CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.2021.nla.05

Vol. **2**, No. 2, pp. 336-356
June 2021

Note on Finding an Optimal Deflation for Quadratic Matrix Polynomials

Xin Liang*

Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China.

Received 14 May 2020; Accepted 6 October 2020

Abstract. This paper is concerned with the way to find an optimal deflation for the eigenvalue problem associated with quadratic matrix polynomials. This work is a response of the work by Tisseur et al., *Linear Algebra Appl., 435:464-479, 2011*, and solves one of open problems raised by them. We build an equivalent unconstrained optimization problem on eigenvalues of a hyperbolic quadratic matrix polynomial of order 2, and develop a technique that transforms the quadratic matrix polynomial to an equivalent one that is easy to solve. Numerical tests are given to illustrate several properties of the problem.

AMS subject classifications: 65F35, 65F15

Key words: deflation, quadratic matrix polynomials, hyperbolic, eigenvalue optimization.

1 Introduction

Given a quadratic matrix polynomial

$$Q(\lambda)=\lambda^2 M+\lambda C+K,$$

where $M,C,K\in\mathbb{R}^{n\times n}$ with M nonsingular. Its associated quadratic eigenvalue problem is

$$Q(\lambda)x=0,\quad y^{\mathrm{H}}Q(\lambda)=0,$$

where λ is an eigenvalue and x,y are its corresponding (right) eigenvector and left eigenvector respectively. An eigenvalue is of positive type, if $y^{\mathrm{H}}Q'(\lambda)x=y^{\mathrm{H}}(2\lambda M+C)x>0$; An eigenvalue is of negative type, if $y^{\mathrm{H}}Q'(\lambda)x=y^{\mathrm{H}}(2\lambda M+C)x<0$.

Suppose that $\lambda(Q)$, the spectra of $Q(\lambda)$, is $\{\lambda_1,\cdots,\lambda_{2n}\}$. Deflating two distinct eigenvalues λ_1,λ_2 is to construct a new quadratic matrix polynomial

$$\widetilde{Q}(\lambda)=\begin{bmatrix}Q_d(\lambda) & \\ & q(\lambda)\end{bmatrix}=\lambda^2\begin{bmatrix}M_d & \\ & m\end{bmatrix}+\lambda\begin{bmatrix}C_d & \\ & c\end{bmatrix}+\begin{bmatrix}K_d & \\ & k\end{bmatrix}$$

*Corresponding author. *Email address:* liangxinslm@tsinghua.edu.cn (X. Liang)

such that $\lambda(q)=\{\lambda_1,\lambda_2\}$, $\lambda(Q_d)=\{\lambda_3,\cdots,\lambda_{2n}\}$. Usually, in the two eigenvalues, one is of positive type, and the other is of negative type. If M,C,K are symmetric, and $\lambda_1\in\lambda(Q)$ but λ_1 is nonreal, then $\overline{\lambda_1}\in\lambda(Q)$, and in this case, it is usually required to deflate this conjugate pair together.

The deflation technique is very useful and popular in computing the eigenvalues of a matrix, so that it is hoped to be used for computing the eigenvalues of quadratic matrix polynomials. However, as far as we know, not many works discussed on this topic. Meini [5] discussed a deflation method coupled in her so-called "shift-and-deflate" technique. Tisseur et al. [6] presented a general way to deflate two distinct eigenvalues of quadratic matrix polynomials if the corresponding eigenvectors are given.

Here we briefly describe the idea of the method introduced by Tisseur et al. First they developed a method to deflate a quadratic polynomial for two given eigenvalues whose eigenvectors are parallel. Then they invented a way to transform a quadratic polynomial for two given eigenvalues whose eigenvectors are nonparallel into a new one that has two eigenvalues whose eigenvectors are parallel, i.e., transform this case into the solved case.

The new quadratic matrix polynomial produced by the deflation may have a significantly large condition number compared to the original quadratic matrix polynomial. For the special case that M,C,K are symmetric, Tisseur et al. gave an optimal choice to minimize the condition number for the parallel case, but for the nonparallel case, in [6, Section 3], they reported:

> Identifying which solution minimizes the condition number $\kappa_2(T)=\|T\|_2\|\|T^{-1}\|_2$ remains an open problem.

Here T is a related transformation matrix, of which the detailed form will be given below.

The aim of this paper is to solve this problem. First this problem is formulated and simplified in Section 2, which induces a constrained optimization problem on the eigenvalues of a hyperbolic quadratic matrix polynomial of order 2. Next we parameterize (or equivalently nondimensionalize) it and obtain an unconstrained optimization problem in Section 3. Then we calculate the gradient and the Hessian matrix of the objective function in Section 4. Then we make several numerical tests to show the properties of this problem and suggest a technique that transforms it to an equivalent problem whose objective function is easy to solve, as is shown in Section 5. Finally some concluding remarks is given in Section 6.

Notation. Throughout this paper, I_n (or simply I if its dimension is clear from the context) is the $n\times n$ identity matrix. For any scalar, vector, or matrix X, $\Re X$ and $\Im X$ are its real part and imaginary part respectively; while $\|X\|_2$ and $\|X\|_\infty$ are its spectral norm and sum-of-row norm. For any matrix X, $\lambda(X)$ represents its spectra, and $\lambda_*(X)$ represents the set consisting of all its nonzero eigenvalues. For any real symmetric matrix X, $X\succ 0$ ($X\succeq 0$) means that X is positive (semi-)definite, and $X\prec 0$ ($X\preceq 0$) if $-X\succ 0$ ($-X\succeq 0$).

By $X\otimes Y$ we denote the Kronecker product of two matrices X,Y. For any set S, $S^{\perp}$ is its orthogonal complement.

2 Simplify the problem

First we state the problem.

Given a symmetric quadratic matrix polynomial $Q(\lambda)=\lambda^2M+\lambda C+K$ with M nonsingular, where (λ_1,x_1) and (λ_2,x_2) are its two eigenpairs of opposite types. Write

$$(\Lambda,X)=\begin{cases}\left(\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix},\begin{bmatrix}x_1 & x_2\end{bmatrix}\right), & \text{if } \lambda_1 \text{ and } \lambda_2 \text{ are real,}\\ \left(\begin{bmatrix}\Re\lambda_1 & \Im\lambda_1\\ -\Im\lambda_1 & \Re\lambda_1\end{bmatrix},\begin{bmatrix}\Re x_1 & \Im x_1\end{bmatrix}\right), & \text{if } \lambda_1=\overline{\lambda_2} \text{ are nonreal and } x_1=\overline{x_2}.\end{cases}$$

If x_1,x_2 are nonparallel vectors, then the transformation T is

$$T=I_{2n}+\begin{bmatrix}ab^{\mathrm{T}} & ad^{\mathrm{T}}\\ af^{\mathrm{T}} & ah^{\mathrm{T}}\end{bmatrix},$$

where

$$a=\frac{Xp}{\|Xp\|_2},\qquad \begin{bmatrix}b & f & d & h\end{bmatrix}=\left(I_n-\frac{zz^{\mathrm{T}}}{z^{\mathrm{T}}z}\right)BA^{+}+U(I_4-AA^{+})+\frac{zw^{\mathrm{T}}}{z^{\mathrm{T}}z}=:V\in\mathbb{R}^{n\times 4}.$$

Here U can be any matrix with $z^{\mathrm{T}}U=0$, and

$$A=\frac{1}{2}\begin{bmatrix}2\alpha_M & \alpha_C & 0\\ \alpha_C & 2\alpha_K & 0\\ 0 & 2\alpha_M & \alpha_C\\ 0 & \alpha_C & 2\alpha_K\end{bmatrix},\qquad B=-\begin{bmatrix}Ma & Ca & Ka\end{bmatrix},$$

where $\alpha_M,\alpha_C,\alpha_K$ are $a^{\mathrm{T}}Ma,a^{\mathrm{T}}Ca,a^{\mathrm{T}}Ka$ respectively. Also $z\in\mathbb{R}^n$, $w\in\mathbb{R}^4$ are given by

$$z=X\Lambda p-\frac{e_\ell^{\mathrm{T}}X\Lambda p-1}{e_\ell^{\mathrm{T}}a}a,\qquad w=\frac{1}{e_\ell^{\mathrm{T}}a}\begin{bmatrix}e_\ell^{\mathrm{T}}X\Lambda p-1\\ e_\ell^{\mathrm{T}}a\|Xp\|_2\\ e_\ell^{\mathrm{T}}X\Lambda q\\ e_\ell^{\mathrm{T}}Xq-1\end{bmatrix},$$

where e_ℓ is a column of I_n that makes $|e_\ell^{\mathrm{T}}a|=\|a\|_\infty$, and

$$p=\frac{\gamma}{\lambda_1-\lambda_2}\begin{bmatrix}1\\ -1\end{bmatrix},\quad \gamma=1 \text{ or i to make } p \text{ real},\qquad q=\Lambda p-(\lambda_1+\lambda_2)p.$$

The optimization problem to be solved is

$$\min_{U\in\mathbb{R}^{n\times 4}:z^{\mathrm{T}}U=0}\kappa_2(T). \tag{2.1}$$

Besides, a,V,A,B,z,w have the properties below, which has been shown in [6]:

$$z^{\mathrm{T}}V=w^{\mathrm{T}},\quad VA=B,\quad w^{\mathrm{T}}A=z^{\mathrm{T}}B; \tag{2.2}$$

$$a^{\mathrm{T}}B+a_0^{\mathrm{T}}A=0\quad\text{where}\quad a_0=\begin{bmatrix}1&0&0&1\end{bmatrix}^{\mathrm{T}}; \tag{2.3}$$

$$A\text{ has full column rank.} \tag{2.4}$$

In the following, we will simplify the problem.

Define a linear mapping

$$\begin{aligned}\mathrm{rs}:\qquad \mathbb{R}^{n\times 2m}&\to\mathbb{R}^{2n\times m}\\ Z=\begin{bmatrix}Z_1&Z_2\end{bmatrix}&\mapsto\mathrm{rs}(Z)=\begin{bmatrix}Z_1\\Z_2\end{bmatrix},\end{aligned}$$

which satisfies

$$\mathrm{rs}(XY)=(I_2\otimes X)\mathrm{rs}(Y),\quad\forall X,Y. \tag{2.5}$$

Let

$$A_0=I_2\otimes a=\begin{bmatrix}a&0\\0&a\end{bmatrix},\quad B_0=\mathrm{rs}(V)=\begin{bmatrix}b&f\\d&h\end{bmatrix},$$

and then

$$T=I_{2n}+A_0B_0^{\mathrm{T}}.$$

To solve the optimization problem equation (2.1), we need to compute $\|T\|_2$ and $\|T^{-1}\|_2$ first, or equivalently, the minimal and maximal eigenvalues of TT^{T}. Note that

$$\begin{aligned}TT^{\mathrm{T}}&=I_{2n}+A_0B_0^{\mathrm{T}}+B_0A_0^{\mathrm{T}}+A_0B_0^{\mathrm{T}}B_0A_0^{\mathrm{T}}\\&=I_{2n}+\begin{bmatrix}A_0&B_0\end{bmatrix}\begin{bmatrix}B_0^{\mathrm{T}}B_0&I_2\\I_2&0\end{bmatrix}\begin{bmatrix}A_0&B_0\end{bmatrix}^{\mathrm{T}}.\end{aligned}$$

Since $A_0^{\mathrm{T}}A_0=(a^{\mathrm{T}}a)I_2=I_2$, noticing $\lambda_*(XY)=\lambda_*(YX),\forall X,Y$,

$$\begin{aligned}\lambda_*(TT^{\mathrm{T}}-I_{2n})&=\lambda_*\left(\begin{bmatrix}A_0&B_0\end{bmatrix}\begin{bmatrix}B_0^{\mathrm{T}}B_0&I_2\\I_2&0\end{bmatrix}\begin{bmatrix}A_0&B_0\end{bmatrix}^{\mathrm{T}}\right)\\&=\lambda_*\left(\begin{bmatrix}B_0^{\mathrm{T}}B_0&I_2\\I_2&0\end{bmatrix}\begin{bmatrix}A_0&B_0\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}A_0&B_0\end{bmatrix}\right)\end{aligned}$$

$$
\begin{aligned}
&=\lambda_*\left(\begin{bmatrix} B_0^{\mathrm{T}}B_0 & I_2 \\ I_2 & 0 \end{bmatrix}\begin{bmatrix} I_2 & A_0^{\mathrm{T}}B_0 \\ B_0^{\mathrm{T}}A_0 & B_0^{\mathrm{T}}B_0 \end{bmatrix}\right)\\
&=\lambda_*\left(\begin{bmatrix} B_0^{\mathrm{T}}B_0 & I_2 \\ I_2 & 0 \end{bmatrix}\begin{bmatrix} I_2 & 0 \\ B_0^{\mathrm{T}}A_0 & I_2 \end{bmatrix}\begin{bmatrix} I_2 & 0 \\ 0 & B_0^{\mathrm{T}}B_0-B_0^{\mathrm{T}}A_0A_0^{\mathrm{T}}B_0 \end{bmatrix}\begin{bmatrix} I_2 & A_0^{\mathrm{T}}B_0 \\ 0 & I_2 \end{bmatrix}\right)\\
&=\lambda_*\left(\begin{bmatrix} I_2 & A_0^{\mathrm{T}}B_0 \\ 0 & I_2 \end{bmatrix}\begin{bmatrix} B_0^{\mathrm{T}}B_0 & I_2 \\ I_2 & 0 \end{bmatrix}\begin{bmatrix} I_2 & 0 \\ B_0^{\mathrm{T}}A_0 & I_2 \end{bmatrix}\begin{bmatrix} I_2 & 0 \\ 0 & B_0^{\mathrm{T}}B_0-B_0^{\mathrm{T}}A_0A_0^{\mathrm{T}}B_0 \end{bmatrix}\right)\\
&=\lambda_*\left(\begin{bmatrix} B_0^{\mathrm{T}}B_0+B_0^{\mathrm{T}}A_0+A_0^{\mathrm{T}}B_0 & B_0^{\mathrm{T}}B_0-B_0^{\mathrm{T}}A_0A_0^{\mathrm{T}}B_0 \\ I_2 & 0 \end{bmatrix}\right).
\end{aligned}
$$

Due to the relationship between the eigenvalues of a quadratic matrix polynomial and its linearization, $\lambda_*(TT^{\mathrm{T}}-I_{2n})=\lambda(\widetilde{H})$, where

$$\widetilde{H}(\lambda)=-\lambda^2 I_2+\lambda(B_0^{\mathrm{T}}B_0+B_0^{\mathrm{T}}A_0+A_0^{\mathrm{T}}B_0)+(B_0^{\mathrm{T}}B_0-B_0^{\mathrm{T}}A_0A_0^{\mathrm{T}}B_0).$$

Note that

$$
\begin{aligned}
H(\lambda)&:=-\widetilde{H}(\lambda-1)\\
&=(\lambda-1)^2I_2-(\lambda-1)(B_0^{\mathrm{T}}B_0+B_0^{\mathrm{T}}A_0+A_0^{\mathrm{T}}B_0)-(B_0^{\mathrm{T}}B_0-B_0^{\mathrm{T}}A_0A_0^{\mathrm{T}}B_0)\\
&=\lambda^2I_2-\lambda(B_0^{\mathrm{T}}B_0+B_0^{\mathrm{T}}A_0+A_0^{\mathrm{T}}B_0+2I_2)+(I_2+B_0^{\mathrm{T}}A_0+A_0^{\mathrm{T}}B_0+B_0^{\mathrm{T}}A_0A_0^{\mathrm{T}}B_0)\\
&=\lambda^2I_2-\lambda(B_0^{\mathrm{T}}B_0+B_0^{\mathrm{T}}A_0+A_0^{\mathrm{T}}B_0+2I_2)+(I_2+A_0^{\mathrm{T}}B_0)^{\mathrm{T}}(I_2+A_0^{\mathrm{T}}B_0). \qquad (2.6)
\end{aligned}
$$

We can see the eigenvalues of H are also the eigenvalues of TT^{T}, and the other eigenvalues of TT^{T} are 1. Since $A_0^{\mathrm{T}}A_0=I_2$,

$$H(1)=-B_0^{\mathrm{T}}B_0+B_0^{\mathrm{T}}A_0A_0^{\mathrm{T}}B_0=-B_0^{\mathrm{T}}(I_{2n}-A_0A_0^{\mathrm{T}})B_0\preceq 0. \qquad (2.7)$$

Hence $H(\lambda)$ is a semi-hyperbolic quadratic matrix polynomial (see, e.g., [2]), which implies

$$\lambda(H)=\{\lambda_i,\, i=1,2,3,4\}\quad \text{satisfying}\quad 0\le\lambda_1\le\lambda_2\le 1\le\lambda_3\le\lambda_4.$$

Here $\lambda_1\ge 0$ is guaranteed by the fact λ_1 is an eigenvalue of TT^{T}. Thus,

$$\kappa_2(T)=\sqrt{\frac{\lambda_4}{\lambda_1}}.$$

Now the problem is to compute the maximal and minimal eigenvalues of H. Note that by Eqs. (2.6) and (2.7),

$$H(\lambda)=\lambda^2I_2-\lambda\left[(I_2+A_0^{\mathrm{T}}B_0)^{\mathrm{T}}(I_2+A_0^{\mathrm{T}}B_0)+I_2-H(1)\right]+(I_2+A_0^{\mathrm{T}}B_0)^{\mathrm{T}}(I_2+A_0^{\mathrm{T}}B_0). \qquad (2.8)$$

The next step is to calculate $A_0^{\mathrm{T}}B_0$ and $B_0^{\mathrm{T}}B_0$.

First, we simplify V. Write

$$a_\perp = \frac{1}{2}\begin{bmatrix}\alpha_C & -2\alpha_M & 2\alpha_K & -\alpha_C\end{bmatrix}^{\mathrm{T}}.$$

Then it is easy to check $a_\perp^{\mathrm{T}} A = 0$. By Eq. (2.4), namely the fact A has full column rank, $A^+ = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$, and $I_4 - AA^+ = \frac{a_\perp a_\perp^{\mathrm{T}}}{a_\perp^{\mathrm{T}} a_\perp}$. Thus, by Eq. (2.2),

$$\begin{aligned} V &= BA^+ - \frac{zz^{\mathrm{T}}}{z^{\mathrm{T}}z}BA^+ + U\frac{a_\perp a_\perp^{\mathrm{T}}}{a_\perp^{\mathrm{T}} a_\perp} + \frac{zw^{\mathrm{T}}}{z^{\mathrm{T}}z} \\ &= BA^+ + \frac{zw^{\mathrm{T}}}{z^{\mathrm{T}}z}(I_4 - AA^+) + U\frac{a_\perp a_\perp^{\mathrm{T}}}{a_\perp^{\mathrm{T}} a_\perp} \\ &= BA^+ + \left(\frac{zw^{\mathrm{T}}}{z^{\mathrm{T}}z} + U\right)\frac{a_\perp a_\perp^{\mathrm{T}}}{a_\perp^{\mathrm{T}} a_\perp}. \end{aligned}$$

Note that

$$\{Ua_\perp : z^{\mathrm{T}}U = 0\} = \{z\}^\perp =: \mathcal{U}.$$

Then

$$V = BA^+ + va_\perp^{\mathrm{T}}, \tag{2.9}$$

where

$$v = \frac{w^{\mathrm{T}}a_\perp}{a_\perp^{\mathrm{T}} a_\perp z^{\mathrm{T}}z}z + u, \quad u \in \mathcal{U},$$

or equivalently,

$$v \in \{v \in \mathbb{R}^n : \ v^{\mathrm{T}}za_\perp^{\mathrm{T}}a_\perp = w^{\mathrm{T}}a_\perp\}. \tag{2.10}$$

Then we simplify $A_0^{\mathrm{T}}B_0$. By Eq. (2.3),

$$a^{\mathrm{T}}BA^+ = -a_0^{\mathrm{T}}AA^+ = -a_0^{\mathrm{T}}\left(I_4 - \frac{a_\perp a_\perp^{\mathrm{T}}}{a_\perp^{\mathrm{T}} a_\perp}\right) = -a_0^{\mathrm{T}}. \tag{2.11}$$

Thus,

$$\begin{aligned} A_0^{\mathrm{T}}B_0 &= (I_2 \otimes a)^{\mathrm{T}}\mathrm{rs}(V) && \text{by Eq. (2.5)} \\ &= \mathrm{rs}(a^{\mathrm{T}}V) && \text{by Eq. (2.9)} \\ &= \mathrm{rs}(a^{\mathrm{T}}BA^+ + a^{\mathrm{T}}va_\perp^{\mathrm{T}}) && \text{by Eq. (2.11)} \\ &= \mathrm{rs}(-a_0^{\mathrm{T}}) + a^{\mathrm{T}}v\,\mathrm{rs}(a_\perp^{\mathrm{T}}). \end{aligned}$$

Write

$$S = \frac{1}{2}\begin{bmatrix}2\alpha_M & \alpha_C \\ \alpha_C & 2\alpha_K\end{bmatrix}, \qquad J = \begin{bmatrix}0 & -1 \\ 1 & 0\end{bmatrix}.$$

It is easy to see that $\mathrm{rs}(a_\perp^{\mathrm{T}}) = SJ$. Therefore,

$$A_0^{\mathrm{T}} B_0 = -I_2 + a^{\mathrm{T}} v S J. \tag{2.12}$$

Next we simplify $B_0^{\mathrm{T}} B_0$. By Eq. (2.5),

$$B_0 = \mathrm{rs}(V) = \mathrm{rs}(B(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}) + \mathrm{rs}(va_\perp^{\mathrm{T}}) = (I_2 \otimes B(A^{\mathrm{T}}A)^{-1})\,\mathrm{rs}(A^{\mathrm{T}}) + (I_2 \otimes v)\,\mathrm{rs}(a_\perp^{\mathrm{T}}).$$

It is easy to see that $\mathrm{rs}(A^{\mathrm{T}}) = (p_0 \otimes I_2)S$ where $p_0 = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^{\mathrm{T}} = e_1 + e_3$. Thus

$$B_0 = \mathrm{rs}(V) = (I_2 \otimes B(A^{\mathrm{T}}A)^{-1})(p_0 \otimes I_2)S + (I_2 \otimes v)SJ =: (I_2 \otimes v)SJ + DS,$$

where

$$D = (I_2 \otimes B(A^{\mathrm{T}}A)^{-1})(p_0 \otimes I_2). \tag{2.13}$$

Besides, by Eq. (2.11),

$$(I_2 \otimes a^{\mathrm{T}})DS = (I_2 \otimes a^{\mathrm{T}})\,\mathrm{rs}(BA^+) = \mathrm{rs}(a^{\mathrm{T}}BA^+) = -I_2. \tag{2.14}$$

Then

$$\begin{aligned} B_0^{\mathrm{T}} B_0 &= [(I_2 \otimes v)SJ + DS]^{\mathrm{T}}[(I_2 \otimes v)SJ + DS] \\ &= -JS(I_2 \otimes v^{\mathrm{T}}v)SJ + SD^{\mathrm{T}}(I_2 \otimes v)SJ - JS(I_2 \otimes v^{\mathrm{T}})DS + SD^{\mathrm{T}}DS \\ &= -v^{\mathrm{T}}vJS^2J + SD^{\mathrm{T}}(I_2 \otimes v)SJ - JS(I_2 \otimes v^{\mathrm{T}})DS + SD^{\mathrm{T}}DS, \end{aligned}$$

and by Eqs. (2.7) and (2.12),

$$\begin{aligned} H_1 &:= -JH(1)J \\ &= -[v^{\mathrm{T}}v - (a^{\mathrm{T}}v)^2]S^2 - JSD^{\mathrm{T}}(I_2 \otimes v)S + S(I_2 \otimes v^{\mathrm{T}})DSJ + JSD^{\mathrm{T}}DSJ + I_2 - a^{\mathrm{T}}v(JS - SJ). \end{aligned}$$

Turn back to Eq. (2.8). Substituting Eq. (2.12),

$$\begin{aligned} H(\lambda) &= \lambda^2 I_2 - \lambda[I_2 - (a^{\mathrm{T}}v)^2 JS^2J - H(1)] - (a^{\mathrm{T}}v)^2 JS^2J \\ &= -J\left(\lambda^2 I_2 - \lambda[I_2 + (a^{\mathrm{T}}v)^2S^2 - H_1] + (a^{\mathrm{T}}v)^2S^2\right)J. \end{aligned}$$

To sum up, the optimization problem equation (2.1) is equivalent to:

$$\min_{v \in \mathbb{R}^n : v^{\mathrm{T}} z a_\perp^{\mathrm{T}} a_\perp = w^{\mathrm{T}} a_\perp} \kappa_2(T) = \min_{v \in \mathbb{R}^n : v^{\mathrm{T}} z a_\perp^{\mathrm{T}} a_\perp = w^{\mathrm{T}} a_\perp} \sqrt{\frac{\lambda_4}{\lambda_1}}, \tag{2.15}$$

where $0 \le \lambda_1 \le \lambda_2 \le 1 \le \lambda_3 \le \lambda_4$ are four eigenvalues of the hyperbolic quadratic polynomial

$$-JH(\lambda)J = \lambda^2 I_2 - \lambda[I_2 + (a^{\mathrm{T}}v)^2S^2 - H_1] + (a^{\mathrm{T}}v)^2S^2$$

and

$$H_1 := I_2 - [v^{\mathrm{T}}v - (a^{\mathrm{T}}v)^2]S^2 - JSD^{\mathrm{T}}(I_2 \otimes v)S + S(I_2 \otimes v^{\mathrm{T}})DSJ + JSD^{\mathrm{T}}DSJ - a^{\mathrm{T}}v(JS - SJ). \tag{2.16}$$

3 Parameterize the problem

First we give a lemma to show the relationship between hyperbolic quadratic matrix polynomial and its eigenvalues.

Lemma 3.1. *For any two hyperbolic quadratic matrix polynomial $H^{(i)}(\lambda)=\lambda^2 I+\lambda B^{(i)}+C$ satisfying $B^{(i)}\preceq 0$, $C\succeq 0$, let $\lambda_{\min}^{(i)}$ and $\lambda_{\max}^{(i)}$ be their minimal and maximal eigenvalues respectively. If $B^{(1)}\succeq B^{(2)}$, then $\lambda_{\min}^{(1)}\geq\lambda_{\min}^{(2)}$, $\lambda_{\max}^{(1)}\leq\lambda_{\max}^{(2)}$.*

Proof. Since the polynomial is hyperbolic, for any vector x satisfying $x^{\mathrm{T}}x=1$, we can well define

$$\rho_{\pm}(x):=\frac{1}{2}\left(-x^{\mathrm{T}}Bx\pm\sqrt{(x^{\mathrm{T}}Bx)^2-4x^{\mathrm{T}}Cx}\right).$$

Then

$$\frac{\partial\rho_+(x)}{\partial(x^{\mathrm{T}}Bx)}=\frac{1}{2}\left(-1+\frac{2x^{\mathrm{T}}Bx}{2\sqrt{(x^{\mathrm{T}}Bx)^2-4x^{\mathrm{T}}Cx}}\right)=\frac{-\rho_+(x)}{\sqrt{(x^{\mathrm{T}}Bx)^2-4x^{\mathrm{T}}Cx}}\leq 0.$$

Thus, $B^{(1)}\succeq B^{(2)}$ implies $x^{\mathrm{T}}B^{(1)}x\geq x^{\mathrm{T}}B^{(2)}x$ for all x. Then, $\rho_+^{(1)}(x)\leq\rho_+^{(2)}(x)$. As a result, by the Courant-Fischer min-max theorem of hyperbolic quadratic matrix polynomials [1], $\lambda_{\max}^{(1)}=\max\rho_+^{(1)}(x)\leq\max\rho_+^{(2)}(x)=\lambda_{\max}^{(2)}$.

Similarly,

$$\frac{\partial\rho_-(x)}{\partial(x^{\mathrm{T}}Bx)}=\frac{1}{2}(-1-\frac{2x^{\mathrm{T}}Bx}{2\sqrt{(x^{\mathrm{T}}Bx)^2-4x^{\mathrm{T}}Cx}})=\frac{\rho_-(x)}{\sqrt{(x^{\mathrm{T}}Bx)^2-4x^{\mathrm{T}}Cx}}\geq 0.$$

Then, $\rho_-^{(1)}(x)\geq\rho_-^{(2)}(x)$, and $\lambda_{\min}^{(1)}=\min\rho_-^{(1)}(x)\geq\min\rho_-^{(2)}(x)=\lambda_{\min}^{(2)}$. □

We may continue simplifying the optimization problem.

First, we know the minimal v must lie in $\{a,Ma,Ca,Ka\}$. Otherwise, let

$$v=v_0+v_\perp\quad\text{where}\quad v_0\in\{a,Ma,Ca,Ka\},\ v_\perp\in\{a,Ma,Ca,Ka\}^\perp.$$

Then $a^{\mathrm{T}}v=a^{\mathrm{T}}v_0$, $B^{\mathrm{T}}v=B^{\mathrm{T}}v_0$, which implies $(I_2\otimes v^{\mathrm{T}})D=(I_2\otimes v_0^{\mathrm{T}})D$ by Eq. (2.13). However, $v^{\mathrm{T}}v=v_0^{\mathrm{T}}v_0+v_\perp^{\mathrm{T}}v_\perp\geq v_0^{\mathrm{T}}v_0$. By Lemma 3.1, $\kappa_2(T;v)\geq\kappa_2(T;v_0)$.

Let

$$B_a=B+a\begin{bmatrix}\alpha_M & \alpha_C & \alpha_K\end{bmatrix},\quad \widetilde{B}_a=B_a(A^{\mathrm{T}}A)^{-1}. \tag{3.1}$$

It is easy to check $a^{\mathrm{T}}B_a=0$ and $B^{\mathrm{T}}B_a=B_a^{\mathrm{T}}B_a$. Since the minimal v must lie in $\{a,Ma,Ca,Ka\}$, we can write

$$v=\xi a+\widetilde{B}_a\widetilde{y}_B,\quad\text{where}\quad \xi\in\mathbb{R},\ \widetilde{y}_B\in\mathbb{R}^3. \tag{3.2}$$

Thus, by Eq. (2.14),

$$a^{\mathrm{T}}v=\xi,\quad v^{\mathrm{T}}v=\xi^2+\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}}\widetilde{B}_a\widetilde{y}_B,\quad (I_2\otimes v^{\mathrm{T}})DS=-\xi I_2+(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}})DS. \tag{3.3}$$

Substituting Eq. (3.3) into Eq. (2.16),

$$H_1 = I_2 - \widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}}\widetilde{B}_a\widetilde{y}_B S^2 - JSD^{\mathrm{T}}(I_2\otimes\widetilde{B}_a\widetilde{y}_B)S + S(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}})DSJ + JSD^{\mathrm{T}}DSJ.$$

Clearly,

$$I_n = aa^{\mathrm{T}} + \widetilde{B}_a\widetilde{B}_a^{+} + B_\perp B_\perp^{+},$$

where $B_\perp$ is a basis of $\{a, Ma, Ca, Ka\}^\perp$. Thus,

$$\begin{aligned}
DSJ &= (I_2\otimes[aa^{\mathrm{T}} + \widetilde{B}_a\widetilde{B}_a^{+} + B_\perp B_\perp^{+}])DSJ \\
&\qquad\qquad \text{by Eq. (2.13)} \\
&= (I_2\otimes aa^{\mathrm{T}})DSJ + (I_2\otimes[\widetilde{B}_a\widetilde{B}_a^{+} + B_\perp B_\perp^{+}])(I_2\otimes B(A^{\mathrm{T}}A)^{-1})(p_0\otimes I_2)SJ \\
&\qquad\qquad \text{by Eq. (2.14)} \\
&= -(I_2\otimes a)J + (I_2\otimes\widetilde{B}_a\widetilde{B}_a^{+}B(A^{\mathrm{T}}A)^{-1})(p_0\otimes I_2)SJ \\
&\qquad\qquad \text{by Eq. (3.1)} \\
&= -(I_2\otimes a)J + \left(I_2\otimes\widetilde{B}_a\widetilde{B}_a^{+}(\widetilde{B}_a - a\begin{bmatrix}\alpha_M & \alpha_C & \alpha_K\end{bmatrix}(A^{\mathrm{T}}A)^{-1})\right)(p_0\otimes I_2)SJ \\
&\qquad\qquad \text{by } a^T B_a = a^T\widetilde{B}_a = 0 \\
&= -(I_2\otimes a)J + (I_2\otimes\widetilde{B}_a)(p_0\otimes I_2)SJ.
\end{aligned}$$

Then

$$\begin{aligned}
-JSD^{\mathrm{T}}DSJ &= -J(I_2\otimes a^{\mathrm{T}}a)J - JS(p_0^{\mathrm{T}}\otimes I_2)(I_2\otimes\widetilde{B}_a^{\mathrm{T}}\widetilde{B}_a)(p_0\otimes I_2)SJ \\
&= I_2 - JS(p_0^{\mathrm{T}}\otimes I_2)(I_2\otimes\widetilde{B}_a^{\mathrm{T}}\widetilde{B}_a)(p_0\otimes I_2)SJ \\
&=: I_2 + SD_a^{\mathrm{T}}D_aS,
\end{aligned}$$

where $D_a = (I_2\otimes\widetilde{B}_a)(p_0\otimes I_2)SJS^{-1}$, and by $a^T\widetilde{B}_a = 0$,

$$\begin{aligned}
-S(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}})DSJ &= S(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}}a)J - S(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}}\widetilde{B}_a)(p_0\otimes I_2)SJ \\
&= -S(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}}\widetilde{B}_a)(p_0\otimes I_2)SJ \\
&= -S(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}})D_aS,
\end{aligned}$$

which gives

$$\begin{aligned}
H_1 &= I_2 - \widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}}\widetilde{B}_a\widetilde{y}_B S^2 + S(I_2\otimes\widetilde{y}_B^{\mathrm{T}}\widetilde{B}_a^{\mathrm{T}})D_aS + SD_a^{\mathrm{T}}(I_2\otimes\widetilde{B}_a\widetilde{y}_B)S - I_2 - SD_a^{\mathrm{T}}D_aS \\
&= -S(D_a - (I_2\otimes\widetilde{B}_a\widetilde{y}_B))^{\mathrm{T}}(D_a - (I_2\otimes\widetilde{B}_a\widetilde{y}_B))S \\
&= -SD_y^{\mathrm{T}}D_yS,
\end{aligned} \tag{3.4}$$

where $D_y = D_a - (I_2\otimes\widetilde{B}_a\widetilde{y}_B)$. Note that

$$\begin{aligned}
D_a &= (I_2\otimes\widetilde{B}_a)(p_0\otimes I_2)SJS^{-1} \\
&= (I_2\otimes B_a(A^{\mathrm{T}}A)^{-1})(p_0\otimes I_2)SJS^{-1} = (I_2\otimes B_a)R,
\end{aligned}$$

where calculation tells

$$R=(I_2\otimes(A^{\mathrm{T}}A)^{-1})(p_0\otimes I_2)SJS^{-1}=\delta\begin{bmatrix}r & -e_1\\ e_3 & r-e_2\end{bmatrix},$$

in which $I_3=\begin{bmatrix}e_1 & e_2 & e_3\end{bmatrix}$, and

$$\delta=\frac{4}{4\alpha_K\alpha_M-\alpha_C^2},\quad r=\frac{1}{2(2\alpha_M^2+2\alpha_K^2+\alpha_C^2)}\begin{bmatrix}2\alpha_C(\alpha_M+\alpha_K)\\ 4\alpha_K^2+\alpha_C^2\\ -2\alpha_C(\alpha_M+\alpha_K)\end{bmatrix}.$$

Thus, writing $y_B=(A^{\mathrm{T}}A)^{-1}\widetilde{y}_B$ and $y=r-\frac{1}{\delta}y_B$,

$$D_y=(I_2\otimes B_a)R-(I_2\otimes B_a y_B)=\delta(I_2\otimes B_a)\begin{bmatrix}y & -e_1\\ e_3 & y-e_2\end{bmatrix}.$$

Then, letting $B_W=B_a^{\mathrm{T}}B_a=B^{\mathrm{T}}B-\begin{bmatrix}\alpha_M & \alpha_C & \alpha_K\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\alpha_M & \alpha_C & \alpha_K\end{bmatrix}$,

$$\begin{aligned}D_y^{\mathrm{T}}D_y&=\delta^2\begin{bmatrix}y^{\mathrm{T}}B_Wy & (e_3-e_1)^{\mathrm{T}}B_Wy\\ (e_3-e_1)^{\mathrm{T}}B_Wy & y^{\mathrm{T}}B_Wy-2e_2^{\mathrm{T}}B_Wy\end{bmatrix}+\delta^2\begin{bmatrix}e_3^{\mathrm{T}}B_We_3 & -e_2^{\mathrm{T}}B_We_3\\ -e_2^{\mathrm{T}}B_We_3 & e_2^{\mathrm{T}}B_We_2+e_1^{\mathrm{T}}B_We_1\end{bmatrix}\\ &=:\delta^2(W_y+W_e).\end{aligned}\tag{3.5}$$

Note that $\begin{bmatrix}y & -e_1\\ e_3 & y-e_2\end{bmatrix}$ has full column rank, because otherwise $y\parallel e_1,(y-e_2)\parallel e_3$, which is a contradiction. Thus $D_y^{\mathrm{T}}D_y\succ 0$ and then by Eq. (3.4) $H_1=-SD_y^{\mathrm{T}}D_yS\prec 0$, which implies $\lambda_2<1<\lambda_3$.

On the other hand, by Eq. (3.2),

$$v=\xi a+B_ay_B=\xi a+\delta B_a(r-y),\quad \xi\in\mathbb{R},\quad y\in\mathbb{R}^3.$$

Then the only constraint on v, namely Eq. (2.10), is

$$\frac{w^{\mathrm{T}}a_\perp}{a_\perp^{\mathrm{T}}a_\perp}=v^{\mathrm{T}}z=\xi a^{\mathrm{T}}z+\delta z^{\mathrm{T}}B_ar-\delta z^{\mathrm{T}}B_ay,$$

or equivalently,

$$z_a^{\mathrm{T}}y=\xi+\beta,$$

where

$$z_a=\frac{\delta B_a^{\mathrm{T}}z}{a^{\mathrm{T}}z},\qquad \beta=z_a^{\mathrm{T}}r-\frac{w^{\mathrm{T}}a_\perp}{a^{\mathrm{T}}za_\perp^{\mathrm{T}}a_\perp}.$$

Also,

$$-JH(\lambda)J=\lambda^2I_2-\lambda\left(I_2+\delta^2S(W_y+W_e)S+\xi^2S^2\right)+\xi^2S^2.$$

To sum up, the constrained optimization problem Eq. (2.1), or Eq. (2.15), is equivalent to this unconstrained optimization problem:

$$\min_{y\in\mathbb{R}^3}\kappa_2(T)=\min_{y\in\mathbb{R}^3}\sqrt{\frac{\lambda_4}{\lambda_1}} \tag{3.6}$$

where $0\le\lambda_1\le\lambda_2<1<\lambda_3\le\lambda_4$ are four eigenvalues of the hyperbolic quadratic polynomial

$$-JH(\lambda)J=\lambda^2I_2-\lambda[I_2+\delta^2SW_eS+\delta^2SW_yS+(z_a^{\mathrm{T}}y-\beta)^2S^2]+(z_a^{\mathrm{T}}y-\beta)^2S^2, \tag{3.7}$$

and

$$W_y=(y^{\mathrm{T}}B_Wy)I_2+\begin{bmatrix}0 & (e_3-e_1)^{\mathrm{T}}B_Wy\\ (e_3-e_1)^{\mathrm{T}}B_Wy & -2e_2^{\mathrm{T}}B_Wy\end{bmatrix}.$$

4 Analyze the problem

It might be possible to have an explicit solution of the unconstrained optimization problem Eq. (3.6), but it must be complicated and difficult to use in practice. Note that $H(\lambda)$ is of order 2, we can directly use some numerical optimization method to reach the minimal point $y_{\min}$ and the minimum $\kappa_2(T;y_{\min})$.

If $z_a^{\mathrm{T}}y=\beta$, then $-JH(\lambda)J=\lambda^2I_2-\lambda(I_2-H_1)$, which implies that $\lambda_1=\lambda_2=0$ and $\kappa_2(T)=+\infty$. Hence the whole domain is split into two connected domains: $\{y:z_a^{\mathrm{T}}y<\beta\}$ and $\{y:z_a^{\mathrm{T}}y>\beta\}$. Then we may use optimization methods to try to attain the local minimum in both of the two domains.

Most optimization methods require the gradient and the Hessian of the objective function, and we calculate it theoretically.

First try to obtain those for the eigenvalues. For any eigenpair of Eq. (3.7), it holds that $-JH(\lambda)Jx=0$, and without loss of generally, $x^{\mathrm{T}}x=1$. Let $W=\delta^2(W_e+W_y)$, $\xi=z_a^{\mathrm{T}}y-\beta$, and then

$$-JH(\lambda)J=\lambda^2I_2-\lambda(I_2+SWS+\xi^2S^2)+\xi^2S^2. \tag{4.1}$$

Taking differential on both sides gives

$$-J\mathrm{d}(H(\lambda))Jx-JH(\lambda)J\mathrm{d}x=0, \tag{4.2}$$

and

$$0=x^{\mathrm{T}}JH(\lambda)J\mathrm{d}x=-x^{\mathrm{T}}J\mathrm{d}(H(\lambda))Jx.$$

Thus,

$$0=-x^{\mathrm{T}}\Big(2\lambda\mathrm{d}\lambda I_2-\mathrm{d}\lambda[I_2+SWS+\xi^2S^2]-\lambda[S\mathrm{d}WS+2\xi\mathrm{d}\xi S^2]+2\xi\mathrm{d}\xi S^2\Big)x.$$

Write $t:=Sx=\begin{bmatrix}\tau_1 & \tau_2\end{bmatrix}^{\mathrm{T}}$, and noticing $x^{\mathrm{T}}x=1$,

$$0=2\lambda\mathrm{d}\lambda-\mathrm{d}\lambda[1+t^TWt+\xi^2t^Tt]-\lambda[t^{\mathrm{T}}\mathrm{d}Wt+2\xi\mathrm{d}\xi t^{\mathrm{T}}t]+2\xi\mathrm{d}\xi t^{\mathrm{T}}t,$$

and then

$$\mathrm{d}\lambda=\frac{\lambda t^{\mathrm{T}}\mathrm{d}Wt+2(\lambda-1)\xi\mathrm{d}\xi t^{\mathrm{T}}t}{2\lambda-[1+t^{T}Wt+\xi^{2}t^{T}t]}.$$

Note that $\mathrm{d}\xi=z_a^T\mathrm{d}y$ and by Eq. (3.5),

$$\begin{aligned}\mathrm{d}Wt&=\delta^2\mathrm{d}\left((y^{\mathrm{T}}B_Wy)I_2+\begin{bmatrix}0&(e_3-e_1)^{\mathrm{T}}B_Wy\\(e_3-e_1)^{\mathrm{T}}B_Wy&-2e_2^{\mathrm{T}}B_Wy\end{bmatrix}\right)t\\&=\delta^2t\mathrm{d}\left(y^{\mathrm{T}}B_Wy\right)+\delta^2\begin{bmatrix}0&(e_3-e_1)^{\mathrm{T}}B_W\mathrm{d}y\\(e_3-e_1)^{\mathrm{T}}B_W\mathrm{d}y&-2e_2^{\mathrm{T}}B_W\mathrm{d}y\end{bmatrix}t\\&=2\delta^2ty^{\mathrm{T}}B_W\mathrm{d}y+\delta^2\begin{bmatrix}\tau_2(e_3-e_1)^{\mathrm{T}}\\\tau_1(e_3-e_1)^{\mathrm{T}}-2\tau_2e_2^{\mathrm{T}}\end{bmatrix}B_W\mathrm{d}y\\&=\delta^2(2ty^{\mathrm{T}}+E_t^T)B_W\mathrm{d}y,\end{aligned}$$

where

$$E_t:=\begin{bmatrix}\tau_2(e_3-e_1)&\tau_1(e_3-e_1)-2\tau_2e_2\end{bmatrix}.$$

Then

$$t^T\mathrm{d}Wt=2\delta^2\left[(t^{\mathrm{T}}t)y+\frac{1}{2}E_tt\right]^{\mathrm{T}}B_W\mathrm{d}y=2\delta^2[(t^{\mathrm{T}}t)y+\tau_1\tau_2(e_3-e_1)-\tau_2^2e_2]^{\mathrm{T}}B_W\mathrm{d}y,$$

and we have

$$\mathrm{d}\lambda=\frac{2\lambda\delta^2[(t^{\mathrm{T}}t)y+\tau_1\tau_2(e_3-e_1)-\tau_2^2e_2]^{\mathrm{T}}B_W+2(\lambda-1)t^Tt\xi z_a^T}{2\lambda-[1+t^TWt+\xi^2t^Tt]}\mathrm{d}y,$$

which implies

$$\nabla\lambda=2\frac{\lambda\delta^2B_W[(t^{\mathrm{T}}t)y+\tau_1\tau_2(e_3-e_1)-\tau_2^2e_2]+(\lambda-1)t^Tt\xi z_a}{2\lambda-[1+t^TWt+\xi^2t^Tt]}.$$

Writing $\tau_W:=t^TWt$, then

$$\nabla\lambda=2\frac{\lambda\delta^2B_W[(t^{\mathrm{T}}t)y+\tau_1\tau_2(e_3-e_1)-\tau_2^2e_2]+(\lambda-1)t^Tt\xi z_a}{2\lambda-[1+\tau_W+\xi^2t^Tt]},$$

and

$$\mathrm{d}(\nabla\lambda)=\frac{\partial\nabla\lambda}{\partial\lambda}\mathrm{d}\lambda+\frac{\partial\nabla\lambda}{\partial\xi}\mathrm{d}\xi+\frac{\partial\nabla\lambda}{\partial\tau_1}\mathrm{d}\tau_1+\frac{\partial\nabla\lambda}{\partial\tau_2}\mathrm{d}\tau_2+\frac{\partial\nabla\lambda}{\partial\tau_W}\mathrm{d}\tau_W.\tag{4.3}$$

Let

$$\sigma:=2\lambda-[1+t^TWt+\xi^2t^Tt],\quad g:=\lambda\delta^2B_W\left[(t^Tt)y+\frac{1}{2}E_tt\right]+(\lambda-1)t^Tt\xi z_a,\tag{4.4}$$

then

$$\nabla\lambda=2\frac{g}{\sigma}.\tag{4.5}$$

Putting

$$\begin{aligned}
&\frac{\partial\nabla\lambda}{\partial\lambda}=\frac{2}{\sigma^2}\left((\delta^2B_W\left[(t^{\mathrm{T}}t)y+\frac{1}{2}E_tt\right]+t^Tt\xi z_a)\sigma-2g\right),\\
&\frac{\partial\nabla\lambda}{\partial\xi}=\frac{2}{\sigma^2}\left((\lambda-1)t^Ttz_a\sigma+2t^Tt\xi g\right),\\
&\frac{\partial\nabla\lambda}{\partial\tau_1}=\frac{2}{\sigma^2}\left((\lambda\delta^2B_W[2\tau_1y+\tau_2(e_3-e_1)]+2(\lambda-1)\tau_1\xi z_a)\sigma+\xi^22\tau_1g\right),\\
&\frac{\partial\nabla\lambda}{\partial\tau_2}=\frac{2}{\sigma^2}\left((\lambda\delta^2B_W[2\tau_2y+\tau_1(e_3-e_1)-2\tau_2e_2]+2(\lambda-1)\tau_2\xi z_a)\sigma+\xi^22\tau_2g\right),\\
&\frac{\partial\nabla\lambda}{\partial\tau_W}=\frac{2}{\sigma^2}g,\\
&\mathrm{d}\tau_W=2t^TW\mathrm{d}t+t^T\mathrm{d}Wt,
\end{aligned}$$

into Eq. (4.3) gives

$$\begin{aligned}
\mathrm{d}(\nabla\lambda)=\frac{2}{\sigma^2}\Bigg(&\left(\left(\delta^2B_W\left[(t^{\mathrm{T}}t)y+\frac{1}{2}E_tt\right]+t^Tt\xi z_a\right)\sigma-2g\right)2\frac{g^T}{\sigma}\mathrm{d}y\\
&+\left((\lambda-1)t^Ttz_a\sigma+2t^Tt\xi g\right)z_a^T\mathrm{d}y\\
&+2\sigma\lambda\delta^2B_Wyt^T\mathrm{d}t+\sigma\lambda\delta^2B_WE_t\mathrm{d}t+2\sigma(\lambda-1)\xi z_at^T\mathrm{d}t+2\xi^2gt^T\mathrm{d}t\\
&+g\left(2t^TW\mathrm{d}t+2\delta^2\left[(t^Tt)y+\frac{1}{2}E_tt\right]^TB_W\mathrm{d}y\right)\Bigg)\\
=:\frac{2}{\sigma^2}&\left(G_y\mathrm{d}y+G_t\mathrm{d}t\right),
\end{aligned}\tag{4.6}$$

where

$$\begin{aligned}
G_y=&\left(\left(\delta^2B_W[\left(t^{\mathrm{T}}t\right)y+\frac{1}{2}E_tt\right]+t^Tt\xi z_a\right)\sigma-2g\right)2\frac{g^T}{\sigma}\\
&+\left((\lambda-1)t^Ttz_a\sigma+2t^Tt\xi g\right)z_a^T+2g\delta^2\left[(t^Tt)y+\frac{1}{2}E_tt\right]^TB_W\\
&\qquad\text{by }\delta^2B_W\left[(t^Tt)y+\frac{1}{2}E_tt\right]=\frac{1}{\lambda}[g-(\lambda-1)t^Tt\xi z_a]\\
=&\left(\frac{4}{\lambda}-\frac{4}{\sigma}\right)gg^T+\frac{2}{\lambda}t^Tt\xi(z_ag^T+gz_a^T)+\sigma(\lambda-1)t^Ttz_az_a^T\\
&\qquad\text{by }-x^TJH(\lambda)Jx=\lambda^2-\lambda(2\lambda-\sigma)+\xi^2t^Tt=0\Rightarrow\sigma-\lambda=-\frac{\xi^2t^Tt}{\lambda}
\end{aligned}$$

$$
\begin{aligned}
&=\frac{4}{\sigma\lambda}\frac{\xi^2 t^T t}{-\lambda}gg^T+\frac{2}{\lambda}t^T t\xi(z_a g^T+gz_a^T)+\sigma(\lambda-1)t^T t z_a z_a^T\\
&=\sigma t^T t\left(-\left(z_a-\frac{2\xi}{\sigma\lambda}g\right)\left(z_a-\frac{2\xi}{\sigma\lambda}g\right)^T+\lambda z_a z_a^T\right),
\end{aligned}
$$

and

$$
G_t=2gt^T W+2(\sigma\lambda\delta^2 B_W y+\sigma(\lambda-1)\xi z_a+\xi^2 g)t^T+\sigma\lambda\delta^2 B_W E_t. \tag{4.7}
$$

Then we calculate $\mathrm{d}t$. By Eq. (4.2), $\mathrm{d}x$ satisfies $-JH(\lambda)J\mathrm{d}x=J\mathrm{d}(H(\lambda))Jx$, which gives

$$
\begin{aligned}
-JH(\lambda)J\mathrm{d}x&=J\mathrm{d}(H(\lambda))Jx\\
&=-2\lambda\mathrm{d}\lambda x+\mathrm{d}\lambda[I_2+SWS+\xi^2S^2]x+\lambda S\mathrm{d}WSx+2\lambda\xi\mathrm{d}\xi S^2x-2\xi\mathrm{d}\xi S^2x\\
&=(1-2\lambda)\mathrm{d}\lambda x+\mathrm{d}\lambda[SWS+\xi^2S^2]x+\lambda S\mathrm{d}WSx+2(\lambda-1)\xi\mathrm{d}\xi S^2x.
\end{aligned}
$$

Note that $x^T Jx=0$, $x^T\mathrm{d}x=0$ and $x\in\mathbb{R}^2$. Hence $\mathrm{d}x=\eta Jx$ holds for some η. Moreover, since x is an eigenvector of the matrix $-JH(\lambda)J$ corresponding to 0, Jx has to be its eigenvector corresponding to a positive eigenvalue ω, and

$$
\begin{aligned}
\omega=(Jx)^T(-JH(\lambda)J)(Jx)&=(Jx)^T[\lambda^2 I_2-\lambda(I_2+SWS+\xi^2S^2)+\xi^2S^2](Jx)\\
&=\lambda^2-\lambda(1+s^T Ws+\xi^2 s^T s)+\xi^2 s^T s,
\end{aligned}
$$

where $s:=SJx=SJS^{-1}t$. Thus

$$
(Jx)^T(-JH(\lambda)J)(\eta Jx)=-\mathrm{d}\lambda x^T J[SWS+\xi^2S^2]x-\lambda x^T JS\mathrm{d}WSx-2(\lambda-1)\xi\mathrm{d}\xi x^T JS^2x,
$$

and

$$
\begin{aligned}
\eta&=\frac{-\mathrm{d}\lambda x^T J[SW+\xi^2S]t-\lambda x^T JS\mathrm{d}Wt-2(\lambda-1)\xi\mathrm{d}\xi x^T JSt}{\omega}\\
&=\frac{-2\frac{g^T\mathrm{d}y}{\sigma}x^T J[SW+\xi^2S]t-\lambda x^T JS(2\delta^2 ty^{\mathrm{T}}B_W\mathrm{d}y+\delta^2E_t^T B_W\mathrm{d}y)-2(\lambda-1)\xi z_a^T\mathrm{d}yx^T JSt}{\omega}\\
&=\frac{-2\frac{x^T J[SW+\xi^2S]t}{\sigma}g^T-2\lambda\delta^2(x^T JSt)y^{\mathrm{T}}B_W-\lambda\delta^2x^T JSE_t^T B_W-2(\lambda-1)\xi(x^T JSt)z_a^T}{\omega}\mathrm{d}y\\
&=\frac{2\frac{s^T Wt+\xi^2s^T t}{\sigma}g^T+2\lambda\delta^2s^T ty^{\mathrm{T}}B_W+\lambda\delta^2s^T E_t^T B_W+2(\lambda-1)\xi s^T tz_a^T}{\omega}\mathrm{d}y.
\end{aligned}
$$

Since by Eq. (4.7),

$$
\begin{aligned}
g_t:=G_t s&=2gt^T Ws+2(\sigma\lambda\delta^2B_W y+\sigma(\lambda-1)\xi z_a+\xi^2g)t^T s+\sigma\lambda\delta^2B_W E_t s\\
&=2(t^T Ws+\xi^2t^T s)g+\sigma[2\lambda\delta^2t^T sB_W y+2(\lambda-1)t^T s\xi z_a+\lambda\delta^2B_W E_t s],
\end{aligned}
$$

we have

$$
\eta=\frac{g_t^T}{\sigma\omega}\mathrm{d}y. \tag{4.8}
$$

Moreover,

$$\begin{aligned} g_t &= 2(t^T W s+\xi^2 t^T s)g+\sigma[2\lambda\delta^2 t^T s B_W y+2(\lambda-1)t^T s\xi z_a+\lambda\delta^2 B_W E_t s] \qquad \text{by Eq. (4.4)}\\ &= 2(t^T W s+\xi^2 t^T s)g+\sigma\Big[2\frac{t^T s}{t^T t}\Big(g-\lambda\delta^2 B_W\frac{1}{2}E_t t\Big)+\lambda\delta^2 B_W E_t s\Big]\\ &= 2\Big(t^T W s+\xi^2 t^T s+\frac{\sigma t^T s}{t^T t}\Big)g+\frac{\sigma}{t^T t}\lambda\delta^2 B_W E_t[(t^T t)s-(t^T s)t]. \end{aligned} \tag{4.9}$$

Note that

$$(t^T t)s-(t^T s)t=[(t^T t)I_2-tt^T]s=-Jtt^T Js=-Jt(x^T SJSJx)=-Jtx^T\Big(-\frac{4}{\delta}I_2\Big)x=\frac{4}{\delta}Jt,$$

and

$$\begin{aligned} 0=x^T(-J^T H(\lambda)J)Jx &= x^T(\lambda^2 I_2-\lambda[I_2+SWS+\xi^2 S^2]+\xi^2 S^2)Jx\\ &= -\lambda[t^T W s+\xi^2 t^T s]+\xi^2 t^T s, \end{aligned}$$

which gives

$$t^T W s=\frac{(1-\lambda)\xi^2 t^T s}{\lambda}. \tag{4.10}$$

Note that by Eq. (4.4)

$$\xi^2 t^T t+\sigma\lambda=2\lambda^2-\lambda[1+t^T W t+\xi^2 t^T t]+\xi^2 t^T t=\lambda^2. \tag{4.11}$$

Substituting Eqs. (4.10) and (4.11) into Eq. (4.9) gives

$$\begin{aligned} g_t=G_t s &= 2\Big(\frac{\xi^2}{\lambda}t^T s+\frac{\sigma t^T s}{t^T t}\Big)g+\frac{4\sigma}{t^T t}\lambda\delta B_W E_t Jt\\ &= 2t^T s\Big(\frac{\xi^2}{\lambda}+\frac{\sigma}{t^T t}\Big)g+\frac{4\sigma}{t^T t}\lambda\delta B_W E_t Jt\\ &= 2\frac{\lambda}{t^T t}[t^T s g+2\sigma\delta B_W E_t Jt]. \end{aligned}$$

Noticing by Eq. (4.8)

$$\mathrm{d}t=S\,\mathrm{d}x=\eta SJx=\eta s=\left(\frac{g_t^T}{\sigma\omega}\mathrm{d}y\right)s,$$

Eq. (4.6) becomes

$$\mathrm{d}(\nabla\lambda)=\frac{2}{\sigma^2}\left(\sigma t^T t\left(\lambda z_a z_a^T-\Big(z_a-\frac{2\xi}{\sigma\lambda}g\Big)\Big(z_a-\frac{2\xi}{\sigma\lambda}g\Big)^T\right)\mathrm{d}y+\frac{g_t g_t^T}{\sigma\omega}\mathrm{d}y\right),$$

and

$$\nabla^2\lambda=\frac{2}{\sigma^2}\left(\sigma t^T t\left(\lambda z_a z_a^T-\Big(z_a-\frac{2\xi}{\sigma\lambda}g\Big)\Big(z_a-\frac{2\xi}{\sigma\lambda}g\Big)^T\right)+\frac{g_t g_t^T}{\sigma\omega}\right). \tag{4.12}$$

Following this, we may easily have $\nabla\lambda_4, \nabla\lambda_1$ and then

$$\nabla\kappa = \frac{\lambda_1\nabla\lambda_4 - \lambda_4\nabla\lambda_1}{2\lambda_1^2\kappa} = \frac{\kappa}{2}\left(\frac{\nabla\lambda_4}{\lambda_4} - \frac{\nabla\lambda_1}{\lambda_1}\right), \tag{4.13}$$

with $\kappa = \kappa_2(T)$, which can be used in the optimization methods. Moreover,

$$\begin{aligned}\nabla^2\kappa &= \left(\frac{\nabla\lambda_4}{\lambda_4} - \frac{\nabla\lambda_1}{\lambda_1}\right)\frac{\nabla\kappa^T}{2} + \frac{\kappa}{2}\left(\frac{\nabla^2\lambda_4}{\lambda_4} - \frac{\nabla\lambda_4\nabla\lambda_4^T}{\lambda_4^2} - \frac{\nabla^2\lambda_1}{\lambda_1} + \frac{\nabla\lambda_1\nabla\lambda_1^T}{\lambda_1^2}\right)\\ &= \frac{\kappa}{2}\left(\frac{2\nabla\kappa\nabla\kappa^T}{\kappa^2} + \frac{\nabla^2\lambda_4}{\lambda_4} - \frac{\nabla\lambda_4\nabla\lambda_4^T}{\lambda_4^2} - \frac{\nabla^2\lambda_1}{\lambda_1} + \frac{\nabla\lambda_1\nabla\lambda_1^T}{\lambda_1^2}\right)\\ &= \frac{\kappa}{2}\left(\frac{\nabla^2\lambda_4}{\lambda_4} - \frac{\nabla^2\lambda_1}{\lambda_1} - \frac{2\nabla\lambda_1\nabla\kappa^T}{\lambda_1\kappa} - \frac{2\nabla\kappa\nabla\lambda_1^T}{\kappa\lambda_1} - \frac{2\nabla\kappa\nabla\kappa^T}{\kappa^2}\right),\end{aligned} \tag{4.14}$$

of which the last equality holds for $\frac{\nabla\lambda_4}{\lambda_4} = 2\frac{\nabla\kappa}{\kappa} + \frac{\nabla\lambda_1}{\lambda_1}$ by Eq. (4.13).

5 Solve the problem

Usually optimization methods converge to the local minimal point or stationary point, which would not be the global minimal point. It is well known that optimization methods performing on convex functions always converge to the unique global minimal point. Hence we would like to know whether the objective function is convex in its connected domain. However, numerical tests do not support that.

Example 5.1 (Convex Test)**.** All the numerical tests in the paper are implemented in MATLAB R2017a. We randomly generate three symmetric matrices M, C, K (not necessarily positive definite), where $M = U^{\mathrm{T}}\Lambda U$ in which U is generated by the MATLAB code `rand(10)`, and Λ is diagonal and each diagonal entry follows the uniform distribution on $[-0.5, 0.5]$. Then we pick its two eigenpairs of different types, and calculate the related unconstrained optimization problem equation (3.6).

Then we randomly generate 1000 points where any coordinate follows the uniform distribution $[-100, 100]$. For each point we compute its Hessian matrix $\nabla^2\kappa$, and then compute its minimal eigenvalue $\lambda_{\min}$. Here the gradient $\nabla\kappa$ and the Hessian $\nabla^2\kappa$ are generated according to Eq. (4.13) and Eq. (4.14) respectively. Specially, if the point (nearly) locates in the hyperplane $z_a^{\mathrm{T}} y = \beta$, then we simply discard the pair, or equivalently let the minimal eigenvalue be $+\infty$. Then we find the minimal $\lambda_{\min}$ of the 1000 pairs.

The procedure above will be repeated 1000 times. The distribution of the 1000 minimal $\lambda_{\min}$ is shown below.

$(-10^8, -10^6)$	$(-10^6, -10^4)$	$(-10^4, -10^3)$	$(-10^3, -10^2)$	$(-10^2, -10^0)$
27	294	351	255	73

The numerical test suggests us that the objective function is hardly convex.

Although the objective function may be not convex, the global minimal point can still be converged if it is pseudoconvex. A function $f\colon S\subset\mathbb{R}^n\to\mathbb{R}$ is called pseudoconvex, if $f(y_1)<f(y_2)\Rightarrow(y_1-y_2)^{\mathrm{T}}\nabla f(y_2)<0$ for any $y_1,y_2\in S$. A function $f(y)$ pseudoconvex on S has this property: if $y_0\in S$ is a stationary point, then y_0 is a global minimal point; if y_0 is a local minimal point, then y_0 is a global minimal point. If $\nabla^2 f(y)$ is positive definite on the subspace $\{\nabla f(y)\}^{\perp}$, then f is pseudoconvex. In detail, we will judge whether $Z^H\nabla^2\kappa Z\succ 0$, where Z is an orthonormal basis matrix of the orthogonal complement of $\nabla\kappa$, and $Z^T\nabla\kappa=0$. Specially $Z=I_3$ if $\nabla\kappa=0$. More information can be found in classic textbooks (e.g. [3]), or this early paper [4].

Numerical tests imply the objective function $\kappa_2(T)$ with respect to y is pseudoconvex in its connected domain for a few instances.

Example 5.2 (Pseudoconvex Test)**.** The whole process is nearly the same as in Example 5.1, but rather than the minimal eigenvalue of $\nabla^2\kappa$, we calculate the minimal eigenvalue $\widetilde{\lambda}_{\min}$ of $Z^H\nabla^2\kappa Z$.

The distribution of the 1000 minimal $\widetilde{\lambda}_{\min}$ is shown below.

$(-10^6,-10^2)$	$(-10^2,-10^{-2})$	$(-10^{-2},-10^{-6})$	$(-10^{-6},-10^{-11})$	$(10^{-13},10^{-8})$
97	595	275	26	7

The numerical test suggests us that the objective function is pseudoconvex for a few instances only.

Although in most instances the objective function is likely not to be pseudoconvex, for any given problem, namely given the quadratic matrix polynomial $Q(\lambda)$ and its two eigenpairs to be deflated, we are able to transform it into an equivalent problem whose associated tiny-scale unconstrained optimization problem with a pseudoconvex objective function.

Recall the problem setting in the beginning of Section 2. Note that a is determined by the eigenvectors only. Thus, performing an affine transformation on the eigenvalue does not change either a or the eigenvectors to be deflated. In details, the problem

$$\widetilde{Q}(\lambda)=\frac{1}{\nu_1^2}Q(\nu_1\lambda+\nu_2)=\lambda^2M+\lambda\frac{2\nu_2M+C}{\nu_1}+\frac{\nu_2^2M+\nu_2C+K}{\nu_1^2}=:\lambda^2M+\lambda\widetilde{C}+\widetilde{K} \tag{5.1}$$

will be treated instead. However,

$$\widetilde{\alpha}_C=a^T\widetilde{C}a=\frac{2\nu_2\alpha_M+\alpha_C}{\nu_1},\quad \widetilde{\alpha}_K=a^T\widetilde{K}a=\frac{\nu_2^2\alpha_M+\nu_2\alpha_C+\alpha_K}{\nu_1^2}.$$

It is not difficult to make $\widetilde{\alpha}_C=0,|\alpha_M|=|\widetilde{\alpha}_K|$. In fact,

$$\nu_2=-\frac{\alpha_C}{2\alpha_M},\quad \nu_1=\frac{\sqrt{|4\alpha_M\alpha_K-\alpha_C^2|}}{2\alpha_M}\neq 0.$$

Moreover, write $\chi=\text{sign}(4\alpha_M\alpha_K-\alpha_C^2)$, and then $\widetilde{\alpha}_K=\chi\alpha_M$.

Since the deflation works on eigenvectors only, the tiny-scale unconstrained optimization problem must have the same form, except the terms involving α_C,α_K. Hence in the following, we will directly use α_C,α_K rather than $\widetilde{\alpha}_C,\widetilde{\alpha}_K$, and all the notation will be kept. Now

$$\delta=\frac{1}{\alpha_M^2\chi},\quad S=\alpha_M\begin{bmatrix}1 & \\ & \chi\end{bmatrix}.$$

So $|\alpha_M|=\frac{1}{\sqrt{|\delta|}}$. The involving hyperbolic quadratic polynomial, namely Eq. (4.1), is

$$-JH(\lambda)J=\lambda^2I_2-\lambda[I_2+SWS+\zeta^2S^2]+\zeta^2S^2=\lambda^2I_2-\lambda\Big[\Big(1+\frac{\zeta^2}{\chi\delta}\Big)I_2+SWS\Big]+\frac{\zeta^2}{\chi\delta}I_2.$$

Since I_2 and SWS can be simultaneously diagonalizable by an orthonormal matrix, that hyperbolic quadratic polynomial is diagonalizable by an orthonormal matrix, which implies: 1) an eigenvector corresponding to a positive-type eigenvalue has to be that corresponding to a negative-type eigenvalue; 2) the eigenvectors corresponding to two positive-type eigenvalues are orthogonal. There are two cases to be considered.

Case 1: λ_1,λ_4 cannot share a same eigenvector, which implies that λ_1,λ_3 share a same eigenvector x_1 and λ_2,λ_4 share a same eigenvector x_4 satisfying $x_4^Tx_1=0$.

Since $x_1^Tx_1=x_4^Tx_4=1$, without loss of generality, assume $x_1=Jx_4$. Then $t_4^Tt_4=x_4^TS^2x_4=\frac{1}{\chi\delta}$, and similarly $t_1^Tt_1=\frac{1}{\chi\delta}$. Thus, $\lambda_1\lambda_3=\frac{\zeta^2}{\chi\delta}=\lambda_2\lambda_4$ and then $\frac{\lambda_1}{\lambda_2}=\frac{\lambda_4}{\lambda_3}$, which implies $\lambda_1=\lambda_2$, $\lambda_3=\lambda_4$. This tells λ_1,λ_4 share a same eigenvector vector, a contradiction.

Case 2: λ_1,λ_4 share a same eigenvector. Since $x_1^Tx_1=x_4^Tx_4=1$, without loss of generality, assume $x_1=x_4$, then $t_1=t_4$, $s_1=s_4$. Then $t_4^Ts_4=x_4^TS^2Jx_4=0$, $t_4^Tt_4=x_4^TS^2x_4=\frac{1}{\chi\delta}$. Also, $\sigma_4=-\sigma_1=\lambda_4-\lambda_1$, $g_{t_4}=g_{t_1}=4\chi\delta^2\sigma_4\lambda_4B_WE_{t_4}Jt_4$ by Eq. (4.9).

By Eqs. (4.5) and (4.13),

$$\begin{aligned}\frac{\nabla\kappa}{\kappa}&=\frac{1}{2}\Big(\frac{\nabla\lambda_4}{\lambda_4}-\frac{\nabla\lambda_1}{\lambda_1}\Big)=\frac{g_4}{\sigma_4\lambda_4}-\frac{g_1}{\sigma_1\lambda_1}\\&=\delta^2B_W\Big[\frac{t_4^Tt_4}{\sigma_4}y+\frac{1}{2\sigma_4}E_{t_4}t_4\Big]+\Big(1-\frac{1}{\lambda_4}\Big)\frac{t_4^Tt_4}{\sigma_4}\xi z_a\\&\quad-\delta^2B_W\Big[\frac{t_1^Tt_1}{\sigma_1}y+\frac{1}{2\sigma_1}E_{t_1}t_1\Big]-\Big(1-\frac{1}{\lambda_1}\Big)\frac{t_1^Tt_1}{\sigma_1}\xi z_a\\&=\delta^2B_W\Big[\Big(\frac{1}{\chi\delta\sigma_4}-\frac{1}{\chi\delta\sigma_1}\Big)y+\frac{1}{2\sigma_4}E_{t_4}t_4-\frac{1}{2\sigma_1}E_{t_4}t_4\Big]\\&\quad+\Big[\frac{1}{\chi\delta\sigma_4}-\frac{1}{\chi\delta\sigma_1}-\frac{1}{\chi\delta\lambda_4\sigma_4}+\frac{1}{\chi\delta\lambda_1\sigma_1}\Big]\xi z_a\\&=\frac{1}{\chi\delta\sigma_4}\Big(\delta^2B_W[2y+\chi\delta E_{t_4}t_4]+\Big[2-\Big(\frac{1}{\lambda_4}+\frac{1}{\lambda_1}\Big)\Big]\xi z_a\Big).\end{aligned}$$

By $Z^T\frac{\nabla\kappa}{\kappa}=0$,

$$\delta^2 Z^T B_W[2y+\chi\delta E_{t_4}t_4]=\Big[\frac{1}{\lambda_4}+\frac{1}{\lambda_1}-2\Big]\xi Z^T z_a,$$

and then

$$\begin{aligned}\frac{2\xi}{\sigma_4\lambda_4}Z^T g_4&=\frac{2\xi}{\sigma_4\lambda_4}\left(\lambda_4\delta^2 Z^T B_W\Big[\frac{1}{\chi\delta}y+\frac{1}{2}E_{t_4}t_4\Big]+\frac{\lambda_4-1}{\chi\delta}\xi Z^T z_a\right)\\&=\frac{2\xi}{\sigma_4\lambda_4}\left(\frac{\lambda_4}{2\chi\delta}\Big[\frac{1}{\lambda_4}+\frac{1}{\lambda_1}-2\Big]\xi Z^T z_a+\frac{\lambda_4-1}{\chi\delta}\xi Z^T z_a\right)\\&=\frac{\xi^2}{\chi\delta\sigma_4\lambda_4}\left(\frac{\lambda_4}{\lambda_1}-1\right)Z^T z_a\\&=\frac{\lambda_1\lambda_4}{(\lambda_4-\lambda_1)\lambda_4}\left(\frac{\lambda_4}{\lambda_1}-1\right)Z^T z_a\qquad\text{by }\frac{\xi^2}{\chi\delta}=\lambda_1\lambda_4\\&=Z^T z_a.\end{aligned}$$

Similarly, $\frac{2\xi}{\sigma_1\lambda_1}Z^T g_1=Z^T z_a$. Thus, by Eqs. (4.12) and (4.14),

$$Z^T\nabla^2\kappa Z=\frac{\kappa}{2}Z^T\left(\frac{\nabla^2\lambda_4}{\lambda_4}-\frac{\nabla^2\lambda_1}{\lambda_1}\right)Z=\kappa Z^T\left(\frac{2}{\chi\delta\sigma_4}z_a z_a^T+\frac{1}{\sigma_4^3}\Big[\frac{1}{\lambda_4\omega_4}+\frac{1}{\lambda_1\omega_1}\Big]g_{t_4}g_{t_4}^T\right)Z\succeq 0.$$

Moreover, in the generic case, except several critical points

$$\operatorname{rank}(Z^T\nabla^2\kappa Z)=\operatorname{rank}(Z^T\begin{bmatrix}z_a & g_{t_4}\end{bmatrix})=2,$$

which implies $Z^T\nabla^2\kappa Z\succ 0$.

To sum up, the objective function of this transformed problem equation (5.1) is pseudoconvex. Then, optimization methods would effectively attain the global minimum in either of the two connected domains. When using optimization method to solve the problem, we may start at one initial y with $z_a^{\mathrm{T}}y>\beta$ and another initial y with $z_a^{\mathrm{T}}y<\beta$. As long as the problem is pseudoconvex, the two numerical solutions come out fast. After that, we may pick the better one in the two. Please note that this transforming technique is implemented for the pseudoconvexity, so it would produce a structure-preserving transformation for the deflation whose condition is worse than that of the untransformed one.

Example 5.3 (Convergence Test)**.** The unconstrained optimization problem equation (3.6) is built as in Example 5.1.

Then we randomly generate 100 points where any coordinate follows the uniform distribution $[-100,100]$. Then we use each point as the initial point and perform both the quasi-Newton method and the trust-region method to observe the convergence behavior. All the options of the methods are set in default by MATLAB. Among those, the tolerance of the difference of the function value is 10^{-6}, and the iteration will stop if the number of iterations reaches 300.

The procedure above will be repeated 1000 times. Two rounds of the procedure are made for the untransformed problem and the transformed problem. The data of the numerical tests are shown below.

problem	method	domain	min attained	iterations			func. evaluations		
				max	min	mean	max	min	mean
untransformed	trust-region	$z_a^{\mathrm{T}}y>\beta$	502	300	10	28.2	301	11	29.2
		$z_a^{\mathrm{T}}y<\beta$	498	300	11	28.5	301	12	29.5
	quasi-Newton	$z_a^{\mathrm{T}}y>\beta$	489	68	1	13.1	280	14	43.3
		$z_a^{\mathrm{T}}y<\beta$	511	70	1	13.0	187	13	43.2
transformed	trust-region	$z_a^{\mathrm{T}}y>\beta$	468	300	4	28.0	301	5	29.5
		$z_a^{\mathrm{T}}y<\beta$	532	300	5	28.4	301	6	29.4
	quasi-Newton	$z_a^{\mathrm{T}}y>\beta$	496	65	1	13.4	300	8	39.2
		$z_a^{\mathrm{T}}y<\beta$	504	70	1	13.5	300	10	39.1

From the data, we know that generally the quasi-Newton method uses fewer iterations but more function evaluations in the sense of the average case and the worst case. However, the trust-region method needs the (numerical) Hessian. Since the number of function evaluations needed by the quasi-Newton method does not exceed twice of that need by the trust-region method, the quasi-Newton method is faster than the trust-region methods according to the data.

It seems that in many cases the minimums are attained in different connected domain by the two methods. Actually in fact, the minimums are nearly the same, which can be illustrated by the distribution of the relative difference of the 1000 pairs of minimums obtained by trust-region and quasi-Newton, namely $\frac{\kappa_{\text{trust-region}}-\kappa_{\text{quasi-Newton}}}{\kappa_{\text{trust-region}}}$ given below.

problem	$(-2,-1)$	$(-1,-10^{-2})$	$(-10^{-2},0)$	$(0,10^{-2})$	$(10^{-2},1)$
untransformed	0	153	44	115	688
transformed	1	33	39	365	562

From the data, we can see the minimums found by the trust-region method is between half and twice those found by the quasi-Newton method, which would produce little affect on the condition number; in most cases, the quasi-Newton method is a little better than the trust-region.

On the other hand, from the data above, we know performing on the transformed problem costs slightly less than performing on the untransformed problem. However, if we compare the minimums found for the untransformed and transformed problems, namely $\frac{\kappa_{\text{untransformed}}}{\kappa_{\text{transformed}}}$ give below,

$(10^{-5},0.1)$	$(0.1,1)$	$(1,10)$	$(10,100)$	$(100,1000)$	$(1000,10^5)$
30	123	299	396	135	17

we may discover that in most cases performing on the transformed problems is likely to find out a much better structure-preserving transformation than performing on the untransformed problems.

To sum up, a good choice is to use the quasi-Newton method to solve the unconstrained optimization problem Eq. (3.6) induced by the transformed quadratic matrix polynomial.

Besides, from the data, we can see the minimum has no preference on the connected domain, which makes us have to compute on them both.

6 Concluding remarks

We have built a tiny-scale unconstrained optimization problem and suggested using optimization methods to solve it, which make it successful to find an optimal deflation for symmetric quadratic matrix polynomials. This would be a good answer to the question asked by Tisseur et al. [6]. However, it is quite natural to ask whether this method can be used for asymmetric problems. It is likely that some analogues would hold, but to ensure this is left for future work. In addition, investigating the theoretical solution to the unconstrained optimization problem would be also valuable to consider in future.

Acknowledgments

Part of the work was done when the author worked in Max Planck Institute for Dynamics of Complex Technical Systems. The author would like to thank Prof. Peter Benner and Prof. Sara Grundel for the discussion, and thank the anonymous referees for the comments that spur the author to survey further.

References

[1] X. Liang and R.-C. Li. The hyperbolic quadratic eigenvalue problem. *Forum of Mathematics, Sigma*, 3(e13), 2015.

[2] K. Veselić. Note on interlacing for hyperbolic quadratic pencils. In J.Behrndt, K.-H. Förster, and C. Trunk (Eds.), *Recent Advances in Operator Theory in Hilbert and Krein Spaces*, 198:305-307, 2010.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*, Cambridge University Press, 2004.

[4] O. L. Mangasarian. Pseudo-convex functions. *SIAM J. Control*, 3(2):281–290, 1965.

[5] B. Meini. A "shift-and-deflate" technique for quadratic matrix polynomials. *Linear Algebra Appl.*, 438:1946–1961, 2013.

[6] F. Tisseur, S. D. Garvey, and C. Munro. Deflating quadratic matrix polynomials with structure preserving transformations. *Linear Algebra Appl.*, 435:464–479, 2011.