SMITH METHOD FOR PROJECTED LYAPUNOV AND STEIN EQUATIONS

Bo YU*[1], Hung-Yuan FAN,[2], Eric King-Wah CHU[3]

We consider the projected Lyapunov and Stein equations arising in model order reduction and optimal control of descriptor systems. The projected Lyapunov equation is transformed to an equivalent projected Stein equation then solved by a generalized Smith iterative method. For a projected general Stein equation with a singular matrix "E", a double Cayley transform is devised to remove the singularity, and then the generalized Smith method is applied. Numerical examples are provided to demonstrate the feasibility and efficiency of our approach.

Keywords: Cayley transform, descriptor system, double Cayley transform, projected Lyapunov equation, projected Stein equation.
MSC2000: 15A24, 65F99, 93C05

1 Introduction

Consider the projected Lyapunov equation (PLE) [1, 4, 23, 24, 25]:

$$EXA^\top + AXE^\top + Q = 0, \quad X = P_r X P_r^\top, \tag{1}$$

where $Q = P_l B B^\top P_l^\top, E, A, X \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ with $m \ll n$ (m far less than n) and P_r, P_l are the spectral projectors onto the right and left deflating subspaces of the matrix pencil $\lambda E - A$ corresponding to the finite eigenvalues along the deflating subspaces associated with the infinite eigenvalues. The definition of the deflating subspace for a matrix pencil, similar to the notion of the invariant subspace for a single matrix, is stated as follows.

Definition 1 ([8]). *Let A and B be $n \times n$ matrices. The k-dimensional subspace $S \subseteq \mathbb{R}^n$ is a deflating subspace for the matrix pencil $A-\lambda B$ if the subspace $\{Ax+By : x, y \in S\}$ has dimension k or less.*

[1]School of Science, Hunan University of Technology, Zhuzhou, 412008, P.R. China. e-mail: wenyubwenyub@aliyun.com. (∗ Corresponding Author)
[2]Department of Mathematics, National Taiwan Normal University, Taipei 116, Taiwan. e-mail: hyfan@math.ntnu.edu.tw.
[3]School of Mathematical Sciences, Monash University, 9 Rainforest Walk, Victoria 3800, Australia. e-mail: eric.chu@monash.edu.

The PLE (1) mainly arises from the periodic descriptor systems [22] of model order reduction and optimal control [1, 4, 16] and can be viewed as a generalization of the Lyapunov equation

$$XA^\top + AX + Q = 0. \tag{2}$$

There is an abundance of literature for the computation of the solution of Lyapunov equations [2, 3, 5, 6, 7, 10, 11, 13, 14, 17, 18, 19, 20, 26]. In general, the Cayley transformation is used for converting (2) into the so-called Stein equation [14]

$$AXA^\top - X + Q = 0, \tag{3}$$

and then the Smith method [21]

$$X_{k+1} = Q + AX_kA^\top \tag{4}$$

with $X_0 = Q$ is employed for calculating the corresponded solution when some stability condition in descriptor systems is available.

Definition 2 ([16]). *A matrix A (or matrix pencil $A - \lambda B$) is c-stable if its all eigenvalues (or finite eigenvalues) lie in the open left half complex plane and d-stable if its all eigenvalues (or finite eigenvalues) lie inside the unit circle in complex plane.*

It is known in [14] that if A in (2) is c-stable, then the Cayley-transformed matrix A in (3) becomes d-stable. So the Smith method (4) is definitely convergent. For the PLE (1) in periodic descriptor systems, we similarly assume that $\lambda E - A$ is c-stable, making the PLE uniquely solvable [24, 25]. Some numerical methods have been considered successfully for the solution of PLE [1, 24, 25, 25] and the associated Stein equations [4, 5, 16]. However, the Smith method (4) seems merely to be linearly convergent when computing the corresponded solution. Additionally, for the projected general Stein equation (PGSE)

$$AXA^\top - EXE^\top + Q = 0, \quad X = P_rXP_r^\top, \quad Q = P_lBB^\top P_l^\top, \tag{5}$$

direct application of Smith method might be invalid as E is possibly singular. In this paper, we design a double Cayley transform to remove the singularity of E and give a generalized Smith method that can retain the quadratic convergence.

The rest of the paper is organized as follows. Section 2 reviews the Cayley transform of PLE into projected Stein equation and gives a generalized version of the Smith method. The concrete algorithm and detailed implementations for large-scale projected Stein equation are shown in Section 3. The mutual transformation between the PLE and the PGSE is stated in Section 4 and the double Cayley transformation follows in Section 5. Numerical experiments are reported in the last section to show the efficiency of the proposed algorithm for computing the solution of large-scale PLE and PGSE.

2 Cayley Transform and Generalized Smith Method

By the Cayley transform with $\gamma > 0$, PLE (1) is equivalent to

$$(A+\gamma E)X(A+\gamma E)^\top - (A-\gamma E)X(A-\gamma E)^\top = 2\gamma(EXA^\top + AXE^\top) = -2\gamma Q,$$

or the projected Stein equation (PSE)

$$\mathcal{S}(X) := \widetilde{A}X\widetilde{A}^\top - X + \widetilde{Q} = 0, \quad X = P_r X P_r^\top, \tag{6}$$

with $A_\gamma := A-\gamma E$ nonsingular, $\widetilde{A} := A_\gamma^{-1}A_{-\gamma} = I+2\gamma A_\gamma^{-1}E$, $\widetilde{Q} := 2\gamma A_\gamma^{-1}QA_\gamma^{-\top} = \widetilde{B}\widetilde{B}^\top$ and $\widetilde{B} := \sqrt{2\gamma}A_\gamma^{-1}P_l B$.

Insert $X = \widetilde{A}X\widetilde{A}^\top + \widetilde{Q}$ into $X = P_r X P_r^\top$ in (6), we have

$$X = P_r\left[\sum_{j=0}^{\infty}\widetilde{A}^j\widetilde{Q}(\widetilde{A}^\top)^j\right]P_r^\top = \sum_{j=0}^{\infty}P_r\widetilde{A}^j\widetilde{Q}(\widetilde{A}^\top)^j P_r^\top. \tag{7}$$

Proposition 3. *If $\widetilde{A}$ in* (6) *is semi-d-stable (with all finite eigenvalues on or inside the unit circle $\mathcal{D}$) and P_r is associated with the d-stable sub-spectrum, the sum in* (7) *is convergent.*

Proof. Let

$$\widetilde{A} = [P_{r1}, \widetilde{P}_{r1}]\begin{bmatrix}\Lambda_s & 0\\ 0 & \Lambda_1\end{bmatrix}[P_{r2}, \widetilde{P}_{r2}]^\top$$

be the eigen-decomposition of $\widetilde{A}$, where diagonal matrices $\Lambda_s \in \mathbb{R}^{r_1\times r_1}$ and $\Lambda_1 \in \mathbb{R}^{(n-r_1)\times(n-r_1)}$ contain respectively the d-stable and unimodular eigenvalues of $\widetilde{A}$, the corresponded eigenvector matrix is $[P_{r1}, \widetilde{P}_{r1}]$ with sub-blocks $P_{r1} \in \mathbb{R}^{n\times r_1}$, $\widetilde{P}_{r1} \in \mathbb{R}^{n\times(n-r_1)}$ and its inverse denoting by $[P_{r2}, \widetilde{P}_{r2}]^\top$ with sub-blocks $P_{r2} \in \mathbb{R}^{n\times r_1}$, $\widetilde{P}_{r2} \in \mathbb{R}^{n\times(n-r_1)}$ (i.e. $[P_{r1}, \widetilde{P}_{r1}]\times[P_{r2}, \widetilde{P}_{r2}]^\top = [P_{r2}, \widetilde{P}_{r2}]^\top \times [P_{r1}, \widetilde{P}_{r1}] = I_n$). Now set $P_r = P_{r1}P_{r2}^\top$ and filter off the unimodular sub-spectrum, we have

$$\begin{aligned} X &= P_r X P_r^\top = P_{r1}P_{r2}^\top\left[\sum_{j=0}^{\infty}\widetilde{A}^j\widetilde{Q}(\widetilde{A}^\top)^j\right]P_{r2}P_{r1}^\top \\ &= P_{r1}\left[\sum_{j=0}^{\infty}\Lambda_s^j P_{r2}^\top\widetilde{Q}P_{r2}\left(\Lambda_s^\top\right)^j\right]P_{r1}^\top. \end{aligned} \tag{8}$$

Obviously, the above sum converges as the spectral radius of Λ_s less than one [1, 4, 24, 25]. □

Furthermore, the generalized Smith method (GSM) can be deduced from (8) as follows. Let $X_0 = P_r\widetilde{Q}P_r^\top$ and $A_0 = \widetilde{A}$ with $\widetilde{Q}$ and $\widetilde{A}$ defined in (6), consider the iteration

$$A_{k+1} = A_k^2, \quad X_{k+1} = X_k + P_r A_k X_k A_k^\top P_r^\top \quad (k \geq 0), \tag{9}$$

with $A_k v$ constructed recursively, not explicitly. Again, the projections P_r and $P_r^\top$ in (9) are filtering off components associated with the unimodular subspectrum so that $\{X_k\}$ converges.

It is no difficult to show that (or see [12] for example), if

$$Z_k = P_r \left[\sum_{j=0}^{k} \widetilde{A}^j \widetilde{Q} (\widetilde{A}^j)^\top \right] P_r^\top \quad (k \geq 0),$$

and $X_0 = Z_0 = P_r \widetilde{Q} P_r^\top$, the k-th iteration of GSM (9) has

$$X_k = P_r \left[\sum_{j=0}^{2^k-1} \widetilde{A}^j \widetilde{Q} (\widetilde{A}^j)^\top \right] P_r^\top = Z_{2^k-1} \quad (k \geq 0),$$

which implies a faster quadratic convergence.

3 Algorithm for Large-Scale Problems

For large-scale PLE (1), we assume A and E are sparse, $m \ll n$ and products of A_γ, P_r, P_l and their transposes with $v \in \mathbb{C}^n$ can be computed efficiently in $O(n)$ complexity. Products of A_γ^{-1} and given vectors can be obtained in n_s flops for one solve of the linear system defined by A_γ, where n_s is reasonably small for structured A. Then the GSM (9) can be applied to solve PLE or PGSE, after appropriate Cayley or double Cayley (see Section 5) transforms, in $O(2^k n) + O(2^k n_s)$ flops at the kth iteration. Except in extraordinarily bad examples, the method quickly converges in k iterations, the operation count will be relative low, with acceptably small values for the coefficient 2^k. If the solution X is numerically low-ranked (i.e. for the given tolerance $\tau > 0$, $rank_\tau(X) \leq c_\tau$ for a constant c_τ independent of n), we have the following result with its neglected proof similar to that of [14, Thm 2.1].

Theorem 4. *Let the GMS algorithm converge in k iterations, according to a given accuracy tolerance. For a given tolerance $\tau > 0$, the projected Stein equation has a numerically low-ranked solution X relative to n when $2^k = O(1)$.*

The relationship between some Krylov subspaces and the numerically low-ranked solution in (8) is given in [14]. Similarly, the GSM (9) keeps adding diminishing low-ranked components to the approximate solution, whose growth of the rank can be controlled by the truncation and compression procedure, implemented using QR decomposition by column pivoting as follows.

3.1 Truncation and Compression

Recall $\widetilde{Q} = \widetilde{B}\widetilde{B}^\top$, the GSM (9) for large-scale system is essential the growth of the numerical ranks in $X_k = C_k T_k C_k^\top$, with T_k being symmetric and invertible. As the

GSM converges, increasingly smaller but higher-rank components are added to C_k. Apparent from (9), the growth in the sizes and ranks of these iterates is potentially exponential. To reduce the dimensions of C_k, we shall compress their columns by orthogonalization. As in [6, 13, 14], consider the economic QR decompositions with column pivoting:

$$C_k = Q_k U_k + \widetilde{Q}_k \widetilde{U}_k, \quad \|\widetilde{U}_k\| \leq \tau.$$

From here on, all norms are the 2-norm. Here τ is some small tolerance controlling the compression and truncation process, n_k is the number of columns in C_k bounded from above by $l_{\max}$ and its rank satisfies $r_k^{(b)} := \operatorname{rank} C_k \leq n_k \leq l_{\max} \ll n$. Also $Q_k \in \mathbb{R}^{n \times r_k^{(b)}}$ is unitary and $U_k \in \mathbb{R}^{r_k^{(b)} \times n_k}$ is full-rank and upper triangular. We have

$$X_k = C_k T_k C_k^\top = Q_k \left(U_k T_k U_k^\top \right) Q_k^\top + O(\tau), \tag{10}$$

and we should replace C_k and T_k by the leaner Q_k and $U_k T_k U_k^\top$. As a result, we ignore the $O(\tau)$ term, control the growth of $r_k^{(b)}$ while sacrificing a hopefully negligible $O(\tau)$ bit of accuracy. We also restrict the widths of C_k, now relabelled $l_k = r_k^{(b)}$ after the compression and truncation in (10), by setting a reasonable upper limit $l_{\max}$.

One beneficial consequence of the truncation and compression process, for the orthogonality of C_k, is the simplified evaluation:

$$\|X_k\| = \|C_k T_k C_k^\top\| + O(\tau) = \|T_k\| + O(\tau). \tag{11}$$

3.2 Computational Issues and Algorithms

The concrete algorithm for solving the transformed PSE (6) is described in Algorithm 1. Note the computational issues, flop counts and error analysis are analogous to those in [14], only with some modifications for the filtering by P_r. From our experience, the shift parameter γ in the Cayley transform is not critical and can be chosen efficiently by trial-and-error.

Algorithm 1 (GSM for Large-Scale Projected Stein Equations)

Input: $\widetilde{A} \in \mathbb{R}^{n \times n}$, $\widetilde{B} \in \mathbb{R}^{n \times l}$; positive tolerances τ and ϵ, and $l_{\max}$;
Output: $C_\epsilon \in \mathbb{R}^{n \times l_\epsilon}$ and $T_\epsilon = T_\epsilon^\top \in \mathbb{R}^{l_\epsilon \times l_\epsilon}$, with $C_\epsilon T_\epsilon C_\epsilon^\top$ approximating X;
Compute the QR decomposition $\widetilde{B} = CR$, $R \in \mathbb{R}^{l \times l}$;
Set $k = 0$, $\widetilde{r}_0 = 2\epsilon$; $A_0 = \widetilde{A}$, $C_0 = P_r C$ and $T_0 = RR^\top$;
Compute $h = \|\widetilde{Q}\| = \|T_0\|$;
Do until convergence:
  Compute $r_k = \|\mathcal{S}(X_k)\|$, $h_k = \|X_k\|$ and $m_k = \|\widetilde{A} X_k \widetilde{A}^\top\|$;
  If the relative residual $\tilde{r}_k = |r_k/(h_k + m_k + h)| < \epsilon$,
    Set $C_\epsilon = C_k$ and $T_\epsilon = T_k$;
    Exit
  End If
  Compute $C_{k+1} = [C_k,\ P_r A_k^\top C_k]$, $T_{k+1} = T_k \oplus T_k$, with $A_{k+1} = A_k^2$;
  Compress C_{k+1}, using the tolerance τ, and modify T_{k+1} accordingly;
  Set $k \leftarrow k + 1$;
End Do

For the iterate X_k, the residual r_k and relative residual $\widetilde{r}_k$ for (6) are defined as

$$r_k := \|\mathcal{S}(X_k)\|, \quad \widetilde{r}_k := \frac{r_k}{\|\widetilde{Q}\| + \|\widetilde{A}X_k\widetilde{A}^\top\| + \|X_k\|}. \tag{12}$$

Remark 1. (i) The direction sum $A \oplus B$ denotes the matrix diag$\{A, B\}$.

(ii) Note $X_k = C_k T_k C_k^\top$ with $C_k \in \mathbb{R}^{n\times l_k}$ and $T_k \in \mathbb{R}^{l_k\times l_k}$, then the residual $\mathcal{S}(X_k) = \widehat{C}_k\widetilde{T}_k\widehat{C}_k^\top$ with $\widehat{C}_k = [\widetilde{B}, C_k, \widetilde{A}C_k] \in \mathbb{R}^{n\times(2l_k+l)}$ and $\widetilde{T}_k = I_l \oplus (-T_k) \oplus T_k \in \mathbb{R}^{(2l_k+l)\times(2l_k+l)}$, the QR decomposition on $\widehat{C}_k$ with $O(n)$ complexity is used for the computation of r_k. After a similar orthogonalization procedure as in Section 3.1, the norms of $\mathcal{S}(X_k)$ can be obtained efficiently as in (11). Also as $\widetilde{A}X_k\widetilde{A}^\top = \widetilde{A}C_kT_k(\widetilde{A}C_k)^\top$, similar comments hold for m_k.

(iii) The width of C_{k+1} nearly doubled that of C_k, so $l_{\max} = 2^k l$. For a given compression accuracy τ in our numerical experiments, the actual rank of the compressed C_k is much less than $l_{\max}$.

(iv) The shift parameter γ in the Cayley transform is not critical and can be chosen efficiently by trial-and-error.

(v) To avoid computing and storing the large A_{k+1}, Algorithm 1 works with the thin and tall matrix $A_k^\top C_k$ at step k, requiring two products $A_{k-1}^\top C_k$ and $A_{k-1}^\top(A_{k-1}^\top C_k)$ for a given C_k. For each column of C_k, we also require at most $O(2^k)$ products $A_\gamma^{-1}v$ (for some vector v), or at most $O(2^k)$ linear solves associating with A_γ. Fortunately, k is normally small from the fast convergence of the algorithm (see also in Section 6). Also, the last iteration in the Algorithm 1 is virtually free because there is no need to prepare C_{k+1}. This together with Theorem 3.1 imply that Algorithm 1 is expect to be of $O(n) + O(n_s)$ computational complexity for the kth iteration for large-scale problems. We have the following result with an analogous proof to [14, Cor 6.1].

Proposition 5. *Let the Algorithm 1 method converge after k iterations to an approximate solution X_k with rank$(X_k) \le 2^k l = O(1)$, according to a given accuracy tolerance. Then the Algorithm 1 has an $O(n)$ computational complexity and memory requirement.*

4 Projected General Stein Equations

Consider the projected general Stein equation (5). This equation can be solved numerically by the GSM (9) when E is nonsingular and well-conditioned, after being transformed to PSE in (6). Otherwise, the direct application of GSM (9) might be invalid as the singularity of E. We will devise a double Cayley transform in Section 5 to remove the singularity. Before that, a preliminary theorem is given as follows (the general complex conjugate transpose with $(\cdot)^H$ is described and γ should be chosen to be real for the special case with $(\cdot)^\top$).

Theorem 6. *(i) Assume that $\lambda E - A$ is c-stable. The uniquely solvable PLE:*

$$AXE^H + EXA^H + Q = 0, \quad X = P_r X P_r^H,$$

is equivalent to the uniquely solvable PGSE:

$$A_{-\overline{\gamma}} X A_{-\overline{\gamma}}^H - A_\gamma X A_\gamma^H + 2\Re e(\gamma) Q = 0, \quad X = P_r X P_r^H,$$

with the real part $\Re e(\gamma) > 0$. *Furthermore,* $A_\gamma := A - \gamma E$ *is c-stable (and nonsingular) with* $A_\gamma^{-1} A_{-\overline{\gamma}}$ *being semi-d-stable, with the original infinite eigenvalues mapped to unity.*

(ii) Assume that $\lambda E - A$ *is d-stable. The uniquely solvable PGSE:*

$$AXA^H - EXE^H + Q = 0, \quad X = P_r X P_r^H,$$

is equivalent to the uniquely solvable PLE:

$$A_{-\gamma} X A_\gamma^H + A_\gamma X A_{-\gamma}^H + 2Q = 0, \quad X = P_r X P_r^H,$$

with $|\gamma| = 1$. *Furthermore,* A_γ *is nonsingular with* $\sigma(A_\gamma^{-1} A_{-\gamma}) \subset \mathbb{C}_- \cup \{1\}$, *with the original infinite eigenvalues mapped to unity, here* $\mathbb{C}_-$ *is the open left half complex plane.*

Proof. The results can be deduced from the following equalities:
(for $\Re e(\gamma) > 0$)

$$\begin{aligned} & A_{-\overline{\gamma}} X A_{-\overline{\gamma}}^H - A_\gamma X A_\gamma^H \\ = \; & (A + \overline{\gamma} E) X (A + \overline{\gamma} E)^H - (A - \gamma E) X (A - \gamma E)^H \\ = \; & 2\Re e(\gamma)(EXA^H + AXE^H) = -2\Re e(\gamma) Q, \end{aligned}$$

and (for $|\gamma| = 1$)

$$\begin{aligned} & A_{-\gamma} X A_\gamma^H + A_\gamma X A_{-\gamma}^H \\ = \; & (A + \gamma E) X (A - \gamma E)^H + (A - \gamma E) X (A + \gamma E)^H \\ = \; & 2(AXA^H - EXE^H) = -2Q. \end{aligned}$$

The invertibility of A_γ and the stability of $\lambda A_\gamma - A_{-\gamma}$ or $\lambda A_\gamma - A_{-\overline{\gamma}}$ can be easily verified by routine manipulation. The solvability of PLEs and PSEs comes from standard classical results for linear matrix equations. □

From the above theorem, the PLE (1) is transformed to PSE (6) and the GSM (9) is applied, as discussed previously. As for general PGSE (5) with E possibly singular, it is first transformed into PLE, as suggested in (ii) in Theorem 6, then followed by a transform back to an equivalent PGSE with nonsingular E and finally reduced to a PSE and solved by the GSM. Such a process is called the double Cayley transform in next section.

5 Double Cayley Transform

We devise a double Cayley transform to remove the singularity of E or A, then the GSM (9) can apply. We consider the more general complex cases for the transforms PGSE→PLE→PGSE and PLE→PGSE→PLE. Note that the double Cayley transforms here only involve linear combinations of A and E with minimal costs, especially when the transformed pencils are not explicitly formed or stored.

5.1 PGSE→PLE→PGSE

Applying firstly (ii) with $\gamma = e^{i\theta}$ and then (i) with $\Re(\gamma) > 0$ in Theorem 6, the PGSE

$$AXA^H - EXE^H + Q = 0, \quad X = P_r X P_r^H,$$

is equivalent to the PGSE:

$$\widehat{A}X\widehat{A}^H - \widehat{E}X\widehat{E}^H + 4\Re e(\gamma)Q = 0, \quad X = P_r X P_r^H,$$

where $\widehat{A} \equiv (1+\bar{\gamma})A + e^{i\theta}(1-\bar{\gamma})E$ and $\widehat{E} \equiv (1-\gamma)A + e^{i\theta}(1+\gamma)E$. Clearly, the matrix pencil $\lambda\widehat{E} - \widehat{A}$ returns to the original $2(e^{i\theta}\lambda E - A)$ when $\gamma = 1$. But for any positive $\gamma \neq 1$, any eigenvalue λ of $\lambda E - A$ are mapped to $\widehat{\lambda}$ of $\lambda\widehat{E} - \widehat{A}$ with

$$\widehat{\lambda} = \frac{(1+\bar{\gamma})\lambda + e^{i\theta}(1-\bar{\gamma})}{(1-\gamma)\lambda + e^{i\theta}(1+\gamma)}, \quad (\lambda \text{ finite}); \tag{13}$$

$$\widehat{\lambda}_u = \frac{1+\bar{\gamma}}{1-\gamma}, \quad (\lambda \text{ infinite}). \tag{14}$$

Let $\gamma = \gamma_r + \gamma_i i$. The denominator $(1-\gamma)\lambda + e^{i\theta}(1+\gamma)$ in (13) will not vanish. Otherwise,

$$\begin{aligned} \lambda &= e^{i\theta}\frac{1+\gamma}{1-\gamma} = e^{i\theta}\frac{1-|\gamma|^2+2\gamma_i i}{1+|\gamma|^2-2\gamma_r}, \\ |\lambda|^2 &= \frac{1+|\gamma|^4+2\gamma_i^2-2\gamma_r^2}{1+|\gamma|^4+2\gamma_i^2-2\gamma_r^2-\delta} \geq 1. \end{aligned} \tag{15}$$

Moreover, it follows from (14) that

$$\begin{aligned} \widehat{\lambda}_u &\equiv \frac{1+\bar{\gamma}}{1-\gamma} = \frac{1-\gamma_r^2+\gamma_i^2+2\gamma_r\gamma_i i}{1+|\gamma|^2-2\gamma_r}, \\ |\widehat{\lambda}_u|^2 &= \frac{(1-\gamma_r^2+\gamma_i^2)^2+4\gamma_r^2\gamma_i^2}{(1-\gamma_r^2+\gamma_i^2)^2+4\gamma_r^2\gamma_i^2-\delta}, \end{aligned}$$

where $\delta \equiv 4\gamma_r\left[(\gamma_r-1)^2+\gamma_i^2\right] \geq 0$, implying that $|\widehat{\lambda}_u|^2 \geq 1$, and $\widehat{\lambda}_u$ is unstable. Therefore, the original infinite eigenvalues are now mapped to the unstable $\widehat{\lambda}_u$ but the original d-stable finite eigenvalues λ are mapped to the d-stable eigenvalues $\widehat{\lambda}$. However, this does not affect the convergence of the GSM (9), as the unstable component are filtered off by the projection P_r, similar to the infinite components being filtered off for the original PSE.

The parameter γ is not hard to choose. For the real cases, the choice of $\gamma > 0$ seems noncritical according to our numerical experiments except in the neigbourhood of zero, one or infinity. For γ near zero or infinity, $\widehat{\lambda}$ will be near 1, implying the slow convergence of GSM (9). If γ is too near 1, the $\widehat{\lambda}_u$ will be very unstable with large magnitude but $\widehat{\lambda} \approx \lambda$ will be more d-stable.

5.2 PLE→PGSE→PLE

Let γ_1 be the shift for the previous PLE→PGSE step in Theorem 6. Besides $\Re e(\gamma_1) > 0$, the new constraint $\Re e(\gamma_1 e^{i\theta}) \geq 0$ is required. Also a complex shift $\gamma_2 = e^{i\theta} \neq \pm 1$ in Theorem 6 has to be chosen for the PGSE→PLE step.

Applying firstly (i) with $\gamma = \gamma_1$ and $\Re e(\gamma) > 0$ and then (ii) with $\gamma_2 = e^{i\theta} \neq \pm 1$ in Theorem 6, the PLE

$$AXE^H + EXA^H + Q = 0, \quad X = P_r X P_r^H,$$

is equivalent to the PLE:

$$\widehat{A}X\widehat{E}^H + \widehat{E}X\widehat{A}^H + 4\Re e(\gamma)Q = 0, \quad X = P_r X P_r^H,$$

where $\Re e(\gamma) > 0$ preserves whatever definiteness existed in Q and $\widehat{A} \equiv (1+e^{i\theta})A + (\overline{\gamma} - \gamma e^{i\theta})E$ and $\widehat{E} \equiv (1-e^{i\theta})A + (\overline{\gamma} + \gamma e^{i\theta})E$. The eigenvalues λ of $\lambda E - A$ are mapped to $\widehat{\lambda}$ of $\lambda\widehat{E} - \widehat{A}$ and

$$\begin{aligned} \widehat{\lambda} &= \frac{(1+e^{i\theta})\lambda + (\overline{\gamma} - \gamma e^{i\theta})}{(1-e^{i\theta})\lambda + (\overline{\gamma} + \gamma e^{i\theta})}, \quad (\lambda \text{ finite}); \\ \widehat{\lambda}_u &= \frac{1+e^{i\theta}}{1-e^{i\theta}}, \quad (\lambda \text{ infinite}). \end{aligned}$$

The eigenvalues of $\widehat{E}$ or $(1-e^{i\theta})\lambda + (\overline{\gamma} + \gamma e^{i\theta})$ does not vanish as that implies

$$\lambda = \frac{\overline{\gamma} + \gamma e^{i\theta}}{e^{i\theta} - 1} = \frac{2i\Im m\{\gamma(1-e^{i\theta})\}}{|e^{i\theta} - 1|^2},$$

which is purely imaginary and not c-stable. Also note that the original infinite eigenvalues are now mapped to the purely imaginary and unstable $\widehat{\lambda}_u$ but the original c-stable finite eigenvalues λ are mapped to the c-stable finite eigenvalues $\widehat{\lambda}$, and this does not affect the solution process as the unstable component are filtered off P_r.

Again, a feasible choice of γ, which is positive in the real cases, can be found by simple try-and-error. Note that $\lambda\widehat{E} - \widehat{A}$ will return to the original matrix pencil $2(\gamma\lambda E - A)$ or $2(\lambda A - \gamma E)$ when $\gamma_2 = \pm 1$. In practice, the choice of γ seems noncritical but it has to be kept away from zero or infinity, as the latter drives $\widehat{\lambda}$ towards the unstable $(1+e^{i\theta})/(1-e^{i\theta})$ or $(1-e^{i\theta})/(1+e^{i\theta})$.

6 Numerical Examples

In this section, we illustrate the effectiveness of Algorithm 1 (and the double Cayley transform) for large-scale PLEs (and PGSEs). Algorithm 1 was coded in MATLAB 2010b on a 64-bit PC with 3.4 GHz Intel Core i3 processor and 8G RAM. The machine accuracy is reflected by `eps` $= 2.22 \times 10^{-16}$ in MATLAB. With the relative residual $\widetilde{r}_k$ in (12) for the PSE (6) in Example 1 (or after the Cayley transform

from the PLE in Example 2), the stopping criterion is $\widetilde{r}_k \leq \epsilon$ for a small tolerance $\epsilon > 0$. We denote the relative accuracy of X_k by

$$d_k \equiv \frac{\|X_{k+1} - X_k\|}{\|X_{k+1}\|},$$

and the width of B_k at the kth iteration m_k. For the last iteration in both examples, the relative accuracy d_k was not computed as the iteration terminated before X_{k+1} can be formed. We recorded the sub-total CPU time in $t_k = \sum_{i=1}^{k} \delta t_i$ with δt_i being the CPU-time required for the ith iteration.

Example 1. Consider the descriptor system $E\dot{x}(t) = Ax(t) + Bu(t)$ with

$$E = \begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

and the sub-block E_{11} is nonsingular. This system arises from a fully linearized 1-D heat equation

$$\frac{\partial T(x,t)}{\partial t} = \alpha \frac{\partial^2 T(x,t)}{\partial^2 x} + u(x,t),$$

which describes the heat transfer T along x at time t [9].

In the corresponding PGSE in (5), the descriptor matrix E arises from the finite difference discretization with the weighted Crank-Nicolson formula [15]. It contains some zero rows from the zero boundary conditions. The matrix A is sparse and the vector $B \in \mathbb{R}^{n\times 1}$ is random. Let the matrix $A_{21}E_{11}^{-1}E_{12} - A_{22}$ be nonsingular, then the projection P_l and P_r are given by

$$P_l = \begin{bmatrix} I & (A_{12} - A_{11}E_{11}^{-1}E_{12})(A_{21}E_{11}^{-1}E_{12} - A_{22})^{-1} \\ 0 & 0 \end{bmatrix}$$

$$P_r = \begin{bmatrix} I - E_{11}^{-1}E_{12}(A_{21}E_{11}^{-1}E_{12} - A_{22})^{-1}A_{21} & -E_{11}^{-1}E_{12}(A_{21}E_{11}^{-1}E_{12} - A_{22})^{-1}A_{22} \\ (A_{21}E_{11}^{-1}E_{12} - A_{22})^{-1}A_{21} & I + (A_{21}E_{11}^{-1}E_{12} - A_{22})^{-1}A_{22} \end{bmatrix}.$$

For $n = 10000, 50000$ and 100000, we set the truncation tolerance $\tau = 10^{-15}$ and the maximum dimension of the subspace $l_{\max} = 100$. We stopped the iteration after $\widetilde{r}_k \leq \epsilon = 9.0 \times 10^{-15}$. The numerical results are summarized in Tables 1–3.

Table.1 Example 1 ($n = 10000$ with $\gamma = 0.4$, $\tau = 10^{-15}$, $l_{\max} = 100$)

k	d_k	r_k	$\widetilde{r}_k$	l_k	δt_k	t_k
1	1.93e−01	4.05e−02	6.27e−02	2	0.007	0.007
2	3.77e−02	7.90e−03	1.08e−02	4	0.018	0.025
3	1.14e−03	3.04e−04	4.06e−04	8	0.044	0.069
4	2.19e−06	4.58e−07	6.10e−07	16	0.142	0.211
5	5.05e−12	1.05e−12	1.40e−12	30	0.443	0.654
6	—	1.95e−15	2.60e−15	53	0.343	0.997

We can see from tables that the solutions are low-ranked when $\tilde{r}_k$ attained the prescribed accuracy. The relative residuals are decreasing quadratically, indicating the fast convergence.

Table.2 Example 1 ($n = 50000$, $\gamma = 0.45$, $\tau = 10^{-15}$, $l_{\max} = 100$)

k	d_k	r_k	$\widetilde{r}_k$	l_k	δt_k	t_k
1	1.56e−01	1.78e−01	5.30e−02	2	0.037	0.038
2	2.48e−02	2.82e−02	7.60e−03	4	0.100	0.138
3	6.30e−04	7.17e−04	1.89e−04	8	0.266	0.464
4	4.11e−07	4.67e−07	1.23e−07	16	0.803	1.267
5	1.77e−13	1.98e−13	5.23e−14	30	2.961	4.228
6	—	3.13e−14	8.28e−15	53	2.141	6.369

Table.3 Example 1 ($n = 100000$, $\gamma = 0.4$, $\tau = 10^{-15}$, $l_{\max} = 100$)

k	d_k	r_k	$\widetilde{r}_k$	l_k	δt_k	t_k
1	1.39e−01	4.08e−02	6.28e−02	2	0.084	0.086
2	3.78e−02	7.96e−03	1.08e−02	4	0.195	0.281
3	1.46e−03	3.06e−03	4.06e−04	8	0.514	0.795
4	2.20e−06	4.61e−06	6.11e−07	16	1.635	2.430
5	5.06e−12	1.05e−11	1.40e−12	30	6.008	8.438
6	—	2.37e−14	3.15e−15	53	4.462	12.900

Example 2. A multi-body damped mass-spring system with holonomic constraints can be described by a nonlinear differential-algebraic equation of the first order. Linearization around an equilibrium state leads to the descriptor system $E\dot{x} = Ax + Bu$ with

$$E = \begin{bmatrix} I & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & I & 0 \\ -K & -D & -G^T \\ G & 0 & 0 \end{bmatrix},$$

where M is the positive definite mass matrix, K the stiffness matrix, D the damping matrix and G the matrix of constraints. Let G be of full row rank and

$$\begin{aligned} P_l &= \begin{bmatrix} \Pi & 0 & \Pi M^{-1} D G_1 \\ \Pi^T D(I-\Pi) & \Pi^T & \Pi^T(K - D\Pi M^{-1} D) G_1 \\ 0 & 0 & 0 \end{bmatrix}, \\ P_r &= \begin{bmatrix} \Pi & 0 & 0 \\ \Pi M^{-1} D(I-\Pi) & \Pi & 0 \\ -G_1^T(K\Pi - D\Pi M^{-1} D(I-\Pi)) & -G_1^T D\Pi & 0 \end{bmatrix}, \end{aligned}$$

with $G_1 = M^{-1}G^T(GM^{-1}G^T)^{-1}$ and $\Pi = I - M^{-1}G^T(GM^{-1}G^T)^{-1}G$.

This damped mass-spring system has g masses. The ith mass of weight m_i is connected to the $(i+1)$-th mass by a spring and a damper with constants k_i and d_i, respectively, and also to the ground by another spring and damper with constants δ_i and κ_i respectively. The first mass is connected to the last one by a rigid bar, which can be influenced by a control. In the PLE (1) for $g = 500, 5000, 50000$ and

the state space dimensions $n = 1001, 10001, 100001$, $B \in \mathbb{R}^{n\times 1}$ is random. The system parameters are $m_1 = ... = m_g$ and

$$k_1 = ... = k_g - 1 = k = 2, \qquad \kappa_1 = ... = \kappa_g = \kappa = 4,$$
$$d_1 = ... = d_g - 1 = d = 3, \qquad \delta_1 = ... = \delta_g = \delta = 7.$$

We set the truncation tolerance $\tau = 10^{-15}$ for X_k and the maximum dimension of the subspace $l_{\max} = 200$. We stopped the iteration after $\widetilde{r}_k \le \epsilon = 9.0\times 10^{-15}$. The numerical results are summarised in Tables 4–6.

Table 4. Example 2 ($n = 10001$, $\gamma = 0.2$, $\tau = 10^{-15}$, $l_{\max} = 200$)

k	d_k	r_k	$\widetilde{r}_k$	l_k	δt_k	t_k
1	4.45e−01	2.47e+03	1.57e−01	2	0.069	0.071
2	2.24e−01	1.23e+03	6.07e−02	4	0.130	0.201
3	5.52e−02	2.93e+02	1.26e−02	8	0.268	0.469
4	3.24e−03	1.70e+01	7.06e−04	16	0.703	1.172
5	1.11e−05	5.81e−02	2.40e−06	32	2.181	3.353
6	1.34e−10	6.98e−07	2.89e−11	63	7.430	10.783
7	—	1.13e−10	6.36e−15	117	1.182	11.965

Table 5. Example 2 ($n = 50001$, $\gamma = 0.30$, $\tau = 10^{-15}$, $l_{\max} = 200$)

k	d_k	r_k	$\widetilde{r}_k$	l_k	δt_k	t_k
1	6.24e−01	1.29e+04	1.79e−01	2	0.326	0.335
2	1.72e−01	1.39e+03	1.43e−02	4	0.661	0.996
3	6.91e−02	1.95e+03	1.72e−02	8	1.588	2.584
4	5.14e−03	1.44e+02	1.21e−03	16	4.561	7.145
5	2.81e−05	7.95e−01	6.68e−06	32	14.494	21.639
6	8.35e−10	2.25e−05	1.89e−10	64	49.686	71.325
7	—	7.04e−10	5.91e−15	119	7.409	78.734

Table 6. Example 2 ($n = 100001$, $\gamma = 0.35$, $\tau = 10^{-15}$, $l_{\max} = 200$)

k	d_k	r_k	$\widetilde{r}_k$	l_k	δt_k	t_k
1	5.89e−01	1.69e+04	1.23e−01	2	0.706	0.810
2	1.66e−01	3.13e+03	1.61e−02	4	1.417	2.227
3	7.66e−02	1.97e+03	8.88e−03	8	3.489	5.716
4	5.51e−03	1.19e+02	5.04e−04	16	9.965	15.681
5	4.46e−05	3.76e−01	1.58e−06	32	32.929	48.610
6	3.19e−09	1.66e−04	7.00e−10	64	116.690	165.300
7	—	1.10e−09	4.65e−15	120	16.358	181.658

The results show that the relative residuals achieve an accuracy of 10^{-15} in about 12, 79 and 182 seconds with $\mathrm{rank}(X_k)$ equal 117, 119 and 120 for $n = 10001$, 50001 and 100001, respectively.

7 Conclusions

We have presented a generalized Smith method for the projected Lyapunov and Stein equations, of computational complexity $O(n) + O(n_s)$ for large-scale problems. A double Cayley transform is proposed for a PGSE with possible singular E, removing the singularity before the GSM is applied. Numerical experiments show that the proposed algorithm and the double Cayley technique are efficient for computing the solution of the equations.

Acknowledgements

Part of the work occurred when the first author visited Monash University. The first author was supported partly by the NSF of China (11301170), Natural Science Foundation of Hunan Province (14JJ2114, 2017JJ2071) and the Excellent Youth Foundation of Hunan Educational Department (17B071, 17C0466).

R E F E R E N C E S

[1] *P. Benner, M. S. Hossain and T. Stykel,* Low-rank iterative methods for periodic projected Lyapunov equations and their application in model reduction of periodic descriptor systems, *Numer. Alg.*, **67** (2014) 669–690.

[2] *P. Benner, J. R. Li and T. Penzl,* Numerical solution of large Lyapunov equations, Riccati equations, and linear-quadratic control problems, *Numer. Linear Algebra Appl.*, **15** (2008) 755–777.

[3] *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.

[4] *E. K.-W. Chu, H.-Y. Fan and W.-W. Lin,* Projected generalized discrete-time periodic Lyapunov equations and balanced realization of periodic descriptor systems, *SIAM J. Matrix Anal. Appl.*, **29** (3) (2007) 982–1006.

[5] *E. K.-W. Chu, H.-Y. Fan, W.-W. Lin and C.-S. Wang,* A structure-preserving doubling algorithm for periodic discrete-time algebraic Riccati equations, *Int. J. Control*, **77** (2004) 767–788.

[6] *E. K.-W. Chu and P. C.-Y. Weng.* Large-scale discrete-time algebraic Riccati equations — doubling algorithm and error analysis, *J. Comp. Appl. Maths.* **277** (2015) 115–126.

[7] *N. Dong, B. Yu,* On the tripling algorithm for large-scale nonlinear matrix equations with low rank structure, *J. Comput. Appl. Math.*, **288** (2015), 18–32.

[8] *G. H. Golub and C. F. Van Loan*, Matrix Computations, (3rd Edition), Johns Hopkins University Press, Baltimore, MD, 1996.

[9] *Y. Jaluria and K. Torrance*, Computational Heat Transfer, 2nd Edn., Taylor & Francis, New York, 2003.

[10] *K. Jbilou,* ADI preconditioned Krylov methods for large Lyapunov matrix equations, *Linear Algebra Appl.*, **432** (2010) 2473–2485.

[11] *J.-R. Li and J. White,* Low rank solution of Lyapunov equations, *SIAM J. Matrix Anal. Appl.*, **24** (1) (2002) 260–280.

[12] *T.-X. Li , P. C.-Y. Weng, E. K.-W. Chu and W.-W. Lin*, Large-scale Stein and Lyapunov equations, Smith method, and applications, *Numer. Algor.*, **63** (2013) 727-752.

[13] *T.-X. Li, E. K.-W. Chu, W.-W. lin and P. C.-Y. Weng,* Solving large-scale continuous-time algebraic Riccati equations by doubling, *J. Comp. Appl. Math.*, **237** (2013) 373–383.

[14] *T.-X. Li, P. C.-Y. Weng, E. K.-W. Chu and W.-W. Lin,* Large-scale Stein and Lyapunov equations, Smith method, and applications, *Numer. Alg.* **63** (2013) 727–752.

[15] *P. Mccorquodale, P. Colella and J. Johansen,* A Cartesian grid embedded boundary method for the heat equation on irregular domains, *J. Comp. Physics*, **173** (2001) 620–635.

[16] *V. Mehrmann and T. Stykel,* Balanced truncation model reduction for large-scale systems in descriptor form, in: Lecture Notes in Computational Science and Engineering, Vol. **45**, Springer-Verlag, Berlin, Heidelberg, 2005, 83-15.

[17] *T. Penzl,* Numerical solution of generalized Lyapunov equations, *Advances in Comp. Maths.*, **8** (1998) 33–48.

[18] *T. Penzl,* A cyclic low-rank Smith method for large sparse Lyapunov equations, *SIAM J. Sci. Comput.*, **21** (1999/2000) 1401–1418.

[19] *J. Sabino,* Solution of large-scale Lyapunov equations via the block modified Smith method, PhD thesis, Rice University, Houston, 2006.

[20] *V. Simoncini,* A new iterative method for solving large-scale Lyapunov matrix equations, *SIAM J. Sci. Comput.*, **29** (2007) 1268–1288.

[21] *R. A. Smith,* Matrix equation $XA + BX = C$, *SIAM J. Appl. Math.* **16** (1968) 198–201.

[22] *J. Sreedhar and P. Van Dooren,* Periodic descriptor systems: solvability and conditionability, *IEEE Trans. Automat. Control*, **44** (1999) 310–313.

[23] *T. Stykel,* Stability and inertia theorems for generalized Lyapunov equations, *Linear Algebra Appl.*, **355** (2002) 297–314.

[24] *T. Stykel,* Low-rank iterative methods for projected generalized Lyapunov equations, *Electron. Trans. Numer. Anal.*, **30** (2008) 187–202.

[25] *T. Stykel and V. Simoncini,* Krylov subspace methods for projected Lyapunov equations, *Appl. Numer. Maths.*, **62** (2012) 35–50.

[26] *B. Yu, D.-H. Li and N. Dong,* Low memory and low complexity iterative schemes for a nonsymmetric algebraic Riccati equation arising from transport theory, *J. Comput. Appl. Math.*, **250** (2013), 175–189.

[27] *L. Zhang, J. Lam and Q. Zhang,* Lyapunov and Riccati equations for discrete-time descriptor systems, *IEEE Trans. Automat. Control*, **44** (1999) 2134–2139.