in \mathbb{C}^{m_0 \times m_0^j}$ form an orthonormal basis of the eigenspace corresponding to d_j. Then for $\eta > 0$ sufficiently small and $j = 1, \ldots \ell$

$$\lambda_{j,\eta}^{(k)} = \lambda_0 - \frac{\lambda_0}{d_j}\eta + O(\eta^2),\ k = 1, \ldots, m_0^j,\ \text{and } y_{j,\eta} = Y\xi_j + O(\eta) \tag{3.8}$$

are perturbed eigenvalues and a basis of the corresponding invariant subspace of $\varphi_\eta(\lambda)$, respectively.

Proof. Since $P(\lambda_0)Y = \lambda_0\varphi_0(\lambda_0)Y = 0$ with $Y^*Y = I_{m_0}$ and $|\lambda_0| = 1$, we have

$$0^* = (P(\lambda_0)Y)^* = \frac{1}{\lambda_0^2}Y^*(\lambda_0^2 A^\top - \lambda_0 Q + A).$$

It follows that Y forms an orthonormal basis for left eigenvectors of $P(\lambda)$ corresponding to λ_0. From (3.7), we obtain that the column vectors of

$$\mathcal{Y}_R = \begin{bmatrix} Y \\ QY - \lambda_0 A^\top Y \end{bmatrix} \text{ and } \mathcal{Y}_L = \begin{bmatrix} Y \\ -\overline{\lambda}_0 Y \end{bmatrix}$$

form a basis of left and right eigenspaces of $\mathcal{M} - \lambda\mathcal{L}$ corresponding to λ_0, respectively. Since λ_0 is semisimple, the matrix

$$[Y^*, -\lambda_0 Y^*]\mathcal{L}\begin{bmatrix} Y \\ QY - \lambda_0 A^\top Y \end{bmatrix} = -Y^*(2\lambda_0 A^\top - Q)Y = -Y^*P'(\lambda_0)Y$$

is nonsingular. Let

$$\widetilde{\mathcal{Y}}_R = -\mathcal{Y}_R(Y^*P'(\lambda_0)Y)^{-1},\ \ \widetilde{\mathcal{Y}}_L = \mathcal{Y}_L.$$

Then we have

$$\widetilde{\mathcal{Y}}_L^*\mathcal{L}\widetilde{\mathcal{Y}}_R = I_{m_0} \text{ and } \widetilde{\mathcal{Y}}_L^*\mathcal{M}\widetilde{\mathcal{Y}}_R = \lambda_0 I_{m_0}. \tag{3.9}$$

For $\eta > 0$, sufficiently small, we consider the perturbed equation of $P(\lambda)$ by

$$P(\lambda) - \lambda i\eta I = \lambda^2 A^\top - \lambda(Q + i\eta I) + A = \lambda\varphi_\eta(\lambda).$$

Let $\mathcal{M}_\eta = \left[\begin{smallmatrix} A & 0 \\ Q + i\eta I & -I \end{smallmatrix}\right]$. Then $\mathcal{M}_\eta - \lambda\mathcal{L}$ is a linearization of $\lambda\varphi_\eta(\lambda)$. By (3.9) and [24, Chapter VI, Theorem 2.12] there are $\widehat{\mathcal{Y}}_R$ and $\widehat{\mathcal{Y}}_L$ such that $[\widetilde{\mathcal{Y}}_R\ \widehat{\mathcal{Y}}_R]$ and $[\widetilde{\mathcal{Y}}_L\ \widehat{\mathcal{Y}}_L]$ are nonsingular and

$$\begin{bmatrix} \widetilde{\mathcal{Y}}_L^* \\ \widehat{\mathcal{Y}}_L^* \end{bmatrix}\mathcal{M}\left[\widetilde{\mathcal{Y}}_R\ \widehat{\mathcal{Y}}_R\right] = \begin{bmatrix} \lambda_0 I_{m_0} & 0 \\ 0 & \widehat{\mathcal{M}} \end{bmatrix},\quad \begin{bmatrix} \widetilde{\mathcal{Y}}_L^* \\ \widehat{\mathcal{Y}}_L^* \end{bmatrix}\mathcal{L}\left[\widetilde{\mathcal{Y}}_R\ \widehat{\mathcal{Y}}_R\right] = \begin{bmatrix} I_{m_0} & 0 \\ 0 & \widehat{\mathcal{L}} \end{bmatrix}.$$

Then, by [24, Chapter VI, Theorem 2.15] there exist matrices $\Delta_1(\eta) = O(\eta)$ and $\Delta_2(\eta) = O(\eta^2)$ such that the column vectors of $\widetilde{\mathcal{Y}}_R + \Delta_1(\eta)$ span the right eigenspace of $(\mathcal{M}_\eta, \mathcal{L})$ corresponding to $(\lambda_0 I_{m_0} + \eta E_{11} + \Delta_2(\eta), I_{m_0})$, where

$$\begin{aligned} E_{11} = \widetilde{\mathcal{Y}}_L^*\begin{bmatrix} 0 & 0 \\ iI & 0 \end{bmatrix}\widetilde{\mathcal{Y}}_R &= \lambda_0 Y^*(iI)Y(Y^*P'(\lambda_0)Y)^{-1} \\ &= -\lambda_0(iY^*(2\lambda_0 A^\top - Q)Y)^{-1}. \end{aligned} \tag{3.10}$$

The matrix $iY^*(2\lambda_0 A^\top - Q)Y$ in (3.10) is Hermitian since

$$\begin{aligned} iY^*(2\lambda_0 A^\top - Q)Y &= iY^*\varphi_0(\lambda_0)Y + i\lambda_0 Y^* A^\top Y - i\overline{\lambda}_0 Y^* AY \\ &= i\lambda_0 Y^* A^\top Y + (i\lambda_0 Y^* A^\top Y)^*. \end{aligned} \tag{3.11}$$

Let d_j for $j = 1, \ldots, \ell$ be the distinct eigenvalues of $iY^*(2\lambda_0 A^\top - Q)Y$ with multiplicities m_0^j, and let $\xi_j \in \mathbb{C}^{m_0 \times m_0^j}$ form an orthonormal basis of the eigenspace corresponding to d_j. Then we have

$$\Phi^* E_{11} \Phi = \operatorname{diag}\left(\frac{-\lambda_0}{d_1} I_{m_0^1}, \ldots, \frac{-\lambda_0}{d_\ell} I_{m_0^\ell} \right),$$

where $\Phi = [\xi_1, \ldots, \xi_\ell] \in \mathbb{C}^{m_0 \times m_0}$. It follows that $\lambda_0 I_{m_0} + \eta E_{11} + \Delta_2(\eta)$ is similar to

$$\lambda_0 I_{m_0} + \operatorname{diag}\left(\frac{-\lambda_0}{d_1} \eta I_{m_0^1}, \ldots, \frac{-\lambda_0}{d_\ell} \eta I_{m_0^\ell} \right) + \Delta_3(\eta)$$

for some $\Delta_3(\eta) = O(\eta^2)$. Then for each $j \in \{1, 2, \ldots, \ell\}$, the perturbed eigenvalues $\lambda_{j,\eta}^{(k)}$, $k = 1, \ldots, m_0^j$, and a basis of the corresponding invariant subspace of $\mathcal{M}_\eta - \lambda \mathcal{L}$ with $\lambda_{j,\eta}^{(k)}|_{\eta=0} = \lambda_0$ can be expressed by

$$\lambda_{j,\eta}^{(k)} = \lambda_0 - \frac{\lambda_0}{d_j}\eta + O(\eta^2), \ k = 1, \ldots, m_0^j, \tag{3.12a}$$

and

$$\zeta_{j,\eta} = \mathcal{Y}_R \xi_j + O(\eta). \tag{3.12b}$$

The second equation in (3.8) follows from (3.12b). ∎

LEMMA 3.2. *Let Z_η be the solution of the equation*

$$Z_\eta - R_\eta^* Z_\eta R_\eta = \eta W_\eta \text{ for } \eta > 0, \tag{3.13}$$

where $W_\eta \in \mathbb{C}^{m \times m}$ is positive definite, $R_\eta = e^{i\theta} I_m + \eta E_\eta$ with $\theta \in [0, 2\pi]$ fixed, and $E_\eta \in \mathbb{C}^{m \times m}$ is uniformly bounded such that $\rho(R_\eta) < 1$. Then Z_η is positive definite. Furthermore, if Z_η converges to Z_0 and W_η converges to a positive definite matrix W_0 as $\eta \to 0^+$, then Z_0 is also positive definite.

Proof. Since $\rho(R_\eta) < 1$ and ηW_η is positive definite, it is well known that Z_η is uniquely determined by (3.13) and is positive definite.

Since E_η is bounded, we have from (3.13) that

$$\begin{aligned} \eta W_\eta &= Z_\eta - \left(e^{-i\theta} I_m + \eta E_\eta^*\right) Z_\eta \left(e^{i\theta} I_m + \eta E_\eta\right) \\ &= -\eta e^{i\theta} E_\eta^* Z_\eta - \eta e^{-i\theta} Z_\eta E_\eta + O(\eta^2). \end{aligned}$$

This implies that

$$W_\eta = -e^{i\theta} E_\eta^* Z_\eta - e^{-i\theta} Z_\eta E_\eta + O(\eta). \tag{3.14}$$

If Z_η converges to Z_0 as $\eta \to 0^+$, then Z_0 is positive semidefinite. To prove that Z_0 is positive definite, it suffices to show that Z_0 is nonsingular. Suppose that $x \in \mathbb{C}^m$

such that $Z_0x = 0$. Then we have $Z_\eta x \to 0$ and $x^*Z_\eta \to 0$ as $\eta \to 0^+$. Multiplying (3.14) by x^* and x from the left and right, respectively, we have

$$x^*W_\eta x = -e^{i\theta}x^*E_\eta^*Z_\eta x - e^{-i\theta}x^*Z_\eta E_\eta x + O(\eta) \to 0 \ \text{ as } \eta \to 0^+.$$

Thus $x = 0$ because W_η converges to W_0 and W_0 is positive definite. It follows that Z_0 is positive definite. ∎

THEOREM 3.3. *The number of eigenvalues (counting multiplicities) of $\varphi_0(\lambda)$ on $\mathbb{T}$ must be even, say, $2m$. Let $X = \lim_{\eta\to 0^+} X_\eta$ be invertible and write $X = X_R + iX_I$ with $X_R = X_R^\top, X_I = X_I^\top \in \mathbb{R}^{n\times n}$. Then*

(i) $\operatorname{rank}(X_I) \le m$;

(ii) $\operatorname{rank}(X_I) = m$ *if all eigenvalues of $\varphi_0(\lambda)$ on $\mathbb{T}$ are semisimple and $\|S_\eta - S\|_2 = O(\eta)$ for $\eta > 0$ sufficiently small, where $S_\eta = X_\eta^{-1}A$ and $S = X^{-1}A$;*

(iii) $\operatorname{rank}(X_I) = m$ *if all eigenvalues of $\varphi_0(\lambda)$ on $\mathbb{T}$ are semisimple and each unimodular eigenvalue of multiplicity m_j is perturbed to m_j eigenvalues (of $\varphi_\eta(\lambda)$) inside the unit circle or to m_j eigenvalues outside the unit circle.*

Proof. Consider the real quadratic pencil $P(\lambda) = \lambda\varphi_0(\lambda) = \lambda^2A^\top - \lambda Q + A$. So $P(\lambda)$ and $\varphi_0(\lambda)$ have the same eigenvalues on $\mathbb{T}$. If $\lambda_0 \neq \pm 1$ is an eigenvalue of $P(\lambda)$ on $\mathbb{T}$ with multiplicity m_0, then so is $\overline{\lambda}_0$. Thus the total number of nonreal eigenvalues of $P(\lambda)$ on $\mathbb{T}$ must be even. Now the quadratic pencil $P_\eta(\lambda) = \lambda^2A^\top - \lambda(Q + i\eta I) + A$ is $\top$-palindromic, and it has no eigenvalues on $\mathbb{T}$ for any $\eta \neq 0$ [12]. If 1 (or -1) is an eigenvalue of $P(\lambda)$ with multiplicity r and Q in $P(\lambda)$ is perturbed to $Q + i\eta I$, then half of these r eigenvalues are perturbed to the inside of $\mathbb{T}$ and the other half are perturbed to the outside of $\mathbb{T}$. This means that r must be even. Thus the total number of eigenvalues of $\varphi_0(\lambda)$ on $\mathbb{T}$ is also even and is denoted by $2m$.

(i) By $X_\eta + A^\top X_\eta^{-1}A = Q_\eta$ we have

$$\begin{aligned} i(Q_\eta^* - Q_\eta) &= i(X_\eta^* - X_\eta) - iA^\top(X_\eta^{-1} - X_\eta^{-*})A \\ &= i(X_\eta^* - X_\eta) - (X_\eta^{-1}A)^* i(X_\eta^* - X_\eta)(X_\eta^{-1}A). \end{aligned}$$

Thus

$$K_\eta - S_\eta^* K_\eta S_\eta = 2\eta I, \tag{3.15}$$

where $K_\eta = i\left(X_\eta^* - X_\eta\right) = 2X_{\eta,I}$. Note that the eigenvalues of $S_\eta = X_\eta^{-1}A$ are the eigenvalues of $P_\eta(\lambda)$ inside $\mathbb{T}$. Since $X = \lim_{\eta\to 0^+} X_\eta$ is invertible, we have $S = X^{-1}A = \lim_{\eta\to 0^+} S_\eta$. Let

$$S = V_0 \begin{bmatrix} R_{0,1} & 0 \\ 0 & R_{0,2} \end{bmatrix} V_0^{-1} \tag{3.16}$$

be a spectral resolution of S, where $R_{0,1} \in \mathbb{C}^{m\times m}$ and $R_{0,2} \in \mathbb{C}^{(n-m)\times(n-m)}$ are upper triangular with $\sigma(R_{0,1}) \subseteq \mathbb{T}$ and $\sigma(R_{0,2}) \subseteq \mathbb{D} \equiv \{\lambda \in \mathbb{C} |\ |\lambda| < 1\}$, and $V_0 = [V_{0,1}, V_{0,2}]$ with $V_{0,1} \in \mathbb{C}^{n\times m}$ and $V_{0,2} \in \mathbb{C}^{n\times(n-m)}$ having unit column vectors. It follows from [24, Chapter V, Theorem 2.8] that there is a nonsingular matrix $V_\eta = [V_{\eta,1}, V_{\eta,2}]$ with $V_{\eta,1} \in \mathbb{C}^{n\times m}$ and $V_{\eta,2} \in \mathbb{C}^{n\times(n-m)}$ such that

$$S_\eta = V_\eta \begin{bmatrix} R_{\eta,1} & 0 \\ 0 & R_{\eta,2} \end{bmatrix} V_\eta^{-1} \tag{3.17}$$

and $R_{\eta,1} \to R_{0,1}$, $R_{\eta,2} \to R_{0,2}$, and $V_\eta \to V_0$, as $\eta \to 0^+$.

From (3.15) and (3.17) we have

$$V_\eta^* K_\eta V_\eta - \begin{bmatrix} R_{\eta,1}^* & 0 \\ 0 & R_{\eta,2}^* \end{bmatrix} V_\eta^* K_\eta V_\eta \begin{bmatrix} R_{\eta,1} & 0 \\ 0 & R_{\eta,2} \end{bmatrix} = 2\eta V_\eta^* V_\eta. \tag{3.18}$$

Let

$$H_\eta = V_\eta^* K_\eta V_\eta = \begin{bmatrix} H_{\eta,1} & H_{\eta,3} \\ H_{\eta,3}^* & H_{\eta,2} \end{bmatrix}, \quad V_\eta^* V_\eta = \begin{bmatrix} W_{\eta,1} & W_{\eta,3} \\ W_{\eta,3}^* & W_{\eta,2} \end{bmatrix}. \tag{3.19}$$

Then (3.18) becomes

$$H_{\eta,1} - R_{\eta,1}^* H_{\eta,1} R_{\eta,1} = 2\eta W_{\eta,1}, \tag{3.20a}$$

$$H_{\eta,2} - R_{\eta,2}^* H_{\eta,2} R_{\eta,2} = 2\eta W_{\eta,2}, \tag{3.20b}$$

$$H_{\eta,3} - R_{\eta,1}^* H_{\eta,3} R_{\eta,2} = 2\eta W_{\eta,3}. \tag{3.20c}$$

As $\eta \to 0^+$, $R_{\eta,1} \to R_{0,1}$ with $\rho(R_{0,1}) = 1$, $R_{\eta,2} \to R_{0,2}$ with $\rho(R_{0,2}) < 1$, and $W_{\eta,2}$ and $W_{\eta,3}$ are bounded. So we have $H_{\eta,2} \to 0$ from (3.20b) and $H_{\eta,3} \to 0$ from (3.20c). It follows from (3.19) that $K_\eta = 2X_{\eta,I}$ converges to $K_0 = 2X_I$ with $\operatorname{rank}(X_I) \le m$.

(ii) Suppose that eigenvalues of $\varphi_0(\lambda)$ on $\mathbb{T}$ are semisimple and $\|S_\eta - S\|_2 = O(\eta)$ for $\eta > 0$ sufficiently small. Then we will show that $H_{\eta,1}$ in (3.20a) converges to $H_{0,1}$ with $\operatorname{rank}(H_{0,1}) = m$. Let $\lambda_1, \dots, \lambda_r \in \mathbb{T}$ be the distinct semisimple eigenvalues of S with multiplicities $m_1, \dots, m_r$, respectively. Then (3.16) can be written as

$$S = V_0 \begin{bmatrix} D_{0,1} & 0 \\ 0 & R_{0,2} \end{bmatrix} V_0^{-1},$$

where $D_{0,1} = \operatorname{diag}\{\lambda_1 I_{m_1}, \dots, \lambda_r I_{m_r}\}$, $V_0 = [V_{0,\lambda_1}, \dots, V_{0,\lambda_r}, V_{0,2}]$, and $\sum_{i=1}^r m_i = m$. Now $S_\eta = S + (S_\eta - S)$ with $\|S_\eta - S\|_2 = O(\eta)$. By repeated application of [24, Chapter V, Theorem 2.8] there is a nonsingular matrix $V_\eta = [V_{\eta,\lambda_1}, \dots, V_{\eta,\lambda_r}, V_{\eta,2}] \in \mathbb{C}^{n\times n}$ such that

$$S_\eta = V_\eta \begin{bmatrix} D_{0,1} + \eta E_{\eta,1} & 0 \\ 0 & R_{0,2} + \eta E_{\eta,2} \end{bmatrix} V_\eta^{-1}$$

and $V_\eta \to V_0$ as $\eta \to 0^+$, where $E_{\eta,1} = \operatorname{diag}\{E_{m_1,\eta}^1, \dots, E_{m_r,\eta}^1\}$ with $E_{m_j,\eta}^1 \in \mathbb{C}^{m_j \times m_j}$ and $E_{\eta,2} \in \mathbb{C}^{(n-m)\times(n-m)}$ are such that $\|E_{m_j,\eta}^1\|_2 = O(1)$ for $j = 1, \dots, r$ and $\|E_{\eta,2}\|_2 = O(1)$.

The equation (3.20a) can then be written as

$$H_{\eta,1} - (D_{0,1} + \eta E_{\eta,1})^* H_{\eta,1} (D_{0,1} + \eta E_{\eta,1}) = 2\eta W_{\eta,1}. \tag{3.21}$$

Since $D_{0,1} + \eta E_{\eta,1}$ is a block diagonal matrix and all eigenvalues of its jth diagonal block converge to λ_j, with λ_j's distinct numbers on $\mathbb{T}$, we have

$$H_{\eta,1} = \operatorname{diag}\{H_{m_1,\eta}^1, \dots, H_{m_r,\eta}^1\} + O(\eta),$$

where $\operatorname{diag}\{H_{m_1,\eta}^1, \dots, H_{m_r,\eta}^1\}$ is the block diagonal of $H_{\eta,1}$. Then (3.21) gives

$$H_{m_j,\eta}^1 - (\lambda_j I_{m_j} + \eta E_{m_j,\eta}^1)^* H_{m_j,\eta}^1 (\lambda_j I_{m_j} + \eta E_{m_j,\eta}^1) = 2\eta W_{m_j,\eta}^1 \text{ for } j = 1, \dots, r,$$

where $W_{m_j,\eta}^1$ is the jth diagonal block of $W_{\eta,1}$. Since $W_{\eta,1}$ is positive definite and converges to a positive definite matrix, $W_{m_j,\eta}^1$, $j = 1, \dots, r$, are also positive definite

and converge to positive definite matrices. For $\eta > 0$, we have $\rho(\lambda_j I_{m_j} + \eta E^1_{m_j,\eta}) < 1$ for $j = 1, \ldots, r$ since $\rho(S_\eta) < 1$. By the assumption that X_η converges to X, we have that $H^1_{m_j,\eta}$ converges to $H^1_{m_j,0}$ for $j = 1, \ldots, r$. From Lemma 3.2, we obtain that $H^1_{m_j,0}$ is positive definite for $j = 1, \ldots, r$. Hence, $H_{\eta,1}$ converges to $H_{0,1}$ with $\text{rank}(H_{0,1}) = m$. It follows from (3.19) that $K_\eta = 2X_{\eta,I}$ converges to $K_0 = 2X_I$ with $\text{rank}(X_I) = m$.

(iii) It suffices to show that $\|S_\eta - S\|_2 = O(\eta)$ for $\eta > 0$ sufficiently small. Since X is a solution of $X + A^\top X^{-1} A = Q$, we have

$$\mathcal{M} \begin{bmatrix} I \\ X \end{bmatrix} = \mathcal{L} \begin{bmatrix} I \\ X \end{bmatrix} S,$$

where the pencil $(\mathcal{M}, \mathcal{L})$ is defined in (3.5) . Under the condition in (iii), the column space of $\left[\begin{smallmatrix} I \\ X \end{smallmatrix}\right]$ is a simple eigenspace of $(\mathcal{M}, \mathcal{L})$, in the terminology of [24]. It follows from [24, Chapter VI, Theorems 2.12 and 2.15] that

$$\begin{bmatrix} A & 0 \\ Q + i\eta I & -I \end{bmatrix} \begin{bmatrix} I + \eta F_{\eta,1} \\ X + \eta F_{\eta,2} \end{bmatrix} = \begin{bmatrix} 0 & I \\ A^\top & 0 \end{bmatrix} \begin{bmatrix} I + \eta F_{\eta,1} \\ X + \eta F_{\eta,2} \end{bmatrix} (S + \eta E_\eta),$$

where $F_{\eta,1}, F_{\eta,2}, E_\eta \in \mathbb{C}^{n \times n}$ with $\max\{\|F_{\eta,1}\|_2, \|F_{\eta,2}\|_2, \|E_\eta\|_2\} \le c$ for $\eta > 0$ sufficiently small and $c > 0$. It is easily seen that

$$\begin{aligned} X_\eta &= (X + \eta F_{\eta,2})(I + \eta F_{\eta,1})^{-1}, \\ S_\eta &= X_\eta^{-1} A = (I + \eta F_{\eta,1})(S + \eta E_\eta)(I + \eta F_{\eta,1})^{-1}. \end{aligned}$$

It follows that $\|S_\eta - S\|_2 = O(\eta)$ for $\eta > 0$ sufficiently small. ∎

Remark 3.1. Without the additional conditions in Theorem 3.3(ii) or (iii), $\text{rank}\,(X_I)$ could be much smaller than m. Consider the example with $A = I_n$ and $Q = 2I_n$. Then $\varphi_0(\lambda)$ has all $2n$ eigenvalues at 1 with partial multiplicities 2. Thus $m = n$, but it is easy to see that $\text{rank}\,(X_I) = 0$. For this example, we have $\|S_\eta - S\|_2 = O(\eta^{1/2})$ for $\eta > 0$ sufficiently small. We also know that the $2n$ eigenvalues of $\varphi_0(\lambda)$ at 1 are perturbed to n eigenvalues inside the unit circle and n eigenvalues outside the unit circle.

COROLLARY 3.4. *If $\varphi_0(\lambda)$ has no eigenvalues on $\mathbb{T}$, then X is real symmetric. Furthermore,* $\text{In}(X) = \text{In}\,(-\varphi_0(1))$. *Here* $\text{In}(W)$ *denotes the inertia of a matrix W.*

Proof. From Theorem 3.3, it is easy to see that X is a real symmetric matrix. Since X is real, $S = X^{-1}A$ is a real matrix. By setting $\lambda = 1$ in (3.4) we get $\varphi_0(1) = -(I - S^\top)X(I - S)$. Hence, $\text{In}(X) = \text{In}(-\varphi_0(1))$. ∎

COROLLARY 3.5. *If all eigenvalues of $\varphi_0(\lambda)$ are on $\mathbb{T}$ and are simple, then X_I is positive definite.*

Proof. The proof is by Theorem 3.3(iii) immediately. ∎

4. An SA. As explained in [12] and also in this paper, the required solution $X = G^{-1}_{L,S}$ is a particular weakly stabilizing solution of (3.3) and is given by $X = \lim_{\eta \to 0^+} X_\eta$, where X_η is the unique stabilizing solution of (1.4). We will call this particular solution *the weakly stabilizing solution* of (3.3). It can be approximated by X_η for a small η. For a fixed $\eta > 0$, X_η can be computed efficiently by the DA studied in [12] for all energy values.

In this section we will develop an SA that for most cases can find the weakly stabilizing solution of (3.3) more efficiently and more accurately than the DA by working on the equation (3.3) directly.

Consider the pencil $(\mathcal{M},\mathcal{L})$ given by (3.5). The simple relation $\mathcal{M}\left[\begin{smallmatrix} I \\ X \end{smallmatrix}\right] = \mathcal{L}\left[\begin{smallmatrix} I \\ X \end{smallmatrix}\right] X^{-1}A$ shows that the weakly stabilizing solution of (3.3) is obtained by $X = X_2X_1^{-1}$, where $\left[\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right]$ forms (or, more precisely, the columns of $\left[\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right]$ form) a basis for the invariant subspace of $(\mathcal{M},\mathcal{L})$ corresponding to its eigenvalues inside $\mathbb{T}$ and its eigenvalues on $\mathbb{T}$ that would be perturbed to the inside of $\mathbb{T}$ when Q is replaced by Q_η with $\eta > 0$.

We now assume that all unimodular eigenvalues $\lambda \neq \pm 1$ of $(\mathcal{M},\mathcal{L})$ are semisimple and the eigenvalues ± 1 (if they exist) have partial multiplicities 2. This assumption seems to hold generically. Under this assumption, for computing the weakly stabilizing solution we need to include all linearly independent eigenvectors associated with the eigenvalues ± 1 and use Theorem 3.1 to determine which half of the unimodular eigenvalues $\lambda \neq \pm 1$ should be included to compute the required invariant subspace.

We may use the QZ algorithm to determine this invariant subspace, but it would be better to exploit the structure of the pencil $(\mathcal{M},\mathcal{L})$. We will use the same approach as in [15] to develop an SA to find a basis for the desired invariant subspace of $(\mathcal{M},\mathcal{L})$ and then compute the weakly stabilizing solution of (3.3). The algorithm is still based on the $(S+S^{-1})$-transform in [19] and Patel's algorithm in [23], but some new issues need to be addressed here.

It is well known that $(\mathcal{M},\mathcal{L})$ is a symplectic pair, i.e., $(\mathcal{M},\mathcal{L})$ satisfies $\mathcal{M}\mathcal{J}\mathcal{M}^\top = \mathcal{L}\mathcal{J}\mathcal{L}^\top$, where $\mathcal{J} = \left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$. Furthermore, the eigenvalues of $(\mathcal{M},\mathcal{L})$ form reciprocal pairs $(\nu, 1/\nu)$, where we allow $\nu = 0, \infty$. We define the $(S+S^{-1})$-transform [19] of $(\mathcal{M},\mathcal{L})$ by

$$\mathcal{K} := \mathcal{M}\mathcal{J}\mathcal{L}^\top\mathcal{J} + \mathcal{L}\mathcal{J}\mathcal{M}^\top\mathcal{J} = \begin{bmatrix} Q & A - A^\top \\ A^\top - A & Q \end{bmatrix}, \tag{4.1a}$$

$$\mathcal{N} := \mathcal{L}\mathcal{J}\mathcal{L}^\top\mathcal{J} = \begin{bmatrix} A & 0 \\ 0 & A^\top \end{bmatrix}. \tag{4.1b}$$

Then $\mathcal{K}$ and $\mathcal{N}$ are both skew-Hamiltonian, i.e., $\mathcal{K}\mathcal{J} = \mathcal{J}\mathcal{K}^\top$ and $\mathcal{N}\mathcal{J} = \mathcal{J}\mathcal{N}^\top$. The relationship between eigenvalues of $(\mathcal{M},\mathcal{L})$ and $(\mathcal{K},\mathcal{N})$ and their Kronecker structures has been studied in [19, Theorem 3.2]. We will first extend that result to allow unimodular eigenvalues for $(\mathcal{M},\mathcal{L})$. The following preliminary result is needed.

LEMMA 4.1. *Let $N_r(\lambda) := \lambda I_r + N_r$, where N_r is the nilpotent matrix with $N_r(i,i+1) = 1$, $i = 1,\dots,r-1$, and zeros elsewhere. Let $\overset{eq.}{\sim}$ denote the equivalence between two matrix pairs. (Two matrix pairs (Y_1,Y_2) and (Z_1,Z_2) are called equivalent if there are nonsingular matrices U and V such that $UY_1V = Z_1$ and $UY_2V = Z_2$.) Then*

(i) *for $\lambda \neq 0, \pm 1$, $(N_r(\lambda) + N_r(\lambda)^{-1}, I_r) \overset{eq.}{\sim} (N_r(\lambda + 1/\lambda), I_r)$;*

(ii) *$(N_r^2 + I_r, N_r) \overset{eq.}{\sim} (I, N_r)$.*

Proof. (i) Since $\lambda \neq 0$, one can show that $N_r(\lambda)^{-1} \equiv [t_{j-i}]$ and $N_r(\lambda) + N_r(\lambda)^{-1} \equiv [s_{j-i}]$ are Toeplitz upper triangular with $t_k = (-1)^k\lambda^{-(k+1)}$ for $k = 0,1,\dots,r-1$, as well as $s_0 = \lambda + 1/\lambda$, $s_1 = 1 - \lambda^{-2}$, and $s_k = t_k$ for $k = 2,\dots,r-1$. Since $\lambda \neq \pm 1$, $s_1 = 1 - \lambda^{-2}$ is nonzero. It follows that $(N_r(\lambda) + N_r(\lambda)^{-1}, I_r) \overset{eq.}{\sim} (N_r(\lambda + 1/\lambda), I_r)$.

(ii) $(I + N_r^2, N_r) \overset{eq.}{\sim} (I_r, N_r(I + N_r^2)^{-1}) \overset{eq.}{\sim} (I_r, N_r - N_r^3 + N_r^5 - \cdots) \overset{eq.}{\sim} (I_r, N_r)$. ∎

THEOREM 4.2. *Suppose that $(\mathcal{M},\mathcal{L})$ has eigenvalues $\{\pm 1\}$ with partial multiplicities 2. Let $\gamma = \lambda + 1/\lambda$ ($\lambda = 0, \infty$ permitted). Then λ and $1/\lambda$ are eigenvalues of $(\mathcal{M},\mathcal{L})$ if and only if γ is a double eigenvalue of $(\mathcal{K},\mathcal{N})$. Furthermore, for $\lambda \neq \pm 1$ (i.e., $\gamma \neq \pm 2$) γ, λ and $1/\lambda$ have the same size Jordan blocks, i.e., they have the same partial multiplicities; for $\lambda = \pm 1$, $\gamma = \pm 2$ are semisimple eigenvalues of $(\mathcal{K},\mathcal{N})$.*

Proof. By the results on Kronecker canonical form for a symplectic pencil (see [20]) and by our assumption, there are nonsingular matrices $\mathcal{X}$ and $\mathcal{Y}$ such that

$$\mathcal{Y}\mathcal{M}\mathcal{X} = \begin{bmatrix} J & D \\ 0 & I_n \end{bmatrix}, \quad \mathcal{Y}\mathcal{L}\mathcal{X} = \begin{bmatrix} I_n & 0 \\ 0 & J \end{bmatrix}, \tag{4.2}$$

where $J = J_1 \oplus J_s \oplus J_0$, $J_1 = I_p \oplus (-I_q)$, J_s is the direct sum of Jordan blocks corresponding to nonzero eigenvalues λ_j of $(\mathcal{M}, \mathcal{L})$, where $|\lambda_j| < 1$ or $\lambda_j = e^{i\theta_j}$ with $\mathrm{Im}(\lambda_j) > 0$, and J_0 is the direct sum of nilpotent blocks corresponding to zero eigenvalues, and $D = I_p \oplus I_q \oplus 0_{n-r}$ with $r = p + q$.

Let $\mathcal{X}^{-1}\mathcal{J}\mathcal{X}^{-\top} = \left[\begin{smallmatrix} X_1 & X_2 \\ -X_2^\top & X_3 \end{smallmatrix}\right]$, where $X_i \in \mathbb{C}^{n\times n}$, $i = 1, 2, 3$. Thus $X_1^\top = -X_1$ and $X_3^\top = -X_3$. Using (4.2) in $\mathcal{M}\mathcal{J}\mathcal{M}^\top = \mathcal{L}\mathcal{J}\mathcal{L}^\top$ we get

$$JX_1J^\top - DX_2^\top J^\top + JX_2D + DX_3D = X_1, \tag{4.3a}$$

$$JX_2 + DX_3 = X_2J^\top, \tag{4.3b}$$

$$JX_3J^\top = X_3. \tag{4.3c}$$

Let $J_{s,0} = J_s \oplus J_0$. We partition X_3 and X_2 by $X_3 = \left[\begin{smallmatrix} X_{3,1} & X_{3,2} \\ -X_{3,2}^\top & X_{3,3} \end{smallmatrix}\right]$ and $X_2 = \left[\begin{smallmatrix} X_{2,1} & X_{2,2} \\ X_{2,3} & X_{2,4} \end{smallmatrix}\right]$, respectively, where $X_{2,1}, X_{3,1} \in \mathbb{C}^{r\times r}$. Comparing the diagonal blocks in (4.3b) we get

$$J_1X_{2,1} + X_{3,1} = X_{2,1}J_1^\top, \tag{4.4a}$$

$$J_{s,0}X_{2,4} = X_{2,4}J_{s,0}^\top. \tag{4.4b}$$

From (4.4a) we see that $X_{3,1}$ has the form $X_{3,1} = \left[\begin{smallmatrix} 0_p & \omega \\ -\omega^\top & 0_q \end{smallmatrix}\right]$. From (4.3c) we have $[I_p \oplus (-I_q)]X_{3,1}[I_p \oplus (-I_q)] = X_{3,1}$. It follows that $\omega = 0$ and thus $X_{3,1} = 0$. From (4.3c) we also have $J_1X_{3,2}J_{s,0}^\top = X_{3,2}$ and $J_{s,0}X_{3,3}J_{s,0}^\top = X_{3,3}$, from which we get $X_{3,2} = 0$ and $X_{3,3} = 0$. So we have $X_3 = 0$. Then (4.3b) becomes $JX_2 = X_2J^\top$, from which we get

$$X_{2,1} = \eta_p \oplus \eta_q, \quad X_{2,2} = X_{2,3}^\top = 0, \tag{4.5}$$

where $\eta_p \in \mathbb{C}^{p\times p}$ and $\eta_q \in \mathbb{C}^{q\times q}$. Moreover, $X_{2,1}$ and $X_{2,4}$ are nonsingular by the nonsingularity of $\mathcal{X}^{-1}\mathcal{J}\mathcal{X}^{-\top}$. Substituting (4.3b) into (4.3a) we get

$$X_1 = JX_1J^\top - DJX_2^\top + X_2J^\top D \equiv JX_1J^\top + V, \tag{4.6}$$

where $V = X_2J^\top D - DJX_2^\top = \left[\begin{smallmatrix} V_1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ with $V_1 = (\eta_p - \eta_p^\top) \oplus (\eta_q^\top - \eta_q)$ by (4.5). Partition $X_1 = \left[\begin{smallmatrix} X_{1,1} & X_{1,2} \\ -X_{1,2}^\top & X_{1,3} \end{smallmatrix}\right]$ with $X_{1,1} \in \mathbb{C}^{r\times r}$. From the equations for the $(1,2)$ and $(2,2)$ blocks of (4.6) we get $X_{1,2} = 0$ and $X_{1,3} = 0$, respectively. Furthermore, from the equation for the $(1,1)$ block in (4.6) we get $X_{1,1} = \xi_p \oplus \xi_q$ with $\xi_p \in \mathbb{C}^{p\times p}$ and $\xi_q \in \mathbb{C}^{q\times q}$, and we also get $\eta_p = \eta_p^\top$ and $\eta_q = \eta_q^\top$, i.e., $X_{2,1}^\top = X_{2,1}$.

From (4.2), (4.4), and Lemma 4.1(ii) we now have

$$
\begin{aligned}
(\mathcal{K},\mathcal{N}) &\overset{eq.}{\sim} (\mathcal{M}\mathcal{J}\mathcal{L}^\top + \mathcal{L}\mathcal{J}\mathcal{M}^\top, \mathcal{L}\mathcal{J}\mathcal{L}^\top)\\
&\overset{eq.}{\sim}\left(\begin{bmatrix} (J_1X_{1,1}+X_{1,1}J_1)\oplus 0_{n-r} & \big((J_1^2+I_r)X_{2,1}\big)\oplus\big((J_{s,0}^2+I_{n-r})X_{2,4}\big)\\ \big(X_{2,1}(J_1^2+I_r)\big)\oplus\big(X_{2,4}^\top((J_{s,0}^2)^\top+I_{n-r})\big) & 0_n\end{bmatrix},\right.\\
&\qquad\left.\begin{bmatrix} X_{1,1}\oplus 0_{n-r} & J_1X_{2,1}\oplus J_{s,0}X_{2,4}\\ X_{2,1}J_1\oplus X_{2,4}^\top J_{s,0}^\top & 0_n\end{bmatrix}\right)\\
&\overset{eq.}{\sim}\left(\begin{bmatrix} 2\xi_p\oplus(-2\xi_q) & 2I_r\\ 2I_r & 0_r\end{bmatrix}\oplus(J_s+J_s^{-1})\oplus I\oplus(J_s+J_s^{-1})\oplus I,\right.\\
&\qquad\left.\begin{bmatrix} \xi_p\oplus\xi_q & I_p\oplus(-I_q)\\ I_p\oplus(-I_q) & 0_r\end{bmatrix}\oplus I\oplus J_0\oplus I\oplus J_0\right)\\
&\overset{eq.}{\sim}\big(2I_{2p}\oplus(-2I_{2q})\oplus(J_s+J_s^{-1})\oplus I\oplus(J_s+J_s^{-1})\oplus I, I_{2p}\oplus I_{2q}\oplus I\oplus J_0\oplus I\oplus J_0\big).
\end{aligned}
$$

The proof is completed by using Lemma 4.1(i). ∎

4.1. Development of SA. It is helpful to keep in mind that the transform $\lambda\to\gamma$ achieves the following:

$$\{0,\infty\}\to\infty,\ \mathbb{T}\to[-2,2],\ \mathbb{R}\setminus\{\pm1,0\}\to\mathbb{R}\setminus[-2,2],\ \mathbb{C}\setminus(\mathbb{R}\cup\mathbb{T})\to\mathbb{C}\setminus\mathbb{R}.$$

By Theorem 4.2 and our assumption on the unimodular eigenvalues of $(\mathcal{M},\mathcal{L})$, all eigenvalues of $(\mathcal{K},\mathcal{N})$ in $[-2,2]$ are semisimple.

Based on the Patel approach [23] we first reduce $(\mathcal{K},\mathcal{N})$ to a block triangular matrix pair of the form

$$\mathcal{U}^\top\mathcal{K}\mathcal{Z}=\begin{bmatrix} K_1 & K_2\\ 0 & K_1^\top\end{bmatrix},\ \mathcal{U}^\top\mathcal{N}\mathcal{Z}=\begin{bmatrix} N_1 & N_2\\ 0 & N_1^\top\end{bmatrix}, \tag{4.7}$$

where K_1 and $N_1\in\mathbb{R}^{n\times n}$ are upper Hessenberg and upper triangular, respectively, K_2 and N_2 are skew symmetric, $\mathcal{U}$ and $\mathcal{Z}\in\mathbb{R}^{2n\times 2n}$ are orthogonal satisfying $\mathcal{U}^\top\mathcal{J}\mathcal{Z}=\mathcal{J}$. By the QZ algorithm, we have orthogonal matrices Q_1 and Z_1 such that

$$Q_1K_1Z_1=K_{11},\quad Q_1N_1Z_1=N_{11}, \tag{4.8}$$

where K_{11} and N_{11} are quasi-upper and upper triangular, respectively.

From (4.7) and (4.8) we see that the pair (K_{11},N_{11}) contains half the eigenvalues of $(\mathcal{K},\mathcal{N})$. We now reduce (K_{11},N_{11}) to the quasi-upper and upper block triangular matrix pair

$$\widetilde{U}^\top K_{11}\widetilde{Z}=\begin{bmatrix} I_{m_0} & \widehat{K}_{01} & \widehat{K}_{02} & \cdots & \widehat{K}_{0r}\\ & \Gamma_1 & \widehat{K}_{12} & \cdots & \widehat{K}_{1r}\\ & & \Gamma_2 & \ddots & \vdots\\ & & & \ddots & \widehat{K}_{r-1r}\\ & & & & \Gamma_r\end{bmatrix},$$

$$\widetilde{U}^\top N_{11}\widetilde{Z}=\operatorname{diag}\{\Gamma_0,I_{m_1},I_{m_2},\dots,I_{m_r}\},$$

where $m_0+m_1+\cdots+m_r=n$, Γ_0 is strictly upper triangular, $\Gamma_1=\operatorname{diag}\{g_1,\dots,g_{m_1}\}$ with $g_i\in[-2,2]$, and $\sigma(\Gamma_j)=\{\gamma_j\}\subseteq\mathbb{R}\setminus[-2,2]$ or $\sigma(\Gamma_j)=\{\gamma_j,\overline{\gamma}_j\}\subseteq\mathbb{C}\setminus\mathbb{R}$ with

$\sigma(\Gamma_j) \cap \sigma(\Gamma_i) = \emptyset$, $i \neq j$, $i, j = 2, \ldots, r$. Let

$$\widehat{K}_i = [\widehat{K}_{ii+1}, \ldots, \widehat{K}_{ir}], \quad i = 0, \ldots, r-1,$$

$$\widehat{\Gamma}_j = \begin{bmatrix} \Gamma_j & \widehat{K}_j \\ 0 & \widehat{\Gamma}_{j+1} \end{bmatrix}, \ \widehat{\Gamma}_r = \Gamma_r, \ j = 1, \ldots, r-1.$$

By solving some Sylvester equations

$$\begin{aligned} \Gamma_0 Z \widehat{\Gamma}_1 - Z &= \widehat{K}_0, \\ Z \widehat{\Gamma}_{i+1} - \Gamma_i Z &= \widehat{K}_i, \ i = 1 \ldots, r-1, \end{aligned} \tag{4.9}$$

we can reduce (K_{11}, N_{11}) to the quasi-upper and upper block diagonal matrix pair

$$\begin{aligned} \widehat{U}^\top K_{11} \widehat{Z} &= \operatorname{diag}\{I_{m_0}, \Gamma_1, \Gamma_2, \ldots, \Gamma_r\}, \\ \widehat{U}^\top N_{11} \widehat{Z} &= \operatorname{diag}\{\Gamma_0, I_{m_1}, I_{m_2}, \ldots, I_{m_r}\}. \end{aligned} \tag{4.10}$$

The procedure here is very much like the block diagonalization described in section 7.6.3 of [8]. Partition $\widehat{Z} = [\widehat{Z}_0, \widehat{Z}_1, \ldots, \widehat{Z}_r]$ with $\widehat{Z}_i \in \mathbb{R}^{n \times m_i}$ according to the block sizes of (4.10). It holds that

$$K_{11} \widehat{Z}_0 \Gamma_0 = N_{11} \widehat{Z}_0, \quad K_{11} \widehat{Z}_j = N_{11} \widehat{Z}_j \Gamma_j, \quad j = 1, \ldots, r. \tag{4.11}$$

It follows that for Z_1 from (4.8), $\mathcal{Z}(:, 1:n)(Z_1 \widehat{Z}_j)$ forms a basis for an invariant subspace of $(\mathcal{K}, \mathcal{N})$ for $j = 0, 1, \ldots, r$. In particular, the columns of $\mathcal{Z}(:, 1:n)(Z_1 \widehat{Z}_1)$ are real eigenvectors of $(\mathcal{K}, \mathcal{N})$ corresponding to real eigenvalues in $[-2, 2]$. We then need to get a suitable invariant subspace of $(\mathcal{M}, \mathcal{L})$ from each of these invariant subspaces for $(\mathcal{K}, \mathcal{N})$.

We start with two lemmas about solving the quadratic equation $\gamma = \lambda + 1/\lambda$ in the matrix form.

LEMMA 4.3. *Given a real quasi-upper triangular matrix*

$$\Gamma_s = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ & \ddots & \vdots \\ 0 & & \gamma_{mm} \end{bmatrix}, \tag{4.12}$$

where γ_{ii} is 1×1 or 2×2 block with $\sigma(\gamma_{ii}) \subseteq \mathbb{C} \setminus [-2, 2]$, $i = 1, \ldots, m$, the quadratic matrix equation

$$\Lambda_s^2 - \Gamma_s \Lambda_s + I = 0 \tag{4.13}$$

of Λ_s is uniquely solvable with Λ_s being real quasi-upper triangular with the same block form as Γ_s in (4.12) and $\sigma(\Lambda_s) \subseteq \mathbb{D} \equiv \{\lambda \in \mathbb{C} | \ |\lambda| < 1\}$.

Proof. Let

$$\Lambda_s = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ & \ddots & \vdots \\ 0 & & \lambda_{mm} \end{bmatrix}$$

have the same block form as Γ_s. We first solve the diagonal blocks $\{\lambda_{ii}\}_{i=1}^m$ of Λ_s from the quadratic equation $\lambda^2 - \gamma_{ii}\lambda + I_{[i]} = 0$, where $[i]$ denotes the size of γ_{ii}. Note that the scalar equation

$$\lambda^2 - \gamma\lambda + 1 = 0 \tag{4.14}$$

has no solutions on $\mathbb{T}$ for $\gamma \in \mathbb{C} \setminus [-2, 2]$. It always has one solution inside $\mathbb{T}$ and the other outside $\mathbb{T}$.

For $i = 1, \ldots, m$, if $\gamma_{ii} \in \mathbb{R} \setminus [-2, 2]$, then $\lambda_{ii} \in (-1, 1)$ is uniquely solved from (4.14) with $\gamma = \gamma_{ii}$. If $\gamma_{ii} \in \mathbb{R}^{2\times 2}$ with $\gamma_{ii} z = \gamma z$ for $z \neq 0$ and $\gamma \in \mathbb{C} \setminus \mathbb{R}$, then $\gamma_{ii} = [z, \overline{z}]\mathrm{diag}\,\{\gamma, \overline{\gamma}\}\,[z, \overline{z}]^{-1}$ and the required solution is $\lambda_{ii} = [z, \overline{z}]\mathrm{diag}\,\{\lambda, \overline{\lambda}\}\,[z, \overline{z}]^{-1} \in \mathbb{R}^{2\times 2}$, where $\lambda \in \mathbb{D}$ is uniquely solved from (4.14).

For $j > i$, comparing the (i, j) block on both sides of (4.13) and using $\lambda_{ii} - \gamma_{ii} = -\lambda_{ii}^{-1}$, we get

$$\lambda_{ij}\lambda_{jj} - \lambda_{ii}^{-1}\lambda_{ij} = \gamma_{ij}\lambda_{jj} + \sum_{\ell=i+1}^{j-1} (\gamma_{i\ell} - \lambda_{i\ell})\lambda_{\ell j}.$$

Since $\sigma(\lambda_{ii}^{-1}) \cap \sigma(\lambda_{jj}) = \emptyset$, $i, j = 1, \ldots, m$, the strictly upper triangular part of Λ_s can be determined by the following recursive formula:

For $d = 1, \ldots, m-1$,
  For $i = 1, \ldots, m-d$, $j = i + d$,
    $A := \lambda_{jj}^{\top} \otimes I_{[i]} - I_{[j]} \otimes \lambda_{ii}^{-1}$,
    $b := \gamma_{ij}\lambda_{jj} + \sum_{\ell=i+1}^{j-1}(\gamma_{i\ell} - \lambda_{i\ell})\lambda_{\ell j}$,
    $\lambda_{ij} = \mathrm{vec}^{-1}(A^{-1}\mathrm{vec}(b))$,
  end i,
end d.

Here $\otimes$ denotes the Kronecker product, vec is the operation of stacking the columns of a matrix into a vector, and vec^{-1} is its inverse operation. ∎

We note that (4.13) is a special case of the palindromic matrix equation studied recently in [16]. The desired solution Λ_s could also be obtained by applying the general formula in [16, Theorem 5], which involves the computation of $(\Gamma_s^{-1})^2$ and a matrix square root. However, for our special equation, the procedure given in the proof of Lemma 4.3 is more direct and numerically advantageous.

LEMMA 4.4. *Given a strictly upper triangular matrix $\Gamma_0 = [\gamma_{ij}] \in \mathbb{R}^{e\times e}$, the quadratic matrix equation*

$$\Gamma_0\Lambda_0^2 - \Lambda_0 + \Gamma_0 = 0 \text{ in } \Lambda_0 = [\lambda_{ij}] \in \mathbb{R}^{e\times e} \tag{4.15}$$

with Λ_0 being strictly upper triangular is uniquely solvable.

Proof. From (4.15) the matrix Λ_0 is uniquely determined by $\lambda_{i,i+j} = \gamma_{i,i+j}$, $i = 1, \ldots, e-2$, $j = 1, 2$, $\lambda_{e-1,e} = \gamma_{e-1,e}$, and

For $j = 3, \ldots, e$,
  For $i = 1, \ldots, e-j+1$,
    $\widehat{\lambda}_{i,i+j-1} = \sum_{\ell=i+1}^{i+j-2} \lambda_{i,\ell}\lambda_{\ell,i+j-1}$,
  end i,
  For $i = 1, \ldots, e-j$,
    $\lambda_{i,i+j} = \gamma_{i,i+j} + \sum_{\ell=i+1}^{i+j-2} \gamma_{i,\ell}\widehat{\lambda}_{\ell,i+j}$,
  end i,
end j. ∎

THEOREM 4.5. *Let Z_s form a basis for an invariant subspace of $(\mathcal{K}, \mathcal{N})$ corresponding to Γ_s with $\sigma(\Gamma_s) \subseteq \mathbb{C} \setminus [-2, 2]$, i.e., $\mathcal{K}Z_s = \mathcal{N}Z_s\Gamma_s$. Suppose that Λ_s solves $\Gamma_s = \Lambda_s + \Lambda_s^{-1}$ as in Lemma 4.3 with $\sigma(\Lambda_s) \subseteq \mathbb{D} \setminus \{0\}$. If the columns of $\mathcal{J}(\mathcal{L}^{\top}\mathcal{J}Z_s\Lambda_s - \mathcal{M}^{\top}\mathcal{J}Z_s)$ are linearly independent, then they form a basis for a stable invariant subspace of $(\mathcal{M}, \mathcal{L})$ corresponding to Λ_s.*

Proof. Since

$$\mathcal{K}Z_s - \mathcal{N}Z_s\Gamma_s = (\mathcal{M}\mathcal{J}\mathcal{L}^{\top} + \mathcal{L}\mathcal{J}\mathcal{M}^{\top})\mathcal{J}Z_s - \mathcal{L}\mathcal{J}\mathcal{L}^{\top}\mathcal{J}Z_s(\Lambda_s + \Lambda_s^{-1}) = 0,$$

we have $\mathcal{M}\mathcal{J}\left(\mathcal{L}^{\top}\mathcal{J}Z_s\Lambda_s - \mathcal{M}^{\top}\mathcal{J}Z_s\right) = \mathcal{L}\mathcal{J}\left(\mathcal{L}^{\top}\mathcal{J}Z_s\Lambda_s - \mathcal{M}^{\top}\mathcal{J}Z_s\right)\Lambda_s$. ∎

Remark 4.1. Let

$$Z_s = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}, \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \mathcal{J}(\mathcal{L}^\top \mathcal{J} Z_s \Lambda_s - \mathcal{M}^\top \mathcal{J} Z_s),$$

where each of X_1, X_2, Z_1, and Z_2 has n rows. Then direct computation gives $X_1 = Z_2\Lambda_s - Z_1$. However, it is more convenient to get $X_2 = QX_1 - A^\top X_1 \Lambda_s$ from $\mathcal{M}\left[\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right] = \mathcal{L}\left[\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right]\Lambda_s$.

We now explain how we can get eigenvectors of $(\mathcal{M}, \mathcal{L})$ from eigenvectors of $(\mathcal{K}, \mathcal{N})$ corresponding to eigenvalues in $[-2, 2]$.

THEOREM 4.6. *Let $v = \left[\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right]$ be a real eigenvector of $(\mathcal{K}, \mathcal{N})$ corresponding to eigenvalue $\gamma \in [-2, 2]$. Let λ be a solution of (4.14) and let $u_1 = \lambda v_2 - v_1$, $u_2 = Qv_1 - \lambda A^\top v_1$. Then $u = \left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right]$ is an eigenvector of $(\mathcal{M}, \mathcal{L})$ corresponding to eigenvalue λ if $u \neq 0$. Moreover, we indeed have $u \neq 0$ for each $\gamma \in (-2, 2)$.*

Proof. The first part is proved by direct computation as in the proof of Theorem 4.5. For the second part, we simply note that λ is not real when $\gamma \in (-2, 2)$, and then $u_1 \neq 0$ since v is a nonzero real vector. ∎

Remark 4.2. When $\gamma \in (-2, 2)$ is an eigenvalue of $(\mathcal{K}, \mathcal{N})$ with multiplicity $2k$ (or an eigenvalue of (K_{11}, N_{11}) with multiplicity k), we can use Theorem 4.6 to get k eigenvectors of $(\mathcal{M}, \mathcal{L})$ corresponding to eigenvalue λ from the k linearly independent eigenvectors of $(\mathcal{K}, \mathcal{N})$ we have already obtained. However, there is no guarantee that the k eigenvectors of $(\mathcal{K}, \mathcal{N})$ so obtained are also linearly independent.

When A is singular, $(\mathcal{M}, \mathcal{L})$ has eigenvalues at 0 and ∞. The following result will then be needed.

THEOREM 4.7. *Let $Z_\infty \in \mathbb{R}^{2n\times m}$ span an infinite invariant subspace of $(\mathcal{K}, \mathcal{N})$ corresponding to (I, Γ_0), where Γ_0 is strictly upper triangular, i.e., $\mathcal{N}Z_\infty = \mathcal{K}Z_\infty\Gamma_0$. Suppose that Λ_0 solves $\Gamma_0\Lambda_0^2 - \Lambda_0 + \Gamma_0 = 0$ as in Lemma 4.4 with Λ_0 being strictly upper triangular. If the columns of $\mathcal{J}(\mathcal{L}^\top \mathcal{J} Z_\infty \Lambda_0 - \mathcal{M}^\top \mathcal{J} Z_\infty)$ are linearly independent, then they form a basis for a zero invariant subpace of $(\mathcal{M}, \mathcal{L})$ corresponding to (Λ_0, I).*

Proof. Since $\Gamma_0 = \Lambda_0(I + \Lambda_0^2)^{-1}$, we have

$$\begin{aligned} \mathcal{N}Z_\infty(I + \Lambda_0^2) &= \mathcal{M}\mathcal{J}\mathcal{M}^\top \mathcal{J} Z_\infty + \mathcal{L}\mathcal{J}\mathcal{L}^\top \mathcal{J} Z_\infty \Lambda_0^2 = \mathcal{K}Z_\infty\Lambda_0 \\ &= \mathcal{L}\mathcal{J}\mathcal{M}^\top \mathcal{J} Z_\infty \Lambda_0 + \mathcal{M}\mathcal{J}\mathcal{L}^\top \mathcal{J} Z_\infty \Lambda_0, \end{aligned}$$

and then $\mathcal{M}\mathcal{J}(\mathcal{L}^\top \mathcal{J} Z_\infty \Lambda_0 - \mathcal{M}^\top \mathcal{J} Z_\infty) = \mathcal{L}\mathcal{J}(\mathcal{L}^\top \mathcal{J} Z_\infty \Lambda_0 - \mathcal{M}^\top \mathcal{J} Z_\infty)\Lambda_0$. ∎

We can now present an SA for the computation of the weakly stabilizing solution of (3.3).

SA.

Input: $A \in \mathbb{R}^{n\times n}$, $Q = Q^\top \in \mathbb{R}^{n\times n}$.

Output: The weakly stabilizing solution X of $X + A^\top X^{-1} A = Q$.

Step 1: Form the matrix pair $(\mathcal{K}, \mathcal{N})$ as in (4.1);

Step 2: Reduce $(\mathcal{K}, \mathcal{N})$ as in (4.7): $\mathcal{K} \leftarrow \mathcal{U}^\top \mathcal{K} \mathcal{Z} = \left[\begin{smallmatrix} K_1 & K_2 \\ 0 & K_1^\top \end{smallmatrix}\right]$, $\mathcal{N} \leftarrow \mathcal{U}^\top \mathcal{N} \mathcal{Z} = \left[\begin{smallmatrix} N_1 & N_2 \\ 0 & N_1^\top \end{smallmatrix}\right]$, where K_1 and $N_1 \in \mathbb{R}^{n\times n}$ are upper Hessenberg and upper triangular, respectively, $\mathcal{U}$ and $\mathcal{Z}$ are orthogonal satisfying $\mathcal{U}^\top \mathcal{J} \mathcal{Z} = \mathcal{J}$ (see a pseudo code in the appendix of [15]); Apply QZ algorithm to get (4.8);

Step 3: Compute eigenmatrix pairs $K_{11}\widehat{Z}_0\Gamma_0 = N_{11}\widehat{Z}_0$, $K_{11}\widehat{Z}_j = N_{11}\widehat{Z}_j\Gamma_j$ of (K_{11}, N_{11}), $j = 1, \ldots, r$, as in (4.11) by solving the Sylvester equations in (4.9) to get (4.10); $[\widehat{Z}_0, \widehat{Z}_1, \ldots, \widehat{Z}_r] \leftarrow Z_1[\widehat{Z}_0, \widehat{Z}_1, \ldots, \widehat{Z}_r]$, where Z_1 is from (4.8);

TABLE 4.1
Total number of flops for SA.

	Flops
Step 2	$\frac{170}{3}n^3$(Patel algorithm) + $39n^3$(QZ algorithm)
Step 3	$2n^3$(assume $m_0 = 0$, $m_i = 2$, $i = 1, \ldots, r$.)
Step 4	$4n^3$
Step 5	Between $4n^3$(no eigenvalues are on $\mathbb{T}$) and $8n^3$(all eigenvalues are on $\mathbb{T}$)
Step 6	$\frac{32}{3}n^3$
Total	$\approx 120n^3$

Step 4: $Z_\infty = \mathcal{Z}(:, 1:n)\widehat{Z}_0 \equiv \begin{bmatrix} Z_{\infty,1} \\ Z_{\infty,2} \end{bmatrix}$, $Z_j = \mathcal{Z}(:, 1:n)\widehat{Z}_j \equiv \begin{bmatrix} Z_{j,1} \\ Z_{j,2} \end{bmatrix}$, $j = 1, \ldots, r$;

Step 5-1: Use Lemma 4.4 to solve the strictly upper triangular Λ_0 for $\Gamma_0\Lambda_0^2 - \Lambda_0 + \Gamma_0 = 0$;
Compute $X_{\infty,1} = Z_{\infty,2}\Lambda_0 - Z_{\infty,1}$, $X_{\infty,2} = QX_{\infty,1} - A^\top X_{\infty,1}\Lambda_0$;
Set $X_1 \leftarrow X_{\infty,1}$, $X_2 \leftarrow X_{\infty,2}$ (by Theorem 4.7);

Step 5-2: For $k = 1, \ldots, m_1$,
Solve $\lambda = e^{i\theta}$ with $\mathrm{Im}(\lambda) \geq 0$ from (4.14) with $\gamma = g_k$;
Compute $x_1 = z_2 e^{i\theta} - z_1$, $x_2 = Qx_1 - e^{i\theta}A^\top x_1$, where $z_1 = Z_{1,1}e_k$, $z_2 = Z_{1,2}e_k$ (e_k is the kth column of the identity matrix);
If $\mathrm{Im}(e^{i\theta}x_1^* A^\top x_1) > 0$, then $x_1 \leftarrow \overline{x}_1$, $x_2 \leftarrow \overline{x}_2$;
Set $X_1 \leftarrow [X_1|x_1]$, $X_2 \leftarrow [X_2|x_2]$ (by Theorem 4.6 and Theorem 3.1);
end k;

Step 5-3: For $j = 2, \ldots, r$,
Use Lemma 4.3 to solve $\Lambda_j = \Lambda_s$ for (4.13) with $\Gamma_s = \Gamma_j$;
Compute $X_{j,1} = Z_{j,2}\Lambda_j - Z_{j,1}$, $X_{j,2} = QX_{j,1} - A^\top X_{j,1}\Lambda_j$;
Set $X_1 \leftarrow [X_1|X_{j,1}]$, $X_2 \leftarrow [X_2|X_{j,2}]$ (by Theorem 4.5);
end j;

Step 6: Compute $X = X_2 X_1^{-1}$.

In Step 5-2 of SA we have assumed that the nonsingular Hermitian matrix $iY^*(2\lambda_0 A^\top - Q)Y$ in Theorem 3.1 (which is the matrix in (3.11)) is definite. This assumption is the same as the assumption in Theorem 3.3(iii). Under this assumption we do not need to form the matrix Y but only need to check the sign of its diagonal element determined by any (normalized) eigenvector. We could have used Theorem 3.1 to choose the right eigenvectors whether the Hermitian matrix is definite or not, but this would increase computational work. We have thus chosen to use the simpler Step 5-2. We can always check in the end whether the imaginary part of the computed X is (nearly) positive semidefinite. If not, we can use a more complicated Step 5-2 according to Theorem 3.1 to recompute X.

To find the weakly stabilizing solution of (3.3), the total number of flops for SA is roughly $120n^3$, where a flop denotes a multiplication or an addition in real arithmetic. A more detailed counting is given in Table 4.1.

4.2. Comparison of SA with other methods. The required solution in the nano application is $X = \lim_{\eta \to 0^+} X_\eta$. By now, we have available five methods in two categories. In the first category, we have three methods that compute X_η for a small η as an approximation to X. They are FPI, Newton's method (NM), and the DA. These methods are discussed in [12]. In the second category, we have two methods that compute X directly. They are QZ and SA. In this paper we have

further studied FPI. We now know when and why it works well. It has been shown by numerical experiments in [12] that NM often converges to an undesired solution or even diverges. We now also have a better understanding of this. Basically, X_η is the unique solution of (3.1) with a positive definite imaginary part and, starting with an initial guess with a positive definite imaginary part, the iterates produced by NM often do not have a positive definite imaginary part (while the iterates from FPI always do). So NM is really not a contender for solving our specific problem and could be dropped from the first category. We also mentioned earlier that the convergence of FPI (say, (2.1) with $c = \frac{1}{2}$) is usually very slow for at least some of the energy values of interest. Therefore we only need to compare SA with DA and QZ, as general purpose methods.

Strictly speaking, SA and QZ are applicable only under some generic assumptions since they rely on our Theorem 3.1 to pick the right nonreal unimodular eigenvalues of $\varphi_0(\lambda)$ to get the desired weakly stabilizing solution. The difficulty associated with nongeneric eigenstructure of $\varphi_0(\lambda)$ is avoided in DA by the introduction of $\eta > 0$. More precisely, this difficulty is only concealed since the relation between $\|X_\eta - X\|$ and η is not clear in nongeneric cases. In the generic case, we often have $\|X_\eta - X\| = O(\eta)$, but the constant hidden in the big O notation could still be big. The smallest η that we have seen in the nano literature is 10^{-6}, although we have also experimented with $\eta = 10^{-10}$ for DA in [12]. In the experiments there, DA typically requires 26 iterations for $\eta = 10^{-6}$, and nearly 40 iterations for $\eta = 10^{-10}$. Note that DA requires $\frac{104}{3}n^3$ real flops each iteration. So in terms of flop counts, four DA iterations is already more expensive than SA. Moreover, since we take $\eta = 0$ directly in SA, the accuracy achieved by SA is usually much better than that achieved by DA with $\eta = 10^{-6}$. But we also note that DA is much easier to use. If we take η to be very small in DA, then DA could also have stability problems since the matrices to be inverted in DA iterations may also be very ill-conditioned in that case.

One potential problem with SA is that we cannot rule out the possibility that the column vectors generated in Steps 5-1, 5-2, and 5-3 of SA are linearly dependent (in exact arithmetic). In that case, SA fails. Note that QZ does not have this problem, but it requires about $440n^3$ flops, much more than the $120n^3$ flops required for SA. The accuracy achieved by QZ is not necessarily better than that from SA since QZ does not exploit the structure of the problem. Moreover, for QZ we may encounter the uncomfortable situation where the number of computed eigenvalues inside $\mathbb{T}$ does not match the number of computed eigenvalues outside $\mathbb{T}$.

Finally, we explain why X_1 is unlikely to be singular (in exact arithmetic) in Step 6 of SA. For this we need to explain why the column vectors generated in Steps 5-1, 5-2, and 5-3 of SA are unlikely to be linearly dependent. Take Step 5-3, for example, which is based on Theorem 4.5. Take $Z_s = Z_j$ and $\Gamma_s = \Gamma_j$ $(j = 2, \ldots, r)$; we need to examine whether the matrix $\mathcal{L}^\top \mathcal{J} Z_s \Lambda_s - \mathcal{M}^\top \mathcal{J} Z_s$ in Theorem 4.5 is of full column rank. We start with the following result.

THEOREM 4.8. *Let X_s, $X_u \in \mathbb{C}^{2n\times m}$ $(m \le n)$ form bases for stable and unstable invariant subspaces of $(\mathcal{L}^\top, \mathcal{M}^\top)$, respectively, corresponding to Λ_s and Λ_s^{-1}, where $\sigma(\Lambda_s) \subseteq \mathbb{D} \setminus \{0\}$, i.e.,*

$$\mathcal{L}^\top X_s = \mathcal{M}^\top X_s \Lambda_s, \quad \mathcal{L}^\top X_u = \mathcal{M}^\top X_u \Lambda_s^{-1}. \tag{4.16}$$

Then $\mathcal{J}^\top X_s$ and $\mathcal{J}^\top X_u$ span two linearly independent eigenspaces of $(\mathcal{K}, \mathcal{N})$ corresponding to $\Lambda_s + \Lambda_s^{-1}$.

Proof. From $\mathcal{L}\mathcal{J}\mathcal{L}^\top = \mathcal{M}\mathcal{J}\mathcal{M}^\top$ and (4.16) we have

$$\begin{aligned}\mathcal{K}\mathcal{J}^\top X_s &= \mathcal{M}\mathcal{J}\mathcal{L}^\top X_s + \mathcal{L}\mathcal{J}\mathcal{M}^\top X_s = \mathcal{M}\mathcal{J}\mathcal{M}^\top X_s\Lambda_s + \mathcal{L}\mathcal{J}\mathcal{L}^\top X_s\Lambda_s^{-1}\\ &= \mathcal{L}\mathcal{J}\mathcal{L}^\top X_s(\Lambda_s + \Lambda_s^{-1}) = \mathcal{N}\mathcal{J}^\top X_s(\Lambda_s + \Lambda_s^{-1}).\end{aligned}$$

Similarly, $\mathcal{K}\mathcal{J}^\top X_u = \mathcal{N}\mathcal{J}^\top X_u(\Lambda_s + \Lambda_s^{-1})$. ∎

Now we can write $Z_s = \mathcal{J}^\top X_s C_s + \mathcal{J}^\top X_u C_u$, where C_s and C_u are coefficient matrices. From (4.16) we have

$$\begin{aligned}\mathcal{L}^\top\mathcal{J}Z_s\Lambda_s - \mathcal{M}^\top\mathcal{J}Z_s &= \mathcal{L}^\top(X_sC_s + X_uC_u)\Lambda_s - \mathcal{M}^\top(X_sC_s + X_uC_u)\\ &= \left[\mathcal{L}^\top X_s,\ \mathcal{L}^\top X_u\right]\begin{bmatrix} C_s\Lambda_s - \Lambda_s^{-1}C_s \\ C_u\Lambda_s - \Lambda_s C_u \end{bmatrix}. \qquad (4.17)\end{aligned}$$

Since $\sigma(\Lambda_s) \subseteq \mathbb{D}\setminus\{0\}$, the linear mapping $\mathcal{T}$ defined by $\mathcal{T}(W) = W\Lambda_s - \Lambda_s^{-1}W$ is a bijection on $\mathbb{C}^{m\times m}$. Thus, the matrix $C_s\Lambda_s - \Lambda_s^{-1}C_s = \mathcal{T}(C_s)$ is generically nonsingular. When $\mathcal{T}(C_s)$ is nonsingular, $\mathcal{L}^\top\mathcal{J}Z_s\Lambda_s - \mathcal{M}^\top\mathcal{J}Z_s$ is also of full column rank since the columns of $\left[\mathcal{L}^\top X_s,\ \mathcal{L}^\top X_u\right]$ are linear independent. The latter assertion is clear when A is nonsingular in (3.5) but can also be shown by using the Kronecker form of $(\mathcal{L}^\top, \mathcal{M}^\top)$ when A is singular.

From the above discussions, we have the following practical strategy for computing the required solution X with reasonable accuracy. We first use SA. In the unlikely event that SA fails, we use QZ. If QZ also fails, then we use DA with a small $\eta > 0$ and might have to accept an approximation with lower accuracy.

5. Numerical results. All numerical experiments are carried out using MATLAB R2008b with IEEE double-precision floating-point arithmetic (*eps* $\approx 2.22\times 10^{-16}$) on the Linux system. We first illustrate the positivity of $X_{\eta,I}$ for $\eta > 0$, where $X_{\eta,I}$ is the imaginary part of the stabilizing solution X_η of (3.1), as well as the rank of $X_I = \lim_{\eta\to 0^+} X_{\eta,I}$. To measure the accuracy of the computed X_η we use the relative residual

$$\mathrm{RRes}_\eta = \frac{\|X_\eta + A^\top X_\eta^{-1}A - Q_\eta\|}{\|X_\eta\| + \|A\|^2\|X_\eta^{-1}\| + \|Q_\eta\|}, \qquad (5.1)$$

where $\|\cdot\|$ is the spectral norm.

Example 5.1. We randomly generate a real matrix A and a real symmetric matrix Q of dimension 6. Then we use the invariant subspace method (the QZ algorithm) to compute the complex symmetric stabilizing solution X_η of (3.1) with $\eta = 10^{-4}$, 10^{-8}, 10^{-12}, respectively. When $\eta = 0$, $\varphi_0(\lambda)$ has $2m = 6$ eigenvalues on $\mathbb{T}$, given by

$$\Lambda = \{-0.80913 \pm 0.58763i,\ 0.64993 \pm 0.76000i,\ -0.13000 \pm 0.99151i\}.$$

By Theorem 3.1 we determine that

$$\Lambda^s = \{-0.80913 + 0.58763i,\ 0.64993 + 0.76000i,\ -0.13000 - 0.99151i\}$$

is such that the perturbed eigenvalues of $\varphi_\eta(\lambda)$ $(\eta > 0)$ associated with each $\lambda^s \in \Lambda^s$ are inside the unit circle. Then we compute the weakly stabilizing solution X of (3.3) by using the invariant subspace corresponding to stable eigenvalues and eigenvalues in Λ^s. The numerical results are shown in Table 5.1, where $X_0 = X$.

We know from section 2 that $X_{\eta,I}$ is positive definite and we know from Theorem 3.3(iii) that $\mathrm{rank}(X_I) = m = 3$. These are confirmed by the numerical results shown in Table 5.2, where $X_{0,I} = X_I$.

TABLE 5.1
Relative residuals and $\|X_\eta - X_\eta^\top\|/\|X_\eta\|$.

η	10^{-4}	10^{-8}	10^{-12}	0
RRes_η	1.17×10^{-15}	1.51×10^{-15}	1.69×10^{-15}	1.59×10^{-15}
$\|X_\eta - X_\eta^\top\|/\|X_\eta\|$	9.88×10^{-15}	7.47×10^{-15}	1.12×10^{-14}	1.14×10^{-14}

TABLE 5.2
The eigenvalues of $X_{\eta,I}$.

η	The eigenvalues of $X_{\eta,I}$
10^{-4}	$33.938,\ 6.4240,\ 9.5171,\ 6.98 \times 10^{-4},\ 1.00 \times 10^{-4},\ 1.34 \times 10^{-4}$
10^{-8}	$33.937,\ 6.4232,\ 9.5134,\ 6.98 \times 10^{-8},\ 1.00 \times 10^{-8},\ 1.34 \times 10^{-8}$
10^{-12}	$33.937,\ 6.4232,\ 9.5134,\ 7.01 \times 10^{-12},\ 1.01 \times 10^{-12},\ 1.36 \times 10^{-12}$
0	$33.937,\ 6.4232,\ 9.5134,\ 2.54 \times 10^{-16},\ -1.94 \times 10^{-15},\ -3.71 \times 10^{-17}$

We now present some numerical comparison of SA and QZ for the computation of the weakly stabilizing solution of (3.3). To measure the accuracy of a computed solution X of (3.3) we use the relative residual RRes defined as in (5.1) with X_η and Q_η replaced by X and Q, respectively.

Example 5.2. As in [12], we consider a semi-infinite Hamiltonian operator for a heterostructured semiconductor of the form

$$H(\psi, \vec{x}) = -\nabla \frac{\hbar}{2\varepsilon(\vec{x})} \nabla\psi + V(\vec{x})\psi,\ \ \vec{x} \equiv \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \Omega, \tag{5.2}$$

where $\Omega \equiv \Omega_1 \cup \Omega_2$ with

$$\begin{cases} \Omega_1 = ([-9,-1] \cup [1,9]) \times (-\infty, 0], \\ \Omega_2 = [-1,1] \times (-\infty, 0], \end{cases}$$

$\hbar$ is the reduced Planck constant, ψ is the associated wave function, $\varepsilon(\vec{x})$ is the electron effective mass with

$$\varepsilon(\vec{x}) = \begin{cases} \varepsilon_1, & \vec{x} \in \Omega_1, \\ \varepsilon_2, & \vec{x} \in \Omega_2, \end{cases}$$

and $V(\vec{x}) = \omega x_1^2$ is the potential energy.

Let $T_r = \mathrm{Trid}(-1, 4, -1)$ be the tridiagonal matrix of dimension r. We use the classical five-point central finite difference method to discretize the Hamiltonian operator (5.2) on uniform grid points in Ω with mesh size h. Then the corresponding matrices A and Q in (3.3) are of the forms

$$A = -\left[\delta_1 I_\ell \oplus \left(\frac{\delta_1 + \delta_2}{2}\right) \oplus \delta_2 I_m \oplus \left(\frac{\delta_1 + \delta_2}{2}\right) \oplus \delta_1 I_\ell\right]$$

and $Q = \mathcal{E}I - B$ with

$$\begin{aligned} B = {} & \delta_1 T_\ell \oplus (2(\delta_1 + \delta_2)) \oplus \delta_2 T_m \oplus (2(\delta_1 + \delta_2)) \oplus \delta_1 T_\ell \\ & - \delta_1 \left(e_{\ell+1} e_\ell^\top + e_\ell e_{\ell+1}^\top\right) - \delta_2 \left(e_{\ell+2} e_{\ell+1}^\top + e_{\ell+1} e_{\ell+2}^\top\right) \\ & - \delta_2 \left(e_{t+1} e_t^\top + e_t e_{t+1}^\top\right) - \delta_1 \left(e_{t+2} e_{t+1}^\top + e_{t+1} e_{t+2}^\top\right) \\ & + \omega h^2 \mathrm{diag}\left((1-c)^2, (2-c)^2, \ldots, (n-c)^2\right), \end{aligned}$$

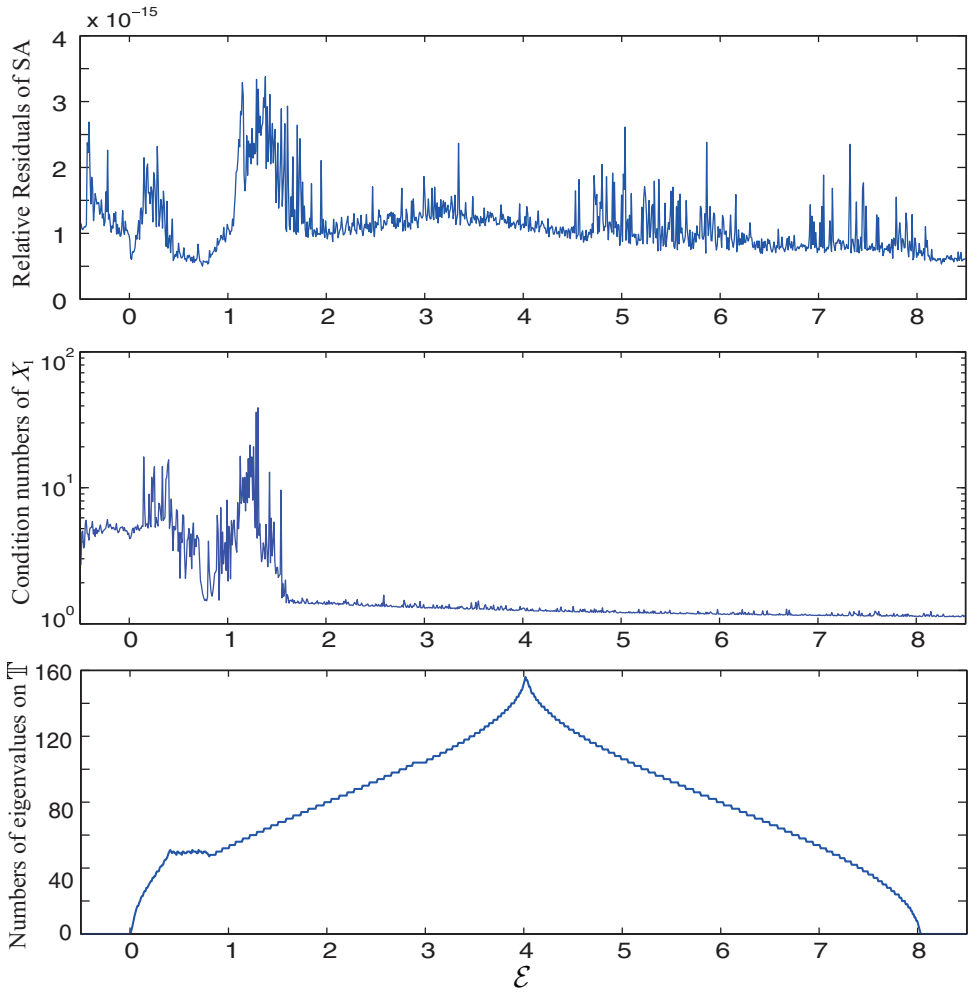

FIG. 1. *Relative residuals of SA, condition numbers of X_1, and numbers of eigenvalues of $(\mathcal{M}, \mathcal{L})$ on $\mathbb{T}$ that are used in the computation of X (so they are halves of the total numbers).*

where $\delta_i = \hbar/2h^2\varepsilon_i$ $(i = 1, 2)$, e_j denotes the jth column vector of the identity matrix, ℓ and m are the numbers of grid points on the x_1-axis in $(-9, -1)$ and $(-1, 1)$, respectively, $t = \ell + m + 1$, $n = 2\ell + m + 2$, $c = (n + 1)/2$, and $\oplus$ denotes the direct sum of two matrices.

In our test we take $\ell = 79$, $m = 19$, $\delta_1 = 1$, $\delta_2 = 0.1$, and $\omega = 5 \times 10^{-4}$, then the matrix size of A is $n = 2\ell + m + 2 = 179$. Let $\Delta = [-0.5, 8.5]$. We divide Δ into p subintervals using $p + 1$ equally spaced nodes $\mathcal{E}_i$, $i = 0, 1, \ldots, p$. We now choose $p = 1000$ and run SA and QZ for each $\mathcal{E}_i$. Recall that for both algorithms X is computed from $X = X_2 X_1^{-1}$ in the end. In Figure 1, we plot the relative residual and the condition number of X_1 (with each column of X_1 normalized, here and elsewhere) for SA. We also plot the number of eigenvalues of $(\mathcal{M}, \mathcal{L})$ on $\mathbb{T}$ that are used in the computation of X (so it is half of the total number). In Figure 2, we plot the relative residual and the condition number of X_1 for QZ. From the figures we can see that the

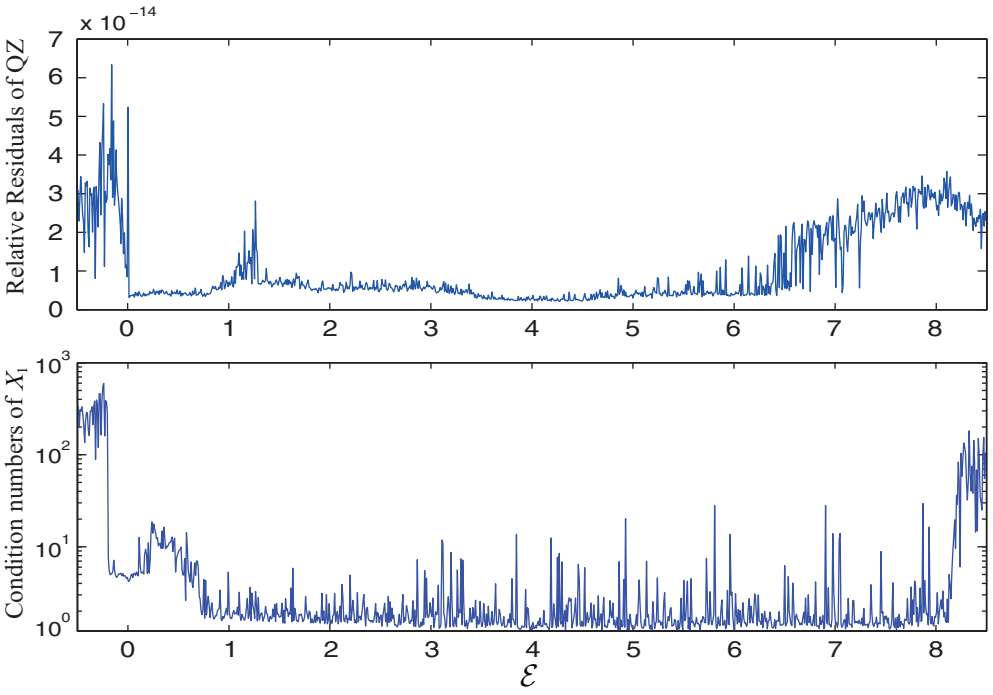

FIG. 2. *Relative residuals of QZ and condition numbers of* X_1.

condition numbers are not very large and the accuracy of the computed X is high for both methods, with that from SA slightly better.

In Example 5.2 the matrix A is nonsingular and so Step 5-1 in SA is never used. We now construct an example with a singular A.

Example 5.3. We construct 10×10 matrices A and Q for (3.3) as follows. $A = \text{rand}(10,10) * \text{diag}(a_1, \ldots, a_5, 0, 0, 0, 0, 0) * \text{rand}(10,10)$, where $a_i = 10 * \text{rand}(1)$, for $i = 1, \ldots, 5$, and $Q = \mathcal{E} I_{10} - B$ with $B = (B_0 + B_0^\top)/2$ and $B_0 = 10 * (\text{rand}(10,10) - 0.5 * \text{ones}(10,10))$. We run SA and QZ for $\mathcal{E} = 0.01i$ for $i = 0, \ldots, 1000$. The results are shown in Figures 3 and 4. We see that SA and QZ have roughly the same accuracy for this example.

6. Conclusions. In this paper we have further studied a nonlinear matrix equation arising in nano research. We have proved general convergence results on fixed-point iterations for (1.4). It is also shown that the required solution of (1.4) is the unique solution with a positive definite imaginary part. Our analysis has also shown that the convergence of these methods is usually very slow when the size of matrices in (1.4) is large. So the use of these simple methods is recommended only when the matrix size is small, which is the case when the equation is obtained from layer-based models. We have also studied (3.3) directly, which is the equation we obtain by letting $\eta = 0$ in (1.4). We have shown which half of the unimodular eigenvalues of $P(\lambda)$ in (3.6) should be included for computing the required weakly stabilizing solution of (3.3) using subspace methods, and we have also determined the rank of the imaginary part of this solution in terms of the number of unimodular eigenvalues of $P(\lambda)$. We have presented a structure-preserving algorithm for (3.3) that is nearly four times more efficient than the QZ algorithm. The SA and the QZ algorithm often provide

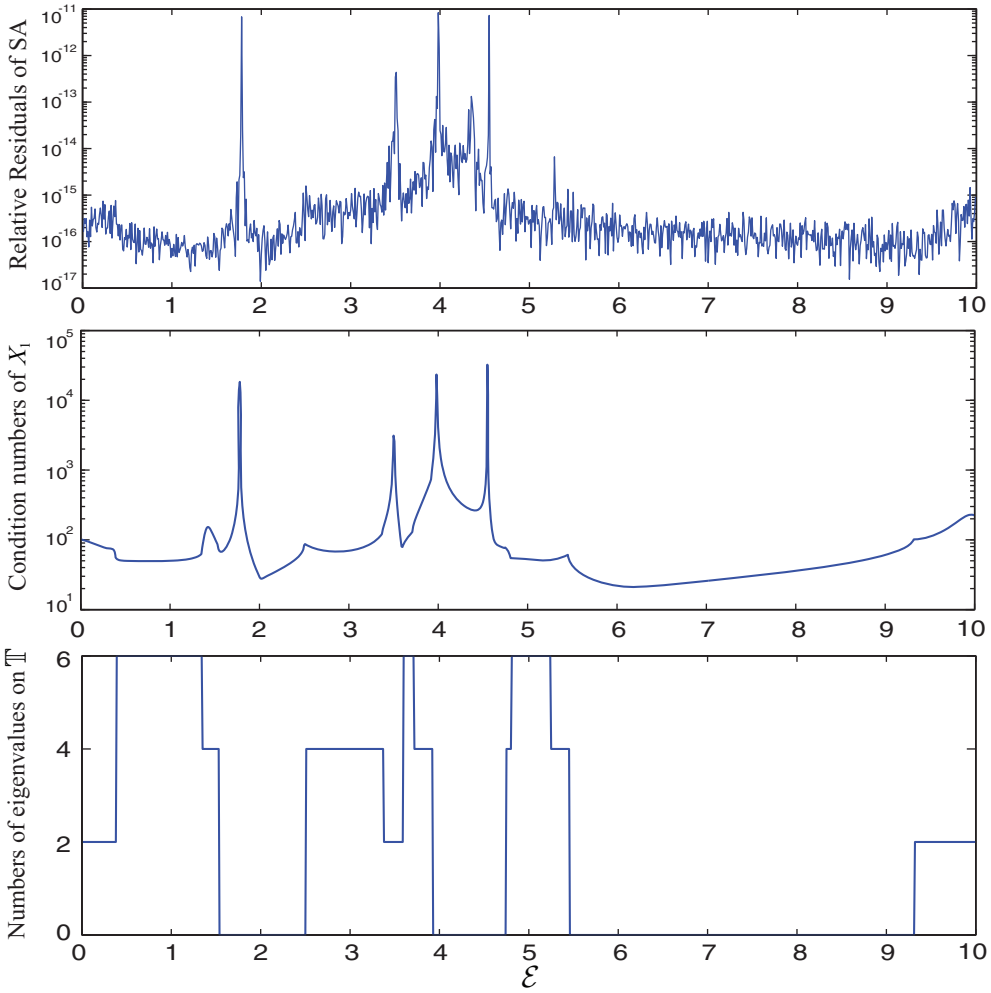

FIG. 3. *Relative residuals of SA, condition numbers of X_1, and numbers of eigenvalues of $(\mathcal{M}, \mathcal{L})$ on $\mathbb{T}$.*

very good accuracy. But the accuracy would suffer if the matrix X_1 used at the end of the algorithms happens to be very ill-conditioned. Newton's method cannot be used as a correction method since the Fréchet derivative at the solution is always singular when $P(\lambda)$ has unimodular eigenvalues. When SA and QZ fail we may use DA on (1.4) with a small $\eta > 0$ but may need to increase η if DA also runs into numerical difficulties. It is possible to improve the accuracy of an approximation from DA by using the FPI (2.1) with $c = \frac{1}{2}$, although at a high computational cost when the matrix size is large.

While SA is based on some theoretical analysis and is usually much more efficient than existing methods, we do not have a stability analysis of the algorithm. In fact, we are unable to rule out the possibility of breakdown, although we have explained that breakdown is very unlikely. Further work is needed to design an efficient algorithm with guaranteed stability for this special nonlinear matrix equation.

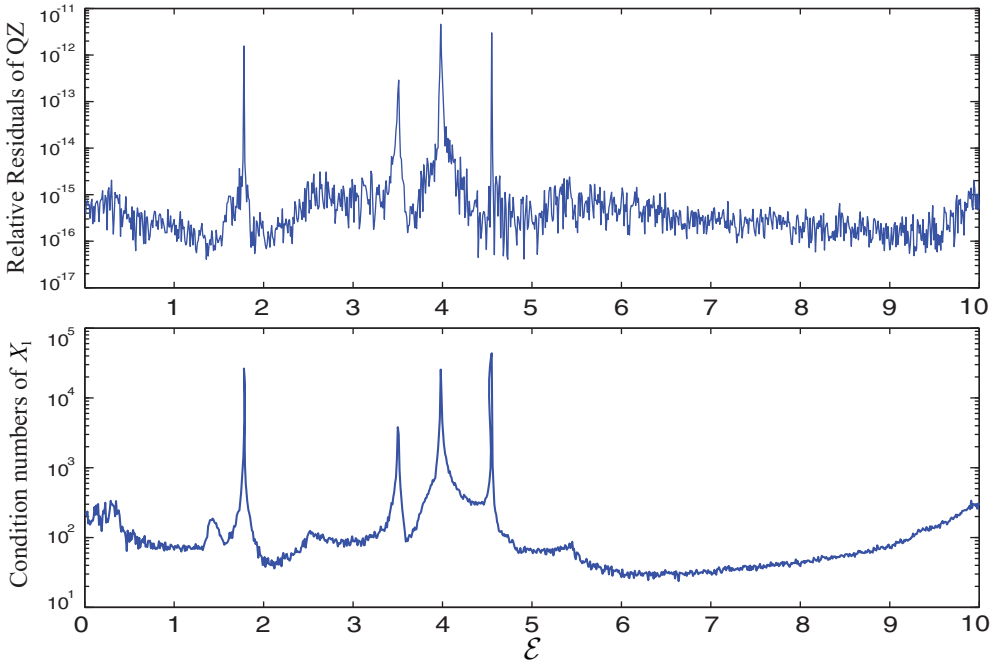

FIG. 4. *Relative residuals of QZ and condition numbers of* X_1.

Acknowledgment. The authors thank the referees for their helpful comments.

REFERENCES

[1] I. APPELBAUM, T. WANG, J. D. JOANNOPOULOS, AND V. NARAYANAMURTI, *Ballistic hot-electron transport in nanoscale semiconductor heterostructures: Exact self-energy of a three-dimensional periodic tight-binding Hamiltonian*, Phys. Rev. B, 69 (2004), 165301.

[2] A. G. BASKAKOV, *Dichotomy of the spectrum of non-self-adjoint operators*, Sibirsk. Mat. Zh., 32 (1991), pp. 24–30 (in Russian); Siberian Math. J., 32 (1992), pp. 370–375 (in English).

[3] C.-Y. CHIANG, E. K.-W. CHU, C.-H. GUO, T.-M. HUANG, W.-W. LIN, AND S.-F. XU, *Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 227–247.

[4] S. DATTA, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press, Cambridge, UK, 1995.

[5] S. DATTA, *Nanoscale device modeling: The Green's function method*, Superlattices and Microstructures, 28 (2000), pp. 253–278.

[6] C. J. EARLE AND R. S. HAMILTON, *A fixed point theorem for holomorphic mappings*, in Global Analysis, American Mathematical Society, Providence, RI, 1970, pp. 61–65.

[7] J. C. ENGWERDA, A. C. M. RAN, AND A. L. RIJKEBOER, *Necessary and sufficient conditions for the existence of a positive definite solution of the matrix equation* $X + A^*X^{-1}A = Q$, Linear Algebra Appl., 186 (1993), pp. 255–275.

[8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[9] C.-H. GUO, *On Newton's method and Halley's method for the principal pth root of a matrix*, Linear Algebra Appl., 432 (2010), pp. 1905–1922.

[10] C.-H. GUO, Y.-C. KUO, AND W.-W. LIN, *Complex symmetric stabilizing solution of the matrix equation* $X + A^TX^{-1}A = Q$, Linear Algebra Appl., 435 (2011), pp. 1187–1192.

[11] C.-H. GUO AND P. LANCASTER, *Iterative solution of two matrix equations*, Math. Comp., 68 (1999), pp. 1589–1603.

[12] C.-H. GUO AND W.-W. LIN, *The matrix equation* $X + A^\top X^{-1}A = Q$ *and its application in nano research*, SIAM J. Sci. Comput., 32 (2010), pp. 3020–3038.

[13] U. Haagerup and S. Thorbjørnsen, *A new application of random matrices:* $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ *is not a group*, Ann. Math., 162 (2005), pp. 711–775.
[14] L. A. Harris, *Fixed points of holomorphic mappings for domains in Banach spaces*, Abstr. Appl. Anal., 2003, pp. 261–274.
[15] T.-M. Huang, W.-W. Lin, and J. Qian, *Structure-preserving algorithms for palindromic quadratic eigenvalue problems arising from vibration of fast trains*, SIAM J. Matrix Anal. Appl., 30 (2009), pp. 1566–1592.
[16] B. Iannazzo and B. Meini, *Palindromic matrix polynomials, matrix functions and integral representations*, Linear Algebra Appl., 434 (2011), pp. 174–184.
[17] D. L. John and D. L. Pulfrey, *Green's function calculations for semi-infinite carbon nanotubes*, Physica Status Solidi B – Basic Solid State Physics, 243 (2006), pp. 442–448.
[18] A. Kletsov, Y. Dahnovsky, and J. V. Ortiz, *Surface Green's function calculations: A nonrecursive scheme with an infinite number of principal layers*, J. Chem. Phys., 126 (2007), 134105.
[19] W.-W. Lin, *A new method for computing the closed-loop eigenvalues of a discrete-time algebraic Riccati equation*, Linear Algebra Appl., 96 (1987), pp. 157–180.
[20] W.-W. Lin, V. Mehrmann, and H. Xu, *Canonical forms for Hamiltonian and symplectic matrices and pencils*, Linear Algebra Appl., 302/303 (1999), pp. 469–533.
[21] W.-W. Lin and S.-F. Xu, *Convergence analysis of structure-preserving doubling algorithms for Riccati-type matrix equations*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 26–39.
[22] B. Meini, *Efficient computation of the extreme solutions of* $X + A^*X^{-1}A = Q$ *and* $X - A^*X^{-1}A = Q$, Math. Comp., 71 (2002), pp. 1189–1204.
[23] R. V. Patel, *On computing the eigenvalues of a symplectic pencil*, Linear Algebra Appl., 188/189 (1993), pp. 591–611.
[24] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[25] J. Tomfohr and O. F. Sankey, *Theoretical analysis of electron transport through organic molecules*, J. Chem. Phys., 120 (2004), pp. 1542–1554.
[26] H. Wielandt, *On eigenvalues of sums of normal matrices*, Pacific J. Math., 5 (1955), pp. 633–638.
[27] X. Zhan and J. Xie, *On the matrix equation* $X + A^\top X^{-1}A = I$, Linear Algebra Appl., 247 (1996), pp. 337–345.