MATHEMATICS OF COMPUTATION
https://doi.org/10.1090/mcom/3398
Article electronically published on January 9, 2019

DOUBLING ALGORITHM FOR THE DISCRETIZED BETHE-SALPETER EIGENVALUE PROBLEM

ZHEN-CHEN GUO, ERIC KING-WAH CHU, AND WEN-WEI LIN

Abstract. The discretized Bethe-Salpeter eigenvalue problem arises in the Green's function evaluation in many body physics and quantum chemistry. Discretization leads to a matrix eigenvalue problem for $H \in \mathbb{C}^{2n\times 2n}$ with a Hamiltonian-like structure. After an appropriate transformation of H to a standard symplectic form, the structure-preserving doubling algorithm, originally for algebraic Riccati equations, is extended for the discretized Bethe-Salpeter eigenvalue problem. Potential breakdowns of the algorithm, due to the ill condition or singularity of certain matrices, can be avoided with a double-Cayley transform or a three-recursion remedy. A detailed convergence analysis is conducted for the proposed algorithm, especially on the benign effects of the double-Cayley transform. Numerical results are presented to demonstrate the efficiency and the structure-preserving nature of the algorithm.

1. Introduction

The Bethe-Salpeter equation (BSE) [29] arises in the Green's function evaluation in many body physics, which is the state-of-the-art model to describe electronic excitation and molecule absorption [6,13–15,20–28,32,33]. In the quantum chemistry and material science communities, the optical absorption spectrum of the BSE is an important and powerful tool for the characterization of different materials. In particular, the comparison of the computed and measured spectra helps to interpret experimental data and validate corresponding theories and models. It is generally known that a good agreement between the theory and the experimental data can only be achieved by taking into account the interacting electron-hole pairs or *excitons*. This is the case for the BSE which is derived from the coupling of the electrons and their corresponding holes.

After discretization, the BSE becomes the Bethe-Salpeter eigenvalue problem (BS-EVP):

$$Hx \equiv \begin{bmatrix} A & B \\ -\overline{B} & -\overline{A} \end{bmatrix} x = \lambda x, \tag{1.1}$$

for $x \neq 0$, where $A, B \in \mathbb{C}^{n\times n}$ satisfy $A^{\mathsf{H}} = A$, $B^{\mathsf{T}} = B$. Here $(\cdot)^{\mathsf{H}}$ and $(\cdot)^{\mathsf{T}}$ denote the conjugate transpose and the transpose of matrices, respectively. It can be shown [4] that any eigenvalue λ comes in quadruplets $\{\pm\lambda, \pm\overline{\lambda}\}$ (except for the

Received by the editor November 20, 2017, and, in revised form, May 6, 2018, July 25, 2018, and August 21, 2018.

2010 *Mathematics Subject Classification.* Primary 15A18, 65F15.

Key words and phrases. Bethe-Salpeter eigenvalue problem, Cayley transform, doubling algorithm.

degenerate cases when λ is purely real or imaginary, or zero). Further details on the BS-EVP can be found in [3, 5, 30] and the references therein.

In principle, all possible excitation energies and absorption spectra are sought although some excitations are more probable than others. The associated likelihood is measured by the spectral density or the density of states of H, defined as the number of eigenvalues per unit energy interval:

$$\phi(\omega) = \frac{1}{2n} \sum_{j=1}^{2n} \delta(\omega - \lambda_j),$$

where δ is the Dirac-delta function and $\lambda_j \in \lambda(H)$, the spectrum of H. Also of interest is the optical absorption spectrum:

$$\epsilon^{+}(\omega) = \sum_{j=1}^{n} \frac{(d_r^{\mathsf{H}} x_j)(y_j^{\mathsf{H}} d_l)}{y_j^{\mathsf{H}} x_j} \delta(\omega - \lambda_j),$$

where x_j and y_j are, respectively, the right- and left-eigenvectors corresponding to $\lambda_j > 0$, and d_r and d_l are the dipole vectors. Evidently, to estimate these quantities, we require *all* the eigenvalues λ_j and the associated eigenvectors x_j and y_j. To complicate computations further, A and B are often high in dimensions (for systems with many occupied and unoccupied states) and generally dense.

In spite of the significance of the BS-EVP (1.1), only a few publications exist on its numerical solution, all under *additional* assumptions. Some remarkable discoveries have been made in [3, 5, 30] under the condition that ΓH is positive definite with $\Gamma = \mathrm{diag}(I_n, -I_n)$, which makes all eigenvalues of H real. Few general and efficient methods have been proposed to solve the BS-EVP (1.1), where some complex eigenvalues can exist for H. Low-rank or tensor approximations [3, 5] have been applied to handle the high computational demand but these techniques require additional structures on H. Based on the equivalence of the BS-EVP and a real Hamiltonian eigenvalue problem, Shao et al. [30] put forward an efficient parallel approach to compute the eigenpairs corresponding to all the positive eigenvalues. Remarkable contributions have also been made for the numerical solution of the related linear response eigenvalue problem [1, 2].

Contributions. We solve the *general* BS-EVP (1.1), without assuming ΓH being positive definite. We propose a doubling algorithm (DA) for the BS-EVP in two recursions. To deal with potential but generically rare breakdowns, we design the double-Cayley transform (DCT) and a three-recursion remedy. The DCT reverses at worst two steps of the DA if there exist some complex eigenvalues and not at all if all eigenvalues are real. In the rare occasions that the DCT fails, the more expensive three-recursion remedy can be applied, without changing the convergence radius. Our DA preserves the special structure of the eigenpairs.

Organization. Some preliminaries are presented in Section 2 and our method is developed in Section 3. We present some illustrative numerical results in Section 4 before the conclusions in Section 5. Appendix A contains three technical lemmas and Appendix B contains the proofs for some results in Section 3.

2. Preliminaries

We denote the column space, the null space, the spectrum, and the set of singular values by $\mathcal{R}(\cdot)$, $\mathcal{N}(\cdot)$, $\lambda(\cdot)$, and $\sigma(\cdot)$, respectively. By $M \oplus N$ or $\operatorname{diag}(M, N)$, we denote $\begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}$; similarly, we define $\bigoplus_j M_j$. The MATLAB expression $M(k:l, s:t)$ denotes the submatrix of M containing elements in rows k to l and columns s to t. The ith column of the identity matrix I is e_i, $J \equiv \begin{bmatrix} & I_n \\ -I_n & \end{bmatrix}$, $\Gamma \equiv \begin{bmatrix} I_n & \\ & -I_n \end{bmatrix}$, and $\Pi \equiv \begin{bmatrix} & I_n \\ I_n & \end{bmatrix}$.

Definition 1. The matrix pair $(M,\, L)$ with $M, L \in \mathbb{C}^{2n\times 2n}$ is a symplectic pair if and only if $MJM^{\mathsf{T}} = LJL^{\mathsf{T}}$.

Definition 2. The matrix pair $(M,\, L)$ is in the first standard symplectic form (SSF-1) if and only if $M = \begin{bmatrix} E & 0 \\ F & I_n \end{bmatrix} \in \mathbb{C}^{2n\times 2n}$ and $L = \begin{bmatrix} I_n & K \\ 0 & E^{\mathsf{T}} \end{bmatrix} \in \mathbb{C}^{2n\times 2n}$, with $E, F \equiv F^{\mathsf{T}}, K \equiv K^{\mathsf{T}} \in \mathbb{C}^{n\times n}$.

Definition 3. Let $M, L \in \mathbb{C}^{2n\times 2n}$ and denote $\mathcal{N}(M, L)$

$$\equiv \left\{ [M_*,\, L_*] : M_*, L_* \in \mathbb{C}^{2n\times 2n},\ \operatorname{rank}([M_*,\, L_*]) = 2n,\ [M_*,\, L_*][L^{\mathsf{T}},\, -M^{\mathsf{T}}]^{\mathsf{T}} = 0 \right\},$$

which is nonempty. The action $(M,\, L) \longrightarrow (\widetilde{M},\, \widetilde{L}) = (M_*M,\, L_*L)$ is called a *doubling transformation* of (M, L) for some $[M_*,\, L_*] \in \mathcal{N}(M, L)$.

Next we consider the properties of the doubling transformation.

Lemma 4 ([19, Theorem 2.1])**.** *Let $(\widetilde{M},\, \widetilde{L})$ be the result of a doubling transformation of $(M,\, L)$, where $M, L, \widetilde{M}, \widetilde{L} \in \mathbb{C}^{2n\times 2n}$. We have*

1. *$(\widetilde{M},\, \widetilde{L})$ is a symplectic pair provided that $(M,\, L)$ is one; and*
2. *if $MU = LUR$ and $MVS = LV$ for some $U, V \in \mathbb{C}^{2n\times l}$ and $R, S \in \mathbb{C}^{l\times l}$, then $\widetilde{M}U = \widetilde{L}UR^2$ and $\widetilde{M}VS^2 = \widetilde{L}V$.*

In other words, doubling transformations preserve symplecticity and deflating subspaces as well as square eigenvalues of matrix pairs.

Lemma 5. *It holds that $H\Pi = -\Pi\overline{H}$ and $\Gamma H \Gamma = H^{\mathsf{H}}$.*

Proof. It can be verified directly. □

Lemma 6. *Assume that $HZ = ZS$ with $Z \in \mathbb{C}^{2n\times l}$ and $S \in \mathbb{C}^{l\times l}$; then we have $H(\Pi\overline{Z}) = (\Pi\overline{Z})(-\overline{S})$ and $(Z^{\mathsf{H}}\Gamma)H = S^{\mathsf{H}}(Z^{\mathsf{H}}\Gamma)$.*

Proof. The results directly follow from Lemma 5. □

If S in Lemma 6 possesses the spectrum $\lambda(S) = \{\lambda, \ldots, \lambda\}$ (repeated l times), Lemmas 5 and 6 imply that $-\lambda$, $\overline{\lambda}$, and $-\overline{\lambda}$ are also the eigenvalues of H with the same algebraic and geometric multiplicities. If $HX_j = X_jS_j$ with $X_j \in \mathbb{C}^{2n\times l_j}$ and $S_j \in \mathbb{C}^{l_j\times l_j}$ $(j = 1, 2)$, Lemma 6 further implies that $(X_2^{\mathsf{H}}\Gamma X_1)S_1 = X_2^{\mathsf{H}}\Gamma H X_1 = S_2^{\mathsf{H}}(X_2^{\mathsf{H}}\Gamma X_1)$ and $(X_2^{\mathsf{T}}\Pi\Gamma)X_1S_1 = (X_2^{\mathsf{T}}\Pi\Gamma)HX_1 = (-S_2^{\mathsf{T}})(X_2^{\mathsf{T}}\Pi\Gamma X_1)$, or equivalently

$$(X_2^{\mathsf{H}}\Gamma X_1)S_1 - S_2^{\mathsf{H}}(X_2^{\mathsf{H}}\Gamma X_1) = 0 = (X_2^{\mathsf{T}}\Pi\Gamma X_1)S_1 + S_2^{\mathsf{T}}(X_2^{\mathsf{T}}\Pi\Gamma X_1).$$

Apparently, when $\lambda(S_1)\cap\lambda(\overline{S}_2)=\emptyset$, we have $X_2^{\mathsf{H}}\Gamma X_1=0$; when $\lambda(S_1)\cap\lambda(-S_2)=\emptyset$, we have $X_2^{\mathsf{T}}\Pi\ \Gamma X_1=0$. By Lemmas 5 and 6, we can then deduce the eigendecomposition result of H for the convergence proof.

First we assume that there is no purely imaginary nor zero eigenvalues for H. Let

$$\begin{aligned}\lambda(H)=\{&\underbrace{\lambda_1,\ldots,\lambda_1}_{l_1},\underbrace{\overline{\lambda}_1,\ldots,\overline{\lambda}_1}_{l_1},\underbrace{-\overline{\lambda}_1,\ldots,-\overline{\lambda}_1}_{l_1},\underbrace{-\lambda_1,\ldots,-\lambda_1}_{l_1},\ldots,\\ &\underbrace{\lambda_s,\ldots,\lambda_s}_{l_s},\underbrace{\overline{\lambda}_s,\ldots,\overline{\lambda}_s}_{l_s},\underbrace{-\overline{\lambda}_s,\ldots,-\overline{\lambda}_s}_{l_s},\underbrace{-\lambda_s,\ldots,-\lambda_s}_{l_s},\\ &\underbrace{\lambda_{s+1},\ldots,\lambda_{s+1}}_{l_{s+1}},\underbrace{-\lambda_{s+1},\ldots,-\lambda_{s+1}}_{l_{s+1}},\ldots,\underbrace{\lambda_t,\ldots,\lambda_t}_{l_t},\underbrace{-\lambda_t,\ldots,-\lambda_t}_{l_t}\},\end{aligned}$$

where $\lambda_j\in\mathbb{C}$ ($\lambda_j\neq\lambda_k$ for $j\neq k$) with (i) $\Re(\lambda_j)\Im(\lambda_j)\neq 0$ and $\Re(\lambda_j)<0$ for $j=1,\ldots,s$, and (ii) $\Im(\lambda_j)=0$ and $\lambda_j<0$ for $j=s+1,\ldots,t$. We have the following.

Lemma 7. *Suppose that no purely imaginary nor zero eigenvalues exist for H. Then there exist*

$$\begin{aligned}X&=[X_1,Y_1,\cdots,X_s,Y_s;\ X_{s+1},\cdots,X_t]\in\mathbb{C}^{2n\times n},\\ S&=\operatorname{diag}(S_1,R_1,\ldots,S_s,R_s;\ S_{s+1},\ldots,S_t)\in\mathbb{C}^{n\times n},\end{aligned}$$

with $X_j\in\mathbb{C}^{2n\times l_j}$, $S_j\in\mathbb{C}^{l_j\times l_j}$, $\lambda(S_j)=\{\lambda_j,\ldots,\lambda_j\}$ $(j=1,\ldots,t)$, $Y_j\in\mathbb{C}^{2n\times l_j}$, $R_j\in\mathbb{C}^{l_j\times l_j}$ and $\lambda(R_j)=\{\overline{\lambda}_j,\ldots,\overline{\lambda}_j\}$ $(j=1,\ldots,s)$, where X is of full column rank, such that

$$H[X,\,\Pi\overline{X}]=[X,\,\Pi\overline{X}]\operatorname{diag}(S,\,-\overline{S}),\qquad [X,\,\Pi\overline{X}]^{\mathsf{H}}\Gamma[X,\,\Pi\overline{X}]=\operatorname{diag}(D,\,-\overline{D}),$$

where $D=\operatorname{diag}(D_1,\ldots,D_s;\ D_{s+1},\ldots,D_t)$,

$$\begin{aligned}D_j&=\begin{bmatrix}0 & X_j^{\mathsf{H}}\Gamma Y_j\\ Y_j^{\mathsf{H}}\Gamma X_j & 0\end{bmatrix}\in\mathbb{C}^{2l_j\times 2l_j}\quad (j=1,\ldots,s),\\ D_j&=X_j^{\mathsf{H}}\Gamma X_j\in\mathbb{C}^{l_j\times l_j}\quad (j=s+1,\ldots,t).\end{aligned}$$

Next consider the case when there exist some purely imaginary eigenvalues for H. We further assume that the partial multiplicities (the sizes of the Jordan blocks) of H associated with the purely imaginary eigenvalues are all even. Let $\mathrm{i}\omega_1,\cdots,\mathrm{i}\omega_q$ be the different purely imaginary eigenvalues with Jordan blocks $J_{2p_{r,j}}(\mathrm{i}\omega_j)\in\mathbb{C}^{2p_{r,j}\times 2p_{r,j}}$ for $r=1,\cdots,l_j$ and $j=1,\cdots,q$. Then there exist $W_{r,j},Z_{r,j}\in\mathbb{C}^{2n\times p_{r,j}}$ such that

$$\begin{aligned}&H\left[W_{1,1},Z_{1,1};\cdots;W_{l_1,1},Z_{l_1,1}\middle|\cdots\middle|W_{1,q},Z_{1,q};\cdots;W_{l_q,q},Z_{l_q,q}\right]\\ &=\left[W_{1,1},Z_{1,1};\cdots;W_{l_1,1},Z_{l_1,1}\middle|\cdots\middle|W_{1,q},Z_{1,q};\cdots;W_{l_q,q},Z_{l_q,q}\right]\cdot\left[\bigoplus_{j=1}^{q}\bigoplus_{r=1}^{l_j}J_{2p_{r,j}}(\mathrm{i}\omega_j)\right].\end{aligned}$$

With $X\in\mathbb{C}^{2n\times n_1}$ and $S\in\mathbb{C}^{n_1\times n_1}$ and by Lemma 7, we obtain

$$H\left[X,W_\omega,\Pi\overline{X},Z_\omega\right]=\left[X,W_\omega,\Pi\overline{X},Z_\omega\right]\widetilde{S},\tag{2.1}$$

where $n_1 + \sum_{j=1}^{q}\sum_{r=1}^{l_j} p_{r,j} = n$, and

$$\begin{aligned} W_\omega &= \left[W_{1,1}, \cdots, W_{l_1,1}; \cdots; W_{1,q}, \cdots, W_{l_q,q}\right], \\ Z_\omega &= \left[Z_{1,1}, \cdots, Z_{l_1,1}; \cdots; Z_{1,q}, \cdots, Z_{l_q,q}\right], \\ J_\omega &= \bigoplus_{j=1}^{q}\bigoplus_{r=1}^{l_j} J_{p_{r,j}}(\mathrm{i}\omega_j), \qquad \Omega_\omega = \bigoplus_{j=1}^{q}\bigoplus_{r=1}^{l_j} e_{p_{r,j}} e_1^{\mathsf{T}}, \end{aligned}$$

$$J_{2p_{r,j}}(\mathrm{i}\omega_j) \equiv \begin{bmatrix} J_{p_{r,j}}(\mathrm{i}\omega_j) & e_{p_{r,j}} e_1^{\mathsf{T}} \\ 0 & J_{p_{r,j}}(\mathrm{i}\omega_j) \end{bmatrix}, \quad \widetilde{S} \equiv \begin{bmatrix} S & & & \\ & J_\omega & & \Omega_\omega \\ & & -\overline{S} & \\ & & & J_\omega \end{bmatrix}.$$

3. Doubling algorithm

We now construct the DA for the BS-EVP, similar to the structure-preserving doubling algorithm (SDA) for algebraic Riccati equations in [7, 8, 17, 18].

3.1. **Initial symplectic pencil.**

We transform H to a symplectic pair (M, L) in the SSF-1 à la Cayley.

Lemma 8. *For $\alpha \in \mathbb{R}$, the matrix pair $(H + \alpha I_{2n}, H - \alpha I_{2n})$ is symplectic.*

Proof. The result can be deduced from $(HJ)^{\mathsf{T}} = HJ$. □

Theorem 9. *Select $\alpha \in \mathbb{R}$ such that both $\alpha I_n - A$ and $R \equiv I_n - (\alpha I_n - \overline{A})^{-1}\overline{B}(\alpha I_n - A)^{-1}B$ are nonsingular. There exists a nonsingular matrix $G \in \mathbb{C}^{2n\times 2n}$ such that $[G(H + \alpha I_n), G(H - \alpha I_n)]$ is a symplectic pair in* SSF-1*, with*

$$M_\alpha \triangleq G(H + \alpha I_n) = \begin{bmatrix} E_\alpha & 0 \\ F_\alpha & I_n \end{bmatrix}, \quad L_\alpha \triangleq G(H - \alpha I_n) = \begin{bmatrix} I_n & \overline{F}_\alpha \\ 0 & \overline{E}_\alpha \end{bmatrix}, \tag{3.1}$$

where $E_\alpha, F_\alpha \in \mathbb{C}^{n\times n}$ satisfy $E_\alpha^{\mathsf{H}} = E_\alpha$ and $F_\alpha^{\mathsf{T}} = F_\alpha$.

Proof. Let $H_\pm \equiv H \pm \alpha I_{2n}$, $A_\pm \equiv A \pm \alpha I_n$,

$$G_1 = \begin{bmatrix} A_-^{-1} & 0 \\ \overline{B}A_-^{-1} & I_n \end{bmatrix}, \quad G_2 = \begin{bmatrix} I_n & A_-^{-1}BR^{-1}\overline{A}_-^{-1} \\ 0 & -R^{-1}\overline{A}_-^{-1} \end{bmatrix},$$

and $G = G_2 G_1$. We obtain

$$\begin{aligned} G_1 H_+ &= \begin{bmatrix} A_-^{-1}A_+ & A_-^{-1}B \\ 2\alpha\overline{B}A_-^{-1} & -\overline{A}_- R \end{bmatrix}, \quad G_2 G_1 H_+ = \begin{bmatrix} E_\alpha & 0 \\ F_\alpha & I_n \end{bmatrix}, \\ G_1 H_- &= \begin{bmatrix} I_n & A_-^{-1}B \\ 0 & -\overline{A}_- R - 2\alpha I_n \end{bmatrix}, \quad G_2 G_1 H_- = \begin{bmatrix} I_n & \overline{F}_\alpha \\ 0 & \overline{E}_\alpha \end{bmatrix}, \end{aligned}$$

with

$$E_\alpha = I_n + 2\alpha\overline{R}^{-1}A_-^{-1}, \quad F_\alpha = -2\alpha\overline{A}_-^{-1}\overline{B}\,\overline{R}^{-1}A_-^{-1}. \tag{3.2}$$

Furthermore, since $A^{\mathsf{H}} = A$ and $B^{\mathsf{T}} = B$, we have

$$\begin{aligned} E_\alpha^{\mathsf{H}} &= I_n + 2\alpha A_-^{-1}R^{-\,\mathsf{T}} = I_n + 2\alpha(A_-^{-1} - B\overline{A}_-^{-1}\overline{B})^{-1} = E_\alpha, \\ F_\alpha^{\mathsf{T}} &= -2\alpha\overline{A}_-^{-1}(I_n - \overline{B}A_-^{-1}B\overline{A}_-^{-1})^{-1}\overline{B}A_-^{-1} = F_\alpha, \end{aligned}$$

i.e., E_α and F_α are Hermitian and complex symmetric, respectively. Lastly, we have

$$(GH_\pm)J(GH_\pm)^\mathsf{T} = \begin{bmatrix} & E_\alpha \\ -\overline{E}_\alpha & \end{bmatrix},$$

implying that $[G(H+\alpha I_n),\, G(H-\alpha I_n)]$ is a symplectic pair in SSF-1. □

The following lemma summarizes the eigenstructure of $(M_\alpha,\, L_\alpha)$ in relation to that of H, neglecting the simple proof.

Lemma 10. *Let*

$$H[X_1^\mathsf{T},\, X_2^\mathsf{T}]^\mathsf{T} = [X_1^\mathsf{T},\, X_2^\mathsf{T}]^\mathsf{T} S \tag{3.3}$$

for some $X_1, X_2 \in \mathbb{C}^{n\times l}$, $S \in \mathbb{C}^{l\times l}$ *and* $\alpha \notin \lambda(H)$. *Then we have*

$$M_\alpha[X_1^\mathsf{T},\, X_2^\mathsf{T}]^\mathsf{T} = L_\alpha[X_1^\mathsf{T},\, X_2^\mathsf{T}]^\mathsf{T} S_\alpha$$

with $S_\alpha \equiv (S-\alpha I_l)^{-1}(S+\alpha I_l)$, *where* $S-\alpha I_l$ *is nonsingular.*

Intrinsically, the DA and the DCT proposed below require $I_n - F_\alpha \overline{F}_\alpha$ and E_α to be nonsingular. Lemma 11 and Theorems 12 and 13 below indicate that a small α, relative to $\|H\|_F$, could achieve such a goal. Moreover, for $\lambda \in \lambda(H)$, we have $(\lambda+\alpha)/(\lambda-\alpha) \in \lambda(S_\alpha)$. For the efficiency of the DA, we desire a small $|(\lambda+\alpha)/(\lambda-\alpha)|$ for $\Re(\lambda) < 0$ and $\alpha > 0$. Hence when $|\alpha| > \rho(H)$ (the spectral radius of H), we desire $|\alpha|$ to be minimized.

Lemma 11. *Let* $\alpha > \|H\|_F$. *Then* $\alpha I_n - A$ *is positive definite and* $R \equiv I_n - (\alpha I_n - \overline{A})^{-1}\overline{B}(\alpha I_n - A)^{-1}B$ *is nonsingular, with* $\|R^{-1}\|_2 \le \left[1 - \|(\alpha I_n - A)^{-1}\|_2^2\|B\|_2^2\right]^{-1}$.

Proof. When $\|A\|_F < \|H\|_F < \alpha$, $\alpha I_n - A$ is positive definite Hermitian. Since $\alpha > \|H\|_F = \sqrt{2(\|A\|_F^2 + \|B\|_F^2)} \ge \|A\|_F + \|B\|_F$, we have $(\alpha - \omega_1)^{-1} \le (\alpha - \|A\|_F)^{-1} < \|B\|_F^{-1}$ with ω_1 being the largest eigenvalue of A. In addition, with $\|(\alpha I_n - A)^{-1}\|_2 = (\alpha-\omega_1)^{-1}$, we have $\|(\alpha I_n - A)^{-1}B\|_2 \le \|(\alpha I_n - A)^{-1}\|_2\|B\|_2 = (\alpha-\omega_1)^{-1}\|B\|_2 < 1$. This implies $\|(\alpha I_n - \overline{A})^{-1}\overline{B}(\alpha I_n - A)^{-1}B\|_2 \le \|(\alpha I_n - A)^{-1}\|_2^2\|B\|_2^2 < 1$ and our results. □

Theorem 12. *As defined in* (3.2), E_α *is nonsingular when* $\alpha > \|H\|_F$.

Proof. Denote the largest and smallest eigenvalues of A by ω_1 and ω_n, respectively. With $\alpha > \|H\|_F$, we have $\|\alpha I_n - A\|_2 = \alpha - \omega_n$ and $\|(\alpha I_n - \overline{A})^{-1}\|_2 = (\alpha-\omega_1)^{-1}$, yielding $\|(\alpha I_n - A) - B(\alpha I_n - \overline{A})^{-1}\overline{B}\|_2 \le (\alpha - \omega_n) + (\alpha-\omega_1)^{-1}\|B\|_2^2$. We also have

$$\begin{aligned}&(\alpha-\omega_n)(\alpha-\omega_1) + \|B\|_2^2 - 2\alpha(\alpha-\omega_1)\\ &= -\left(\alpha + \frac{\omega_n-\omega_1}{2}\right)^2 + \frac{(\omega_1+\omega_n)^2}{4} + \|B\|_2^2 < -\left(\alpha + \frac{\omega_n - \omega_1}{2}\right)^2 + \frac{\alpha^2 - \|A\|_F^2}{2},\end{aligned}$$

as $\alpha^2 > \|H\|_F^2 = 2(\|B\|_F^2 + \|A\|_F^2)$. From the fact that $2\|A\|_F^2 \ge 2(\omega_1^2 + \omega_n^2) \ge (\omega_n - \omega_1)^2$, we obtain

$$\frac{\alpha^2 - \|A\|_F^2}{2} - \left(\alpha + \frac{\omega_n-\omega_1}{2}\right)^2 \le -\frac{(\alpha + \omega_n - \omega_1)^2}{2} < 0.$$

This implies $(\alpha-\omega_n)(\alpha-\omega_1) + \|B\|_2^2 - 2\alpha(\alpha-\omega_1) < 0$. We deduce $\|(\alpha I_n - A) - B(\alpha I_n - \overline{A})^{-1}\overline{B}\|_2 < 2\alpha$, thus $2\alpha \notin \lambda\{(\alpha I_n - A) - B(\alpha I_n - \overline{A})^{-1}\overline{B}\}$. Therefore, $E_\alpha = I_n - 2\alpha\left[(\alpha I_n - A) - B(\alpha I_n - \overline{A})^{-1}\overline{B}\right]^{-1}$ is nonsingular. □

Complementing Theorem 12, we have $\lambda(E_\alpha)$ lying outside $[0,2]$ when $\alpha > \|H\|_F$ because the moduli of all eigenvalues of $\left[(\alpha I_n - A) - B(\alpha I_n - \overline{A})^{-1}\overline{B}\right]^{-1}$ are greater than $(2\alpha)^{-1}$.

Theorem 13. *Assume that $\alpha > \|H\|_F + \sqrt{2\|B\|_2\|H\|_F}$. Then $\|F_\alpha\|_2 < 1$ with F_α defined in* (3.2).

Proof. Let ω_1 be the largest eigenvalue of A. Then it holds that

$$\|(\alpha I_n - A)^{-1}\|_2 = (\alpha - \omega_1)^{-1}, \quad \|F_\alpha\|_2 \le \frac{2\alpha\|B\|_2}{(\alpha-\omega_1)^2 - \|B\|_2^2}.$$

Next we show that $\|B\|_2/[(\alpha-\omega_1)^2 - \|B\|_2^2]$, in the right-hand side of the inequality above, is bounded strictly from above by $(2\alpha)^{-1}$ when $\alpha > \|H\|_F + \sqrt{2\|B\|_2\|H\|_F}$, or equivalently

$$(\alpha-\omega_1)^2 - 2\alpha\|B\|_2 - \|B\|_2^2 > 0. \tag{3.4}$$

If $\|B\|_2 + \omega_1 \le 0$, (3.4) is apparently valid. When $\|B\|_2 + \omega_1 > 0$ and considering the left-hand side of (3.4) as a quadratic function in α, (3.4) holds if and only if $\alpha > \|B\|_2 + \omega_1 + \sqrt{2\|B\|_2(\|B\|_2 + \omega_1)}$. Since $\|H\|_F \ge \|A\|_F + \|B\|_F \ge \|B\|_2 + \omega_1$, our result follows. □

Theorem 13 demonstrates that relative to $\|H\|_F$, a small positive number will be a good candidate for the initial α, such as $\alpha = (1+\sqrt{2})\|H\|_F$. Additionally, when the condition in Theorem 13 is satisfied, E_α and $I_n - F_\alpha\overline{F}_\alpha$ are nonsingular.

Although Theorems 12 and 13 show that a small α (relative to $\|H\|_F$) is sufficient for E_α and $I_n - F_\alpha\overline{F}_\alpha$ to be nonsingular, the minimization of $|(\lambda+\alpha)/(\lambda-\alpha)|$ for an optimal α deserves further consideration, for the fast convergence of the DA. For the optimal α, [11] proposes some remarkable techniques for the suboptimal solution $\alpha_{opt} := \text{argmin}_{\alpha>0}\max_{\Re(\lambda)<0}\left|\frac{\lambda+\alpha}{\lambda-\alpha}\right|$. With some prior knowledge (in $\mathcal{D}$ below) of the eigenvalues of H, [11] essentially solves the following optimization problem:

$$\alpha_{sopt} := \text{argmin}_{\alpha>0}\max_{\zeta\in\mathcal{D}}\left|\frac{\zeta+\alpha}{\zeta-\alpha}\right|, \qquad \{\lambda\in\lambda(H) : \Re(\lambda)<0\} \subset \mathcal{D} \subset \mathbb{C}_-.$$

With $\mathcal{D}$ being an interval, a disk, an ellipse, or a rectangle, [11, Theorem 2.1] considers the suboptimal solution α_{sopt}. The technique can be applied to (3.1) for a suboptimal α when the distance between $\{\lambda \in \lambda(H) : \Re(\lambda) < 0\}$ and the imaginary axis is known.

From now on, we will always assume $\alpha > 0$ such that $\alpha I_{2n} - H$, $\alpha I_n - A$, $I_n - (\alpha I_n - \overline{A})^{-1}\overline{B}(\alpha I_n - A)^{-1}B$, and E_α are nonsingular and also assume that $1 \notin \sigma(F_\alpha)$ (before the discussion in Section 3.3).

3.2. Algorithm.

We now construct a new symplectic pair by applying the doubling action to a given symplectic pair (M, L) in SSF-1 in (3.1); i.e., for $E^{\mathsf{H}} = E \in \mathbb{C}^{n\times n}$, $F^{\mathsf{T}} = F \in \mathbb{C}^{n\times n}$, we have

$$M = \begin{bmatrix} E & 0 \\ F & I_n \end{bmatrix}, \qquad L = \begin{bmatrix} I_n & \overline{F} \\ 0 & \overline{E} \end{bmatrix}. \tag{3.5}$$

Theorem 14. *For M, L in (3.5) with $1 \notin \sigma(F)$, there exists $[M_*,\, L_*] \in \mathcal{N}(M, L)$ such that $(\widetilde{M},\, \widetilde{L}) = (M_* M,\, L_* L)$, from the doubling transformation of $(M,\, L)$, is symplectic. Furthermore, with $\widetilde{E}^{\mathsf{H}} = \widetilde{E}, \widetilde{F}^{\mathsf{T}} = \widetilde{F} \in \mathbb{C}^{n\times n}$, $[\widetilde{M},\, \widetilde{L}]$ retains the* SSF-1*:*

$$\widetilde{M} = \begin{bmatrix} \widetilde{E} & 0 \\ \widetilde{F} & I_n \end{bmatrix}, \qquad \widetilde{L} = \begin{bmatrix} I_n & \overline{\widetilde{F}} \\ 0 & \overline{\widetilde{E}} \end{bmatrix}.$$

Proof. Let

$$M_* = \begin{bmatrix} E + E\overline{F}(I_n - F\overline{F})^{-1}F & 0 \\ \overline{E}(I_n - F\overline{F})^{-1}F & I_n \end{bmatrix}, \qquad L_* = \begin{bmatrix} I_n & E\overline{F}(I_n - F\overline{F})^{-1} \\ 0 & \overline{E}(I_n - F\overline{F})^{-1} \end{bmatrix},$$

where the fact that $I_n - F\overline{F}$ is nonsingular follows from $F = F^{\mathsf{T}}$ and $1 \notin \sigma(F)$. We have $\operatorname{rank}([M_*,\, L_*]) = 2n$ and

$$M_* L = \begin{bmatrix} E(I_n - \overline{F}F)^{-1} & E\overline{F}(I_n - F\overline{F})^{-1} \\ \overline{E}(I_n - F\overline{F})^{-1}F & \overline{E}(I_n - F\overline{F})^{-1} \end{bmatrix} = L_* M,$$

implying that $[M_*,\, L_*] \in \mathcal{N}(M, L)$. Routine manipulations yield

$$M_* M = \begin{bmatrix} E(I_n - \overline{F}F)^{-1}E & 0 \\ F + \overline{E}F(I_n - \overline{F}F)^{-1}E & I_n \end{bmatrix}, \qquad L_* L = \begin{bmatrix} I_n & \overline{F} + E\overline{F}(I_n - F\overline{F})^{-1}\overline{E} \\ 0 & \overline{E}(I_n - F\overline{F})^{-1}\overline{E} \end{bmatrix}.$$

With $\widetilde{E} = E(I_n - \overline{F}F)^{-1}E$ and $\widetilde{F} = F + \overline{E}F(I_n - \overline{F}F)^{-1}E$, the result follows. □

If we initially take $M_0 = M_\alpha$ and $L_0 = L_\alpha$ (from (3.1)), indicating that $E_0 = E_\alpha$ and $F_0 = F_\alpha$ (specified in (3.2)), then successive doubling transformations in Theorem 14 produce a sequence of symplectic pairs (M_k, L_k) provided that $(I_n - \overline{F}_k F_k)$ are nonsingular for $k \geq 0$. Specifically, we have a well-defined *doubling iteration*, provided that $1 \notin \sigma(F_k)$: (for $k = 0, 1, \ldots$)

$$E_{k+1} = E_k(I_n - \overline{F}_k F_k)^{-1}E_k, \qquad F_{k+1} = F_k + \overline{E}_k F_k(I_n - \overline{F}_k F_k)^{-1}E_k. \tag{3.6}$$

Assuming (3.3) with $S_\alpha \equiv (S - \alpha I_l)^{-1}(S + \alpha I_l)$, Lemmas 4 and 10 imply

$$M_k \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = L_k \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} S_\alpha^{2^k}, \quad M_k = \begin{bmatrix} E_k & 0 \\ F_k & I_n \end{bmatrix}, \quad L_k = \begin{bmatrix} I_n & \overline{F}_k \\ 0 & \overline{E}_k \end{bmatrix}. \tag{3.7}$$

Now we analyze the computational complexity of the DA. By *cop* we denote a basic arithmetic operation on two complex numbers. Computing $E_0 \equiv E_\alpha$ and $F_0 \equiv F_\alpha$ requires $42\frac{1}{3}n^3$ cops and $2n^3$ cops, respectively, where the spectral decomposition of A and the SVD of $R = I_n - (\alpha I_n - \overline{A})^{-1}\overline{B}(\alpha I_n - A)^{-1}B$ are involved. For one iteration (3.6), except $21n^3$ cops for the SVD of F_k, E_{k+1} and F_{k+1} need $4n^3$ cops and $5n^3$ cops, respectively. In total, the DA requires $\mathcal{O}(n^3)$ cops, where the exact counts depends on the number of iterations.

The DA in (3.6) has two iterative formulae for E_k and F_k. Interestingly, the SDAs for Riccati equations and quadratic palindromic eigenvalue problems [7–9] have three, those for nonsymmetric algebraic Riccati equations [17, 18] have four, while the PDA for the linear palindromic eigenvalue problem [16] has one.

Convergence. Without loss of generality, we assume for the moment that $1 \notin \sigma(F_k)$ for all $k = 0, 1, \cdots$. When $1 \in \sigma(F_k)$ for some k, Theorem 19 and Corollary 20 below demonstrate that the following convergence results still hold. We also assume technically that X_1 and $[X_1, \Psi_{11}]$, respectively, are nonsingular in Theorems 15 and 16 below.

Theorem 15. *Assume that H possesses no purely imaginary eigenvalue and*

$$H[X_1^\top,\, X_2^\top\,]^\top = [X_1^\top,\, X_2^\top\,]^\top S$$

with $X_1, X_2, S \in \mathbb{C}^{n\times n}$, where $\lambda(S)$ is in the interior of the left half plane. Then for $\{E_k\}$ and $\{F_k\}$ generated by (3.6), we have $\lim_{k\to\infty} E_k = 0$ and $\lim_{k\to\infty} F_k = -X_2X_1^{-1}$, both converging quadratically.

Proof. Let $S_\alpha \equiv (S - \alpha I_n)^{-1}(S + \alpha I_n)$. Note that the spectral radius of S_α is less than 1 when $\alpha > 0$. The proof is similar to that of [19, Corollary 3.2]. □

The following theorem illustrates the linear convergence of the DA when purely imaginary eigenvalues exist. The proof is analogous to that for [12, Theorem 4.2].

Let the Jordan decompositions of $J_{2p_{r,j}}(\mathrm{i}\omega_j + \alpha)[J_{2p_{r,j}}(\mathrm{i}\omega_j - \alpha)]^{-1}$ be

$$J_{2p_{r,j}}(\mathrm{i}\omega_j + \alpha)[J_{2p_{r,j}}(\mathrm{i}\omega_j - \alpha)]^{-1} = Q_{r,j}J_{2p_{r,j}}(\mathrm{e}^{\mathrm{i}\theta_j})Q_{r,j}^{-1}$$

for $r = 1, \cdots, l_j$ and $j = 1, \cdots, q$. Denote $W_\omega = [W_{1,\omega}^\top, W_{2,\omega}^\top]^\top$, $Z_\omega = [Z_{1,\omega}^\top, Z_{2,\omega}^\top]^\top$ and, for $s', t' = 1, 2$,

$$Q_{r,j} = \begin{bmatrix} Q_{r,j}^{(11)} & Q_{r,j}^{(12)} \\ Q_{r,j}^{(21)} & Q_{r,j}^{(22)} \end{bmatrix}, \quad Q^{(s't')} := \bigoplus_{j=1}^{q}\bigoplus_{r=1}^{l_j} Q_{r,j}^{(s't')};$$

$$\Psi_{11} \equiv W_{1,\omega}Q^{(11)} + Z_{1,\omega}Q^{(21)}, \quad \Psi_{21} \equiv W_{2,\omega}Q^{(11)} + Z_{2,\omega}Q^{(21)},$$
$$\Psi_{12} \equiv W_{1,\omega}Q^{(12)} + Z_{1,\omega}Q^{(22)}, \quad \Psi_{22} \equiv W_{2,\omega}Q^{(12)} + Z_{2,\omega}Q^{(22)}.$$

Theorem 16. *Assume that the partial multiplicities of H associated with the purely imaginary eigenvalues are all even, and H has the eigen-decomposition specified in (2.1). With $X = [X_1^\top, X_2^\top]^\top$ and $[X_1, \Psi_{11}]$ being nonsingular, we have $\lim_{k\to\infty} E_k = 0$ and $\lim_{k\to\infty} F_k = -[X_2, \Psi_{21}][X_1, \Psi_{11}]^{-1}$, both converging linearly.*

Proof. By (2.1) and Lemmas 4 and 10, we have

$$M_k \begin{bmatrix} X_1 & W_{1,\omega} & \overline{X_2} & Z_{1,\omega} \\ X_2 & W_{2,\omega} & \overline{X_1} & Z_{2,\omega} \end{bmatrix} = L_k \begin{bmatrix} X_1 & W_{1,\omega} & \overline{X_2} & Z_{1,\omega} \\ X_2 & W_{2,\omega} & \overline{X_1} & Z_{2,\omega} \end{bmatrix} \widetilde{S}_\alpha^{2^k}, \tag{3.8}$$

where $\widetilde{S}_\alpha = (\widetilde{S} + \alpha I_{2n})(\widetilde{S} - \alpha I_{2n})^{-1}$ with $\widetilde{S}$ from (2.1). Let Π_ω be the permutation matrix satisfying

$$\Pi_\omega \operatorname{diag}\left\{ S, -\overline{S}; \bigoplus_{j=1}^{q}\bigoplus_{r=1}^{l_j} J_{2p_{r,j}}(\mathrm{i}\omega_j)\right\}\Pi_\omega^\top = \widetilde{S},$$

and denote $\mathcal{D} \equiv \operatorname{diag}\left\{I_{n_1}, I_{n_1}; \bigoplus_{j=1}^{q}\bigoplus_{r=1}^{l_j} Q_{r,j}\right\}$, $J_{\omega,\theta} = \bigoplus_{j=1}^{q}\bigoplus_{r=1}^{l_j} J_{p_{r,j}}(\mathrm{e}^{\mathrm{i}\theta_j})$, and $S_\alpha := (S + \alpha I_{n_1})(S - \alpha I_{n_1})^{-1}$, it holds that

$$\widetilde{S}_\alpha = \left(\Pi_\omega \mathcal{D}\Pi_\omega^\top\right)\begin{bmatrix} S_\alpha & & & \\ & J_{\omega,\theta} & & \Omega_\omega \\ & & \overline{S}_\alpha^{-1} & \\ & & & J_{\omega,\theta} \end{bmatrix}\left(\Pi_\omega \mathcal{D}^{-1}\Pi_\omega^\top\right).$$

This further implies

$$\widetilde{S}_{\alpha}^{2^k} = \left(\Pi_\omega \mathcal{D} \Pi_\omega^{\mathsf{T}}\right) \begin{bmatrix} S_\alpha^{2^k} & & & \\ & J_{\omega,\theta}^{2^k} & & \Omega_{\omega,\theta,k} \\ & & \overline{S}_\alpha^{-2^k} & \\ & & & J_{\omega,\theta}^{2^k} \end{bmatrix} \left(\Pi_\omega \mathcal{D}^{-1} \Pi_\omega^{\mathsf{T}}\right) \tag{3.9}$$

with $\Omega_{\omega,\theta,k} = \bigoplus_{j=1}^{q} \bigoplus_{r=1}^{l_j} J_{2p_{r,j}}^{2^k}(\mathrm{e}^{\mathrm{i}\theta_j})(1 : p_{r,j}, p_{r,j}+1 : 2p_{r,j})$. By Lemma 26, for $r = 1, \cdots, l_j$ and $j = 1, \cdots, q$, $J_{2p_{r,j}}^{2^k}(\mathrm{e}^{\mathrm{i}\theta_j})(1 : p_{r,j}, p_{r,j}+1 : 2p_{r,j})$ are nonsingular for sufficiently large k, and so $\Omega_{\omega,\theta,k}$ is. Besides, it follows from Lemma 26 that

$$\|\Omega_{\omega,\theta,k}^{-1} J_{\omega,\theta}^{2^k}\|_2 \le \frac{1}{2^k} \max_{r,j} \left\{ \frac{\sqrt{p_{r,j}}(p_{r,j}+2)(p_{r,j}-1)!}{2(p_{r,j}-1)! - \sqrt{p_{r,j}}} \right\} = \mathcal{O}(2^{-k}),$$

$$\|J_{\omega,\theta}^{2^k} \Omega_{\omega,\theta,k}^{-1} J_{\omega,\theta,k}^{2^k}\|_2 \le \frac{1}{2^k} \max_{r,j} \left\{ \frac{\sqrt{p_{r,j}}(p_{r,j}+2)^2(p_{r,j}-1)!}{2(2(p_{r,j}-1)! - \sqrt{p_{r,j}})} \right\} = \mathcal{O}(2^{-k}).$$

By (3.8) and (3.9) we have

$$\begin{aligned} & M_k \begin{bmatrix} X_1 & W_{1,\omega} & \overline{X_2} & Z_{1,\omega} \\ X_2 & W_{2,\omega} & \overline{X_1} & Z_{2,\omega} \end{bmatrix} \left(\Pi_\omega \mathcal{D} \Pi_\omega^{\mathsf{T}}\right) \\ = & L_k \begin{bmatrix} X_1 & W_{1,\omega} & \overline{X_2} & Z_{1,\omega} \\ X_2 & W_{2,\omega} & \overline{X_1} & Z_{2,\omega} \end{bmatrix} \left(\Pi_\omega \mathcal{D} \Pi_\omega^{\mathsf{T}}\right) \cdot \begin{bmatrix} S_\alpha^{2^k} & & & \\ & J_{\omega,\theta}^{2^k} & & \Omega_{\omega,\theta,k} \\ & & \overline{S}_\alpha^{-2^k} & \\ & & & J_{\omega,\theta}^{2^k} \end{bmatrix}, \end{aligned}$$

equivalent to

$$E_k X_1 = (X_1 + \overline{F}_k X_2) S_\alpha^{2^k}, \tag{3.10a}$$

$$E_k \Psi_{11} = (\Psi_{11} + \overline{F}_k \Psi_{21}) J_{\omega,\theta}^{2^k}, \tag{3.10b}$$

$$E_k \overline{X}_2 = (\overline{X_2} + \overline{F}_k \overline{X}_1) \overline{S}_\alpha^{-2^k}, \tag{3.10c}$$

$$E_k \Psi_{12} = (\Psi_{11} + \overline{F}_k \Psi_{21}) \Omega_{\omega,\theta,k} + (\Psi_{12} + \overline{F}_k \Psi_{22}) J_{\omega,\theta}^{2^k}, \tag{3.10d}$$

$$F_k \Psi_{11} + \Psi_{21} = \overline{E}_k \Psi_{21} J_{\omega,\theta}^{2^k}, \tag{3.10e}$$

$$F_k \Psi_{12} + \Psi_{22} = \overline{E}_k \Psi_{21} \Omega_{\omega,\theta,k} + \overline{E}_k \Psi_{22} J_{\omega,\theta}^{2^k}. \tag{3.10f}$$

Now substituting (3.10b) into (3.10d), (3.10e) into (3.10f), and post-multiplying $\Omega_{\omega,\theta,k}^{-1} J_{\omega,\theta}^{2^k}$ on both right-hand sides give

$$E_k \Psi_{12} \Omega_{\omega,\theta,k}^{-1} J_{\omega,\theta}^{2^k} = E_k \Psi_{11} + \left(\Psi_{12} + \overline{F}_k \Psi_{22}\right) J_{\omega,\theta}^{2^k} \Omega_{\omega,\theta,k}^{-1} J_{\omega,\theta}^{2^k},$$

$$\left(F_k \Psi_{12} + \Psi_{22}\right) \Omega_{\omega,\theta,k}^{-1} J_{\omega,\theta}^{2^k} = F_k \Psi_{11} + \Psi_{21} + \overline{E}_k \Psi_{22} J_{\omega,\theta}^{2^k} \Omega_{\omega,\theta,k}^{-1} J_{\omega,\theta}^{2^k}.$$

Combining the above two equalities with (3.10a) and (3.10c) brings

$$
\begin{aligned}
E_k = {} & [X_1, -\Psi_{12}] \begin{bmatrix} S_\alpha^{2^k} & \\ & J_{\omega,\theta}^{2^k}\Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1} \\
& + \overline{F}_k[X_2, -\Psi_{22}] \begin{bmatrix} S_\alpha^{2^k} & \\ & J_{\omega,\theta}^{2^k}\Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1} \\
& + E_k[0, \Psi_{12}] \begin{bmatrix} S_\alpha^{2^k} & \\ & \Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1},
\end{aligned}
\tag{3.11}
$$

$$
\begin{aligned}
F_k = {} & [-X_2, -\Psi_{21}][X_1, \Psi_{11}]^{-1} \\
& + \overline{E}_k[X_2, -\Psi_{22}] \begin{bmatrix} S_\alpha^{2^k} & \\ & J_{\omega,\theta}^{2^k}\Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1} \\
& + [0, \Psi_{22}] \begin{bmatrix} S_\alpha^{2^k} & \\ & \Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1} \\
& + F_k[0, \Psi_{12}] \begin{bmatrix} S_\alpha^{2^k} & \\ & \Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1}.
\end{aligned}
\tag{3.12}
$$

We then write

$$
\begin{aligned}
\eta_1 &:= \left\| [0, \Psi_{12}\Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k}][X_1, \Psi_{11}]^{-1} \right\|_2, \\
\eta_2 &:= \left\| [X_1, -\Psi_{12}] \begin{bmatrix} S_\alpha^{2^k} & \\ & J_{\omega,\theta}^{2^k}\Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1} \right\|_2, \\
\eta_3 &:= \left\| [X_2, -\Psi_{22}] \begin{bmatrix} S_\alpha^{2^k} & \\ & J_{\omega,\theta}^{2^k}\Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k} \end{bmatrix} [X_1, \Psi_{11}]^{-1} \right\|_2, \\
\eta_4 &:= \left\| [-X_2, -\Psi_{21}][X_1, \Psi_{11}]^{-1} \right\|_2, \quad \eta_5 := \left\| [0, \Psi_{22}\Omega_{\omega,\theta,k}^{-1}J_{\omega,\theta}^{2^k}][X_1, \Psi_{11}]^{-1} \right\|_2,
\end{aligned}
$$

which satisfy $\eta_j = \mathcal{O}(2^{-k})$ $(j \neq 4)$. Taking the 2-norm on both sides of (3.11) and (3.12) gives (for sufficiently large k)

$$
\|E_k\|_2 \le \frac{\eta_2}{1-\eta_1} + \|F_k\|_2 \frac{\eta_3}{1-\eta_1}, \qquad \|F_k\|_2 \le \frac{\eta_4}{1-\eta_1} + \frac{\eta_5}{1-\eta_1} + \|E_k\|_2 \frac{\eta_3}{1-\eta_1},
$$

implying that both sequences $\{E_k\}$ and $\{F_k\}$ are bounded from above. Taking limits on (3.11) and (3.12) yields $\lim_{k\to\infty} E_k = 0$ and

$$
\lim_{k\to\infty} F_k = -[X_2, \Psi_{21}][X_1, \Psi_{11}]^{-1},
$$

where both converge linearly. □

Next assume that we have acquired a sympletic pair (M_k, L_k) with $\|E_k\|_F < \mathbf{u}$, where $\mathbf{u}$ is some small tolerance. The question is then how to compute the eigenvalues and eigenvectors of H from E_k and F_k. Without loss of generality, we just show the details for the case without purely imaginary eigenvalues.

Denote the error $Z_k \equiv F_k + X_2 X_1^{-1}$ (Theorem 15 and (3.6) suggest $\|Z_k\|_F < \mathbf{u}$), where $X_1, X_2 \in \mathbb{C}^{n\times n}$ satisfy $H\left[X_1^{\mathsf{T}}, X_2^{\mathsf{T}}\right]^{\mathsf{T}} = \left[X_1^{\mathsf{T}}, X_2^{\mathsf{T}}\right]^{\mathsf{T}} S$ with $\lambda(S) \subseteq \mathbb{C}_-$. We have

$$H\begin{bmatrix} I_n \\ -F_k \end{bmatrix} = \begin{bmatrix} I_n \\ -F_k \end{bmatrix} X_1SX_1^{-1} + \begin{bmatrix} 0 \\ Z_k \end{bmatrix} X_1SX_1^{-1} - H\begin{bmatrix} 0 \\ Z_k \end{bmatrix}. \tag{3.13}$$

Pre- and post-multiplying $\left[I_n,\, -F_k^{\mathsf{H}}\right]$ and $(I_n + F_k^{\mathsf{H}}F_k)^{-1}$, respectively, to (3.13), we have

$$\begin{aligned}&(I_n + F_k^{\mathsf{H}}F_k)X_1SX_1^{-1}(I_n + F_k^{\mathsf{H}}F_k)^{-1}\\ &= \left\{\left[I_n,\, -F_k^{\mathsf{H}}\right] H \left[I_n,\, -F_k^{\mathsf{T}}\right]^{\mathsf{T}} + (F_k^{\mathsf{H}}Z_kX_1SX_1^{-1} + BZ_k + F_k^{\mathsf{H}}\overline{A}Z_k)\right\}(I_n + F_k^{\mathsf{H}}F_k)^{-1}.\end{aligned}$$

Accordingly, we can take the eigenvalues of $H_k \equiv \left[I_n,\, -F_k^{\mathsf{H}}\right] H \left[I_n,\, -F_k^{\mathsf{T}}\right]^{\mathsf{T}} (I_n + F_k^{\mathsf{H}}F_k)^{-1}$ to approximate $\lambda(S)$ (the stable subspectrum of H). By the generalized Bauer-Fike theorem [31], when the eigenvalues $\lambda_p(S)$ have Jordan blocks of maximum size m, there exists an eigenvalue $\lambda_q(H_k)$ such that

$$\begin{aligned}\frac{|\lambda_p(S) - \lambda_q(H_k)|^m}{[1 + |\lambda_p(S) - \lambda_q(H_k)|]^{m-1}} &\le \Upsilon\|(F_k^{\mathsf{H}}Z_kX_1SX_1^{-1} + BZ_k + F_k^{\mathsf{H}}\overline{A}Z_k)(I_n + F_k^{\mathsf{H}}F_k)^{-1}\|_2\\ &\le \Upsilon\|F_k^{\mathsf{H}}Z_kX_1SX_1^{-1} + BZ_k + F_k^{\mathsf{H}}\overline{A}Z_k\|_2\end{aligned}$$

for some $\Upsilon > 0$ associated with S. So we can approximate $\lambda(S)$ by $\lambda(H_k)$.

3.3. **Double-Cayley transform.**

When $1 \in \sigma(F_{k_0})$ (a generically rare occasion) for some $k_0 > 1$ (or the condition in Theorem 14 is violated), we cannot construct the new symplectic pair $(M_{k_0+1},\, L_{k_0+1})$ via the doubling transformation in (3.6). In this section, we divert the DA from this potential interruption using a DCT. We shall also prove the efficiency of the technique, not requiring a restart with a new α. It is worthwhile to point out that the DCT may be applied when $I - \overline{F}_{k_0}F_{k_0}$ is ill-conditioned. In practice, we may set a tolerance $\mathbf{u}$ and once the singular values [10] of F_{k_0} satisfy $\min_{\sigma\in\sigma(F_{k_0})}|\sigma - 1|/\max_{\sigma\in\sigma(F_{k_0})}|\sigma - 1| < \mathbf{u}$, the DCT is then applied.

Theorem 17. *Let $\vartheta \in \{-1, 1\}$ and $\beta \in \mathbb{R}$. Provided that $\vartheta \notin \lambda(E_{k_0})$, then*

(a) *$Z = \vartheta I_n - E_{k_0} + \vartheta\overline{F}_{k_0}(\vartheta\overline{E}_{k_0} - I_n)^{-1}F_{k_0}$ is nonsingular;*

(b) *$(\widehat{H} + \beta\vartheta I_{2n})[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} = (\widehat{H} - \beta\vartheta I_{2n})[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}}(\vartheta S_\alpha^{2^{k_0}})$ with $\widehat{A} = \beta\vartheta I_n - 2\beta Z^{-1}$, $\widehat{B} = (\beta I_n - \vartheta\widehat{A})\overline{F}_{k_0}(\overline{E}_{k_0} - \vartheta I_n)^{-1}$ and $\widehat{H} = \begin{bmatrix} \widehat{A} & \widehat{B} \\ -\overline{\widehat{B}} & -\overline{\widehat{A}} \end{bmatrix}$; and*

(c) *$\widehat{A}$ is Hermitian and $\widehat{B}$ is symmetric.*

(The proof is deferred to Appendix B.1.)

Theorem 17 implies $\widehat{H}[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} = \beta\vartheta[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}}(S_\alpha^{2^{k_0}} + \vartheta I_l)(S_\alpha^{2^{k_0}} - \vartheta I_l)^{-1}$, hence each eigenvalue λ of H corresponds to an eigenvalue μ of $\widehat{H}$:

$$\mu = f(\lambda) \triangleq \beta\vartheta\cdot\frac{(\lambda+\alpha)^{2^{k_0}} + \vartheta(\lambda-\alpha)^{2^{k_0}}}{(\lambda+\alpha)^{2^{k_0}} - \vartheta(\lambda-\alpha)^{2^{k_0}}}. \tag{3.14}$$

More specifically, for $\lambda \in \lambda(H)$, we have

$$\begin{cases} \{\mu,\ \overline{\mu} = f(\overline{\lambda}),\ -\mu = f(-\lambda),\ -\overline{\mu} = f(-\overline{\lambda})\} \subseteq \lambda(\widehat{H}) & \text{if } \Re(\lambda)\Im(\lambda) \ne 0;\\ \{\mu,\ -\mu = f(-\lambda)\} \subseteq \lambda(\widehat{H}) & \text{if } \Im(\lambda) = 0;\\ \{\mu,\ \overline{\mu} = f(\overline{\lambda})\} \subseteq \lambda(\widehat{H}) & \text{if } \Re(\lambda) = 0.\end{cases}$$

In addition, $\mu \in \lambda(\widehat{H})$ is purely imaginary if $\lambda \in \lambda(H)$ is so. Equivalently, there exist no purely imaginary eigenvalues for $\widehat{H}$ when there is none for H.

Next select $\gamma \in \mathbb{R}$ with $\gamma I_n - \widehat{A}$ and $I_n - \left(\gamma I_n - \overline{\widehat{A}}\right)^{-1} \overline{\widehat{B}} \left(\gamma I_n - \widehat{A}\right)^{-1} \widehat{B}$ being nonsingular. Theorem 9 could then be applied to $\widehat{A}$ and $\widehat{B}$, which are defined in Theorem 17, to obtain a new SSF-1 derived from $\widehat{H}$. Thus, we have

$$M_{k_0+1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = L_{k_0+1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \left[\beta\vartheta(S_\alpha^{2^{k_0}} + \vartheta I_l)(S_\alpha^{2^{k_0}} - \vartheta I_l)^{-1} + \gamma I_l\right] \cdot \left[\beta\vartheta(S_\alpha^{2^{k_0}} + \vartheta I_l)(S_\alpha^{2^{k_0}} - \vartheta I_l)^{-1} - \gamma I_l\right]^{-1},$$

with

$$M_{k_0+1} = \begin{bmatrix} E_{k_0+1} & 0 \\ F_{k_0+1} & I_n \end{bmatrix}, \qquad L_{k_0+1} = \begin{bmatrix} I_n & \overline{F}_{k_0+1} \\ 0 & \overline{E}_{k_0+1} \end{bmatrix},$$

$$E_{k_0+1} = I_n - 2\gamma \left[(\gamma I_n - \widehat{A}) - \widehat{B}(\gamma I_n - \overline{\widehat{A}})^{-1}\overline{\widehat{B}}\right]^{-1},$$

$$F_{k_0+1} = -2\gamma \left(\gamma I_n - \overline{\widehat{A}}\right)^{-1} \overline{\widehat{B}} \left[(\gamma I_n - \widehat{A}) - \widehat{B}(\gamma I_n - \overline{\widehat{A}})^{-1}\overline{\widehat{B}}\right]^{-1}.$$

We call the above transform from (M_{k_0}, L_{k_0}) to (M_{k_0+1}, L_{k_0+1}), both symplectic, a *DCT*. Accordingly, with $\delta_\lambda \triangleq (\lambda+\alpha)(\lambda-\alpha)^{-1}$, $|\delta_\lambda| < 1$ and $\varpi \triangleq (\beta-\vartheta\gamma)(\beta\vartheta+\gamma)^{-1}$, an eigenvalue μ of $\widehat{H}$ (in (3.14)) would be transformed into an eigenvalue ν of (M_{k_0+1}, L_{k_0+1}) via the following formula: (for $\lambda \in \lambda(H)$)

$$\begin{aligned} \nu \equiv \nu(\mu) &= \frac{\mu+\gamma}{\mu-\gamma} \\ &= \frac{\beta\vartheta[(\lambda+\alpha)^{2^{k_0}} + \vartheta(\lambda-\alpha)^{2^{k_0}}] + \gamma[(\lambda+\alpha)^{2^{k_0}} - \vartheta(\lambda-\alpha)^{2^{k_0}}]}{\beta\vartheta[(\lambda+\alpha)^{2^{k_0}} + \vartheta(\lambda-\alpha)^{2^{k_0}}] - \gamma[(\lambda+\alpha)^{2^{k_0}} - \vartheta(\lambda-\alpha)^{2^{k_0}}]} = \vartheta \cdot \frac{\varpi + \delta_\lambda^{2^{k_0}}}{1 + \varpi\delta_\lambda^{2^{k_0}}}. \end{aligned}$$

One may consider the condition number of $I_n - \overline{F}_{k_0+1}F_{k_0+1}$, or equivalently, the difference between 1 and $\sigma(F_{k_0+1})$. Obviously, $\sigma(F_{k_0})$ depends on γ. Without loss of generality we assume $\vartheta = 1$; then with $\gamma = \beta(\kappa^{2^{k_0}} + 1)(\kappa^{2^{k_0}} - 1)^{-1}$(where κ is to be specified), we have

$$F_{k_0+1} = -\frac{\kappa^{2^{k_0}}+1}{\kappa^{2^{k_0}}-1} \left(\frac{\overline{Z}}{\kappa^{2^{k_0}}-1} + I_n\right)^{-1} F_{k_0}(E_{k_0} - I_n)^{-1} \cdot \left[\left(\frac{Z}{\kappa^{2^{k_0}}-1} + I_n\right) - \overline{F}_{k_0}(\overline{E}_{k_0} - I_n)^{-1} \left(\frac{\overline{Z}}{\kappa^{2^{k_0}}-1} + I_n\right)^{-1} F_{k_0}(E_{k_0} - I_n)^{-1}\right]^{-1} Z.$$

Thus we can choose some κ to make $I_n - \overline{F}_{k_0+1}F_{k_0+1}$ well conditioned. We leave the issue of an optimal κ (or γ) for the future, while making random choices in our numerical experiments. Theorem 19 and Corollary 20 below illustrate that κ characterizes the convergence rate and does not have to be large.

With $\gamma > 0$ and $\Re(\mu) < 0$, we have $|\nu(\mu)| < 1$. The following lemma reveals more.

Lemma 18. *Provided that $\vartheta\beta, \gamma > 0$, then each ν corresponding to a nonpurely imaginary eigenvalue $\lambda \in \lambda(H)$ with $\Re(\lambda) < 0$ satisfies $|\nu| < 1$.*

(The proof of Lemma 18 can be found in Appendix B.2.)

Lemma 18 demonstrates that for $\lambda \in \lambda(H)$ satisfying $\Im(\lambda) \neq 0$, the DCT maps half of these λ to some values inside of the unit circle and the other half outside. Next we consider the detailed relationship between ν and $\varrho = \delta_\lambda^{2^{k_0}}$, which is vital for the convergence of the DA coupled with the DCT.

Obviously, when $\vartheta\beta, \gamma > 0$, we have $|\varpi| < 1$. Taking $\gamma = \beta(\kappa^{2^{k_0}} + \vartheta)(\vartheta\kappa^{2^{k_0}} - 1)^{-1} > 0$ with $\kappa > 1$, we obtain $\varpi = -\kappa^{-2^{k_0}}$ and

$$\nu = \vartheta \cdot \frac{\delta_\lambda^{2^{k_0-1}} - \kappa^{-2^{k_0-1}}}{1 - \delta_\lambda^{2^{k_0-1}}\kappa^{-2^{k_0-1}}} \cdot \frac{\delta_\lambda^{2^{k_0-1}} + \kappa^{-2^{k_0-1}}}{1 + \delta_\lambda^{2^{k_0-1}}\kappa^{-2^{k_0-1}}}. \tag{3.15}$$

With the new formula (3.15) of ν, then under the assumptions in Lemma 18 the following theorem gives a sharp bound for those $|\nu|$ corresponding to λ which satisfies $\Im(\lambda) \neq 0$ and $|\delta_\lambda| < 1$.

Theorem 19. *Assume that λ is not a purely imaginary eigenvalue of H, $\vartheta\beta > 0$, and $\kappa \geq 2$. Then we have $|\nu| \leq \max\left\{|\delta_\lambda|^{2^{k_0-2}},\ \kappa^{-2^{k_0-2}}\right\}$.*

(The proof is in Appendix B.3.)

For a real $\lambda \in \lambda(H)$, we can obtain a better result, with the power 2^{k_0-2} replaced by 2^{k_0} in the following corollary.

Corollary 20. *Let $\kappa > 1$ and $\vartheta\beta, \alpha > 0$; then for $\lambda < 0$ ($\lambda \in \lambda(H)$), we have $|\nu| \leq \max\left\{|\delta_\lambda|^{2^{k_0}},\ \kappa^{-2^{k_0}}\right\}$.*

Proof. Let $\phi \equiv \operatorname{arctanh} \delta_\lambda^{2^{k_0}}$; then $\phi = \frac{1}{2}\ln\left[\frac{(\lambda-\alpha)^{2^{k_0}}+(\lambda+\alpha)^{2^{k_0}}}{(\lambda-\alpha)^{2^{k_0}}-(\lambda+\alpha)^{2^{k_0}}}\right] > 0$ since $\lambda < 0$, and $\psi \equiv \operatorname{arctanh}(-\kappa^{-2^{k_0}}) = -\frac{1}{2}\ln\left(\frac{\kappa^{2^{k_0}}+1}{\kappa^{2^{k_0}}-1}\right) < 0$. From the definition of ν, we have $\nu = \vartheta\tanh(\phi + \psi)$. Because $\tanh(\omega) = (\mathrm{e}^{\omega} - \mathrm{e}^{-\omega})(\mathrm{e}^{\omega} + \mathrm{e}^{-\omega})^{-1}$, $\tanh(-\omega) = -\tanh(\omega)$ and $\tanh(\omega)$ is nondecreasing with respect to $\omega \in \mathbb{R}$. Then when $\phi \geq |\psi|$ we have $0 \leq |\nu| = \tanh(\phi + \psi) \leq \tanh(\phi)$. Otherwise for $\phi < |\psi|$, we have $|\nu| = \tanh(-\psi - \phi) < \tanh(-\psi) = \kappa^{-2^{k_0}}$. Hence, the result holds. □

To sum up, we propose the DCT to avoid the potential interruption of the DA caused by $1 \in \sigma(F_{k_0})$ for some k_0. We have conducted a detailed analysis on the eigenvalue ν of the new symplectic pair (M_{k_0+1}, L_{k_0+1}), producing a sharp bound of $|\nu|$ in Theorem 19 relative to $|\delta_\lambda|^{2^{k_0-2}}$. Furthermore, Theorem 19 and Corollary 20 imply that a double-Cayley step reverses the convergence *at worst by two steps in general and not at all when λ is real.* This guarantees the convergence of the DA when the DCT is only occasionally called for. Similar comments apply when there exist some singular value $\sigma \in \sigma(F_{k_0})$ close to unity, meaning that $I - \overline{F}_{k_0}F_{k_0}$ is ill-conditioned, and the double-Cayley remedy is applied.

Note that the DCT is applicable when $\vartheta \notin \lambda(E_{k_0})$ with $\vartheta \in \{-1, 1\}$. In the rare occasions when the condition is violated, the three-recursion remedy proposed in Section 3.4 will be employed.

We construct a special type of examples to show the need for the DCT.

Example 3.1. Let $A = \overline{U}\Lambda_A U^{\mathsf{T}} \in \mathbb{C}^{5\times5}$ and $B = \overline{U}\Sigma_B U^{\mathsf{H}} \in \mathbb{C}^{5\times5}$ with U being some unitary matrix generated by functions `randn` and `qr` in MATLAB, that is, $[U, R] = \mathtt{qr}(\mathtt{randn}(5,5) + \mathtt{i}\mathtt{randn}(5,5))$. Then by Theorem 9 we have $E_0 = \overline{U}\Lambda_0 U^{\mathsf{T}}$

and $F_0 = U\Sigma_0 U^{\mathsf{T}}$, where $\Lambda_0 = I_5 + 2\alpha\left(I_5 - \Sigma_B\overline{\Sigma_B}(\alpha I_5 - \Lambda_A)^{-2}\right)^{-1}(\Lambda_A - \alpha I_5)^{-1}$ and $\Sigma_0 = -2\alpha\overline{\Sigma_B}\left((\Lambda_A - \alpha I_5)^2 - \Sigma_B\overline{\Sigma_B}\right)^{-1}$. Now applying the doubling iteration (3.6) gives

$$\begin{aligned} E_k &= \overline{U}\Lambda_k U^{\mathsf{T}}, & \Lambda_k &= \Lambda_{k-1}^2(I_5 - \Sigma_{k-1}\overline{\Sigma_{k-1}})^{-1}, \\ F_k &= U\Sigma_k U^{\mathsf{T}}, & \Sigma_k &= \Sigma_{k-1} + \Lambda_{k-1}^2\Sigma_{k-1}(I_5 - \Sigma_{k-1}\overline{\Sigma_{k-1}})^{-1}. \end{aligned}$$

Provided that $I_5 + \Lambda_k$ is nonsingular, we reversely deduce that

$$\Sigma_{k-1} = \Sigma_k(I_5 + \Lambda_k)^{-1}, \qquad \Lambda_{k-1}^2 = \Lambda_k\left(I_5 - \Sigma_k\overline{\Sigma_k}(I_5 + \Lambda_k)^{-2}\right).$$

Moreover, when $\Lambda_k\left(I_5 - \Sigma_k\overline{\Sigma_k}(I_5 + \Lambda_k)^{-2}\right)$ is positive definite, we have

$$\Lambda_{k-1} = \Lambda_k^{\frac{1}{2}}\left(I_5 - \Sigma_k\overline{\Sigma_k}(I_5 + \Lambda_k)^{-2}\right)^{\frac{1}{2}}.$$

Specifically, if $\Lambda_k > 0$ and $0 < \Sigma_k < I_5 + \Lambda_k$, then the above inverse procedure can proceed without breakdown and eventually gives

$$\begin{aligned} \Lambda_A &= \alpha I_5 + 2\alpha(\Lambda_0 - I_5)^{-1}\left(I_5 - \Sigma_0\overline{\Sigma_0}(\Lambda_0 - I_5)^{-2}\right)^{-1}, \\ \Sigma_B &= -2\alpha\overline{\Sigma_0}\left((\Lambda_0 - I_5)^2 - \Sigma_0\overline{\Sigma_0}\right)^{-1}. \end{aligned}$$

By the above generating procedure, we initially set $k = 4$,

$$\begin{aligned} \Lambda_k &= \operatorname{diag}(10^{-3}, 10^{-3}, 10^{-3}, 10^{-3}, 10^{-6}), \\ \Sigma_k &= \operatorname{diag}(0.5, 0.5, 0.5, 0.5, 1), \end{aligned}$$

and $\alpha = 1$ to compute H. Note that the next doubling step breaks down for E_4 and F_4 since $1 \in \sigma(F_4)$ and the DCT is required to carry the DA forward. For the DCT, which is applied to E_4 and F_4, we take $\vartheta = 1$, $\beta = 2$, and $\kappa = 2$, implying $\gamma = 2(2^{16}+1)(2^{16}-1)^{-1}$. Relative to iteration j, Figure 1 plots the 2−norm of Λ_j (or $\|E_j\|_2$). The DCT corresponds to the dashed line; data marked with circles correspond to those obtained by reversed iterations; and data marked with asterisks correspond to those acquired by the forward step, after the DCT. It shows that $\|\Lambda_j\|_2$ (or $\|E_j\|_2$) convergences to 0, indicating the effectiveness of the DCT.

3.4. **Three-recursion remedy.** This subsection is devoted to resolving the issue that the DCT fails. In particular, one may apply the three-recursion remedy from this section when more than two step reversions occur with some complex eigenvalues.

Let $Z = Z^{\mathsf{T}} \in \mathbb{C}^{n\times n}$ (which may be chosen randomly) and let $I_n + \overline{F}_k Z$ be nonsingular. Write $P_k = (I_n + \overline{F}_k Z)^{-1}E_k$, $G_k = (I_n + \overline{F}_k Z)^{-1}\overline{F}_k$, and $H_k = (F_k + Z) - E_k^{\mathsf{T}}Z(I + \overline{F}_k Z)^{-1}E_k$. The following lemma shows how we transform the two recursions for E_k and F_k to three.

Lemma 21. *For the decomposition* (2.1) *it holds that*

$$\begin{aligned} &\begin{bmatrix} P_k & 0 \\ H_k & I_n \end{bmatrix}\begin{bmatrix} I & 0 \\ -Z & I_n \end{bmatrix}\begin{bmatrix} X_1 & W_{1,\omega} & \overline{X}_2 & Z_{1,\omega} \\ X_2 & W_{2,\omega} & \overline{X}_1 & Z_{2,\omega} \end{bmatrix} \\ &= \begin{bmatrix} I_n & G_k \\ 0 & P_k^{\mathsf{T}} \end{bmatrix}\begin{bmatrix} I & 0 \\ -Z & I_n \end{bmatrix}\begin{bmatrix} X_1 & W_{1,\omega} & \overline{X}_2 & Z_{1,\omega} \\ X_2 & W_{2,\omega} & \overline{X}_1 & Z_{2,\omega} \end{bmatrix}\widetilde{S}_\alpha^{2^k}, \end{aligned} \tag{3.16}$$

where $X_1, X_2, W_{1,\omega}, W_{2,\omega}, Z_{1,\omega}, Z_{2,\omega}$, and $\widetilde{S}_\alpha^{2^k}$ are defined as in (3.8).

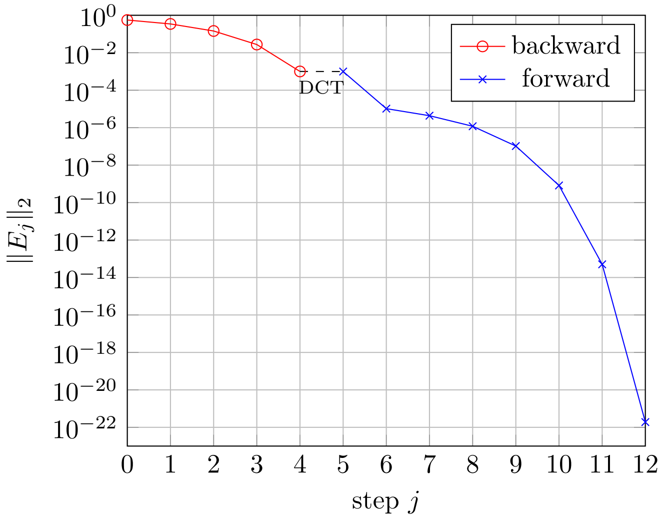

FIGURE 1. 2−norm of E_j relative to iteration step

Proof. Define $\Phi = \begin{bmatrix} (I_n + \overline{F}_k Z)^{-1} & 0 \\ -E_k^{\mathsf{T}} Z (I_n + \overline{F}_k Z)^{-1} & I_n \end{bmatrix}$; then we deduce that

$$\Phi \begin{bmatrix} E_k & 0 \\ F_k & I_n \end{bmatrix} \begin{bmatrix} I_n & 0 \\ Z & I_n \end{bmatrix} = \begin{bmatrix} P_k & 0 \\ H_k & I_n \end{bmatrix}, \quad \Phi \begin{bmatrix} I_n & \overline{F}_k \\ 0 & \overline{E}_k \end{bmatrix} \begin{bmatrix} I_n & 0 \\ Z & I_n \end{bmatrix} = \begin{bmatrix} I_n & G_k \\ 0 & P_k^{\mathsf{T}} \end{bmatrix}.$$

With $\begin{bmatrix} I_n & 0 \\ Z & I_n \end{bmatrix}^{-1} = \begin{bmatrix} I_n & 0 \\ -Z & I_n \end{bmatrix}$, the result follows from (3.8). □

Since $F_k^{\mathsf{T}} = F_k$ and $Z^{\mathsf{T}} = Z$, we have $G_k^{\mathsf{T}} = G_k$ and $H_k^{\mathsf{T}} = H_k$. Applying the doubling algorithms [19] for CARE and DARE, provided that $(I_n - G_{k+j-1}H_{k+j-1})^{-1}$ are well-defined for $j \geq 1$, we formulate the three recursions for P_{k+j}, G_{k+j}, and H_{k+j}:

$$\begin{aligned} P_{k+j} &= P_{k+j-1}(I_n - G_{k+j-1}H_{k+j-1})^{-1}P_{k+j-1}, \\ G_{k+j} &= G_{k+j-1} + P_{k+j-1}(I_n - G_{k+j-1}H_{k+j-1})^{-1}G_{k+j-1}P_{k+j-1}^{\mathsf{T}}, \\ H_{k+j} &= H_{k+j-1} + P_{k+j-1}^{\mathsf{T}}H_{k+j-1}(I_n - G_{k+j-1}H_{k+j-1})^{-1}P_{k+j-1}, \end{aligned} \tag{3.17}$$

where $G_{k+j}^{\mathsf{T}} = G_{k+j}$ and $H_{k+j}^{\mathsf{T}} = H_{k+j}$. It is worthwhile to point out that when $I_n - G_{k+j}H_{k+j}$ is singular or ill-conditioned, we can always randomly choose some other $Z^{\mathsf{T}} = Z \in \mathbb{C}^{n\times n}$ and construct $\Psi \in \mathbb{C}^{2n\times 2n}$ such that

$$\Psi \begin{bmatrix} P_{k+j} & 0 \\ H_{k+j} & I_n \end{bmatrix} \begin{bmatrix} I_n & 0 \\ Z & I_n \end{bmatrix} = \begin{bmatrix} \widetilde{P}_{k+j} & 0 \\ \widetilde{H}_{k+j} & I_n \end{bmatrix}, \quad \Psi \begin{bmatrix} I_n & G_{k+j} \\ 0 & P_{k+j}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ Z & I_n \end{bmatrix} = \begin{bmatrix} I_n & \widetilde{G}_{k+j} \\ 0 & \widetilde{P}_{k+j}^{\mathsf{T}} \end{bmatrix}.$$

With $I_n - G_{k+j}H_{k+j}$ being well-conditioned for all $j \geq 0$, the following two theorems demonstrate the convergence of the three recursions specified in (3.17).

Theorem 22. *Upon the assumption in Theorem* 15*, it holds that* $\lim_{k\to\infty} P_k = 0$ *and* $\lim_{k\to\infty} H_k = Z - X_2X_1^{-1}$*, both converging quadratically.*

Proof. The results follow from the fact

$$\begin{aligned} &\begin{bmatrix} P_k & 0 \\ H_k & I_n \end{bmatrix} \begin{bmatrix} X_1 & \overline{X}_2 \\ X_2 - ZX_1 & \overline{X}_1 - Z\overline{X}_2 \end{bmatrix} \\ &= \begin{bmatrix} I_n & G_k \\ 0 & P_k^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} X_1 & \overline{X}_2 \\ X_2 - ZX_1 & \overline{X}_1 - Z\overline{X}_2 \end{bmatrix} \begin{bmatrix} S_\alpha^{2^k} & \\ & \overline{S}_\alpha^{-2^k} \end{bmatrix} \end{aligned}$$

and $\lim_{k\to\infty} S_\alpha^{2^k} = 0$. We omit the details, as in [19, Corollary 3.2]. □

Theorem 23. *Under the assumption in Theorem* 16*, it holds that* $\lim_{k\to\infty} P_k = 0$ *and* $\lim_{k\to\infty} H_k = Z - X_2 X_1^{-1}$*, both converging linearly.*

Proof. By (3.16) and similar to the proof of Theorem 16, we obtain the result. □

4. Numerical results

We illustrate the performance of the DA with some test examples, three of which from discretized Bethe-Salpeter equations and one generated by the `randn` command in MATLAB. We also apply `eig` in MATLAB (as in $\texttt{eig}(H)$ and $\texttt{eig}(\Gamma H, \Gamma)$) and Algorithm 1 in [30] for comparison. Computing $\texttt{eig}(\Gamma H, \Gamma)$ is based on the equivalence of $Hx = \lambda x$ and $\begin{bmatrix} A & B \\ \overline{B} & \overline{A} \end{bmatrix} x = \lambda \begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix} x$. No DCT or three-recursion remedy is required. All algorithms are implemented in MATLAB 2012b on a 64-bit PC with an Intel Core i7 processor at 3.4 GHz and 8G RAM.

Example 4.1. We consider three examples from the discretized Bethe-Salpeter equations for naphthalene ($C_{10}H_8$), gallium arsenide (GaAs), and boron nitride (BN). The dimensions of the corresponding H associated with $C_{10}H_8$, GaAs, and BN are, respectively, 64, 256 and 4608. All eigenpairs of H are computed.

Using $\texttt{eig}(H)$ as the baseline for comparison, we present the relative accuracy of the computed eigenvalues and the relative execution time (eTime) of the other three algorithms, all averaged over 50 trials. For the relative accuracy, we compute $\text{prec} = \log_{10}[\max_j |(\lambda_j - \widehat{\lambda}_j)/\lambda_j|]$, where λ_j and $\widehat{\lambda}_j$ are the computed eigenvalues by the $\texttt{eig}(H)$ command and one of the methods, respectively. The residuals

$$\frac{\|H - [X, \Pi\overline{X}]\operatorname{diag}(S, -\overline{S})[X, \Pi\overline{X}]^{-1}\|_F}{\|H\|_F}, \quad \frac{\|Y^{\mathsf{H}} H X - \Lambda\|_F}{\|H\|_F},$$

for the DA, $\texttt{eig}(\Gamma H, \Gamma)$ and [30, Algorithm 1] are displayed, with Y and X being, respectively, the left and right eigenvector matrices and Λ the diagonal matrix containing the eigenvalues of H (see [30] for details). Also, the numbers of iterations required for the DA averaged over 50 trails are presented. For the DA all α's in the 50 trails are generated by the function `randn`. The results are tabulated in Table 1.

TABLE 1. Numerical results for Example 4.1

		$C_{10}H_8$	
	DA	algorithm 1 in [30]	$\mathtt{eig}(\Gamma H, \Gamma)$
prec	-13.97	-13.92	-13.95
residual	8.14×10^{-16}	2.60×10^{-15}	1.71×10^{-15}
eTime	7.958×10^{-1}	5.764×10^{-1}	3.792×10^{-1}
iteration	6.84	–	–
		GaAs	
	DA	algorithm 1 in [30]	$\mathtt{eig}(\Gamma H, \Gamma)$
prec	-13.74	-13.54	-13.75
residual	6.86×10^{-16}	6.33×10^{-15}	5.07×10^{-15}
eTime	5.881×10^{-1}	3.587×10^{-1}	3.533×10^{-1}
iteration	8.46	–	–
		BN	
	DA	algorithm 1 in [30]	$\mathtt{eig}(\Gamma H, \Gamma)$
prec	-13.11	-13.12	-13.04
residual	7.50×10^{-16}	2.54×10^{-14}	1.63×10^{-14}
eTime	6.610×10^{-1}	4.754×10^{-1}	4.843×10^{-1}
iteration	7.44	–	–

Table 1 demonstrates that all three methods produce comparable results in terms of the relative accuracy. The DA spends slightly more time than the other methods but produces more accurate solutions with smaller residuals.

Example 4.2. The test example, randomly generated by the command `randn` in MATLAB, is designed to illustrate the structure-preserving property, a distinct feature of the DA. We have $A = \mathrm{diag}\{A_1, A_2, A_3\}$ and $B = \mathrm{diag}\{B_1, B_2, B_3\}$ with

$$A_1 = \begin{bmatrix} 2.6361 & 1.0378 \times 10^{1} & 5.0751 \times 10^{-2} \\ 1.0378 \times 10^{1} & 5.2431 \times 10^{-2} & -4.6067 \times 10^{-1} \\ 5.0751 \times 10^{-2} & -4.6067 \times 10^{-1} & -1.6892 \times 10^{-2} \end{bmatrix},$$

$$A_2 = \begin{bmatrix} -4.0549 \times 10^{-1} & -3.7710 + 2.7569\mathrm{i} \\ -3.7710 - 2.7569\mathrm{i} & -4.0549 \times 10^{-1} \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 3.6378 \times 10^{-1} & 2.7293 \times 10^{-1} + 3.5908\mathrm{i} \\ 2.7293 \times 10^{-1} - 3.5908\mathrm{i} & 3.6378 \times 10^{-1} \end{bmatrix},$$

$$B_1 = \begin{bmatrix} -2.6361 & -1.0375 \times 10^{1} & -5.1181 \times 10^{-2} \\ -1.0375 \times 10^{1} & -5.3457 \times 10^{-2} & 5.0988 \times 10^{-1} \\ -5.1181 \times 10^{-2} & 5.0988 \times 10^{-1} & 4.2022 \times 10^{-3} \end{bmatrix},$$

$$B_2 = \begin{bmatrix} 1.2343 \times 10^{-1} - 3.8788\mathrm{i} \times 10^{-1} & 3.7566 - 2.7464\mathrm{i} \\ 3.7566 - 2.7464\mathrm{i} & 4.0704 \times 10^{-1} + 6.0156\mathrm{i} \times 10^{-5} \end{bmatrix},$$

$$B_3 = \begin{bmatrix} 3.6148 \times 10^{-1} - 5.5211\mathrm{i} \times 10^{-2} & -2.7152 \times 10^{-1} - 3.5722\mathrm{i} \\ -2.7152 \times 10^{-1} - 3.5722\mathrm{i} & -3.6567 \times 10^{-1} + 5.9265\mathrm{i} \times 10^{-5} \end{bmatrix}.$$

The spectrum of H is

$$\begin{aligned} \lambda(H) = \quad & \{\pm 4.1204 \times 10^{-3}, \quad \pm 4.1204 \times 10^{-3}, \quad \pm 4.1204 \times 10^{-3}, \\ & \pm 4.0549 \times 10^{-1} \pm 5.9927\mathrm{i} \times 10^{-5}, \quad \pm 3.6378 \times 10^{-1} \pm 5.8959\mathrm{i} \times 10^{-5}\}. \end{aligned}$$

Note that the algebraic and the geometric multiplicities of $\pm 4.1204 \times 10^{-3}$ are 3 and 1, respectively. The DA, $\mathtt{eig}(H)$ and $\mathtt{eig}(\Gamma H, \Gamma)$ produce the eigenvalues λ_D,

λ_E and λ_{Ge}, respectively,

$$\begin{aligned}
\lambda_D = \ & \{\pm 4.1092 \times 10^{-3}, \quad \pm 4.1092 \times 10^{-3}, \quad \pm 4.1092 \times 10^{-3}, \\
& \ \pm 4.0549 \times 10^{-1} \pm 5.9927\mathrm{i} \times 10^{-5}, \quad \pm 3.6378 \times 10^{-1} \pm 5.8959\mathrm{i} \times 10^{-5}\}, \\
\lambda_E = \ & \{ \ \ 4.1137 \times 10^{-3} - 1.1615\mathrm{i} \times 10^{-5}, \quad 4.1136 \times 10^{-3} + 1.1614\mathrm{i} \times 10^{-5}, \\
& \ \ \ 4.1338 \times 10^{-3} + 1.2681\mathrm{i} \times 10^{-9}, \\
& \ -4.1136 \times 10^{-3} - 1.1649\mathrm{i} \times 10^{-5}, \quad -4.1136 \times 10^{-3} + 1.1650\mathrm{i} \times 10^{-5}, \\
& \ -4.1338 \times 10^{-3} - 1.3011\mathrm{i} \times 10^{-9}, \\
& \ \pm 4.0549 \times 10^{-1} \pm 5.9927\mathrm{i} \times 10^{-5}, \quad \pm 3.6378 \times 10^{-1} \pm 5.8959\mathrm{i} \times 10^{-5}\}, \\
\lambda_{Ge} = \ & \{ \ \ 4.1272 \times 10^{-3} - 1.1919\mathrm{i} \times 10^{-5}, \quad 4.1272 \times 10^{-3} - 1.1919\mathrm{i} \times 10^{-5}, \\
& \ \ \ 4.1272 \times 10^{-3} - 1.1919\mathrm{i} \times 10^{-5}, \\
& \ -4.1272 \times 10^{-3} + 1.1851\mathrm{i} \times 10^{-5}, \quad -4.1272 \times 10^{-3} + 1.1851\mathrm{i} \times 10^{-5}, \\
& \ -4.1272 \times 10^{-3} + 1.1851\mathrm{i} \times 10^{-5}, \\
& \ \pm 4.0549 \times 10^{-1} \pm 5.9927\mathrm{i} \times 10^{-5}, \quad \pm 3.6378 \times 10^{-1} \pm 5.8959\mathrm{i} \times 10^{-5}\}.
\end{aligned}$$

Although all three methods produce computed eigenvalues of low relative accuracy, with $prec_D = -2.5680$, $prec_E = -2.4862$ and $prec_{Ge} = -2.4764$, the DA preserves the distinct eigenstructure of H. All eigenvalues from the DA appear in quadruples $\{\lambda, \overline{\lambda}, -\lambda, -\overline{\lambda}\} \subseteq \lambda(H)$, unless when $\Im(\lambda) = 0$ then in pairs $\{\lambda, -\lambda\} \subseteq \lambda(H)$. The low accuracy (in the order of $\pm 4.1204 \times 10^{-3}$) of the computed eigenvalues from the methods are attributed to the defective eigenvalues. Note that Algorithm 1 in [30] failed because the required assumption $\Gamma H > 0$ is not satisfied.

5. Conclusions

We propose a doubling algorithm for the discretized Bethe-Salpeter eigenvalue problem, where the Hamiltonian-like matrix H is first transformed to a symplectic pair with special structure; then $E_k = E_k^{\mathsf{H}}$ and $F_k = F_k^{\mathsf{T}}$ are computed iteratively. Theorems are proved on the quadratic convergence of the algorithm if no purely imaginary eigenvalues exist (and linear convergence if the partial multiplicities of purely imaginary eigenvalues are all even). The double-Cayley transform is designed to deal with any potential but generically rare breakdowns when 1 is in or close to $\sigma(F_k)$ for some k. We also prove that at most two steps of retrogression occur (for complex eigenvalues of H, but none for real ones). In addition, a three-recursion remedy is put forward when the double-Cayley transform fails. Numerical examples have been presented to illustrate the efficiency and the distinct structure-preserving nature of the doubling method. The optimal choice of α and the removal of the invertibility assumption of X_1 (or $[X_1, \Psi_{11}]$ if purely imaginary eigenvalues exist) will be left for future research.

Appendix A. Useful lemmas

The lemmas are required in Section 3.

Lemma 24. *Given $\omega, \zeta \in \mathbb{R}$, it holds that*

(a) $|\tanh(-\omega + \mathrm{i}\zeta)|^2 = |\tanh(\omega + \mathrm{i}\zeta)|^2 = [\mathrm{e}^{2\omega} + \mathrm{e}^{-2\omega} - 2\cos(2\zeta)][\mathrm{e}^{2\omega} + \mathrm{e}^{-2\omega} + 2\cos(2\zeta)]^{-1}$*;*

(b) $|\tanh(\omega + \mathrm{i}\zeta)|^2 < 1$ *when* $\cos(2\zeta) > 0$*; and*

(c) *for* $\cos(2\zeta) > 0$, $|\tanh(\omega + \mathrm{i}\zeta)|^2$ *is monotonically nondecreasing with respect to ω when $\omega \geq 0$, and monotonically nonincreasing otherwise.*

Proof. Simple computations lead to the two results (a) and (b), and we omit the details here. For (c), we have $\partial|\tanh(\omega+\mathrm{i}\zeta)|^2/\partial\omega = [8(\mathrm{e}^{2\omega}-\mathrm{e}^{-2\omega})\cos(2\zeta)][(\mathrm{e}^{2\omega}+\mathrm{e}^{-2\omega}+2\cos(2\zeta))^2]^{-1}$. Since $\cos(2\zeta)>0$, the result follows. □

Lemma 25. *Define* $f(\xi)=(\xi-\tau)^2+\xi^2$*; then for* $0\le\xi\le\frac{\tau}{2}$*, we have*

(a) $f(\xi)=f(\tau-\xi)$*;*
(b) $f(\xi)\ge f(\eta)\ge\frac{\tau}{\sqrt{2}}$ *for all* η *with* $\frac{\tau}{2}\ge\eta\ge\xi$*; and*
(c) $f(\xi)\ge f(\eta)\ge\frac{1}{\sqrt{2}}$ *for all* η *with* $\tau-\xi\ge\eta\ge\frac{\tau}{2}$*.*

Proof. From the fact that (ξ,ξ) and $(1-\xi,1-\xi)$ are two symmetrical points with respect to the line $g(\omega)=-\omega+\tau$, the result follows with details omitted. □

Lemma 26 ([12, Lemma 4.3])**.** *For the Jordan block* $J_{2p}(\mathrm{e}^{\mathrm{i}\theta})$*, it holds that* $J_{2p}^{2^k}(\mathrm{e}^{\mathrm{i}\theta})(1:p,p+1:2p)$ *are nonsingular for sufficiently large* k *and satisfy*

$$\|\big(J_{2p}^{2^k}(\mathrm{e}^{\mathrm{i}\theta})(1:p,p+1:2p)\big)^{-1}\big(J_p^{2^k}(\mathrm{e}^{\mathrm{i}\theta})\big)\|_2\le\frac{1}{2^k}\frac{\sqrt{p}(p+2)(p-1)!}{2(p-1)!-\sqrt{p}}=\mathcal{O}(2^{-k}),$$

$$\|\big(J_p^{2^k}(\mathrm{e}^{\mathrm{i}\theta})\big)\big(J_{2p}^{2^k}(\mathrm{e}^{\mathrm{i}\theta})(1:p,p+1:2p)\big)^{-1}\big(J_p^{2^k}(\mathrm{e}^{\mathrm{i}\theta})\big)\|_2$$
$$\le\frac{1}{2^k}\frac{\sqrt{p}(p+2)^2(p-1)!}{2(2(p-1)!-\sqrt{p})}=\mathcal{O}(2^{-k}).$$

Appendix B. Proofs for double-Cayley transform

This section contains some tedious details for the DCT.

B.1. Proof of Theorem 17.

Lemma 27. *Assume that the doubling iteration* (3.6) *does not break off for all* $k<k_0$*. If* E_0 *is nonsingular, so are* E_k $(0<k\le k_0)$*.*

Proof. This directly follows from $E_{k+1}=E_k(I_n-\overline{F}_kF_k)^{-1}E_k$ in (3.6). □

Obviously, Lemma 27 suggests that M_{k_0} and L_{k_0}, defined in (3.7), are both nonsingular and so is

$$L_{k_0}^{-1}M_{k_0}=\begin{bmatrix}E_{k_0}-\overline{F}_{k_0}\overline{E}_{k_0}^{-1}F_{k_0} & -\overline{F}_{k_0}\overline{E}_{k_0}^{-1}\\ \overline{E}_{k_0}^{-1}F_{k_0} & \overline{E}_{k_0}^{-1}\end{bmatrix}.$$

Since $L_{k_0}^{-1}M_{k_0}[X_1^{\mathsf{T}},\,X_2^{\mathsf{T}}]^{\mathsf{T}}=[X_1^{\mathsf{T}},\,X_2^{\mathsf{T}}]^{\mathsf{T}}S_\alpha^{2^{k_0}}$, the fact that $\{0,\alpha\}\not\subset\lambda(H)$ implies $L_{k_0}^{-1}M_{k_0}\pm I_{2n}$ are nonsingular. Consequently, we can prove Theorem 17 as follows.

Proof of Theorem 17. For (a) with $\vartheta\notin\lambda(E_{k_0})$, $E_{k_0}-\vartheta I_n$ is nonsingular and so is

$$K\triangleq\begin{bmatrix}I_n & \overline{F}_{k_0}(I_n-\vartheta\overline{E}_{k_0})^{-1}\\ 0 & (\overline{E}_{k_0}^{-1}-\vartheta I_n)^{-1}\end{bmatrix}.$$

In addition, pre-multiplying $L_{k_0}^{-1}M_{k_0}$ by K gives

$$K(L_{k_0}^{-1}M_{k_0}-\vartheta I_{2n})=\begin{bmatrix}E_{k_0}-\vartheta I_n+\vartheta\overline{F}_{k_0}(I_n-\vartheta\overline{E}_{k_0})^{-1}F_{k_0} & 0\\ (I_n-\vartheta\overline{E}_{k_0})^{-1}F_{k_0} & I_n\end{bmatrix},$$

implying that $Z=\vartheta I_n-E_{k_0}+\vartheta\overline{F}_{k_0}(\vartheta\overline{E}_{k_0}-I_n)^{-1}F_{k_0}$ is nonsingular.

For (b), manipulations show that $\widehat{H} = \beta\vartheta(L_{k_0}^{-1}M_{k_0} - \vartheta I_n)^{-1}(L_{k_0}^{-1}M_{k_0} + \vartheta I_n)$. Then $M_{k_0}\ [X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} = L_{k_0}[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} S_\alpha^{2^{k_0}}$ implies

$$(L_{k_0}^{-1}M_{k_0}-\vartheta I_n)^{-1}(L_{k_0}^{-1}M_{k_0}+\vartheta I_n)[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} = [X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}}(S_\alpha^{2^{k_0}}-\vartheta I_n)^{-1}(S_\alpha^{2^{k_0}}+\vartheta I_n),$$

leading to $\widehat{H}[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} = [X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}}[\beta\vartheta(S_\alpha^{2^{k_0}} - \vartheta I_n)^{-1}(S_\alpha^{2^{k_0}} + \vartheta I_n)]$. The result follows from the resulting equalities

$$\begin{aligned}(\widehat{H} + \beta\vartheta I_n)[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} &= [X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}}[2\beta\vartheta(S_\alpha^{2^{k_0}} - \vartheta I_n)^{-1}S_\alpha^{2^{k_0}}],\\ (\widehat{H} - \beta\vartheta I_n)[X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}} &= [X_1^{\mathsf{T}},\, X_2^{\mathsf{T}}]^{\mathsf{T}}[2\beta(S_\alpha^{2^{k_0}} - \vartheta I_n)^{-1}].\end{aligned}$$

For (c), $\widehat{A}^{\mathsf{H}} = \widehat{A}$ directly follows from its definition and the facts that $E_{k_0}^{\mathsf{H}} = E_{k_0}$ and $F_{k_0}^{\mathsf{T}} = F_{k_0}$. For the symmetry of $\widehat{B}$, observe that

$$\begin{aligned}\widehat{B} &= 2\beta\vartheta Z^{-1}\overline{F}_{k_0}(\overline{E}_{k_0} - \vartheta I_n)^{-1}\\ &= 2\beta\vartheta(E_{k_0}-\vartheta I_n)^{-1}[I_n+\vartheta\overline{F}_{k_0}(\vartheta\overline{E}_{k_0}-I_n)^{-1}F_{k_0}(\vartheta I_n-E_{k_0})^{-1}]^{-1}\overline{F}_{k_0}(\vartheta I_n-\overline{E}_{k_0})^{-1}\\ &= 2\beta\vartheta(E_{k_0}-\vartheta I_n)^{-1}\overline{F}_{k_0}[I_n+\vartheta(\vartheta\overline{E}_{k_0}-I_n)^{-1}F_{k_0}(\vartheta I_n-E_{k_0})^{-1}\overline{F}_{k_0}]^{-1}(\vartheta I_n-\overline{E}_{k_0})^{-1}\\ &= 2\beta\vartheta(E_{k_0} - \vartheta I_n)^{-1}\overline{F}_{k_0}\overline{Z}^{-1} = 2\beta\vartheta(E_{k_0} - \vartheta I_n)^{-1}\overline{F}_{k_0}Z^{-\mathsf{T}} = \widehat{B}^{\mathsf{T}}.\end{aligned}$$

The proof is complete. □

B.2. **Proof of Lemma 18.**

Proof. Let $\xi + \mathrm{i}\eta = \varrho = \delta_\lambda^{2^{k_0}}$. We then have $|\varrho| = |\delta_\lambda|^{2^{k_0}}$ and $|\xi| \le |\delta_\lambda|^{2^{k_0}}$. Consequently, from the definition of ν we deduce that

$$\begin{aligned}|\nu|^2 &= \frac{(\xi^2+\eta^2)(\beta\vartheta+\gamma)^2 + (\beta-\vartheta\gamma)^2 + 2\vartheta\xi(\beta^2-\gamma^2)}{(\beta\vartheta+\gamma)^2 + (\beta-\vartheta\gamma)^2(\xi^2+\eta^2) + 2\vartheta\xi(\beta^2-\gamma^2)}\\ &= \frac{|\delta_\lambda|^{2^{k_0+1}} + 2\xi\varpi + \varpi^2}{|\delta_\lambda|^{2^{k_0+1}}\varpi^2 + 2\xi\varpi + 1}. \qquad \text{(B.1)}\end{aligned}$$

Since $\vartheta\beta, \gamma > 0$ and the function defined in (B.1) is (i) monotonically nondecreasing with respect to ξ when $\beta > \vartheta\gamma$ or (ii) monotonically nonincreasing otherwise, we obtain

$$|\nu|^2 \le \begin{cases} \frac{|\delta_\lambda|^{2^{k_0}}(|\delta_\lambda|^{2^{k_0}}+2\varpi)+\varpi^2}{|\delta_\lambda|^{2^{k_0}}(2\varpi+|\delta_\lambda|^{2^{k_0}}\varpi^2)+1} & \text{if} \quad \beta > \vartheta\gamma;\\[2ex] \frac{|\delta_\lambda|^{2^{k_0}}(|\delta_\lambda|^{2^{k_0}}-2\varpi)+\varpi^2}{|\delta_\lambda|^{2^{k_0}}(-2\varpi+|\delta_\lambda|^{2^{k_0}}\varpi^2)+1} & \text{if} \quad \beta < \vartheta\gamma; \end{cases}$$

which is equivalent to

$$|\nu|^2 \le \frac{|\delta_\lambda|^{2^{k_0}}(|\delta_\lambda|^{2^{k_0}} + 2|\varpi|) + \varpi^2}{|\delta_\lambda|^{2^{k_0}}(2|\varpi| + |\delta_\lambda|^{2^{k_0}}\varpi^2) + 1} = \left(\frac{|\delta_\lambda|^{2^{k_0}} + |\varpi|}{|\delta_\lambda|^{2^{k_0}}|\varpi| + 1}\right)^2.$$

Obviously, $(|\delta_\lambda|^{2^{k_0}} + |\varpi|)(|\delta_\lambda|^{2^{k_0}}|\varpi| + 1)^{-1} < 1$ from $|\varpi| = |\beta - \vartheta\gamma|/(\vartheta\beta + \gamma) < 1$ and $|\delta_\lambda| < 1$, thus the result follows. □

B.3. **Proof of Theorem 19.**

Proof. Denote $\xi + \mathrm{i}\eta = \delta_\lambda^{2^{k_0-1}}$ and define

$$\begin{aligned}\phi &= \operatorname{arctanh} \delta_\lambda^{2^{k_0-1}} \\ &= \frac{1}{2}\ln\left|\frac{(\lambda-\alpha)^{2^{k_0-1}}+(\lambda+\alpha)^{2^{k_0-1}}}{(\lambda-\alpha)^{2^{k_0-1}}-(\lambda+\alpha)^{2^{k_0-1}}}\right| + \frac{\mathrm{i}}{2}\arg\left[\frac{(\lambda-\alpha)^{2^{k_0-1}}+(\lambda+\alpha)^{2^{k_0-1}}}{(\lambda-\alpha)^{2^{k_0-1}}-(\lambda+\alpha)^{2^{k_0-1}}}\right], \\ \psi &= \operatorname{arctanh} \kappa^{-2^{k_0-1}} = \frac{1}{2}\left[\ln(1+\sqrt{|\varpi|}) - \ln(1-\sqrt{|\varpi|})\right].\end{aligned}$$

We deduce that

$$\arg\left[\frac{(\lambda-\alpha)^{2^{k_0-1}}+(\lambda+\alpha)^{2^{k_0-1}}}{(\lambda-\alpha)^{2^{k_0-1}}-(\lambda+\alpha)^{2^{k_0-1}}}\right] = \arctan\frac{2\eta}{1-\xi^2-\eta^2} \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

Specifically, $\arg\left[\frac{(\lambda-\alpha)^{2^{k_0-1}}+(\lambda+\alpha)^{2^{k_0-1}}}{(\lambda-\alpha)^{2^{k_0-1}}-(\lambda+\alpha)^{2^{k_0-1}}}\right] = 0$ when $\lambda \in \mathbb{R}$. Moreover, by the definitions of ϕ and ψ, routine manipulations show that

$$\nu = \vartheta \tanh(\phi - \psi)\tanh(\phi + \psi)$$

with

$$\phi \pm \psi = \frac{1}{2}\ln\left[\frac{\sqrt{\gamma+\vartheta\beta} \pm \sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta} \mp \sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}}\right] + \frac{\mathrm{i}}{2}\arctan\frac{2\eta}{1-\xi^2-\eta^2}.$$

With $\gamma = \beta\frac{\kappa^{2^{k_0}}+\vartheta}{\vartheta\kappa^{2^{k_0}}-1}$ and $\cos\left(\arctan\frac{2\eta}{1-\xi^2-\eta^2}\right) > 0$, we have

$$\begin{cases} \ln\left(\frac{\sqrt{\gamma+\vartheta\beta}+\sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta}-\sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}}\right) \geq 0, & \text{if } \frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2} \geq 1; \\ \ln\left(\frac{\sqrt{\gamma+\vartheta\beta}-\sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta}+\sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}}\right) < 0 & \text{otherwise.}\end{cases}$$

From Lemma 24 and $[(1+\xi)^2+\eta^2][(1-\xi)^2+\eta^2]^{-1} \geq 1 \Leftrightarrow \xi \geq 0$, we obtain

$$|\nu| < \begin{cases} \big|\tanh(\phi-\psi)\big|, & \text{if } \xi > 0; \\ \big|\tanh(\phi+\psi)\big|, & \text{if } \xi < 0.\end{cases}$$

Now assume that $\xi > 0$ and consider two distinct cases.

(i) When

$$\sqrt{\frac{(1-\xi)^2+\eta^2}{(1+\xi)^2+\eta^2}} \leq \frac{\sqrt{\gamma+\vartheta\beta}-\sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta}+\sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}} < 1$$

or

$$\frac{\sqrt{\gamma+\vartheta\beta}-\sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta}+\sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}} \geq 1,$$

we have

$$\ln\left[\sqrt{\frac{(1-\xi)^2+\eta^2}{(1+\xi)^2+\eta^2}}\right] \leq \ln\left[\frac{\sqrt{\gamma+\vartheta\beta}-\sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta}+\sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}}\right] < 0$$

or

$$0 < \ln\left[\frac{\sqrt{\gamma+\vartheta\beta}-\sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta}+\sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}}\right] < \ln\left[\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}}\right].$$

Hence by (c) and (b) in Lemma 24, it is apparent that

$$
\begin{aligned}
|\nu|^2 &< |\tanh(\phi-\psi)|^2\\
&\le \left|\tanh\left\{\frac{1}{2}\ln\left[\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}}\right]+\frac{\mathrm{i}}{2}\arctan\frac{2\eta}{1-\xi^2-\eta^2}\right\}\right|^2\\
&= |\tanh(\phi)|^2 = |\delta_\lambda|^{2^{k_0}},
\end{aligned}
$$

implying that $|\nu|<|\delta_\lambda|^{2^{k_0-1}}$.

(ii) When

$$
\frac{\sqrt{\gamma+\vartheta\beta}-\sqrt{\vartheta\gamma-\beta}}{\sqrt{\gamma+\vartheta\beta}+\sqrt{\vartheta\gamma-\beta}}\sqrt{\frac{(1+\xi)^2+\eta^2}{(1-\xi)^2+\eta^2}} < \sqrt{\frac{(1-\xi)^2+\eta^2}{(1+\xi)^2+\eta^2}} < 1,
$$

we define $\widehat{\xi}+\mathrm{i}\widehat{\eta}=\delta_\lambda^{2^{k_0-2}}$ and without loss of generality assume that $\widehat{\xi}>0$, which satisfies $\widehat{\xi}>|\widehat{\eta}|$ for $0<\xi=\widehat{\xi}^2-\widehat{\eta}^2$. Similar to (i), we obtain

$$
|\nu| < |\tanh(\phi-\psi)| = |\tanh(\widehat{\phi}-\widehat{\psi})\tanh(\widehat{\phi}+\widehat{\psi})| < |\tanh(\widehat{\phi}-\widehat{\psi})|,
$$

where $\widehat{\phi}=\operatorname{arctanh}\delta_\lambda^{2^{k_0-2}}$ and $\widehat{\psi}=\operatorname{arctanh}\kappa^{-2^{k_0-2}}$. Since $\xi=\widehat{\xi}^2-\widehat{\eta}^2>0$ and $|\widehat{\xi}|^2+|\widehat{\eta}|^2=|\delta_\lambda|^{2^{k_0-1}}$, we have $\widehat{\xi}^2>\frac{1}{2}|\delta_\lambda|^{2^{k_0-1}}$, leading to

$$
\begin{aligned}
|\nu|^2 &< |\tanh(\widehat{\phi}-\widehat{\psi})|^2\\
&= \frac{\dfrac{\kappa^{2^{k_0-2}}-1}{\kappa^{2^{k_0-2}}+1}\cdot\dfrac{1+|\delta_\lambda|^{2^{k_0-1}}+2\widehat{\xi}}{1-|\delta_\lambda|^{2^{k_0-1}}}+\dfrac{\kappa^{2^{k_0-2}}+1}{\kappa^{2^{k_0-2}}-1}\cdot\dfrac{1+|\delta_\lambda|^{2^{k_0-1}}-2\widehat{\xi}}{1-|\delta_\lambda|^{2^{k_0-1}}}-2}{\dfrac{\kappa^{2^{k_0-2}}-1}{\kappa^{2^{k_0-2}}+1}\cdot\dfrac{1+|\delta_\lambda|^{2^{k_0-1}}+2\widehat{\xi}}{1-|\delta_\lambda|^{2^{k_0-1}}}+\dfrac{\kappa^{2^{k_0-2}}+1}{\kappa^{2^{k_0-2}}-1}\cdot\dfrac{1+|\delta_\lambda|^{2^{k_0-1}}-2\widehat{\xi}}{1-|\delta_\lambda|^{2^{k_0-1}}}+2}.
\end{aligned}
$$

Since $|\tanh(\widehat{\phi}-\widehat{\psi})|^2$ is monotonically nonincreasing with respect to $\widehat{\xi}$, taking $\widehat{\xi}=\frac{1}{\sqrt{2}}|\delta_\lambda|^{2^{k_0-2}}$ in the above formula yields

$$
|\nu|^2 < |\tanh(\widehat{\phi}-\widehat{\psi})|^2 < \frac{1+|\delta_\lambda|^{2^{k_0-1}}\kappa^{2^{k_0-1}}-\sqrt{2}\kappa^{2^{k_0-2}}|\delta_\lambda|^{2^{k_0-2}}}{\kappa^{2^{k_0-1}}+|\delta_\lambda|^{2^{k_0-1}}-\sqrt{2}\kappa^{2^{k_0-2}}|\delta_\lambda|^{2^{k_0-2}}}
$$

$$
= \kappa^{-2^{k_0-1}}\cdot\left[\frac{(2^{-1/2}\kappa^{2^{k_0-2}}|\delta_\lambda|^{2^{k_0-2}}-1)^2+2^{-1}\kappa^{2^{k_0-1}}|\delta_\lambda|^{2^{k_0-1}}}{(2^{-1/2}\kappa^{-2^{k_0-2}}|\delta_\lambda|^{2^{k_0-2}}-1)^2+2^{-1}\kappa^{-2^{k_0-1}}|\delta_\lambda|^{2^{k_0-1}}}\right] \tag{B.2}
$$

$$
= |\delta_\lambda|^{2^{k_0-1}}\cdot\left[\frac{(\kappa^{2^{k_0-2}}-2^{-1/2}|\delta_\lambda|^{-2^{k_0-2}})^2+2^{-1}|\delta_\lambda|^{-2^{k_0-1}}}{(\kappa^{2^{k_0-2}}-2^{-1/2}|\delta_\lambda|^{2^{k_0-2}})^2+2^{-1}|\delta_\lambda|^{2^{k_0-1}}}\right]. \tag{B.3}
$$

Obviously for $\kappa\ge 2$, we obtain $(\kappa^{-1}|\delta_\lambda|)^{2^{k_0-2}}<1/2$. Hence, by Lemma 25, when either

(a) $2^{-1/2}\,(|\delta_\lambda|\kappa)^{2^{k_0-2}}\le\frac{1}{2}$, i.e., $(|\delta_\lambda|\kappa)^{2^{k_0-2}}\le 1/\sqrt{2}$; or

(b) $\frac{1}{2}<2^{-1/2}\,(|\delta_\lambda|\kappa)^{2^{k_0-2}}\le 1-2^{-1/2}|\delta_\lambda|^{2^{k_0-2}}\kappa^{-2^{k_0-2}}$, i.e.,

$$
(|\delta_\lambda|\kappa)^{2^{k_0-2}}\ge 1/\sqrt{2}, \qquad |\delta_\lambda|^{2^{k_0-2}}(\kappa^{2^{k_0-2}}+\kappa^{-2^{k_0-2}})\le\sqrt{2},
$$

the quantity in the square brackets in (B.2) would be no greater than 1. This indicates that $|\nu|^2\le\kappa^{-2^{k_0-1}}$ or $|\nu|<\kappa^{-2^{k_0-2}}$.

When

$$(|\delta_\lambda|\kappa)^{2^{k_0-2}} \geq 1/\sqrt{2}, \qquad |\delta_\lambda|^{2^{k_0-2}}(\kappa^{2^{k_0-2}} + \kappa^{-2^{k_0-2}}) > \sqrt{2},$$

which imply $|\delta_\lambda|^{2^{k_0-2}} > \sqrt{2}/(\kappa^{2^{k_0-2}} + \kappa^{-2^{k_0-2}})$, we obtain

$$\text{(B.4)} \qquad |\delta_\lambda|^{2^{k_0-2}} + |\delta_\lambda|^{-2^{k_0-2}} < \frac{\sqrt{2}\kappa^{2^{k_0-2}}}{\kappa^{2^{k_0-1}}+1} + \frac{\kappa^{2^{k_0-1}}+1}{\sqrt{2}\kappa^{2^{k_0-2}}} < \sqrt{2}\kappa^{2^{k_0-2}},$$

where the first "<" follows from the fact that the function $f(x) = x + x^{-1}$ is monotonically decreasing when $x < 1$. Thus, the assumption $\kappa \geq 2$ and (B.4) together affirm that $2^{-1/2}|\delta_\lambda|^{2^{k_0-2}} < 2^{-1}\kappa^{2^{k_0-2}}$ and $2^{-1/2}|\delta_\lambda|^{-2^{k_0-2}} \leq \kappa^{2^{k_0-2}} - 2^{-1/2}|\delta_\lambda|^{2^{k_0-2}}$. Again using Lemma 25, we know that the quantity in the square brackets in (B.3) is no greater than 1, suggesting that the value of the right-hand side of (B.3) will be no greater than $|\delta_\lambda|^{2^{k_0-1}}$, or equivalently $|\nu| < |\delta_\lambda|^{2^{k_0-2}}$.

Consequently, the result holds for the case when $\xi > 0$. The $\xi < 0$ case can be proved similarly and we omit the details. □

Acknowledgment

We thank Professor Ren-Cang Li for his kindness in providing three test problems in Example 4.1.

References

[1] Z. Bai and R.-C. Li, *Minimization principles for the linear response eigenvalue problem I: Theory*, SIAM J. Matrix Anal. Appl. **33** (2012), no. 4, 1075–1100, DOI 10.1137/110838960. MR3023465

[2] Z. Bai and R.-C. Li, *Minimization principles for the linear response eigenvalue problem II: Computation*, SIAM J. Matrix Anal. Appl. **34** (2013), no. 2, 392–416, DOI 10.1137/110838972. MR3046810

[3] P. Benner, S. Dolgov, V. Khoromskaia, and B. N. Khoromskij, *Fast iterative solution of the Bethe-Salpeter eigenvalue problem using low-rank and QTT tensor approximation*, J. Comput. Phys. **334** (2017), 221–239, DOI 10.1016/j.jcp.2016.12.047. MR3606226

[4] P. Benner, H. Fassbender, and C. Yang, *Some remarks on the complex J-symmetric eigenvalue problem*, Preprint, MPIMD/15-12, Max Planck Institute Magdeburg, 2015 (available at `www.mpi-magdeburg.mpg.de/preprints`).

[5] P. Benner, V. Khoromskaia, and B. N. Khoromskij, *A reduced basis approach for calculation of the Bethe-Salpeter excitation energies using low-rank tensor factorizations*, Molecular Phys., **114** (2016) 1148–1161.

[6] M. E. Casida, *Time-dependent density-functional response theory for molecules*, Recent Advances in Density Functional Methods, Part I, D.P. Chong (Ed.), World Scientific, Singapore, 155 (1995) 1207–1216.

[7] E. K.-W. Chu, H.-Y. Fan, and W.-W. Lin, *A structure-preserving doubling algorithm for continuous-time algebraic Riccati equations*, Linear Algebra Appl. **396** (2005), 55–80, DOI 10.1016/j.laa.2004.10.010. MR2112199

[8] E. K.-W. Chu, H.-Y. Fan, W.-W. Lin, and C.-S. Wang, *Structure-preserving algorithms for periodic discrete-time algebraic Riccati equations*, Internat. J. Control **77** (2004), no. 8, 767–788, DOI 10.1080/00207170410001714988. MR2072208

[9] E. K.-W. Chu, T.-M. Hwang, W.-W. Lin, and C.-T. Wu, *Vibration of fast trains, palindromic eigenvalue problems and structure-preserving doubling algorithms*, J. Comput. Appl. Math. **219** (2008), no. 1, 237–252, DOI 10.1016/j.cam.2007.07.016. MR2437709

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996. MR1417720

[11] T.-M. Huang, R.-C. Li, W.-W. Lin, and L. Lu., *Optimal parameters for doubling algorithms*, Technical Report 2017-03, Department of Mathematics, University of Texas at Arlington, May 2017. Available at http://www.uta.edu/math/preprint/.

[12] T.-M. Huang and W.-W. Lin, *Structured doubling algorithms for weakly stabilizing Hermitian solutions of algebraic Riccati equations*, Linear Algebra Appl. **430** (2009), no. 5-6, 1452–1478, DOI 10.1016/j.laa.2007.08.043. MR2490689

[13] V. Khoromskaia, B. N. Khoromskij, and R. Schneider, *Tensor-structured factorized calculation of two-electron integrals in a general basis*, SIAM J. Sci. Comput. **35** (2013), no. 2, A987–A1010, DOI 10.1137/120884067. MR3040965

[14] S. Körbel, P. Boulanger, I. Duchemin, X. Blase, M. AL Marques, and S. Botti, *Benchmark many-body GW and Bethe-Salpeter calculations for small transition metal molecules*, J. Chemical Theory Comp., **10** (2014) 3934–3943.

[15] X. Leng, F. Jin, M. Wei, and Y. Ma, *GW method and Bethe-Salpeter equation for calculating electronic excitations*, Wiley Interdisciplinary Reviews: Computation Molecular Science, Wiley Online Library, 2016.

[16] T. Li, C.-Y. Chiang, E. K.-w. Chu, and W.-W. Lin, *The palindromic generalized eigenvalue problem $A^*x = \lambda Ax$: numerical solution and applications*, Linear Algebra Appl. **434** (2011), no. 11, 2269–2284, DOI 10.1016/j.laa.2009.12.020. MR2776795

[17] T. Li, E. K.-w. Chu, J. Juang, and W.-W. Lin, *Solution of a nonsymmetric algebraic Riccati equation from a one-dimensional multistate transport model*, IMA J. Numer. Anal. **31** (2011), no. 4, 1453–1467, DOI 10.1093/imanum/drq034. MR2846762

[18] T. Li, E. K.-w. Chu, J. Juang, and W.-W. Lin, *Solution of a nonsymmetric algebraic Riccati equation from a two-dimensional transport model*, Linear Algebra Appl. **434** (2011), no. 1, 201–214, DOI 10.1016/j.laa.2010.09.006. MR2737242

[19] 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), no. 1, 26–39, DOI 10.1137/040617650. MR2218940

[20] G. Onida, L. Reining, and A. Rubio, *Electronic excitations: density-functional versus many-body Green's-function approaches*, Rev. Mod. Phys., **74** (2002) 601–659.

[21] R.M. Parrish, E.G. Hohenstein, N. Schunck, C. Sherrill, and T. J. Martinez, *Exact tensor hypercontraction: A universal technique for the resolution of matrix elements of local finite-range N-body potentials in many-body quantum problems*, Phys. Rev. Lett., **111** (2013) 132505.

[22] Y. Ping, D. Rocca, and G. Galli, *Electronic excitations in light absorbers for photoelectrochemical energy conversion: First principles calculations based on many body perturbation theory*, Chem. Soc. Rev., **42** (2013) 2437–2469.

[23] S. Reine, T. Helgaker, and R. Lindh, *Multi-electron integrals*, WIREs Comput. Mol. Sci., **2** (2012) 290–303.

[24] L. Reining, V. Olevano, A. Rubio, and G. Onida, *Excitonic effects in solids described by time-dependent density functional theory*, Phys. Rev. Lett., **88** (2002) 066404.

[25] E. Ribolini, J. Toulouse, and A. Savin, *Electronic excitation energies of molecular systems from the Bethe-Salpeter equation: Example of H2 molecule*, Concepts and Methods in Modern Theoretical Chemistry, S. Ghosh and P. Chattaraj (eds), Vol. 1: Electronic Structure and Reactivity, 367 (2013) 367–390.

[26] E. Ribolini, J. Toulouse, and A. Savin, *Electronic excitations from a linear-response range-separated hybrid scheme*, Molecular Phys., **111** (2013) 1219–1234.

[27] M. Rohlfing and S. G. Louie, *Electron-hole excitations and optical spectra from first principles*. Phys. Rev. B, **62** (2000) 4927–4944.

[28] E. Runge and E. Gross, *Density-function theory for time-dependent systems*, Phys. Rev. Lett., **52** (1984) 997–1000.

[29] E. E. Salpeter and H. A. Bethe, *A relativistic equation for bound-state problems*, Physical Rev. (2) **84** (1951), 1232–1242. MR0052996

[30] M. Shao, F. H. da Jornada, C. Yang, J. Deslippe, and S. G. Louie, *Structure preserving parallel algorithms for solving the Bethe-Salpeter eigenvalue problem*, Linear Algebra Appl. **488** (2016), 148–167, DOI 10.1016/j.laa.2015.09.036. MR3419779

[31] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Computer Science and Scientific Computing, Academic Press, Inc., Boston, MA, 1990. MR1061154

[32] R. E. Stratmann, G. E. Scuseria, and M. J. Frisch, *An efficient implementation of time-dependent density-functional theory for the calculation of excitation energies of large molecules*, J. Chem. Phys., **109** (1998) 8218–8224.

[33] S. Wilson, *Universal basis sets and Cholesky decomposition of the two-electron integral matrix*, Comput. Phys. Commun., **58** (1990) 71–81.

Department of Mathematics, Nanjing University, Nanjing 210093, People's Republic of China
Email address: guozhch06@gmail.com

School of Mathematics, Monash University, 9 Rainforest Walk, Victoria 3800, Australia
Email address: eric.chu@monash.edu

Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan
Email address: wwlin@math.nctu.edu.tw