Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices

by

Lars Grasedyck, Wolfgang Hackbusch, and Boris N. Khoromskij

Preprint no.: 62 2002

Solution of Large Scale Algebraic Matrix Riccati Equations by Use of Hierarchical Matrices

L. Grasedyck, W. Hackbusch, B.N. Khoromskij, Leipzig

Abstract

In previous papers, a class of hierarchical matrices ($\mathcal{H}$-matrices) is introduced which are data-sparse and allow an approximate matrix arithmetic of almost optimal complexity. Here, we investigate a new approach to exploit the $\mathcal{H}$-matrix structure for the solution of large scale Lyapunov and Riccati equations as they typically arise for optimal control problems where the constraint is a partial differential equation of elliptic type. This approach leads to an algorithm of linear-logarithmic complexity in the size of the matrices.

AMS Subject Classification: *65F05, 65F30, 65F50*
Key words: *Hierarchical matrices, data-sparse approximations, formatted matrix operations, fast solvers, Lyapunov equations, Riccati equations, control problems*

1 Introduction

1.1 Overview

In 1980, Roberts [17] published a method to solve the algebraic matrix Riccati equation by use of the matrix sign function. Since the method basically involves the inversion, addition and multiplication of matrices, one expects a cubic complexity in the size of the matrices.

In this paper we consider the same method but make use of a special matrix representation, the so-called $\mathcal{H}$-matrices, instead of the standard matrix representation. Our analysis consists of two parts:

1. We prove that the solution of the algebraic matrix Riccati equation can be approximated in the $\mathcal{H}$-matrix format. This existence result indicates the possibility to apply the $\mathcal{H}$-matrix arithmetic. Moreover, we prove that the matrices in Roberts method can be approximated by matrices in $\mathcal{H}$-matrix representation.
2. We develop an efficient numerical scheme to compute an $\mathcal{H}$-matrix approximation to the solution of the algebraic matrix Riccati equation with almost linear complexity in the size of the matrices, i.e., $\mathcal{O}(n \log^q n)$ for $n \times n$ matrices.

This article contains six sections: the current section gives a short overview. The second section introduces a linear quadratic optimal control problem leading to a Riccati equation. The solution procedure based on the matrix sign function is introduced in Section 3. In Section 4 we investigate the structure of the matrices appearing in the solution procedure and observe that $\mathcal{H}$-matrices are a good choice for an (approximate) representation of the matrices. The influence of the approximation error in the numerical scheme will be analysed in Section 5 while the numerical results in the last section show the behaviour of our method applied to two model problems.

1.2 Lyapunov and Riccati Equation

An equation of the form

$$A^T X + XA - XFX + G = 0$$

for given $A, G, F \in \mathbb{R}^{n \times n}$ and unknown $X \in \mathbb{R}^{n \times n}$ is called *(algebraic matrix) Riccati equation.* For $F = 0$ the equation simplifies to a so-called *Lyapunov equation.* The standard approach to solve a Riccati equation is to apply Newton's method resulting in a series of Lyapunov equations.

1.3 Large scale Lyapunov Equations

A fixed Lyapunov equation can, e.g., be solved by the Bartels-Stewart algorithm [2], which is of complexity $\mathcal{O}(n^3)$. When dealing with large scale Lyapunov or Riccati equations (i.e., n is considerably large) one is interested in reducing the complexity for a certain class of matrices A, F, G.

Rosen and Wang [18] assume that the matrix A stems from the discretisation of a partial differential equation of elliptic type, while G is allowed to be arbitrary. Then it is possible to apply multigrid techniques and solve the Lyapunov equation with $\mathcal{O}(n^2)$ operations.

Penzl [15] assumes that A is symmetric positive definite and that G is a symmetric positive semidefinite matrix of low rank $k_G = \mathcal{O}(1)$. Then the eigenvalues of the solution X decay exponentially, such that the solution can be approximated by a matrix of low rank. The low rank structure can be utilised (e.g., in the Smith method [16]) to compute the solution with $\mathcal{O}(n)$ operations. However, one has to solve sparse linear systems of equations in each step.

Let us consider a simple example: the matrix $A \in \mathbb{R}^{n \times n}$ is assumed to be the symmetric stiffness matrix from the Ritz-Galerkin discretisation (e.g., linear finite elements) of a partial differential operator of elliptic type. Then the solution X to $A^T X + XA + I = 0$ (I is the identity) is $X = -\frac{1}{2}A^{-1}$. Here, the matrix $G = I$ is not of low rank, but the solution X can still be represented in a suitable format that is explained in the following section.

1.4 $\mathcal{H}$-Matrices

In previous papers ([6],[7],[9],[10],[11]) a class of hierarchical matrices ($\mathcal{H}$-matrices) has been introduced that allows a sparse approximation to large, dense stiffness matrices arising in boundary element method or finite element method applications. In the FEM case, it is the inverse of the stiffness matrix that is dense and can be approximated by an $\mathcal{H}$-matrix ([3]).

We consider matrices over the (product) index set $I \times J$. The product index set $I \times J$ is partitioned into blocks $r \times s \subset I \times J$, where the blocks $r \times s$ are nodes of a so-called *$\mathcal{H}$-tree* $T_{I \times J}$. Each of those blocks allows for a low rank representation of the corresponding matrix block. The maximal rank of the matrix blocks is denoted by k.

The definition of the $\mathcal{H}$-tree $T_{I \times J}$ and the set $\mathcal{M}_{\mathcal{H},k}(T_{I \times J})$ of $\mathcal{H}$-matrices can be found in [6] and [7]. Here, it suffices to know that the leaves of the $\mathcal{H}$-tree $T_{I \times J}$ form a partition of $I \times J$ and a matrix M belongs to $\mathcal{M}_{\mathcal{H},k}(T_{I \times J})$ if the rank of M restricted to a leaf of $T_{I \times J}$ is bounded by k.

Since $\mathcal{H}$-matrices of fixed (block-wise) rank k corresponding to the $\mathcal{H}$-tree $T_{I \times J}$ are not a linear subspace of $\mathbb{R}^{I \times J}$, some kind of projection of the sum, product and inverse into the set of $\mathcal{H}$-matrices is necessary. For the (exact) sum of $\mathcal{H}$-matrices one can calculate a best approximation (in the Frobenius norm $\|M\|_F^2 = \sum_{i,j} M_{ij}^2$) in $\mathcal{M}_{\mathcal{H},k}(T_{I \times J})$. This is called the *formatted* addition ($\oplus$). For the product and inverse of $\mathcal{H}$-matrices the *formatted* multiplication $\odot$ and inversion $\widetilde{Inv}$ (introduced in [9],[6],[7]) is some approximation but not necessarily a best approximation.

Associated to the $\mathcal{H}$-tree $T_{I \times J}$ are

- the depth $p > 1$ of the $\mathcal{H}$-tree,
- the sparsity constant C_{sp} and
- the idempotency constant C_{id}.

The depth p of the $\mathcal{H}$-tree is typically proportional to $\log(|I| + |J|)$ ($|I|$ denotes the number of elements in I). The constants C_{sp} and C_{id} are defined and estimated in [6] and [7] (a constant similar to C_{sp} was also used in [11]). Based upon these three values one can estimate the complexity of the standard arithmetic operations for $\mathcal{H}$-matrices.

Theorem 1.1 *Let $k \in \mathbb{N}$ denote the blockwise rank, $n := |I|$, $m := |J|$ and $T_{I\times J}$ be an $\mathcal{H}$-tree with sparsity constant C_{sp} and depth $p > 1$. Then the storage requirements $N_{\mathcal{H},store}$ and computational complexity $N_{\mathcal{H}\cdot v}$ of the matrix vector multiplication and $N_{\mathcal{H}\oplus\mathcal{H}}$ of the formatted addition for matrices belonging to $\mathcal{M}_{\mathcal{H},k}(T_{I\times J})$ are bounded by*

$$\begin{aligned} N_{\mathcal{H},store} &\le C_{\mathrm{sp}} k(n+m)p, \\ N_{\mathcal{H}\cdot v} &\le 2C_{\mathrm{sp}} k(n+m)p, \\ N_{\mathcal{H}\oplus\mathcal{H}} &\le 20 C_{\mathrm{sp}} k^2 (n+m)p + 368 C_{\mathrm{sp}} k^3 (n+m). \end{aligned}$$

Let $T_{I\times J}$ be an $\mathcal{H}$-tree with idempotency constant C_{id}. Then the computational complexity of the formatted multiplication $N_{\mathcal{H}\odot\mathcal{H}}$ and the formatted inversion $N_{\widetilde{Inv}(\mathcal{H})}$ of matrices belonging to $\mathcal{M}_{\mathcal{H},k}(T_{I\times J})$ ($n = m$ for the inversion) can be estimated by

$$\begin{aligned} N_{\mathcal{H}\odot\mathcal{H}} &\le 394 C_{\mathrm{sp}}^2 C_{\mathrm{id}} k^2 (n+m) p \max\{k, p\}, \\ N_{\widetilde{Inv}(\mathcal{H})} &\le N_{\mathcal{H}\odot\mathcal{H}}. \end{aligned}$$

Proof. [9], [6] and [7]. ■

Remark 1.2 *A matrix $M \in \mathbb{R}^{n\times m}$ of rank at most k can be represented in factorised form $M = AB^T$ with matrices $A \in \mathbb{R}^{n\times k}$ and $B \in \mathbb{R}^{m\times k}$. We call this an $\mathrm{R}(k)$-representation of the matrix M. For $k \ll n, m$, this is an efficient way to store and evaluate the matrix M.*

The (exact) multiplication of an $\mathrm{R}(k)$-matrix with an arbitrary matrix yields again an $\mathrm{R}(k)$-matrix. The (exact) multiplication of two $\mathcal{H}$-matrices belonging to $\mathcal{M}_{\mathcal{H},k}(T_{I\times J})$ as in Theorem 1.1 yields an $\mathcal{H}$-matrix belonging to $\mathcal{M}_{\mathcal{H},\tilde{k}}(T_{I\times J})$ with $\tilde{k} := C_{\mathrm{sp}} C_{\mathrm{id}} kp$. The computational cost for the exact multiplication is $\mathcal{O}(k^2(n+m)p^2)$ (see [6], [7]).

2 Problem Description

2.1 The Autonomous Linear Quadratic Optimal Control Problem

Let $n, n_y, n_u \in \mathbb{N}$, $x_0 \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n_u}$ and $C \in \mathbb{R}^{n_y\times n}$. The autonomous linear quadratic optimal control problem is to find $u \in L_2(0,\infty;\mathbb{R}^{n_u})$ minimising the quadratic performance index

$$J(u) := \int_0^\infty \left(y(t)^T y(t) + u(t)^T u(t) \right) \mathrm{d}t \qquad (2.1)$$

for the solution $x \in L_2(0,\infty;\mathbb{R}^n)$ of the differential equation

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \qquad t \in (0,\infty), \\ y(t) &= Cx(t), \\ x(0) &= x_0. \end{aligned}$$

Theorem 2.1 ([13] Existence of a linear state feedback solution) *If (A, B) is stabilisable and (A, C) detectable, that is there exist matrices $K_{stab} \in \mathbb{R}^{n_u \times n}$ and $K_{det} \in \mathbb{R}^{n_y \times n}$ such that $A + BK_{stab}$ and $A^T + C^T K_{det}$ are stability matrices (spectrum in the left complex halfplane), then the optimal control u exists and can be realised in linear state feedback form as*

$$u(t) = -B^T X x(t), \quad t \in (0, \infty),$$

where $X \in \mathbb{R}^{n \times n}$ is the (in the set of symmetric positive semidefinite matrices) unique solution of the algebraic matrix Riccati equation

$$A^T X + XA - XFX + G = 0$$

for the matrices $F := BB^T$ and $G := C^T C$.

In general, the structure of the matrix A can be arbitrary and the stability and detectability is neither easy to check nor always given. In the applications that we are aiming at, the matrix A will be the spatial discretisation of some partial differential operator of elliptic type. Therefore, A is a stability matrix and thus the system stabilisable and detectable.

2.2 The Algebraic Matrix Riccati Equation

According to Theorem 2.1, we seek a symmetric positive semidefinite solution $X \in \mathbb{R}^{n \times n}$ of the algebraic matrix Riccati equation

$$A^T X + XA - XFX + G = 0. \tag{2.2}$$

Here, $A \in \mathbb{R}^{n \times n}$ is a stability matrix and $F, G \in \mathbb{R}^{n \times n}$ are symmetric positive semidefinite. The rank of the matrices $F = BB^T$ and $G = C^T C$ is bounded by n_u and n_y, where $B^T \in \mathbb{R}^{n_u \times n}$ and $C \in \mathbb{R}^{n_y \times n}$.

Remark 2.2 *If A stems from the discretisation of some partial differential operator, then the dimension n grows with decreasing mesh size. The ranks n_u, n_y of the matrices F and G on the other hand can be independent of the discretisation. In that case one can assume $n_u, n_y \ll n$, which will lead to an (approximate) low rank representation of the solution X. Our method can exploit this low rank structure, but is not restricted to this case.*

Remark 2.3 *Let $A = M^{-1}\hat{A}$ and $F = M^{-1}\hat{F}M^{-1}$ with a symmetric positive definite matrix M, a symmetric negative definite matrix $\hat{A}$ and a symmetric positive semidefinite matrix $\hat{F}$ (as it typically occurs for finite element discretisations, see Section 6). Then the algebraic matrix Riccati equation (2.2) reads*

$$\hat{A}M^{-1}X + XM^{-1}\hat{A} - XM^{-1}\hat{F}M^{-1}X + G = 0. \tag{2.3}$$

If we multiply by $W := M^{-\frac{1}{2}}$ from the left and right of (2.3), the equation transforms into

$$W\hat{A}W\, WXW + WXW\, W\hat{A}W - WXW\, W\hat{F}W\, WXW + WGW = 0. \tag{2.4}$$

If we define $\tilde{A} := W\hat{A}W$, $\tilde{F} := W\hat{F}W$ and $\tilde{G} := WGW$ then we can solve the transformed algebraic matrix Riccati equation

$$\tilde{A}A\tilde{X} + \tilde{X}\tilde{A} - \tilde{X}\tilde{F}\tilde{X} + \tilde{G} = 0$$

where all matrices $\tilde{A}, \tilde{F}, \tilde{G}$ are symmetric, and gain the solution X by $X := M^{\frac{1}{2}}\tilde{X}M^{\frac{1}{2}}$. The eventual low rank structure of the matrices F, G is preserved and the symmetry of $\tilde{A}$ can be beneficial. Since the calculation of $M^{\frac{1}{2}}$ and $M^{-\frac{1}{2}}$ is rather expensive, the transformation to the symmetric case is usually omitted.

3 Solution Strategy

There is a variety of strategies for solving algebraic matrix Riccati equations for matrices of a certain structure. Basically, one can either try to solve the (nonlinear) equation (2.2) directly, or one can apply Newton's method to simplify the equation to a linear one. The latter results in a series of Lyapunov equations and is almost always the method of choice for solving sparse large scale Riccati equations. The method that we propose is essentially based on the sign-iteration due to Roberts [17]. It can be applied to the Lyapunov as well as the Riccati equation and is neither limited to some low rank structure of F and G nor does depend on the sparsity of A (but we adopt the data-sparsity of A). However, the analysis in this paper is only done for the case that F is of low rank, which corresponds to a low dimensional control u. The motivation for our particular choice of the solution process is to minimise the total number of matrix inversions which we consider as a suitable complexity unit in the overall cost estimate.

3.1 Newton's Method Applied to the Riccati Equation

The Newton iteration

$$X_{i+1} \quad \text{solves} \quad (A - FX_i)^T X_{i+1} + X_{i+1}(A - FX_i) + X_i F X_i + G = 0 \tag{3.1}$$

converges (locally) quadratically to the solution X of the Riccati equation (2.2), if F and G are symmetric and the initial guess X_0 stabilises $(A, -F)$ (see, e.g., [12]). This is the case if A is a stability matrix and the initial guess is chosen as $X_0 := 0$. In this case, the solution to the Lyapunov equation (3.1) is explicitly given as (see [14])

$$X_{i+1} = \int_0^\infty \exp(t(A - FX_i)^T)(X_i F X_i + G)\exp(t(A - FX_i))\mathrm{d}t. \tag{3.2}$$

Moreover, all matrices $A - FX_i$ are again stability matrices but typically not symmetric, while $X_i F X_i + G$ is symmetric positive semidefinite.

3.2 Solving the Riccati or Lyapunov Equation by Use of the Matrix Sign Function

Definition 3.1 (Matrix sign function) *We define the matrix sign function as*

$$\operatorname{sign} : \{M \in \mathbb{C}^{n\times n} \mid \forall \lambda \in \sigma(M) : \Re e(\lambda) \neq 0\} \to \mathbb{C}^{n\times n}, \qquad M \mapsto \frac{1}{\pi i}\oint_\Gamma (\xi I - M)^{-1}\mathrm{d}\xi - I,$$

where Γ is an arbitrary path of index 1 around the eigenvalues of M with positive real part and I denotes the $n \times n$ identity matrix.

Example 3.2 (Matrix sign of a diagonalisable matrix) *Let $M \in \mathbb{C}^{n\times n}$ be a matrix that is diagonalisable, $M = TDT^{-1}$, $T \in \mathbb{R}^{n\times n}$, $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and $\Re e(\lambda_j) \neq 0$ for all $j \in \{1, \dots, n\}$. Let Γ be a path of index 1 around the eigenvalues of M with positive real part. For each of the eigenvalues λ_j there holds*

$$\frac{1}{\pi i}\oint_\Gamma (\xi - \lambda_j)^{-1}\mathrm{d}\xi - 1 = \begin{cases} 2 - 1 & = & 1 & \text{if } \Re e(\lambda_j) > 0, \\ 0 - 1 & = & -1 & \text{if } \Re e(\lambda_j) < 0. \end{cases}$$

We can compute $\operatorname{sign}(M)$ *as follows:*

$$\begin{aligned} \operatorname{sign}(M) &= \operatorname{sign}(TDT^{-1}) \\ &= \frac{1}{\pi i}\oint_\Gamma (\xi I - TDT^{-1})^{-1}\mathrm{d}\xi - I \quad = \quad \frac{1}{\pi i}\oint_\Gamma T(\xi I - D)^{-1}T^{-1}\mathrm{d}\xi - I \\ &= T\operatorname{diag}(s_1, \dots, s_n)\, T^{-1}, \quad s_j = \begin{cases} 1 & \text{if } \Re e(\lambda_j) > 0, \\ -1 & \text{if } \Re e(\lambda_j) < 0. \end{cases} \end{aligned}$$

An algorithm to solve certain Riccati equations by use of the matrix sign function is presented in [17] and we summarise the main result in

Theorem 3.3 (Representation by the matrix sign function) *Let $A \in \mathbb{R}^{n\times n}$ be a stability matrix, $F, G \in \mathbb{R}^{n\times n}$ symmetric positive semidefinite. Then the stabilising solution X of (2.2) satisfies*

$$\begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix} X = - \begin{bmatrix} N_{12} \\ N_{22} \end{bmatrix}, \tag{3.3}$$

where the matrices $N_{11}, N_{12}, N_{21}, N_{22} \in \mathbb{R}^{n\times n}$ are

$$\begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} := \operatorname{sign}\left(\begin{bmatrix} A^T & G \\ F & -A \end{bmatrix} \right) - \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

and $\begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix}$ is of full rank n. In the Lyapunov case $F = 0$, this simplifies to

$$X = \frac{1}{2} N_{12}.$$

A method to solve (3.3) efficiently is presented in the last part of Section 5. A simple method to calculate the sign function of a matrix S is Newton's method applied to the equation $X^2 = I$ with initial guess $X_0 := S$, as it is described in [17].

Theorem 3.4 (Newton's method to calculate the matrix sign function) *Let $S \in \mathbb{R}^{n_s \times n_s}$ be a matrix whose spectrum does not intersect the imaginary axis. Then the iteration*

$$S_0 := S, \qquad S_{i+1} := \frac{1}{2}(S_i + S_i^{-1}) \tag{3.4}$$

converges (locally quadratically) to the sign of S.

Lemma 3.5 (Global convergence of Newton's method) *Let $S \in \mathbb{R}^{n_s \times n_s}$ be a matrix whose spectrum $\sigma(S)$ does not intersect the imaginary axis. Let $\|\cdot\|_2$ denote the spectral norm of a matrix and*

$$\mu := \max \left\{ 1 + |\Re e(\lambda)| + |\Re e(\lambda)|^{-1} + \frac{|\Im m(\lambda)|}{|\Re e(\lambda)|} \;\middle|\; \lambda \in \sigma(S) \right\}.$$

Then the minimal number of iterations i_{min} of (3.4) necessary to get

$$\forall \lambda \in \sigma(S_{i_{min}}): \qquad |\lambda^2 - 1| \le \varepsilon \tag{3.5}$$

for a given $\varepsilon \in (0,1)$ is bounded by $i_{min} = \mathcal{O}(\log(\mu)^2 + \log(\log(1/\varepsilon)))$. If $S = TJT^{-1}$ is a Jordan decomposition of S then the minimal number of iterations j_{min} of (3.4) necessary to get

$$\|S_{j_{min}} - \operatorname{sign}(S)\|_2 \le \varepsilon$$

is bounded by

$$j_{min} = \mathcal{O}\left(\log(\mu)^2 + \log(\log(1/\varepsilon + \operatorname{cond}(T))) \right).$$

If the spectral values $\lambda \in \sigma(S)$ fulfil $|\Re e(\lambda)| \ge |\Im m(\lambda)|$ then the number of iterations j_{min} necessary to get $\|S_{j_{min}} - \operatorname{sign}(S)\|_2 \le \varepsilon$ is bounded by $\mathcal{O}(\log(\rho) + \log(\log(1/\varepsilon + \rho)))$, where $\rho := \max_{\lambda \in \sigma(S)}(|\lambda| + |\lambda^{-1}|)$.

Proof. Equation (3.5)
To prove (3.5) , we analyse the convergence for each $\lambda_0 = x_0 + iy_0 \in \sigma(S)$ separately. The corresponding spectral values for the operators S_j are defined by the sequence

$$\begin{aligned} \lambda_{j+1} &:= \frac{1}{2}(\lambda_j + \lambda_j^{-1}), \\ x_{j+1} &:= \Re e(\lambda_{j+1}) = \frac{1}{2}x_j(1 + \frac{1}{x_j^2 + y_j^2}), \\ y_{j+1} &:= \Im m(\lambda_{j+1}) = \frac{1}{2}y_j(1 - \frac{1}{x_j^2 + y_j^2}). \end{aligned}$$

Throughout the proof we will make use of the following basic facts:

$$|y_j/x_j| \leq |y_{j-1}/x_{j-1}| \qquad (j \in \mathbb{N}), \tag{3.6}$$

$$\frac{1}{8} \leq x_j^2 + y_j^2 \leq 8 \Rightarrow |y_{j+1}/x_{j+1}| \leq \frac{7}{9}|y_j/x_j|, \tag{3.7}$$

$$x_j^2 + y_j^2 \geq 1/2 \Rightarrow |y_{j+1}| \leq \frac{1}{2}|y_j|. \tag{3.8}$$

We distinguish between three phases of convergence:

Phase 1 (Convergence towards equilibrium)
The first phase $P_1 = \{0, \ldots, j_1\}$ is defined by the condition $x_j^2 < y_j^2$ for all $j \in P_1$ and $x_{j_1+1}^2 \geq y_{j_1+1}^2$. From (3.6) we conclude that $x_{j_1+j}^2 \geq y_{j_1+j}^2$ for all $j \in \mathbb{N}$. We want to prove that after $\mathcal{O}(\log(\mu)^2)$ iterations we leave Phase 1 and x_{j_1+1} is bounded from above and below.

Start of Phase 1

Case 1: $x_0^2 + y_0^2 > 8$. As long as $x_j^2 + y_j^2 > 8$ the iterates decrease by a factor of 1/3:

$$\begin{aligned} |x_{j+1}| &\leq \frac{1}{2}|x_j|(1 + 1/8) \leq \frac{9}{16}|x_j|, \\ |y_{j+1}| &= \frac{1}{2}|y_j|(1 - 1/(x_j^2 + y_j^2)) \;\; \leq \;\; \frac{1}{2}|y_j|, \\ x_{j+1}^2 + y_{j+1}^2 &\leq \frac{81}{256}x_j^2 + \frac{1}{4}y_j^2 \;\; \leq \;\; \frac{1}{3}(x_j^2 + y_j^2). \end{aligned} \tag{3.9}$$

Since $x_{j+1}^2 + y_{j+1}^2$ is also bounded from below (by 4/3) we come to Case 3 after $\log(x_0^2 + y_0^2)$ iterations.

Case 2: $x_0^2 + y_0^2 < \frac{1}{8}$. Then $|y_1| = \frac{1}{2}|y_0(1 - 1/(x_0^2 + y_0^2))| \geq \frac{1}{4}|y_0|^{-1} \geq 1/2$, therefore we come to Case 1 or Case 3 in the first iteration. x_1 is bounded from below and above by

$$\frac{1}{2}(|x_0| + |x_0|^{-1}) \;\geq\; |x_1| = \frac{1}{2}|x_0|(1 + \frac{1}{x_0^2 + y_0^2}) \;\geq\; \frac{1}{2}|x_0|,$$

while $|y_1| \leq \frac{|y_0|}{|x_0|}|x_1| \leq \mu|x_0|^{-1}$. If $x_1^2 + y_1^2 > 8$ (Case 1) then we will need $\mathcal{O}(\log(\mu))$ iterations to get to Case 3.

Case 3: $x_0^2 + y_0^2 \in [\frac{1}{8}, 8]$.

Core of Phase 1. Without loss of generality, we assume $x_0^2 + y_0^2 \in [\frac{1}{8}, 8]$ (at most $\mathcal{O}(\log(\mu))$ iterations are necessary to ensure this). Whenever the iterates $x_j^2 + y_j^2$ are contained in $[\frac{1}{8}, 8]$ the ratio $|y_{j+1}/x_{j+1}|$ decreases by $\frac{7}{9}$ (see (3.7)). After $\mathcal{O}(\log(\mu))$ of these steps we will therefore leave Phase 1. At last we show that every $\mathcal{O}(\log(\mu))$ iterations one of the iterates is contained in $[\frac{1}{8}, 8]$.

Let $x_j^2 + y_j^2 \in [\frac{1}{8}, 8]$ and $x_{j+1}^2 + y_{j+1}^2 \notin [\frac{1}{8}, 8]$. We prove that after at most $j' = \mathcal{O}(\log(\mu))$ steps we either get back to $x_{j+j'}^2 + y_{j+j'}^2 \in [\frac{1}{8}, 8]$ or we leave Phase 1. In the latter case we bound $x_{j+j'}^2$ from below and above.

Case 1: $x_{j+1}^2 + y_{j+1}^2 > 8$. Since

$$\begin{aligned} y_{j+1}^2 &= \frac{1}{4} y_j^2 (1 - 1/(x_j^2 + y_j^2))^2 \leq \frac{1}{4} 8 \cdot 49 \leq 98, \\ x_{j+1}^2 &= \frac{1}{4} x_j^2 (1 + 1/(x_j^2 + y_j^2))^2 \leq \frac{1}{4} 8 \cdot 81 \leq 162, \end{aligned}$$

we will either leave Phase 1 with $|x_{j+j'}| \in (2, 13)$ or we will arrive after 4 steps (see (3.9)) at $x_{j+j'}^2 + y_{j+j'}^2 \in [\frac{1}{8}, 8]$.

Case 2: $x_{j+1}^2 + y_{j+1}^2 < \frac{1}{8}$. As in the start of Phase 1, Case 2, we get either $y_{j+2}^2 \geq \frac{1}{4}$ or $x_{j+2}^2 \geq \frac{1}{4}$. We will now bound x_{j+1} from below and x_{j+2}, y_{j+2} from above.

$$\begin{aligned} |x_{j+1}| &\geq \frac{1}{2}|x_j| \geq \frac{1}{2}\frac{|x_0|}{|y_0|}|y_j| \geq \frac{1}{8}\mu^{-1}, \\ |x_{j+2}| &\leq \frac{1}{2}(|x_{j+1}| + |x_{j+1}|^{-1}|) \leq 5\mu, \\ |y_{j+2}| &\leq \mu |x_{j+2}| \leq 5\mu^2. \end{aligned}$$

After at most $\mathcal{O}(\log(\mu))$ iterations one of the iterates $j + j'$ fulfils $x_{j+j'}^2 + y_{j+j'}^2 \in [\frac{1}{8}, 8]$.

The number of iterations necessary to leave Phase 1 is $\mathcal{O}(\log(\mu)^2)$.

Phase 2 (Linear convergence)

From Phase 1 we have $x_{j_1+1}^2 \geq y_{j_1+1}^2$ and $|x_{j_1+1}| \in [\frac{1}{8}\mu^{-1}, 5\mu]$. It follows $|x_{j_1+2}| \in [\frac{1}{2}, 5\mu]$ and after $\mathcal{O}(\log(\mu))$ steps we get $|x_{j_2'}| \in [\frac{1}{2}, 2]$. After another $\mathcal{O}(1)$ steps (3.8) yields $|y_{j_2}| < \frac{1}{8}$ and $||x_{j_2}| - 1| < \frac{1}{8}$. After $\mathcal{O}(\log(\mu))$ iterations we leave Phase 2.

Phase 3 (Quadratic convergence)

Phase 3 is defined by the condition $|y_{j_2}| < \frac{1}{8}$ and $||x_{j_2}| - 1| < \frac{1}{8}$. We prove

$$\begin{aligned} ||x_{j+1}| - 1| &\leq 2 \max\{||x_j| - 1|, |y_j|\}^2, \\ |y_{j+1}| &\leq 2 \max\{||x_j| - 1|, |y_j|\}^2 \end{aligned}$$

for all $j \geq j_2$. Let $q := \max\{||x_j| - 1|, |y_j|\}$. Then

$$\begin{aligned} ||x_{j+1}| - 1| &= |\frac{1}{2}|x_j|(1 + \frac{1}{x_j^2 + y_j^2}) - 1| \\ &= |\frac{1}{2}|x_j| - 1 + \frac{1}{2}|x_j|^{-1} \ + \frac{1}{2}\frac{|x_j|}{x_j^2 + y_j^2} - \frac{1}{2}|x_j|^{-1}| \\ &\leq |\frac{1}{2}|x_j| - 1 + \frac{1}{2}\sum_{\nu=0}^{\infty}(1 - |x_j|)^{\nu}| + \frac{1}{2}\frac{y_j^2}{|x_j|^3 + |x_j| y_j^2} \\ &\leq \frac{1}{2}(1 - |x_j|)^2 \frac{1}{|x_j|} + \frac{3}{4} y_j^2 \quad \leq \quad 2q^2, \\ |y_{j+1}| &= \frac{1}{2}|y_j||1 - \frac{1}{x_j^2 + y_j^2}| = \frac{1}{2}\frac{||y_j| x_j^2 + |y_j|^3 - |y_j||}{x_j^2 + y_j^2} \\ &\leq \frac{4}{7}|y_j^3| + \frac{4}{7}(|x_j^2 - 1||y_j|) \quad \leq \quad \frac{1}{14}|y_j|^2 + \frac{4}{7}||x_j| - 1||x_j + 1||y_j| \\ &\leq \frac{1}{14}|y_j^2 + \frac{60}{49}||x_j| - 1||y_j| \quad \leq \quad 2q^2. \end{aligned}$$

Consequently we get $\max\{||x_{j_2+j}| - 1|, |y_{j_2+j}|\} \leq 2^{2^j} 8^{-2^j} = 4^{-2^j}$, which ensures $|\lambda_{j_{min}}^2 - 1| \leq \varepsilon$ for $j_{min} = \mathcal{O}(\log(\mu)^2 + \log(\log(1/\varepsilon)))$.

Let $S = TJT^{-1}$ be a Jordan decomposition of S, where

$$J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_l \end{bmatrix}, \qquad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix},$$

consists of l Jordan blocks J_i. We define the series $J^{(0)} := J, J^{(j+1)} := \frac{1}{2}(J^{(j)} + (J^{(j)})^{-1})$ which preserves the block-diagonal structure. We fix a single Jordan block J_i with corresponding eigenvalue λ and define the sequence

$$\lambda_{j+1} := \frac{1}{2}(\lambda_j + \lambda_j^{-1}), \quad \lambda_0 := \lambda.$$

We define the upper-diagonal structure (of the same size as J)

$$U(\alpha, \beta) := \begin{bmatrix} \alpha & \beta & & \\ & \ddots & \ddots & \\ & & \ddots & \beta \\ & & & \alpha \end{bmatrix}.$$

Let $J^{(j)} = U(\lambda_j, \beta_j)$. Then it holds

$$J^{(j+1)} = U(\lambda_{j+1}, \beta_{j+1}), \qquad \beta_{j+1} = \frac{1}{2}\beta_j(1 - \lambda_j^{-2})$$

such that $|\beta_{j+1}| \leq \max\{1, |\lambda_j^{-2}|\}|\beta_j|$ and

$$|\beta_j| \leq \eta^j, \quad \eta := \max_{i=1,\ldots,j}(1 + |\lambda_i^{-2}|).$$

As in the first part of the proof (core of Phase 1, Case 2) the norm of λ_j^{-1} is bounded by 8μ such that $|\beta_j| \leq (8\mu)^j$ (this is only a rough estimate). After i_{min} steps of iteration (3.4) we get

$$S_{i_{min}} = T \begin{bmatrix} J_1^{(i_{min})} & & \\ & \ddots & \\ & & J_l^{(i_{min})} \end{bmatrix} T^{-1}, \qquad J_i^{(i_{min})} = U(\lambda_{i_{min}}, \beta_{i_{min}}),$$

with $|\beta_{i_{min}}| \leq (8\mu)^{i_{min}}$. From (3.5) it follows

$$|\beta_{i_{min}+1}| \leq 2\varepsilon|\beta_{i_{min}}|,$$

such that after i_{min} steps of the iteration we get

$$|\beta_{2i_{min}}| \leq (16\varepsilon\mu)^{i_{min}}.$$

Taking $\tilde{\varepsilon} := \varepsilon/(16\mu\,\mathrm{cond}_2(T))$ instead of ε yields

$$\|S_{2i_{min}} - \mathrm{sign}(S)\|_2 \leq \varepsilon.$$

■

4 Structure of the Matrices Involved

4.1 Sylvester Equation

Before we formulate the theorems in detail, we first outline the basic idea. Let $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{m\times m}$, $G \in \mathbb{C}^{m\times n}$. The spectra $\sigma(A), \sigma(B)$ of A, B are assumed to be contained in the sets (see Figure 1)

$$\sigma(A) \subset S_A := \{x + iy \in \mathbb{C} \mid \lambda_{A,1} < x < \lambda_{A,2}, |y| \leq \mu\} \tag{4.1}$$

$$\sigma(B) \subset S_B := \{x + iy \in \mathbb{C} \mid \lambda_{B,1} < x < \lambda_{B,2}, |y| \leq \mu\}, \tag{4.2}$$

where

$$\lambda_{A,2} + \lambda_{B,2} < -3. \tag{4.3}$$

The assumption (4.3) is only needed to simplify the basic idea, later we will only need $\lambda_{A,2} + \lambda_{B,2} < 0$. In order to express the matrix exponentials $\exp(tA)$, $\exp(tB)$ by the Dunford-Cauchy representation

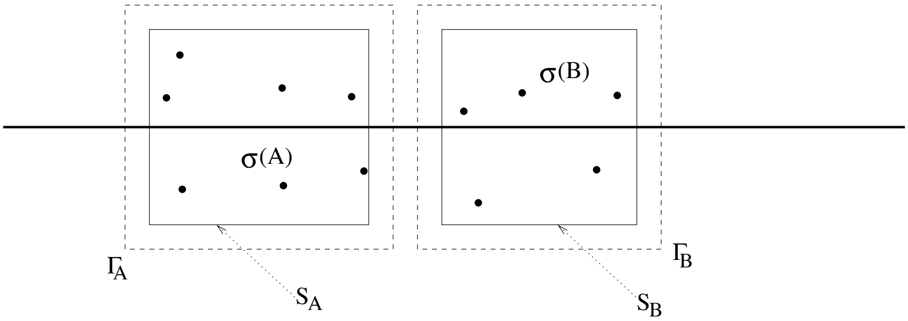

Figure 1: The spectrum $\sigma(A)$ of A is contained in S_A, that of B in S_B. The sets Γ_A, Γ_B have a distance of at least 1 to S_A, S_B.

$$\exp(tA) = \frac{1}{2\pi i}\oint_{\Gamma_A} \exp(\xi t)(\xi I - A)^{-1}\mathrm{d}\xi,$$

$$\exp(tB) = \frac{1}{2\pi i}\oint_{\Gamma_B} \exp(\eta t)(\eta I - B)^{-1}\mathrm{d}\eta,$$

we define closed paths around S_A, S_B:

$$\Gamma_A := \Big\{a + ib \Big| \quad (a \in [\lambda_{A,1} - 1, \lambda_{A,2} + 1] \wedge b \in \{-\mu - 1, \mu + 1\}) \\ \vee \ (a \in \{\lambda_{A,1} - 1, \lambda_{A,2} + 1\} \wedge b \in [-\mu - 1, \mu + 1]) \Big\},$$

$$\Gamma_B := \Big\{a + ib \Big| \quad (a \in [\lambda_{B,1} - 1, \lambda_{B,2} + 1] \wedge b \in \{-\mu - 1, \mu + 1\}) \\ \vee \ (a \in \{\lambda_{B,1} - 1, \lambda_{B,2} + 1\} \wedge b \in [-\mu - 1, \mu + 1]) \Big\}.$$

The paths are chosen such that $\mathrm{dist}(\Gamma_A, \sigma(A)) \geq 1$ and $\mathrm{dist}(\Gamma_B, \sigma(B)) \geq 1$. From (4.3) we conclude that the unique solution $X \in \mathbb{C}^{m\times n}$ to the Sylvester equation

$$BX + XA + G = 0 \tag{4.4}$$

is (cf. [14])

$$X = \int_0^\infty \exp(tB)G\exp(tA)\mathrm{d}t. \tag{4.5}$$

Insertion of the Dunford-Cauchy representation yields

$$X = -\frac{1}{4\pi^2}\oint_{\Gamma_A}\oint_{\Gamma_B}(\xi I - A)^{-1}G(\eta I - B)^{-1}\int_0^\infty \exp(t(\xi+\eta))\mathrm{d}t\mathrm{d}\xi\mathrm{d}\eta. \tag{4.6}$$

If we replace $\int_0^\infty \exp(t(\xi+\eta))\mathrm{d}t$ by a suitable quadrature formula $\sum_{j=-k}^{k}\omega_j \exp(t_j(\xi+\eta))$ (with t_j, ω_j independent of $\xi+\eta$), then the modified solution reads

$$\begin{aligned}\tilde{X} &:= -\frac{1}{4\pi^2}\oint_{\Gamma_A}\oint_{\Gamma_B}(\xi I - A)^{-1}G(\eta I - B)^{-1}\sum_{j=-k}^{k}\omega_j \exp(t_j(\xi+\eta))\mathrm{d}\xi\mathrm{d}\eta \qquad (4.7)\\ &= \sum_{j=-k}^{k}\omega_j \exp(t_j B)G\exp(t_j A).\end{aligned}$$

The error $\|X - \tilde{X}\|$ is estimated in the following theorem preceded by two auxiliary lemmata.

Lemma 4.1 *Let $M \in \mathbb{C}^{n\times n}$, $z \in \mathbb{C}$ with $\operatorname{dist}(z, \sigma(M)) \geq 1$.*

1. *If M is symmetric, then $\|(zI - M)^{-1}\|_2 \leq 1$.*
2. *If $M = TDT^{-1}$, $D = \operatorname{diag}(d_1, \ldots, d_n)$, then $\|(zI - M)^{-1}\|_2 \leq \operatorname{cond}_2(T)$.*

Proof. Let $M = TDT^{-1}$, $D = \operatorname{diag}(d_1, \ldots, d_n)$. Then

$$\|(zI - M)^{-1}\|_2 = \|T(zI - D)^{-1}T^{-1}\|_2 \leq \operatorname{cond}_2(T)\|(zI - D)^{-1}\|_2 \leq \operatorname{cond}_2(T).$$

If M is symmetric, then $\operatorname{cond}_2(T) = 1$. ■

Lemma 4.2 (Stenger) *Let $z \in \mathbb{C}$ with $\Re e(z) \leq -1$. Then for each $k \in \mathbb{N}$ the points and weights*

$$t_j := \log\left(\exp(jk^{-1/2}) + \sqrt{1 + \exp(2jk^{-1/2})}\right), \tag{4.8}$$

$$\omega_j := (k + k\exp(-2jk^{-1/2}))^{-1/2}, \qquad j = -k, \ldots, k, \tag{4.9}$$

fulfil

$$\left|\int_0^\infty \exp(tz)\mathrm{d}t - \sum_{j=-k}^{k}\omega_j \exp(t_j z)\right| \leq C_{\text{sinc}}\exp(|\Im m(z)|/\pi)\exp(-\sqrt{k}), \tag{4.10}$$

where the constant C_{sinc} does not depend upon k, z.

Proof. The function $t \mapsto \exp(tz)$ is holomorphic in $\mathbb{C}$ and satisfies [20, (4.2.59)] with $C_2 = 1, \alpha = 1, \beta = 1$. The points t_j are the z_k from [20, Example 4.2.11] and the ω_j are the weights in [20, (4.2.60)] (with $d := \pi^{-1}$, $h = k^{-1/2}$ and $n = N = M = k$). Applying [20, Example 4.2.11] yields the estimate

$$\left|\int_0^\infty \exp(tz)\mathrm{d}t - \sum_{j=-k}^{k}\omega_j \exp(t_j z)\right| \leq C_3\exp(-\sqrt{k})$$

with a constant C_3 depending upon z. Finally, we estimate the constant C_3. As in [20, Example 4.2.11] we define

$$\phi(z) := \log(\sinh(z)), \qquad \phi'(z) = \frac{1}{\tanh(z)},$$

such that ϕ is a conformal map of $\mathcal{D}_d := \{z \in \mathbb{C} : |\arg(\sinh(z))| < d\}$ (the analyticity domain for the integrand) onto the strip $D_d := \{z \in \mathbb{C} : |\Im m(z)| < d\}$. For the function $\mathcal{G}(t,z) := \exp(tz)$ satisfying the conditions of [20, Example 4.2.11] the quadrature rule of Stenger can be analysed by use of the splitting

$$\left| \int_0^\infty \mathcal{G}(t,z)\mathrm{d}t - h \sum_{j=-k}^{k} \frac{\mathcal{G}(t_j,z)}{\phi'(t_j)} \right| \leq \left| \int_0^\infty \mathcal{G}(t,z)\mathrm{d}t - h \sum_{j=-\infty}^{\infty} \frac{\mathcal{G}(t_j,z)}{\phi'(t_j)} \right| + \left| h \sum_{j=-\infty}^{-k-1} \frac{\mathcal{G}(t_j,z)}{\phi'(t_j)} \right| + \left| h \sum_{j=k+1}^{\infty} \frac{\mathcal{G}(t_j,z)}{\phi'(t_j)} \right|.$$

Due to [20], each term in the above error estimate can be bounded by

$$\frac{1}{3} C_{\mathrm{sinc}} \sup_{\Re e(\xi)>0, |\Im m(\xi)| \leq d} |\mathcal{G}(\xi,z)| \exp(-\sqrt{k}),$$

where the constant C_{sinc} does not depend upon k, z. From the estimate

$$|\mathcal{G}(\xi,z)| = \exp(\Re e(z)\Re e(\xi) - \Im m(\xi)\Im m(z)) \overset{\Re e(z)\leq 0, \Re e(\xi)>0}{\leq} \exp(|\Im m(z)|/\pi)$$

we obtain $C_3 \leq C_{\mathrm{sinc}} \exp(|\Im m(z)|/\pi)$. ■

Theorem 4.3 (Representation of the solution to the Sylvester equation) *Let* $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{m\times m}$, $G \in \mathbb{C}^{n\times m}$. *The spectra* $\sigma(A), \sigma(B)$ *of* A, B *are assumed to be contained in the sets* S_A, S_B *defined in* (4.1) , (4.2) , *where we assume* $\lambda_{A,2} + \lambda_{B,2} < 0$. *We define*

$$\begin{aligned} \alpha &:= 3|\lambda_{A,2} + \lambda_{B,2}|^{-1}, \\ c_A &:= \max\{\|(zI - \alpha A)^{-1}\|_2 \mid z \in \mathbb{C}, \mathrm{dist}(z, \sigma(\alpha A)) \geq 1\}, \\ c_B &:= \max\{\|(zI - \alpha B)^{-1}\|_2 \mid z \in \mathbb{C}, \mathrm{dist}(z, \sigma(\alpha B)) \geq 1\}. \end{aligned}$$

For each $k \in \mathbb{N}$ *and* $j \in \{-k, \dots, k\}$ *let* $t_j \in [0, 1 + \sqrt{k}]$ *and* $\omega_j \in (0,1)$ *denote the points and weights from* (4.8) , (4.9) . *If* X *is the unique solution to* (4.4) , *then the matrix*

$$\tilde{X} := \sum_{j=-k}^{k} \alpha \omega_j \exp(\alpha t_j B) G \exp(\alpha t_j A)$$

and the constant

$$C(A,B) := c_A c_B C_{\mathrm{sinc}} \exp(\frac{2 + 2\alpha\mu}{\pi}) \frac{\alpha}{4\pi^2} (8 + \alpha(\mu + \lambda_{A,2} - \lambda_{A,1}))(8 + \alpha(\mu + \lambda_{B,2} - \lambda_{B,1}))$$

fulfil

$$\|X - \tilde{X}\|_2 \leq C(A,B) \|G\|_2 \exp(-\sqrt{k}).$$

Proof. If X is a solution to (4.4) then X satisfies

$$(\alpha B)X + X(\alpha A) + \alpha G = 0.$$

The spectra of $\alpha A, \alpha B$ fulfil $\alpha\lambda_{A,2} + \alpha\lambda_{B,2} \leq -3$ and we can define the sets Γ_A, Γ_B as

$$\begin{aligned} \Gamma_A := \Big\{ a + ib \Big| \quad & (a \in [\alpha\lambda_{A,1} - 1, \alpha\lambda_{A,2} + 1] \wedge b \in \{-\alpha\mu - 1, \alpha\mu + 1\}) \\ & \vee \ (a \in \{\alpha\lambda_{A,1} - 1, \alpha\lambda_{A,2} + 1\} \wedge b \in [-\alpha\mu - 1, \alpha\mu + 1]) \Big\}, \\ \Gamma_B := \Big\{ a + ib \Big| \quad & (a \in [\alpha\lambda_{B,1} - 1, \alpha\lambda_{B,2} + 1] \wedge b \in \{-\alpha\mu - 1, \alpha\mu + 1\}) \\ & \vee \ (a \in \{\alpha\lambda_{B,1} - 1, \alpha\lambda_{B,2} + 1\} \wedge b \in [-\alpha\mu - 1, \alpha\mu + 1]) \Big\} \end{aligned}$$

such that $\mathrm{dist}(\Gamma_A, \sigma(\alpha A)) \geq 1$, $\mathrm{dist}(\Gamma_B, \sigma(\alpha B)) \geq 1$, $\Re e(\xi + \eta) \leq -1$ for $\xi \in \Gamma_A, \eta \in \Gamma_B$, and $\sigma(\alpha A) \subset \Gamma_A$, $\sigma(\alpha B) \subset \Gamma_B$. Let $k \in \mathbb{N}$ and t_j, ω_j be the points and weights from Lemma 4.2. We estimate the approximation error by

$$\begin{aligned}
\|X - \tilde{X}\|_2 \overset{(4.6),(4.7)}{=} & \Bigg\| - \frac{1}{4\pi^2} \oint_{\Gamma_A} \oint_{\Gamma_B} (\xi I - \alpha A)^{-1} \alpha G (\eta I - \alpha B)^{-1} \\
& \left(\int_0^\infty \exp(t(\xi + \eta)) \mathrm{d}t - \sum_{j=1}^{k} \omega_j \exp(t_j(\xi + \eta)) \right) \mathrm{d}\xi \mathrm{d}\eta \Bigg\|_2 \\
& \overset{|\Im m(\xi+\eta)| \leq 2+2\alpha\mu}{\leq} \frac{1}{4\pi^2} \oint_{\Gamma_A} \oint_{\Gamma_B} c_A \alpha \|G\|_2 c_B C_{\mathrm{sinc}} \exp(\frac{2 + 2\alpha\mu}{\pi}) \exp(-\sqrt{k}) \mathrm{d}\xi \mathrm{d}\eta \\
& \leq C(A, B) \|G\|_2 \exp\left(-\sqrt{k}\right).
\end{aligned}$$

■

Corollary 4.4 (R(k)-approximation to the solution of the Sylvester equation) *We use the same notation as in Theorem 4.3. Let k_G denote the rank of G. Then the minimal rank k_X needed to approximate the solution X to* (4.4) *up to an error of $\|\tilde{X} - X\|_2 \leq \varepsilon$, $\varepsilon \in (0, 1)$, by an* $\mathrm{R}(k_X)$*-matrix $\tilde{X}$ is bounded by*

$$k_X \leq k_G \log(C(A, B)\|G\|_2 \varepsilon^{-1})^2. \tag{4.11}$$

Since $\|G\|_2 = \|AX - XB\|_2 \leq (\|A\|_2 + \|B\|_2)\|X\|_2$ we also get the estimate for the relative error

$$\|\tilde{X} - X\|_2 \leq \varepsilon \|X\|_2$$

with X of rank $k_X \leq k_G \log(C(A, B)(\|A\|_2 + \|B\|_2)\varepsilon^{-1})^2$.

Remark 4.5 *If we consider only the dependency on ε in* (4.11) *then we have $k_X = \mathcal{O}(\log(1/\varepsilon)^2)$, while in [8] the estimate $k_X = \mathcal{O}(\log(1/\varepsilon))$ is established. However, the desired accuracy ε, the size of the matrices and the spectrum of the matrices is typically not independent. If we assume that the desired accuracy ε is of the size* $\log(1/\varepsilon) = \mathcal{O}(q)$*, the size of the matrices n and m is* $\log(n + m) = \mathcal{O}(q)$*, the norm of the matrices is* $\log(\|A\| + \|B\|) = \mathcal{O}(q)$*, the distance λ between the spectrum of A and that of $-B$ is bounded by* $\log(1/\lambda) = \mathcal{O}(q)$ *and the maximum of the imaginary part of the eigenvalues of A and B is bounded by $\mathcal{O}(\lambda)$, then the estimate* (4.11) *and the one from [8] read*

$$k_X = \mathcal{O}(q^2),$$

which coincides with the estimate from Penzl [15] for the symmetric Lyapunov case.

Lemma 4.6 (Approximation of the operator exponential) *Let $\mu \in \mathbb{R}_{\geq 0}$ and $A \in \mathbb{C}^{n \times n}$ with spectrum $\sigma(A) \subset \{x + iy \in \mathbb{C} \mid x \geq 2$ and $|y| \leq \mu\}$. For the parabola*

$$\Gamma_A := \{\frac{1}{2}(\mu + 1)^{-2}\eta^2 + \frac{1}{2} + i\eta \mid \eta \in (-\infty, \infty)\} \tag{4.12}$$

and the interior $\Omega_A := \{\xi + i\eta \mid \eta \in (-\infty, \infty)$ and $\xi > \frac{1}{2}(\mu + 1)^{-2}\eta^2 + \frac{1}{2}\}$ we define the so-called strong P-positivity constant

$$M := \sup_{z \in \mathbb{C} \setminus \Omega_A} \|(zI - A)^{-1}\|_2 (1 + \sqrt{|z|}).$$

Then the matrix exponential $\exp(-A)$ *can be approximated by a linear combination of resolvents, i.e., for each $k_{\mathrm{E}} \in \mathbb{N}$ there exist points $z_j \in \mathbb{C} \setminus \Omega_A$ and weights $w_j \in \mathbb{C}$ such that*

$$\| \exp(-A) - \sum_{j=-k_{\mathrm{E}}}^{k_{\mathrm{E}}} w_j (z_j I - A)^{-1} \|_2 \leq M \exp\left(4(\mu + 1)^2 - (\mu + 1)^{2/3} k_{\mathrm{E}}^{2/3}\right). \tag{4.13}$$

Proof. We want to apply [5, Theorem 2.4]. The integration parabola is defined by

$$\Gamma_b = \{\frac{a}{4}\eta^2 + b + i\eta \mid \eta \in (-\infty, \infty)\},$$

where $a := \frac{1}{2}(\mu+1)^{-2}$ and $b := 2 - \frac{3}{2}(\mu+1)^2$. In the following we further estimate the expression appearing in [5, (2.12)]. We choose the parameter $k := 4$ and get

$$\begin{aligned} b(k) &= 2 - (k-1)/(4a) = b, \\ d &= (1 - \frac{1}{\sqrt{k}})\frac{k}{2a} = 2(\mu+1)^2, \\ s &= ((2\pi d)^2 a/k)^{1/3} = (2\pi^2(\mu+1)^2)^{1/3}, \\ c &= M_1 \exp(d^2 a/k + d - b) = M_1 \exp((\mu+1)^2/2 + 2(\mu+1)^2 - 2 + \frac{3}{2}(\mu+1)^2) \\ &= M_1 \exp(4(\mu+1)^2/2 - 2). \end{aligned}$$

To estimate the constant

$$M_1 = \sup_{z \in \mathbb{C}, |\Im m(z)| \le 2(\mu+1)^2} \frac{|2az/k - i|}{1 + \sqrt{|az^2/k + b - iz|}}, \tag{4.14}$$

we distinguish between two cases for $z = x + iy$:

1. If $|x| \le 6(\mu+1)^2$ then $M_1 \le |2az/k - i| = \sqrt{(\frac{1}{4}(\mu+1)^{-2}|x|)^2 + (\frac{1}{4}(\mu+1)^{-2}|y| + 1)^2} \le 3$.
2. If $|x| > 6(\mu+1)^2$ then we estimate the numerator in (4.14) by

$$|2az/k - i| = |\frac{1}{4}(\mu+1)^{-2}x + i(\frac{1}{4}(\mu+1)^{-2}y - 1)| \le \frac{1}{4}(\mu+1)^{-2}|x| + 3/2 \le 2 + \frac{|x|}{4(\mu+1)}.$$

The denominator can be bounded from below if we consider only the real part:

$$|az^2/k + b - iz| \ge |\frac{1}{8}(\mu+1)^{-2}x^2 + 2 - \frac{1}{8}(\mu+1)^{-2}y^2 - \frac{3}{2}(\mu+1)^2 + y|.$$

From

$$\begin{aligned} &| - \frac{1}{8}(\mu+1)^{-2}y^2 - \frac{3}{2}(\mu+1)^2 + y| \le 4(\mu+1)^2 \quad \text{and} \\ &\qquad |\frac{1}{8}(\mu+1)^{-2}x^2 + 2| \ge \frac{1}{72}(\mu+1)^{-2}x^2 + \frac{1}{9}(\mu+1)^{-2}x^2 \ge \frac{1}{72}(\mu+1)^{-2}x^2 + 4(\mu+1)^2 \end{aligned}$$

we get $1 + \sqrt{|az^2/k + b - iz|} \ge 1 + \sqrt{\frac{1}{72}(\mu+1)^{-2}x^2} \ge 2/3 + \frac{1}{12}(\mu+1)^{-1}|x|$. Therefore $M_1 \le 3$.

The error estimate [5, (2.12)] reads

$$\| \exp(-A) - \sum_{j=-k_{\mathrm{E}}}^{k_{\mathrm{E}}} w_j (z_j I - A)^{-1} \|_2 \le M\sqrt{\pi}\, 3 \exp(4(\mu+1)^2 - 2) \left[\frac{Z_1}{N_1} + \frac{Z_2}{N_2} \right], \tag{4.15}$$

where

$$\begin{aligned} Z_1 &= 2\sqrt{k} \exp(-s(k_{\mathrm{E}}+1)^{2/3}) = 4\exp(-(2\pi^2(\mu+1)^2)^{1/3}(k_{\mathrm{E}}+1)^{2/3}) \\ &\le 4\exp(-((2\pi^2)^{1/3} - 1)(\mu+1)^{2/3}) \exp(-(\mu+1)^{2/3} k_{\mathrm{E}}^{2/3}) \\ &\le 4\exp(-\frac{3}{2}(\mu+1)^{2/3}) \exp(-(\mu+1)^{2/3} k_{\mathrm{E}}^{2/3}), \end{aligned}$$

$$
\begin{aligned}
&Z_2 = k\exp(-s(k_{\mathrm{E}}+1)^{2/3}) = Z_1,\\
&N_1 = \sqrt{a}(1-\exp(-s(k_{\mathrm{E}}+1)^{2/3})) \geq (\mu+1)^{-1}\frac{1}{\sqrt{2}}(1-\exp(-4)) \geq \frac{1}{2}(\mu+1)^{-1},\\
&N_2 = (k_{\mathrm{E}}+1)^{1/3}(2\pi d k a^2)^{1/3} = (k_{\mathrm{E}}+1)^{1/3}(4\pi(\mu+1)^{-2})^{1/3} \geq 2(\mu+1)^{-1}.
\end{aligned}
$$

Inserting these bounds in (4.15) , yields

$$
\begin{aligned}
\left[\frac{Z_1}{N_1}+\frac{Z_2}{N_2}\right] &\leq \frac{3}{2}(\mu+1)4\exp\left(-\frac{3}{2}(\mu+1)^{2/3}\right)\exp\left(-(\mu+1)^{2/3}k_{\mathrm{E}}^{2/3}\right)\\
&\leq 6\exp(-\frac{3}{2})\exp\left(-(\mu+1)^{2/3}k_{\mathrm{E}}^{2/3}\right),
\end{aligned}
$$

$$
\begin{aligned}
&\|\exp(-A) - \sum_{j=-k_{\mathrm{E}}}^{k_{\mathrm{E}}} w_j(z_jI-A)^{-1}\|_2\\
&\qquad\leq M\sqrt{\pi}3\exp\left(4(\mu+1)^2-2\right)6\exp(-\frac{3}{2})\exp\left(-(\mu+1)^{2/3}k_{\mathrm{E}}^{2/3}\right)\\
&\qquad\leq M\exp\left(-(\mu+1)^{2/3}k_{\mathrm{E}}^{2/3}+4(\mu+1)^2\right).
\end{aligned}
$$

■

Corollary 4.7 (Approximation of the solution to the Sylvester equation by a sum of resolvents) *Let $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{m\times m}$, $G \in \mathbb{C}^{n\times m}$, $\lambda_1 < \lambda_2 \in \mathbb{R}_{<0}, \mu \in \mathbb{R}_{\geq 0}$. The spectra $\sigma(A)$, $\sigma(B)$ of A, B are assumed to be contained in*

$$S := \{x+iy \in \mathbb{C} \mid \lambda_1 < x < \lambda_2 \text{ and } |y| < \mu\}.$$

We define the constants

$$
\begin{aligned}
\alpha &:= \frac{3}{2}|\lambda_2|^{-1},\\
c &:= \max_{C\in\{\alpha A,\alpha B\}} \max\{\|(zI-C)^{-1}\|_2 \mid z\in\mathbb{C}, \mathrm{dist}(z,\sigma(C))\geq 1\},\\
c_e &:= \max_{C\in\{A,B\}}\max_{t\geq 0}\|\exp(tC)\|_2.
\end{aligned}
$$

For each $k := (2k_{\mathrm{E}}+1)^2(2k_I+1)$ with $k_I, k_{\mathrm{E}} \in \mathbb{N}$ and

$$k_{\mathrm{E}} \geq \left((\log(M)+2)(\frac{\mu(1+\sqrt{k_I})}{2|\lambda_2|}+1)^{-2/3} + 4(\frac{\mu(1+\sqrt{k_I})}{2|\lambda_2|}+1)^{4/3}\right)^{3/2} \tag{4.16}$$

we define the integration parabola Γ, the interior Ω and the strong P-positivity constant M by

$$
\begin{aligned}
\Gamma &:= \{\frac{1}{2}((1+\sqrt{k})\alpha\mu+1)^{-2}\eta^2+\frac{1}{2}+i\eta \mid \eta\in(-\infty,\infty)\},\\
\Omega &:= \{\xi+i\eta \mid \eta\in(-\infty,\infty) \text{ and } \xi > \frac{1}{2}((1+\sqrt{k})\alpha\mu+1)^{-2}\eta^2+\frac{1}{2}\},\\
M &:= \max_{t\in[0,1+\sqrt{k_I}]}\max_{C\in\{A,B\}}\sup_{z\in\mathbb{C}\setminus\Omega}\|((z-2)I+\alpha tC)^{-1}\|_2(1+\sqrt{|z|}).
\end{aligned}
$$

Then there exist points $t_j \in [0,\infty)$ and weights $w_j \in \mathbb{C}$, $j \in \{1,\ldots,k\}$, such that the solution X to (4.4) can be approximated by a matrix $\tilde{X} = \sum_{j=1}^{k} w_j(z_jI-B)^{-1}G(z_jI-A)^{-1}$ with

$$
\begin{aligned}
\|X-\tilde{X}\|_2 &\leq c^2 C_{\mathrm{sinc}}\exp((2+2\alpha\mu)/\pi)\frac{3\|G\|_2}{32\pi^2|\lambda_2|}(13+3|\lambda_2|^{-1}(\mu-\lambda_1))^2\exp\left(-\sqrt{k_I}\right) \qquad (4.17)\\
&+ c_e M\frac{9\|G\|_2}{2|\lambda_2|}(2k_I+1)\exp\left(2+4(\frac{3+3\sqrt{k_I}}{|\lambda_2|}\mu+1)^2-(\frac{3+3\sqrt{k_I}}{|\lambda_2|}\mu+1)^{2/3}k_E^{2/3}\right).
\end{aligned}
$$

If A and B are symmetric then

$$\|X - \tilde{X}\|_2 \le 3C_{\mathrm{sinc}} \exp(2/\pi)\|G\|_2 \frac{|\lambda_1|^2}{|\lambda_2|^3} \exp(-\sqrt{k_I}) + M(9k_I + 9/2)|\lambda_2|^{-1}\|G\|_2 \exp(6 - k_E^{2/3}).$$

Proof. Let X be the unique solution to (4.4) . According to Theorem 4.3, the matrix

$$Y := \sum_{j=-k_I}^{k_I} \alpha\omega_j \exp(\alpha t_j B) G \exp(\alpha t_j A)$$

fulfils

$$\begin{aligned}\|X - Y\|_2 &\le c^2 C_{\mathrm{sinc}} \exp((2 + 2\alpha\mu)/\pi) \frac{\alpha\|G\|_2}{4\pi^2}(8 + \alpha(\mu + \lambda_2 - \lambda_1))^2 \exp(-\sqrt{k_I}) \\ &= c^2 C_{\mathrm{sinc}} \exp((2 + 2\alpha\mu)/\pi) \frac{3\|G\|_2}{32\pi^2|\lambda_2|}(13 + 3|\lambda_2|^{-1}(\mu - \lambda_1))^2 \exp(-\sqrt{k_I})\end{aligned}$$

which produces the first term in (4.17) . The spectra of the matrices $-\alpha t_j A + 2I, -\alpha t_j B + 2I$ are contained in $\{x + iy \in \mathbb{C} \mid x > 2 \text{ and } |y| < \alpha(1 + \sqrt{k_I})\mu\}$. Application of Lemma 4.6 for the matrices $-\alpha t_j A + 2I, -\alpha t_j B + 2I$ instead of A and $\alpha(1 + \sqrt{k_I})\mu$ instead of μ yields

$$\begin{aligned}\| \exp(\alpha t_j A - 2I) - &\overbrace{\sum_{i=-k_{\mathrm{E}}}^{k_{\mathrm{E}}} \tilde{w}_{j,i}(\tilde{z}_{j,i} I - (-\alpha t_j A + 2I))^{-1}}^{=:E_{A,j}} \|_2 \\ &\le M \exp\left(4(\alpha(1 + \sqrt{k_I})\mu + 1)^2 - (\alpha(1 + \sqrt{k_I})\mu + 1)^{2/3} k_{\mathrm{E}}^{2/3}\right)\end{aligned}$$

and the same for B instead of A. For the matrix

$$\tilde{X} := \sum_{j=-k_I}^{k_I} \alpha\omega_j E_{B,j} \exp(2I) G \exp(2I) E_{A,j}$$

we get the error estimate

$$\begin{aligned}\|Y - \tilde{X}\|_2 = \|Y - &\sum_{j=-k_I}^{k_I} \alpha\omega_j \exp(2I) \exp(\alpha t_j B - 2I) G \exp(2I) E_{A,j} \\ &+ \sum_{j=-k_I}^{k_I} \alpha\omega_j \exp(2I) \exp(\alpha t_j B - 2I) G \exp(2I) E_{A,j} - \tilde{X}\|_2 \\ &\le (2k_I + 1)\alpha c_e \|G\|_2 e^2 M \exp\left(4(\alpha(1 + \sqrt{k_I})\mu + 1)^2 - (\alpha(1 + \sqrt{k_I})\mu + 1)^{2/3} k_{\mathrm{E}}^{2/3}\right) \\ &\quad + (2k_I + 1)\alpha \max_{j=-k_I}^{k_I} \|E_{A,j}\|_2 \|G\|_2 e^2 M \exp\left(4(\alpha(1 + \sqrt{k_I})\mu + 1)^2 - (\alpha(1 + \sqrt{k_I})\mu + 1)^{2/3} k_{\mathrm{E}}^{2/3}\right) \\ &\overset{(4.16)}{\le} (2k_I + 1)\alpha 3 c_e \|G\|_2 e^2 M \exp\left(4(\alpha(1 + \sqrt{k_I})\mu + 1)^2 - (\alpha(1 + \sqrt{k_I})\mu + 1)^{2/3} k_{\mathrm{E}}^{2/3}\right) \\ &= c_e M \frac{9\|G\|_2}{2|\lambda_2|}(2k_I + 1) \exp\left(2 + 4(\frac{3 + 3\sqrt{k_I}}{2|\lambda_2|}\mu + 1)^2 - (\frac{3 + 3\sqrt{k_I}}{2|\lambda_2|}\mu + 1)^{2/3} k_{\mathrm{E}}^{2/3}\right).\end{aligned}$$

If A and B are both symmetric, then we can apply Lemma 4.1 and get $c = 1$, $c_e = 1$, $\mu = 0$. ■

Corollary 4.8 ($\mathcal{H}$-matrix approximation to the solution of the Sylvester equation) *We use the same notation as in Corollary 4.7. We assume*

$$G \in \mathcal{M}_{\mathcal{H},k_G}(T_{I \times J})$$

and that for $\delta \in \mathbb{R}_{>0}$ all resolvents $(z_jI-A)^{-1}, (z_jI-B)^{-1}$ can be approximated by an $\mathcal{H}$-matrix $A^{(j)} \in \mathcal{M}_{\mathcal{H},k_A}(T_{J\times J})$, $B^{(j)} \in \mathcal{M}_{\mathcal{H},k_B}(T_{I\times I})$ with

$$\|(z_jI-A)^{-1} - A^{(j)}\|_2 \le \delta, \quad \|(z_jI-B)^{-1} - B^{(j)}\|_2 \le \delta \tag{4.18}$$

(for a more detailed analysis concerning the existence of $\mathcal{H}$-matrix approximants the reader is referred to [3]). If we define the approximate solution

$$X_{\mathcal{H}} := \sum_{j=1}^{k} w_j B^{(j)} G A^{(j)},$$

then the approximation error is of the size

$$\|X - X_{\mathcal{H}}\|_2 = \mathcal{O}\left(\exp(-\sqrt{k_I}) + \exp(-k_E^{2/3}) + \delta\right)$$

(neglecting linear terms in k_I) while Remark 1.2 yields

$$X_{\mathcal{H}} \in \mathcal{M}_{\mathcal{H},k_X}(T_{I\times J}), \qquad k_X = \mathcal{O}\left(\log(n+m)^2(k_A+k_B)k_G k_E^2 k_I\right).$$

If the rank needed for the $\mathcal{H}$-approximants is $k_A = k_B = \mathcal{O}(\log(1/\delta))$ then

$$\|X - X_{\mathcal{H}}\|_2 = \mathcal{O}(\delta) \quad \textit{for} \quad k_X = \mathcal{O}(k_G \log(n+m)^2 \log(1/\delta)^6).$$

Since $\|G\|_2 = \|AX - XB\|_2 \le (\|A\|_2 + \|B\|_2)\|X\|_2$, we also get the above estimates for the relative error.

4.2 Lyapunov Equation

The Lyapunov equation

$$A^TX + XA + G = 0 \tag{4.19}$$

for $A, G \in \mathbb{R}^{n\times n}$ is a special Sylvester equation (4.4) for $B := A^T$. Let the spectrum $\sigma(A)$ of A be contained in the left complex halfplane and let X denote the unique solution to (4.19) .

If G is of low rank $k_G > 0$, then Remark 4.5 proves that for each $\varepsilon \in (0,1)$ there exists a matrix X_R of rank $\mathcal{O}(\log(1/\varepsilon))$ such that $\|X - X_R\|_2/\|X\|_2 \le \varepsilon$. For the ease of presentation we neglect the constants.

If G is an element of the $\mathcal{H}$-matrix class $\mathcal{M}_{\mathcal{H},k_G}(T_{I\times I})$ and if the resolvents $(zI-A)^{-1}$ in (4.18) can be approximated by an $\mathcal{H}$-matrix up to an error of $\varepsilon \in (0,1)$ with blockwise rank $k_A = \mathcal{O}(\log(1/\varepsilon))$, then Corollary 4.8 proves that there exists a matrix $X_{\mathcal{H}} \in \mathcal{M}_{\mathcal{H},k_X}(T_{I\times I})$ of blockwise rank $k_X = \mathcal{O}(\log(1/\varepsilon)^6)$ such that $\|X - \tilde{X}\|_2/\|X\|_2 \le \varepsilon$. Again we neglect the constants.

These results are a generalisation of the ones from [8] and [15] (for the R(k)-matrix case) to the $\mathcal{H}$-matrix case. With the assumption that $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and G of low rank k_G, the author of [15] was able to prove that the singular values $\lambda_1 \ge \ldots \ge \lambda_n \ge 0$ of X are bounded by

$$\frac{\lambda_{m\cdot k_G+1}}{\lambda_1} \le (\prod_{j=0}^{m-1} \frac{\kappa_{m,j}-1}{\kappa_{m,j}+1})^2, \qquad \kappa_{m,j} := \mathrm{cond}_2(A)^{\frac{2j+1}{2m}}. \tag{4.20}$$

Remark 4.9 *In order to compare our estimate to the result by Penzl [15] we have to analyse* (4.20) *. We assume that the spectral condition of A is larger than* 1*. Let $\varepsilon \in (0,1)$ and*

$$m := \lceil (\log_2(\mathrm{cond}_2(A)) + 1)(\log_2(1/\varepsilon) + 1) \rceil.$$

It follows for $j \le \lceil \log_2(1/\varepsilon) \rceil$ that $\kappa_j = (\mathrm{cond}_2(A))^{\frac{2j+1}{2m}} \le (\mathrm{cond}_2(A))^{1/\log_2(\mathrm{cond}_2(A))} \le 2$ and

$$\left(\prod_{j=0}^{m-1} \frac{\kappa_{m,j}-1}{\kappa_{m,j}+1}\right)^2 \le \left(\prod_{j=0}^{\lceil \log_2(1/\varepsilon)\rceil} \frac{1}{3}\right)^2 \le 3^{-2\log_2(1/\varepsilon)} \le \varepsilon.$$

This proves that the solution X to (4.4) *can be approximated up to a relative error of ε by a matrix $\tilde{X}$ of rank $\mathcal{O}(k_C \log_2(\mathrm{cond}_2(A)) \log_2(1/\varepsilon)))$.*

A conclusion of all the previous results is that the solution X to the Lyapunov equation can be approximated by an $\mathcal{H}$-matrix (or R(k)-matrix) if G is an $\mathcal{H}$-matrix (or R(k)-matrix). An algorithm to compute the $\mathcal{H}$-matrix (or R(k)-matrix) approximation efficiently will be presented in the following sections.

4.3 Riccati Equation

Let $A, F, G \in \mathbb{R}^{n\times n}$ and let the spectrum $\sigma(A)$ of A be contained in the left complex halfplane. A solution X of the Riccati equation

$$A^T X + XA - XFX + G = 0 \tag{4.21}$$

can (for theoretical considerations) be regarded as a solution of the Lyapunov equation

$$A^T X + XA + \tilde{G} = 0, \qquad \tilde{G} := G - XFX.$$

If F and G are of rank $k_F, k_G \ll n$ then $\tilde{G}$ is of rank at most $k_{\tilde{G}} = k_F + k_G$ and we can apply Remark 4.5 to prove that the rank k_X that is necessary to approximate X up to a relative error of ε in the set of R(k)-matrices is $k_X = \mathcal{O}(\log(1/\varepsilon))$.

If F is of rank $k_F \ll n$ and $G \in \mathcal{M}_{\mathcal{H},k_G}(T_{I\times I})$, $k_G \ll n$, then $\tilde{G} \in \mathcal{M}_{\mathcal{H},k_G+k_F}(T_{I\times I})$ and we can apply Corollary 4.8 to prove that the blockwise rank k_X that is necessary to approximate X up to a relative error of ε in the set $\mathcal{M}_{\mathcal{H},k_X}(T_{I\times I})$ is $k_X = \mathcal{O}(\log(1/\varepsilon)^6)$.

Note that we make use of the low rank of the matrix F, while G can be an $\mathcal{H}$-matrix. For the general situation that F is not of low rank, the algorithms presented later are still applicable, but we cannot bound the (blockwise) rank needed to approximate the solution to (4.21) .

4.4 Matrix Sign Function

In the last two subsections we have seen that the solution X to the Riccati or Lyapunov equation can (under moderate assumptions) be represented as an $\mathcal{H}$-matrix. The following two questions arise:

- Does the matrix sign$\left(\begin{bmatrix} A^T & G \\ F & -A \end{bmatrix}\right)$ consist of $\mathcal{H}$-matrix substructures ?
- Do the iterates S_i from (3.4) bear any specific structure ?

We assume that the matrices $F, G \in \mathbb{R}^{n\times n}$ are symmetric, F is of low rank $k_F := \mathrm{rank}(F)$ and $A - FX$ is a stability matrix (X solves (4.21)). We define the matrix

$$S := \begin{bmatrix} A^T & G \\ F & -A \end{bmatrix}.$$

For the solution X to (4.21) there holds

$$S = \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} -(A - FX) & -F \\ 0 & (A - FX)^T \end{bmatrix} \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1}$$

and it follows (see [17] and use $\operatorname{sign}(S)S = S\operatorname{sign}(S)$)

$$\begin{aligned}
\operatorname{sign}(S) &= \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \operatorname{sign}\left(\begin{bmatrix} -(A-FX) & -F \\ 0 & (A-FX)^T \end{bmatrix}\right) \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1} \\
&= \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} I & Z \\ 0 & -I \end{bmatrix} \begin{bmatrix} 0 & I \\ -I & X \end{bmatrix} \\
&= \left[\begin{array}{c|c} -XZ-I & 2X+XZX \\ \hline -Z & I+ZX \end{array}\right], \qquad (4.22)
\end{aligned}$$

where Z satisfies the Lyapunov equation

$$AZ + ZA^T + \tilde{G} = 0, \qquad \tilde{G} := -FXZ - ZXF - 2F. \qquad (4.23)$$

According to Remark 4.5, the matrix Z can be approximated by a low rank matrix $\tilde{Z}$ with $k_Z := \operatorname{rank}(\tilde{Z}) = \mathcal{O}(k_F \log(1/\varepsilon))$ and $\|Z - \tilde{Z}\|_2 \leq \varepsilon$. A direct conclusion is

Corollary 4.10 (Approximation of $\operatorname{sign}(S)$**)** *The matrix* $\operatorname{sign}(S)$ *can be approximated by a matrix*

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}, \qquad K_{11} + I, K_{21}, K_{22} - I, K_{12} - 2X \in \mathrm{R}(k_Z).$$

We denote the index set by $I := \{1, \ldots, n\}$*. The matrices* K_{11}, K_{21} *and* K_{22} *are contained in* $\mathcal{M}_{\mathcal{H},k_Z}(T_{I\times I})$ *and* K_{12} *is contained in the same space (*$\mathcal{H}$*-matrix or low rank matrix) as* X *but with rank increased by* k_Z*. The approximation error* $\|Z - \tilde{Z}\|_2$ *leads to the estimate*

$$\|\operatorname{sign}(S) - K\|_2 \leq \varepsilon(1 + \|X\|_2^2) = \mathcal{O}(\varepsilon).$$

Lemma 4.11 *Let* $A_0 := A - FX$ *and* $F_0 := F$*. For each* $i \in \mathbb{N}_0$ *we define*

$$A_{i+1} := \frac{1}{2}\left(A_i + A_i^{-1}\right) \qquad \textit{and} \qquad (4.24)$$

$$F_{i+1} := \frac{1}{2}\left(F_i + A_i^{-1} F_i A_i^{-T}\right). \qquad (4.25)$$

Ultimately A_i *converges to* $-I$ *(because* A_0 *is a stability matrix) and* F_i *converges to* $-Z$*. The iterates* F_i *can be approximated by a matrix* $\tilde{F}_i$ *of rank at most* $2k_Z$*, such that*

$$\|F_i - \tilde{F}_i\|_2 \leq \varepsilon \|A_i\|_2.$$

Proof. Since $Y := -\frac{1}{2}Z$ solves (due to (4.23)) the Lyapunov equation $(A - FX)Y + Y(A - FX)^T + F = 0$, it holds

$$\tilde{S}_0 := \begin{bmatrix} A - FX & F \\ 0 & -(A-FX)^T \end{bmatrix} = \begin{bmatrix} Y & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} -(A-FX)^T & 0 \\ 0 & A - FX \end{bmatrix} \begin{bmatrix} Y & -I \\ I & 0 \end{bmatrix}^{-1}.$$

For the matrix

$$\tilde{S}_0 = \begin{bmatrix} A_0 & F_0 \\ 0 & -A_0^T \end{bmatrix} = \begin{bmatrix} Y & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} -A_0^T & 0 \\ 0 & A_0 \end{bmatrix} \begin{bmatrix} Y & -I \\ I & 0 \end{bmatrix}^{-1}$$

we perform the Newton iteration (3.4), $\tilde{S}_{i+1} := \frac{1}{2}(\tilde{S}_i + \tilde{S}_i^{-1})$, and get

$$\tilde{S}_i = \begin{bmatrix} A_i & F_i \\ 0 & -A_i^T \end{bmatrix} = \begin{bmatrix} Y & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} -A_i^T & 0 \\ 0 & A_i \end{bmatrix} \begin{bmatrix} Y & -I \\ I & 0 \end{bmatrix}^{-1} = \begin{bmatrix} A_i & -YA_i^T - A_iY \\ 0 & A_i^T \end{bmatrix}.$$

It follows $F_i = -YA_i^T - A_iY = \frac{1}{2}(ZA_i^T + A_iZ)$ and for the matrix $\tilde{F}_i := \frac{1}{2}(\tilde{Z}A_i^T + A_i\tilde{Z})$ of rank $2k_Z$ we obtain

$$\|F_i - \tilde{F}_i\|_2 \leq \frac{1}{2}(\|Z - \tilde{Z}\|_2 \|A_i^T\|_2 + \|A_i\|_2 \|Z - \tilde{Z}\|_2) \leq \varepsilon \|A_i\|_2.$$

■

Lemma 4.12 *Let $A_0 \in \mathbb{C}^{n\times n}$ be a regular diagonalisable matrix whose spectrum $\sigma(A_0)$ does not intersect the imaginary axis. Let $A_0 = TD_0T^{-1}$ be a diagonalisation of A_0, $d := \max_{\lambda\in\sigma(A)} |\lambda + \lambda^{-1}|$. For all $i \in \mathbb{N}$ we define*

$$A_i := \frac{1}{2}\left(A_{i-1} + A_{i-1}^{-1}\right).$$

If for all $x + iy \in \sigma(A)$ the imaginary part is bounded by $|y| \le |x|$, then the norms of A_i, A_i^{-1} can be bounded by

$$\|A_i^{-1}\|_2 \le \sqrt{2}\,\mathrm{cond}_2(T), \qquad \|A_i\|_2 \le \left(2^{-i}d + \sqrt{2}\left(1 - 2^{-i}\right)\right)\mathrm{cond}_2(T). \tag{4.26}$$

If $\sigma(A_0) \subset \mathbb{R}$, then

$$\|A_i^{-1}\|_2 \le \mathrm{cond}_2(T), \qquad \|A_i\|_2 \le \left(1 + 2^{-i}\,(d-1)\right)\mathrm{cond}_2(T). \tag{4.27}$$

Proof. We define for $i \in \mathbb{N}$ the matrices

$$D_i := \frac{1}{2}(D_{i-1} + D_{i-1}^{-1}),$$

such that $A_i = TD_iT^{-1}$. Let $\|M\|_T := \|T^{-1}MT\|_2$. Then $\|A_i\|_T = \|D_i\|_2$ and $\|A_i\|_2 \le \mathrm{cond}_2(T)\|A_i\|_T$.

Case (4.26) **:** If $|y| \le |x|$ for all $x + iy \in \sigma(A)$ then this implies $|\Im m(d_i)| \le |\Re e(d_i)|$ for $d_i = (D_0)_{ii}$. For all subsequent $d_i = (D_j)_{ii}$ with $x + iy = (D_{j-1})_{ii}$ it follows

$$d_i = \frac{1}{2}(x + iy + \frac{x - iy}{x^2 + y^2}) = x + \frac{x}{x^2 + y^2} + i(y - \frac{y}{x^2 + y^2})$$

which implies $|\Im m(d_i)| \le |\Re e(d_i)|$.

We prove $\|D_i^{-1}\|_2 \le \sqrt{2}$ for all $i \in \mathbb{N}$ and by induction

$$\|D_i\|_2 \le 2^{-i}d + \sum_{j=1}^{i-1} 2^{-j}\sqrt{2},$$

which is fulfilled for $i = 1$. Let $i \in \mathbb{N}$ and $D_{i-1} = \mathrm{diag}(d_1, \dots, d_n)$.

$$\|D_i^{-1}\|_2 = 2\max_{i=1,\dots,n} |d_i + d_i^{-1}|^{-1} \overset{|\Im m(d_i)|\le|\Re e(d_i)|}{\le} 2/\sqrt{2} = \sqrt{2}.$$

$$\|D_i\|_2 = \|\frac{1}{2}(D_{i-1} + D_{i-1}^{-1})\|_2 \le \frac{1}{2}\sqrt{2} + \frac{1}{2}(2^{-i+1}d + \sum_{j=1}^{i-2} 2^{-j}\sqrt{2}) = 2^{-i}d + \sum_{j=1}^{i-1} 2^{-j}\sqrt{2}.$$

Case (4.27) **:** Same as above but all $(D_j)_{ii}$ are real-valued and thus $2\max_{i=1,\dots,n} |d_i + d_i^{-1}|^{-1} \le 1$. ∎

Theorem 4.13 (Newton iteration for $\mathrm{sign}(S)$ with low rank G) *Let the matrices $F, G \in \mathbb{R}^{n\times n}$ be symmetric, F of rank k_F and G of rank k_G. Let $\varepsilon \in (0, 1)$ and $A \in \mathbb{R}^{n\times n}$. We assume:*

1. *The solution X to (4.21) can be approximated by a matrix $\tilde{X}$ of rank k_X such that $\|X - \tilde{X}\|_2 \le \varepsilon$.*
2. *$A - FX$ is a stability matrix.*
3. *Each of the matrices A_i from (4.24) in Lemma 4.11 can be approximated by a matrix $\tilde{A}_i \in \mathcal{M}_{\mathcal{H},k_A}(T_{I\times I})$ such that $\|A_i - \tilde{A}_i\|_2 \le \varepsilon$.*

Then each iterate S_i of the iteration (3.4) *can be approximated by a matrix*

$$K^{(i)} = \begin{bmatrix} K_{11}^{(i)} & K_{12}^{(i)} \\ K_{21}^{(i)} & K_{22}^{(i)} \end{bmatrix}, \qquad K_{11}^{(i)}, K_{22}^{(i)} \in \mathcal{M}_{\mathcal{H},k_A+2k_Z}(T_{I\times I}), \quad K_{12}^{(i)} \in \mathrm{R}(2k_X+2k_Z), \quad K_{21}^{(i)} \in \mathrm{R}(2k_Z)$$

and the approximation error is bounded by

$$\|S_i - K^{(i)}\|_2 \leq \varepsilon(1 + \|A_i\|_2(2 + \|X\|_2 + \|X\|_2^2)).$$

Proof. Using the notation from Lemma 4.11, the statement holds for

$$S_0 = \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} -(A-FX) & -F \\ 0 & (A-FX)^T \end{bmatrix} \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1} = \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} (-\tilde{S}_0) \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1}$$

and we can conclude

$$\begin{aligned} S_i &= \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} (-\tilde{S}_i) \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} -A_i & -F_i \\ 0 & A_i^T \end{bmatrix} \begin{bmatrix} 0 & I \\ -I & X \end{bmatrix} \\ &= \left[\begin{array}{c|c} A_i^T + XF_i & -A_i^T X - XA_i - XF_iX \\ \hline F_i & -A_i - F_iX \end{array}\right]. \end{aligned}$$

We define $K_{11}^{(i)} := \tilde{A}_i^T + X\tilde{F}_i$, $K_{22}^{(i)} := -\tilde{A}_i - \tilde{F}_iX$, $K_{12}^{(i)} := -A_i^T\tilde{X} - \tilde{X}A_i - X\tilde{F}_iX$, $K_{21}^{(i)} := \tilde{F}_i$ and get the error estimate

$$\begin{aligned} \|S_i - K^{(i)}\|_2 \leq\, & \|A_i - \tilde{A}_i\|_2 + \|X\|_2\|\tilde{F}_i - F_i\|_2 \\ & + \max\{\|F_i - \tilde{F}_i\|_2, \quad 2\|A_i\|_2\|X - \tilde{X}\|_2 + \|\tilde{F}_i - F_i\|_2\|X\|_2^2\} \\ \leq\, & \varepsilon + \varepsilon\|X\|_2\|A_i\|_2 + \max\{\varepsilon\|A_i\|_2, \quad 2\varepsilon\|A_i\|_2 + \varepsilon\|A_i\|_2\|X\|_2^2\} \\ =\, & \varepsilon(1 + 2\|A_i\|_2 + \|A_i\|_2\|X\|_2 + \|A_i\|_2\|X\|_2^2). \end{aligned}$$

From Lemma 4.11 the rank of $\tilde{F}_i$ is bounded by $2k_Z$. Due to the ideal property (see Remark 1.2) of R(k)-matrices, we have $\mathrm{rank}(K_{12}^{(i)}) \leq 2k_X + 2k_Z$. Since $\tilde{A}_i \in \mathcal{M}_{\mathcal{H},k_A}(T_{I\times I})$, it follows that $K_{11}^{(i)}, K_{22}^{(i)} \in \mathcal{M}_{\mathcal{H},k_A+2k_Z}(T_{I\times I})$ ■

Theorem 4.14 (Newton iteration for $\mathrm{sign}(S)$ **with** $\mathcal{H}$**-matrix** G**)** *Let the matrices $F, G \in \mathbb{R}^{n\times n}$ be symmetric, F of rank k_F and $G \in \mathcal{M}_{\mathcal{H},k_G}(T_{I\times I})$. Let $\varepsilon \in (0,1)$ and $A \in \mathbb{R}^{n\times n}$. The depth of $T_{I\times I}$ is denoted by p and the constants describing the sparsity and idempotency are $C_{\mathrm{sp}}, C_{\mathrm{id}}$. We assume:*

1. *The solution X to* (4.21) *can be approximated by $\tilde{X} \in \mathcal{M}_{\mathcal{H},k_X}(T_{I\times I})$ such that $\|X - \tilde{X}\|_2 \leq \varepsilon$.*
2. *$A - FX$ is a stability matrix.*
3. *Each of the matrices A_i from* (4.24) *can be approximated by a matrix $\tilde{A}_i \in \mathcal{M}_{\mathcal{H},k_A}(T_{I\times I})$ such that $\|A_i - \tilde{A}_i\|_2 \leq \varepsilon$.*

Then each iterate S_i of the iteration (3.4) *can be approximated by a matrix*

$$K^{(i)} = \begin{bmatrix} K_{11}^{(i)} & K_{12}^{(i)} \\ K_{21}^{(i)} & K_{22}^{(i)} \end{bmatrix}, \qquad K_{11}^{(i)}, K_{22}^{(i)} \in \mathcal{M}_{\mathcal{H},k'}(T_{I\times I}), \quad K_{12}^{(i)} \in \mathrm{R}(3k_X), \quad K_{21}^{(i)} \in \mathrm{R}(2k_Z)$$

with $k' := 2C_{\mathrm{sp}}C_{\mathrm{id}}p\max\{k, k_X\} + 2k_Z$ such that the approximation error is bounded by

$$\|S_i - K^{(i)}\|_2 \leq \varepsilon(1 + \|F_i\|_2(1 + \|X\|_2) + 2\|A_i\|_2).$$

Proof. Using the notation from Lemma 4.11, the statement holds for

$$S_0 = \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} -(A-FX) & -F \\ 0 & (A-FX)^T \end{bmatrix} \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1} = \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix} (-\tilde{S}_0) \begin{bmatrix} X & -I \\ I & 0 \end{bmatrix}^{-1}$$

and we can conclude, as in the previous Theorem,

$$S_i = \left[\begin{array}{c|c} A_i^T + XF_i & -A_i^T X - XA_i - XF_iX \\ \hline F_i & -A_i - F_iX \end{array}\right].$$

We define the matrices $K_{11}^{(i)} := \tilde{A}_i^T + X\tilde{F}_i$, $K_{22}^{(i)} := -\tilde{A}_i - \tilde{F}_iX$, $K_{12}^{(i)} := -\tilde{A}_i^T\tilde{X} - \tilde{X}\tilde{A}_i - X\tilde{F}_iX$, $K_{21}^{(i)} := \tilde{F}_i$ and get the error estimate

$$\|S_i - K_i\|_2 \leq \varepsilon(1 + 2\|A_i\|_2 + \|A_i\|_2\|X\|_2 + \|A_i\|_2\|X\|_2^2).$$

From Lemma 4.11 the rank of $\tilde{F}_i$ is bounded by $2k_Z$. Due to the ideal property of R(k)-matrices, we get $K_{11}^{(i)}, K_{22}^{(i)} \in \mathcal{M}_{\mathcal{H},k_A+2k_Z}(T_{I\times I})$. Remark 1.2 ensures $\tilde{X}\tilde{A}_i, \tilde{A}_i^T\tilde{X} \in \mathcal{M}_{\mathcal{H},k''}(T_{I\times I})$ with $k'' := C_{\mathrm{sp}}C_{\mathrm{id}}p\max\{k, k_X\}$. Since $\operatorname{rank}(X\tilde{F}_iX) \leq 2k_Z$, we get $K_{12}^{(i)} \in \mathcal{M}_{\mathcal{H},k'}(T_{I\times I})$ with $k' := 2k'' + 2k_Z$. ■

Remark 4.15 *In Theorem 4.13 and Theorem 4.14 we need three assumptions:*

1. *The solution X can be approximated in a suitable format, namely by an* R(k)*-matrix or an $\mathcal{H}$-matrix $\tilde{X}$. This has already been investigated in Remark 4.5 and Corollary 4.8.*
2. *$A - FX$ is a stability matrix. This can be assumed because we seek a stabilising solution X.*
3. *Each of the matrices A_i can be approximated by an $\mathcal{H}$-matrix. If A is the (sparse) stiffness matrix from the (finite element or finite difference) discretisation of an elliptic partial differential operator, then $A - FX$ belongs to $\mathcal{M}_{\mathcal{H},k_F}(T_{I\times I})$. The set of $\mathcal{H}$-matrices was chosen such that the inverse A^{-1} to A can be approximated by an $\mathcal{H}$-matrix $\widetilde{A^{-1}}$ with moderate blockwise rank. Since the matrix A_1 is a rank $2k_F$-perturbation of $\frac{1}{2}(A + A^{-1})$, we can approximate A_1 by an $\mathcal{H}$-matrix. Moreover, the A_i can be regarded as the discretisation of an elliptic pseudo-differential operator which makes it plausible that they can again be approximated by an $\mathcal{H}$-matrix.*

5 Using $\mathcal{H}$-Matrices for the Solution

In the last section we have used the matrix exponential in order to prove that, if the matrix G is an $\mathcal{H}$-matrix or R(k)-matrix, then the solution X to the Riccati or Lyapunov equation is an $\mathcal{H}$-matrix or R(k)-matrix. The representation of the solution X used in Corollary 4.4 and Corollary 4.8 leads to an algorithm where one can insert $\mathcal{H}$-matrix arithmetics to get a fast solver for the Lyapunov equation. For the solution of the Riccati equation (4.21) one has to deal with a series of Lyapunov equations (4.19), where one can exploit the fact that the matrices A_ν appearing in the Lyapunov equation in the ν-th Newton step are rank-k_F-perturbations of A.

An entirely different approach for the solution of the Riccati equation is to use the algorithm of Subsection 3.2 with the formatted $\mathcal{H}$-matrix arithmetics. We already know that the iterates in (3.4) can be approximated by blockwise R(k)-matrices and $\mathcal{H}$-matrices, but the influence of the approximation error in the numerical scheme has to be analysed. It turns out that scaling strategies (since $\operatorname{sign}(S) = \operatorname{sign}(\alpha S)$ for $\alpha > 0$ one can choose a scaling parameter α in each step to accelerate convergence) are not advisable.

5.1 Application of the Matrix Sign Function

The formatted $\mathcal{H}$-matrix operations $(\oplus, \odot, \widetilde{\mathrm{Inv}})$ introduce some kind of "rounding" error that has not yet been regarded. Our main concern is the iterative scheme (3.4) to compute the matrix sign function. Since there will be $\mathcal{O}(\log(n))$ steps, the rounding errors could be amplified such that the approximate solution does not approximate the solution.

We start the (exact) Newton iteration with the matrix $S_0 := \begin{bmatrix} A^T & G \\ F & -A \end{bmatrix}$ and define the (exact) iterates for $i \in \mathbb{N}_0$ as $S_{i+1} := \frac{1}{2}(S_i + S_i^{-1})$. The exact starting matrix is replaced by some approximation $\tilde{S}_0 \in \mathcal{M}_{\mathcal{H},k}(T)$ where T is, as we have derived in Theorem 4.13 and Theorem 4.14, the partitioning

$$T = \begin{bmatrix} T_{I\times I} & I \times I \\ I \times I & T_{I\times I} \end{bmatrix} \quad \text{(if } G \text{ has low rank)}, \qquad T = \begin{bmatrix} T_{I\times I} & T_{I\times I} \\ I \times I & T_{I\times I} \end{bmatrix} \quad \text{(if } G \text{ is an } \mathcal{H}\text{-matrix)},$$

for the index set $I = \{1, \ldots, n\}$. For the sake of simplicity we assume that the rank k in the blockwise low rank structures is always the same, which could be enforced by taking the maximum of all ranks appearing in Theorem 4.13 and Theorem 4.14.

The (inexact) iterates are defined as

$$\tilde{S}_{i+1} := \frac{1}{2}(\tilde{S}_i \oplus \widetilde{\mathrm{Inv}}(\tilde{S}_i)) \quad \in \mathcal{M}_{\mathcal{H},k}(T), \tag{5.1}$$

where $\oplus$ and $\widetilde{\mathrm{Inv}}$ are the formatted $\mathcal{H}$-matrix addition and inversion in the set $\mathcal{M}_{\mathcal{H},k}(T)$. The accuracy δ, ρ of the formatted addition and inversion, respectively, can be controlled by the blockwise rank k of the $\mathcal{H}$-matrices. Typically, we have $k = \mathcal{O}(\log(1/\delta) + \log(1/\rho))$.

Theorem 5.1 (Error propagation) *Let $\rho, \delta > 0$, $i_{max} \in \mathbb{N}$ and for all $i = 0, \ldots, i_{max}$*

$$\|\tilde{S}_i^{-1} - \widetilde{\mathrm{Inv}}(\tilde{S}_i)\|_2 \leq \delta \qquad (\mathcal{H}\textit{-matrix inversion error}), \tag{5.2}$$

$$\left\| \left(\tilde{S}_i + \widetilde{\mathrm{Inv}}(\tilde{S}_i)\right) - \left(\tilde{S}_i \oplus \widetilde{\mathrm{Inv}}(\tilde{S}_i)\right) \right\|_2 \leq \rho \qquad (\mathcal{H}\textit{-matrix addition error}). \tag{5.3}$$

We define the error amplification coefficients

$$c_0 := \|\tilde{S}_0 - S\|_2(\delta + \rho)^{-1}.$$
$$c_{i+1} := \frac{1}{2}(1 + c_i + c_i \frac{\|S_i^{-1}\|_2^2}{1 - c_i(\rho + \delta)\|S_i^{-1}\|_2})$$

by induction for $i \in \mathbb{N}_0$ and assume that

$$c_i(\rho + \delta)\|S_i^{-1}\|_2 < 1 \tag{5.4}$$

for all $i = 0, \ldots, i_{max}$. Then the distance of the inexact iterate to the exact iterate can be bounded by

$$\|\tilde{S}_i - S_i\|_2 \leq c_i(\rho + \delta). \tag{5.5}$$

Proof. By induction, where (5.5) for $i = 0$ is fulfilled due to the definition of c_0. For the induction step $i \to i + 1$ we define

$$E_i := S_i - \tilde{S}_i,$$
$$D_i := \widetilde{\mathrm{Inv}}(\tilde{S}_i) - \tilde{S}_i^{-1},$$
$$R_i := (\tilde{S}_i + \widetilde{\mathrm{Inv}}(\tilde{S}_i)) - (\tilde{S}_i \oplus \widetilde{\mathrm{Inv}}(\tilde{S}_i)).$$

Then the (inexact) iterate in step $i+1$ can be written as

$$\begin{aligned}
\tilde{S}_{i+1} &= \frac{1}{2}(\tilde{S}_i \oplus \widetilde{\mathrm{Inv}}(\tilde{S}_i)) \quad = \quad \frac{1}{2}(\tilde{S}_i + \widetilde{\mathrm{Inv}}(\tilde{S}_i) + R_i) \\
&= \frac{1}{2}(S_i - E_i + (S_i - E_i)^{-1} + D_i + R_i) \\
&\overset{(5.4)}{=} \frac{1}{2}(S_i - E_i + S_i^{-1}\sum_{\nu=0}^{\infty}(E_i S_i^{-1})^{\nu} + D_i + R_i) \\
&= S_{i+1} + \frac{1}{2}(-E_i + D_i + R_i + \sum_{\nu=1}^{\infty}(E_i S_i^{-1})^{\nu}).
\end{aligned}$$

Using the Definition of c_i and (5.4), (5.5) we get

$$\begin{aligned}
\|\tilde{S}_{i+1} - S_{i+1}\|_2 &\le \frac{1}{2}(c_i(\rho+\delta) + \delta + \rho + \|S_i^{-1}\|_2 \sum_{\nu=1}^{\infty}(c_i(\rho+\delta)\|S_i^{-1}\|_2)^{\nu}) \\
&= \frac{1}{2}(1 + c_i + c_i\|S_i^{-1}\|_2^2 \sum_{\nu=0}^{\infty}(c_i(\rho+\delta)\|S_i^{-1}\|_2)^{\nu})(\rho+\delta) \\
&= \frac{1}{2}(1 + c_i + c_i \frac{\|S_i^{-1}\|_2^2}{1 - c_i(\rho+\delta)\|S_i^{-1}\|_2})(\rho+\delta) \quad = \quad c_{i+1}(\rho+\delta).
\end{aligned}$$

■

Corollary 5.2 *We use the notation from Theorem 5.1 and define*

$$s := \max_{i\in\mathbb{N}_0} \|S_i^{-1}\|_2.$$

We fix a number of iterations $i_{max} \in \mathbb{N}$ and assume

$$\forall i \in \{1, \dots, i_{max}\} \qquad \rho + \delta \le \frac{1}{2}(c_0 + i + 1)^{-2} s^{-4i-3}.$$

Then the error amplification coefficients c_i, $i \in \{0, \dots, i_{max}\}$, can be bounded by

$$c_i \le (c_0 + i)s^{2i}. \tag{5.6}$$

To achieve $\|\tilde{S}_{i_{max}} - S_{i_{max}}\|_2 \le \varepsilon$ one has to take

$$\rho + \delta \le \min\{\varepsilon(c_0 + i_{max})^{-1} s^{-2i_{max}},\ \frac{1}{2}(c_0 + i_{max} + 1)^{-2} s^{-4i_{max}-3}\}.$$

If we assume that the rank k needed to gain a relative error ξ in the $\mathcal{H}$-matrix arithmetic is proportional to $\log(\xi^{-1})$, then the rank k needed to get an overall accuracy of ε (error due to the formatted $\mathcal{H}$-matrix arithmetics and due to the error propagation) is $k = \mathcal{O}(\log(\varepsilon^{-1}) + \log(i_{max}) + i_{max}\log(s))$.

Proof. We prove (5.6) by induction. The case $i = 0$ is obvious. Since $\lim_{i\to\infty} S_i = \lim_{i\to\infty} S_i^{-1}$, we get $s \ge 1$. For the induction step we have to show

$$\frac{1}{2}(1 + c_i + c_i \frac{\|S_i^{-1}\|_2^2}{1 - c_i(\rho+\delta)\|S_i^{-1}\|_2}) \le (c_0 + i + 1)s^{2i+2}.$$

Estimating $\|S_i^{-1}\|_2 \le s$ and multiplying both sides by $1 - c_i(\rho+\delta)s$ it suffices to prove

$$\frac{1}{2}((1 + c_i)(1 - c_i(\rho+\delta)s) + c_i s^2) \le (c_0 + i + 1)s^{2i+2}(1 - c_i(\rho+\delta)s).$$

The left side can be bounded by

$$\frac{1}{2}((1+c_i)(1-c_i(\rho+\delta)s)+c_is^2) \le \frac{1}{2}+\frac{1}{2}c_i+\frac{1}{2}c_is^2 \le \frac{1}{2}+c_is^2 \le \frac{1}{2}+(c_0+i)s^{2i+2},$$

the right-hand side fulfils

$$(c_0+i+1)s^{2i+2}(1-c_i(\rho+\delta)s) \ge 1+(c_0+i)s^{2i+2}-(c_0+i+1)s^{2i+2}c_i(\rho+\delta)s.$$

Therefore, we have to ensure

$$\frac{1}{2} \ge (c_0+i+1)s^{2i+2}c_i(\rho+\delta)s = (c_0+i+1)(c_0+i)s^{4i+3}(\rho+\delta)$$

which is true due to the assumption $\rho+\delta \le \frac{1}{2}(c_0+i+1)^{-2}s^{-4i-3}$.

The bound on the error amplification coefficients $c_{i_{max}}$ yields

$$\|\tilde{S}_i - S_i\|_2 \overset{(5.5)}{\le} (c_0+i_{max})s^{2i_{max}}(\rho+\delta),$$

which gives the last assertion of the corollary. ■

Remark 5.3 *Theorem 5.1 is only a worst case estimate. In practice the error amplification is almost negligible. However, in the literature (see, e.g., [1]) an acceleration technique by scaling is proposed for Newton's method to calculate the sign of a matrix. Iteration (3.4) is replaced by*

$$S_{i+1} := \frac{1}{2}(\alpha S_i + \alpha^{-1}S_i^{-1}), \quad \alpha > 0,$$

which is equivalent to one Newton step with the matrix αS_i instead of S_i (both have the same sign). In the case $\alpha < 1$ the norm of the inverse is amplified by a factor of α^{-1} and estimate (5.6) indicates the consequences: the error is amplified by α^{-2} for each Newton step where the scaling technique is used. Therefore, scaling is only an option in the first step if the initial error is $c_0 = 0$.

From Theorem 5.1 the bound for the distance of the inexact iterates $\tilde{S}_i$ of (5.1) to the exact iterates S_i grows with increasing i, therefore it could happen that Newton's method converges while (5.1) is divergent. The next Lemma ensures at least local quadratic convergence for the inexact iterates.

Lemma 5.4 (Local convergence) *We use the notation from Theorem 5.1, including assumptions (5.2), (5.3), (5.4). We define*

$$\sigma := \max_{i\in\mathbb{N}_0} \|\tilde{S}_i + \tilde{S}_i^{-1}\|_2 \ge 2,$$

and assume

$$q := \|\tilde{S}_0^2 - I\|_2 \le \frac{1}{4}, \tag{5.7}$$

$$\rho+\delta \le \frac{1}{8}\sigma^{-1}. \tag{5.8}$$

Then it holds for $i \in \mathbb{N}$ that

$$\|\tilde{S}_i^2 - I\|_2 \le q^{2^i} + \sigma(\rho+\delta).$$

Proof. The case $i = 0$ is true due to the definition of q. We define

$$\begin{aligned} E_i &:= I - \tilde{S}_i^2, \\ D_i &:= \widetilde{\mathrm{Inv}}(\tilde{S}_i) - \tilde{S}_i^{-1}, \\ R_i &:= (\tilde{S}_i + \widetilde{\mathrm{Inv}}(\tilde{S}_i)) - (\tilde{S}_i \oplus \widetilde{\mathrm{Inv}}(\tilde{S}_i)). \end{aligned}$$

The induction step $i \to i+1$ follows. First we get

$$\begin{aligned}
\tilde{S}_{i+1}^2 &= \left(\frac{1}{2}\tilde{S}_i + \frac{1}{2}\tilde{S}_i^{-1} + \frac{1}{2}(D_i + R_i)\right)^2 \\
&= \frac{1}{4}\tilde{S}_i^2 + \frac{1}{4}\tilde{S}_i^{-2} + \frac{1}{2}I + \frac{1}{2}(\tilde{S}_i + \tilde{S}_i^{-1})(D_i + R_i)^2 + \frac{1}{4}(D_i + R_i)^2 \\
&= \frac{3}{4}I - \frac{1}{4}E_i + \frac{1}{4}\sum_{\nu=0}^{\infty} E_i^\nu + \frac{1}{2}(\tilde{S}_i + \tilde{S}_i^{-1})(D_i + R_i)^2 + \frac{1}{4}(D_i + R_i)^2 \\
&= I + \frac{1}{4}\sum_{\nu=2}^{\infty} E_i^\nu + \frac{1}{2}(\tilde{S}_i + \tilde{S}_i^{-1})(D_i + R_i)^2 + \frac{1}{4}(D_i + R_i)^2,
\end{aligned}$$

$$\|\tilde{S}_{i+1}^2 - I\|_2 \le \frac{1}{4}\frac{\|E_i\|_2^2}{1 - \|E_i\|_2} + \frac{1}{2}\sigma(\rho + \delta) + \frac{1}{4}(\rho + \delta) \le \frac{(q^{2^i} + \sigma(\rho + \delta))^2}{1 - q^{2^i} - \sigma(\rho + \delta)} + \frac{1}{2}\sigma(\rho + \delta) + \frac{1}{4}(\rho + \delta).$$

We have to prove

$$\frac{1}{4}\frac{(q^{2^i} + \sigma(\rho + \delta))^2}{1 - q^{2^i} - \sigma(\rho + \delta)} + \frac{1}{2}\sigma(\rho + \delta) + \frac{1}{4}(\rho + \delta) \le q^{2^{i+1}} + \sigma(\rho + \delta).$$

Multiplying both sides by $1 - q^{2^i} - \sigma(\rho + \delta)$, we get for the left side

$$\mathrm{LS} := \frac{1}{4}(q^{2^i} + \sigma(\rho + \delta))^2 + \frac{1}{2}(1 - q^{2^i} - \sigma(\rho + \delta))\sigma(\rho + \delta) + \frac{1}{4}(1 - q^{2^i} - \sigma(\rho + \delta))(\rho + \delta)$$

and for the right side

$$\mathrm{RS} := (q^{2^{i+1}} + \sigma(\rho + \delta))(1 - q^{2^i} - \sigma(\rho + \delta)).$$

The left side can be estimated by

$$\mathrm{LS} \le \frac{1}{4}q^{2^{i+1}} + \frac{1}{2}\sigma(\rho + \delta) + \frac{1}{4}(\rho + \delta) \le \frac{1}{4}q^{2^{i+1}} + \frac{5}{8}\sigma(\rho + \delta),$$

while the right side can be estimated by

$$\mathrm{RS} = q^{2^{i+1}} - q^{2^i}q^{2^{i+1}} - q^{2^{i+1}}\sigma(\rho + \delta) + \sigma(\rho + \delta) - (\sigma(\rho + \delta))^2 - \sigma(\rho + \delta)q^{2^i} \ge \frac{1}{2}q^{2^{i+1}} + \frac{5}{8}\sigma(\rho + \delta).$$

■

Corollary 5.5 (Stopping criterion) *In Theorem 5.1 we have seen that the attainable accuracy decreases as the number of iterations increases. Therefore, one has to stop the iteration as soon as possible. On the other hand, the convergence is locally quadratic (Lemma 5.4), such that stopping just before the quadratic convergence would lead to an approximate solution that is not even close to the solution. This dilemma can be overcome by the following simple criterion:*

- *In each step of Newton's method calculate the approximate spectral norm $\eta_i := \|\tilde{S}_i^2 - I\|_2$ (the power iteration takes $\mathcal{O}(\log(n))$ steps to determine η_i up to 10 percent relative error).*
- *After each iteration i compute*
$$\theta_i := \frac{\eta_{i-1}}{\eta_i} \quad (\approx q^{-2^i}).$$
- *If the convergence rate θ_i stays smaller than $\frac{3}{2}$, then stop after a fixed number of iterations (e.g., 5). The convergence is dominated by the $\mathcal{H}$-matrix errors ρ, δ.*
- *If the convergence rate θ_i grows larger than $\frac{3}{2}$, then stop if the rate θ_i decreases, that is $\theta_i < \frac{3}{4}\max_{j=1,\dots,i}\theta_j$.*

So far we have investigated how to compute the sign of the matrix $S = \begin{bmatrix} A^T & G \\ F & -A \end{bmatrix}$. The last step is to compute the solution X that solves equation (3.3):

$$\begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix} X = -\begin{bmatrix} N_{12} \\ N_{22} \end{bmatrix}, \qquad \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} := \operatorname{sign}\left(\begin{bmatrix} A^T & G \\ F & -A \end{bmatrix}\right) - \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

According to (4.22), we know

$$N_{11} = -XZ - 2I, N_{12} = 2X + XZX,$$
$$N_{21} = -Z, N_{22} = ZX,$$

where Z satisfies (4.23) and can be approximated by a rank-k_Z matrix $\tilde{Z}$ and $N_{11} \approx -X\tilde{Z} - 2I$. Therefore it is advisable to compute a low rank approximation R_{11} to $(N_{11} + 2I)$ and a low rank approximation R_{22} to N_{22} (this conversion is not expensive because N_{11} and N_{22} are given in the $\mathcal{H}$-matrix format). Equation (3.3) reads

$$\begin{bmatrix} R_{11} - 2I \\ N_{21} \end{bmatrix} X = -\begin{bmatrix} N_{12} \\ R_{22} \end{bmatrix},$$

If $R_{11} - 2I$ is invertible (this is easy to check, because R_{11} is of low rank), one can directly apply the Sherman-Morrison-Woodbury formula to compute

$$X := (2I - R_{11})^{-1} N_{12} = \frac{1}{2} N_{12} + \frac{1}{4} U (I - \frac{1}{2} V^T U)^{-1} V^T N_{12}, \quad UV^T := R_{11}$$

(the invertibility of $R_{11} - 2I$ is equivalent to the invertibility of the $2k_Z \times 2k_Z$-matrix $I - \frac{1}{2}V^T U$). If $R_{11} - 2I$ is not invertible, we solve the normalised equation

$$\begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix}^T \begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix} X = -\begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix}^T \begin{bmatrix} N_{12} \\ N_{22} \end{bmatrix},$$

where the matrix on the left side is

$$M_{\mathrm{LS}} := 4I + 2ZX + 2XZ + ZXXZ + ZZ = 4I + 2ZX + (2X + ZXX + Z)Z$$

and can be approximated by an $\mathrm{R}(2k_Z)$ matrix plus $4I$. Using the low rank approximations R_{11}, R_{22} we get

$$M_{\mathrm{LS}} = 4I - 2R_{11} - 2R_{11}^T + R_{11}^T R_{11} + N_{21}^T N_{21} =: 4I + R_{LS}.$$

The right side is

$$M_{\mathrm{RS}} := -R_{11}^T N_{12} + 2N_{12} - N_{21} R_{22} = Z(2XX + XXZX + 2ZX) + XZX + 4X$$

which can be approximated by an $\mathrm{R}(2k_Z)$ matrix plus $4X$. Using a low rank representation UV^T of R_{LS} we get the solution

$$X := \frac{1}{4} M_{\mathrm{RS}} - \frac{1}{16} U (I + \frac{1}{4} V^T U)^{-1} V^T M_{\mathrm{RS}}.$$

Remark 5.6 (Complexity) *In order to estimate the overall complexity of our method to solve the algebraic matrix Riccati equation, we first summarise the necessary steps of the algorithm:*

1. *Compute the matrices A, F, G. This involves choosing a proper discretisation scheme and computation of the entries. The achievable accuracy (for a fixed blockwise rank) for the solution of the Riccati equation may depend upon the discretisation error.*
2. *Store the matrix $S_0 := \begin{bmatrix} A^T & G \\ F & -A \end{bmatrix}$ in the $\mathcal{H}$-matrix format $\mathcal{M}_{\mathcal{H},k}(T)$ described at the beginning of Subsection 5.1. It may be necessary to convert subblocks of the matrices A, F, G to fit the $\mathcal{H}$-matrix format, but typically the entries of the $\mathcal{H}$-matrix format can be computed directly.*

3. *Compute* $\operatorname{sign}(S_0)$ *by Newton's method* (5.1) , $S_{i+1} := \frac{1}{2}(S_i \oplus \widetilde{Inv}(S_i))$, *using the formatted* $\mathcal{H}$*-matrix arithmetics. The number of iterations necessary depends on the spectrum of the matrix* $A_0 = A - FX$ *and is typically proportional to* $\log(\operatorname{cond}(A))$ *(see Lemma 3.5).*

4. *Solve the equation* (3.3) , *which basically involves the inversion of a low rank perturbation of the identity.*

The complexity for the first two steps depends on the discretisation scheme, but it should be negligible, e.g., proportional to the storage requirements of the $\mathcal{H}$*-matrix* S_0.

In the third step, we have to compute the formatted sum and inverse of $\mathcal{H}$*-matrices in* $\mathcal{M}_{\mathcal{H},k}(T)$ *which is of complexity* $\mathcal{O}(n\log(n)^2k^2)$. *Assuming that the number of iterations is proportional to* $\log(n)$, *this amounts to* $\mathcal{O}(n\log(n)^3k^2)$.

The last step is again negligible.

5.2 Error Estimation

Let $\tilde{X}$ be an approximate solution to (4.21) and denote the exact solution by X. We want to estimate the relative error $\|X-\tilde{X}\|_2/\|X\|_2$ of the approximation, but the exact solution X is not available.

If we define $R(\tilde{X}) := A^T\tilde{X} + \tilde{X}A - \tilde{X}F\tilde{X} + G$ then the difference $Z := \tilde{X} - X$ fulfils the equation

$$(A - F\tilde{X})^T Z + Z(A - F\tilde{X}) - ZFZ - R(\tilde{X}) = 0.$$

So far we have not gained anything, because in order to determine Z (or $\varepsilon := \|Z\|_2/\|X\|_2$) we have to solve again a Riccati equation. The crux is that it is sufficient to determine Z up to a relative error of $\frac{1}{2}$: let $\varepsilon < \frac{1}{2}$ and $\tilde{Z}$ be an approximation to Z with

$$\|\tilde{Z} - Z\|_2 \le \frac{1}{2}\|Z\|_2.$$

Then it holds

$$\begin{aligned} \frac{\|\tilde{Z}\|_2}{\|\tilde{X}\|_2} &\le \frac{\|\tilde{Z}-Z\|_2 + \|Z\|_2}{\|X\|_2 - \|\tilde{X}-X\|_2} \;\le\; \frac{3}{2}(1-\varepsilon)^{-1}\frac{\|Z\|_2}{\|X\|_2} \;\le\; 3\frac{\|Z\|_2}{\|X\|_2}, \\ \frac{\|\tilde{Z}\|_2}{\|\tilde{X}\|_2} &\ge \frac{\frac{1}{2}\|Z\|_2}{(1+\varepsilon)\|X\|_2} \;\ge\; \frac{1}{3}\frac{\|Z\|_2}{\|X\|_2}. \end{aligned}$$

The relative error $\|\tilde{X} - X\|_2/\|X\|_2$ of the approximate solution $\tilde{X}$ is therefore bounded by

$$\frac{1}{3}\frac{\|\tilde{Z}\|_2}{\|\tilde{X}\|_2} \quad\le\quad \frac{\|\tilde{X}-X\|_2}{\|X\|_2} \quad\le\quad 3\frac{\|\tilde{Z}\|_2}{\|\tilde{X}\|_2}.$$

This error estimator leads again to the task of solving a Riccati equation (only up to a relative error of 1/2), but simply taking $\|R(\tilde{X})\|/\|A\|$ or a similar easier computable value does not give a reliable estimate for the relative error.

6 Numerical Examples

The numerical examples in this section serve two purposes: in the one-dimensional example we can compare our results to the ones gained in the literature. Since the structure of the matrices is rather simple, there are plenty of methods available, but many of them depend on the special one-dimensional structure. In the two-dimensional example the matrix G from the Riccati equation (4.21) is not of low rank, the differential operator has "jumping coefficients" and the structure of the matrix A is not as simple as in the one-dimensional case.

6.1 The One-Dimensional Low-Rank Model Problem

We consider the linear quadratic optimal control problem of the one-dimensional heat flow: the goal is to minimise

$$J(u) := \int_0^\infty \left(y(t)^2 + u(t)^2\right) \mathrm{d}t$$

for $u \in L_2(0,\infty)$, where y is defined by the differential equation

$$\begin{array}{lcll} \frac{\partial}{\partial t} x(t,\xi) & = & \frac{\partial^2}{\partial \xi^2} x(t,\xi) + b(\xi)u(t), & \xi \in (0,1),\ t \in (0,\infty), \\ x(t,\xi) & = & 0, & \xi \in \{0,1\}, t \in (0,\infty), \\ x(0,\xi) & = & x_0(\xi), & \xi \in (0,1), \\ y(t) & = & \int_{0.2}^{0.3} x(t,\xi) d\xi, & t \in (0,\infty). \end{array}$$

The starting value $x_0 \in L_2(0,1)$ is given and

$$b(\xi) := \begin{cases} 1 & \xi \in (0.2, 0.3), \\ 0 & \text{otherwise.} \end{cases}$$

The differential equation is discretised by finite differences on a uniform mesh of $(0,1)$ with n inner grid-points and mesh width $h = (n+1)^{-1}$. If we define the matrices

$$A_{ij} := \begin{cases} -2h^{-2} & i = j \\ h^{-2} & |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}, \qquad B_{i1} := \begin{cases} 1 & ih \in [0.2, 0.3] \\ 0 & \text{otherwise} \end{cases},$$

$$C_{1j} := \int_{0.2}^{0.3} \phi_j(x) \mathrm{d}x, \qquad i,j \in \{1,\dots,n\},$$

where ϕ_i denotes the i-th Lagrange basis function for the interpolation, then the discrete problem is the autonomous linear quadratic optimal control problem of Subsection 2.1 with $n_u = n_y = 1$ (this implies $\operatorname{rank}(B) = \operatorname{rank}(C) = 1$).

The $\mathcal{H}$-tree for the index set $I = \{1,\dots,n\}$ is defined in [9, Section 5]. For the iterates of Newton's method to compute the matrix sign function of $S := \left[\begin{smallmatrix} A^T & G \\ F & -A \end{smallmatrix}\right]$ we use the format from Theorem 4.13 depicted in Figure 2 with $\mathcal{M}_{\mathcal{H},k}(T_{I\times I})$-matrices in the two diagonal blocks and R(20)-matrices in the two (largest) off-diagonal blocks. The relative error $\|X - \tilde{X}\|_2 / \|X\|_2$ for the approximate R(20)-matrix solution $\tilde{X}$ can be seen in Table 1 and the singular values of $\tilde{X}$ are depicted in Figure 3. A blockwise singular value decomposition of sign(S) is depicted in Figure 3.

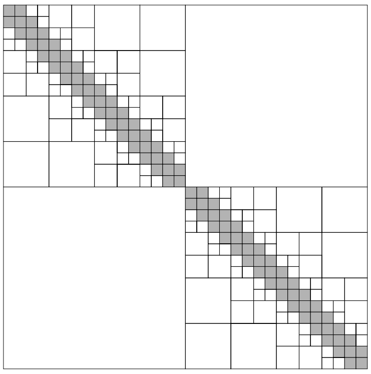

Figure 2: The structure of the matrices S_i. The grey boxes are full matrices, the white squares are R(k)-matrices.

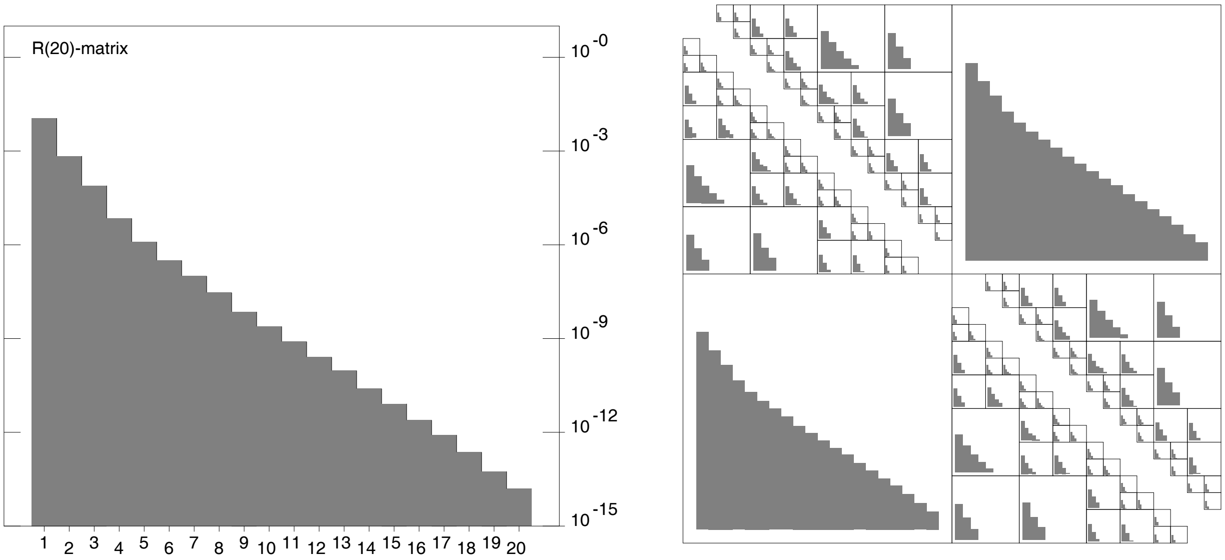

Figure 3: *Left:* the first (largest) 20 singular values of the solution X in logarithmic scale from 10^{-15} to 10^0. *Right:* a blockwise singular value decomposition of $\operatorname{sign}(S)$. Depicted are the first (largest) k singular values of each R(k)-block in logarithmic scale from 10^{-15} to 10^0 as in the left image. Singular values smaller than 10^{-15} are plotted as 10^{-15}.

If we use full matrices instead of $\mathcal{H}$-matrices in the two diagonal blocks and R(20)-matrices in the two (largest) off-diagonal blocks, then we would need about 15 years to compute the solution for $n = 65536$ degrees of freedom (by use of the matrix sign function) on a SUN QUASAR. The complexity is cubic in n.

The results by Rosen and Wang [18] computed on a SUN SPARC 600 for $n = 101$ degrees of freedom took 2062 seconds. This extrapolates to approximately 27 years for $n = 65536$ degrees of freedom, which is equivalent to roughly one year on a SUN QUASAR. The complexity is quadratic in n.

For $n = 65536$ degrees of freedom we need about 6 hours to compute the solution X up to a relative error of 10^{-3} (cf. Table 1). The complexity is almost linear in n.

The storage requirements for the (exact) solution for $n = 65536$ degrees of freedom would be 35 Gigabytes, and due to the quadratic dependency on n we are not able to store or compute the exact solution X for larger n. An R(20) representation of X takes only 20 Megabytes (linear dependency on n).

Penzl [16] computed a low rank approximation $\tilde{X}$ to the solution of the Lyapunov equation (4.19) with matrices A, G similar to the ones in our example. The work for $n = 10000$ degrees of freedom amounted to 10^8 flops, whereas the exact solution by the Bartels-Stewart algorithm would take 10^{13} flops. Since the CPU time is not mentioned, we cannot directly compare those results to the ones from our method.

6.2 The Two-Dimensional $\mathcal{H}$-Matrix Model Problem

In the previous section we have compared our method to existing methods for a low rank approximation of the solution X to the Riccati equation. Now, we want to give an example where the matrix G in the Riccati equation is *not* of low rank, but an $\mathcal{H}$-matrix of full (global) rank. While our method can exploit the $\mathcal{H}$-matrix structure and computes an approximation to the solution with linear-logarithmic complexity in the size of the matrices, there are no known algorithms in the literature that can achieve a similar efficiency.

rank k	number of degrees of freedom n				
	256	1024	4096	16384	65536
k=1	1.5_{10}-1	1.3_{10}-1	2.5_{10}-0	divergent	divergent
k=2	2.6_{10}-4	4.2_{10}-4	1.2_{10}-3	5.6_{10}-4	6.7_{10}-4
k=3	1.2_{10}-5	1.3_{10}-5	1.5_{10}-5	2.3_{10}-5	3.9_{10}-5
k=4	9.1_{10}-8	1.1_{10}-7	1.0_{10}-6	1.8_{10}-6	6.2_{10}-7
k=5	4.6_{10}-9	1.1_{10}-8	1.5_{10}-8	3.0_{10}-8	3.1_{10}-8
k=6	3.7_{10}-10	2.4_{10}-10	4.9_{10}-10	5.9_{10}-10	1.7_{10}-9
Newton steps	14	17	20	23	26
time [sec.], k=2	8.5	67	462	3033	18263
time [sec.], full	17.7	1814	$\approx 1.1_{10}5$	$\approx 7.4_{10}6$	$\approx 4.8_{10}8$

Table 1: The table contains the relative error $\varepsilon := \|\tilde{X} - X\|_2 / \|X\|_2$ for increasing rank k and n degrees of freedom. The number of Newton steps to compute $\operatorname{sign}(S)$ are $\frac{5}{3}\log_2(n)$. In the last two rows we compare the time in seconds (on a SUN QUASAR with 450 Mhz) needed to compute the solution for the $\mathcal{H}$-matrix approach (rank $k = 2$) and the full matrix approach.

The following example is the two-dimensional optimal control of the heat equation. The goal is to minimise

$$J(u) := \int_0^\infty \left(y(t)^2 + u(t)^2\right) \mathrm{d}t$$

for $u \in L_2(0,\infty)$ where y is defined by the differential equation

$$\begin{array}{rcll}
\frac{\partial}{\partial t}x(t,\xi) & = & \left(\frac{\partial_1}{\partial\xi_1} + \frac{\partial_2}{\partial\xi_2}\right)\left(\sigma(\xi)\left[\begin{array}{c}\frac{\partial}{\partial\xi_1}x(t,\xi)\\ \frac{\partial}{\partial\xi_2}x(t,\xi)\end{array}\right]\right) + b(\xi)u(t), & \xi \in (0,1)^2, t \in (0,\infty),\\
x(t,\xi) & = & 0, & \xi \in [0,1]^2 \setminus (0,1)^2, t \in (0,\infty),\\
x(0,\xi) & = & x_0(\xi), & \xi \in (0,1)^2,\\
y(t) & = & \left(\int_{(0,1)^2} x(t,\xi)^2 \mathrm{d}\xi\right)^{1/2}, & t \in (0,\infty).
\end{array}$$

The starting value $x_0 \in L_2((0,1)^2)$ is given and the functions b, σ are (see Figure 4)

$$\sigma(\xi) := \left\{\begin{array}{ll} 10 & \xi \in [0,1] \times [\frac{3}{8},\frac{5}{8}] \\ 0.1 & \xi \in [\frac{3}{8},\frac{5}{8}] \times \left([0,\frac{3}{8}) \cup (\frac{5}{8},1]\right) \\ 1.0 & \text{otherwise} \end{array}\right., \qquad b(\xi) := \left\{\begin{array}{ll} 1 & \xi \in [0,\frac{1}{8}] \times [\frac{3}{8},\frac{5}{8}] \\ 0 & \text{otherwise} \end{array}\right. .$$

The differential equation is discretised (in the weak or variational formulation) using the space of nodal affine finite elements on a uniform triangulation of $(0,1)^2$ with n inner grid points. We denote the (Lagrange) basis functions by ϕ_i $(i = 1,\dots,n)$. The resulting discrete problem is the autonomous linear quadratic optimal control problem from Section 2 with $n_u = 1, n_y = n$ and matrices A, B, C defined as follows. The entries of the mass matrix and the stiffness matrix are

$$E_{ij} := \int_{(0,1)^2} \phi_i(\xi)\phi_j(\xi)d\xi, \qquad \tilde{A}_{ij} := \int_{(0,1)^2} \sigma(\xi)\langle\nabla\phi_j(\xi), \nabla\phi_i(\xi)\rangle d\xi$$

for $i, j \in \{1,\dots,n\}$. Both E and $\tilde{A}$ are symmetric positive definite, but $\tilde{A}$ is ill-conditioned. The entries of the discrete right-hand side $\tilde{B}$ are

$$\tilde{B}_{i1} := \int_{(0,1)^2} b(\xi)\phi_i(\xi)d\xi, \qquad i = 1,\dots,n.$$

Finally, the matrices A, B, C are

$$A := -E^{-1}\tilde{A}, \quad B := E^{-1}\tilde{B}, \quad C := E^{1/2}.$$

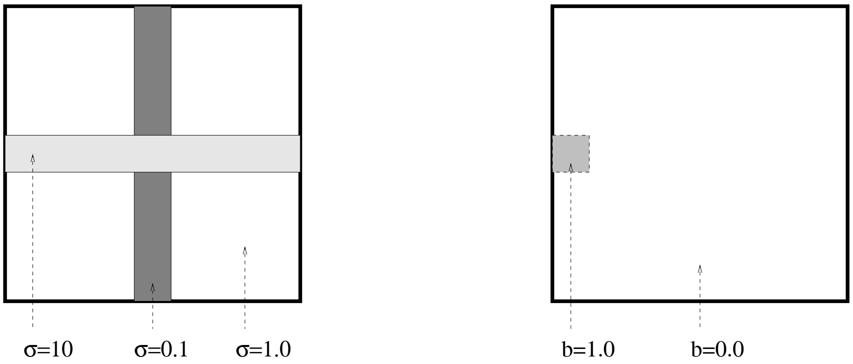

Figure 4: The functions b and σ on $(0,1)^2$.

We store the matrices $\tilde{A}$ and $G = C^T C = E$ in the $\mathcal{H}$-matrix format based on the $\mathcal{H}$-tree $T_{I\times I}$ that was established for this two-dimensional uniform triangulation of $[0,1]^2$ in [10]. Similarly, appropriate $\mathcal{H}$-trees can be constructed for arbitrary triangulations (see [7]).

The singular values of the solution X are depicted in Figure 5, where one can see that the singular values do not decay rapidly as it was the case in the one-dimensional example with low rank G.

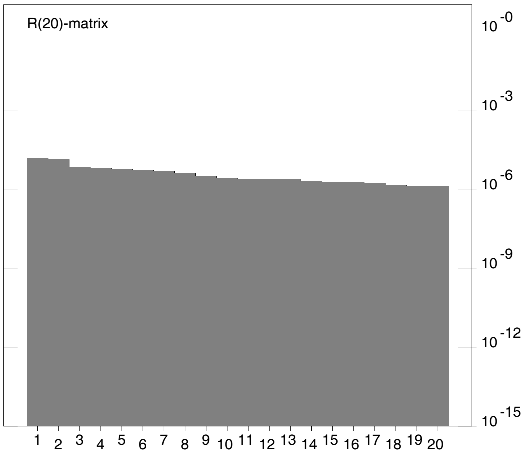

Figure 5: The first 20 singular values of the solution X in logarithmic scale from 10^{-15} to 10^0.

For the iterates of Newton's method to compute the matrix sign function of $S := \left[\begin{smallmatrix} A^T & G \\ F & -A \end{smallmatrix}\right]$ we use the format from Theorem 4.13 depicted in Figure 6 with $\mathcal{M}_{\mathcal{H},k}(T_{I\times I})$-matrices in three of the blocks of S and an R($2k$)-matrix in the lower left block. The relative error $\|X - \tilde{X}\|_2 / \|X\|_2$ for the approximate $\mathcal{H}$-matrix solution $\tilde{X}$ can be seen in Table 2 and a blockwise singular value decomposition of X is depicted in Figure 6.

The mass matrix E has to be inverted, which is done in the set $\mathcal{M}_{\mathcal{H},k}(T_{I\times I})$. By taking the formatted inverse $\widetilde{Inv}(E)$ instead of E^{-1} we introduce an error (besides the discretisation error), but since E is well conditioned this error is rather small (see [3]). The matrix $F = BB^T$ is of rank 1 but again we use the formatted inverse $\widetilde{Inv}(E)$ to define an approximation to F.

We compute an approximation $\tilde{X}$ to the solution X of (4.21) as described in Section 5 with rank k for the blockwise rank of the $\mathcal{H}$-matrices and rank $2k$ for the rank of the lower left block that corresponds to the

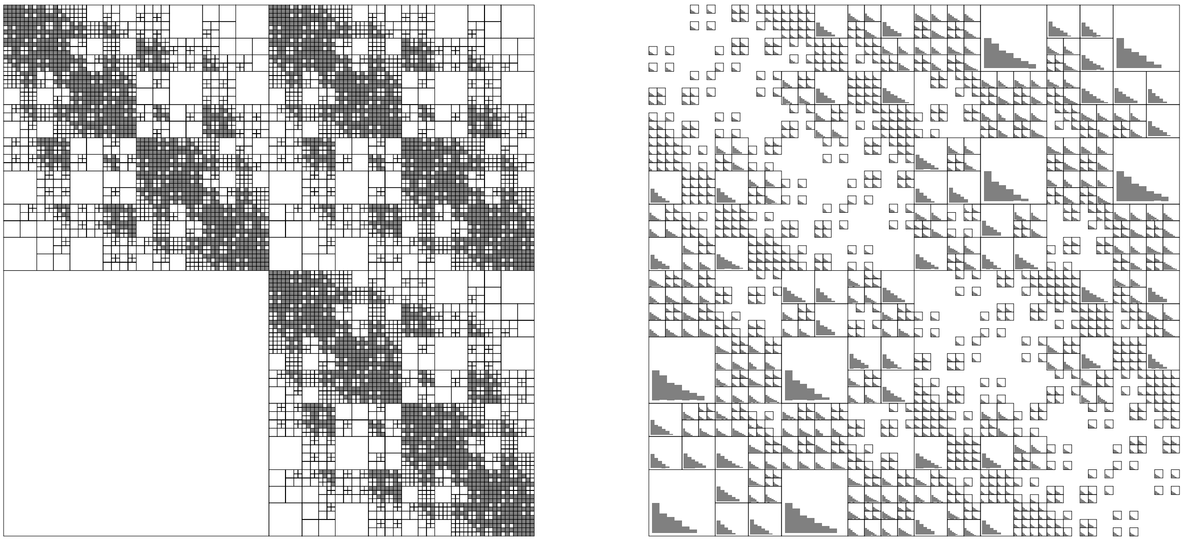

Figure 6: *Left:* the structure of the matrices S_i. The grey boxes are full matrices, the white squares are R(k)-matrices. *Right:* a blockwise singular value decomposition of X. Depicted are the first (largest) 8 singular values of each R(k)-block in logarithmic scale from 10^{-15} to 10^0 as in Figure 3. Singular values smaller than 10^{-15} are plotted as 10^{-15}.

iterates F_i. The results for $k = 1, \dots, 7$ can be seen in Table 2, where we estimate the relative error ε by

$$\varepsilon = \|X - \tilde{X}\|_2 / \|X\|_2 \approx \|X_{\mathcal{H}} - \tilde{X}\|_2 / \|X_{\mathcal{H}}\|_2$$

for an $\mathcal{H}$-matrix approximation $X_{\mathcal{H}}$ computed with rank $k = 8$.

rank k	number of degrees of freedom n			
	256	1024	4096	16384
k=1	4.1$_{10}$-3	1.5$_{10}$-2	2.0$_{10}$-2	7.2$_{10}$-2
k=2	1.6$_{10}$-4	2.1$_{10}$-3	7.6$_{10}$-3	2.3$_{10}$-2
k=3	7.4$_{10}$-5	3.4$_{10}$-4	1.6$_{10}$-3	8.1$_{10}$-3
k=4	1.4$_{10}$-5	1.1$_{10}$-4	4.0$_{10}$-4	6.0$_{10}$-4
k=5	3.1$_{10}$-6	2.2$_{10}$-5	2.0$_{10}$-4	2.6$_{10}$-4
k=6	8.4$_{10}$-7	6.9$_{10}$-6	5.2$_{10}$-5	7.5$_{10}$-5
k=7	5.7$_{10}$-7	1.6$_{10}$-6	1.2$_{10}$-5	2.0$_{10}$-5
Newton steps	10	11	12	13
time [sec]	20	570	5784	38613

Table 2: The Table presents the relative error ε. Last but one row: number of Newton steps to compute sign(S). Last row: time in seconds to compute the rank $k = 2$ solution on a SUN QUASAR with 450 Mhz.

References

[1] Z. Bai, J. Demmel: *Design of a parallel nonsymmetric eigenroutine toolbox, Part I,* Proc. of the 6th SIAM Conf. on parallel processing for scientific computing, eds. R. F. Sincovec et al. (1993).

[2] R. H. Bartels, G. W. Stewart: *Solution of the equation $AX + XB = C$,* Comm. ACM 15, 820–26 (1972).

[3] M. Bebendorf, W. Hackbusch: *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients,* Report 21, Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig 2002.

[4] G. H. Golub, C. F. Van Loan: *Matrix Computations,* Johns Hopkins University Press (1996).

[5] I. Gavrilyuk, W. Hackbusch, B. N. Khoromskij: *$\mathcal{H}$-matrix approximation for the operator exponential with applications,* Numer. Math., to appear.

[6] L. Grasedyck: *Theorie und Anwendungen Hierarchischer Matrizen,* PhD thesis, University Kiel, Germany, 2001.

[7] L. Grasedyck, W. Hackbusch: *Construction and arithmetics of $\mathcal{H}$-matrices,* in preparation.

[8] L. Grasedyck: *Existence of a low rank or $\mathcal{H}$-matrix approximant to the solution of a Sylvester equation,* Preprint No. 2 (2002), University Kiel, Germany.

[9] W. Hackbusch: *A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part I: Introduction to $\mathcal{H}$-matrices,* Computing 62 (1999), 89-108.

[10] W. Hackbusch, B. Khoromskij: *A sparse $\mathcal{H}$-matrix arithmetic. Part II: Application to multi-dimensional problems,* Computing 64 (2000), 21-47.

[11] W. Hackbusch, B. Khoromskij: *A sparse $\mathcal{H}$-matrix arithmetic: general complexity estimates,* J. Comp. Appl. Math. 125 (2000), 479-501.

[12] D. L. Kleinman: *On an iterative technique for Riccati equations computation,* IEEE Trans. Automat. Control, AC-13 (1968), 114-115.

[13] H. W. Knobloch, H. Kwakernaak: *Lineare Kontrolltheorie,* Springer-Verlag Berlin, Heidelberg, New York (1985).

[14] P. Lancaster: *Explicit solutions of linear matrix equations,* SIAM Review 12 (1970), 544-566.

[15] T. Penzl: *Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case,* Systems and Control Letters 40 (2000), 139-144.

[16] T. Penzl: *A cyclic low rank Smith method for large sparse Lyapunov equations with applications in model reduction and optimal control,* Preprint 98-6, TU Chemnitz, 1998.

[17] J. D. Roberts: *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function,* Internat. J. Control 32 (1980), 677-687.

[18] J. I. G. Rosen, C. Wang: *A multilevel technique for the approximate solution of operator Lyapunov and algebraic Riccati equations,* Siam J. Numer. Anal. 32 (1995), 514-541.

[19] D. L. Russell: *Mathematics of finite dimensional control systems: Theory and design,* Lecture Notes in Pure and Applied Mathematics, 43, Marcel Dekker, New York, 1979.

[20] F. Stenger: *Numerical methods based on Sinc and analytic functions,* Springer, New York 1993.

Lars Grasedyck, Wolfgang Hackbusch, Boris N. Khoromskij
Max-Planck-Institute for Mathematics in the Sciences
Inselstr. 22-26
D-04103 Leipzig
Germany
{lgr,wh,bokh}@mis.mpg.de