Numer. Math. (2007) 106:511–528
DOI 10.1007/s00211-007-0072-3

Numerische Mathematik

Averaging operators for exponential splittings

Marko Huhtanen

Received: 18 September 2006 / Revised: 26 February 2007 / Published online: 31 March 2007

Abstract Averaging operations are considered in connection with exponential splitting methods. Toeplitz plus Hankel related matrices are resplit by applying appropriate averaging operators leading to a hierarchy of structured matrices. With the resulting parts, the option of using exponential splitting methods becomes available. A related, seemingly important group of unitary unipotents is looked at. Based on a formula due to Lenard, a very fast iterative method to find the nearest Toeplitz plus Hankel matrix in the Frobenius norm is devised.

Mathematics Subject Classification (2000) 65F30 · 65L05

1 Introduction

Only rarely a square matrix can be split such that the exponentials of the parts are readily computable, while requiring a negligible amount of storage. By readily we mean in $O(n)$, or perhaps $O(n \log^k n)$ operations, for a small $k \in \mathbb{N}$. Then, in large scale time integration, the option of using exponential splitting methods becomes available. A prime example of this consists of matrices that can be represented as the sum of a circulant and a diagonal matrix (see, e.g., [10]). In this paper, we consider other such instances by applying averaging operators and splittings arising in this manner.

Averagings of matrices has been considered in [4,5] for pinching and truncating of matrices. In a more general context, the averaging operation (see Sect. 2.1 for its definition) can be viewed as a general procedure to generate structured matrices, as well as splittings, in an algebraic way. Occasionally related matrices admit a rapid

M. Huhtanen (✉)
Institute of Mathematics, Helsinki University of Technology, Box 1100, 02015 Hut, Finland
e-mail: Marko.Huhtanen@hut.fi

computation of their exponential, or offer an opportunity to apply exponential splitting methods. This is exactly the case with the sum of a circulant and a diagonal matrix. Completely analogously, we resplit Toeplitz plus Hankel matrices in terms of two averaging operators such that the exponentials of the parts are readily computable. For this we employ appropriate unitary unipotents involving the Fourier matrix, hence allowing us to look at the set of Toeplitz plus Hankel matrices as an element of a more general hierarchy of structured matrices having similar properties. Small rank and diagonal perturbations of Toeplitz plus Hankel matrices are also looked at. For applications where such matrices appear, see [6] (and its forward citations in MathSciNet).

In connection with these considerations we propose a very fast iterative method, based on the formula introduced by Lenard [14], to inexpensively approximate a given matrix with a Toeplitz plus Hankel matrix in the Frobenius norm. This is a natural problem while dealing with two different averaging operators in the special case of Toeplitz plus Hankel matrices. The formula of Lenard has undoubtedly other applications and deserves more attention.

This paper is organized as follows. In Sect. 2, the averaging operator is introduced together with the two most central examples. For more averaging operators, these two examples give rise to a seemingly important group of unitary unipotents. With appropriate averagings, a resplitting of Toeplitz plus Hankel matrices is constructed. This approach supports a general hierarchy of structured matrices which is then described. In Sect. 3, these ideas are used to apply exponential splitting methods with Toeplitz plus Hankel matrices. Small rank perturbations, which typically can be expected to result from boundary conditions in applications, are allowed. Diagonal perturbations can also be handled. This is done with Krylov subspace methods. In Sect. 4, an iterative method for finding a nearest Toeplitz plus Hankel matrix is devised.

2 Averagings, unipotents and splitting Toeplitz plus Hankel matrices

We look at the general procedure of averaging of matrices and consider it in connection with splitting Toeplitz plus Hankel related matrices.

2.1 Averaging of matrices

Let $J \in \mathbb{C}^{n\times n}$ be a unipotent[1] matrix, i.e.,

$$J^k = I, \tag{2.1}$$

the identity, for some $k \in \mathbb{N}$. (The study of this matrix equation is also of separate interest [1].) For convenience, k is assumed to be the smallest natural number

[1] There is another definition of unipotency requiring $(I - J)^k = 0$ instead.

satisfying (2.1). Motivated by [4], the averaging of $M \in \mathbb{C}^{n\times n}$ with respect to J means forming

$$A = \frac{1}{k}\sum_{j=0}^{k-1} J^j M J^{-j}. \tag{2.2}$$

Clearly, A commutes with J. The range of this averaging operator, which we call the set (or algebra) of averagings with respect to J, is the centralizer of J, i.e., the set of matrices commuting with it. In case J is unitary, the averaging belongs to the relatively scarce family of readily computable, as well as useful linear operators on $\mathbb{C}^{n\times n}$. Then $\|A\| \le \|M\|$ in any unitarily invariant norm $\|\cdot\|$.

To give the two most central examples of unitary unipotents with $k = n$, let J_c be the unitary circulant matrix with ones on the first sub-diagonal and at the position $(1, n)$. Then A is circulant. For the other one, in [4] one is concerned with the diagonal matrix

$$J_d = e^D, \quad \text{where} \quad D = \frac{2\pi i}{n}\mathrm{diag}(0, 1, 2, \ldots, n-1),$$

in which case A equals the diagonal matrix having the diagonal of M.

Example 1 In the sections that follow we are simultaneously interested in the ranges of two different averaging operators. In the case of J_c and J_d, the sum of the ranges obviously consists of circulant plus diagonal matrices.

For a familiar example with a small k (that also connects these two above averagings), take J to be the Fourier matrix F_n to have $k = 4$. Recall also that with an involution $k = 2$; see [18] for its application to computing the matrix exponential.

Averaging is a smoothing operation in the following sense.

Theorem 2.1 *Let $J \in \mathbb{C}^{n\times n}$ be unitary and unipotent. Then (2.2) is the nearest matrix in the Frobenius norm to $M \in \mathbb{C}^{n\times n}$ from the centralizer of J.*

Proof Let J be unitarily diagonalized as $J = U\Lambda U^*$ with the eigenvalues ordered counterclockwise in Λ. Since A commutes with J, it forces U^*AU to be block-diagonal with the blocks of size the number of equaling eigenvalues in Λ. By the unitary invariance of the Frobenius norm, we may look at the difference $U^*MU - U^*AU$ from which the claim follows since in the blocks the matrix is not altered. □

To generate more averaging operators from the ones available, obviously a power of a unipotent is unipotent. Also, the Kronecker product of two unipotents is again unipotent.

Example 2 In practice the *preaverages* $A_l = \frac{1}{l}\sum_{j=0}^{l-1} J^j M J^{-j}$ for $l \le k$ are also of some interest through the splitting $M = A_l + (M - A_l)$. The case $l = 2$ appears in the displacement structure considerations to deal with Toeplitz-like matrices; see [11] and references therein. Then one is concerned with the rank of the part $M - A_l$ with J being J_c.

The antilinear analogy of (2.2) is defined as follows. Denote by τ the standard conjugation operator on $\mathbb{C}^n$, i.e., $\tau x = \overline{x}$. Let $J \in \mathbb{C}^{n\times n}$ be such that $(J\tau)^k = I$. For this, k must be even such that $(J\overline{J})^{k/2} = I$, i.e., the matrix $J\overline{J}$ is unipotent. Then the antilinear averaging of M means forming

$$A = \frac{1}{k}\sum_{j=0}^{k-1}(J\tau)^j M(J\tau)^{-j}.$$

Observe that $(J\tau)^{-1} = \overline{J}^{-1}\tau$ and $J\tau A = AJ\tau$, i.e., $J\overline{A} = AJ$. In case J is unitary, again $\|A\| \le \|M\|$ holds in any unitarily invariant norm $\|\cdot\|$, and an analogy of Theorem 2.1 can be readily proved by replacing the centralizer of J with the set of matrices commuting with the operator $J\tau$. For instance, with J_c one gets "conjugate circulant" matrices, i.e., then the diagonals of A appear cyclically as with circulant matrices, except that now on each diagonal we have $a_{s+1,t+1} = \overline{a_{s,t}}$ for the consecutive entries. (See [2] for conjugate Toeplitz matrices.)

2.2 A group of unitary unipotents

For more averaging operators in terms of J_c and J_d, we look more carefully at a seemingly important group of unitary unipotents appearing in computational harmonic analysis[2] [13] as well as in approximation techniques in $C^\star$-algebras [7, p. 174]. (Observe that, in general, the product of two unipotents is not unipotent.) To this end we are concerned with the products $J_c^j J_d^l$ with $1 \le j, l \le n-1$. For the order of the terms we have

$$J_c^j J_d^l = e^{-2\pi ijl/n} J_d^l J_c^j, \tag{2.3}$$

i.e., J_c^j and J_d^l are $e^{-2\pi ijl/n}$-commutative; see [8] for this concept with canonical forms and interesting historical remarks. This kind of relaxed commutativity is rare. In fact, if two invertible $U, V \in \mathbb{C}^{n\times n}$ satisfy

$$UV = \mu VU$$

for some $\mu \in \mathbb{C}$, then $U^j V U^{-j} = \mu^j V$ for any $j \in \mathbb{Z}$. Hence, the finiteness of the spectrum of V forces $\mu = e^{i\theta}$ with $\frac{n}{2\pi}\theta \in \mathbb{N} \cup \{0\}$. If λ is an eigenvalue of V, then so is $e^{ki\theta}\lambda$ for any $k \in \mathbb{Z}$. Thus, the smallest possible non-zero value for θ is $2\pi/n$, in which case all the eigenvalues of V are distinct and located on a circle centered at the origin. This is the case with J_c and J_d.

Observe that $J_c^j J_d^l$ is necessarily unipotent since (2.3) yields

$$(J_c^j J_d^l)^n = e^{-\frac{(n-1)n}{2}2\pi il/n} I = \pm I,$$

[2] In computational harmonic analysis J_c is called a "translation" and J_d a "modulation".

so that $k \leq 2n$. To recover the exact value of k, there is a finite step algorithm for the eigenvalues and eigenvectors that builds on the Euclidean algorithm for finding the greatest common divisor of j and n during the process. To describe this, suppose $\lambda \neq 0$, and perform Gaussian elimination steps on $J_c^j J_d^l - \lambda I$ to reach its echelon form. Except its last j columns, the echelon form is diagonal with $-\lambda$ on its diagonal. These last j columns can be given in j-by-j blocks with the exponentials of diagonal matrices.

For this, let $n = jt + r$ with $t \in \mathbb{N}$ and $0 \leq r < j$. We may assume $j \leq \lfloor \frac{n}{2} \rfloor$, otherwise consider the transpose multiplied by $e^{\frac{-i2\pi jl}{n}}$. Starting from the top, for $s = 2, \ldots, t-1$, the blocks can be computed to be $\frac{1}{\lambda^{s-1}} e^{M_s}$, where

$$M_s = \frac{i2\pi l}{n} \left(\operatorname{diag}(0, s, 2s, \ldots, (j-1)s) + \frac{js(s-3)}{2} I \right).$$

The tth block is $\frac{1}{\lambda^{t-1}} e^{M_t} - \lambda K^r$, where K is the nilpotent forward shift having ones on its first sub-diagonal. Then, once the Gaussian elimination steps have been completed for the first $n-j$ columns, the j-by-j block in the lower right corner of the eliminated $J_c^j J_d^l - \lambda I$ is

$$J_c^{j-r} e^{i\pi l(n-r)(t-3)/n} \left(\frac{e^{D_1}}{\lambda^t} \oplus \frac{e^{D_2}}{\lambda^{t-1}} \right) - \lambda I, \tag{2.4}$$

where J_c is now of size j-by-j and

$$D_1 = \frac{i2\pi l(t-1)}{n}(0, 1, 2, \ldots, r-1) \quad \text{and} \quad D_2 = \frac{i2\pi lt}{n}(r, r+1, \ldots, j-1).$$

The simplest case occurs when j divides n, i.e., when the Euclidean algorithm stops just after one step.

Theorem 2.2 *Suppose $1 \leq j, l \leq n-1$ and j divides n. Let a be the greatest common divisor of j and l. Then the eigenvalues of $J_c^j J_d^l$ are*

$$\lambda = e^{i\pi l(1-\frac{j}{n})} e^{\frac{i2\pi ar}{n}}$$

with $r = 0, 1, \ldots, \frac{n}{b} - 1$, where b is the greatest common divisor of a and n.

Proof Because j divides n, we have $r = 0$ so that $J_c^{j-r} = I$ and only D_2 appears in (2.4). With these the spectrum of $J_c^j J_d^l$ is obtained once the j diagonal entries of (2.4) are set to equal zero. This gives us the eigenvalues[3] in a closed form

$$\lambda = e^{i\pi l(1-\frac{3j}{n})} e^{i2\pi(lm+jp)/n} \tag{2.5}$$

[3] With this it is then a simple task to find the eigenvectors with back substitution.

for the $(m+1)$th diagonal entry with $m = 0, 1, \ldots, j-1$, while $p \in \mathbb{Z}$. By elementary divisibility theory, $ls + jp$ has exactly the values that are divisible by the greatest common divisor of l and j. □

If j does not divide n, then by taking the transpose of (2.4) yields

$$J_c^r e^{i\pi l(n-r)(t-3)/n} \left(\frac{e^{D_2}}{\lambda^{t-1}} \oplus \frac{e^{D_1}}{\lambda^t} \right) - \lambda I$$

which is still structured similarly to $J_c^j J_d^l - \lambda I$. With this the elimination steps can be performed analogously in a closed form, followed by a transposition. This is repeated until the appearing power of J_c is the identity, i.e., the greatest common divisor of j and n has been obtained. Then what remains is to construct the eigenvalues by elementary calculations from the diagonal elements of the diminished block like in (2.5).

2.3 Splitting Toeplitz plus Hankel matrices and their generalizations

For circulant matrices fundamental linear algebra computations are very inexpensive with the FFT. Even though also Toeplitz matrices have many special properties, it seems that no substantial savings can be achieved in the basic task of computing the matrix exponential. For the possibility to apply splitting methods, we look from the outset at the more general set of Toeplitz plus Hankel matrices and resplit them in terms of two appropriate averaging operators.

To this end, denote by B the backward identity, i.e., the permutation matrix having ones on the diagonal joining the left lower corner with the right upper corner. Then consider matrices that can be represented as the sum

$$C_1 + BC_2 \tag{2.6}$$

with C_1 and C_2 circulant matrices. Analogously, look at matrices with C_1 and C_2 replaced with skew-circulant matrices S_1 and S_2. By the fact that the parts of any Toeplitz plus Hankel matrix

$$M = T + H \in \mathbb{C}^{n\times n}$$

can be decomposed as

$$T = C_1 + S_1 \quad \text{and} \quad H = B(C_2 + S_2),$$

by choosing circulant and skew-circulant matrices appropriately, we look at the splitting

$$M = (C_1 + BC_2) + (S_1 + BS_2) \tag{2.7}$$

after reshuffling the terms. It is these parts that are of interest to us.

Theorem 2.3 *Let B be the backward identity and n even. The set of matrices*

$$C_1 + BC_2, \quad \text{with } C_1 \text{ and } C_2 \text{ circulant matrices},$$

is a $C^\star$-subalgebra of $\mathbb{C}^{n\times n}$. It equals the set of averagings with respect to the unitary unipotent $J = F_n^* D F_n$ with $k = n/2$, where F_n is the Fourier matrix and

$$D = \operatorname{diag}(1, \omega, \omega^2, \ldots, \omega^{\frac{n}{2}-1}, 1, \omega^{\frac{n}{2}-1}, \omega^{\frac{n}{2}-2}, \ldots, \omega) \tag{2.8}$$

with $\omega = e^{4\pi i/n}$.

Proof For the backward identity B we have $BC = C^T B$ for any circulant matrix C. Also, B is an involution, i.e., $B^2 = I$ holds. Hence, we obviously have a vector space over $\mathbb{C}$ and, for the product, let C_j be circulant matrices for $j = 1, \ldots, 4$. Then

$$(C_1 + BC_2)(C_3 + BC_4) = C_1C_3 + C_2^T C_4 + B(C_2C_3 + C_1^T C_4).$$

Thus, we have an algebra which is also closed under taking the Hermitian transpose. Therefore, we are dealing with a $C^\star$-algebra.

For the second part of the claim, consider a matrix carrying the structure $C_1 + BC_2$ with circulant matrices C_1 and C_2. Take the Fourier matrix F_n so that $F_nC_1F_n^* = \operatorname{diag}(d_1^1, d_2^1, \ldots, d_n^1) = D_1$ is diagonal. Moreover, we have

$$F_nBC_2F_n^* = \begin{bmatrix} d_1^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & d_2^2 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & 0 \\ 0 & d_n^2 & \cdots & 0 \end{bmatrix} = D_2, \tag{2.9}$$

i.e., we obtain a matrix whose (1,1)-entry and the first sub-anti-diagonal entries are possibly nonzero. Consequently, for $j = 1, 2, \ldots, \lfloor \frac{n-1}{2} \rfloor$, the orthogonal subspaces $\operatorname{span}\{e_1, e_{n/2+1}\}$ and $\operatorname{span}\{e_{j+1}, e_{n+1-j}\}$ are invariant for $F_n(C_1 + BC_2)F_n^* = D_1 + D_2$. From this it follows that the claimed D provides the commuting unipotent once we set $J = F_n^* D F_n$. □

Once the construction is clear, for skew-circulant matrices we have an analogous claim such that the set of matrices

$$S_1 + BS_2, \quad \text{with } S_1 \text{ and } S_2 \text{ skew-circulant matrices}, \tag{2.10}$$

is a $C^\star$-subalgebra of $\mathbb{C}^{n\times n}$. With n even, take $\sigma = e^{\pi i/n}$ and set

$$\Omega^{1/2} = \operatorname{diag}(1, \sigma, \sigma^2, \ldots, \sigma^{n-1}). \tag{2.11}$$

Then $(F_n\overline{\Omega}^{1/2})S_1(F_n\overline{\Omega}^{1/2})^* = \operatorname{diag}(\hat{d}_1^1, \hat{d}_2^1, \ldots, \hat{d}_n^1) = \hat{D}_1$ whereas

$$(F_n\overline{\Omega}^{1/2})BS_2(F_n\overline{\Omega}^{1/2})^* = \begin{bmatrix} 0 & 0 & \cdots & \hat{d}_1^2 \\ 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \hat{d}_{n-1}^2 & \cdots & 0 \\ \hat{d}_n^2 & 0 & \cdots & 0 \end{bmatrix} = \hat{D}_2, \tag{2.12}$$

i.e., only the anti-diagonal entries can be nonzero. Therefore the orthogonal subspaces $\operatorname{span}\{e_j, e_{n+1-j}\}$, for $j = 1, 2, \ldots, n$, are invariant for

$$(F_n\overline{\Omega}^{1/2})(S_1 + BS_2)(F_n\overline{\Omega}^{1/2})^*.$$

As a result, to have the set of matrices (2.10), the averaging is now taken with respect to the unitary unipotent $\hat{J} = (F_n\overline{\Omega}^{1/2})^*\hat{D}(F_n\overline{\Omega}^{1/2})$, where

$$\hat{D} = \operatorname{diag}(1, \omega, \omega^2, \ldots, \omega^{\frac{n}{2}-1}, \omega^{\frac{n}{2}-1}, \omega^{\frac{n}{2}-2}, \ldots, \omega, 1) \tag{2.13}$$

with $\omega = e^{4\pi i/n}$ and $k = n/2$.

In conclusion, besides circulant plus diagonal matrices, important families of matrices can be represented as the sum of appropriate averagings. Observe that neither the set of Toeplitz nor the set of Hankel matrices equals the averagings with respect to a unipotent. (Proof: The range of an averaging operator is an algebra, by being the centralizer of a unipotent. The set of Toeplitz (or Hankel) matrices is not an algebra.) As opposed to circulant plus diagonal matrices, now $k = n/2$ for the unipotents involved, reflecting the fact that dimension of the subspaces consisting of Toeplitz plus Hankel matrices is approximately twice the dimension of the subspace consisting of circulant plus diagonal matrices. In particular, a general matrix structure of Toeplitz plus Hankel-type is now readily available as follows.

Definition 2.4 Let $J = F_n^* D F_n$ and $\hat{J} = (F_n\overline{\Omega}^{1/2})^*\hat{D}(F_n\overline{\Omega}^{1/2})$ with F_n the Fourier matrix and $\Omega^{1/2}$ defined in (2.11). With n even, if D and $\hat{D}$ are diagonal unipotents with $k = n/2$, then the sum of the ranges of the averaging operations with J and $\hat{J}$ is a Toeplitz plus Hankel-type matrix structure.

This should make the general hierarchy clear once k is allowed to vary. Decreasing the value of k yields us less and less structured matrices. Namely, with $k = n$ we have the set of Toeplitz matrices (any Toplitz matrix is the sum of a circulant and a skew-circulant matrix) while with $k = 1$ we have the whole $\mathbb{C}^{n\times n}$. For large values of k the computation of the exponentials (and any basic linear algebra operations) of the parts is inexpensive with the FFT. This is illustrated in the section that follows with Toeplitz plus Hankel matrices and their small rank perturbations.

3 Exponential splittings

In large scale problems applying the exponential to a vector is computationally a very demanding task. In case matrix–vector products are inexpensive, Krylov subspace methods can be executed at least in the Hermitian case. In non-Hermitian problems their behaviour is poorly understood—a fact that is reflected by the largely missing preconditioning ideas.

Exponential splitting methods occasionally yield another option to carry out the time integration. Then the application of the exponentials of the parts to a vector must be inexpensive, combined with negligible storage requirements. As is well known, this is the case with the sum of a circulant and a diagonal matrix. With the splitting of Sect. 2.3, we can now apply exponential splitting methods with Toeplitz plus Hankel matrices.

3.1 Exponential splittings for Toeplitz plus Hankel matrices and their perturbations

For a Toeplitz plus Hankel matrix M, consider its resplitting (2.7). To find the exponential of the part $C_1 + BC_2$, use

$$e^{C_1+BC_2} = F_n^* e^{D_1+D_2} F_n,$$

where D_1 is diagonal and D_2 is defined in (2.9). With this it takes approximately the same number of floating point operations to have the exponential as it takes to compute the exponential of a circulant matrix. The storage consumption is of the same order. This is due to the fact that finding $e^{D_1+D_2}$ reduces to computing the exponentials of 2-by-2 matrices by using the invariant subspaces constructed in the proof of Theorem 2.3. Moreover, with $D_1 + D_2$ the computation of the norm of $C_1 + BC_2$, which is needed for choosing the step length in the splitting methods, costs only $O(n)$ floating point operations.

Analogously, the exponential of $S_1 + BS_2$ can be computed in $O(n \log n)$ floating point operations with $O(n)$ storage by using (2.12) and the FFT.

The following lemmas follows readily.

Lemma 3.1 *Let $C_2 = \alpha I$ with $\alpha \in \mathbb{C}$ in (2.7). Then $d_1^2 = d_2^2 = \cdots = d_n^2 = \alpha$ in (2.9).*

Lemma 3.2 *Let $S_2 = \beta I$ with $\beta \in \mathbb{C}$ in (2.7). Then $\hat{d}_1^2 = \hat{d}_2^2 = \cdots = \hat{d}_n^2 = \beta$ in (2.12).*

In particular, the splitting (2.7) is not unique. With $\alpha, \beta \in \mathbb{C}$ we can equally well take

$$M = \underbrace{(C_1 + \alpha I) + B(C_2 + \beta I)}_{K} + \underbrace{(S_1 - \alpha I) + B(S_2 - \beta I)}_{L} \tag{3.1}$$

and still preserve the necessary structure for inexpensive computations, i.e., we have the freedom to distribute I and B between K and L. It appears reasonable to require

$$\min_{\alpha,\beta} \max\{\|K(\alpha,\beta)\|_2, \|L(\alpha,\beta)\|_2\}$$

motivated by the aim at maximizing the step length $\triangle t$. (For error analysis of splitting methods, see [10,12].) For a computationally more tractable alternative, employ the Frobenius norm and look at

$$\min_{\alpha,\beta} \left(\|K(\alpha,\beta)\|_F^2 + \|L(\alpha,\beta)\|_F^2 \right).$$

With n even set $\alpha = s + it$ and $\beta = u + iv$ and solve

$$\begin{bmatrix} 2n & 0 & 2 & 0 \\ 0 & 2n & 0 & 2 \\ 2 & 0 & 2n & 0 \\ 0 & 2 & 0 & 2n \end{bmatrix} \begin{bmatrix} s \\ t \\ u \\ v \end{bmatrix} = -\sum_{j=1}^{n} \begin{bmatrix} \mathrm{Re}\,(d_j^1 + \hat{d}_j^1) \\ \mathrm{Im}\,(d_j^1 + \hat{d}_j^1) \\ \mathrm{Re}\,(d_j^2 + \hat{d}_j^2) \\ \mathrm{Im}\,(d_j^2 + \hat{d}_j^2) \end{bmatrix} \tag{3.2}$$

to locate the unique critical point. Consuming only $O(n)$ floating point operations, this yields an inexpensive way to choose α and β in (3.1).

In practice one often encounters perturbations of Toeplitz plus Hankel matrices. Toeplitz plus diagonal matrices belong to this category; see [16] and references therein. For another example, boundary conditions typically cause small rank perturbations to structured matrices arising in applications [6]. To deal with these, assume more generally that

$$M = F + T + H \tag{3.3}$$

with a perturbation F while T is a Toeplitz and H a Hankel matrix. Then split M as $M = F + K + L$, where K and L are defined in (3.1) with the parameters α and β chosen according to (3.2). Using these three parts we can take

$$e^{\frac{\triangle t}{2}L} e^{\frac{\triangle t}{2}K} e^{\triangle t F} e^{\frac{\triangle t}{2}K} e^{\frac{\triangle t}{2}L} \tag{3.4}$$

to approximate $e^{\triangle t A}$. By inspecting the Taylor expansions involved, we have a second order method such that with $F = 0$ this reduces to the classical Strang splitting method $e^{\frac{\triangle t}{2}L} e^{\triangle t K} e^{\frac{\triangle t}{2}L}$ [17]. To employ (3.4), an inexpensive way to apply the exponential of F is needed since with the techniques just introduced we can handle the exponentials of K and L. For diagonal matrices the application is obvious while the case of small rank is considered in the section that follows.

3.2 Krylov subspace methods and low degree splittings

Recall that if F is of small rank, then it is of low degree by the fact that

$$\deg(F) \leq \operatorname{rank}(F) + 1,$$

where $\deg(F)$ denotes the degree of the minimal polynomial of F. Let $v \in \mathbb{C}^n$ and consider first computing $e^F v$ for a low degree matrix F.

For this purpose, generate an orthonormal basis $q_1, q_2, \ldots, q_k$ of the Krylov subspace

$$\mathcal{K}(F; v) = \operatorname{span}\{v, Fv, F^2 v, \ldots\}$$

with the Arnoldi method. Since

$$k = \dim(\mathcal{K}(F; v)) \leq \deg(F),$$

for small k this is economical. Then $FQ_k = Q_k H_k$, where H_k is a Hessenberg matrix and $Q_k = [q_1\, q_2 \ldots q_k] \in \mathbb{C}^{n\times k}$. Computing e^{H_k} explicitly with direct methods yields us the formula

$$e^F v = Q_k e^{H_k} Q_k^* v. \tag{3.5}$$

In principle, this can be combined with exponential splitting methods due to the following low degree splitting.

Theorem 3.3 *Suppose* $M \in \mathbb{C}^{n\times n}$ *with* $\deg(M) = n$. *Then* $M = F_1 + F_2$ *with* $\deg(F_1) \leq 2$ *and* $\deg(F_2) \leq 4$.

Proof Since the degree of the minimal polynomial of M is the dimension of $\mathbb{C}^n$, we have $M = XCX^{-1}$ for a companion matrix C and an invertible $X \in \mathbb{C}^{n\times n}$. Split C as $C = C_1 + C_2$ with C_1 having ones at the positions $(2j, 2j-1)$ for $j = 1, 2, \ldots,$ while the other entries of C_1 are zeros. Hence $\deg(XC_1X^{-1}) = \deg(C_1) = 2$.

Now $C_2 = C - C_1$ has ones at the positions $(2j+1, 2j)$ for $j = 1, 2, \ldots,$ while its other entries are zeros, except for its last columns. Hence C_2 and thereby XC_2X^{-1} is the sum of a rank-one matrix and a matrix of degree two. By [9, Proposition 2.6] we can conclude that the degree of C_2 is at most four. □

At present this combination of Krylov subspace and splitting methods is of theoretical interest only since we do not know how to inexpensively split a non-derogatory matrix M as the sum of two low degree matrices.

Consider next using the formula (3.5) with the splitting (3.3).

Example 3 Suppose $F = D + G$ with a diagonal matrix D having j distinct eigenvalues and G of small rank. Then

$$\deg(F) \leq j(\operatorname{rank}(G) + 1)$$

by [9, Proposition 2.6]. A matrix–vector product with F costs $O((\text{rank}\,(G)+1)n)$ operations and therefore applying the formula (3.5) is inexpensive for moderate values of j and rank (G). Observe though that in applying the exponential splitting method (3.4), Q_k needs to be generated at each step.

The case of F being merely of small rank is much less costly since Q_k needs to computed just once. With this the formula (3.5) can be used to compute $e^F c$ for any $c \in \mathbb{C}^n$, by making the generically valid assumption[4]

$$\dim\left(\mathcal{K}(F;v)\right) = \deg(F). \tag{3.6}$$

Namely, by splitting c as $c = c_1 + c_2$ with $c_1 \in \mathcal{K}(F;v)$ and c_2 in the orthogonal complement of $\mathcal{K}(F;v)$, we have

$$e^F c = Q_k e^{H_k} Q_k^* c_1 + e^F c_2. \tag{3.7}$$

The first part is obtained inexpensively. To have the second part, find $w \in \mathbb{C}^n$ solving the linear system $Fw = Fc_2$. If (3.6) is satisfied, then $Fc_2 \in \mathcal{K}(F;v)$ and the solution is given by $w = Q_k \hat{w}$, where $\hat{w}$ solves the k-by-k linear system $H_k \hat{w} = Q_k^* F c_2$. This yields us then

$$e^F c_2 = c_2 - w + e^F w = c_2 - w + Q_k e^{H_k} Q_k^* w \tag{3.8}$$

by the fact that

$$e^F c_2 = c_2 + \left(\sum_{j=1}^{\infty} \frac{F^{j-1}}{j!}\right) F c_2 = c_2 + \left(\sum_{j=1}^{\infty} \frac{F^{j-1}}{j!}\right) F w = c_2 + \left(-I + \sum_{j=0}^{\infty} \frac{F^j}{j!}\right) w.$$

From this we can conclude that, for F of small rank, the most significant cost of applying e^F comes from generating Q_k. Since this needs to be done only once, thereafter applying the exponential splitting method (3.4) is very fast with the formula (3.5).

4 Approximating with Toeplitz plus Hankel matrices

The sum of the ranges of two averaging operators is a subspace of $\mathbb{C}^{n\times n}$. In view of the preceding considerations, we need to recover the parts of their elements. For the case of Toeplitz plus Hankel matrices, in what follows an iterative method is devised to approximate $A \in \mathbb{C}^{n\times n}$ with a Toeplitz plus Hankel matrix. The method can be used to recover whether a given large matrix is actually a Toeplitz plus Hankel matrix, or nearly so, unless it is known in advance. For another application, in preconditioning

[4] This is a simplifying assumption which can be circumvented.

dense linear systems such approximations are of use because linear systems involving Toeplitz plus Hankel matrices can be solved fast.

Using the Frobenius norm and the respective inner product, an obvious approach to solve the problem consists of a generation of an orthonormal basis of the subspace of Toeplitz plus Hankel matrices. Then what remains is to compute the Fourier coefficients of A to have the nearest element. For large n this approach is not attractive even though there is an orthonormal basis for the subspace consisting of Toeplitz matrices, as well as for Hankel matrices, readily available; see Appendix. In spite of this, the computation of an orthonormal basis for the sum of the subspaces appears to consume a lot of storage and is relatively expensive.

As an alternative, we look at an iterative method that allows us to stop the computations prematurely if sufficient accuracy is achieved. Hence the overall cost of computations can be controlled without losing the approximation computed so far. Also, the method consumes a very small amount of storage.

To this end we employ the following formula of Lenard [14] in finite dimensions. Suppose P and Q are two orthogonal projections on a finite dimensional Hilbert space such that the intersection of the ranges of P and Q is $\{0\}$. Then

$$R = (I - Q)(I - PQ)^{-1}P + (I - P)(I - QP)^{-1}Q \tag{4.1}$$

gives the orthogonal projector onto the sum of the ranges of P and Q. This formula is of particular interest in case there is an orthonormal basis of the range of P as well as of the range of Q separately readily available, but forming an orthonormal basis of the sum of the ranges is not quite feasible, for instance, due to storage restrictions.

Having

$$\|PQ\|_2 < 1 \tag{4.2}$$

is a necessary and sufficient condition on the existence of R. For our purposes it is crucial to look at the product PQ more carefully in terms of its singular values.[5] Clearly, the nonzero singular values of PQ equal the nonzero singular values of P restricted to the range of Q.

With these preparations, consider applying the formula (4.1) with P being the orthogonal projection on $\mathbb{C}^{n\times n}$ onto the set of Toeplitz and Q onto the set of Hankel matrices. Let us do this with respect to the standard inner product

$$(A, B) = \text{trace}\,(B^*A), \quad \text{for } A, B \in \mathbb{C}^{n\times n}, \tag{4.3}$$

on $\mathbb{C}^{n\times n}$. As an obstacle to using the formula (4.1), it is readily seen that the intersection of the ranges of P and Q is spanned by E_1 and E_2, where E_1 (resp. E_2) has ones on the even (resp. odd) diagonals and zeros on the odd (resp. even) diagonals. Since E_1 and E_2 are orthogonal, we can conclude that exactly the two largest singular values

[5] The singular values of PQ give the cosines of the canonical angles between the ranges of P and Q, a concept of importance, for instance, in perturbation theory for pairs of subspaces; see [3, p. 201].

of PQ equal one. Therefore, the formula of Lenard cannot be applied to our problem as such because the condition (4.2) is not satisfied and the required inverses do not exist. To fix this, we proceed by making a small rank perturbation to P (alternatively, to Q) without changing the sum of the ranges, to "speed-up" the convergence.

For this purpose, assume for simplicity that n is even so that the Frobenius norms of E_1 and E_2 equal being $\frac{n}{\sqrt{2}}$. Then replace P in (4.1) with the orthogonal projector

$$\hat{P} = P - \frac{2}{n^2}(\cdot, E_1)E_1 - \frac{2}{n^2}(\cdot, E_2)E_2,$$

i.e., then the subspace spanned by E_1 and E_2 is deflated from the subspace spanned by Toeplitz matrices. Hence, for $A \in \mathbb{C}^{n\times n}$ an application of $\hat{P}$ can be written as

$$\hat{P}A = PA - \frac{2}{n^2}\left(\sum_{j,k=1}^{n/2}(a_{2j-1,2k-1} + a_{2j,2k})E_1 + \sum_{j,k=1}^{n/2}(a_{2j,2k-1} + a_{2j-1,2k})E_2\right)$$

with

$$PA = \text{Toeplitz}(t_{-n+1}, t_{-n+2}, \ldots, t_{n-2}, t_{n-1}), \tag{4.4}$$

where t_k is the average value of the entries of A on its kth diagonal. With this the computation of $\hat{P}A$ consumes $O(n^2)$ floating point operations. The action of Q remains unaltered, being defined analogously to (4.4) with respect to the anti-diagonals of A. After these changes, the sum of the ranges of $\hat{P}$ and Q still equals the sum of the ranges of P and Q, as required.

With these changes the formula of Lenard can be used. The most time-consuming part in its application to a matrix $A \in \mathbb{C}^{n\times n}$ with $\hat{P}$ and Q consists of solving the matrix equations

$$(I - \hat{P}Q)X = \hat{P}A \tag{4.5}$$

and

$$(I - Q\hat{P})Y = QA. \tag{4.6}$$

It is now these equations that we intend to solve iteratively since we only need to perform $O(n^2)$ computations once to find the right-hand sides $\hat{P}A$ and QA. Thereafter the actual iterations can be organized to cost only $O(n)$ floating point operations per step since we only need to apply $\hat{P}$ (resp. Q) to Hankel (resp. Toeplitz) matrices repeatedly. This reduction is due to the following theorem.

Theorem 4.1 *Let P be the orthogonal projection on $\mathbb{C}^{n\times n}$ onto the set of Toeplitz matrices. If $H \in \mathbb{C}^{n\times n}$ is a Hankel matrix, then computing PH and $\hat{P}H$ costs $3n-3$ and $5n-1$ floating point operations, respectively.*

Proof Denote the anti-diagonal entries of H by h_j, for $-n+1, \ldots, n-1$, negative indeces appearing below the main anti-diagonal. Suppose n is even (odd is treated similarly). To have PH, compute

$$\begin{array}{rcccl}
\hat{t}_{-n+1} = & & \hat{t}_{n-1} & = & h_0, \\
\hat{t}_{-n+2} = & & \hat{t}_{n-2} & = & h_{-1} + h_1, \\
\hat{t}_{-n+3} = & & \hat{t}_{n-3} & = & \hat{t}_{-n+1} + h_{-2} + h_2, \\
\hat{t}_{-n+4} = & & \hat{t}_{n-4} & = & \hat{t}_{-n+2} + h_{-3} + h_3, \\
\quad \vdots & & & \vdots & \\
\hat{t}_{-1} = & & \hat{t}_1 & = & \hat{t}_{-3} + h_{-n+2} + h_{n-2}, \\
\hat{t}_0 = & \hat{t}_{-2} + h_{-n+1} + h_{n-1} & & &
\end{array}$$

which takes $2(n-1)$ floating point operations. Then set $t_{-k} = t_k = \hat{t}_k/(n-k)$, for $k = 0, 1, \ldots, n-2$. In all this takes $3(n-3)$ floating point operations.

Observe that $\sum_{j,k=1}^{n/2}(a_{2j-1,2k-1} + a_{2j,2k}) = 2(\hat{t}_{-n+2} + \hat{t}_{-n+4} \cdots + \hat{t}_{-2}) + \hat{t}_0$ and $\sum_{j,k=1}^{n/2}(a_{2j,2k-1} + a_{2j-1,2k}) = 2(\hat{t}_{-n+3} + \hat{t}_{-n+3} \cdots + \hat{t}_{-1})$ in case A is a Hankel matrix. Thus, collecting the floating point operations, it takes $5n-1$ to have $\hat{P}H$. □

Analogously, it costs only $3n-3$ floating point operations to compute QT for a Toeplitz matrix T.

Because of this, combined with the fact that $\|\hat{P}Q\|_2 < 1$, we propose using simple iterations such as the Neumann series or the Tchebyshev iteration for solving (4.5) and (4.6). These methods, albeit simple, have the advantage that they consume a very little storage. Also, they do not require performing inner products. In view of Theorem 4.1, performing inner products can be regarded as too expensive consuming $O(n^2)$ floating point operations. The success of this approach obviously depends completely on the speed of convergence of these methods.

To estimate the speed of convergence, we need $\|\hat{P}Q\|_2$, i.e., the third singular value of PQ. (Observe that everything applies to $\|Q\hat{P}\|_2$ as well.) This is obviously equivalent to finding the 2-norm of $\hat{P}$, or the third singular value of P, restricted to the set of Hankel matrices. See Appendix for a matrix representation to this end. With this, based on numerical experiments, we conjecture that its limit value is 0.5. This is clearly an impressive speed. More importantly, for very moderate values the limit is nearly attained; see Table 1 for the behaviour of the third singular value computed numerically and rounded to four digits, versus the dimension n. Observe that with $n = 100$ the third singular value is already 0.5002.

Table 1 The dimension n and the behaviour of the third singular value of PQ, computed with MATLAB

$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 100$	$n = 200$	$n = 300$
0.5156	0.5038	0.5017	0.5009	0.5006	0.5002	0.5000	0.5000

5 Conclusions

Exponential splitting methods rely on very exceptional splittings of the matrix. Averaging operators involving two carefully chosen, different unitary unipotents provide an algebraic way to look at such splittings. Toeplitz plus Hankel matrices, analogously to circulant plus diagonal matrices, serve as an example of this approach. The approach leads to a hierarchy of matrices with similar properties.

In connection with Toeplitz plus Hankel matrices it is natural to consider approximating with them. An iterative method based on a formula of Lenard leads to very fast convergence, with moderate storage requirements. Based on numerical experiments, it is conjectured that the asymptotic convergence factor (of mere simple iterations) approaches 0.5 as the dimension n grows to infinity.

Acknowledgment Supported by the Academy of Finland.

Appendix

Using the notation of Sect. 4, in this appendix a matrix representation for P restricted to Hankel matrices with its range restricted to Toeplitz matrices is given. For an orthonormal basis among Toeplitz matrices, we use the MATLAB [15] notation and set

$$\begin{aligned} q_1 &= \operatorname{diag}([1], n-1) \\ q_2 &= \operatorname{diag}\left(\frac{1}{\sqrt{2}}[1\ 1], n-2\right) \\ q_3 &= \operatorname{diag}\left(\frac{1}{\sqrt{3}}[1\ 1\ 1], n-3\right) \\ &\vdots \\ q_{2n-1} &= \operatorname{diag}([1], -n+1), \end{aligned}$$

i.e., we start from the $(n-1)$st diagonal and proceed downwards. With B of appropriate size denoting the backward identity, among Hankel matrices we use the orthonormal basis

$$\begin{aligned} e_1 &= B\operatorname{diag}\left(\frac{1}{\sqrt{n}}[1\ 1\ \cdots\ 1]\right) \\ e_2 &= B\operatorname{diag}\left(\frac{1}{\sqrt{n-1}}[1\ 1\ \cdots\ 1], 1\right) \\ e_3 &= B\operatorname{diag}\left(\frac{1}{\sqrt{n-2}}[1\ 1\ \cdots\ 1], 2\right) \\ &\vdots \\ e_n &= B\operatorname{diag}([1], n-1) \end{aligned}$$

$$
\begin{aligned}
e_{n+1} &= B\,\mathrm{diag}\left(\frac{1}{\sqrt{n-1}}[1\ 1\ \cdots\ 1], -1\right)\\
e_{n+2} &= B\,\mathrm{diag}\left(\frac{1}{\sqrt{n-2}}[1\ 1\ \cdots\ 1], -2\right)\\
&\vdots\\
e_{2n-2} &= B\,\mathrm{diag}([1], -n+1),
\end{aligned}
$$

i.e., we start from the main anti-diagonal and move from there first upwards. Once done, then downwards, excluding the main anti-diagonal.

Suppose n is even. Then the matrix representation is $\begin{bmatrix} v & V & V \\ 0 & v^T B & v^T B \\ Bv & BV & BV \end{bmatrix}$ with v of size $(n-1)$-by-1 and V of size $(n-1)$-by-$(n-1)$. For the entries,

$$
v = \left[\frac{1}{\sqrt{n}}\ 0\ \frac{1}{\sqrt{3n}}\ 0\ \frac{1}{\sqrt{5n}}\ \cdots\ 0\ \frac{1}{\sqrt{(n-1)n}}\right]^T.
$$

For $j = 1, 2, \ldots, n-1$ the columns of V are

$$
\left[\mathbf{0}\ \frac{1}{\sqrt{(j+1)(n-j)}}\ 0\ \frac{1}{\sqrt{(j+3)(n-j)}}\ 0\ \frac{1}{\sqrt{(j+5)(n-j)}}\ 0\cdots\right]^T,
$$

with the last entry of the column being 0 for j odd and $\mathbf{0}$ denoting the zero vector of size 1-by-j.

References

1. Arnold, V.I.: Fermat dynamics, matrix arithmetics, finite circles, and the finite Lobachevsky planes. Funct. Anal. Appl. **38**(1), 1–13 (2004)
2. Barnett, S., Gover, M.J.: Some extensions of Hankel and Toeplitz matrices. Linear Multilinear Algebra **14**(1), 45–65 (1983)
3. Bhatia, R.: Matrix Analysis Graduate Texts in Mathematics, 169. Springer, New York (1997)
4. Bhatia, R.: Pinching, trimming, truncating, and averaging of matrices. Am. Math. Mon. **107**(7), 602–608 (2000)
5. Bhatia, R., Kahan, W., Li, R.-C.: Pinchings and norms of scaled triangular matrices. Linear Multilinear Algebra **50**(1), 15–21 (2002)
6. Chan, R.N., Ng, M.K.: Conjugate gradient methods for Toeplitz systems. SIAM Rev. **38**, 427–482 (1996)
7. Davidson, K.: C^*-Algebras by Example. Fields Institute Monograph Series, vol. 6 (1996)
8. Holtz, O., Mehrmann, V., Schneider, H.: Potter, Wielandt, and Drazin on the matrix equation $AB = \omega BA$: new answers to old questions. Am. Math. Mon. **111**(8), 655–667 (2004)
9. Huhtanen, M., Nevanlinna, O.: Minimal decompositions and iterative methods. Numer. Math. **86**, 257–281 (2000)
10. Jahnke, T., Lubich, C.: Error bounds for exponential operator splittings. BIT **40**, 735–744 (2000)
11. Kailath, T., Olshevsky, V.: Displacement structure approach to discrete-trigonometric-transform based preconditioners of G. Strang type and of T. Chan type. SIAM J. Matrix Anal. Appl. **26**, 706–734 (2005)
12. Kozlov, R., Kvaerno, A., Owren, B.: The behaviour of the local error in splitting methods applied to stiff problems. J. Comput. Phys. **195**, 576–593 (2004)

13. Lawrence, J., Pfander, G.E., Walnut, D.: Linear independence of Gabor systems in finite dimensional vector spaces. J. Fourier Anal. Appl. **11**(6), 715–726 (2005)
14. Lenard, A.: The numerical range of projections. J. Funct. Anal. **10**, 410–413 (1972)
15. Mathworks, Matlab, www.mathworks.com/products/matlab
16. Ng, M.K., Michael, Bai, Z.-Z.: A hybrid preconditioner of banded matrix approximation and alternating direction implicit iteration for symmetric sinc-Galerkin linear systems. Linear Algebra Appl. **366**, 317–335 (2003)
17. Strang, G.: On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5**, 506–517 (1968)
18. Zanna, A., Munthe-Kaas, H.: Generalized polar decompositions for the approximation of the matrix exponential. SIAM J. Matrix Anal. Appl. **23**, 840–862 (2001)