SIAM J. MATRIX ANAL. APPL.
Vol. 39, No. 4, pp. 1564–1590




# SOLVING RANK-STRUCTURED SYLVESTER AND LYAPUNOV EQUATIONS*

STEFANO MASSEI†, DAVIDE PALITTA‡, AND LEONARDO ROBOL§

**Abstract.** We consider the problem of efficiently solving Sylvester and Lyapunov equations of medium and large scale, in case of rank-structured data, i.e., when the coefficient matrices and the right-hand side have low-rank off-diagonal blocks. This comprises problems with banded data, recently studied in [A. Haber and M. Verhaegen, *Automatica J. IFAC*, 73 (2016), pp. 256–268; D. Palitta and V. Simoncini, *Numerical Methods for Large-Scale Lyapunov Equations with Symmetric Banded Data*, preprint, arxiv, 1711.04187, 2017], which often arise in the discretization of elliptic PDEs. We show that, under suitable assumptions, the quasiseparable structure is guaranteed to be numerically present in the solution, and explicit novel estimates of the numerical rank of the off-diagonal blocks are provided. Efficient solution schemes that rely on the technology of hierarchical matrices are described, and several numerical experiments confirm the applicability and efficiency of the approaches. We develop a MATLAB toolbox that allows easy replication of the experiments and a ready-to-use interface for the solvers. The performances of the different approaches are compared, and we show that the new methods described are efficient on several classes of relevant problems.

**Key words.** Sylvester equation, Lyapunov equation, banded matrices, quasiseparable matrices, off-diagonal singular values, $\mathcal{H}$-matrices

**AMS subject classifications.** 15A06, 15A24, 65D32, 65F10, 93C20

**DOI.** 10.1137/17M1157155

## 1. Introduction.

We consider the problem of solving Sylvester equations of the form

$$AX + XB = C, \tag{1}$$

where $A \in \mathbb{R}^{n_A \times n_A}$, $B \in \mathbb{R}^{n_B \times n_B}$, $C \in \mathbb{R}^{n_A \times n_B}$, and $A$, $B$ are symmetric positive definite and rank-structured. More precisely, we assume that the matrices $A$, $B$, and $C$ are *quasiseparable*, i.e., their off-diagonal blocks have low rank. For the sake of simplicity, throughout the paper we assume $C$ to be square, that is, $n_A = n_B \equiv n$, but our results can be easily extended to the case of different $n_A$ and $n_B$.

Sylvester equations arise in different settings, such as problems of control [1, 9], discretization of PDEs [16, 34], block-diagonalization [20, Chapter 7.1.4], and many others. The Lyapunov equation, that is, (1) with $B = A$, is of particular interest due to its important role in control theory [1]. The symmetric and positive definite constraint is not strictly necessary in our analysis, and some relaxations involving the field of values will be presented.

Even in the case of sparse $A$, $B$, and $C$, the solution $X$ to (1) is, in general, dense, and it cannot be easily stored for large-scale problems. To overcome this numerical

*Received by the editors November 15, 2017; accepted for publication (in revised form) by D. Szyld August 20, 2018; published electronically October 23, 2018.
http://www.siam.org/journals/simax/39-4/M115715.html
**Funding:** This work was supported by the GNCS/INdAM project "Metodi numerici avanzati per equazioni e funzioni di matrici con struttura" by the Region of Tuscany (PAR-FAS 2007–2013), and by MIUR, the Italian Ministry of Education, Universities and Research (FAR) within the Call FAR–FAS 2014 (MOSCARDO Project: ICT technologies for structural monitoring of age-old constructions based on wireless sensor networks and drones, 2016–2018).
†EPF, Lausanne, Switzerland (stefano.massei@epfl.ch).
‡Dipartimento di Matematica, Università di Bologna, Bologna, Italy (davide.palitta3@unibo.it).
§ISTI-CNR, Pisa, Italy (leonardo.robol@isti.cnr.it).

difficulty, the right-hand side is often supposed to be low rank, i.e., $C = C_1 C_2^T$ with $C_1, C_2 \in \mathbb{R}^{n\times k}$, $k \ll n$. In this case, under some suitable assumptions on the spectra of $A$ and $B$, it is possible to prove that the solution $X$ is numerically low rank [2, 7, 21, 36] so that it can be well approximated by a low-rank matrix $X \approx UV^T$. The low-rank property of $X$ justifies the solution of these kinds of equations by the so-called low-rank methods, which directly compute and store only the factors $U$, $V$. A large amount of work in this direction has been carried out in recent years; see, e.g., [42] and the references therein. However, in many cases the known term $C$ is not low rank. It is very easy to construct a simplified example to show that low-rank methods have no hope of being effective in this more general context. Consider (1) with $A = B = I$ and $C = 2I$ where $I$ denotes the identity matrix. It is immediate to check that the solution is $X = I$, and therefore every approximation $UV^T \approx X$ which is not full rank needs to satisfy $\|UV^T - X\|_2 \geqslant 1$. Obviously, this example has no practical relevance from the computational point of view, since a Lyapunov equation with diagonal data needs to have a diagonal solution, which can be computed in $O(n)$ time and represented in $O(n)$ storage. Nevertheless, it shows that even if all the coefficients and the solution $X$ are full rank, they can indeed be very structured. One might wonder if also banded structures are preserved. This is not true, in general, since banded matrices are not an algebra (in contrast to what is true for diagonal ones), but approaches which exploit the banded properties of $A$, $B$, $C$ and, to a certain extent, of the solution $X$, have been recently proposed by Haber and Verhaegen in [25] and by Palitta and Simoncini in [35]. The preservation of a banded structure in the solution is strictly connected with the conditioning of $A$ and $B$. Unless they are both ill-conditioned, the solution $X$ of (1) is well approximated by a banded matrix $\widetilde{X}$. Otherwise, it has been shown that $X$ can be represented by a couple $(X_B, S_m)$, $X \approx X_B + S_m S_m^T$, where $X_B$ is banded and $S_m$ is low rank so that a low memory allocation is still required; see [35].

In this work, we consider a more general structure, the so-called quasiseparability, which is often numerically present in $X$ when we have it in $A$, $B$, and $C$, so that a low memory requirement is demanded for storing the solution. Informally, a matrix is said to be quasiseparable if its off-diagonal blocks are low-rank matrices, and the quasiseparable rank is defined as the maximum of the ranks of the off-diagonal blocks. We say that a matrix is numerically quasiseparable when the above property holds only up to a certain $\epsilon$, i.e., only few singular values of each off-diagonal block are above a fixed threshold.

A simple yet meaningful example arises from the context of PDEs: consider the differential equation

$$\begin{cases} -\Delta u = \log\left(\tau + |x-y|\right), & (x,y) \in \Omega, \\ u(x,y) \equiv 0, & (x,y) \in \partial\Omega, \end{cases} \qquad \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \tag{2}$$

where $\Omega$ is the rectangular domain $[0,1] \times [0,1]$ and $\tau > 0$. The discretization by centered finite differences of (2) with $n$ nodes in each direction, $(x_i, y_j)$, $i, j = 1, \ldots, n$, yields the following Lyapunov equation:

$$AX + XA = C, \qquad A, C \in \mathbb{R}^{n\times n},$$

$$C_{i,j} = \log\left(\tau + |x_i - y_j|\right), \qquad A = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

$$h := \tfrac{1}{n-1},$$

The fact that $A$ is banded implies that it is quasiseparable, and also $C$ shares this

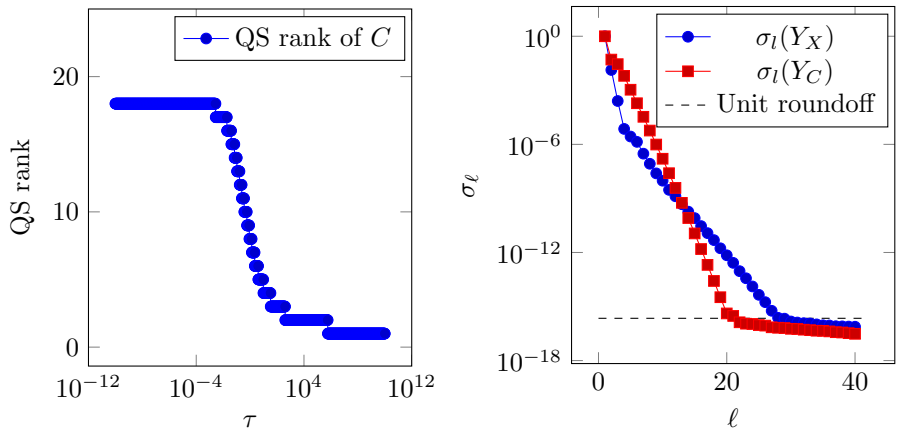


FIG. 1. *On the left, the maximum numerical ranks of the off-diagonal blocks of the right-hand side $C$ for different values of $\tau$ and $n = 300$, using a threshold of $10^{-14}$ for truncation. On the right, we set $\tau = 10^{-4}$, and the singular values of the off-diagonal blocks $Y_C := C(\frac{n}{2}+1:n, 1:\frac{n}{2})$ and $Y_X := X(\frac{n}{2}+1:n, 1:\frac{n}{2})$ rescaled by the 2-norm of the two blocks, respectively, are reported. The black dashed line indicates the machine precision $2.22 \cdot 10^{-16}$.*

property. Indeed, this follows from the fact that the modulus function it is not regular in the whole domain, but it is analytic when the sign of $x-y$ is constant. This happens in the subdomains corresponding to the off-diagonal blocks. Separable approximation (and thus low rank) can be obtained by expanding the source $\log(\tau + |x+y|)$ in the Chebyshev basis. The approximation of these kinds of functions has been previously investigated in [28, Chapter 9]. In Figure 1 (on the right) we have reported the decay of the singular values of one off-diagonal block of $C$ and $X$ for the case of $\tau = 10^{-4}$ and $n = 300$. In this case the numerical quasiseparable rank of the right-hand side $C$ and the solution $X$ does not exceed 20 and 30, respectively. This property holds for any $\tau > 0$: in Figure 1 (on the left) we have checked the quasiseparable rank of the matrix $C$ for various values of $\tau$, and one can see that it is uniformly bounded. The rank is higher when $\tau$ is small, because the function is "less regular," and tends to 1 as $\tau \to \infty$, because the off-diagonal blocks tend to a constant in this case.

The problem of solving linear matrix equations whose coefficients are represented as $\mathcal{H}$-matrices has already been addressed in [21, 22]. In [5], the authors consider the case of Lyapunov equations with $\mathcal{H}$-matrix coefficients and low-rank right-hand side. Recently, in [10, 11] the use of hierarchical matrices in the cyclic reduction iteration for solving quadratic matrix equations has been deeply studied. We will exploit the framework of $\mathcal{H}$-matrices to store quasiseparable matrices and to perform matrix operations at an almost linear cost (up to logarithmic factors).

In this paper, we compare the use of hierarchical matrices in the matrix sign iteration and in the estimation of an integral formula for solving (1). The latter approach, suggested but not numerically tested in [21, 22], relies on evaluating the closed formula [40]

$$X = \int_0^{+\infty} e^{-At} C e^{-Bt} dt \tag{3}$$

by combining a numerical integrating scheme and rational approximations for the matrix exponential. We employ (3) for our purpose, but different closed forms of $X$

are available in the literature. See, e.g., [42]. Starting with $\mathcal{H}$-matrix representations of $A$, $B$, and $C$, formula (3) can be efficiently approximated exploiting $\mathcal{H}$-arithmetic. To the best of our knowledge, this technique has been exploited only theoretically for computing $X$ in the $\mathcal{H}$-matrix framework. On the other hand, exponential sums are widely used as an approximation tool in the solution of tensor Sylvester equations [14, 15].

The representation (3) has already been used in [21, 22] as a theoretical tool to estimate the quasiseparable rank of the solution, but the derived bounds may be very pessimistic and are linked with the convergence of the integral formula, which cannot be easily made explicit. We improve these estimates by developing a theoretical analysis which relies on some recent results [7], exploited also in [11], where the numerical rank of the solution $X$ is determined by estimating the exponential decay in the singular values of its off-diagonal blocks.

The paper is organized as follows; in section 2 we introduce the notion of quasiseparability and we deliver the technical tools for analyzing the preservation of the structure in the solution $X$. In particular, we provide bounds for the off-diagonal singular values of $X$ and we show some numerical experiments in order to validate them. In section 3, hierarchically off-diagonal low-rank (HODLR) matrices are introduced and their impact on the computational effort for handling matrix operations is described. The two algorithms for solving (1) are presented in section 4. In particular, in section 4.1 we recall the sign function method presented in [22], whereas the procedure used for the numerical approximation of (3) is illustrated in section 4.2. Both approaches are based on the use of HODLR arithmetic. We address the solution of certain generalized Lyapunov and Sylvester equations in section 5. In section 6 we perform numerical tests on instances of (1) coming from both artificially crafted models and real-world problems where the quasiseparable structure is present. Finally, in section 7 we draw some concluding remarks.

**2. Quasiseparable structure in the solution.** The main purpose of this section is to prove that, under some reasonable assumptions on the spectrum of $A$ and $B$, the solution $X$ to the matrix equation (1) needs to be quasiseparable if $A$, $B$, and $C$ are quasiseparable. Throughout the paper we indicate with $\sigma_1(M) \leq \sigma_2(M) \leq \dots$ the ordered singular values of the matrix $M$.

**2.1. Quasiseparability structures.** The literature on quasiseparable (or semiseparable) matrices—see Figure 2—is rather large, and the term is often used to denote slightly different objects. Therefore, also in the spirit of making this paper as self-contained as possible, we recall the definition of quasiseparable matrices that we will use throughout the paper. We refer the reader to [19, 46, 47, 48] and the references therein for a complete survey about quasiseparable and semiseparable structures.

DEFINITION 2.1. *A matrix $A$ is* quasiseparable *of order $k$ if the maximum of the ranks of all its submatrices contained in the strictly upper or lower part is less than or equal to $k$.*

*Example* 2.2. A banded matrix—see Figure 3—with bandwidth $k$ is quasiseparable of order (at most) $k$. In particular, diagonal matrices are quasiseparable of order 0, tridiagonal matrices are quasiseparable of order 1, and so on.

**2.2. Zolotarev problems and off-diagonal singular values.** We are interested in exploiting the quasiseparable rank in numerical computations. In many cases,

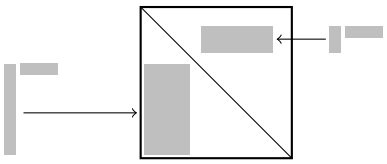

FIG. 2. *Pictorial description of the quasiseparable structure; the off-diagonal blocks can be represented as low-rank outer products.*

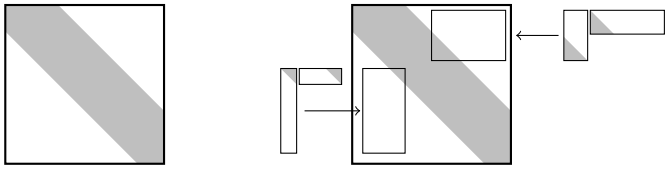

FIG. 3. *Graphic description of the quasiseparability of banded matrices; in grey, the nonzero entries.*

the request of the exact preservation of a certain structure is too strong, and it cannot be guaranteed. However, for computational purposes, we are satisfied if the property holds in an approximate way, i.e., if our data are well approximated by structured ones. This can be rephrased by asking that the off-diagonal blocks of the solution $X$ of (1) have a low numerical rank. More precisely, given a generic off-diagonal block of the sought solution $X$, we want to prove that only a limited number of its singular values are larger than $\epsilon \cdot \|X\|_2$, where $\epsilon$ is a given threshold. This kind of analysis has been already performed in [10, 11, 33] for studying the numerical preservation of quasiseparability when solving quadratic matrix equations and computing matrix functions. See also the Ph.D. thesis [32] for more details.

In order to formalize this approach, we extend a result that provides bounds for the singular values of the solution of (1) when the right-hand side has low rank. The latter is based on an old problem considered by Zolotarev at the end of the 19th century [50], which concerns rational approximation in the complex plane. The following version can be found, along with the proof, in [7, Theorem 2.1] or in a similar form in [11, Theorem 4.2].

THEOREM 2.3. *Let $X$ be an $n \times n$ matrix that satisfies the relation $AX + XB = C$, where $C$ is of rank $k$ and $A, B$ are normal matrices. Let $E, F$ be two disjoint sets containing the spectra of $A$ and $-B$, respectively. Then, the following upper bound on the singular values of $X$ holds:*

$$\frac{\sigma_{1+k\ell}(X)}{\sigma_1(X)} \leqslant Z_\ell(E,F) := \inf_{r(x)\in\mathcal{R}_{\ell,\ell}} \frac{\max_{x\in E}|r(x)|}{\min_{y\in F}|r(y)|}, \qquad \ell \geqslant 1,$$

*where $\mathcal{R}_{\ell,\ell}$ is the set of rational functions of degree at most $(\ell, \ell)$.*

Theorem 2.3 provides useful information only if one manages to choose the sets $E$ and $F$ as well separated. In general it is difficult to explicitly bound $Z_\ell(E, F)$, but some results exist for specific choices of domains, especially when $E$ and $F$ are real intervals; see, for instance, [7, 23]. The combination of these results with Theorem 2.3 proves the well-known fact that a Sylvester equation with positive definite coefficients and with a low-rank right-hand side has a numerically low-rank solution.

LEMMA 2.4. *Let $A, B$ be symmetric positive definite matrices with spectra contained in $[a, b]$, $0 < a < b$. Consider the Sylvester equation $AX + XB = C$, with $C$ of rank $k$. Then the solution $X$ satisfies*

$$\frac{\sigma_{1+k\ell}(X)}{\sigma_1(X)} \leqslant 4\rho^{-2\ell},$$

*where $\rho = \exp(\frac{\pi^2}{2\mu(\frac{b}{a})})$ and $\mu(\cdot)$ is the Grötzsch ring function*

$$\mu(\lambda) := \frac{\pi}{2}\frac{K(\sqrt{1-\lambda^2})}{K(\lambda)}, \qquad K(\lambda) := \int_0^1 \frac{1}{(1-t^2)(1-\lambda^2 t^2)}dt.$$

*Proof.* Applying Theorem 2.3 with $E = [a, b]$ and $F = [-b, -a]$, we get

$$\frac{\sigma_{1+k\ell}(X)}{\sigma_1(X)} \leqslant Z_\ell(E, F).$$

Using Corollary 3.2 in [7] for bounding $Z_\ell(E, F)$, we get the claim. ∎

*Remark* 2.5. A slightly weaker bound which does not involve elliptic functions is the following [7]:

$$Z_\ell([a, b], [-b, -a]) \leqslant 4\rho^{-2\ell}, \qquad \rho = \exp\left(\frac{\pi^2}{2\log\left(4\frac{b}{a}\right)}\right), \qquad 0 < a < b < \infty.$$

It is easy to see that in case of Lyapunov equations with symmetric positive definite coefficients we can replace the quantity $\frac{b}{a}$ with the condition number of $A$.

COROLLARY 2.6. *Let $A$ be a symmetric positive definite matrix with condition number $\kappa_A$, and consider the Lyapunov equation $AX + XA = C$, with $C$ of rank $k$. Then the solution $X$ satisfies*

$$\frac{\sigma_{1+k\ell}(X)}{\sigma_1(X)} \leqslant 4\rho^{-2\ell},$$

*where $\rho = \exp(\frac{\pi^2}{2\mu(\kappa_A)})$ and $\mu(\cdot)$ is defined as in Lemma 2.4.*

We are now interested in proving that the solution of a Sylvester equation with low-order quasiseparable data is numerically quasiseparable. An analogous task was addressed in [21]. The approach developed by the authors can be used for estimating either the rank of $X$ in the case of a low-rank right-hand side or the rank of the off-diagonal blocks of $X$ when the coefficients are hierarchical matrices. In particular, it has been shown that if the coefficients are efficiently represented by means of the hierarchical format, then the solution also shares this property. The estimates provided in [21] exploit the convergence of a numerical integrating scheme for evaluating the closed integral formula (3). These bounds are however quite implicit and are more pessimistic than the estimates provided in [36] and in [43] for the case of a low-rank right-hand side (which is the setting where all the previous results are applicable).

Here, we directly characterize the off-diagonal singular values of the solution applying Theorem 2.3 blockwise.

THEOREM 2.7. *Let $A$ and $B$ be symmetric positive definite matrices of quasiseparable rank $k_A$ and $k_B$, respectively, and suppose that the spectra of $A$ and $B$ are both*

*contained in the interval* $[a, b]$. *Then, if* $X$ *solves the Sylvester equation* $AX + XB = C$, *with* $C$ *of quasiseparable rank* $k_C$, *a generic off-diagonal block* $Y$ *of* $X$ *satisfies*

$$\frac{\sigma_{1+k\ell}(Y)}{\sigma_1(Y)} \leqslant 4\rho^{-2\ell},$$

*where* $k := k_A + k_B + k_C$, $\rho = \exp(\frac{\pi^2}{2\mu(\frac{b}{a})})$, *and* $\mu(\cdot)$ *is defined as in Lemma* 2.4.

*Proof.* Consider the following block partitioning for the Lyapunov equation:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} + \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix},$$

where the off-diagonal blocks—in each matrix—do not involve any elements of the main diagonal and all the dimensions are compatible. Without loss of generality we can consider the case $Y = X_{21}$. Observe that, writing the above system blockwise, we get the following relation:

$$A_{21}X_{11} + A_{22}X_{21} + X_{21}B_{11} + X_{22}B_{21} = C_{21}.$$

In particular, the block $X_{21}$ solves the Sylvester equation

$$A_{22}X_{21} + X_{21}B_{11} = C_{21} - A_{21}X_{11} - X_{22}B_{21},$$

in which the right-hand side has (standard) rank bounded by $k$. Since $A_{22}$ and $B_{11}$ are principal submatrices of symmetric positive definite matrices, they are again symmetric positive definite and such that $\kappa_2(A_{22}) \leqslant \frac{b}{a}$, and $\kappa_2(B_{11}) \leqslant \frac{b}{a}$. Therefore, using Lemma 2.4, we get the claim. ∎

*Remark* 2.8. In the case where $A$, $B$, and $C$ are banded with bandwidth $k_A$, $k_B$, and $k_C$, respectively, one can refine the bound given in Theorem 2.7 by using $k := \max\{k_A + k_B, k_C\}$. Indeed, $A_{21}$ being the off-diagonal block of a banded matrix, it has a row generator with nonzero entries only in the first $k_A$ rows. Analogously, the nonzero entries of the column generator corresponding to $B_{21}$ are located in its last $k_B$ rows. Finally, nonzero entries of $C_{21}$ are in the $k_C$ diagonals located in the upper right corner. Therefore, the matrix $C_{21} - A_{21}X_{11} - X_{22}B_{21}$ has nonzero elements only on the first $k_A$ rows, in the last $k_B$ columns, and in the $k_C$ upper right corner diagonals; see Figure 4. This provides the upper bound $\max\{k_A + k_B, k_C\}$ for its rank.

In Figure 5 we compare the bound given in Theorem 2.7 with the off-diagonal singular values of the solution. In this experiment, the matrix $C \in \mathbb{R}^{n\times n}$, $n = 300$, is diagonal with random entries and $A = B = MM^T$ where $M \in \mathbb{R}^{n\times n}$ is bidiagonal with ones on the main diagonal and random elements—chosen in $(0, 1)$—in the subdiagonal. The theoretical bound manages to describe the superlinear decay of the off-diagonal singular values. On the other hand, there is a significant gap between this estimate and the real behavior of the singular values. This is due to the fact that we are bounding the quantity $Z_\ell(E, F)$ where $E$ and $F$ are the convex hull of the spectra of $A$ and $-B$, respectively, instead of considering the Zolotarev problem directly on the discrete spectra. This is done in order to find explicit bounds, but it can cause an overestimation as outlined in [6].

A key property in the proof of Theorem 2.7 is the fact that submatrices of positive definite matrices are better conditioned than the original ones. This is an instance of a more general situation, which we can use to characterize the solution of Sylvester equations with nonnormal coefficients.

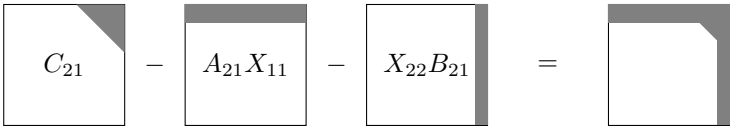


FIG. 4. *Sparsity structure of the equation for the off-diagonal block $X_{21}$ when $A$, $B$, and $C$ are banded matrices. As described in Remark* 2.8 *the rank of the right-hand side is bounded by* $\max\{k_A + k_B, k_C\}$.

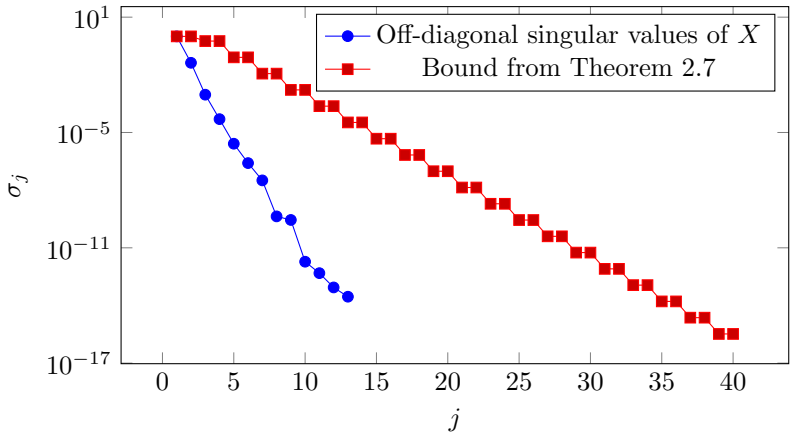


FIG. 5. *Off-diagonal singular values in the solution $X$ to* (1) *where $C$ is a random diagonal matrix and $A = B = MM^T$ with $M$ bidiagonal matrix with ones on the main diagonal and random elements—chosen in* $(0, 1)$*—in the subdiagonal. The dimension of the matrices is $n \times n$ with $n = 300$. The blue dots represent the most significant singular values of the off-diagonal block $X(\frac{n}{2} + 1 : n, 1 : \frac{n}{2})$. The red squares represent the theoretical bound given by Theorem* 2.12. *(Color available online.)*

DEFINITION 2.9. *Given an $n \times n$ square matrix $A$, we say that its* field of values *is the subset of the complex plane defined as follows:*

$$\mathcal{W}(A) := \left\{ \frac{x^H A x}{x^H x} \;\middle|\; x \in \mathbb{C}^n \backslash \{0\} \right\}.$$

One can easily check that for a normal matrix, being unitarily diagonalizable, the field of values is just the convex hull of the eigenvalues. For a general matrix, we know that the spectrum is contained in $\mathcal{W}(A)$, but the latter can be strictly larger than the convex hull of the former.

LEMMA 2.10. *Let $P$ be an orthogonal projection, i.e., an $n \times k$ matrix, $k < n$, with orthonormal columns. Then, for any matrix $A$, $\mathcal{W}(P^H A P) \subseteq \mathcal{W}(A)$. In particular, the field of values of any principal submatrix of $A$ is contained in $\mathcal{W}(A)$.*

*Proof.* The result directly comes by observing that

$$\max_{y \in \mathbb{C}^k} \frac{y^H P^H A P y}{y^H y} = \max_{y \in \mathbb{C}^k} \frac{y^H P^H A P y}{y^H P^H P y} \overbrace{\leqslant}^{x = Py} \max_{x \in \mathbb{C}^n} \frac{x^H A x}{x^H x}.$$

∎

LEMMA 2.11 (Crouzeix [18]). *Let $A$ be any $n \times n$ matrix, and let $f(z)$ be a holomorphic function defined on $\mathcal{W}(A)$. Then,*

$$\|f(A)\|_2 \leqslant \mathcal{C} \max_{z \in \mathcal{W}(A)} |f(z)|,$$

*where $\mathcal{C}$ is a universal constant smaller than or equal to $1+\sqrt{2}$.*

The above result is conjectured to be true with $\mathcal{C}=2$ and in this form is often referred to as the *Crouzeix conjecture* [17]. Lemmas 2.10–2.11 can be exploited to obtain a generalization of Theorem 2.7.

THEOREM 2.12. *Let $A, B$ be matrices of quasiseparable rank $k_A$ and $k_B$, respectively, and such that $\mathcal{W}(A) \subseteq E$ and $\mathcal{W}(-B) \subseteq F$. Consider the Sylvester equation $AX + XB = C$, with $C$ of quasiseparable rank $k_C$. Then a generic off-diagonal block $Y$ of the solution $X$ satisfies*

$$\frac{\sigma_{1+k\ell}(Y)}{\sigma_1(Y)} \leqslant \mathcal{C}^2 \cdot Z_\ell(E,F), \qquad k := k_A + k_B + k_C.$$

Other similar extensions of this result can be obtained using the theory of $K$-spectral sets [3].

**2.3. Quasiseparable approximability.** In the previous section we showed that, when the coefficients of the Sylvester equation are quasiseparable, the off-diagonal blocks of the solution $X$ have quickly decaying singular values. We want to show that this property implies the existence of a quasiseparable approximant.

In order to do that, we first introduce the definition of an $\epsilon$-quasiseparable matrix.

DEFINITION 2.13. *We say that $A$ has* $\epsilon$-quasiseparable rank $k$ *if, for every off-diagonal block $Y$, $\sigma_{k+1}(Y) \leqslant \epsilon$. If the property holds for the lower (resp., upper) off-diagonal blocks, we say that $A$ has lower (resp., upper) $\epsilon$-quasiseparable rank $k$.*

*Remark* 2.14. Notice that, if a matrix $A$ has $\epsilon$-quasiseparable rank $k$, then the same property is true for any of its principal submatrices $A'$. In fact, any off-diagonal block $Y$ of $A'$ is also an off-diagonal block of $A$, and therefore $\sigma_{k+1}(Y) \leqslant \epsilon$.

The next step is showing that an $\epsilon$-quasiseparable matrix admits a quasiseparable approximant. First, we need the following technical lemma, where $\oplus$ denotes the direct sum.

LEMMA 2.15. *Let $A$ be a matrix with $\epsilon$-quasiseparable rank $k$, and let $Q$ be any $(k+1)\times(k+1)$ unitary matrix. Then, $(I_{n-k-1} \oplus Q)A$ also has $\epsilon$-quasiseparable rank $k$.*

*Proof.* We prove the result for the lower off-diagonal blocks; the proof for the upper part follows along the same lines. Observe that we can verify the property for every maximal subdiagonal block $Y$; that is, $Y$ involves the first subdiagonal and the lower left corner. If $Y$ is contained in the last $k$ rows, its rank is at most $k$. Otherwise, if $Y$ includes elements from the last $j > k$ rows, we can write $Y = (I_{j-k-1} \oplus Q)\widetilde{Y}$, where $\widetilde{Y}$ is the corresponding subblock of $A$ (these two situations are depicted in Figure 6). Therefore, $\sigma_{k+1}(Y) = \sigma_{k+1}(\widetilde{Y}) \leqslant \epsilon$. ∎

THEOREM 2.16. *Let $A$ be of $\epsilon$-quasiseparable rank $k$ for $\epsilon > 0$. Then, there exists a matrix $\delta A$ of norm bounded by $\|\delta A\|_2 \leqslant 2\sqrt{n}\cdot\epsilon$ so that $A + \delta A$ is $k$-quasiseparable.*

*Proof.* We first show that there exists a perturbation $\delta A_\ell$ of norm bounded by $\sqrt{n}\cdot\epsilon$ that makes every lower off-diagonal block of $A$ of rank $k$.

We prove the result by induction on the dimension of $A$. If $n \leqslant 2k+1$, there is nothing to prove, since $A$ has all the off-diagonal blocks of rank at most $k$. If $n \geqslant 2k+2$, consider the following block partitioning of $A$:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad A_{11} \in \mathbb{C}^{(n-k-1)\times(n-k-1)}, \quad A_{22} \in \mathbb{C}^{(k+1)\times(k+1)}.$$

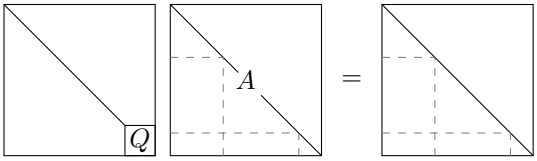


FIG. 6. *Off-diagonal blocks in the matrix* $(I_{n-k-1} \oplus Q)A$. *From the picture one sees that the* $Q$ *acts on the tall block without changing its singular values and that the small one has small rank thanks to the small number of rows.*

Since $\sigma_{k+1}(A_{21}) \leqslant \epsilon$, multiplying on the left by a unitary matrix $I_{n-k-1} \oplus Q^T$, where $Q$ contains the first $k$ left singular vectors of $A_{21}$, yields

$$\widetilde{A} := (I_{n-k-1} \oplus Q^T)A = \begin{bmatrix} \widetilde{A}_1 & v \\ w^T & d \end{bmatrix}, \qquad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad \|w_1\|_2 \leqslant \epsilon, \quad w_2 \in \mathbb{C}^k, \quad d \in \mathbb{C}.$$

Observe that, in view of Lemma 2.15, $\widetilde{A}$ still has $\epsilon$-quasiseparable rank $k$, and, according to Remark 2.14, the same holds for $\widetilde{A}_1$. Therefore, thanks to the induction step, there exists $\delta\widetilde{A}_{\ell,1}$ such that $\widetilde{A}_1 + \delta\widetilde{A}_{\ell,1}$ has lower quasiseparable rank $k$ and $\|\delta\widetilde{A}_{\ell,1}\|_2 \leqslant \sqrt{n-1} \cdot \epsilon$.

Define $\delta A_\ell$ and $\delta\widetilde{A}_\ell$ as follows:

$$\delta A_\ell := (I_{n-k-1} \oplus Q) \underbrace{\begin{bmatrix} \delta\widetilde{A}_{\ell,1} & 0 \\ -z^T & 0 \end{bmatrix}}_{\delta\widetilde{A}_\ell}, \qquad z = \begin{bmatrix} w_1 \\ 0 \end{bmatrix}.$$

Notice that $\|\delta\widetilde{A}_\ell\|_2 \leqslant \sqrt{\|\delta\widetilde{A}_{\ell,1}\|_2^2 + \|z\|_2^2} \leqslant \sqrt{n}\epsilon$. We claim that $A + \delta A_\ell$ is lower $k$-quasiseparable. With a direct computation we get

$$A + \delta A_\ell = (I_{n-k-1} \oplus Q) \underbrace{\begin{bmatrix} \widetilde{A}_1 + \delta\widetilde{A}_{\ell,1} & v \\ w^T - z^T & d \end{bmatrix}}_{\widehat{A}}.$$

The matrix $\widehat{A}$ is lower $k$-quasiseparable. In fact, every subdiagonal block of $\widehat{A}$ is equal to a subblock of $\widetilde{A}_1 + \delta\widetilde{A}_{\ell,1}$, possibly with an additional last row. If the subblock does not involve the last $k+1$ columns, the additional row is zero, and so the rank does not increase. Otherwise, the smallest dimension of the block is less than or equal to $k$, so its rank is at most $k$. Applying Lemma 2.15 once more with $\epsilon = 0$, we get that $A + \delta A_\ell$ is lower $k$-quasiseparable. Notice that it is not restrictive to assume $\delta A_\ell$ is lower triangular. In fact, if this is not the case, one can consider $\mathrm{tril}(\delta A_\ell)$ which still has the same property and has a smaller norm.

Repeating the process with $A^T$, we obtain an upper triangular matrix $\delta A_u$, of norm bounded by $\sqrt{n} \cdot \epsilon$, such that $A + \delta A_u$ is upper $k$-quasiseparable. Therefore, we have that $A + \delta A$ with $\delta A := \delta A_\ell + \delta A_u$ is $k$-quasiseparable, and $\|\delta A\|_2 \leqslant \|\delta A_\ell\|_2 + \|\delta A_u\|_2 \leqslant 2\sqrt{n} \cdot \epsilon$. ∎

*Remark* 2.17. Notice that, for $n \leqslant 2k+1$, the claim of Theorem 2.16 holds by choosing $\delta A = 0$. This means that the constant $2\sqrt{n}$ can be replaced with $2\sqrt{\max\{n-2k-1, 0\}}$.

The above result shows that a matrix with $\epsilon$-quasiseparable rank of $k$ can be well approximated by a matrix with exact quasiseparable rank $k$.

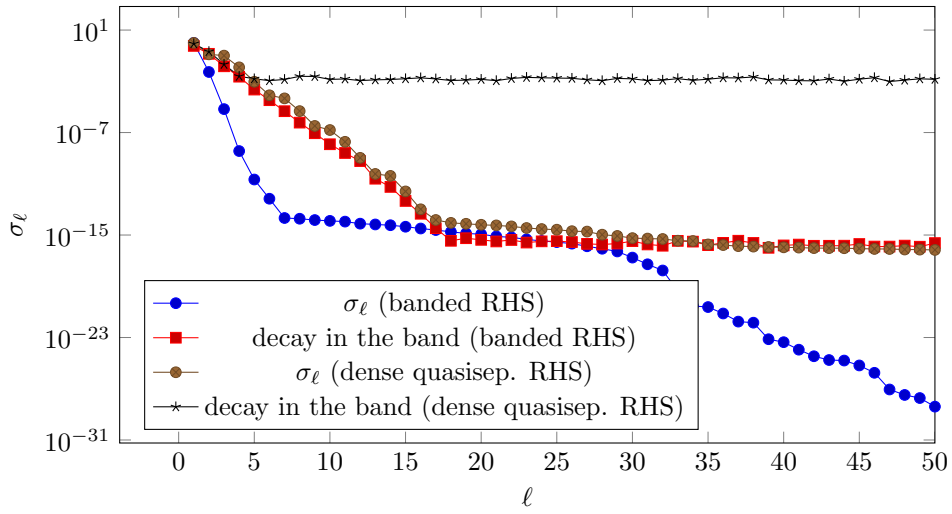


FIG. 7. *We compute* $X_i \in \mathbb{R}^{n\times n}$, $n = 300$, *as the solution of* $AX_i + X_iA = C_i$ *for* $i = 1, 2$, *respectively.* $A$ *is symmetric and tridiagonal with eigenvalues in* $[0.2, +\infty)$ *(positive definite and well-conditioned).* $C_1$ *is tridiagonal symmetric while* $C_2$ *is a dense random symmetric quasiseparable matrix of rank* 1.

**2.4. Preservation of the quasiseparable and banded structures.** The results of the previous section guarantee the presence of a numerical quasiseparable structure in the solution $X$ to (1) when the spectra of $A$ and $-B$ are well separated in the sense of the Zolotarev problem.

The preservation of a banded pattern in the solution has already been treated in [25, 35] in the case of Lyapunov equations with banded data and well-conditioned coefficient matrix. Moreover, in [35], it has been shown that if $A$ is ill-conditioned, the solution $X$ can be written as the sum of a banded matrix and a low-rank one, so that $X$ is quasiseparable. It is worth noticing that the results concerning the preservation of the banded and the banded plus low-rank structures do not require the separation property on the spectra of the coefficient matrices. This means that there are cases—not covered by the results of section 2.2—where the quasiseparability is still preserved.

In order to validate this consideration, we set up some experiments concerning the solution to (1) varying the structure of the coefficients and of the right-hand side. In particular, the features of the solution we are interested in are the distribution of the singular values $\sigma_\ell$ of the off-diagonal block $X(\frac{n}{2}+1 : n, 1 : \frac{n}{2})$[1] and the decay in the magnitude of the elements getting far from the main diagonal. The latter quantity is represented with the distribution of the maximum magnitude along the subdiagonal $\ell$ as $\ell$ varies from 1 to $n$. In all the performed tests we set $n = 300$, and the solution $X$ is computed by the Bartels–Stewart algorithm [4].

**Test 1.** We compute $X_i$ as the solution of $AX_i + X_iA = C_i$ for $i = 1, 2$. The matrix $A$ is chosen symmetric tridiagonal with eigenvalues in $[0.2, +\infty)$; in particular $A$ is positive definite and well-conditioned. The right-hand side $C_1$ is taken tridi-

[1] Notice that, in order to obtain a good hierarchical representation of the given matrices, the same structure needs to be present also in the upper off-diagonal block, and in the smaller off-diagonal blocks obtained in the recursion. Here we check just the larger off-diagonal block for simplicity; in the generic case, one may expect the quasiseparable rank to be given by the rank of this block.

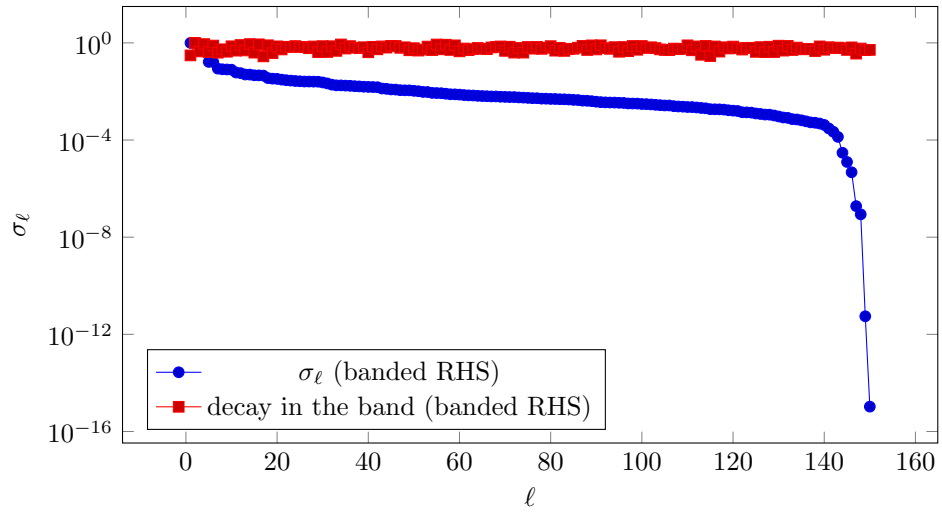


FIG. 8. *We compute the solution $X$ of $AX + XA = C$ and we analyze the off-diagonal block $X(\frac{n}{2}+1 : n, 1 : \frac{n}{2})$. $A = \mathrm{trid}(-1, 2, 1) - 1.99 \cdot I$ (indefinite and ill-conditioned) while $C$ is a random diagonal matrix.*

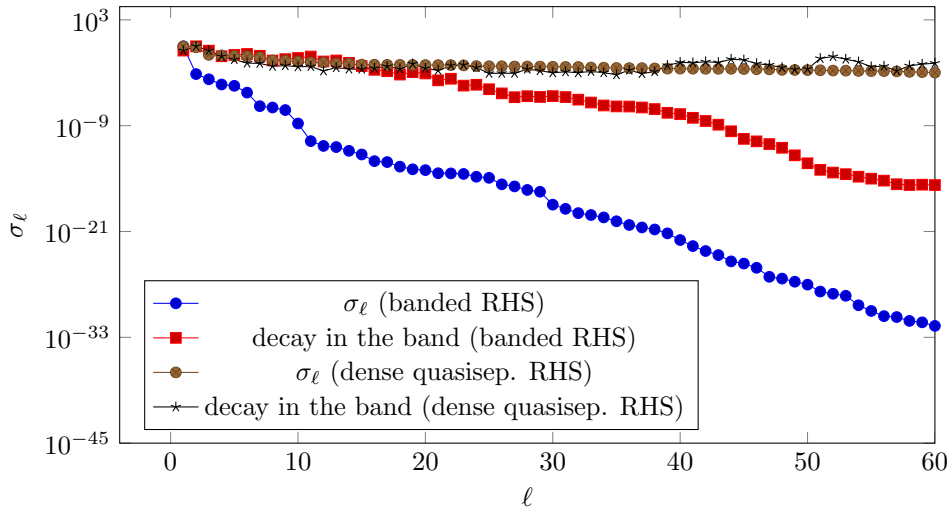


FIG. 9. *We compute $X_i \in \mathbb{R}^{n\times n}$, $n = 300$, as the solution of $AX_i - X_iB = C_i$ for $i = 1, 2$, respectively. $A$ and $B$ are symmetric and tridiagonal with eigenvalues in $[0.2, 14]$ and $[0.5, 14]$ (well conditioned but without separation of the spectra). $C_1$ is tridiagonal symmetric while $C_2$ is a dense random symmetric quasiseparable matrix of rank 1.*

agonal symmetric with random entries while $C_2$ is a random dense symmetric matrix with quasiseparable rank 1. In the first case, results from [25, 35] ensure that—numerically—the banded structure is maintained in the solution, and this is shown in Figure 7. Notice that the decay in the off-diagonal singular values is much stronger than the decay in the bandwidth so that, in this example, it is more advantageous to look at the solution as a quasiseparable matrix instead of a banded one. Theorem 2.7 guarantees the solution

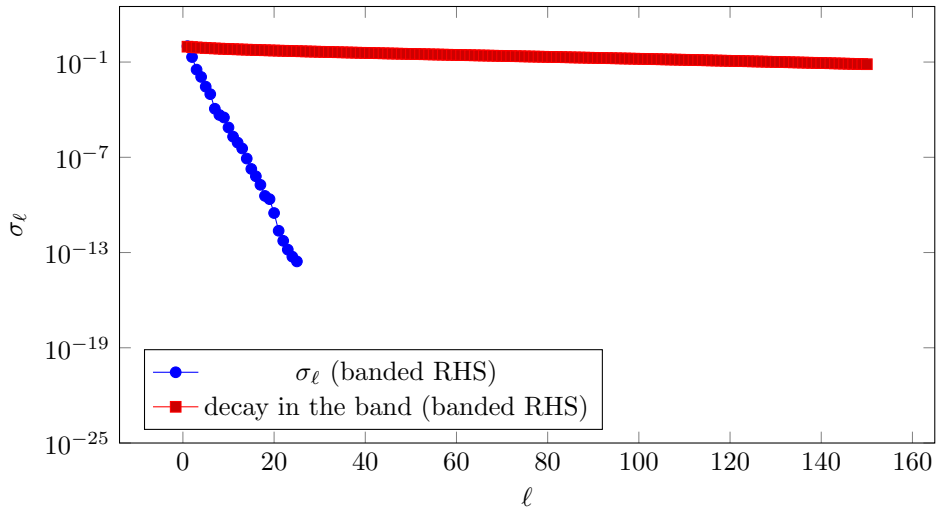


FIG. 10. *We compute the solution $X \in \mathbb{R}^{n\times n}$, $n = 300$, of $AX + XA = C$. $A = \mathrm{trid}(-1, 2, 1)$ (positive definite and ill-conditioned) while $C$ is a random diagonal matrix.*

to be quasiseparable also in the second case, whereas the banded structure is completely lost.

**Test 2.** We compute the solution $X$ of $AX+XA = C$. We consider $A = \mathrm{trid}(-1, 2, 1)-1.99\cdot I$, so that it is indefinite and ill-conditioned, and we set $C$ equal to a random diagonal matrix. As highlighted in Figure 8, neither the quasiseparable nor the band structure is present in the solution $X$.

**Test 3.** We compute $X_i$ as the solution of $AX_i + X_iB = C_i$ for $i = 1, 2$. The matrices $A$ and $-B$ are chosen symmetric and tridiagonal with eigenvalues in $[0.2, 14]$ and $[0.5, 14]$, so both are well conditioned but with interlaced spectra. The right-hand side $C_1$ is chosen tridiagonal symmetric while $C_2$ is set equal to a random dense symmetric matrix with quasiseparable rank 1. The results in Figure 9 suggest that both the structures are preserved in the first case and lost in the second case. Once again, in the case of preservation, the decay in the off-diagonal singular values is stronger than the decay in the bandwidth. Notice that, when present, the quasiseparability of the solution cannot be predicted by means of Theorem 2.7, but results from [25, 35] can be employed to estimate the banded structure of the solution. This test shows how the banded structure is a very particular instance of the more general quasiseparable one.

**Test 4.** We compute the solution $X$ of $AX + XA = C$. We choose $A = \mathrm{trid}(-1, 2, 1)$, so it is positive definite and ill-conditioned, and we set $C$ equal to a random diagonal matrix. Figure 10 clearly shows that quasiseparability is preserved while the banded structure is not present in the solution $X$. In this case, the quasiseparability of the solution can be shown by Theorem 2.7. Equivalently, one can exploit arguments in [35], where it has been shown that the solution can be represented as the sum of a banded matrix and a low-rank one so that $X$ is quasiseparable.

To summarize, the situations where we know that the quasiseparable structure is present in the solution of (1) are as follows:

(i) $A$, $B$, and $C$ are quasiseparable and the spectra of $A$ and $-B$ are well separated;[2]
(ii) $A$, $B$, and $C$ are banded and well-conditioned.

On the other hand, to use the computational approach of section 4 we need the spectra of $A$ and $-B$ to be separated by a line.

**3. HODLR matrices.** An efficient way to store and operate on matrices with an off-diagonal data-sparse structure is to use hierarchical formats. There is a vast literature on this topic. See, e.g., [12, 26, 28] and the references therein. In this work, we rely on a particular subclass of the set of hierarchical representations sometimes called hierarchically off-diagonal low-rank (HODLR), which can be described as follows: letting $A \in \mathbb{C}^{n\times n}$ be a $k$-quasiseparable matrix, we consider the $2 \times 2$ block partitioning

$$A = \begin{bmatrix} A_{11} & A_{22} \\ A_{21} & A_{22} \end{bmatrix}, \qquad A_{11} \in \mathbb{C}^{n_1\times n_1}, \quad A_{22} \in \mathbb{C}^{n_2\times n_2},$$

where $n_1 := \lfloor \frac{n}{2} \rfloor$ and $n_2 := \lceil \frac{n}{2} \rceil$. Since the antidiagonal blocks $A_{12}$ and $A_{21}$ do not involve any element of the main diagonal of $A$, they have rank at most $k$, so they are represented as low-rank outer products. Then, the strategy is applied recursively on the diagonal blocks $A_{11}$ and $A_{22}$. The process stops when the diagonal blocks reach a minimal dimension $n_{\min}$, at which they are stored as full matrices. The procedure is graphically described in Figure 11. If $n_{\min}$ and $k$ are negligible with respect to $n$, then the storage cost is linear-polylogarithmic with respect to the size of the matrix, as briefly summarized in Table 1. HODLR matrices are equivalent to hierarchical matrices with weak admissibility in the classification used in [27].

It is natural to compare the storage required by the HODLR representation and the truncation of banded structures when they are both present in the solution. Consider the following test: we compute the solution $X$ of a Lyapunov equation with a tridiagonal well-conditioned coefficient matrix $A$ and a diagonal right-hand side with random entries. As discussed in the previous section, the solution has a fast decay in the magnitude of the entries as we get far from the main diagonal. We compare the accuracy obtained when the solution $X$ is stored in the HODLR format with different thresholds in the low-rank truncation of the off-diagonal blocks, and when a fixed number of diagonals is memorized. In particular, the accuracy achieved keeping $5k$ diagonals and truncating the SVD of the off-diagonal blocks using thresholds $10^{-k}$, for $k = 0, \dots, 16$, is illustrated in Figure 12. We can see that the two approaches have comparable performances for this example. The experiment is repeated using $A = \mathrm{trid}(-1, 2, -1)$, highlighting the nonfeasibility of the sparse format in this case as the banded structure is not preserved in the solution.

The HODLR format has been studied intensively in the last decade and algorithms with almost linear complexity for computing matrix operations are available; see, e.g., Chapter 3 in [27]. Intuitively, the convenience of using this representation in a procedure is strictly related to the growth of the numerical rank of the off-diagonal blocks in the intermediate results. This can be formally justified with an argument based on the Eckart–Young best approximation property; see Theorem 2.2 in [11].

We relied on `hm-toolbox` for our experiments, which is available at https://github.com/numpi/hm-toolbox and implements HODLR arithmetic.

[2] We consider the spectra to be well separated if Theorem 2.7 can be used to prove the quasiseparability. As we have seen, this also includes cases where the spectra are close, such as when they are separated by a line.

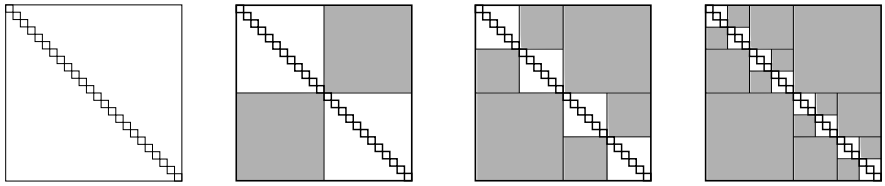

FIG. 11. *The behavior of the block partitioning in the HODLR matrix representation. The blocks filled with grey are low rank matrices represented in a compressed form, and the diagonal blocks in the last step are stored as dense matrices.*

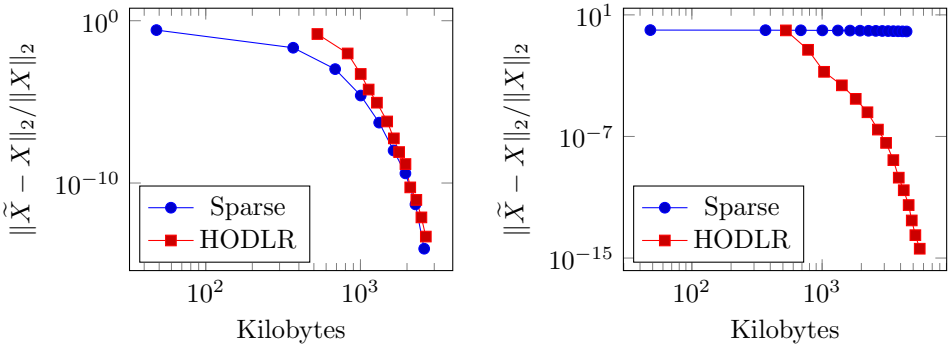


FIG. 12. *We show the accuracy obtained approximating the solution $X$ to a Lyapunov equation keeping a certain number of diagonals and by truncating the HODLR representations with $n_{min} = 50$. The plot reports the accuracy obtained with respect to the memory consumption when $A$ is banded and well conditioned (left), and for $A = \mathrm{trid}(-1, 2, -1)$ (right). The matrices have dimension $n = 2048$; the storage cost for the dense matrix $X$ is 32678 KB.*

**4. Solving the Sylvester equation.** In this section we show how to deal with the issue of solving (1) taking advantage of the quasiseparable structure of the data. We first discuss the matrix sign iteration and then we show how to efficiently evaluate the integral formula (3). Both these algorithms are implemented in the `hm-toolbox`.

**4.1. Matrix sign function.** Here, we briefly recall the matrix sign function iteration, first proposed in the $\mathcal{H}$-format by Grasedyck, Hackbusch, and Khoromskij in [22]. We use HODLR arithmetic in the iteration scheme proposed by Roberts in [39], which relies on the following result.

THEOREM 4.1. *Let $A, B \in \mathbb{C}^{n\times n}$ be positive definite; then the solution $X$ of (1) verifies*

$$X = \frac{1}{2}N_{12}, \qquad \begin{bmatrix} N_{11} & N_{12} \\ 0 & N_{22} \end{bmatrix} := \mathrm{sign}\left(\begin{bmatrix} A & C \\ 0 & -B \end{bmatrix}\right), \tag{4}$$

*and, given a square matrix $M$, we define $\mathrm{sign}(M) := \frac{1}{\pi \mathbf{i}}\int_\gamma (zI - M)^{-1}dz$ with $\gamma$, a path of index 1 around the eigenvalues of $M$ with positive real part.*

The sign function of a square matrix $S := \mathrm{sign}(M)$ can be approximated applying the Newton method to the equation $X^2 - I = 0$ with starting point $S_0 = M$. This requires computing the sequence

$$S_0 = M, \qquad S_{i+1} = \frac{1}{2}(S_i + S_i^{-1}), \tag{5}$$

TABLE 1
*Computational complexity of the HODLR matrix arithmetic. The integer k is the maximum of the quasiseparable ranks of the inputs while n is the size of the matrices.*

| Operation | Computational complexity |
|---|---|
| Matrix-vector multiplication | $O(kn\log(n))$ |
| Matrix-matrix addition | $O(k^2n\log(n))$ |
| Matrix-matrix multiplication | $O(k^2n\log^2(n))$ |
| Matrix inversion | $O(k^2n\log^2(n))$ |
| Solve linear system | $O(k^2n\log^2(n))$ |

which converges to $S$, provided that $M$ has no eigenvalues on the imaginary axis [22]. Rewriting (5) blockwise yields

$$A_{i+1} = \frac{1}{2}(A_i + A_i^{-1}), \quad B_{i+1} = \frac{1}{2}(B_i + B_i^{-1}), \quad C_{i+1} = \frac{1}{2}(A_i^{-1}C_iB_i^{-1} + C_i), \tag{6}$$

where $A_0 = A, B_0 = B$, $C_0 = C$, and $C_{i+1} \to 2X$. As stopping criterion we used the condition

$$\|A_{i+1} - A_i\|_F + \|B_{i+1} - B_i\|_F + \|C_{i+1} - C_i\|_F \leqslant \sqrt{\epsilon},$$

where $\epsilon$ is the selected accuracy. This can be heuristically justified by saying that since Newton is quadratically convergent, if the above quantity is a good estimate of the error of the previous step, then we have already obtained the solution at the required precision.

We implemented the algorithm in [22] that performs the iteration using hierarchical matrix arithmetic. When an appropriate scaling of $A$ and $B$ is performed, convergence is reached in a few steps [30]. The scaling strategy is crucial to keeping the number of iterations of the Newton scheme low, and the scaling parameter $\alpha > 0$ can be optimally chosen at every iteration, as shown in [30]. When the spectra of $A$ and $B$ are real, the optimal choice is $\alpha_i = \sqrt{\|S_i^{-1}\|_2/\|S_i\|_2}$. However, if hierarchical matrix arithmetic is employed, the scaling strategy may introduce a nonnegligible error propagation, as outlined in [22]. We found out that a good trade-off is to scale only in the first iteration. This does not affect the accuracy of the iterative steps if the matrix $S_0$ can be exactly represented in the hierarchical format [22, Remark 5.3] and allows us to keep the number of iterations proportional to $\log(\max\{\kappa(A), \kappa(B)\})$ [22]. For instance, in the case of $A = B$ being the discrete Laplacian operator, which has a condition number that grows as $\mathcal{O}(n^2)$, the latter choice makes the computational cost of the approach $\mathcal{O}(n\log^3(n))$.

**4.2. Solution by means of the integral formula.** We now propose applying a quadrature scheme for evaluating the semi-infinite integral in (3). We perform the change of variable $x = f(\theta) := L \cdot \cot\left(\frac{\theta}{2}\right)^2$, where $\theta$ is the new variable and $L$ is a parameter chosen to optimize the convergence. This is a very common strategy for the approximation of integral over infinite domain, which is discussed in detail by Boyd in [13]. We transform (3) into

$$X = 2L \int_0^\pi \frac{\sin(\theta)}{(1-\cos(\theta))^2}\, e^{-Af(\theta)} C e^{-Bf(\theta)} d\theta, \tag{7}$$

which can be approximated by a Gauss–Legendre quadrature scheme. Other quadrature formulas, such as Clenshaw–Curtis rules, can be employed. However, as discussed

by Trefethen in [45], the difference between Gauss–Legendre and Clenshaw–Curtis formulas is small. Moreover, in most of our tests, Gauss–Legendre schemes showed some slight computational advantages over Clenshaw–Curtis rules as the cost of computing the integration points is negligible.[3]

The quadrature scheme yields an approximation of (7) of the form

$$X \approx \sum_{j=1}^{m} \omega_j \cdot e^{-Af(\theta_j)} C e^{-Bf(\theta_j)}, \tag{8}$$

where $\theta_j$ are the Legendre points and $\omega_j = 2Lw_j \cdot \frac{\sin(\theta_j)}{(1-\cos(\theta_j))^2}$ and $w_j$ are the Legendre weights.

Finally, we numerically approximate the quantities $e^{-Af(\theta_j)}$ and $e^{-Bf(\theta_j)}$, which represent the dominant cost of the algorithm. For this task, we have investigated two rational approximations, which have been implemented in our toolbox.

**Padé.** The matrix exponential $e^A$ can be well approximated by a diagonal Padé approximant of degree $(d, d)$ if $\|A\|$ is small enough.[4] We thus satisfy this condition by using the relation $e^A = (e^{2^{-k}A})^{2^k}$, a tecnhique typically called "scaling and squaring." The Padé approximant is known explicitly for all $d$. See, e.g., [30, Chapter 10]. In this case the evaluation of the matrix exponential requires $2d+3+k$ matrix multiplication and one inversion where $k = \lceil \log_2 \|A\| \rceil$. This strategy is also implemented in the MATLAB function `expm`.

**Chebyshev.** Since $A$ is supposed to be positive definite, the matrix exponential $e^{-tA}$ can be approximated by a rational Chebyshev function that is uniformly accurate for every positive value of $t$, as described by Popolizio and Simoncini in [37]. The rational function is of the form

$$e^x \approx \frac{r_1}{x - s_1} + \cdots + \frac{r_d}{x - s_d}.$$

Given the poles and the weights in the above expansion, this strategy requires $d$ inversions and additions. See, e.g., [35] for a numerical procedure to compute the poles and weights $s_i$, $r_i$.

*Remark* 4.2. In our tests, evaluating the matrix exponential $e^{-f(\theta)A}$ by means of the Padé approximant performs better when $f(\theta)A$ has a moderate norm. When $f(\theta)\|A\|_2$ is large the squaring phase becomes the bottleneck of the computation. In this case we rely on the rational Chebyshev expansion, which has a cost independent of $\|A\|_2$.

The procedure is summarized in Algorithm 1. The evaluations of the matrix exponentials $\text{EXPM}(-f \cdot A)$, $\text{EXPM}(-f \cdot B)$ are performed according to the strategy outlined in Remark 4.2.

**5. Solving certain generalized equations.** The solution of certain generalized Sylvester equations can be recast in terms of standard Sylvester ones. The results in section 2 thus suggest the presence of a quasiseparable structure also in the solution of these kinds of equations. For the sake of simplicity, we focus on generalized

[3] In practice we have precomputed the points for the usual cases so that an explicit computation of them is never carried out in the numerical experiments.

[4] The exact choice of the ball where Padé is accurate enough depends on the desired accuracy and the value of $d$.

**Algorithm 1** Solution of a Sylvester equation by means of the integral formula.

**procedure** LYAP_INTEGRAL($A, B, C, m$) ▷ Solves $AX + XB = C$ with $m$ integration points
  $L \leftarrow 100$ ▷ This can be tuned to optimize the accuracy
  $[w, \theta] \leftarrow$ GAUSSLEGENDREPTS($m$) ▷ Integration points and weights on $[0, \pi]$
  $X \leftarrow 0_{n \times n}$
  **for** $i = 1, \dots, m$ **do**
    $f \leftarrow L \cdot \cot(\frac{\theta_i}{2})^2$
    $X \leftarrow X + w_i \frac{\sin(\theta_i)}{(1-\cos\theta_i)^2} \cdot \text{EXPM}(-f \cdot A) \cdot C \cdot \text{EXPM}(-f \cdot B)$
  **end for**
  $X \leftarrow 2L \cdot X$
**end procedure**

Lyapunov equations, but the approach we are going to present can be easily extended to the Sylvester case as well. We consider equations of the form

$$AX + XA + \sum_{j=1}^{s} M_j X M_j^T = C, \qquad A, X, C, M_j \in \mathbb{R}^{n \times n}, \tag{9}$$

where $A$ is symmetric positive definite, both $A$ and $C$ are quasiseparable, and $M_j$ is low rank for $j = 1, \dots, s$. We generalize Theorem 2.7 to this framework.

COROLLARY 5.1. *Let $A$ be a symmetric positive definite matrix of quasiseparable rank $k_A$, and let $\kappa_A$ be its condition number. Moreover, consider the generalized Lyapunov equation $AX + XA + \sum_{j=1}^{s} M_j X M_j^T = C$, with $M_j$ of rank $r_j$, $j = 1, \dots, s$, and $C$ of quasiseparable rank $k_C$. Then a generic off-diagonal block $Y$ of the solution $X$ satisfies*

$$\frac{\sigma_{1+k\ell}(Y)}{\sigma_1(Y)} \leqslant 4\rho^{-2\ell},$$

*where $k := 2k_A + k_C + \sum_{j=1}^{s} r_j$, $\rho = \exp(\frac{\pi^2}{2\mu(\kappa_A)})$, and $\mu(\cdot)$ is defined as in Lemma 2.4.*

*Proof.* The solution $X$ satisfies $AX + XA = C - \sum_{j=1}^{s} M_j X M_j^T$, where the right-hand side has quasiseparable rank $k_C + \sum_{j=1}^{s} r_j$. By applying Theorem 2.7 to the latter we get the claim. ∎

Equation (9) can be rephrased as an $n^2 \times n^2$ linear system by Kronecker transformations

$$(\mathcal{L} + \mathcal{M})\,\text{vec}(X) = \text{vec}(C), \qquad \mathcal{L} := I \otimes A + A \otimes I, \qquad \mathcal{M} := \sum_{j=1}^{s} M_j \otimes M_j.$$

We assume that $\mathcal{L}$ is invertible, and such that $M_j = U_j V_j^T$ with $U_j, V_j \in \mathbb{R}^{n \times r_j}$, $j = 1, \dots, s$. In particular, the matrix $\mathcal{M} \in \mathbb{R}^{n^2 \times n^2}$ is of rank $r := \sum_{j=1}^{s} r_j^2$, and it can be factorized as

$$\mathcal{M} = UV^T = \left[\begin{array}{c|c|c} U_1 \otimes U_1 & \dots & U_s \otimes U_s \end{array}\right] \cdot \left[\begin{array}{c|c|c} V_1 \otimes V_1 & \dots & V_s \otimes V_s \end{array}\right]^T.$$

Plugging this factorization into the Sherman–Morrison–Woodbury formula, we get

$$\text{vec}(X) = \mathcal{L}^{-1}\text{vec}(C) - \mathcal{L}^{-1}U\left(I_r + V^T\mathcal{L}^{-1}U\right)^{-1}V^T\mathcal{L}^{-1}\text{vec}(C). \tag{10}$$

See, e.g., [8]. As shown in [38, section 4], the solution of (9) by (10) requires the inversion of an $r \times r$ linear system and the solution of $r+1$ Lyapunov equations of the form

$$AZ + ZA = C, \quad AZ_{ij} + Z_{ij}A = \widetilde{U}_{ij} \quad \text{for } i = 1, \dots, s,\ j = 1, \dots, r_i^2,$$

where $\widetilde{U}_{ij} = \text{vec}^{-1}((U_i \otimes U_i)(:, j))$ has rank 1. Since $A$ and $C$ are quasiseparable, the matrix $Z$ can be computed by one of the methods presented in the previous sections, whereas well-established low-rank methods can be employed in computing the $Z_{ij}$'s. In our tests we have used the method based on extended Krylov subspaces discussed in [41]. The procedure is illustrated in Algorithm 2.

**Algorithm 2** Solution of a generalized Lyapunov equation (low-rank $\mathcal{M}$) by (10).

**procedure** SMW_Gen_Lyap($A, C, U_i, V_i$) ▷ Solve $AX + XA + \sum_{i=i}^{s} U_i V_i^T X V_i U_i^T = C$
  $\widehat{X} \leftarrow A\widehat{X} + \widehat{X}A = C$
  **for** $h = 1 : s$ **do**
    $\widehat{X}_h \leftarrow V_h^T \widehat{X} V_h$
  **end for**
  **for** $h = 1 : s$, $i, j = 1 : r_h$ **do**
    $U_{ij}^h \leftarrow U_h(:, i)U_h(:, j)^T$
    $Z_k \leftarrow AZ_k + Z_kA = U_{ij}^h$ ▷ $k := j + (i-1)r_h + \sum_{t=1}^{h-1} r_t^2$
    **for** $m = 1 : s$ **do**
      $W_{mk} \leftarrow V_m^T Z_k V_m$
    **end for**
    $[Z_{1+\sum_{t=1}^{h-1} r_t^2}, \dots, Z_{\sum_{t=1}^{h} r_t^2}] = Z_h^{(u)} Z_h^{(v)T}$
  **end for**
  $R \leftarrow \left(I_r + \left[\text{vec}(W_{mk})\right]_{\substack{m=1,\dots,s \\ k=1,\dots,r}}\right)^{-1}$
  $\widehat{Z} \leftarrow R \cdot [\text{vec}(\widehat{X}_1); \dots; \text{vec}(\widehat{X}_s)]$
  $S \leftarrow \sum_{h=1}^{s} Z_h^{(u)} \cdot \widehat{Z}_{r_h} \cdot Z_h^{(v)T}$ ▷ $\widehat{Z}_{r_h} := \texttt{reshape}(\widehat{Z}(1 + \sum_{i=1}^{h-1} r_i^2 : \sum_{i=1}^{h} r_i^2), r_h, r_h)$
  $X \leftarrow \widehat{X} - S$
  **return** $X$
**end procedure**

Another interesting class of generalized Lyapunov equations consists of (9) with $\rho(\mathcal{L}^{-1}\mathcal{M}) < 1$, where $\rho(\cdot)$ denotes the spectral radius. For these kinds of problems, the matrices $M_j$ do not need to be low rank, but we suppose they all have a small quasiseparable rank. In this case, one can consider the Neumann series expansion of $(\mathcal{L} + \mathcal{M})^{-1}$ as done in [31]. More precisely, it holds that

$$(\mathcal{L} + \mathcal{M})^{-1} = (I + \mathcal{L}^{-1}\mathcal{M})^{-1}\mathcal{L}^{-1} = \sum_{j=0}^{\infty} (-1)^j (\mathcal{L}^{-1}\mathcal{M})^j \mathcal{L}^{-1},$$

so that the solution $X$ to (9) verifies

$$X = \sum_{i=0}^{\infty} Z_i, \quad \text{where} \quad \begin{cases} AZ_0 + Z_0A = C, \\ AZ_{i+1} + Z_{i+1}A = -\sum_{j=1}^{s} M_j Z_i M_j^T. \end{cases} \tag{11}$$

A numerical solution can thus be computed by truncating the series in (11), that is, $X \approx X_\ell := \sum_{i=0}^{\ell} Z_i$, where the number of terms $\ell$ is related to the accuracy of the computed solution. If $\ell$ is moderate, that is, $\rho(\mathcal{L}^{-1}\mathcal{M}) \ll 1$, $X_\ell$ is the sum of a few quasiseparable matrices $Z_i$, and it is thus quasiseparable. Notice that the quasiseparability of the $M_j$'s is necessary to maintain a quasiseparable structure in the right-hand sides $-\sum_{j=1}^{s} M_j Z_i M_j^T$, $i = 0, \dots, \ell - 1$.

**6. Numerical experiments.** An extensive computational comparison among different approaches for quasiseparable Sylvester equations, as well as their implementation, is still lacking in the literature, and in this section we perform some numerical experiments trying to fill this gap. To this end, we employ the MATLAB `hm-toolbox` that we have developed while writing this paper. The toolbox, which includes all the tested algorithms, is now freely available at https://github.com/numpi/hm-toolbox. All the timings reported are relative to MATLAB 2016a run on a machine with a CPU running at 3066 MHz, 12 cores,[5] and 192GB of RAM.

Each of the following sections contains a specific example. Some of these are artificially constructed to describe particular cases; others present real or realistic applications, arising from PDEs. We start by describing the classical Laplacian case and then proceed, comparing our results with a two-dimensional (2D) heat equation arising from practical applications. Eventually, we show how to solve some partial integro-differential equations.

To test the accuracy of our approach we report the relative residual on the linearized system of the computed solution. If $\mathcal{S}$ is the coefficient matrix of the linearized system, we measure the relative residual,

$$r(\mathcal{S}, X) := \frac{\|\mathcal{S} \cdot x - c\|_2}{\|\mathcal{S}\|_F \cdot \|x\|_2}, \qquad x = \mathrm{vec}(X), \qquad c = \mathrm{vec}(C),$$

which can be easily shown to be the relative backward error in the Frobenius norm [29]. When we deal with (standard) Sylvester problems, we have $\mathcal{S} = I \otimes A + B \otimes I$ with $A$ and $B$ symmetric. This allows us to use the—easier to compute—bound

$$\|\mathcal{S}\|_F^2 \geqslant n(\|A\|_F^2 + \|B\|_F^2),$$

so that

$$r(\mathcal{S}, X) = \frac{\|\mathcal{S} \cdot x - c\|_2}{\|\mathcal{S}\|_F \cdot \|x\|_2} = \frac{\|AX + XB - C\|_F}{\|I \otimes A + B \otimes I\|_F \cdot \|X\|_F} \leqslant \frac{\|AX + XB - C\|_F}{\sqrt{n(\|A\|_F^2 + \|B\|_F^2)} \cdot \|X\|_F},$$

and we actually compute and check the right-hand side in the above expression.

In case of a generalized Lyapunov equation, the system matrix is of the form $\mathcal{S} := I \otimes A + B \otimes I + M \otimes M$, and the relative residual norm is bounded using the inequality $\|\mathcal{S}\|_F \geqslant \|I \otimes A + B \otimes I\|_F - \|M\|_F^2$. Notice that this never requires forming the large system matrix $\mathcal{S}$ and can be evaluated using the arithmetic of hierarchical matrices when considering large scale problems.

[5] All the available cores have only been used to run the parallel implementation of the solver based on the integral formula. None of the other solvers exploited the parallelism in the machine.

**6.1. The Laplace equation.** We consider the 2D Laplace equation on the unit square $\Omega = [0,1]^2$:

$$\begin{cases} -\Delta u = f(x,y), & (x,y) \in \Omega, \\ u(x,y) = 0, & (x,y) \in \partial\Omega, \end{cases} \qquad \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

We construct the matrix $A$ representing the finite difference discretization of the second-order derivative in the above equation on an $n \times n$ grid using centered differences, so that we obtain the equation $AX + XA = C$, with

$$A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \qquad h = \frac{1}{n-1},$$

and $C$ contains the samplings of the function $f(x,y)$ on our grid. We consider the case where $f(x,y) = \log(1 + |x-y|)$. As already discussed, the latter choice provides a right-hand side which is numerically quasiseparable. This is due to the fact that in the subdomains corresponding to the off-diagonal blocks, $f$ is analytic and it is well approximated by a sum of a few separable functions. One can also exploit this property in order to retrieve the HODLR representation of $C$; the sampling of a separable function $g(x) \cdot h(y)$ on a square grid provides a matrix of rank 1, and the sampling of $g$ and $h$ yields its generating factors. The computation of the expansion of $f$ in the subdomains has been performed by means of Chebfun2 [44].

Using `hm-toolbox`, the equation can be solved with a few MATLAB instructions, as shown in Figure 13 for the case $n = 2048$. The function `hmoption` can be used to set some options for the toolbox. In this case we set the relative threshold for the off-diagonal truncation to $10^{-12}$ and the minimum size of the blocks to 256. The class `hm` implements the hierarchical structure, and here we initialize it using a sparse tridiagonal matrix. Invoking the `lyap` function uses our implementation specialized for $\mathcal{H}$-matrices.

In this example, we used the sign iteration, which is the default method for the implementation of `lyap`. The quasiseparable rank of the solution (obtained using the function `hmrank`) is 13, which is reasonably small compared to the size of the problem.

In Table 2 and Figure 14 we show the timings for the solution of this problem for different grid sizes. We stress that, since full matrices are never represented, a large amount of RAM is not needed to run the solver. Nevertheless, this is needed when using `lyap` from the MATLAB Control Toolbox, so we have comparisons with the latter only for $n \leqslant 4096$.

The results in Table 2 show that the timings are just a little more than linear in the size of the problem. Figure 14 illustrates that the complexity is in fact $\mathcal{O}(n \log^2 n)$ for the methods that evaluate the integral formula (3).

The approach based on the sign iteration is faster than the one that exploits the integral formula. Nevertheless the latter has a slightly better asymptotic cost since it requires $\mathcal{O}(n \log^2(n))$ flops instead of $\mathcal{O}(n \log^3(n))$. Another advantage of the integral formula is the easy parallelization. In fact, the evaluation of the integrand at the nodes can be carried out in a parallel fashion on different machines or cores. In our tests we used 32 integration nodes, so the maximum gain in the performances can be obtained using 32 cores. The results reported in Table 2 confirm the acceleration of the parallel implementation when using 12 cores.

```
>> n = 2048;
>> hmoption('threshold', 1e-12);
>> hmoption('block-size', 256);
>> f = @(x,y) log(1 + abs(x - y));
>> A = (n-1) 2 * spdiags(ones(n,1) * [ -1 2 -1 ], -1:1, n, n);
>> H = hm('tridiagonal', A);
>> C = hm('chebfun2', f, [-1,1], [-1,1], n);
>> X = lyap(H, C, 'method', 'sign');
>> hmrank(X)
ans =
      13
```

FIG. 13. *Example MATLAB session where the* `hm-toolbox` *is used to compute the solution of a Lyapunov equation involving the* 2*D Laplacian and a numerically quasiseparable right-hand side.*

TABLE 2

*Timings and features of the solution of the Laplacian equation for different grid sizes. For the methods based on the HODLR arithmetic the minimum block size is set to* 256 *and the relative threshold in truncation is* $\epsilon = 10^{-12}$. *For small problems we also report the timings of the* `lyap` *function included in the Control Toolbox in MATLAB. The relative residuals of the Lyapunov equation are reported as well for the different methods. The residuals for the parallel version of the method based on the exponential have been omitted since they coincide with those of the sequential one. In fact, the two algorithms perform exactly the same computations.*

| $n$ | $T_{\text{Sign}}$ | $Res_{\text{Sign}}$ | QS rk | $T_{\text{Exp}}$ | $T_{\text{ParExp}}$ | $Res_{\text{Exp}}$ | QS rk | $T_{\text{lyap}}$ |
|---|---|---|---|---|---|---|---|---|
| 512 | 0.71 | $2.97 \cdot 10^{-12}$ | 13 | 3.69 | 1.52 | $3.92 \cdot 10^{-9}$ | 13 | 0.85 |
| 1,024 | 1.73 | $4.33 \cdot 10^{-12}$ | 14 | 9.37 | 3.21 | $8.71 \cdot 10^{-10}$ | 14 | 7.52 |
| 2,048 | 4.76 | $2.03 \cdot 10^{-11}$ | 13 | 22.78 | 6.34 | $7.21 \cdot 10^{-10}$ | 14 | 80.15 |
| 4,096 | 13.33 | $5.19 \cdot 10^{-11}$ | 15 | 57.15 | 14.51 | $5.73 \cdot 10^{-11}$ | 12 | 523.16 |
| 8,192 | 35.93 | $3.65 \cdot 10^{-11}$ | 13 | 136.42 | 31.82 | $9.23 \cdot 10^{-12}$ | 11 | |
| 16,384 | 92.83 | $1 \cdot 10^{-10}$ | 14 | 334.75 | 70.28 | $3.14 \cdot 10^{-12}$ | 11 | |
| 32,768 | 245.82 | $1.55 \cdot 10^{-10}$ | 16 | 790.28 | 154.65 | $1.42 \cdot 10^{-12}$ | 11 | |
| 65,536 | 609.86 | $1.33 \cdot 10^{-10}$ | 15 | 1,825.2 | 351.82 | $8.86 \cdot 10^{-13}$ | 10 | |
| $1.31 \cdot 10^5$ | 1,474.56 | $1.58 \cdot 10^{-10}$ | 17 | 4,122.17 | 763.05 | $2.03 \cdot 10^{-12}$ | 9 | |

**6.2. The 2D heat equation.** We consider now a case of more practical interest, which has been described and studied by Haber and Verhaegen in [24, 25]. They study a particular discretization for the 2D heat equation that gives rise to a Lyapunov equation with banded matrices. Let $S_m = \text{trid}(1, 0, 1)$ be the $m \times m$ matrix with 1 on the super- and subdiagonal and zeros elsewhere, and let $\mathbf{1}_m \in \mathbb{C}^m$ be the vector with all the entries equal to 1. The resulting Lyapunov equation involves the coefficient matrices

$$A = I_m \otimes (aI_6 + eS_6) + eS_m \otimes I, \qquad C = I_m \otimes (0.2 \cdot \mathbf{1}_6\mathbf{1}_6^T + 0.8I) + 0.1S_m \otimes (\mathbf{1}_6\mathbf{1}_6^T).$$

For the details on how these matrices are obtained from the discretization phase we refer the reader to [24]. The parameters $a$ and $e$ are set to $a = 1.36$ and $e = -0.34$. These two matrices are banded, with bandwidth 6 and 11, respectively. However, a careful look shows that the quasiseparable rank of $A$ is 6, but the one of $C$ is 1: the quasiseparable representation can exploit more structure than the banded one in this problem.

We have solved this problem for different values of $m$, from $m = 128$ to $m = 32768$. For each $m$, the size of the associated matrices $A$ and $P$ is $6m \times 6m$. We have

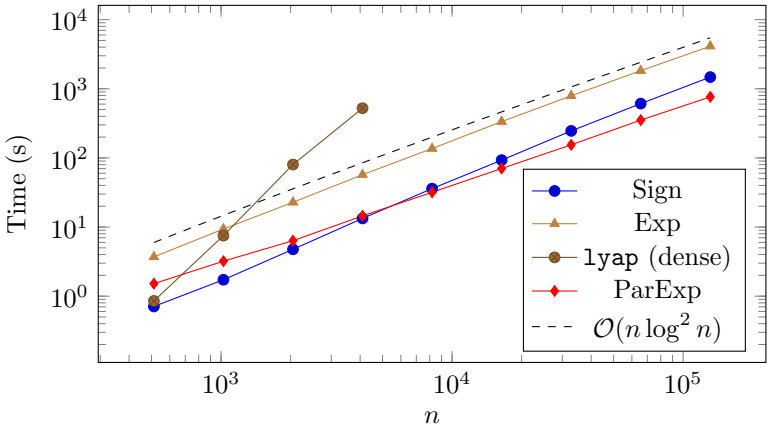


FIG. 14. *Timings for the solution of the Laplacian equation for different grid sizes. The performances of the different algorithms are reported. The dashed line reports the theoretical complexity of* $\mathcal{O}(n\log^2(n))$.

TABLE 3

*Timings and features of the solution of the heat equation for different grid sizes. For the methods based on the HODLR arithmetic the minimum block size is set to* 256 *and the relative threshold in truncation is* $\epsilon = 10^{-12}$. *In this example the quasiseparable rank of the solution coincides for the implementation based on the sign function and on the integral formula, so we have only reported it once.*

| $n$ | $T_{\text{Sign}}$ | $Res_{\text{Sign}}$ | $T_{\text{ParExp}}$ | $Res_{\text{ParExp}}$ | QS rk | $T_{\text{SparseCG}}$ | $Res_{\text{SparseCG}}$ |
|---|---|---|---|---|---|---|---|
| 768 | 1.06 | $8.95 \cdot 10^{-13}$ | 1.96 | $9.44 \cdot 10^{-12}$ | 13 | 1.18 | $2.96 \cdot 10^{-11}$ |
| 1,536 | 2.74 | $1.42 \cdot 10^{-12}$ | 4.99 | $4.92 \cdot 10^{-12}$ | 12 | 2.49 | $2.81 \cdot 10^{-11}$ |
| 3,072 | 8.29 | $9.73 \cdot 10^{-12}$ | 13.12 | $1.53 \cdot 10^{-11}$ | 12 | 4.79 | $2.67 \cdot 10^{-11}$ |
| 6,144 | 19.3 | $4.94 \cdot 10^{-12}$ | 32.21 | $1.08 \cdot 10^{-11}$ | 10 | 9.23 | $2.57 \cdot 10^{-11}$ |
| 12,288 | 48.44 | $4.76 \cdot 10^{-12}$ | 79.46 | $1.36 \cdot 10^{-11}$ | 10 | 18.25 | $2.41 \cdot 10^{-11}$ |
| 24,576 | 117.32 | $4.71 \cdot 10^{-12}$ | 189.84 | $1.80 \cdot 10^{-11}$ | 10 | 36.96 | $3.22 \cdot 10^{-11}$ |
| 49,152 | 277.8 | $1.09 \cdot 10^{-11}$ | 445.03 | $1.62 \cdot 10^{-11}$ | 10 | 67.18 | $3.03 \cdot 10^{-11}$ |
| 98,304 | 589.51 | $3.87 \cdot 10^{-11}$ | 1,092.1 | $2.69 \cdot 10^{-11}$ | 10 | 121.31 | $2.87 \cdot 10^{-11}$ |
| $1.97 \cdot 10^5$ | 1,312.6 | $1.05 \cdot 10^{-10}$ | 2,677.1 | $8.16 \cdot 10^{-11}$ | 9 | 213.08 | $2.75 \cdot 10^{-11}$ |

also compared our implementation to the (sparse) conjugate gradient implemented in matrix form, as proposed in [35]. One can see that, at the $k$th iteration of the conjugate gradient method, the solution in matrix form has a bandwidth proportional to $k$; when the method converges in a few steps, this can provide an accurate banded approximation to the solution in linear time. In fact, this problem is well-conditioned independently of $n$ and therefore is the ideal candidate for the application of this method (as shown in [35]). Additionally, the sparse arithmetic implemented in MATLAB is very efficient, and the computational cost is linear without any logarithmic factor. Table 3 and Figure 15 confirm the predicted $\mathcal{O}(n\log^2 n)$ complexity for the methods that we propose. The timings of the conjugate gradient are comparable to the sign iteration for small dimensions, but then the absence of the $\log^2(n)$ factor in the complexity is a big advantage for the former method.

All the proposed approaches seem to work better in terms of CPU time than the one reported by Haber and Verhaegen in [25], which uses a comparable (although slightly older) CPU. Moreover, their approach delivers only about two digits of accuracy with the selected parameter, while we get solutions with a relative error of about

| $n$ | $Mem_{\text{HODLR}}$ | $Mem_{\text{Sparse}}$ |
|---|---|---|
| 768 | 1,464.8 | 6,015.8 |
| 1,536 | 3,169.8 | 13,122 |
| 3,072 | 6,915.6 | 26,637 |
| 6,144 | 14,023 | 52,400 |
| 12,288 | 29,967 | $1.02 \cdot 10^5$ |
| 24,576 | 63,774 | $1.90 \cdot 10^5$ |
| 49,152 | $1.35 \cdot 10^5$ | $3.64 \cdot 10^5$ |
| 98,304 | $2.86 \cdot 10^5$ | $6.91 \cdot 10^5$ |
| $1.97 \cdot 10^5$ | $5.72 \cdot 10^5$ | $1.31 \cdot 10^6$ |

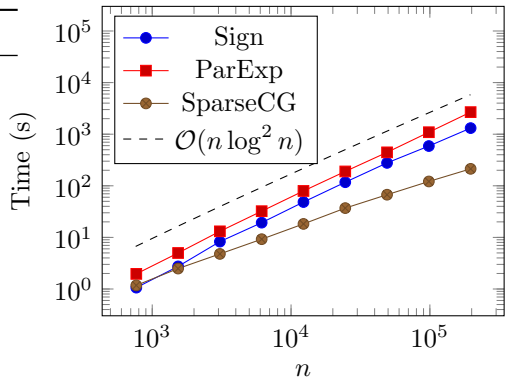


FIG. 15. *On the left, the memory consumption in storing the solution of the heat equation computed with* `ParExp` *and* `SparseCG`*, respectively. The first exploits the HODLR representation while the second makes use of the sparse format. The numerical values reported are in KB (Kilobytes). On the right, timings for the solution of the heat equation.*

$10^{-10}$ in the Frobenius norm.

The table in Figure 15 reports the memory usage when the solution is stored in the HODLR and in the sparse formats. We can see that the method using HODLR matrices, although slower, is more memory efficient compared to the `SparseCG` of a factor of about 2.

**6.3. Partial integro-differential equation.** Here, we consider a generalized Sylvester equation that has the structure described by Corollary 5.1 and arises from the discretization of the following PDE:

$$-\Delta u(x,y) + q(x,y) \int_{[0,1]^2} r(x,y)u(x,y)\,dx\,dy = f(x,y), \qquad (x,y) \in (0,1)^2, \tag{12}$$

where $q(x,y) = q_1(x)q_2(y)$ and $r(x,y) = r_1(x)r_2(y)$ are separable functions, and we assume zero Dirichlet boundary conditions. The discrete operator can be expressed in terms of the matrix equation $AX + XA + M_1 X M_2^T = C$, where $A = (n-1)^2 \cdot \text{trid}(-1, 2, -1)$, $C$ is the sampling of $f$ over the uniform grid $x_j = y_j = \frac{j-1}{n-1}$, $j = 1 \dots, n$, and

$$M_1 = \frac{1}{n-1} \begin{bmatrix} q_1(x_1) \\ q_1(x_2) \\ \vdots \\ q_1(x_{n-1}) \\ q_1(x_n) \end{bmatrix} \begin{bmatrix} \frac{1}{2} r_2(x_1) \\ r_2(x_2) \\ \vdots \\ r_2(x_{n-1}) \\ \frac{1}{2} r_2(x_n) \end{bmatrix}^T, \qquad M_2 = \frac{1}{n-1} \begin{bmatrix} q_2(x_1) \\ q_2(x_2) \\ \vdots \\ q_2(x_{n-1}) \\ q_2(x_n) \end{bmatrix} \begin{bmatrix} \frac{1}{2} r_1(x_1) \\ r_1(x_2) \\ \vdots \\ r_1(x_{n-1}) \\ \frac{1}{2} r_1(x_n) \end{bmatrix}^T.$$

In this experiment we consider $f(x,y) = \log(1 + |x-y|)$, and $q_j(x) = r_j(x) \equiv \sin(3x)$, $j = 1, 2$, so that $M_1 = M_2$. We test Algorithm 2, and the results are reported in Figure 16. The results report the timings of Algorithm 2 using the sign iteration for solving the first quasiseparable Lyapunov equation, and one can see that the timings of the overall procedure are just slightly larger than those reported for the example in section 6.1. Indeed, step 2 of Algorithm 2 is the dominating cost of the whole computation.

| $n$ | $T_{\text{Sign}}$ | $Res_{\text{Sign}}$ | QS rk |
|---|---|---|---|
| 512 | 0.89 | $4.20 \cdot 10^{-12}$ | 12 |
| 1,024 | 2.27 | $5.18 \cdot 10^{-12}$ | 10 |
| 2,048 | 5.33 | $1.11 \cdot 10^{-11}$ | 10 |
| 4,096 | 14.82 | $4.62 \cdot 10^{-11}$ | 11 |
| 8,192 | 37.93 | $3.09 \cdot 10^{-11}$ | 11 |
| 16,384 | 99.63 | $6.49 \cdot 10^{-11}$ | 12 |
| 32,768 | 268.37 | $1.19 \cdot 10^{-10}$ | 14 |
| 65,536 | 656.41 | $1.43 \cdot 10^{-10}$ | 14 |
| $1.31 \cdot 10^5$ | 1,565 | $1.52 \cdot 10^{-10}$ | 14 |

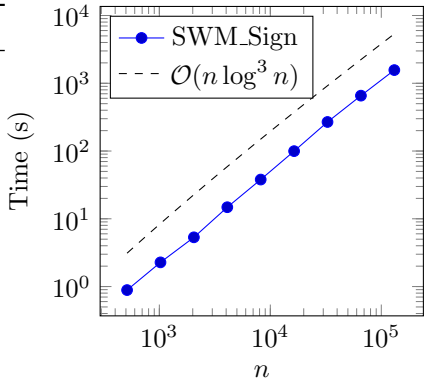


FIG. 16. *On the left, timings and features of the solution of the partial integro-differential equation for different grid sizes. The minimum block size is set to* 256 *and relative threshold in truncation is* $\epsilon = 10^{-12}$. *On the right, we plot timings for the solution of the generalized Lyapunov equation coming from the partial integro-differential equation.*

**7. Final remarks.** We have compared and analyzed two different strategies for the solution of some linear matrix equations with rank-structured data. We have presented some theoretical results that justify the feasibility of the approaches relying on tools from rational approximation. The techniques that we developed can be applied to treat the case of banded matrix coefficients in a natural way, thus providing an alternative approach to the ones presented in [25, 35]. Moreover, our methods still perform well when the conditioning of the coefficients increases. This allows us to cover a wider set of problems related to PDEs, such as those including the Laplacian operator.

Numerical tests confirm the scalability of the approach in treating large-scale instances. Our experiments show that the sign iteration is usually the fastest and most accurate method, although the procedure based on the integral formula can be more effective in parallel environments.

In the case of the asymptotically ill-conditioned coefficients in the matrix equation (such as for the 2D Laplacian), the complexity of the sign iteration is slightly worse than that of the integration formula ($\mathcal{O}(n \log^3 n)$ instead of $\mathcal{O}(n \log^2 n)$). This can make the latter method the most attractive choice for very large-scale problems. Moreover, relying on the arithmetic of hierarchically semiseparable matrices (HSS) [49], in place of HODLR, would likely further improve the proposed approach. This will be subject to future investigation. The main difficulty lies in creating a fast and reliable procedure for the computation of the inverse in HSS format.

**Acknowledgments.** We thank Daniel Kressner for useful discussions and suggestions. All the authors are members of the INdAM Research group GNCS, whose support is gratefully acknowledged.

REFERENCES

[1] A. C. ANTOULAS, *Approximation of Large-Scale Dynamical Systems*, Adv. Des. Control 6, SIAM, Philadelphia, 2005, https://doi.org/10.1137/1.9780898718713.

[2] A. C. Antoulas, D. C. Sorensen, and Y. Zhou, *On the decay rate of Hankel singular values and related issues*, Systems Control Lett., 46 (2002), pp. 323–342, https://doi.org/10.1016/S0167-6911(02)00147-0.

[3] C. Badea and B. Beckermann, *Spectral Sets*, 2nd ed., Handbook of Linear Algebra, Chapman and Hall/CRC, Boca Raton, FL, 2013, pp. 613–638.

[4] R. H. Bartels and G. W. Stewart, *Algorithm* 432: *Solution of the matrix equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820–826.

[5] U. Baur and P. Benner, *Factorized solution of Lyapunov equations based on hierarchical matrix arithmetic*, Computing, 78 (2006), pp. 211–234.

[6] B. Beckermann and A. Gryson, *Extremal rational functions on symmetric discrete sets and superlinear convergence of the ADI method*, Constr. Approx., 32 (2010), pp. 393–428, https://doi.org/10.1007/s00365-010-9087-6.

[7] B. Beckermann and A. Townsend, *On the singular values of matrices with displacement structure*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1227–1248, https://doi.org/10.1137/16M1096426.

[8] P. Benner and T. Breiten, *Low rank methods for a class of generalized Lyapunov equations and related issues*, Numer. Math., 124 (2013), pp. 441–470.

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

[10] D. A. Bini, S. Massei, and L. Robol, *Efficient cyclic reduction for Quasi-Birth–Death problems with rank structured blocks*, Appl. Numer. Math., 116 (2017), pp. 37–46, https://doi.org/10.1016/j.apnum.2016.06.014.

[11] D. A. Bini, S. Massei, and L. Robol, *On the decay of the off-diagonal singular values in cyclic reduction*, Linear Algebra Appl., 519 (2017), pp. 27–53, https://doi.org/10.1016/j.laa.2016.12.027.

[12] S. Börm, L. Grasedyck, and W. Hackbusch, *Hierarchical Matrices*, Lecture Note 21/(2003), Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig, Germany, 2003.

[13] J. P. Boyd, *Exponentially convergent Fourier-Chebyshev quadrature schemes on bounded and infinite intervals*, J. Sci. Comput., 2 (1987), pp. 99–109.

[14] D. Braess and W. Hackbusch, *Approximation of* $1/x$ *by exponential sums in* $[1, \infty)$, IMA J. Numer. Anal., 25 (2005), pp. 685–697.

[15] D. Braess and W. Hackbusch, *On the efficient computation of high-dimensional integrals and the approximation by exponential sums*, in Multiscale, Nonlinear and Adaptive Approximation, Springer, New York, 2009, pp. 39–74.

[16] T. Breiten, V. Simoncini, and M. Stoll, *Low-rank solvers for fractional differential equations*, Electron. Trans. Numer. Anal., 45 (2016), pp. 107–132.

[17] M. Crouzeix, *Numerical range and functional calculus in Hilbert space*, J. Funct. Anal., 244 (2007), pp. 668–690.

[18] M. Crouzeix and C. Palencia, *The numerical range is a* $(1 + \sqrt{2})$*-spectral set*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 649–655, https://doi.org/10.1137/17M1116672.

[19] Y. Eidelman, I. Gohberg, and I. Haimovici, *Separable Type Representations of Matrices and Fast Algorithms, Vol.* 1, Oper. Theory Adv. Appl. 234, Birkhäuser/Springer, Basel, 2014.

[20] G. H. Golub and C. F. Van Loan, *Matrix Computations*, *Vol.* 3, Johns Hopkins University Press, Baltimore, MD, 2012.

[21] L. Grasedyck, *Existence of a low rank or* $\mathscr{H}$*-matrix approximant to the solution of a Sylvester equation*, Numer. Linear Algebra Appl., 11 (2004), pp. 371–389, https://doi.org/10.1002/nla.366.

[22] L. Grasedyck, W. Hackbusch, and B. N. Khoromskij, *Solution of large scale algebraic matrix Riccati equations by use of hierarchical matrices*, Computing, 70 (2003), pp. 121–165, https://doi.org/10.1007/s00607-002-1470-0.

[23] S. Güttel, E. Polizzi, P. T. P. Tang, and G. Viaud, *Zolotarev quadrature rules and load balancing for the FEAST eigensolver*, SIAM J. Sci. Comput., 37 (2015), pp. A2100–A2122, https://doi.org/10.1137/140980090.

[24] A. Haber, *Estimation and Control of Large-Scale Systems with an Application to Adaptive Optics for EUV Lithography*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 2014.

[25] A. Haber and M. Verhaegen, *Sparse solution of the Lyapunov equation for large-scale interconnected systems*, Automatica J. IFAC, 73 (2016), pp. 256–268, https://doi.org/10.1016/j.automatica.2016.06.002.

[26] W. Hackbusch, *A sparse matrix arithmetic based on H-matrices. Part* I: *Introduction to H-matrices*, Computing, 62 (1999), pp. 89–108.

[27] W. Hackbusch, *Hierarchical Matrices: Algorithms and Analysis*, Springer Ser. Comput. Math. 49, Springer, Heidelberg, 2015, https://doi.org/10.1007/978-3-662-47324-5.
[28] W. Hackbusch, B. Khoromskij, and S. A. Sauter, *On $\mathcal{H}^2$-matrices*, in Lectures on Applied Mathematics (Munich, 1999), Springer, Berlin, 2000, pp. 9–29.
[29] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898718027.
[30] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008, https://doi.org/10.1137/1.9780898717778.
[31] E. Jarlebring, G. Mele, D. Palitta, and E. Ringh, *Krylov methods for low-rank commuting generalized Sylvester equations*, Numer. Linear Algebra Appl., to appear.
[32] S. Massei, *Exploiting Rank Structures in the Numerical Solution of Markov Chains and Matrix Functions*, Ph.D. thesis, Scuola Normale Superiore di Pisa, Pisa, Italy, 2017.
[33] S. Massei and L. Robol, *Decay bounds for the numerical quasiseparable preservation in matrix functions*, Linear Algebra Appl., 516 (2017), pp. 212–242, https://doi.org/10.1016/j.laa.2016.11.041.
[34] D. Palitta and V. Simoncini, *Matrix-equation-based strategies for convection-diffusion equations*, BIT, 56 (2016), pp. 751–776, https://doi.org/10.1007/s10543-015-0575-8.
[35] D. Palitta and V. Simoncini, *Numerical Methods for Large-Scale Lyapunov Equations with Symmetric Banded Data*, preprint, https://arxiv.org/abs/1711.04187, 2017.
[36] T. Penzl, *Eigenvalue decay bounds for solutions of Lyapunov equations: The symmetric case*, Systems Control Lett., 40 (2000), pp. 139–144, https://doi.org/10.1016/S0167-6911(00)00010-4.
[37] M. Popolizio and V. Simoncini, *Acceleration techniques for approximating the matrix exponential operator*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 657–683, https://doi.org/10.1137/060672856.
[38] E. Ringh, G. Mele, J. Karlsson, and E. Jarlebring, *Sylvester-based preconditioning for the waveguide eigenvalue problem*, Linear Algebra Appl., 542 (2018), pp. 441–463.
[39] J. D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. Control, 32 (1980), pp. 677–687, https://doi.org/10.1080/00207178008922881.
[40] Y. Saad, *Numerical solution of large Lyapunov equations*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods (Amsterdam, 1989), Progr. Syst. Control Theory 5, Birkhäuser Boston, Boston, MA, 1990, pp. 503–511.
[41] V. Simoncini, *A new iterative method for solving large-scale Lyapunov matrix equations*, SIAM J. Sci. Comput., 29 (2007), pp. 1268–1288, https://doi.org/10.1137/06066120X.
[42] V. Simoncini, *Computational methods for linear matrix equations*, SIAM Rev., 58 (2016), pp. 377–441, https://doi.org/10.1137/130912839.
[43] D. C. Sorensen and Y. Zhou, *Bounds on Eigenvalue Decay Rates and Sensitivity of Solutions to Lyapunov Equations*, Tech. Rep. TR02-07, Dept. of Comp. Appl. Math., Rice Univ., Houston, TX, 2002.
[44] A. Townsend and L. N. Trefethen, *An extension of Chebfun to two dimensions*, SIAM J. Sci. Comput., 35 (2013), pp. C495–C518, https://doi.org/10.1137/130908002.
[45] L. N. Trefethen, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87, https://doi.org/10.1137/060659831.
[46] R. Vandebril, M. V. Barel, G. Golub, and N. Mastronardi, *A bibliography on semiseparable matrices*, Calcolo, 42 (2005), pp. 249–270.
[47] R. Vandebril, M. Van Barel, and N. Mastronardi, *Matrix Computations and Semiseparable Matrices. Eigenvalue and Singular Value Methods, Vol.* 2, Johns Hopkins University Press, Baltimore, MD, 2008.
[48] R. Vandebril, M. Van Barel, and N. Mastronardi, *Matrix Computations and Semiseparable Matrices. Linear Systems, Vol.* 1, Johns Hopkins University Press, Baltimore, MD, 2008.
[49] J. Xia, S. Chandrasekaran, M. Gu, and X. S. Li, *Fast algorithms for hierarchically semiseparable matrices*, Numer. Linear Algebra Appl., 17 (2010), pp. 953–976, https://doi.org/10.1002/nla.691.
[50] E. Zolotarev, *Application of elliptic functions to questions of functions deviating least and most from zero*, Zap. Imp. Akad. Nauk. St. Petersburg, 30 (1877), pp. 1–59.