A generalized structured doubling algorithm for optimal control problems

Volker Mehrmann* Federico Poloni †

December 25, 2010

Abstract

We propose a generalization of the Structured Doubling Algorithm (SDA) to compute invariant subspaces of structured matrix pencils that arise in the context of solving linear quadratic optimal control problems. The new algorithm is designed to attain better accuracy when the classical Riccati equation approach for the solution of the optimal control problem is not well suited because the stable and unstable invariant subspaces are not well separated (due to eigenvalues near or on the imaginary axis) or in the case when the Riccati solution does not exist at all. We analyze the convergence of the method and compare the new method with the classical SDA algorithm as well as some recent structured QR-methods.

Keywords: structured doubling algorithm, optimal control, even pencil, BVD pencil, symplectic pencil, Cayley transformation, disk function method, pencil arithmetic

AMS(MOS) subject classification: 65F15, 65F30, 49J15, 49N10

1 Introduction

The main motivation for the new algorithmic approach that we discuss in this paper is the solution of continuous and discrete time optimal control problems. In order to set the notation, we briefly review these problems.

1.1 Continuous-time optimal control

Consider the linear quadratic optimal control problem to minimize the quadratic cost functional

$$\mathcal{S}(x,u) = \int_{t_0}^{\infty} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt, \tag{1}$$

subject to the linear control system

$$E\dot{x} = Ax + Bu, \qquad x(t_0) = x^0. \tag{2}$$

Here $x(t)$ is the *state*, x^0 is an *initial vector*, $u(t)$ is the *control input* and the coefficient matrices satisfy $E, A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,m}$, $Q = Q^T \in \mathbb{R}^{n,n}$, $R = R^T \in \mathbb{R}^{m,m}$. We assume as usual that $\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix}$ is positive semidefinite. The basic theory for this problem can be found in any book on linear control theory, e.g., [2, 35, 38]. If E is nonsingular, then the problem may easily be reduced to one with $E = I_n$ by inverting E in (2). Therefore, in the following the cases of interest are

*Inst. f. Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany, email: mehrmann@math.tu-berlin.de

†Scuola Normale Superiore; Piazza dei Cavalieri, 7; 56126 Pisa, Italy, email: f.poloni@sns.it

$E = I_n$ and E singular or ill-conditioned; in the latter case, (2) is known as a *descriptor system* or a linear *differential-algebraic equation* (DAE).

Application of the Pontryagin maximum principle [36, 33] leads to the necessary optimality condition given by the two-point boundary value problem

$$\mathcal{E}_c \begin{bmatrix} \dot{\mu} \\ \dot{x} \\ \dot{u} \end{bmatrix} = \mathcal{A}_c \begin{bmatrix} \mu \\ x \\ u \end{bmatrix}, \qquad x(t_0) = x^0, \qquad \lim_{t\to\infty} E^T\mu(t) = 0, \tag{3}$$

with the *matrix pencil*

$$\mathcal{L}(\lambda) := \lambda\mathcal{E}_c - \mathcal{A}_c := \lambda \begin{bmatrix} 0 & E & 0 \\ -E^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & A & B \\ A^T & Q & S \\ B^T & S^T & R \end{bmatrix}. \tag{4}$$

Since $\mathcal{L}(\lambda) = \mathcal{L}(-\lambda)^T$, pencils of this form are called *even pencils*, [30] or para-Hermitian pencils [43].

If R is well-conditioned with respect to inversion, then (3) may be transformed into the problem

$$\hat{\mathcal{E}}_c \begin{bmatrix} \dot{x} \\ \dot{\mu} \end{bmatrix} = \hat{\mathcal{A}}_c \begin{bmatrix} x \\ \mu \end{bmatrix}, \qquad x(t_0) = x^0, \qquad \lim_{t\to\infty} E^T\mu(t) = 0, \tag{5}$$

with associated pencil

$$\begin{aligned} \lambda\hat{\mathcal{E}}_c - \hat{\mathcal{A}}_c =& \lambda \begin{bmatrix} E & 0 \\ 0 & E^T \end{bmatrix} - \begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ -Q + SR^{-1}S^T & -(A - BR^{-1}S^T)^T \end{bmatrix} \\ =:& \lambda \begin{bmatrix} E & 0 \\ 0 & E^T \end{bmatrix} - \begin{bmatrix} F & -G \\ -H & -F^T \end{bmatrix}. \end{aligned} \tag{6}$$

When $E = I$, this pencil reduces to $\lambda I - \mathcal{H}$ with the *Hamiltonian matrix*

$$\mathcal{H} = \begin{bmatrix} F & -G \\ -H & -F^T \end{bmatrix}. \tag{7}$$

The solution of (5) can be obtained in many different ways. For instance, the classical approach that is described in most textbooks and implemented in most design packages (for $E = I$) is to determine first (if it exists) X, the unique positive semidefinite solution of the associated *algebraic Riccati equation*

$$0 = H + XF + F^T X - XGX, \tag{8}$$

and then to obtain the optimal control as a *linear feedback*

$$u(t) = -R^{-1}(B^T X + S^T)x(t). \tag{9}$$

The associated state vector of the closed loop system, i.e., the solution of

$$\dot{x} = (A - BR^{-1}(B^T X + S^T))x(t),\ x(t_0) = x_0 \tag{10}$$

is then *asymptotically stable*, i.e., $\lim_{t\to\infty} x(t) = 0$. For this reason the positive semidefinite solution of (8) is called *c-stabilizing solution*.

It should be noted, however, that the solution of the boundary value problems (3), (5), and thus also the optimal feedback control (9), may exist and be unique even when E is singular or a positive semidefinite solution to the Riccati equation does not exist. For an analysis of the different possibilities see [21, 28, 33]. If E is invertible, then under the presented assumptions on the cost functional, a sufficient condition for the existence and uniqueness of the stabilizing Riccati solution is that the pair (A, B) is *c-stabilizable*, i.e., Rank $([\lambda I - A, B]) = n$ for all $\lambda \in \mathbb{C}$ with $\Re\lambda \geq 0$. If this is the case, then the pencils (7) and (4) have exactly n eigenvalues with negative real part

and n with positive real part. Moreover, if $E = I$, then the n-dimensional invariant subspace $\mathcal{U}$ associated with the eigenvalues with negative real part of (7) spanned by the columns of

$$U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \in \mathbb{R}^{2n,n} \tag{11}$$

is *Lagrangian*, i.e., it satisfies $U^T \mathcal{J} U = 0$, where

$$\mathcal{J} = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix},$$

and I_n is the $n \times n$ identity matrix. If U_1 is invertible, then the c-stabilizing solution of (8) is given by $X = U_2 U_1^{-1}$.

In the more general case that E or R are not invertible, the deflating subspace of (4) associated with the eigenvalues in the closed left half plane has typically a dimension smaller than n and is clearly not Lagrangian. To see this we can use a different representation of the system, see [33, Chapter 9]. Perform a singular value decomposition

$$U^T E V = \begin{bmatrix} \Sigma_q & 0 \\ 0 & 0 \end{bmatrix}$$

with $\Sigma_q \in \mathbb{R}^{q,q}$ invertible and diagonal and partition the vectors and matrices

$$\begin{aligned} U^T A V &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},\ U^T x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ U^T x^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \end{bmatrix}, \\ U^T B &= \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},\ V^T S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix},\ V^T \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},\ V^T Q V = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \end{aligned}$$

accordingly. Performing a congruence transformation with $T = \operatorname{diag}(U^T, V^T, I_r)$ from the left and with T^T from the right to (4) and reordering the blocks, we obtain the system

$$\left[\begin{array}{c|c|ccc} 0 & \Sigma_q & 0 & 0 & 0 \\ \hline -\Sigma_q^T & 0 & 0 & 0 & \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} \dot{\mu}_1 \\ \hline \dot{x}_1 \\ \hline \dot{\mu}_2 \\ \dot{x}_2 \\ \dot{u} \end{array}\right] = \left[\begin{array}{c|c|ccc} 0 & A_{11} & 0 & A_{12} & B_1 \\ \hline A_{11}^T & Q_{11} & A_{21}^T & Q_{12} & S_1 \\ \hline 0 & A_{21} & 0 & A_{22} & B_2 \\ A_{12}^T & Q_{12}^T & A_{22}^T & Q_{22} & S_2 \\ B_1^T & S_1^T & B_2^T & S_2^T & R \end{array}\right] \left[\begin{array}{c} \mu_1 \\ \hline x_1 \\ \hline \mu_2 \\ x_2 \\ u \end{array}\right],$$

with boundary conditions $x_1(t_0) = x_1^0$, $x_2(t_0) = x_2^0$, $A_{21}x_1^0 + A_{22}x_2^0 + B_2 u(t_0) = 0$, $\lim_{t\to\infty} \mu_1(t) = 0$. Introducing

$$\begin{aligned} \tilde{E} &= \Sigma_q,\ \tilde{A} = A_{11},\ \tilde{Q} = Q_{11},\ \tilde{B} = \begin{bmatrix} 0 & A_{12} & B_1 \end{bmatrix},\ \tilde{x} = x_1,\ \tilde{\mu} = \mu_1, \\ \tilde{S} &= \begin{bmatrix} A_{21}^T & Q_{12} & S_1 \end{bmatrix}, \tilde{R} = \begin{bmatrix} 0 & A_{22} & B_2 \\ A_{22}^T & Q_{22} & S_2 \\ B_2^T & S_2^T & R \end{bmatrix},\ \tilde{u} = \begin{bmatrix} \mu_2 \\ x_2 \\ u \end{bmatrix}, \end{aligned} \tag{12}$$

we obtain a new boundary value problem

$$\tilde{\mathcal{E}}_c \begin{bmatrix} \dot{\tilde{\mu}} \\ \dot{\tilde{x}} \\ \dot{\tilde{u}} \end{bmatrix} = \tilde{\mathcal{A}}_c \begin{bmatrix} \tilde{\mu} \\ \tilde{x} \\ \tilde{u} \end{bmatrix}, \qquad \tilde{x}(t_0) = x_1^0, \qquad \lim_{t\to\infty} \tilde{\mu}(t) = 0, \tag{13}$$

with consistency conditions $A_{21}x_1^0 + A_{22}x_2^0 + B_2 u(t_0) = 0$, $x_2(t_0) = x_2^0$, and a transformed *matrix pencil*

$$\lambda \tilde{\mathcal{E}}_c - \tilde{\mathcal{A}}_c := \lambda \begin{bmatrix} 0 & \tilde{E} & 0 \\ -\tilde{E}^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \tilde{A} & \tilde{B} \\ \tilde{A}^T & \tilde{Q} & \tilde{S} \\ \tilde{B}^T & \tilde{S}^T & \tilde{R} \end{bmatrix}. \tag{14}$$

So except for the extra consistency conditions (which are not relevant for the eigenvalue problem), the system has the same form as the one with a nonsingular E.

If E and R are nonsingular then the even pencil (4) has exactly $2n$ finite eigenvalues and we can use the methods discussed below. If this not the case, then the problem corresponds to a singular control problem and a regularization procedure is necessary, which either modifies the cost functional [27] or removes modes that are not impulse controllable from the system, see also [12]. In the following we therefore assume that in (4) E and R are nonsingular.

Currently, the preferred approach to solve the two-point boundary value problem (5) numerically, is to compute the deflating subspace associated with the eigenvalues in the left half plane of (4) via a structured generalized Schur form [6, 10].

In the standard case (5) this would be the real Hamiltonian Schur-form of $\mathcal{H}$, [33], i.e., one determines an *orthogonal and symplectic* matrix $\mathcal{U} \in \mathbb{R}^{2n,2n}$, i.e., $\mathcal{U}^T\mathcal{U} = I_{2n}$ and $\mathcal{U}^T\mathcal{J}\mathcal{U} = \mathcal{J}$, such that

$$\mathcal{U}^T\mathcal{H}\mathcal{U} = \begin{bmatrix} T_{11} & T_{12} \\ 0 & -T_{11}^T \end{bmatrix}, \tag{15}$$

where T_{11} has only eigenvalues with negative real part. Then $\mathcal{U}$ has the form

$$\mathcal{U} = \begin{bmatrix} U_1 & -U_2 \\ U_2 & U_1 \end{bmatrix}, \tag{16}$$

and the first n columns are given by (11) and span the *c-stabilizing invariant subspace* of (7) associated with the n eigenvalues in the left half plane.

The case in which $\mathcal{H}$ has eigenvalues on the imaginary axis arises as well in the applications, e.g., in H_∞ control [7, 18, 20, 44]. In the standard case (5), then one computes the unique *c-semi-stable Lagrangian subspace U* (if it exists), see [21] for existence, uniqueness, and a complete parametrization of all solutions. Similarly, in the general case one computes the unique *c-semi-stable subspace* (if it exists) and uses it to solve the boundary value problem, [6].

1.2 Discrete-time optimal control

There is an almost completely analogous theory for the corresponding discrete-time linear quadratic optimal control problem, see e.g. [19, 35]. In this case one minimizes

$$\sum_{j=0}^{\infty} \left(\begin{bmatrix} x_j \\ u_j \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} x_j \\ u_j \end{bmatrix} \right) \tag{17}$$

subject to the discrete-time linear control system

$$Ex_{j+1} = Ax_j + Bx_j,\ x_0 = x^0. \tag{18}$$

Introducing a sequence of Lagrange multipliers $\mu_j \in \mathbb{R}^n$, $j = 0, 1, 2, \ldots$ and applying again the Pontryagin maximum principle yields a necessary optimality condition given by the discrete-time two-point boundary value problem

$$\mathcal{E}_d \begin{bmatrix} \mu_{j+1} \\ x_{j+1} \\ u_{j+1} \end{bmatrix} = \mathcal{A}_d \begin{bmatrix} \mu_j \\ x_j \\ u_j \end{bmatrix}, \qquad x_0 = x^0, \qquad \lim_{j\to\infty} E^T\mu_j = 0, \tag{19}$$

with the *matrix pencil*

$$\lambda\mathcal{E}_d - \mathcal{A}_d := \begin{bmatrix} 0 & E & 0 \\ A^T & 0 & 0 \\ B^T & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & A & B \\ E^T & Q & S \\ 0 & S^T & R \end{bmatrix}. \tag{20}$$

Pencils with this structure are called *BVD-pencils* in [13].

Again, if R is well-conditioned with respect to inversion, then one can eliminate the variable u_j in the boundary value problem (19), and transform the problem to the boundary value problem

$$\hat{\mathcal{E}}_d \begin{bmatrix} x_{j+1} \\ \mu_{j+1} \end{bmatrix} = \hat{\mathcal{A}}_d \begin{bmatrix} x_j \\ \mu_j \end{bmatrix}, \qquad x_0 = x^0, \qquad \lim_{j\to\infty} E^T \mu_j = 0, \tag{21}$$

with the *reduced BVD-pencil*

$$\begin{aligned} \lambda\hat{\mathcal{E}}_d - \hat{\mathcal{A}}_d &:= \lambda \begin{bmatrix} E^T & -G \\ 0 & F^T \end{bmatrix} - \begin{bmatrix} F & 0 \\ H & E \end{bmatrix} \\ &:= \lambda \begin{bmatrix} E & -BR^{-1}B^T \\ 0 & (A - BR^{-1}S^T)^T \end{bmatrix} - \begin{bmatrix} A - BR^{-1}S^T & 0 \\ Q - SR^{-1}S^T & E^T \end{bmatrix}. \end{aligned} \tag{22}$$

When $E = I$, i.e.,

$$\lambda\hat{\mathcal{E}}_d - \hat{\mathcal{A}}_d := \lambda \begin{bmatrix} I & -G \\ 0 & F^T \end{bmatrix} - \begin{bmatrix} F & 0 \\ H & I \end{bmatrix}, \tag{23}$$

then this pencil is a *symplectic pencil* [19, 33, 34], i.e., it satisfies

$$\hat{\mathcal{E}}_d \mathcal{J} \hat{\mathcal{E}}_d^T = \hat{\mathcal{A}}_d \mathcal{J} \hat{\mathcal{A}}_d. \tag{24}$$

If, furthermore, E and $F = A - BR^{-1}S^T$ are invertible, then one may transform even further and instead solve the boundary value problem associated with the eigenvalue problem for the *symplectic matrix* $\mathcal{S} = \hat{\mathcal{E}}_d^{-1}\hat{\mathcal{A}}_d$.

As in the continuous-time case, this boundary value problem is usually solved via the computation of a deflating subspace of the associated structured pencils (20) or (22), respectively. However, in this case the desired *d-stable deflating subspace* is associated with the eigenvalues inside the unit disk.

The associated (discrete-time) algebraic Riccati equation is given by

$$\begin{aligned} 0 =& Q - SR^{-1}S^T - EX - (A - BR^{-1}S^T)^T X (E^T + BR^{-1}B^T X)^{-1}(A - BR^{-1}S^T) \\ =& H - EX - F^T X (E^T + GX)^{-1} F, \end{aligned}$$

and if the d-stabilizing deflating subspace of (22) (or the d-stabilizing invariant subspace of the corresponding symplectic matrix) associated with the eigenvalues inside the unit disc is n-dimensional and spanned by the columns of a matrix as in (11), then the unique *d-stabilizing positive semidefinite* Riccati solution is given by $X = -U_2 U_1^{-1}$.

For discrete-time problems, analogous results as in the continuous-time case hold when E is singular. It has been shown in [33, Chapter 9] that the resulting discrete system again has the form (20) just replacing every matrix by the matrices in (12) plus some extra consistency conditions which are not relevant for the eigenvalue problem. So also in this case we may assume that E is invertible. The case that R is singular again requires a regularization or the removal of the parts of the system that are not impulse-controllable.

In most cases we expect n eigenvalues in the open unit disk but the case that there are eigenvalues on the unit circle is also of practical relevance in discrete-time robust control, [44].

1.3 Solution methods

In the following we assume that the Hamiltonian matrix has exactly n eigenvalues in the open complex left half plane (therefore due to symmetry of the spectrum of Hamiltonian matrices also n eigenvalues in the open right half plane) and that the c-stabilizing solution to (8) exists and is unique. However, if eigenvalues are close to the imaginary axis, then even if the solution exists and is unique, the computation of the Riccati solution may be highly ill-conditioned.

Similarly, in the discrete-time case we assume that there are exactly n eigenvalues inside the unit circle, and again by symmetry it follows that there are also n eigenvalues outside the unit circle.

The currently best methods to compute the desired stabilizing deflating subspaces for full dense matrices of small and medium size are the structure preserving methods of [6, 15, 32]. They are numerically backward stable and work well, whenever the Hamiltonian Schur form and hence the unique Lagrangian subspace associated with the stable eigenvalues exists and is unique. When there are eigenvalues close to or on the imaginary axis, then these methods perform much better than all other currently available methods. There are analogous methods for the discrete problem that use a palindromic formulation of the symplectic pencil, [37].

An alternative method that also shows good performance for well-conditioned small- to medium-size problems is the structured doubling algorithm (SDA) [1, 26, 17, 16, 29, 25]. This algorithm was originally designed as an iterative algorithm to compute the solution of algebraic Riccati equations. It is quadratically convergent and it is based purely on matrix products and inversions, and therefore well-suited to parallel and multi-threaded implementations. Like the Newton method and the sign function method for the Riccati equation [4, 5, 11, 33], and it has the potential to be used in the context of low rank approximate solutions. Furthermore, it has been shown recently in [25] that SDA converges also in the case where a unique stabilizing solution exists but when there are eigenvalues of $\mathcal{H}$ on the imaginary axis. In this case, however, the convergence rate becomes linear instead of quadratic.

In the case in which the matrix U_1 in (16) or (11) is ill-conditioned with respect to inversion, then the corresponding Riccati solution X becomes very large; the quantities appearing in the SDA grow as well, and ill-conditioning in the intermediate steps often leads to poor performance of the algorithm. Unlike the Newton method, which exhibits a 'self-correcting' behavior in many contexts, the SDA cannot recover from an ill-conditioned intermediate step.

In this paper, we address this issue, and propose a generalization of the structured doubling algorithm that aims to achieve more accuracy in these problematic cases, at the expense of a larger cost per step. Numerical examples are provided that show that the new variant of the structured doubling algorithm achieves better results on several border-line cases.

The paper is structured as follows. In Section 2 we propose a new block-swapping procedure that is based on the pencil arithmetic of [5]. In Section 3 we present the structured doubling algorithm as a block-swapping technique and in Section 4 we show how this approach can be generalized to avoid certain instabilities of the standard structured doubling algorithm. Moreover, we analyze the convergence and structure preservation properties of the new algorithm. The proof of a technical result on the generalized SDA convergence is deferred to Appendix A. In Section 5, we report the results of several numerical experiments that demonstrate the improvements that we have made.

2 Block-swapping in pencil arithmetic

Pencil arithmetic [5] is a technique to extend several common matrix operations, such as matrix sum and matrix product to matrix pencils. One of its many uses is to perform linear algebra operations on matrices of the form $\mathcal{E}^{-1}\mathcal{A}$, while storing and operating on the pair $(\mathcal{E}, \mathcal{A})$. This allows to extend such operations in a numerically stable way even to the case that $\mathcal{E}$ is singular or ill-conditioned.

The basic procedure from pencil arithmetic that we exploit is the following *block-swapping procedure*, see also [39].

Given two matrices $\mathcal{E}, \mathcal{A} \in \mathbb{R}^{\ell,\ell}$, with the property that

$$\begin{bmatrix} \mathcal{A} \\ -\mathcal{E} \end{bmatrix}$$

has full column rank, determine $\widetilde{\mathcal{E}}, \widetilde{\mathcal{A}} \in \mathbb{R}^{\ell,\ell}$ such that

$$\widetilde{\mathcal{A}}\mathcal{E} = \widetilde{\mathcal{E}}\mathcal{A} \text{ or } \begin{bmatrix} \widetilde{\mathcal{E}} & \widetilde{\mathcal{A}} \end{bmatrix} \begin{bmatrix} \mathcal{A} \\ -\mathcal{E} \end{bmatrix} = 0, \tag{25}$$

and Rank $\begin{bmatrix} \widetilde{\mathcal{E}} & \widetilde{\mathcal{A}} \end{bmatrix} = \ell$.

Clearly there is no unique solution to this problem. If the pair $(\widetilde{\mathcal{E}}, \widetilde{\mathcal{A}})$ is a solution, then so is $(S\widetilde{\mathcal{E}}, S\widetilde{\mathcal{A}})$ for any nonsingular matrix $S \in \mathbb{R}^{k,k}$.

A simple approach to make the block-swapping unique, which however fails if $\mathcal{E}$ is singular or ill-conditioned, is to use $(\widetilde{\mathcal{E}}, \widetilde{\mathcal{A}}) = (S\mathcal{A}\mathcal{E}^{-1}, S)$ for a given fixed nonsingular matrix S. A more robust, but also computationally more expensive choice is to compute a QR-decomposition

$$\begin{bmatrix} \mathcal{A} \\ -\mathcal{E} \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix}$$

and to take $(\widetilde{\mathcal{E}}, \widetilde{\mathcal{A}}) = (Q_2^T, Q_4^T)$, see [5].

We have the following sufficient condition for the solvability of the block swapping problem.

Proposition 1. *Let $\mathcal{A}, \mathcal{E} \in \mathbb{R}^{\ell,\ell}$ be given, let $M \in \mathbb{R}^{2\ell,2\ell}$ be a nonsingular matrix, and let*

$$M \begin{bmatrix} \mathcal{A} \\ -\mathcal{E} \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix}.$$

If X is nonsingular, then for any nonsingular $S \in \mathbb{R}^{\ell,\ell}$ the pair $(\tilde{\mathcal{E}}, \tilde{\mathcal{A}})$ defined by

$$\begin{bmatrix} \tilde{\mathcal{E}} & \tilde{\mathcal{A}} \end{bmatrix} = \begin{bmatrix} -SYX^{-1} & S \end{bmatrix} M^{-1}$$

is a solution to the block-swapping problem, i.e., it satisfies (25).

Proof. The result follows by direct computation. □

Note that the simple approach corresponds to $S = I$ and

$$M = \begin{bmatrix} 0 & I_\ell \\ I_\ell & 0 \end{bmatrix},$$

while the more robust approach corresponds to $S = I$ and $M = Q^T$, with Q being the orthogonal factor of the QR-decomposition.

We propose another approach, which allows to preserve special block structures. Partition the matrices appearing in a block-swapping problem as

$$\mathcal{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad \mathcal{E} = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}, \quad \widetilde{\mathcal{A}} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix}, \quad \widetilde{\mathcal{E}} = \begin{bmatrix} \widetilde{E}_{11} & \widetilde{E}_{12} \\ \widetilde{E}_{21} & \widetilde{E}_{22} \end{bmatrix}, \tag{26}$$

where $A_{11}, E_{11}, \widetilde{A}_{11}, \widetilde{E}_{11} \in \mathbb{R}^{n,n}$ and the other blocks are of according sizes.

Proposition 2. *Let $\mathcal{A}$, $\mathcal{E}$ and the blocks $\widetilde{E}_{11}$, $\widetilde{A}_{12}$, $\widetilde{E}_{21}$, $\widetilde{A}_{22}$ be given. If the matrix*

$$\begin{bmatrix} -E_{11} & -E_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{27}$$

is nonsingular, then $\widetilde{\mathcal{A}}$, $\widetilde{\mathcal{E}}$ can be completed to satisfy (25) by setting

$$\begin{bmatrix} \widetilde{A}_{11} & \widetilde{E}_{12} \\ \widetilde{A}_{21} & \widetilde{E}_{22} \end{bmatrix} = \begin{bmatrix} \widetilde{E}_{11} & \widetilde{A}_{12} \\ \widetilde{E}_{21} & \widetilde{A}_{22} \end{bmatrix} \begin{bmatrix} -A_{11} & -A_{12} \\ E_{21} & E_{22} \end{bmatrix} \begin{bmatrix} -E_{11} & -E_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1}.$$

Proof. To obtain the desired solution, we can either set

$$M = \begin{bmatrix} 0 & 0 & I_n & 0 \\ 0 & I_{\ell-n} & 0 & 0 \\ I_n & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{\ell-n} \end{bmatrix}$$

in Proposition 1, or rewrite the block-swapping problem as

$$\left[\begin{array}{cc|cc} \widetilde{A}_{11} & \widetilde{E}_{12} & \widetilde{E}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{E}_{22} & \widetilde{E}_{21} & \widetilde{A}_{22} \end{array}\right] \left[\begin{array}{cc} -E_{11} & -E_{12} \\ A_{21} & A_{22} \\ \hline A_{11} & A_{12} \\ -E_{21} & -E_{22} \end{array}\right] = 0,$$

i.e.,

$$\begin{bmatrix} \widetilde{A}_{11} & \widetilde{E}_{12} \\ \widetilde{A}_{21} & \widetilde{E}_{22} \end{bmatrix} \begin{bmatrix} -E_{11} & -E_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} \widetilde{E}_{11} & \widetilde{A}_{12} \\ \widetilde{E}_{21} & \widetilde{A}_{22} \end{bmatrix} \begin{bmatrix} -A_{11} & -A_{12} \\ E_{21} & E_{22} \end{bmatrix}.$$

□

The block-swapping solution described in Proposition 2 has the same computational cost as the simple solution $(\widetilde{\mathcal{E}}, \widetilde{\mathcal{A}}) = (S\mathcal{A}\mathcal{E}^{-1}, S)$, but it requires the inversion of a different matrix. Thus, when $\mathcal{E}$ is singular or ill-conditioned, while (27) is not, then it is preferable to use this solution instead.

3 Structured Doubling Algorithms

In this section we review some previous work on doubling algorithms.

3.1 The inverse-free disc function method

The inverse-free disc function method is a matrix iteration that computes two special deflating subspaces of a matrix pencil, the *d-stable* deflating subspace relative to all the eigenvalues λ with $|\lambda| < 1$ and the *d-unstable* deflating subspace relative to all the eigenvalues λ with $|\lambda| > 1$. It is an algorithm of the family of *doubling algorithms*, first introduced and analyzed in [1] for discrete-time algebraic Riccati equations.

To generalize the method, we make use of the following 'squaring' result for matrix pencils.

Theorem 3 ([3, 4]). *Let $\lambda\mathcal{E} - \mathcal{A}$ be a regular matrix pencil with $\mathcal{A}, \mathcal{E} \in \mathbb{R}^{\ell,\ell}$, and let $(\widetilde{\mathcal{E}}, \widetilde{\mathcal{A}})$ be a solution of (25). Then the pencil*

$$\lambda\widetilde{\mathcal{E}}\mathcal{E} - \widetilde{\mathcal{A}}\mathcal{A} \tag{28}$$

is regular and has the same right deflating subspaces as $\lambda\mathcal{A} + \mathcal{E}$. Furthermore, the eigenvalues of (28) are the squares of the eigenvalues of the pencil.

This result is not surprising if $\mathcal{E}$ is invertible, since in this case the eigenvalues and right deflating subspaces of $\lambda\mathcal{E} - \mathcal{A}$ correspond to the eigenvalues and right invariant subspaces of $\mathcal{E}^{-1}\mathcal{A}$. Moreover, it is simple to check that (25) implies that

$$(\widetilde{\mathcal{E}}\mathcal{E})^{-1}(\widetilde{\mathcal{A}}\mathcal{A}) = (\mathcal{E}^{-1}\mathcal{A})^2.$$

Thus, the map

$$\mathcal{S}: \ (\mathcal{A}, \mathcal{E}) \mapsto (\widetilde{\mathcal{A}}\mathcal{A}, \widetilde{\mathcal{E}}\mathcal{E}) \tag{29}$$

is a way to extend the concept of squaring to matrix pencils.

If this squaring operation is repeated k times, then one obtains a pencil $\mathcal{S}^k(\mathcal{E}, \mathcal{A})$ with the same deflating subspaces as $\lambda\mathcal{E} - \mathcal{A}$, but every eigenvalue λ_0 of the original pencil $\lambda\mathcal{E} - s\mathcal{A}$ is mapped to $\lambda_0^{2^k}$.

In particular, eigenvalues with $|\lambda| < 1$ rapidly converge to 0, and eigenvalues with $|\lambda| > 1$ rapidly converge to ∞. Making use of this rapid convergence and some scaling to avoid overflow, one can determine deflating subspaces of the original pencil associated with the stable (unstable) eigenvalues by computing the deflating subspaces associated with 0 and ∞ of the pencil associated

Algorithm 1: Inverse-free disc function method

input : A matrix pencil $\mathcal{A}_0 - s\mathcal{E}_0$ without unimodular eigenvalues
output: Matrices U, V whose columns form bases of the d-stable and d-unstable deflating subspace of $\lambda\mathcal{E}_0 - \mathcal{A}_0$, respectively.

$k \longleftarrow 0$;
while *a suitable stopping criterion is not satisfied* **do**
  $(\widetilde{\mathcal{E}_k}, \widetilde{\mathcal{A}_k}) \longleftarrow$ any solution of (25) for given $(\mathcal{A}_k, \mathcal{E}_k)$;
  $\mathcal{A}_{k+1} \longleftarrow \widetilde{\mathcal{A}_k}\mathcal{A}_k$;
  $\mathcal{E}_{k+1} \longleftarrow \widetilde{\mathcal{E}_k}\mathcal{E}_k$;
  $k \longleftarrow k + 1$;
end
$U \longleftarrow$ approximate nullspace of $\mathcal{A}_k$;
$V \longleftarrow$ approximate nullspace of $\mathcal{E}_k$.

with $\mathcal{S}^\ell(\mathcal{A}, \mathcal{E})$. This idea is the basis of the inverse-free disc function method of [4] presented in Algorithm 1.

The numerical stability of Algorithm 1 and the speed of convergence for the sequences $\mathcal{A}_k$ and $\mathcal{E}_k$ depend strongly on the choice of the solution of (25), the distance of the eigenvalues to the unit circle and the angle between the d-stable and the d-unstable subspace.

3.2 The Cayley transform

To apply a the disk function method or any other doubling algorithm in the case of continuous time optimal control problems, typically the matrix pencil is first transformed so that the eigenvalues in the left half problem are moved into the unit disk. This can e.g. be done by the *Cayley transform*

$$\mathcal{C}_\gamma : \mathcal{H} \mapsto (\mathcal{H} - \gamma I)^{-1}(\mathcal{H} + \gamma I),$$

with a given parameter $\gamma > 0$.

The following properties are well known, see e.g. [30, 31].

1. The invariant subspaces of $\mathcal{C}_\gamma(\mathcal{H})$ coincide with those of $\mathcal{H}$.
2. The eigenvalue of $\mathcal{C}_\gamma(\mathcal{H})$ corresponding to an eigenvalue $\lambda \neq \gamma$ of $\mathcal{H}$ is $\mathcal{C}_\gamma(\lambda) = \frac{\lambda+\gamma}{\lambda-\gamma}$, while γ is mapped to ∞.
3. Eigenvalues of $\mathcal{H}$ in the left half-plane (*c-stable eigenvalues*) correspond to eigenvalues of $\mathcal{C}_\gamma(\mathcal{H})$ inside the unit circle (*d-stable eigenvalues*). In particular, the c-stabilizing invariant subspace of $\mathcal{H}$ is the d-stabilizing invariant subspace of $\mathcal{C}_\gamma(\mathcal{H})$.
4. The Cayley transform of a Hamiltonian matrix is a symplectic matrix.

The Cayley transform can easily be extended to matrix pencils [31] as

$$\mathcal{C}_\gamma(\lambda\mathcal{E} - \mathcal{A}) = \lambda(\mathcal{A} - \gamma\mathcal{E}) - (\mathcal{A} + \gamma\mathcal{E});$$

and the properties extend to the eigenvalues and deflating subspaces of the involved pencils.

3.3 Standard structured form and structured doubling algorithm

In this section we introduce he *structured doubling algorithm* (SDA) [1, 16, 17, 26], a variant of the inverse-free disc function method in which the pencils are kept in a special form which makes the iteration cheaper and better-conditioned.

A pencil $\lambda\mathcal{E} - \mathcal{A}$ with $\mathcal{A}, \mathcal{E} \in \mathbb{R}^{\ell,\ell}$ is said to be in *standard structured form (SSF)* if it can be expressed as

$$\mathcal{A} = \begin{bmatrix} K & 0 \\ -L & I_{\ell-n} \end{bmatrix}, \quad \mathcal{E} = \begin{bmatrix} I_n & -M \\ 0 & N \end{bmatrix}, \tag{30}$$

for suitable matrices $K \in \mathbb{R}^{n,n}$, $L \in \mathbb{R}^{\ell-n,n}$, $M \in \mathbb{R}^{n,\ell-n}$, $N \in \mathbb{R}^{\ell-n,\ell-n}$. Note that the notation for the blocks here is different from the customary one in the SDA literature. This is necessary to avoid a conflict with the common notation in control theory as introduced in Section (1).

It is simple to check that if $\ell = n$, then a pencil in SSF is symplectic if and only if $K = N^T$, $L = L^T$ and $M = M^T$. The pencil (22) is in SSF whenever $E = I$. Therefore, this form applies naturally to discrete-time optimal control problems with nonsingular E. On the other hand, by using the Cayley transform we can turn an invariant subspace problem from a continuous-time control problem into one in the form of a discrete-time control problem. It is shown in [16] how to turn the pencil obtained from the Cayley transform of a Hamiltonian matrix $\lambda(\mathcal{H} - \gamma I) - (\mathcal{H} + \gamma I)$ and transform it into SSF.

With the help of Proposition 2, we can perform the squaring transformation $(\tilde{\mathcal{E}}, \tilde{\mathcal{A}}) = \mathcal{S}(\mathcal{E}, \mathcal{A})$ in such a way that $(\tilde{\mathcal{E}}, \tilde{\mathcal{A}})$ is still in SSF. To this end, we need to determine a solution to the swapping problem (25) that is in SSF, i.e., it has the form

$$\widetilde{\mathcal{A}} = \begin{bmatrix} \widetilde{K} & 0 \\ -\widetilde{L} & I_{\ell-n} \end{bmatrix}, \quad \widetilde{\mathcal{E}} = \begin{bmatrix} I_n & -\widetilde{M} \\ 0 & \widetilde{N} \end{bmatrix}. \tag{31}$$

This condition is equivalent to the condition

$$\begin{bmatrix} \widetilde{E}_{11} & \widetilde{A}_{12} \\ \widetilde{E}_{21} & \widetilde{A}_{22} \end{bmatrix} = I_\ell.$$

By Proposition 2 we obtain

$$\begin{bmatrix} \widetilde{K} & \widetilde{M} \\ \widetilde{L} & \widetilde{N} \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} I_n & -M \\ -L & I_{\ell-n} \end{bmatrix}^{-1},$$

and thus

$$\widetilde{\mathcal{A}} = \begin{bmatrix} K(I_n - ML)^{-1} & 0 \\ -N(I_{\ell-n} - LM)^{-1}L & I_n \end{bmatrix}, \quad \widetilde{\mathcal{E}} = \begin{bmatrix} I_n & -K(I_n - ML)^{-1}M \\ 0 & N(I_{\ell-n} - LM)^{-1} \end{bmatrix}. \tag{32}$$

The matrices of the resulting pencil $\lambda\widehat{\mathcal{E}} - \widehat{\mathcal{A}} = \mathcal{S}(\mathcal{E}, \mathcal{A}) = \widetilde{\mathcal{A}}\mathcal{A} - s\widetilde{\mathcal{E}}\mathcal{E}$ are still in SSF, and are given by

$$\begin{aligned} \widehat{\mathcal{A}} &= \begin{bmatrix} K(I_n - ML)^{-1}K & 0 \\ -(L + N(I_{\ell-n} - LM)^{-1}LK) & I_{\ell-n} \end{bmatrix}, \\ \widehat{\mathcal{E}} &= \begin{bmatrix} I_n & -(M + K(I_n - ML)^{-1}MN) \\ 0 & N(I_{\ell-n} - LM)^{-1}N \end{bmatrix}. \end{aligned} \tag{33}$$

Theoretically, the only hypothesis required to carry out these operations is that the matrices $I - ML$ and $I - LM$ are nonsingular, and actually they are nonsingular simultaneously.

Proposition 4. *Let $L \in \mathbb{R}^{n,m}$, $M \in \mathbb{R}^{n,m}$. Then $I_n - ML$ is nonsingular if and only if $I_m - LM$ is nonsingular.*

Proof. It is a basic fact in linear algebra, see e.g. [24, Chapter 4] that the two products LM and ML have the same spectrum, except for the zero eigenvalues. In particular, ML has the eigenvalue 1 if and only if LM does. □

If we apply this doubling transformation repeatedly, then we expect the blocks K, N to converge to 0 and the right nullspaces of the limiting matrices to converge to the *graph subspaces*

$$\begin{bmatrix} I_n \\ L_\infty \end{bmatrix}, \quad \begin{bmatrix} M_\infty \\ I_{\ell-n} \end{bmatrix}, \tag{34}$$

respectively, which in the case of problems arising from optimal control are associated with solutions to algebraic Riccati equations given by L_∞, M_∞.

Algorithm 2 implements this variant of the inverse-free disc method (Algorithm 1) by updating directly the blocks of the involved matrices. If no further properties of the blocks in Algorithm 2

Algorithm 2: Structured Doubling Algorithm

input : K_0, N_0, M_0, L_0 defining a pencil in standard structured form
output: L_∞,M_∞ so that the subspaces in (34) are respectively the canonical semi-d-stable and semi-d-unstable deflating subspaces of the given pencil

$k \longleftarrow 0$;
while *a suitable stopping criterion is not satisfied* **do**
  $\widetilde{K_k} \longleftarrow K_k(I_n - M_k L_k)^{-1}$;
  $\widetilde{N_k} \longleftarrow N_k(I_{\ell-n} - L_k M_k)^{-1}$;
  $M_{k+1} \longleftarrow M_k + \widetilde{K_k} M_k N_k$;
  $L_{k+1} \longleftarrow L_k + \widetilde{N_k} L_k K_k$;
  $K_{k+1} \longleftarrow \widetilde{K_k} K_k$;
  $N_{k+1} \longleftarrow \widetilde{N_k} N_k$;
  $k \longleftarrow k + 1$.
end
$L_\infty = L_k$;
$M_\infty = M_k$.

are assumed, then each step of the algorithm costs $\frac{14}{3}\ell^3 - 8\ell n(\ell - n)$ floating point operations. This amounts to $\frac{64}{3}n^3$ operations in the case $\ell = 2n$.

3.4 Convergence Analysis

To analyze whether Algorithm 2 is well defined, we apply the following result, where for symmetric matrices $M, L \in \mathbb{R}^{n,n}$, we write $M \geq L$ ($M > L$) if $M - L$ is positive semidefinite (positive definite).

Theorem 5 ([29]). *Consider a pencil in standard structured form with $\ell - n = n$, $M_0 = M_0^T \geq 0$, $L_0 = L_0^T \leq 0$, and $K = N^T$. If we apply Algorithm 2 to this pencil, then $I - L_k M_k$ and $I - M_k L_k$ are nonsingular at every step and the factors satisfy*

$$\begin{aligned} &0 \leq M_0 \leq M_1 \leq \cdots \leq M_k \leq \cdots \leq M_\infty, \\ &0 \geq L_0 \geq L_1 \geq \cdots \geq L_k \geq \cdots \geq L_\infty. \end{aligned} \tag{35}$$

In the case of symplectic matrices arising in control theory, the assumptions of this theorem are typically satisfied, [33], so that every step of Algorithm 2 is well-defined. However, the doubling algorithm is also applied in cases when $-L_0$ and M_0 are not symmetric positive semidefinite [25]. In this case, the algorithm may break down when $I - L_k M_k$ becomes singular or ill-conditioned at some step of the algorithm.

It was recently proved in [25] that Algorithm 2 converges even in the presence of unimodular eigenvalues, when a special condition on the unimodular Jordan chains holds. For this result, we need a little more notation.

An $\ell \times \ell$ pencil is called *weakly d-split* if there is an integer $0 \leq r \leq \ell$ such that

- the sum of all partial multiplicities associated with the d-stable eigenvalues is $\ell - r$;
- the sum of all partial multiplicities associated with d-unstable eigenvalues is $\ell - r$;
- the partial multiplicities $2r_j$ of all d-critical (unimodular) eigenvalues are all even (and they sum up to $2r$ if the two previous properties hold).

The *canonical d-semi-stable subspace* of a matrix pencil is then the deflating subspace spanned by all eigenvectors and generalized eigenvectors (Jordan chains) associated with the d-stable eigenvalues and by the first r_j vectors of any Jordan chain of size $2r_j$ associated with a critical eigenvalue. The *canonical d-semi-unstable subspace* is defined analogously. In the context of standard control problems ($E = I$), the canonical d-semi-unstable and d-semi-stable subspaces are the (unique) Lagrangian ones, see e.g. [28].

Theorem 6 ([25]). *Consider Algorithm 2 applied to a weakly d-split pencil in the form* (30). *Suppose that there is no breakdown, and that there exist matrices $M_\infty \in \mathbb{R}^{n,m}, L_\infty \in \mathbb{R}^{m,n}$ such that the columns of the matrices in (34) span a canonical d-semi-stable and d-semi-unstable deflating subspace, respectively. Then, the following asymptotic bounds hold for $k \to \infty$.*

1. $\|K_k\| = O(2^{-k})$,
2. $\|N_k\| = O(2^{-k})$,
3. $\|L_k - L_\infty\| = O(2^{-k})$,
4. $\|M_k - M_\infty\| = O(2^{-k})$.

Moreover, if the pencil has no unimodular eigenvalues, then the convergence is quadratic, and in particular in parts 3. and 4., the term $O(2^{-k})$ can be replaced with $O(\nu^{2^k})$, with

$$\nu = \frac{|\lambda_s|}{|\lambda_u|},$$

where λ_s is a d-stable eigenvalue of maximal modulus, and λ_u a d-unstable eigenvalue of minimal modulus.

Algorithm 2 has very nice convergence properties but it is clear that the convergence is slow whenever there are eigenvalues on or close to the unit circle. Furthermore, the algorithm makes the strong assumption that the d-semi-stable and d-semi-unstable deflating subspaces have a representation as graph subspaces (34), which as we have already noted in the applications from control theory may not be the case, or even if so then this representation may be very ill-conditioned to compute. Then the SDA algorithm is expected to perform badly, as we demonstrate in the numerical section below.

4 Generalized SDA

In this section we construct a variant of Algorithm 2 in which we drop the requirement that the desired invariant subspaces are graph subspaces; i.e., that there are always two identity blocks in (30). This not only allows the treatment of problems where in the limiting deflating subspace one or both of the blocks L_∞, M_∞ become singular, but it also allows to treat the case of singular E as well.

To this aim, we relax the assumptions on the standard structured form, and we work with pencils in the form

$$\mathcal{A} = \begin{bmatrix} K & 0 \\ L^{(1)} & L^{(2)} \end{bmatrix}, \quad \mathcal{E} = \begin{bmatrix} M^{(1)} & M^{(2)} \\ 0 & N \end{bmatrix}, \tag{36}$$

which we call *weak standard structured form* (WSSF). Unlike the standard structured form, there is no unique WSSF pencil that is equivalent to a given regular pencil under left-multiplication by an invertible matrix. In fact, we can left-multiply both $\mathcal{A}$ and $\mathcal{E}$ by any matrix in the form

$$\begin{bmatrix} S^{(1)} & 0 \\ 0 & S^{(2)} \end{bmatrix}$$

without altering the WSSF or the right deflating subspaces. Thus, a normalization is needed to make the representation unique. The choice of such a normalization and its consequences are discussed in the following sections.

4.1 Transforming continuous- to discrete-time problems

As the SDA, our algorithm operates on a pencil arising from the discrete-time setting and the Cayley transform can be used to obtain a discrete-time problem from a continuous-time one. It is shown in [16] how to transform the Cayley transform of a Hamiltonian matrix (7) to a SSF pencil, to which SDA can be applied. However, this approach relies heavily on the symplectic structure and does not generalize immediately to the case in which E is singular. Furthermore, even when we can obtain a factorization to SSF, in the resulting pencil the blocks $L^{(1)}$ and $M^{(2)}$ are in general not symmetric. Thus we do not obtain a complete analogy with (23) and the obtained algorithm is not completely structure-preserving. Instead, we may apply techniques introduced in [41, 42] directly on the unreduced pencil (4). Let us briefly review these results.

Theorem 7 ([41, 42]). *Let $E_c, A_c, B_c, Q_c, S_c, R_c$ define a continuous-time optimal control problem* (4)*, and let a corresponding discrete-time problem $E_d, A_d, B_d, Q_d, S_d, R_d$ be defined by*

$$\begin{bmatrix} E_d & 0 \\ A_d & B_d \end{bmatrix} = \frac{\sqrt{2}}{2} \begin{bmatrix} E_c - A_c & -B_c \\ E_c + A_c & B_c \end{bmatrix} \mathcal{W},$$
$$\begin{bmatrix} Q_d & S_d \\ S_d^* & R_d \end{bmatrix} = \mathcal{W}^* \begin{bmatrix} Q_c & S_c \\ S_c^* & R_c \end{bmatrix} \mathcal{W},$$

where $\mathcal{W} \in C^{n+r,n+r}$ is any nonsingular matrix such that $\begin{bmatrix} E_c - A_c & -B_c \end{bmatrix} \mathcal{W} = \begin{bmatrix} E_d & 0 \end{bmatrix}$ holds (for instance, we may perform a block LQ or LU factorization). If

$$U_d = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$$

is a deflating subspace for the discrete-time problem (20) *(with all the terms subscripted with d) whose associated eigenvalues are not -1 nor ∞, then we have $\mathcal{E}_d U_d T_d = \mathcal{A}_d U_d$ for some matrix T_d, and*

$$U_c \begin{bmatrix} U_1(I + T_d) \\ \mathcal{W}^{-1} U_2 \end{bmatrix}$$

is a deflating subspace for the continuous-time problem (4) *(with all the terms subscripted by c) such that $\mathcal{E}_c U_c T_c = \mathcal{A}_c U_c$, and $T_c = (T_d - I)(T_d + I)^{-1}$ is the inverse Cayley transform of T_d. In particular, the (canonical) d-semi-stable deflating subspace of the discrete-time problem corresponds to the (canonical) d-semi-stable subspace of the continuous one.*

By preprocessing the problem in a suitable way, it is possible to ensure that the condition on the eigenvalues of the deflating subspace ($\lambda \neq -1, \infty$) is satisfied [42, Section 6], although this may require using complex arithmetics even for a real problem.

4.2 Derivation of a generalized structured doubling algorithm

By modifying the derivation of the structured doubling algorithm, we can derive a similar algorithm which maintains only the more general WSSF along the iterates. If we impose that

$$\begin{bmatrix} \widetilde{E}_{11} & \widetilde{A}_{12} \\ \widetilde{E}_{21} & \widetilde{A}_{22} \end{bmatrix} = I_\ell, \tag{37}$$

then Theorem 2 yields

$$\begin{bmatrix} \widetilde{K} & \widetilde{M} \\ \widetilde{L} & \widetilde{N} \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} M^{(1)} & -M^{(2)} \\ -L^{(1)} & L^{(2)} \end{bmatrix}^{-1}.$$

Thus, setting

$$\begin{bmatrix} M^{(1)} & -M^{(2)} \\ -L^{(1)} & L^{(2)} \end{bmatrix}^{-1} =: \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}, \tag{38}$$

we obtain an analogue of (33)

$$\widehat{\mathcal{A}} = \begin{bmatrix} KZ_{11}K & 0 \\ L^{(1)} + NZ_{21}K & L^{(2)} \end{bmatrix}, \quad \widehat{\mathcal{E}} = \begin{bmatrix} M^{(1)} & M^{(2)} + KZ_{12}N \\ 0 & NZ_{22}N \end{bmatrix}. \tag{39}$$

Remark 8. Note that the update relation (39) fits well to the freedom in the normalization in the WSSF. In fact, if we start with a WSSF with a different normalization

$$\begin{bmatrix} S^{(1)} & 0 \\ 0 & S^{(2)} \end{bmatrix} \mathcal{A} = \begin{bmatrix} S^{(1)} & 0 \\ 0 & S^2 \end{bmatrix} \begin{bmatrix} K & 0 \\ L^{(1)} & L^{(2)} \end{bmatrix},$$
$$\begin{bmatrix} S^{(1)} & 0 \\ 0 & S^{(2)} \end{bmatrix} \mathcal{E} = \begin{bmatrix} S^{(1)} & 0 \\ 0 & S^{(2)} \end{bmatrix} \begin{bmatrix} M^{(1)} & M^{(2)} \\ 0 & N \end{bmatrix},$$

then following the same steps we get to

$$\begin{bmatrix} S^{(1)} & 0 \\ 0 & S^{(2)} \end{bmatrix} \widehat{\mathcal{A}}, \begin{bmatrix} S^{(1)} & 0 \\ 0 & S^{(2)} \end{bmatrix} \widehat{\mathcal{E}}. \tag{40}$$

Similarly, if we replace (37) with the more general form

$$\begin{bmatrix} \widetilde{E}_{11} & \widetilde{A}_{12} \\ \widetilde{E}_{21} & \widetilde{A}_{22} \end{bmatrix} = \begin{bmatrix} S^{(1)} & 0 \\ 0 & S^{(2)} \end{bmatrix},$$

then the introduced multiplicative factors can be factored out as in (40).

Therefore, changing the normalization factors before the procedure or introducing a new degree of freedom in (37) is equivalent to carrying out the pencil squaring procedure and normalizing only after the iteration step.

By iterating the pencil transformation as described, we get the following generalized structured doubling (gSDA) Algorithm 3. For the sake of brevity, in the algorithm and the following analysis, we set

$$\mathcal{L}_k := \begin{bmatrix} L_k^{(1)} & L_k^{(2)} \end{bmatrix}, \ \mathcal{M}_k := \begin{bmatrix} M_k^{(1)} & M_k^{(2)} \end{bmatrix}.$$

The iterated application of Remark 8 leads to an observation that is useful in the theoretical analysis.

Remark 9. Let $\mathcal{A}_k - z\mathcal{E}_k$ and $\mathcal{A}'_k - z\mathcal{E}'_k$, for $k = 0, 1, 2, \ldots$ be two sequences obtained with Algorithm 3, the former with normalization factors $S_k^{(1)}, S_k^{(2)}$ and the latter with ${S'}_k^{(1)}, {S'}_k^{(2)}$. Then for each $k = 0, 1, 2, \ldots$

$$\mathcal{A}'_k - z\mathcal{E}'_k = \begin{bmatrix} \Sigma_k^{(1)} & 0 \\ 0 & \Sigma_k^{(2)} \end{bmatrix} (\mathcal{A}_k - z\mathcal{E}_k),$$

with

$$\Sigma_k^{(i)} = {S'}_k^{(i)} \left(S_k^{(i)}\right)^{-1} {S'}_{k-1}^{(i)} \left(S_{k-1}^{(i)}\right)^{-1} \cdots {S'}_0^{(i)} \left(S_0^{(i)}\right)^{-1}, \quad i = 1, 2.$$

A natural choice for the normalization factors is choosing them so that $\mathcal{L}_{k+1}^T$ and $\mathcal{M}_{k+1}^T$ are orthogonal. With this choice we then obtain Algorithm 4.

4.3 Applicability of the generalized structured doubling algorithm

As in the standard SDA, we need to ensure that the matrix to invert in (38) is nonsingular. When $E = I$, then SDA and gSDA differ only by a (nonsingular) normalization factor. In view of Remark 9, the nonsingularity of this matrix is therefore equivalent to the nonsingularity of $I - M_k L_k$ in SDA, which can often be ensured (see e.g. Theorem 5).

In the case that E is singular, then without a proper regularization (see Section 1) it cannot be ensured that the algorithm can be carried out without breakdown.

Algorithm 3: gSDA

input : K_0, L_0, M_0, N_0 defining a pencil in SSF
output: U,V spanning respectively the canonical d-semi-stable and d-semi-unstable subspace
$k \longleftarrow 0$;
while *a suitable stopping criterion is not satisfied* **do**

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \longleftarrow \begin{bmatrix} M_k^{(1)} & -M_k^{(2)} \\ -L_k^{(1)} & L_k^{(2)} \end{bmatrix}^{-1};$$

$M_{k+\frac{1}{2}}^{(1)} \longleftarrow M_k^{(1)}$;
$M_{k+\frac{1}{2}}^{(2)} \longleftarrow M_k^{(2)} + K_k Z_{12} N_k$;
$L_{k+\frac{1}{2}}^{(1)} \longleftarrow L_k^{(1)} + N_k Z_{21} K_k$;
$L_{k+\frac{1}{2}}^{(2)} \longleftarrow L_k^{(2)}$;
$K_{k+\frac{1}{2}} \longleftarrow K_k Z_{11} K_k$;
$N_{k+\frac{1}{2}} \longleftarrow N_k Z_{22} N_k$.
Choose suitable nonsingular normalization factors $S_k^{(1)} \in \mathbb{R}^{n,n}$, $S_k^{(2)} \in \mathbb{R}^{\ell-n,\ell-n}$;
$\mathcal{M}_{k+1} \longleftarrow S_k^{(1)} \mathcal{M}_{k+\frac{1}{2}}$;
$\mathcal{L}_{k+1} \longleftarrow S_k^{(2)} \mathcal{L}_{k+\frac{1}{2}}$;
$K_{k+1} \longleftarrow S_k^{(1)} K_{k+\frac{1}{2}}$;
$N_{k+1} \longleftarrow S_k^{(2)} N_{k+\frac{1}{2}}$;
$k \longleftarrow k+1$;
end
$U \longleftarrow \operatorname{null} \mathcal{L}_k$;
$V \longleftarrow \operatorname{null} \mathcal{M}_k$.

4.4 Convergence of the generalized structured doubling algorithm

The following theorem generalizes the convergence results of the structured doubling algorithm given in [14, 25].

Theorem 10. *Suppose that Algorithm 3 can be applied with no breakdown to a weakly d-split starting pencil, and that the normalization factors $S_k^{(i)}$ are chosen in any way such that $\|\mathcal{L}_k\|$ and $\|\mathcal{M}_k\|$ are bounded. Moreover, let*

$$U = \begin{bmatrix} U^{(1)} \\ U^{(2)} \end{bmatrix} \in \mathbb{R}^{\ell,n}, \quad V = \begin{bmatrix} V^{(1)} \\ V^{(2)} \end{bmatrix} \in \mathbb{R}^{\ell,\ell-n}$$

be any two matrices spanning the canonical d-semi-stable and d-semi-unstable subspaces of the initial pencil (30)*, respectively. Then, the following implications hold.*

1. *if $U^{(1)}$ is nonsingular, then $\|K_k\| = O(2^{-k})$;*
2. *if $\|K_k\| = O(2^{-k})$, then $\|\mathcal{M}_k V\| = O(2^{-k})$;*
3. *if $V^{(2)}$ is nonsingular, then $\|N_k\| = O(2^{-k})$;*
4. *if $\|N_k\| = O(2^{-k})$, then $\|\mathcal{L}_k U\| = O(2^{-k})$;*

If the pencil has no eigenvalues of modulus 1*, then the convergence is quadratic, and in particular*

$$\|\mathcal{M}_k V\| = O\left(\left(\frac{|\lambda_s|}{|\lambda_u|}\right)^{2^k}\right), \ \|\mathcal{L}_k U\| = O\left(\left(\frac{|\lambda_s|}{|\lambda_u|}\right)^{2^k}\right),$$

where λ_s is a stable eigenvalue of maximal modulus, and λ_u is an unstable eigenvalue of minimal modulus.

Algorithm 4: gSDA with orthonormalization

input : K_0, L_0, M_0, N_0 defining a pencil in SSF
output: U,V spanning respectively the canonical d-semi-stable and d-semi-unstable subspace
$k \longleftarrow 0$;
while *a suitable stopping criterion is not satisfied* **do**

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \longleftarrow \begin{bmatrix} M_k^{(1)} & -M_k^{(2)} \\ -L_k^{(1)} & L_k^{(2)} \end{bmatrix}^{-1};$$

$M_{k+\frac{1}{2}}^{(1)} \longleftarrow M_k^{(1)}$;
$M_{k+\frac{1}{2}}^{(2)} \longleftarrow M_k^{(2)} + K_k Z_{12} N_k$;
$L_{k+\frac{1}{2}}^{(1)} \longleftarrow L_k^{(1)} + N_k Z_{21} K_k$;
$L_{k+\frac{1}{2}}^{(2)} \longleftarrow L_k^{(2)}$;
$K_{k+\frac{1}{2}} \longleftarrow K_k Z_{11} K_k$;
$N_{k+\frac{1}{2}} \longleftarrow N_k Z_{22} N_k$;
$[Q_{\mathcal{L}}, R_{\mathcal{L}}] \longleftarrow \mathtt{qr}({\mathcal{L}_{k+\frac{1}{2}}}^T)$;
$\mathcal{L}_{k+1} \longleftarrow Q_{\mathcal{L}}^T$;
$N_{k+1} \longleftarrow R_{\mathcal{L}}^{-T} N_{k+\frac{1}{2}}$;
$[Q_{\mathcal{M}}, R_{\mathcal{M}}] \longleftarrow \mathtt{qr}({\mathcal{M}_{k+\frac{1}{2}}}^T)$;
$\mathcal{M}_{k+1} \longleftarrow Q_{\mathcal{M}}^T$;
$K_{k+1} \longleftarrow R_{\mathcal{M}}^{-T} K_{k+\frac{1}{2}}$;
$N_{k+1} \longleftarrow S_k^{(2)} N_{k+\frac{1}{2}}$;
$k \longleftarrow k+1$;
end
$U \longleftarrow \operatorname{null} \mathcal{L}_k$;
$V \longleftarrow \operatorname{null} \mathcal{M}_k$.

Proof. Since the proof is rather technical, we present it in Appendix A. □

Theorem 10 is somewhat unsatisfactory from our point of view. While the boundedness of the blocks $\mathcal{L}_k$ and $\mathcal{M}_k$ can be achieved with a suitable normalization, the hypothesis that $U^{(1)}$ and $V^{(2)}$ are invertible may not be easily available, with a consequence that we have convergence to a wrong subspace as the following example shows.

Example 11. Consider the pencil

$$\lambda \mathcal{E} - \mathcal{A} = \lambda \begin{bmatrix} 1 & 0 \\ 0 & \lambda_0 \end{bmatrix} - \begin{bmatrix} \lambda_0 & 0 \\ 0 & 1 \end{bmatrix}. \tag{41}$$

for $\lambda_0 > 1$, for which $K_k = N_k = \lambda^{2^k}$, $\mathcal{M}_k = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $\mathcal{L}_k = \begin{bmatrix} 0 & 1 \end{bmatrix}$. Here $\operatorname{null} \mathcal{L}$ is the d-unstable deflating subspace and $\operatorname{null} \mathcal{M}$ is the d-stable one, exactly the opposite of what we want to achieve. It is clear that in this example a simple permutation similarity solves the problem.

In fact, it is easy to understand the problem with the choice of the initial pencil and to provide a fix. For this, consider the (orthogonal) subspace iteration [23, section 7.3], which for a matrix $T \in \mathbb{R}^{n,n}$ is as follows. Starting from an initial subspace $\mathcal{W}_0 = \operatorname{Span} W_0$ one iterates the transformation $W_{k+1} = TW_k$, and orthogonalizes the columns of W_k in order to maintain a numerically stable basis for $\mathcal{W}_k = \operatorname{Span} W_k$. If $w = \dim \mathcal{W}_0$, and if the w dominant eigenvalues of T can be identified (i.e., if we order the eigenvalues of T so that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$, then we must have $|\lambda_w| > |\lambda_{w+1}|$), and $\mathcal{W}_k$ converges linearly to the invariant subspace relative to $\lambda_1, \lambda_2, \ldots, \lambda_w$. However, if the columns of the starting matrix already span an invariant subspace associated with some undesired eigenvalues then clearly the subspace iteration method cannot converge. The same analysis applied to the generalized structured doubling algorithm shows that in

an 'unlucky' situation the columns of our starting matrix may be contained in a deflating subspace. In general, however, this is unlikely if random starting matrices are used. If however, the starting matrices are close to such a bad situation, then we can expect slow convergence in the beginning until the algorithm has overcome this problem.

In order to see what we need to start the algorithm in an appropriate way, consider the following result.

Theorem 12. *Suppose that both $\mathcal{E}_0$ and $\mathcal{A}_0$ are nonsingular and form the nonsingular matrix $\mathcal{T} = \mathcal{E}_0^{-1}\mathcal{A}_0$. Then, in the generalized structured doubling algorithm we have that*

- $\operatorname{null} \mathcal{L}_k$ *coincides with the subspace obtained by applying 2^k steps of orthogonal subspace iteration to $\mathcal{T}^{-1}$, starting with $W_0 = [I_n\ 0]^T \in \mathbb{R}^{\ell,n}$;*
- $\operatorname{null} \mathcal{M}_k$ *coincides with the subspace obtained by applying 2^k steps of orthogonal subspace iteration to $\mathcal{T}$, starting with $W_0 = [0\ I_{\ell-n}]^T \in \mathbb{R}^{\ell,\ell-n}$.*

Proof. Let $U_k \in \mathbb{R}^{\ell,n}$, $V_k \in \mathbb{R}^{\ell,\ell-n}$ be such that $\operatorname{Span} U_k = \operatorname{null} \mathcal{L}_k$, $\operatorname{Span} V_k = \operatorname{null} \mathcal{M}_k$, and let

$$U_k = \begin{bmatrix} U_k^{(1)} \\ U_k^{(2)} \end{bmatrix}.$$

Then,

$$\begin{aligned} \mathcal{M}^{2^k} U_k =& (\mathcal{E}_0^{-1}\mathcal{A}_0)^{2^k} U_k = (\mathcal{E}_k^{-1}\mathcal{A}_k) U_k = \begin{bmatrix} M_k^{(1)} & M_k^{(2)} \\ 0 & N_k \end{bmatrix}^{-1} \begin{bmatrix} K_k & 0 \\ L_k^{(1)} & L_k^{(2)} \end{bmatrix} U_k \\ =& \begin{bmatrix} M_k^{(1)} & M_k^{(2)} \\ 0 & N_k \end{bmatrix}^{-1} \begin{bmatrix} K_k U_k^{(1)} \\ 0 \end{bmatrix} = \begin{bmatrix} \left(M_k^{(1)}\right)^{-1} K_k U_k^{(1)} \\ 0 \end{bmatrix} \subseteq \operatorname{Span} \begin{bmatrix} I_n \\ 0 \end{bmatrix}. \end{aligned}$$

A similar argument starting from $\mathcal{T}^{-2^k} V_0$ yields the second part. □

Theorem 12 can be used to give a more direct proof of Theorem 10 in the special case that the nonsingularity assumptions hold, by considering the behavior of the subspace iteration for $\lambda_w = \lambda_{w+1}$ in the presence of nontrivial Jordan blocks. The nonsingularity assumption, however, are what we wanted to avoid.

It is a well established fact [40] that the presence of rounding errors (by destroying certain Jordan blocks) is usually beneficial for the subspace iteration, as it often makes the method converge to the correct subspace even if the starting subspace is deficient in the directions of some of the leading (generalized) eigenvectors. Something similar happens with the SDA. When $U^{(1)}$ is singular, then the starting subspace $\mathcal{U}_0$ and U are not in generic position (as can easily be checked by computing $U^T\mathcal{U}_0$), but generically this exact orthogonality is lost in finite precision arithmetic. The computational results in Section 5 illustrate this. In view of Theorem 10, convergence to the correct subspace happens if K_k and N_k converge to zero, so this provides a practical criterion to check whether this effect has happened.

4.5 Structure preservation

We have introduced the generalized SDA in its most general form, in which no further structure is imposed on the blocks, as it may turn out useful in other contexts, e.g. [14]. However, in the pencils arising from optimal control, there is further structure to exploit. As we have shown in Section 1, in the context of discrete-time optimal control we obtain BVD-pencils or reduced BVD-pencils. Let us consider now a reduced BVD-pencils in WSSF with a matrix $E \in \mathbb{R}^{n,n}$ that may be singular or nonsingular. We call this an E-BVD pencil if $\ell - n = n$, $M^{(1)} = \left(L^{(2)}\right)^T = E$, $N = K^T$, $M^{(2)} = \left(M^{(2)}\right)^T$, $L^{(1)} = \left(L^{(1)}\right)^T$. Note that when $E = I$ then the pencil is symplectic.

We then have the following result.

Proposition 13. *Consider the generalized structured doubling algorithm applied to an E-BVD starting pencil.*

- *If the normalization factors $S_k^{(1)} = S_k^{(2)} = I$ are used, then at every step of the iteration the resulting pencil is E-BVD.*
- *At every step of the iteration, there exists a nonsingular normalization matrix*

$$\begin{bmatrix} \Sigma_k^{(1)} & 0 \\ 0 & \Sigma_k^{(2)} \end{bmatrix} \tag{42}$$

(with blocks of conformal size) such that a left multiplication by this matrix yields an E-BVD pencil.

Proof. The first part is easily shown by induction, by noticing that

$$\left(\begin{bmatrix} M_k^{(2)} & -M_k^{(1)} \\ -L_k^{(2)} & L_k^{(1)} \end{bmatrix} \right)^{-1} = \begin{bmatrix} Z_{21} & Z_{22} \\ Z_{11} & Z_{12} \end{bmatrix} \tag{43}$$

is symmetric. The second part follows directly from Remark 9. □

We also have a monotonicity result that generalizes Theorem 5.

Theorem 14. *Suppose that the gSDA is applied to an E-BVD starting pencil, with $M_0^{(2)} \leq 0$ and $L_0^{(1)} \geq 0$. If the normalization factors are identities, then $0 \leq L_0^{(1)} \leq L_1^{(1)} \leq \cdots \leq L_k^{(1)} \leq \cdots$ and $0 \geq M_0^{(2)} \geq M_1^{(2)} \geq \cdots \geq M_k^{(2)} \geq \cdots$, i.e., the entries in the two blocks converge monotonically.*

Proof. We prove by induction that $L_k^{(1)} \leq L_{k+1}^{(1)}$. For every $\lambda > 0$, the $(1,1)$ block of the inverse of

$$\begin{bmatrix} M_k^{(2)} & -M_k^{(1)} \\ -L_k^{(2)} & L_k^{(1)} \end{bmatrix} + \lambda \begin{bmatrix} -I_n & 0 \\ 0 & I_n \end{bmatrix}$$

satisfies $(L_k^{(1)} + \lambda I - E^T(M_k^{(2)} - \lambda I)^{-1}E)^{-1} > 0$. Thus, its limit Z_{21} is positive semidefinite, and hence $L_{k+1}^{(1)} - L_k^{(1)} = N_k Z_{21} N_k^T \geq 0$. The second part is proved analogously. □

4.6 The symplectic case

In the case that $E = I$, the starting pencil is symplectic, i.e., $K_0 = N_0^T$, $M_0^{(2)} = \left(M_0^{(2)}\right)^T$, $L_0^{(1)} = \left(L_0^{(1)}\right)^T$, $M^{(1)} = L^{(2)} = I$. In this case, if we choose the normalization $S^{(1)} = S^{(2)} = I$, then the traditional SDA coincides with the gSDA and it preserves the symplectic structure of the pencil at every step. In this case, Algorithm 2 can be slightly simplified, since there is no need to compute K_{k+1} and N_{k+1} separately, nor to invert both $I - M_k L_k$ and $I - L_k M_k$, as the second matrix of both pairs is the transpose of the first.

The following result shows what happens to the symplectic structure under the gSDA.

Theorem 15. *When Algorithm 3 is applied to a symplectic pencil, then for every k we have that $\mathcal{L}_k^T$ and $\mathcal{M}_k^T$ are Lagrangian subspaces, i.e., $L_k^{(2)} {L_k^{(1)}}^T = L_k^{(1)} {L_k^{(2)}}^T$ and $M_k^{(2)} {M_k^{(1)}}^T = M_k^{(1)} {M_k^{(2)}}^T$. Moreover, $L_k^{(2)} K_k^T = N_k {M_k^{(1)}}^T$.*

Proof. By Remark 9, Algorithm 3 differs from traditional SDA (Algorithm 2) only by a normalization factor in the form (42). Direct inspection of the blocks shows that these factors must be $\Sigma_k^{(1)} = {M_k^{(1)}}^{-1}$ and $\Sigma_k^{(2)} = {L_k^{(2)}}^{-1}$. Since in the traditional SDA $K_k = N_k^T$, we get $\left({M_k^{(1)}}^{-1} K_k\right)^T =$

${L_k^{(2)}}^{-1} N_k$, i.e., $L_k^{(2)} K_k^T = N_k {M_k^{(1)}}^T$. Moreover, this normalization corresponds to choosing new bases for the subspaces $\mathcal{L}_k^T$ and $\mathcal{M}_k^T$, without changing the subspaces themselves. Using the iterates of traditional SDA, it is easy to check that the subspaces spanned by $\begin{bmatrix} L_k & I_m \end{bmatrix}^T$ and $\begin{bmatrix} I_n & M_k \end{bmatrix}^T$ are Lagrangian, as this corresponds to L_k and M_k being Hermitian. Since being Lagrangian is a property of the subspace, independent of the choice of basis, this also holds for Algorithm 3. □

The preservation of structure can be exploited in Algorithm 3 to perform a faster (and structured) matrix inversion in (38). Indeed, one sees that

$$\begin{bmatrix} M^{(1)} & -M^{(2)} \\ -L^{(1)} & L^{(2)} \end{bmatrix}^{-1} = \begin{bmatrix} (L^{(2)})^T & (M^{(2)})^T \\ (L^{(1)})^T & (M^{(1)})^T \end{bmatrix}^T \begin{bmatrix} R^{-1} & 0 \\ 0 & -R^T \end{bmatrix},$$

with $R = M^{(1)} {L^{(2)}}^T - M^{(2)} {L^{(1)}}^T$. The final step can be simplified as well, since $\operatorname{null} \mathcal{L} = \mathcal{J}\mathcal{L}^T$.

On the other hand, the normalization choice suggested in Algorithm 3 of using a QR-factorization of $\mathcal{L}_k^T$ and $\mathcal{M}_k^T$) does not guarantee that $N_k = K_k^T$, which would be useful to reduce the computational costs and ensure structure preservation. Ideally, we would like the following properties to hold at the same time.

P1. $N_k = K_k^T$ for each k,

P2. $\mathcal{L}_k$ and $\mathcal{M}_k$ have orthonormal rows.

In fact, the following results show that in general this is not possible.

Proposition 16. *There are no normalization factors $S_k^{(1)}$ and $S_k^{(2)}$ in Algorithm 3 ensuring that for all possible starting pencils* (30) *P1 and P2 hold.*

Proof. The condition $N_k = K_k^T$ means that we must choose $S_k^{(2)} = {S_k^{(1)}}^T = S_k$ at each k. If this and property P2 would hold at the same time, then $S_k \mathcal{L}_{k+\frac{1}{2}} \mathcal{L}_{k+\frac{1}{2}}^T S_k^T = S_k^T \mathcal{M}_{k+\frac{1}{2}} \mathcal{M}_{k+\frac{1}{2}}^T S_k = I$, i.e., $\mathcal{L}_{k+\frac{1}{2}} \mathcal{L}_{k+\frac{1}{2}}^T = S_k^{-1} S_k^{-T}$ and $\mathcal{M}_{k+\frac{1}{2}} \mathcal{M}_{k+\frac{1}{2}}^T = S_k^{-T} S_k^{-1}$. This would imply that $\mathcal{L}_{k+\frac{1}{2}} \mathcal{L}_{k+\frac{1}{2}}^T$ and $\mathcal{M}_{k+\frac{1}{2}} \mathcal{M}_{k+\frac{1}{2}}^T$ have the same eigenvalues, but this does not hold in general. □

So we cannot expect to have properties P1 and P2 at the same time. Asking for P1 yields the traditional SDA, or in general algorithms that tend to have convergence problems in the case where $U^{(1)}$ and $V^{(2)}$ are close to singular. Requiring property P2 yields Algorithm 3, which (as our experiments illustrate) performs better for this class of problems.

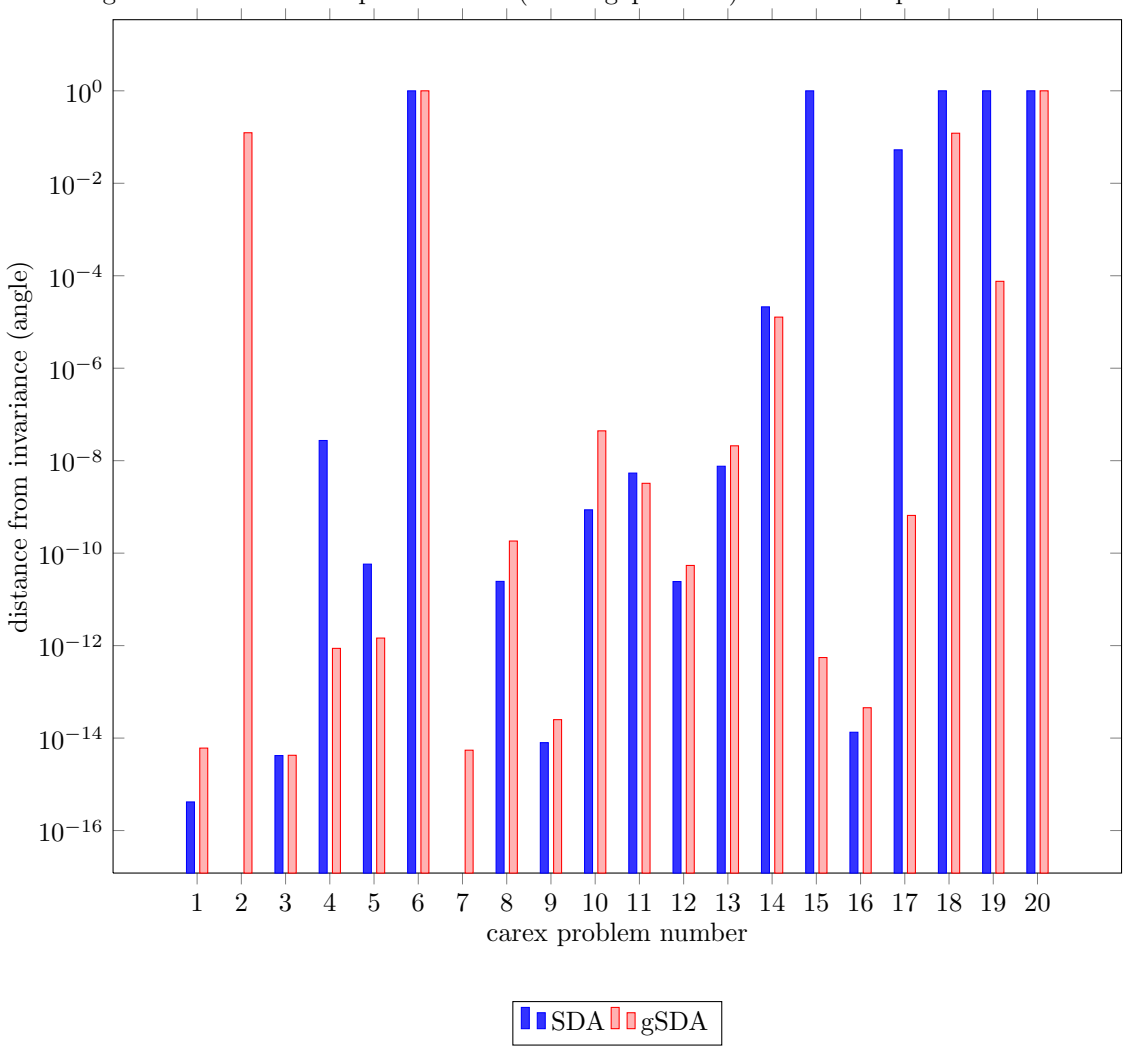
5 Numerical experiments

The examples in the benchmark suites carex [8] and darex [9] are designed for the solution of Riccati equations, i.e., in our language, they assume nonsingular $U^{(1)}$ and $V^{(2)}$ blocks from the beginning. In order to generate examples with numerically singular $U^{(1)}$ and $V^{(2)}$, we chose problems in carex and considered the optimal control problem with Hamiltonian matrix $-\mathcal{H}$ instead of $\mathcal{H}$. This has the effect of switching the c-stable and the c-unstable subspace, i.e., swapping $U^{(1)}$, $V^{(2)}$ with $V^{(1)}$, $U^{(2)}$, respectively. In this way the original carex problems with a singular solution X are transformed into problems for which $V^{(2)}$ is numerically singular. As we see, several of these problems, which are nonetheless *well-posed* optimal control problems, lead to convergence failure or large errors in the SDA.

For each problem in carex (after the sign switch of the Hamiltonian), we computed the canonical c-semi-stable and c-semi-unstable invariant subspace U and V, respectively, with both the SDA and the gSDA. To analyze and compare the quality of the results we can use different measures.

The angle between subspaces is defined by $\theta(\mathcal{U}, \mathcal{V}) = \|P_{\mathcal{U}} - P_{\mathcal{V}}\|$, with $P_{\mathcal{U}}$ and $P_{\mathcal{V}}$ being the orthogonal projectors on $\mathcal{U}$ and $\mathcal{V}$ respectively, and $\|\cdot\|$ denoting the matrix norm induced by the Euclidean norm, see e.g. [22]. Then, for the error in the *gap metric* $\theta(\mathcal{U}, \mathcal{H}\mathcal{U})$, Figure 1 displays

Figure 1: Invariant subspace residual (in the gap metric) for the examples in carex



$\max(\theta(\mathcal{U}, \mathcal{HU}), \theta(\mathcal{V}, \mathcal{HV}))$ for all the problems in the carex benchmark set. In the case of parametric problems, the default values are chosen. A Cayley transform with parameter $\gamma = 1$ was chosen to transform the pencils to symplectic SSF, the only exception being Problem 15. for which $\gamma = 2$ is used, since the SSF for the pencil with $\gamma = 1$ does not exist. For Problem 2. and 7., the SDA failed to converge, while the gSDA converged, though with bad accuracy for Problem 2. For several problems in the set the gSDA was significantly more accurate than the SDA; on the other hand, for some problems the SDA achieved marginally better result (no more than 1 significant digit, two only for Problem 10.).

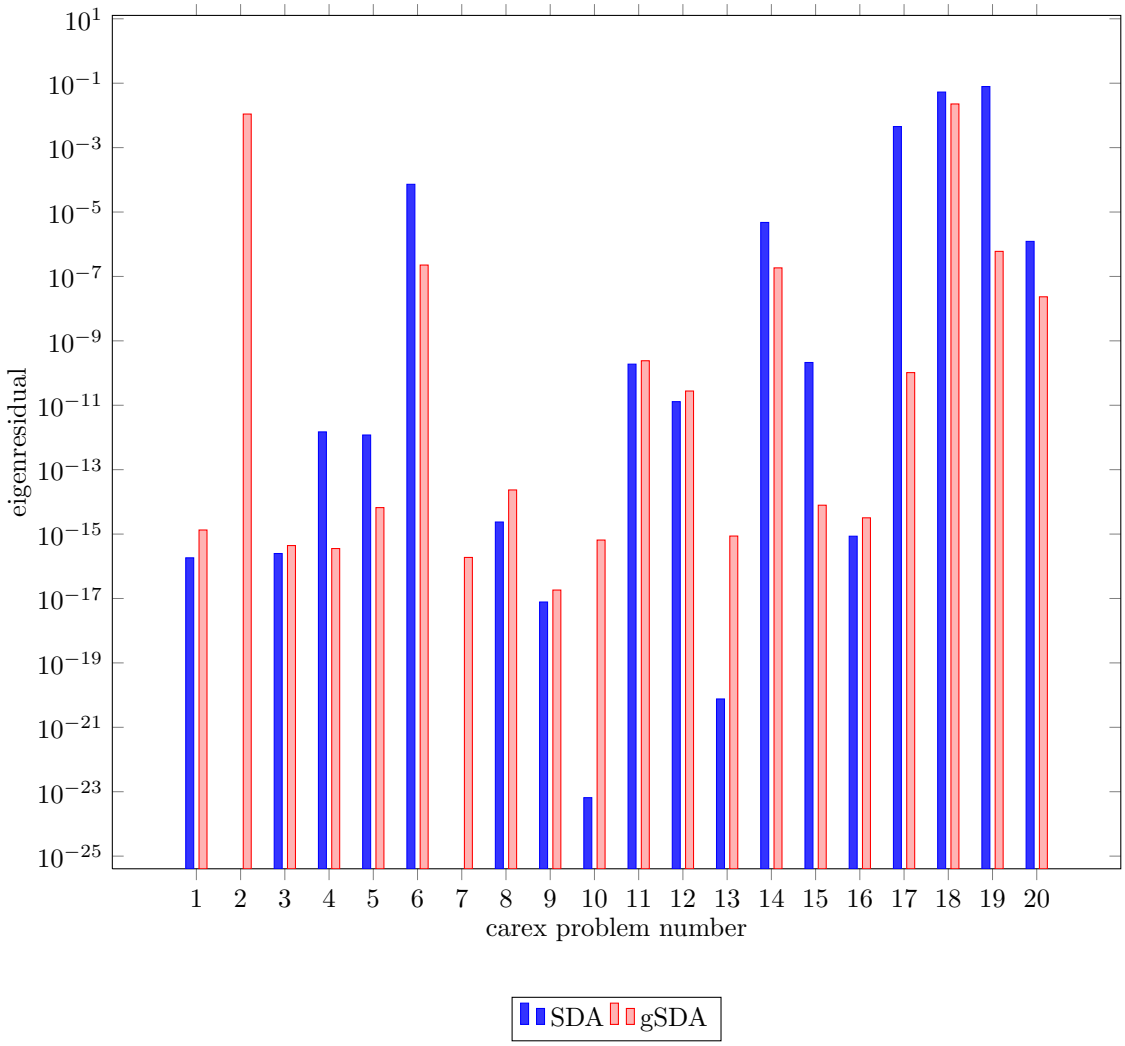
According to [8, table 1], carex Problems 2., 6., 15., 17. and 18. have a solution $\hat{X}$ that is exactly singular, implying that $V^{(2)}$ is singular. For these problems we expect ill-conditioning in our modified problems. Problems 7. and 20. are severely ill-conditioned as well.

To compare the results with the structure preserving Hamiltonian Schur algorithms, [15, 32], we also display in Figure 2 the relative invariant subspace residual

$$\frac{\|(I - UU^T)HU\|_F}{\|H\|_F},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. We point out that, unlike SDA and gSDA, the Hamiltonian

Figure 2: Invariant subspace residual for the examples in carex



Schur algorithms are not affected by the sign change in the Hamiltonian, and produce exactly the same invariant subspaces, apart from the small errors caused by the eigenvalue-swapping subroutines after the reduction to Schur form. Therefore, we may use the figures in [15] in our comparison: for the default values, the Hamiltonian Schur algorithm achieves a relative residual of 10^{-15} or better in all cases. This accuracy is not reached in all cases with the generalized structured doubling algorithm gSDA. There may be several reasons for this, but the major difficulty is that the Cayley transform introduces errors in the data to which the method is applied, while the Schur methods work directly on the Hamiltonian data. This is particularly bad when the problem has eigenvalues near the shift-point of the Cayley transform.

5.1 Convergence history

To provide some insight on the reason for the potential instability remaining in both methods, it is interesting to study the convergence history of the two methods on some of the problems where the block $V^{(2)}$ is singular. In Tables 1 and 2 we display the norms of K_k for both methods at each iteration. For gSDA, the norm of N_k is displayed as well (for the SDA we have $K_k = N_k^T$, thus the two norms coincide). For the SDA, the column labeled cond. displays $\max(\mathrm{cond}(I - M_k L_k), \mathrm{cond}(I - L_k M_k))$; for gSDA, it displays $\max(\mathrm{cond}(R_{\mathcal{L}_k^T}), \mathrm{cond}(R_{\mathcal{M}_k^T}))$; i.e.,

in both cases the maximum condition number among any of the matrices to invert during the iteration.

For Problem 2., numerical errors in the SDA do not alter the singularity of the $V^{(2)}$ block, thus the algorithm diverges with a behavior similar to that of (41): $\operatorname{null} L_k$ contains a basis for the unstable subspace at each step, thus the repeated doubling takes K_k to ∞ instead of to 0. In the gSDA, a similar behavior appears until in step 5 the inversion of an ill-conditioned matrix completely destroys the initial data. The method eventually converges, but to a solution bearing little resemblance to the correct one. Similar behavior arises in Problem 6., where in both methods the growth of K_k and N_k leads to ill-conditioning in an intermediate matrix and complete loss of accuracy. For Problems 15. and 17., the growth in the gSDA is much milder than the one in the SDA, thus the gSDA is able to obtain a meaningful result.

6 Conclusions

The proposed generalization of the structured doubling algorithm represents a compromise between row orthogonality of the blocks in the iterates, as in the inverse-free disc method, and symplectic structure preservation and numerical efficiency, as in the customary structured doubling algorithm.

The generalized framework allows to construct different variants, based on different choices of the normalization factors $S_k^{(1)}$ and $S_k^{(2)}$.

Many of the problems included in the benchmark set carex are problematic for most algorithms to solve Riccati equations. Although our generalization of the structured doubling algorithm cannot handle all of them in a fully satisfactory way, it obtains much better accuracy on a large subset of them. In some cases, a temporary growth of the matrices K_k and N_k during the intermediate steps is detrimental for the accuracy. Though this algorithm cannot be considered a definitive choice for all small- to medium-size control problems, it represents an improvement over the customary structured doubling algorithm and suggests new generalizations and research directions.

7 Acknowledgments

The algorithm exposed in this paper was developed while the second author, F. Poloni, was visiting the Numerical Mathematics group of the Technische Universität in Berlin. The hospitality of the research group and of TU Berlin and the support of Scuola Normale Superiore for the visit are gratefully acknowledged.

A Proof of Theorem 10

Proof. Suppose that there are s Jordan blocks relative to unimodular eigenvalues, each with eigenvalue ω_j of multiplicity $2r_j$, $j = 1, \dots, s$. We partition them as

$$J_{\omega_j, 2r_j} = \begin{bmatrix} J_{\omega_j, r_j} & \Gamma^{(j)} \\ 0 & J_{\omega_j, r_j} \end{bmatrix},$$

where J_{ω, r_j} is a Jordan block with the same eigenvalue and half the algebraic multiplicity, and $\Gamma^{(j)}$ is the $r_j \times r_j$ matrix having 1 in the lower-left corner and 0 elsewhere. Constructing the matrices

$$J_c = \bigoplus_{j=1}^{s} J_{\omega_j, r_j}, \quad \Gamma = \bigoplus_{j=1}^{s} \Gamma^{(j)};$$

it is clear that

$$J_1 := \begin{bmatrix} J_c & \Gamma \\ 0 & J_c \end{bmatrix}$$

is similar to $\bigoplus_{j=1}^{s} J_{\omega_j,2r_j}$, since they differ only by the same row and column exchanges. Let Γ_k be defined so that $J_1^{2^k}$ is partitioned as

$$J_1^{2^k} = \begin{bmatrix} J_c^{2^k} & \Gamma_k \\ 0 & J_c^{2^k} \end{bmatrix}.$$

It is proved in [25, Lemma 4.4] that Γ_k is nonsingular for sufficiently large k, and

$$\Gamma_k^{-1} J_c^{2^k} = O(2^{-k}), \quad J_c^{2^k} \Gamma_k^{-1} J_c^{2^k} = O(2^{-k}). \tag{44}$$

If we start from the Kronecker canonical from of $\lambda\mathcal{E} - \mathcal{A}$, then after some row exchanges we get to

$$\mathcal{W}(\lambda\mathcal{E} - \mathcal{A})\mathcal{Z} = \lambda \begin{bmatrix} I & 0 \\ 0 & E_u \oplus I \end{bmatrix} - \begin{bmatrix} J_s \oplus J_c & 0 \oplus \Gamma \\ 0 & J_u \oplus J_c \end{bmatrix} =: \lambda\mathcal{K}_1 - \mathcal{K}_2,$$

where the matrices are chosen such that $\lambda I - J_s$ is the Kronecker structure relative to the $n - r$ d-stable eigenvalues, $\lambda E_u - J_u$ is the Kronecker structure relative to the $\ell - n - r$ d-unstable eigenvalues (including eigenvalues at infinity if needed). Note that the canonical semi-stable subspace of $\lambda\mathcal{E} - \mathcal{A}$ is spanned by

$$\begin{bmatrix} Z_{11} \\ Z_{21} \end{bmatrix}, \text{ where } \mathcal{Z} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix},$$

and therefore Z_{11} is nonsingular if and only if $U^{(1)}$ is.

Since the two matrices $\mathcal{K}_1$ and $\mathcal{K}_2$ commute, it follows by induction as in [14, Section 2.3] that

$$\mathcal{A}_k \mathcal{Z} \mathcal{K}_1^{2^k} = \mathcal{E}_k \mathcal{Z} \mathcal{K}_2^{2^k}. \tag{45}$$

Note that

$$\mathcal{K}_1^{2^k} = \begin{bmatrix} I & 0 \\ 0 & E_u^{2^k} \oplus I \end{bmatrix}, \ \mathcal{K}_2^{2^k} = \begin{bmatrix} J_s^{2^k} \oplus J_c^{2^k} & 0 \oplus \Gamma_k \\ 0 & J_u^{2^k} \oplus J_c^{2^k} \end{bmatrix}.$$

Multiplying both terms of (45) by $\begin{bmatrix} I_n & 0_{n\times m} \end{bmatrix}$ on the left and

$$\begin{bmatrix} I_n \\ 0 \oplus -\Gamma_k^{-1} J_c^{2^k} \end{bmatrix} \tag{46}$$

on the right, we get

$$K_k \left(Z_{11} - Z_{12}(0 \oplus \Gamma_k^{-1} J_c^{2^k}) \right) = \mathcal{M}_k \mathcal{Z} \begin{bmatrix} J_s^{2^k} \oplus 0 \\ 0 \oplus -J_c^{2^k} \Gamma_k^{-1} J_c^{2^k} \end{bmatrix}.$$

Since $\mathcal{M}_k$ is bounded and $J_s^{2^k}$ decreases exponentially with k, using (44) we see that the right-hand side is $O(2^{-k})$. Hence, if Z_{11} (or, equivalently, $U^{(1)}$) is nonsingular, then we obtain $K_k = O(2^{-k})$.

The eigenvalues of the pencil $\lambda\mathcal{A} - \mathcal{E}$ are the inverses of those of $\lambda\mathcal{E} - \mathcal{A}$, with the same partial multiplicities; thus its Kronecker canonical form can be written in the form

$$\widehat{\mathcal{W}}(\lambda\mathcal{A} - \mathcal{E})\widehat{\mathcal{Z}} = \lambda \begin{bmatrix} I & 0 \\ 0 & \widehat{E}_u \oplus I \end{bmatrix} - \begin{bmatrix} \widehat{J}_s \oplus \widehat{J}_c & 0 \oplus \widehat{\Gamma} \\ 0 & \widehat{J}_u \oplus \widehat{J}_c \end{bmatrix} =: \lambda\widehat{\mathcal{K}_1} - \widehat{\mathcal{K}_2},$$

where $\lambda I - \widehat{J}_s$ is stable and its eigenvalues are the inverses of the d-unstable eigenvalues of $\lambda\mathcal{E} - \mathcal{A}$ (including $\infty^{-1} = 0$ if needed), with the same partial multiplicity.

Note that the canonical d-semi-unstable subspace of $\lambda\mathcal{E} - \mathcal{A}$ is spanned by

$$\begin{bmatrix} \widehat{Z}_{11} \\ \widehat{Z}_{21} \end{bmatrix}, \text{where } \widehat{\mathcal{Z}} = \begin{bmatrix} \widehat{Z}_{11} & \widehat{Z}_{12} \\ \widehat{Z}_{21} & \widehat{Z}_{22} \end{bmatrix},$$

and hence $\widehat{Z}_{21}$ is nonsingular if and only if $V^{(2)}$ is. Proceeding in the same way, we obtain

$$\mathcal{E}_k \widehat{\mathcal{Z}} \widehat{\mathcal{K}}_1^{2^k} = \mathcal{A}_k \widehat{\mathcal{Z}} \widehat{\mathcal{K}}_2^{2^k}, \tag{47}$$

and multiplying by $\begin{bmatrix} 0_{\ell-n\times n} & I_{\ell-n} \end{bmatrix}$ on the left and the matrix in (46) on the right, we get

$$N_k \left(\widehat{Z}_{21} - \widehat{Z}_{22} (0 \oplus \widehat{\Gamma}_k^{-1} \widehat{J}_c^{2^k}) \right) = \mathcal{L}_k \widehat{\mathcal{Z}} \begin{bmatrix} \widehat{J}_s^{2^k} \oplus 0 \\ 0 \oplus -\widehat{J}_c^{2^k} \widehat{\Gamma}_k^{-1} \widehat{J}_c^{2^k} \end{bmatrix},$$

and thus the nonsingularity of $V^{(2)}$ implies $N_k = O(2^{-k})$.

Multiplying (45) by $\begin{bmatrix} 0_{\ell-n\times n} & I_{\ell-n} \end{bmatrix}$ on the left and the matrix in (46) on the right, we get

$$\mathcal{L}_k \begin{bmatrix} Z_{11} \\ Z_{21} \end{bmatrix} - \mathcal{H}_k \begin{bmatrix} Z_{12} \\ Z_{22} \end{bmatrix} (0 \oplus \Gamma_k^{-1} J_c^{2^k}) = N_k \mathcal{Z} \begin{bmatrix} J_s^{2^k} \oplus 0 \\ 0 \oplus -J_c^{2^k} \Gamma_k^{-1} J_c^{2^k} \end{bmatrix}, \tag{48}$$

and thus if $N_k = O(2^k)$, then all terms apart from the first are $O(2^k)$ and it follows that

$$\mathcal{L}_k \begin{bmatrix} Z_{11} \\ Z_{21} \end{bmatrix} = O(2^k).$$

Similarly, multiplying (47) by $\begin{bmatrix} I_n & 0_{n\times\ell-n} \end{bmatrix}$ and (46) yields

$$\mathcal{M}_k \begin{bmatrix} \widehat{Z}_{11} \\ \widehat{Z}_{21} \end{bmatrix} = O(2^k).$$

If there are no d-critical eigenvalues, then the second terms in all the expressions containing a $\oplus$ sign vanish, and one sees that the convergence speed is determined only by $\rho(J_s^{2^k}) = |\lambda_s|^{2^k}$ and $\rho(\widehat{J}_s^{2^k}) = |\lambda_u|^{-2^k}$. □

References

[1] B.D.O. Anderson. Second-order convergent algorithms for the steady-state Riccati equation. *Internat. J. Control*, 28(2):295–306, 1978.

[2] M. Athans and P.L. Falb. *Optimal Control*. McGraw-Hill, New York, 1966.

[3] Z. Bai, J. Demmel, and M. Gu. An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems. *Numer. Math.*, 76(3):279–308, 1997.

[4] P. Benner. *Contributions to the Numerical Solution of Algebraic Riccati Equations and Related Eigenvalue Problems*. Doctoral dissertation, Fakultät für Mathematik, TU Chemnitz-Zwickau, Chemnitz, 1997. Published by Logos-Verlag, Berlin.

[5] P. Benner and R. Byers. An arithmetic for matrix pencils: theory and new algorithms. *Numer. Math.*, 103(4):539–573, 2006.

[6] P. Benner, R. Byers, V. Mehrmann, and H. Xu. Numerical computation of deflating subspaces of skew Hamiltonian/Hamiltonian pencils. *SIAM J. Matrix Anal. Appl.*, 24:165–190, 2002.

[7] P. Benner, R. Byers, V. Mehrmann, and H. Xu. A robust numerical method for the γ-iteration in h_∞-control. *Linear Algebra Appl.*, 425:548–570, 2007.

[8] P. Benner, A. Laub, and V. Mehrmann. A collection of benchmark examples for the numerical solution of algebraic Riccati equations I: the continuous-time case. Technical Report SPC 95-22, Forschergruppe 'Scientific Parallel Computing', Fakultät für Mathematik, TU Chemnitz-Zwickau, 1995.

[9] P. Benner, A. Laub, and V. Mehrmann. A collection of benchmark examples for the numerical solution of algebraic Riccati equations II: the discrete-time case. Technical Report SPC 95-23, Forschergruppe 'Scientific Parallel Computing', Fakultät für Mathematik, TU Chemnitz-Zwickau, 1995.

[10] P. Benner, V. Mehrmann, and H. Xu. A new method for computing the stable invariant subspace of a real Hamiltonian matrix. *J. Comput. Appl. Math.*, 86(1):17–43, 1997.

[11] R. Byers. Solving the algebraic Riccati equation with the matrix sign function. *Linear Algebra Appl.*, 85:267–279, 1987.

[12] R. Byers, T. Geerts, and V. Mehrmann. Descriptor systems without controllability at infinity. *SIAM J. Cont.*, 35:462–479, 1997.

[13] R. Byers, D.S. Mackey, V. Mehrmann, and H. Xu. Symplectic, BVD, and palindromic approaches to discrete-time control problems. In P. Petkov and N. Christov, editors, *Collection of Papers Dedicated to the 60-th Anniversary of Mihail Konstantinov*, pages 81–102. Rodina, Sofia, Bulgaria, 2009.

[14] C.-Y. Chiang, E.K.-W. Chu, C.-H. Guo, T.-M. Huang, W.-W. Lin, and S.-F. Xu. Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case. *SIAM J. Matrix Anal. Appl.*, 31(2):227–247, 2009.

[15] D. Chu, X. Liu, and V. Mehrmann. A numerical method for computing the Hamiltonian Schur form. *Numer. Math.*, 105(3):375–412, 2007.

[16] 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.*, 396:55–80, 2005.

[17] 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(8):767–788, 2004.

[18] D.J. Clements and K. Glover. Spectral factorization via Hermitian pencils. *Linear Algebra Appl.*, 122-124:797–846, 1989.

[19] H. Fassbender. *Symplectic methods for the symplectic eigenproblem.* Kluwer Academic/Plenum Publishers, New York, 2000.

[20] B.A. Francis and J.C. Doyle. Linear control theory with an H_∞ optimality criterion. *SIAM J. Control Optim.*, 25(4):815–844, 1987.

[21] G. Freiling, V. Mehrmann, and H. Xu. Existence, uniqueness and parametrization of lagrangian invariant subspaces. *SIAM J. Matrix Anal. Appl.*, 23:1045–1069, 2002.

[22] I. Gohberg, P. Lancaster, and L. Rodman. *Invariant subspaces of matrices with applications*, volume 51. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[23] G.H. Golub and C.F. Van Loan. *Matrix computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[24] L. Hogben, editor. *Handbook of linear algebra.* Discrete Mathematics and its Applications (Boca Raton). Chapman & Hall/CRC, Boca Raton, FL, 2007.

[25] T.-M. Huang and W.-W. Lin. Structured doubling algorithms for weakly stabilizing Hermitian solutions of algebraic Riccati equations. *Linear Algebra Appl.*, 430(5-6):1452–1478, 2009.

[26] M. Kimura. Convergence of the doubling algorithm for the discrete-time algebraic Riccati equation. *Internat. J. Systems Sci.*, 19(5):701–711, 1988.

[27] P. Kunkel and V. Mehrmann. Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index. *Math. Control, Signals, Sys.*, 20:227–269, 2008.

[28] P. Lancaster and L. Rodman. *Algebraic Riccati equations*. Oxford University Press, Oxford, 1995.

[29] 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(1):26–39 (electronic), 2006.

[30] D.S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Structured polynomial eigenvalue problems: Good vibrations from good linearizations. *SIAM J. Matrix Anal. Appl.*, 28(4):1029–1051, 2006.

[31] V. Mehrmann. A step towards a unified treatment of continuous and discrete time control problems. *Linear Algebra Appl.*, 241–243:749–779, 1996.

[32] V. Mehrmann, C. Schröder, and D. S. Watkins. A new block method for computing the Hamiltonian Schur form. *Linear Algebra Appl.*, 431(3-4):350–368, 2009.

[33] V.L. Mehrmann. *The autonomous linear quadratic control problem*, volume 163 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Berlin, 1991. Theory and numerical solution.

[34] T. Pappas, A.J. Laub, and N.R. Sandell. On the numerical solution of the discrete-time algebraic Riccati equation. *IEEE Trans. Automat. Control*, AC-25:631–641, 1980.

[35] P.H. Petkov, N.D. Christov, and M.M. Konstantinov. *Computational Methods for Linear Control Systems*. Prentice-Hall, Hertfordshire, UK, 1991.

[36] L.S. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishenko. *The Mathematical Theory of Optimal Processes*. Interscience, New York, 1962.

[37] C. Schröder. *Palindromic and Even Eigenvalue Problems - Analysis and Numerical Methods*. Phd thesis, Technical University Berlin, Germany, 2008.

[38] V. Sima. *Algorithms for Linear-Quadratic Optimization*, volume 200 of *Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, NY, 1996.

[39] X. Sun and E.S. Quintana-Ortí. Spectral division methods for block generalized Schur decompositions. *Math. Comp.*, 73(248):1827–1847 (electronic), 2004.

[40] J. H. Wilkinson. *The algebraic eigenvalue problem*. Oxford University Press, Oxford, UK, 1988.

[41] H. Xu. On equivalence of pencils from discrete-time and continuous-time control. *Linear Algebra Appl.*, 414(1):97–124, 2006.

[42] H. Xu. Transformations between discrete-time and continuous-time algebraic Riccati equations. *Linear Algebra Appl.*, 425(1):77–101, 2007.

[43] D.C. Youla. On the factorization of rational matrices. *IRE Trans. Inform. Theory*, 7, 1961.

[44] K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, Upper Saddle River, NJ, 1995.

Table 1: Convergence history for the SDA (left) and the gSDA (right)

(a) carex Problem 2.

k	$\|K_k\|$	cond.
1	4.57e+01	2.20e+00
2	4.13e+02	2.20e+00
3	3.35e+04	2.20e+00
4	2.19e+08	1.30e+00
5	8.49e+15	1.60e+16
6	4.54e+01	Inf

k	$\|K_k\|$	$\|N_k\|$	cond.
1	3.57e+01	1.22e+01	5.47e+00
2	3.22e+02	1.10e+02	1.00e+00
3	2.61e+04	8.88e+03	1.00e+00
4	8.93e+07	4.28e+07	1.94e+00
5	4.33e+00	3.38e+00	9.68e+14
6	1.63e+00	1.86e-15	1.62e+01
7	3.99e-01	3.09e-46	1.00e+00
8	2.38e-02	3.78e-107	1.00e+00
9	8.51e-05	2.23e-228	1.00e+00
10	1.09e-09	0.00e+00	1.00e+00
11	1.77e-19	0.00e+00	1.00e+00
12	4.70e-39	0.00e+00	1.00e+00
13	3.31e-78	0.00e+00	1.00e+00
14	1.64e-156	0.00e+00	1.00e+00
15	4.05e-313	0.00e+00	1.00e+00
16	0.00e+00	0.00e+00	1.00e+00

(b) carex Problem 6.

k	$\|K_k\|$	cond.
1	4.25e+02	2.29e+05
2	1.82e+03	3.12e+06
3	8.83e+04	3.02e+10
4	2.54e+10	3.49e+20
5	3.00e+20	3.61e+43
6	5.16e+16	1.24e+49
7	5.59e+13	1.25e+47
8	4.29e+13	1.90e+44
9	1.50e+09	2.65e+41
10	2.70e-01	1.21e+40
11	9.48e-22	1.21e+40
12	1.17e-62	1.21e+40
13	1.77e-144	1.21e+40
14	4.07e-308	1.21e+40
15	0.00e+00	1.21e+40

k	$\|K_k\|$	$\|N_k\|$	cond.
1	5.15e+01	5.25e+01	9.00e+03
2	7.51e+01	3.32e+02	1.39e+04
3	1.13e+03	7.93e+04	4.70e+06
4	5.34e+02	4.71e+09	1.38e+15
5	1.42e+03	8.84e+16	2.24e+31
6	1.25e+03	2.48e+17	3.08e+31
7	8.39e+04	4.08e+17	1.34e+34
8	2.56e+05	1.24e+15	6.36e+32
9	3.38e+04	2.03e+16	6.91e+30
10	1.09e+03	2.33e+17	5.68e+31
11	9.08e-01	5.68e+18	1.42e+31
12	6.80e-07	6.81e+21	4.28e+22
13	3.87e-19	9.96e+27	2.55e+14
14	1.25e-43	2.14e+40	1.79e+02
15	1.30e-92	9.81e+64	1.00e+00
16	1.42e-190	2.07e+114	1.00e+00
17	0.00e+00	9.23e+212	1.00e+00

Table 2: Convergence history for the SDA (left) and the gSDA (right)

(a) carex Problem 15.

k	$\|K_k\|$	cond.	k	$\|K_k\|$	$\|N_k\|$	cond.
1	1.04e+01	1.03e+02	1	1.51e+00	9.00e+00	4.00e+01
2	8.10e+01	8.65e+01	2	8.32e-01	8.10e+01	8.40e+01
3	6.56e+03	1.42e+02	3	1.05e-01	6.56e+03	6.68e+03
4	4.33e+07	1.15e+13	4	6.62e-04	4.33e+07	4.33e+07
5	2.03e+02	3.02e+19	5	1.35e-08	7.37e+01	1.65e+02
6	6.96e-11	1.61e+18	6	3.84e-18	1.86e-16	1.00e+00
7	1.53e-27	1.61e+18	7	2.86e-37	1.39e-35	1.00e+00
8	2.81e-60	1.61e+18	8	1.58e-75	7.68e-74	1.00e+00
9	7.72e-126	1.61e+18	9	4.80e-152	2.34e-150	1.00e+00
10	5.33e-257	1.61e+18	10	4.45e-305	2.18e-303	1.00e+00
11	0.00e+00	1.61e+18	11	0.00e+00	0.00e+00	1.00e+00

(b) carex Problem 17.

k	$\|K_k\|$	cond.	k	$\|K_k\|$	$\|N_k\|$	cond.
1	1.78e+02	4.42e+02	1	7.86e+01	7.86e+01	2.10e+01
2	6.17e+03	1.54e+06	2	8.42e+02	8.42e+02	1.19e+03
3	7.09e+05	8.14e+10	3	8.93e+03	8.93e+03	8.60e+05
4	7.97e+07	8.11e+15	4	2.31e+04	2.31e+04	4.15e+08
5	5.28e+08	1.19e+19	5	8.77e+03	8.77e+03	2.38e+09
6	4.70e+07	9.40e+18	6	8.01e+02	8.01e+02	8.37e+07
7	8.28e+04	1.25e+19	7	6.65e+00	6.65e+00	9.10e+03
8	2.51e-01	3.72e+19	8	4.58e-04	4.58e-04	1.01e+00
9	3.32e-12	3.72e+19	9	2.17e-12	2.17e-12	1.00e+00
10	5.37e-34	3.72e+19	10	4.88e-29	4.88e-29	1.00e+00
11	1.41e-77	3.72e+19	11	2.47e-62	2.47e-62	1.00e+00
12	9.79e-165	3.72e+19	12	6.33e-129	6.33e-129	1.00e+00
13	0.00e+00	3.72e+19	13	4.15e-262	4.15e-262	1.00e+00
			14	0.00e+00	0.00e+00	1.00e+00