A unifying framework for preconditioners based on fast transforms

by

Kurt Otto

Report No. 187 (revised)/1999

UPPSALA UNIVERSITET
Institutionen för
teknisk databehandling

UPPSALA UNIVERSITY
Department of
Scientific Computing

A UNIFYING FRAMEWORK FOR PRECONDITIONERS BASED ON FAST TRANSFORMS

KURT OTTO[†]

Abstract. The framework is aimed at preconditioners for discretized multidimensional PDE problems, where the underlying grid is structured. Instead of embedding the discretized operator into a matrix, a tensor approach is advocated. Therefore, optimal approximations of complex tensors of even order are considered. By construction subtensors in the approximants are diagonalizable by fast transforms. Theorems describing efficient ways of computing the approximants are derived. As an application, effective preconditioners for the discretized Helmholtz equation with radiation boundary conditions are studied.

Key words. preconditioners, fast transforms, tensor notation, complexity

AMS subject classifications. 65F10, 65T50, 65N22

1. Introduction. Tensors were first utilized to describe the elastic deformation of solids. Actually, the word tensor stems from the Latin word "tensus" meaning stretched. In the beginning of the 20th century, tensor calculus was refined by the Italian mathematicians Ricci and Levi-Cevita. Since then tensor calculus has been an invaluable tool in differential geometry, special and general relativity, and several branches of Physics. The classical way of using tensors is to let them define coordinate invariant linear operators.

In this paper, a tensor notation is advocated from a different point of view. Let $u(x_1, \dots, x_L)$ denote the true solution to a scalar partial differential equation in L dimensions. For a *structured* grid $(x_1^{j_1}, \dots, x_L^{j_L})$, $j_\ell = 1, \dots, m_\ell$, $\ell = 1, \dots, L$, it is convenient to let an approximate solution be denoted by $u^{j_1,\dots,j_L} \approx u(x_1^{j_1}, \dots, x_L^{j_L})$. The approximant u can be conceived as an $\binom{L}{0}$ tensor in $\mathbb{C}^{m_1\times\cdots\times m_L}$ with contravariant components $u^{j_1,\dots,j_L}$. A finite difference approximation of the differential equation will result in a linear system of equations, which can be stated in tensor form

$$Bu = g,$$

or in component form

$$B^{j_1,\dots,j_L}_{k_1,\dots,k_L} u^{k_1,\dots,k_L} = g^{j_1,\dots,j_L}, \quad j_\ell, k_\ell = 1, \dots, m_\ell, \quad \ell = 1, \dots, L.$$

The coefficient tensor $B \in \mathbb{C}^{m_1\times\cdots\times m_L}_{m_1\times\cdots\times m_L}$ will typically be banded. If the solution were a vector field $\boldsymbol{u}$ with m_0 components, the approximant would become a tensor with components $u^{j_0;j_1,\dots,j_L}$ and the coefficient tensor would belong to $\mathbb{C}^{m_0\times m_1\times\cdots\times m_L}_{m_0\times m_1\times\cdots\times m_L}$. The main objective of this paper is to develop a transform based framework for preconditioners to $B \in \mathbb{C}^{m_0\times m_1\times\cdots\times m_L}_{m_0\times m_1\times\cdots\times m_L}$, where B typically is banded in levels $\ell = 1, \dots, L$.

[†]Department of Scientific Computing, Uppsala University, SE-751 04 Uppsala, Sweden (kurt@tdb.uu.se).

1.1. Matrix notation. Let s_diag(A, r) denote the sum of the rth diagonal of a square matrix A, where $r = 0$ corresponds to the main diagonal, a negative value of r refers to the $|r|$th diagonal to the left of the main diagonal, and a positive value identifies the $|r|$th diagonal to the right of the main diagonal. Similarly, s_antidiag(A, r) yields the sum of the rth antidiagonal of A with $r = 0$ referring to the main antidiagonal, negative values of r corresponding to antidiagonals to the left of the main antidiagonal, and positive values of r denoting antidiagonals to the right.

1.2. Tensor notation. Let $A \in \mathbb{C}^{m_1\times\cdots\times m_K}_{m_1\times\cdots\times m_L}$ denote a tensor of type $\binom{K}{L}$ and order $K + L$, which has components $A^{i_1,\ldots,i_K}_{j_1,\ldots,j_L} \in \mathbb{C}$ where the upper and lower indices range as $i_\ell = 1, \ldots, m_\ell$, $\ell = 1, \ldots, K$, and $j_\ell = 1, \ldots, m_\ell$, $\ell = 1, \ldots, L$. Throughout the paper the following *summation convention* is adopted: whenever an expression involving tensor components contains an index that is *repeated* both as subscript and superscript, a summation over the index is understood. An underlined index means that the summation convention is suppressed for that index.

For tensors $A \in \mathbb{C}^{m_1\times\cdots\times m_K}_{m_1\times\cdots\times m_L}$ and $B \in \mathbb{C}^{m_1\times\cdots\times m_L}_{m_1\times\cdots\times m_N}$, the inner product AB is defined as the tensor $C \in \mathbb{C}^{m_1\times\cdots\times m_K}_{m_1\times\cdots\times m_N}$ with components

$$C^{i_1,\ldots,i_K}_{k_1,\ldots,k_N} = A^{i_1,\ldots,i_K}_{j_1,\ldots,j_L} B^{j_1,\ldots,j_L}_{k_1,\ldots,k_N}.$$

The components of the $\binom{L}{L}$ identity tensor I are given by

$$I^{i_1,\ldots,i_L}_{j_1,\ldots,j_L} = \prod_{\ell=1}^{L} \delta^{i_\ell}_{j_\ell}, \quad i_\ell, j_\ell = 1, \ldots, m_\ell,$$

where the Kronecker delta δ^i_j satisfies

$$\delta^i_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The adjoint $A^* \in \mathbb{C}^{m_1\times\cdots\times m_L}_{m_1\times\cdots\times m_K}$ of A has the following components

$$[A^*]^{j_1,\ldots,j_L}_{i_1,\ldots,i_K} = \bar{A}^{i_1,\ldots,i_K}_{j_1,\ldots,j_L}.$$

A tensor Q is unitary if $Q^*Q = QQ^* = I$. Finally, the Frobenius norm of A is given by

$$\|A\|_F^2 = [A^*]^{j_1,\ldots,j_L}_{i_1,\ldots,i_K} A^{i_1,\ldots,i_K}_{j_1,\ldots,j_L}.$$

The analogy with matrix algebra should be quite clear. Column vectors are $\binom{1}{0}$ tensors, row vectors are $\binom{0}{1}$ tensors, and matrices are $\binom{1}{1}$ tensors. The inner product of two $\binom{1}{1}$ tensors is analogous to the matrix product. Using the concepts defined in this section, it is quite easy to generalize several results from matrix algebra to tensor algebra. Here are just a few examples:

- A scalar product $\langle\cdot,\cdot\rangle$ is defined by $\langle x, y\rangle \equiv x^*y$, $\forall x, y \in \mathbb{C}^{m_1\times\cdots\times m_K}$.
- The adjoint A^* satisfies $\langle x, Az\rangle = \langle A^*x, z\rangle$, $\forall x \in \mathbb{C}^{m_1\times\cdots\times m_K}$ and $\forall z \in \mathbb{C}^{m_1\times\cdots\times m_L}$.
- $(AB)^* = B^*A^*$
- The Frobenius norm of a tensor is invariant under inner multiplication by unitary tensors.

2. The normal block framework. The basic idea for a normal block preconditioner with several levels is to apply unitary transforms in all levels but two. For the remaining levels, linear systems are solved, and the preconditioning is completed by inverse transforms. In the polytransform case, different transforms are allowed for different levels. The case where the transforms are the same for the zeroth level is called the monotransform case. The *polynomial* normal block classes generated by [7, eq. (1.2a)] with $q = m = m_1$ and [7, eq. (6.2)] with $q = m_1$ coincide with the monotransform case of the normal block class for $m_0 = 1$, $L = 1$, and $m_0 = n_c$, $L = 2$, respectively. For $q < m_1$, the polynomial normal classes above are subsets of the normal class.

If the operand for a polytransform preconditioner were embedded in a vector, the transform part of the preconditioner could be expressed as a matrix with a hierarchy of L block levels. However, that matrix becomes muddled with permutation matrices, which are needed in order to get the various transforms applied correctly. The structure of the preconditioning operator becomes much clearer when the operand is represented as a tensor of order $L+1$. The standard matrix notation, which has been long prevailing in the numerical linear algebra community, makes designing and coding unnecessarily complicated in this context.

To determine a preconditioner M for a tensor B can be viewed as a minimization problem. We will seek optimal [5] approximations, i.e., approximants that minimize $\|B - M\|_F$.

2.1. Approximants for tensors of order four. Consider the $\binom{2}{2}$ tensor B with components

$$B^{j_0;j_1}_{k_0;k_1}, \quad j_0, k_0 = 1, \ldots, m_0, \quad j_1, k_1 = 1, \ldots, m_1. \tag{2.1}$$

Form the unitary tensor $Q \in \mathbb{C}^{m_0\times m_1}_{m_0\times m_1}$, the components of which satisfy

$$Q^{j_0;j_1}_{k_0;k_1} = 0, \quad \text{if } j_0 \neq k_0. \tag{2.2}$$

Compose a preconditioner M as

$$M = Q\Lambda Q^*, \tag{2.3a}$$

$$\Lambda^{j_0;j_1}_{k_0;k_1} = 0, \quad \text{if } j_1 \neq k_1. \tag{2.3b}$$

THEOREM 2.1. *For B and M defined in* (2.1) *and* (2.3), *the solution to the minimization problem*

$$\min_{\Lambda} \|B - M\|_F$$

is given by

$$\Lambda_{k_0;\underline{j_1}}^{j_0;\underline{j_1}} = [Q^*]_{\underline{j_0};\alpha_1}^{\underline{j_0};\underline{j_1}} B_{k_0;\beta_1}^{j_0;\alpha_1} Q_{\underline{k_0};\underline{j_1}}^{\underline{k_0};\beta_1}.$$

Proof. Exploiting the fact that the Frobenius norm is invariant under inner multiplications by unitary tensors yield

$$\mathcal{F} \equiv \|B - M\|_F = \|Q^*(B - Q\Lambda Q^*)Q\|_F = \|Q^*BQ - \Lambda\|_F.$$

Since the components $\Lambda_{k_0;k_1}^{j_0;j_1}$ are zero for $j_1 \neq k_1$ and due to (2.2), the minimizer of $\mathcal{F}$ is given by

$$\Lambda_{k_0;\underline{j_1}}^{j_0;\underline{j_1}} = \left[Q^*BQ\right]_{k_0;\underline{j_1}}^{j_0;\underline{j_1}} = [Q^*]_{\underline{j_0};\alpha_1}^{\underline{j_0};\underline{j_1}} B_{k_0;\beta_1}^{j_0;\alpha_1} Q_{\underline{k_0};\underline{j_1}}^{\underline{k_0};\beta_1}.$$

□

2.2. Approximants for tensors of higher, even order. Consider $\binom{L+1}{L+1}$ tensors with components

$$B_{k_0;k_1,\ldots,k_L}^{j_0;j_1,\ldots,j_L}, \quad j_\ell, k_\ell = 1, \ldots, m_\ell, \quad \ell = 0, \ldots, L. \tag{2.4}$$

Form the tensor $Q \in \mathbb{C}_{m_0 \times m_1 \times \cdots \times m_L}^{m_0 \times m_1 \times \cdots \times m_L}$ with components

$$Q_{k_0;k_1,\ldots,k_L}^{j_0;j_1,\ldots,j_L} = \delta_{k_1}^{j_1} \prod_{\ell=2}^{L} {}_\ell Q_{k_0;k_\ell}^{j_0;j_\ell}, \tag{2.5}$$

where the tensors ${}_\ell Q$, $\ell = 2, \ldots, L$, satisfy (2.2) and are unitary by construction. Consequently, the tensor Q is also unitary. Compose a preconditioner M as

$$M = Q\Lambda Q^*, \tag{2.6a}$$

$$\Lambda_{k_0;k_1,\ldots,k_L}^{j_0;j_1,\ldots,j_L} = 0, \quad \text{if } j_\ell \neq k_\ell, \quad \ell = 2, \ldots, L. \tag{2.6b}$$

THEOREM 2.2. *For B and M defined in* (2.4) *and* (2.6), *the solution to the minimization problem*

$$\min_{\Lambda} \|B - M\|_F$$

is given by

$$\Lambda_{k_0;k_1,\underline{j_2},\ldots,\underline{j_L}}^{j_0;j_1,\underline{j_2},\ldots,\underline{j_L}} = \left(\prod_{\ell=2}^{L} [{}_\ell Q^*]_{\underline{j_0};\alpha_\ell}^{\underline{j_0};\underline{j_\ell}}\right) B_{k_0;k_1,\beta_2,\ldots,\beta_L}^{j_0;j_1,\alpha_2,\ldots,\alpha_L} \prod_{\ell=2}^{L} {}_\ell Q_{\underline{k_0};\underline{j_\ell}}^{\underline{k_0};\beta_\ell}.$$

Proof. By the nonzero pattern (2.6b) for Λ and by the structure (2.5) of Q, the minimizer is given by

$$\begin{aligned}\Lambda_{k_0;k_1,\underline{j_2},\ldots,\underline{j_L}}^{j_0;j_1,\underline{j_2},\ldots,\underline{j_L}} &= \left[Q^*BQ\right]_{k_0;k_1,\underline{j_2},\ldots,\underline{j_L}}^{j_0;j_1,\underline{j_2},\ldots,\underline{j_L}} \\ &= \left(\prod_{\ell=2}^{L} [{}_\ell Q^*]_{\underline{j_0};\alpha_\ell}^{\underline{j_0};\underline{j_\ell}}\right) B_{k_0;k_1,\beta_2,\ldots,\beta_L}^{j_0;j_1,\alpha_2,\ldots,\alpha_L} \prod_{\ell=2}^{L} {}_\ell Q_{\underline{k_0};\underline{j_\ell}}^{\underline{k_0};\beta_\ell}.\end{aligned}$$

□

2.3. Normal block preconditioners. Since Q is unitary, it follows from (2.6) that $M^{-1} = (Q\Lambda Q^*)^{-1} = Q\Lambda^{-1}Q^{-1}$. Therefore, the normal block solve $u = M^{-1}g$ can be performed as

1. $v = Q^{-1}g$
2. solve $\Lambda z = v$
3. $u = Qz$

An algorithmic formulation is stated below, where $n_\ell = \prod_{i=\ell}^{L} m_i$ is introduced.

Phases 1 and 3: For $\ell = 2, \ldots, L$ and $j_0 = 1, \ldots, m_0$, compute n_1/m_ℓ *independent* transforms of length m_ℓ.

Phase 2: Solve n_2 *independent* systems with coefficient tensors $\Lambda^{:;:,\underline{j_2,\ldots,j_L}}_{:;:,\underline{j_2,\ldots,j_L}} \in \mathbb{C}^{m_0\times m_1}_{m_0\times m_1}$.

If fast algorithms are available for the transforms, the complexity becomes $\mathcal{O}(n_0 \log_2 n_2)$ for phases 1 and 3. The complexity is $n_2\mathrm{a}(m_1, m_0)$ for phase 2.

3. Arithmetic complexity. The number of arithmetic operations required to form a normal block preconditioner, i.e., to compute the minimizer Λ for a tensor B depends on its sparsity. Let $m_{(1)}m_1$ denote the number of nonzero subtensors at level $\ell = 1$, and let $n_{(\ell)}n_\ell$ be the number of nonzero components in $B^{j_0;\ldots,j_{\ell-1},:,\ldots,:}_{k_0;\ldots,k_{\ell-1},:,\ldots,:} \in \mathbb{C}^{m_\ell\times\cdots\times m_L}_{m_\ell\times\cdots\times m_L}$.

3.1. The polytransform case. If B is dense, the most economical way to compute Λ in Theorems 2.1 and 2.2 is to exploit fast transforms (if available). For sparse B, it is preferable to mask off the zero components, and use a direct calculation with the formulas for Λ. The arithmetic complexities for the polytransform case are summarized in Table 3.1.

TABLE 3.1
Arithmetic complexity to form Λ in the polytransform case.

tensor class	$L \geq 2$	$L = 1$
dense	$\mathcal{O}(m_0^2 m_{(1)} n_1 \sum_{\ell=2}^{L} n_\ell \log_2 m_\ell)$	$\mathcal{O}(m_0^2 m_1^2 \log_2 m_1)$
sparse	$\mathcal{O}(m_0^2 m_{(1)} n_1 n_{(2)} n_2)$	$\mathcal{O}(m_0^2 m_1 n_{(1)} n_1)$

3.2. The monotransform case. For certain trigonometric transforms, it is more convenient to find monotransform approximants by using the bases described in [2] and [9], where approximation of Toeplitz matrices is considered. Such a technique has been applied to real matrices [4]. The main purpose of this section is to extend the results in [4, 3] to all the transforms in Appendix A and to complex tensors. The technique to find the minimizer in [4] is based on solving normal equations, which have been derived by differentiating the functional w.r.t. the degrees of freedom for the minimizer. That would not be valid for complex tensors. The normal equations could instead be obtained by projection with the scalar product associated with the Frobenius norm.

We first treat the case $L = 1$ and $m_0 = 1$. Following Boman and Koltracht [2], we define an eigenspace $\mathcal{D}_Q$.

Definition. *Let $Q \in \mathbb{C}^m_m$ be a nonsingular tensor. For a given Q, denote by $\mathcal{D}_Q$ the m-dimensional vector space of all tensors M such that $Q^{-1}MQ$ is diagonal.*

The eigenvalues of a tensor $M \in \mathcal{D}_Q$ can be computed from the jth row z of M *provided* that the jth row of Q is nonzero everywhere [2]. The row vector λ, holding the eigenvalues of M, satisfies

$$\begin{aligned} &\lambda = zQD, \\ &D^i_k = \begin{cases} 1/Q^j_k & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases}, \quad i,k = 1,\dots,m. \end{aligned} \tag{3.1}$$

For all the tensors Q listed in Appendix A, the first row is nonzero, which makes $j = 1$ a viable option in (3.1). The procedure to find the first row is to choose an appropriate basis for $\mathcal{D}_Q$, determine the solution to the normal equations, and extract the row. For the convenience of the reader, we restate results from [2, 4] as lemmas that characterize the bases for $\mathcal{D}_Q$ for some sine and cosine transforms.

Lemma 3.1 (Boman and Koltracht [2]). *Let Q be associated with* ST, *and introduce the following tensors in $\mathbb{R}^m_m$*

$$\begin{aligned} &{}^0R = I, \\ &{}^rR^j_k = \begin{cases} 1 & \textit{if } k - j = -r, \\ 1 & \textit{if } k - j = r, \\ -1 & \textit{if } k + j = r, \\ -1 & \textit{if } k + j = 2m + 2 - r, \\ 0 & \textit{otherwise}, \quad r = 1,\dots,m-1. \end{cases} \end{aligned}$$

Then $\{{}^rR\}^{m-1}_{r=0}$ is a basis for $\mathcal{D}_Q$. Moreover, when $r > 0$ the spectrum of ${}^rR$ is

$$\lambda^r_k = 2\cos\Big(\frac{rk\pi}{m+1}\Big), \quad k = 1,\dots,m.$$

Lemma 3.2 (Boman and Koltracht [2]). *Let Q be associated with* CT-I, *and introduce the following tensors in $\mathbb{R}^m_m$*

$$\begin{aligned} &{}^0R = I, \\ &{}^rR^j_k = \begin{cases} 1 & \textit{if } k - j = -r, \quad j < m \wedge k > 1, \\ \sqrt{2} & \textit{if } k - j = -r, \quad j = m \vee k = 1, \\ 1 & \textit{if } k - j = r, \quad j > 1 \wedge k < m, \\ \sqrt{2} & \textit{if } k - j = r, \quad j = 1 \vee k = m, \\ 1 & \textit{if } k + j = r + 2, \quad j > 1 \wedge k > 1, \\ 1 & \textit{if } k + j = 2m - r, \quad j < m \wedge k < m, \\ 0 & \textit{otherwise}, \quad r = 1,\dots,m-2, \end{cases} \\ &{}^{m-1}R^j_k = \begin{cases} 1 & \textit{if } k + j = m + 1, \\ 0 & \textit{otherwise}. \end{cases} \end{aligned}$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_{DQ}$*, where* $D = \operatorname{diag}(\frac{1}{\sqrt{2}}, 1, \ldots, 1, \frac{1}{\sqrt{2}})$*. Moreover, when* $0 < r < m-1$ *the spectrum of* ${}^rR$ *is*

$$\lambda_k^r = 2\cos\Big(\frac{r(k-1)\pi}{m-1}\Big), \quad k = 1, \ldots, m.$$

LEMMA 3.3 (Kailath and Olshevsky [9]). *Let* Q *be associated with* ST-II*, and introduce the following tensors in* $\mathbb{R}_m^m$

$$\begin{aligned} {}^0R &= I, \\ {}^rR_k^j &= \begin{cases} 1 & \text{if } k-j=-r, \\ 1 & \text{if } k-j=r, \\ -1 & \text{if } k+j=r+1, \\ -1 & \text{if } k+j=2m+1-r, \\ 0 & \text{otherwise}, \quad r=1,\ldots,m-1. \end{cases} \end{aligned}$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_Q$*. Moreover, when* $r > 0$ *the spectrum of* ${}^rR$ *is*

$$\lambda_k^r = 2\cos\Big(\frac{rk\pi}{m}\Big), \quad k = 1, \ldots, m.$$

LEMMA 3.4 (Boman and Koltracht [2]). *Let* Q *be associated with* CT-II*, and introduce the following tensors in* $\mathbb{R}_m^m$

$$\begin{aligned} {}^0R &= I, \\ {}^rR_k^j &= \begin{cases} 1 & \text{if } k-j=-r, \\ 1 & \text{if } k-j=r, \\ 1 & \text{if } k+j=r+1, \\ 1 & \text{if } k+j=2m+1-r, \\ 0 & \text{otherwise}, \quad r=1,\ldots,m-1. \end{cases} \end{aligned}$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_Q$*. Moreover, when* $r > 0$ *the spectrum of* ${}^rR$ *is*

$$\lambda_k^r = 2\cos\Big(\frac{r(k-1)\pi}{m}\Big), \quad k = 1, \ldots, m.$$

LEMMA 3.5 (Boman and Koltracht [2]). *Let* Q *be associated with* ST-III*,* *and introduce the following tensors in* $\mathbb{R}^m_m$

$$
\begin{aligned}
{}^0R &= I,\\
{}^rR^j_k &= \begin{cases} 1 & \text{if } k-j=-r, \quad j<m,\\ \sqrt{2} & \text{if } k-j=-r, \quad j=m,\\ 1 & \text{if } k-j=r, \quad k<m,\\ \sqrt{2} & \text{if } k-j=r, \quad k=m,\\ -1 & \text{if } k+j=r,\\ 1 & \text{if } k+j=2m-r, \quad j<m \wedge k<m,\\ 0 & \text{otherwise}, \quad r=1,\dots,m-1. \end{cases}
\end{aligned}
$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_{DQ}$, *where* $D = \operatorname{diag}(1,\dots,1,\frac{1}{\sqrt{2}})$. *Moreover, when* $r > 0$ *the spectrum of* ${}^rR$ *is*

$$\lambda^r_k = 2\cos\Big(\frac{r(k-\frac{1}{2})\pi}{m}\Big), \quad k = 1,\dots,m.$$

Lemma 3.6 (Kailath and Olshevsky [9]). *Let* Q *be associated with* CT-III, *and introduce the following tensors in* $\mathbb{R}^m_m$

$$
\begin{aligned}
{}^0R &= I,\\
{}^rR^j_k &= \begin{cases} 1 & \text{if } k-j=-r, \quad k>1,\\ \sqrt{2} & \text{if } k-j=-r, \quad k=1,\\ 1 & \text{if } k-j=r, \quad j>1,\\ \sqrt{2} & \text{if } k-j=r, \quad j=1,\\ 1 & \text{if } k+j=r+2, \quad j>1 \wedge k>1,\\ -1 & \text{if } k+j=2m+2-r,\\ 0 & \text{otherwise}, \quad r=1,\dots,m-1. \end{cases}
\end{aligned}
$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_{DQ}$, *where* $D = \operatorname{diag}(\frac{1}{\sqrt{2}},1,\dots,1)$. *Moreover, when* $r > 0$ *the spectrum of* ${}^rR$ *is*

$$\lambda^r_k = 2\cos\Big(\frac{r(k-\frac{1}{2})\pi}{m}\Big), \quad k = 1,\dots,m.$$

Lemma 3.7 (Kailath and Olshevsky [9]). *Let* Q *be associated with* ST-IV, *and introduce the following tensors in* $\mathbb{R}^m_m$

$$
\begin{aligned}
{}^0R &= I,\\
{}^rR^j_k &= \begin{cases} 1 & \text{if } k-j=-r,\\ 1 & \text{if } k-j=r,\\ -1 & \text{if } k+j=r+1,\\ 1 & \text{if } k+j=2m+1-r,\\ 0 & \text{otherwise}, \quad r=1,\dots,m-1. \end{cases}
\end{aligned}
$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_Q$. *Moreover, when* $r > 0$ *the spectrum of* ${}^rR$ *is*

$$\lambda_k^r = 2\cos\Big(\frac{r(k-\frac{1}{2})\pi}{m}\Big), \quad k = 1, \dots, m.$$

LEMMA 3.8 (Kailath and Olshevsky [9]). *Let* Q *be associated with* CT-IV, *and introduce the following tensors in* $\mathbb{R}_m^m$

$$\begin{aligned} {}^0R &= I, \\ {}^rR_k^j &= \begin{cases} 1 & \text{if } k - j = -r, \\ 1 & \text{if } k - j = r, \\ 1 & \text{if } k + j = r + 1, \\ -1 & \text{if } k + j = 2m + 1 - r, \\ 0 & \text{otherwise}, \quad r = 1, \dots, m-1. \end{cases} \end{aligned}$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_Q$. *Moreover, when* $r > 0$ *the spectrum of* ${}^rR$ *is*

$$\lambda_k^r = 2\cos\Big(\frac{r(k-\frac{1}{2})\pi}{m}\Big), \quad k = 1, \dots, m.$$

We now present constructive theorems for various monotransform approximants. The proofs for the different transforms are very similar. We only show the proof for the ST-IV and CT-IV cases. We introduce the notation $\mathcal{T}(u)$ for a symmetric Toeplitz matrix with u as its first row, and $\mathcal{H}(u, w)$ for a Hankel matrix with u as the first row and w as the last row. We also need a shift operator and a reverse shift operator defined as

$$\begin{aligned} \mathrm{s}_\ell(u) &= (u_{\ell+1} \ \cdots \ u_m \ 0 \ \cdots \ 0), \\ \sigma_\ell(u) &= (0 \ \cdots \ 0 \ u_m \ \cdots \ u_{\ell+1}). \end{aligned}$$

Let P be a permutation matrix achieving an odd–even shuffle, i.e., the nonzero elements satisfy

$$\begin{aligned} P_{2j-1}^j &= 1, \quad j = 1, \dots, \lceil m/2 \rceil, \\ P_{2j-2\lceil m/2 \rceil}^j &= 1, \quad j = \lceil m/2 \rceil + 1, \dots, m. \end{aligned}$$

This is an extension of [4, Thm. 1] to the complex case.

THEOREM 3.1 (Chan, Ng, and Wong [4]). *Let* $B \in \mathbb{C}_m^m$ *and* M *be the minimizer of* $\|B - M\|_F$ *over all* $M \in \mathcal{D}_Q$, *where* Q *is associated with* ST. *Denote by* u_1 *the column vector with all* $\lceil m/2 \rceil$ *entries equal to one, by* u_2 *the column vector with all* $\lfloor m/2 \rfloor$ *entries equal to one, and by* e *the first canonical column vector of length* $\lceil m/2 \rceil$. *Then it holds that*

$$M = \mathcal{T}(\mu) - \mathcal{H}\big(\mathrm{s}_2(\mu), \sigma_2(\mu)\big)$$

with

$$\mu = \tfrac{1}{2m+2} f P^* \begin{pmatrix} I + ee^* + u_1 u_1^* & 0 \\ 0 & I + 2u_2 u_2^* \end{pmatrix} P$$

if m is even, and

$$\mu = \tfrac{1}{2m+2} f P^* \begin{pmatrix} I + ee^* + 2u_1 u_1^* & 0 \\ 0 & I + u_2 u_2^* \end{pmatrix} P$$

if m is odd. The entries of the row vector f are

$$f_{r+1} = \mathrm{trace}({}^r R^* B), \quad r = 0, \dots, m-1,$$

where ${}^r R$ is given by Lemma 3.1.

THEOREM 3.2. *Let $B \in \mathbb{C}_m^m$ and M be the minimizer of $\|D(B-M)D^{-1}\|_F$ over all $M \in \mathcal{D}_Q$, where Q is associated with* CT-I. *Denote by u_1 the column vector with all $\lceil m/2 \rceil$ entries equal to one, by u_2 the column vector with all $\lfloor m/2 \rfloor$ entries equal to one, by e_1 and e_ℓ the first and last canonical column vectors of length $\lceil m/2 \rceil$. Then it holds that*

$$M = \big(\mathcal{T}(\mu) + \mathcal{H}\big(\mu, \sigma_0(\mu)\big)\big) D^2$$

with

$$\mu = \tfrac{1}{2m-2} f P^* \begin{pmatrix} I + e_1 e_1^* - \frac{1}{m-1} u_1 u_1^* & 0 \\ 0 & I + e_\ell e_\ell^* - \frac{1}{m-1} u_2 u_2^* \end{pmatrix} P$$

if m is even, and

$$\mu = \tfrac{1}{2m-2} f P^* \begin{pmatrix} I + e_1 e_1^* + e_\ell e_\ell^* - \frac{1}{m-1} u_1 u_1^* & 0 \\ 0 & I - \frac{1}{m-1} u_2 u_2^* \end{pmatrix} P$$

if m is odd. The entries of the row vector f are

$$f_{r+1} = \mathrm{trace}({}^r R^* D B D^{-1}), \quad r = 0, \dots, m-1,$$

where ${}^r R$ and D are given by Lemma 3.2.

This is an extension of results in [3, and their references therein] to the ST-II case (and complex tensors).

THEOREM 3.3 (Chan and Wong [3]). *Let $B \in \mathbb{C}_m^m$ and M be the minimizer of $\|B - M\|_F$ over all $M \in \mathcal{D}_Q$, where Q is associated with* ST-II *or* CT-II. *Denote by u_1 the column vector with all $\lceil m/2 \rceil$ entries equal to one, by u_2 the column vector with all $\lfloor m/2 \rfloor$ entries equal to one, and by e the first canonical column vector of length $\lceil m/2 \rceil$. Letting the upper and lower sign cases correspond to* ST-II *and* CT-II, *respectively, it holds that*

$$M = \mathcal{T}(\mu) \mp \mathcal{H}\big(\mathrm{s}_1(\mu), \sigma_1(\mu)\big)$$

with

$$\mu = \tfrac{1}{2m} f P^* \begin{pmatrix} I + ee^* + 2u_1u_1^* & \pm 2u_1u_2^* \\ \pm 2u_2u_1^* & I + (2 - \frac{2}{m})u_2u_2^* \end{pmatrix} P$$

if m is even, and

$$\mu = \tfrac{1}{2m} f P^* \begin{pmatrix} I + ee^* + (2 - \frac{2}{m})u_1u_1^* & \pm 2u_1u_2^* \\ \pm 2u_2u_1^* & I + 2u_2u_2^* \end{pmatrix} P$$

if m is odd. The entries of the row vector f are

$$f_{r+1} = \operatorname{trace}({}^r\!R^* B), \quad r = 0, \ldots, m-1,$$

where ${}^r\!R$ is given by Lemma 3.3 or Lemma 3.4.

THEOREM 3.4. *Let $B \in \mathbb{C}_m^m$ and M be the minimizer of $\|D(B-M)D^{-1}\|_F$ over all $M \in \mathcal{D}_Q$, where Q is associated with* ST-III *or* CT-III. *Denote by e the first canonical column vector of length m. Then it holds that*

$$M = \begin{cases} \big(\mathcal{T}(\mu) + \mathcal{H}(-\mathrm{s}_2(\mu), \sigma_0(\mu))\big)D^2 & \textit{for } \text{ST-III} \\ \big(\mathcal{T}(\mu) + \mathcal{H}(\mu, -\sigma_2(\mu))\big)D^2 & \textit{for } \text{CT-III} \end{cases}$$

with

$$\mu = \tfrac{1}{2m} f(I + ee^*).$$

The entries of the row vector f are

$$f_{r+1} = \operatorname{trace}({}^r\!R^* D B D^{-1}), \quad r = 0, \ldots, m-1,$$

where ${}^r\!R$ and D are given by Lemma 3.5 or Lemma 3.6.

THEOREM 3.5. *Let $B \in \mathbb{C}_m^m$ and M be the minimizer of $\|B - M\|_F$ over all $M \in \mathcal{D}_Q$, where Q is associated with* ST-IV *or* CT-IV. *Denote by e the first canonical column vector of length m. Letting the upper and lower sign cases correspond to* ST-IV *and* CT-IV, *respectively, it holds that*

$$M = \mathcal{T}(\mu) \pm \mathcal{H}\big(-\mathrm{s}_1(\mu), \sigma_1(\mu)\big)$$

with

$$\mu = \tfrac{1}{2m} f(I + ee^*).$$

The entries of the row vector f are

$$f_{r+1} = \operatorname{trace}({}^r\!R^* B), \quad r = 0, \ldots, m-1,$$

where ${}^r\!R$ is given by Lemma 3.7 or Lemma 3.8.

Proof. Taking ${}^r\!R$ from Lemma 3.7 or Lemma 3.8 yields

$$\mathcal{F} \equiv \|B - M\|_F^2 = \|B - \sum_{r=0}^{m-1} {}^r\!R \mu_{r+1}\|_F^2.$$

Note that $\langle A, B\rangle \equiv \operatorname{trace}(A^*B)\ \forall A, B \in \mathbb{C}_n^m$ is the scalar product that is associated with the Frobenius norm, in the sense that $\langle A, A\rangle = \|A\|_F^2$. To minimize $\mathcal{F}$ is a least squares problem with basis tensors ${}^rR$ labeled $r = 0, \ldots, m-1$. The solution is determined by the normal equations $\mu A = f$, where

$$\begin{aligned} f_{s+1} &= \langle {}^sR, B\rangle = \operatorname{trace}({}^sR^* B), \quad s = 0, \ldots, m-1, \\ A_{s+1}^{r+1} &= \langle {}^sR, {}^rR\rangle = \operatorname{trace}({}^sR^*\, {}^rR), \quad r, s = 0, \ldots, m-1. \end{aligned}$$

From the band structure of the basis tensors, it follows that the matrix A is given by

$$A = 2mI - mee^*.$$

The Sherman–Morrison–Woodbury formula yields

$$A^{-1} = \tfrac{1}{2m}(I + ee^*),$$

leading to

$$\mu = fA^{-1} = \tfrac{1}{2m} f(I + ee^*).$$

The formula for M follows directly from the definitions of ${}^rR$ and

$$M = \sum_{r=1}^{m} {}^{r-1}R\mu_r.$$

$\square$

From the theorems and lemmas in this section, convenient formulas for computing the first row of the minimizers can be identified directly. The following corollary is due to Chan, Ng, and Wong [4].

COROLLARY 3.1. *The first row z of the minimizer M in Theorem 3.1 is given by*

$$\begin{aligned} z_1 &= \tfrac{1}{2m+2}(2f_1 - f_3), \\ z_r &= \tfrac{1}{2m+2}(f_r - f_{r+2}), \quad r = 2, \ldots, m-2, \end{aligned}$$

with

$$\begin{aligned} z_{m-1} &= \tfrac{1}{2m+2}(f_{m-1} + s_o), \\ z_m &= \tfrac{1}{2m+2}(f_m + 2s_e) \end{aligned}$$

if m is even, and

$$\begin{aligned} z_{m-1} &= \tfrac{1}{2m+2}(f_{m-1} + s_e), \\ z_m &= \tfrac{1}{2m+2}(f_m + 2s_o) \end{aligned}$$

if m is odd, where

$$
\begin{aligned}
f_1 &= \mathrm{trace}(B),\\
f_2 &= \mathrm{s\_diag}(B,-1) + \mathrm{s\_diag}(B,1),\\
f_{r+1} &= \mathrm{s\_diag}(B,-r) + \mathrm{s\_diag}(B,r) -\\
&\quad \mathrm{s\_antidiag}(B,-(m+1-r)) - \mathrm{s\_antidiag}(B,m+1-r),\\
&\quad r = 2,\ldots,m-1,\\
s_o &= \sum_{r=1}^{\lceil m/2\rceil} f_{2r-1}, \qquad s_e = \sum_{r=1}^{\lfloor m/2\rfloor} f_{2r}.
\end{aligned}
$$

COROLLARY 3.2. *Let A denote the tensor obtained by scaling all the components in the first and last rows of $B \in \mathbb{C}_m^m$, except the first and last ones, by a factor $\frac{1}{2}$. The first row z of the minimizer M in Theorem 3.2 is given by*

$$
\begin{aligned}
z_1 &= \tfrac{1}{m-1}(f_1 - \tfrac{1}{2m-2}s_o),\\
z_{2r-1} &= \tfrac{1}{m-1}(f_{2r-1} - \tfrac{1}{m-1}s_o), \quad r = 2,\ldots,\left\lfloor \tfrac{m}{2}\right\rfloor,\\
z_{2r} &= \tfrac{1}{m-1}(f_{2r} - \tfrac{1}{m-1}s_e), \quad r = 1,\ldots,\left\lfloor \tfrac{m-1}{2}\right\rfloor,
\end{aligned}
$$

with

$$z_m = \tfrac{1}{m-1}(f_m - \tfrac{1}{2m-2}s_e)$$

if m is even, and

$$z_m = \tfrac{1}{m-1}(f_m - \tfrac{1}{2m-2}s_o)$$

if m is odd, where

$$
\begin{aligned}
f_1 &= \mathrm{trace}(A),\\
f_m &= \mathrm{s\_antidiag}(A,0),\\
f_{r+1} &= \mathrm{s\_diag}(A,-r) + \mathrm{s\_diag}(A,r) +\\
&\quad \mathrm{s\_antidiag}(A,-(m-1-r)) + \mathrm{s\_antidiag}(A,m-1-r),\\
&\quad r = 1,\ldots,m-2,\\
s_o &= \sum_{r=1}^{\lceil m/2\rceil} f_{2r-1}, \qquad s_e = \sum_{r=1}^{\lfloor m/2\rfloor} f_{2r}.
\end{aligned}
$$

COROLLARY 3.3. *The first row z of the minimizer M for the* ST-II *case in Theorem 3.3 is given by*

$$
\begin{aligned}
z_1 &= \tfrac{1}{2m}(2f_1 - f_2 + \tfrac{2}{m}s_e),\\
z_r &= \tfrac{1}{2m}(f_r - f_{r+1} - (-1)^r \tfrac{2}{m}s_e), \quad r = 2,\ldots,m-1,\\
z_m &= \tfrac{1}{2m}(f_m + 2(s_o + s_e) - \tfrac{2}{m}s_e)
\end{aligned}
$$

if m is even, and

$$\begin{aligned}
z_1 &= \tfrac{1}{2m}(2f_1 - f_2 - \tfrac{2}{m}s_o),\\
z_r &= \tfrac{1}{2m}(f_r - f_{r+1} + (-1)^r\tfrac{2}{m}s_o), \quad r = 2,\dots,m-1,\\
z_m &= \tfrac{1}{2m}(f_m + 2(s_o + s_e) - \tfrac{2}{m}s_o)
\end{aligned}$$

if m is odd, where

$$\begin{aligned}
f_1 &= \operatorname{trace}(B),\\
f_{r+1} &= \text{s_diag}(B,-r) + \text{s_diag}(B,r) -\\
&\quad \text{s_antidiag}(B,-(m-r)) - \text{s_antidiag}(B,m-r),\\
&\quad r = 1,\dots,m-1,\\
s_o &= \sum_{r=1}^{\lceil m/2\rceil} f_{2r-1}, \qquad s_e = \sum_{r=1}^{\lfloor m/2\rfloor} f_{2r}.
\end{aligned}$$

COROLLARY 3.4. *The first row z of the minimizer M for the* CT-II *case in Theorem* 3.3 *is given by*

$$\begin{aligned}
z_1 &= \tfrac{1}{2m}(2f_1 + f_2 - \tfrac{2}{m}s_e),\\
z_r &= \tfrac{1}{2m}(f_r + f_{r+1} - \tfrac{2}{m}s_e), \quad r = 2,\dots,m-1,\\
z_m &= \tfrac{1}{2m}(f_m + 2(s_e - s_o) - \tfrac{2}{m}s_e)
\end{aligned}$$

if m is even, and

$$\begin{aligned}
z_1 &= \tfrac{1}{2m}(2f_1 + f_2 - \tfrac{2}{m}s_o),\\
z_r &= \tfrac{1}{2m}(f_r + f_{r+1} - \tfrac{2}{m}s_o), \quad r = 2,\dots,m-1,\\
z_m &= \tfrac{1}{2m}(f_m + 2(s_o - s_e) - \tfrac{2}{m}s_o)
\end{aligned}$$

if m is odd, where

$$\begin{aligned}
f_1 &= \operatorname{trace}(B),\\
f_{r+1} &= \text{s_diag}(B,-r) + \text{s_diag}(B,r) +\\
&\quad \text{s_antidiag}(B,-(m-r)) + \text{s_antidiag}(B,m-r),\\
&\quad r = 1,\dots,m-1,\\
s_o &= \sum_{r=1}^{\lceil m/2\rceil} f_{2r-1}, \qquad s_e = \sum_{r=1}^{\lfloor m/2\rfloor} f_{2r}.
\end{aligned}$$

COROLLARY 3.5. *Let A denote the tensor obtained by scaling all the components in the last row of $B \in \mathbb{C}_m^m$, except the last one, by a factor $\frac{1}{2}$.*

The first row z of the minimizer M for the ST-III *case in Theorem* 3.4 *is given by*

$$
\begin{aligned}
z_1 &= \tfrac{1}{2m}(2f_1 - f_3),\\
z_r &= \tfrac{1}{2m}(f_r - f_{r+2}), \quad r = 2, \ldots, m-2,\\
z_r &= \tfrac{1}{2m}f_r, \quad r = m-1, m,
\end{aligned}
$$

where

$$
\begin{aligned}
f_1 &= \mathrm{trace}(A),\\
f_2 &= \mathrm{s\_diag}(A,-1) + \mathrm{s\_diag}(A,1) + \mathrm{s\_antidiag}(A, m-2),\\
f_{r+1} &= \mathrm{s\_diag}(A,-r) + \mathrm{s\_diag}(A,r) -\\
&\quad \mathrm{s\_antidiag}(A, -(m+1-r)) + \mathrm{s\_antidiag}(A, m-1-r),\\
&\quad r = 2, \ldots, m-1.
\end{aligned}
$$

COROLLARY 3.6. *Let A denote the tensor obtained by scaling all the components in the first row of $B \in \mathbb{C}^m_m$, except the first one, by a factor $\frac{1}{2}$. The first row z of the minimizer M for the* CT-III *case in Theorem* 3.4 *is given by*

$$
z = \tfrac{1}{m}f,
$$

where

$$
\begin{aligned}
f_1 &= \mathrm{trace}(A),\\
f_2 &= \mathrm{s\_diag}(A,-1) + \mathrm{s\_diag}(A,1) + \mathrm{s\_antidiag}(A, -(m-2)),\\
f_{r+1} &= \mathrm{s\_diag}(A,-r) + \mathrm{s\_diag}(A,r) +\\
&\quad \mathrm{s\_antidiag}(A, -(m-1-r)) - \mathrm{s\_antidiag}(A, m+1-r),\\
&\quad r = 2, \ldots, m-1.
\end{aligned}
$$

COROLLARY 3.7. *The first row z of the minimizer M for the* ST-IV *case in Theorem* 3.5 *is given by*

$$
\begin{aligned}
z_1 &= \tfrac{1}{2m}(2f_1 - f_2),\\
z_r &= \tfrac{1}{2m}(f_r - f_{r+1}), \quad r = 2, \ldots, m-1,\\
z_m &= \tfrac{1}{2m}f_m,
\end{aligned}
$$

where

$$
\begin{aligned}
f_1 &= \mathrm{trace}(B),\\
f_{r+1} &= \mathrm{s\_diag}(B,-r) + \mathrm{s\_diag}(B,r) -\\
&\quad \mathrm{s\_antidiag}(B, -(m-r)) + \mathrm{s\_antidiag}(B, m-r),\\
&\quad r = 1, \ldots, m-1.
\end{aligned}
$$

COROLLARY 3.8. *The first row z of the minimizer M for the* CT-IV *case in Theorem* 3.5 *is given by*

$$\begin{aligned} z_1 &= \tfrac{1}{2m}(2f_1 + f_2), \\ z_r &= \tfrac{1}{2m}(f_r + f_{r+1}), \quad r = 2, \ldots, m-1, \\ z_m &= \tfrac{1}{2m} f_m, \end{aligned}$$

where

$$\begin{aligned} f_1 &= \operatorname{trace}(B), \\ f_{r+1} &= \text{s\_diag}(B, -r) + \text{s\_diag}(B, r) + \\ &\quad \text{s\_antidiag}(B, -(m-r)) - \text{s\_antidiag}(B, m-r), \\ &\quad r = 1, \ldots, m-1. \end{aligned}$$

THEOREM 3.9. *Let $B \in \mathbb{C}_m^m$ and $M(B)$ be the minimizer of $\|B - M\|_F$ over all $M \in \mathcal{D}_Q$, where Q is associated with* MST. *Denote by $M_S(B)$ the minimizer of $\|B - M\|_F$ over all $M \in \mathcal{D}_S$, where S is associated with* ST. *Then it holds that*

$$M(B) = D M_S(D^* B D) D^*,$$

where

$$D_k^j = \begin{cases} i^{j+1} & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}, \quad j, k = 1, \ldots, m.$$

REMARK.

THEOREM 3.10 (Chan [5]). *Let $B \in \mathbb{C}_m^m$ and M be the minimizer of $\|B - M\|_F$ over all $M \in \mathcal{D}_Q$, where Q is associated with* FT. *Then the first row z of the circulant minimizer is given by*

$$z = \tfrac{1}{m} f,$$

where

$$\begin{aligned} f_1 &= \operatorname{trace}(B), \\ f_{r+1} &= \text{s\_diag}(B, r) + \text{s\_diag}(B, r-m), \quad r = 1, \ldots, m-1. \end{aligned}$$

For the monotransform case $L \geq 2$, the structure of Q is simplified into

$$Q = I \otimes {}_2Q \otimes \cdots \otimes {}_LQ,$$

where ${}_\ell Q \in \mathbb{C}_{m_\ell}^{m_\ell}$, $\ell = 2, \ldots, L$. The complexity to form Λ becomes

$$m_0^2 m_{(1)} n_1 \Big(\mathrm{a}_L(m_L)/m_L + \mathcal{O}(\log_2 m_L) + \sum_{\ell=2}^{L-1} n_{(\ell+1)} \big(\mathrm{a}_\ell(m_\ell)/m_\ell + \mathcal{O}(\log_2 m_\ell) \big) \Big),$$

where $\mathrm{a}_\ell(\cdot)$ is the complexity given in Table 3.2.

TABLE 3.2
Arithmetic complexity a$(\cdot)$ *to form* z *in the monotransform case.*

tensor class	$L = 1$, $m_0 = 1$
dense	$\mathcal{O}(m^2)$
band	$\mathcal{O}(pm)$
Toeplitz	$\mathcal{O}(m)$
band Toeplitz	$\mathcal{O}(p)$

To compute λ costs $\mathcal{O}(m \log_2 m)$ a.o., except possibly for the FT case.

4. Applications. In order to demonstrate the usefulness of normal block preconditioners and tensor notation, we consider the discretized Helmholtz equation. The revisited problem [13] is a two-dimensional restriction (plane symmetry) of time-harmonic sound waves propagating in a waveguide parametrized by Cartesian coordinates x_1 and x_2. Such problems arise in, e.g.,

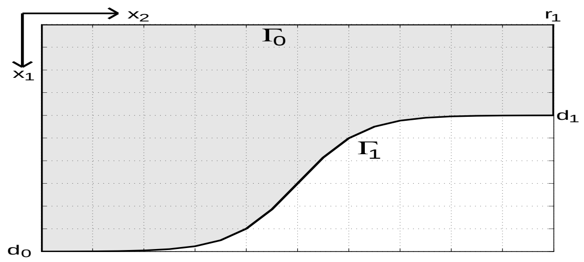

FIG. 4.1. *The physical domain.*

underwater acoustics. We study physical domains that can be mapped onto rectangles, parametrized by logical coordinates ξ_1 and ξ_2, via some orthogonal transformation [11, p. 82]. The governing equation expressed in the logical coordinates becomes

$$-\frac{\partial}{\partial \xi_1}\left(a_1 \frac{\partial u}{\partial \xi_1}\right) - \frac{\partial}{\partial \xi_2}\left(a_2 \frac{\partial u}{\partial \xi_2}\right) - a_0 u = 0, \quad 0 \le \xi_1 \le 1, \quad 0 \le \xi_2 \le 1, \tag{4.1}$$

where u is the scalar phasor [6, p. 337] of the acoustic pressure $\mathrm{Re}(ue^{-i2\pi ft})$ [8, p. 73] and the coefficients are

$$a_1 = \frac{\eta_2}{\eta_1}, \quad a_2 = \frac{\eta_1}{\eta_2}, \quad a_0 = \eta_1 \eta_2 \kappa^2. \tag{4.2}$$

The wavenumber is given by $\kappa = 2\pi f/c$, where f is the frequency, and c is the sound speed. The quantities $\eta_i(\xi_1, \xi_2)$ are the scale factors, i.e., the square root of the diagonal elements in the metric tensor.

The choice of boundary conditions at Γ_0 and Γ_1 depends on whether the walls of the waveguide are acoustically soft or rigid. The following combinations of physical boundary conditions are considered

$$\text{(4.3a)} \qquad \begin{cases} u(0,\xi_2) = 0 \\ \dfrac{\partial u}{\partial \xi_1}(1,\xi_2) = 0 \end{cases}, \quad 0 \le \xi_2 \le 1,$$

$$\text{(4.3b)} \qquad \begin{cases} \dfrac{\partial u}{\partial \xi_1}(0,\xi_2) = 0 \\ \dfrac{\partial u}{\partial \xi_1}(1,\xi_2) = 0 \end{cases}, \quad 0 \le \xi_2 \le 1,$$

$$\text{(4.3c)} \qquad \begin{cases} u(0,\xi_2) = 0 \\ u(1,\xi_2) = 0 \end{cases}, \quad 0 \le \xi_2 \le 1.$$

In underwater acoustics, where Γ_0 is the sea surface (soft wall) and the bottom Γ_1 could be considered rigid, the appropriate boundary conditions are given by (4.3a). For the artificial boundary conditions at $\xi_2 = 1$ (and $\xi_2 = 0$), Dirichlet–to–Neumann (DtN) maps [10] are employed.

Introduce uniform computational grids

$$\xi_2^k = \left(k - \tfrac{3}{2}\right) h_2, \quad k = 1, \ldots, m_2, \quad h_2 = \frac{1}{m_2 - 2}$$

and

$$\text{(4.4a)} \qquad \xi_1^j = \left(j - \tfrac{1}{2}\right) h_1, \quad j = 0, \ldots, m_1 + 1, \quad h_1 = \frac{1}{m_1}$$

or

$$\text{(4.4b)} \qquad \xi_1^j = j h_1, \qquad j = 0, \ldots, m_1 + 1, \quad h_1 = \frac{1}{m_1 + 1}.$$

For accuracy reasons [14, p. 134] mainly, we center the grid around those boundaries where the boundary conditions involve derivatives. Consequently, grid (4.4a) is utilized when the boundary conditions are (4.3a) or (4.3b), and grid (4.4b) matches (4.3c). It might seem awkward that for the Dirichlet boundary condition in (4.3a), the corresponding grid (4.4a) has no point coinciding with $\xi_1 = 0$. This seeming mismatch will be explained later. Let $u^{j,k}$ denote the approximate solution at the point (ξ_1^j, ξ_2^k). We employ a second-order acccurate finite difference method [13] to discretize (4.1) and the DtN maps. The boundary conditions (4.3) are discretized by

$$\text{(4.5a)} \qquad \begin{cases} u^{0,k} = -u^{1,k} \\ u^{m_1+1,k} = u^{m_1,k} \end{cases}, \quad k = 1, \ldots, m_2,$$

$$\text{(4.5b)} \qquad \begin{cases} u^{0,k} = u^{1,k} \\ u^{m_1+1,k} = u^{m_1,k} \end{cases}, \quad k = 1, \ldots, m_2,$$

$$\text{(4.5c)} \qquad \begin{cases} u^{0,k} = 0 \\ u^{m_1+1,k} = 0 \end{cases}, \quad k = 1, \ldots, m_2,$$

which are at least second-order accurate approximations. This results in a linear system of equations, which in component form reads

$$B^{j,k}_{r,s} u^{r,s} = g^{j,k}, \quad j = 1, \dots, m_1, \quad k = 1, \dots, m_2, \tag{4.6}$$

where g is a source term. For indices $k = 2, \dots, m_2 - 1$ corresponding to inner grid lines, the components of the coefficient tensor B are

$$\left.\begin{aligned} B^{1,\underline{k}}_{1,\underline{k}} &= 2a_1^{\frac{1}{2},k} + a_1^{\frac{3}{2},k} + \alpha(a_2^{1,k-\frac{1}{2}} + a_2^{1,k+\frac{1}{2}}) - (h_1)^2 a_0^{1,k} \\ B^{m_1,\underline{k}}_{m_1,\underline{k}} &= a_1^{m_1-\frac{1}{2},k} + \alpha(a_2^{m_1,k-\frac{1}{2}} + a_2^{m_1,k+\frac{1}{2}}) - (h_1)^2 a_0^{m_1,k} \end{aligned}\right\} \text{for (4.5a)},$$

$$\left.\begin{aligned} B^{1,\underline{k}}_{1,\underline{k}} &= a_1^{\frac{3}{2},k} + \alpha(a_2^{1,k-\frac{1}{2}} + a_2^{1,k+\frac{1}{2}}) - (h_1)^2 a_0^{1,k} \\ B^{m_1,\underline{k}}_{m_1,\underline{k}} &= a_1^{m_1-\frac{1}{2},k} + \alpha(a_2^{m_1,k-\frac{1}{2}} + a_2^{m_1,k+\frac{1}{2}}) - (h_1)^2 a_0^{m_1,k} \end{aligned}\right\} \text{for (4.5b)},$$

$$\left.\begin{aligned} B^{1,\underline{k}}_{1,\underline{k}} &= a_1^{\frac{1}{2},k} + a_1^{\frac{3}{2},k} + \alpha(a_2^{1,k-\frac{1}{2}} + a_2^{1,k+\frac{1}{2}}) - (h_1)^2 a_0^{1,k} \\ B^{m_1,\underline{k}}_{m_1,\underline{k}} &= a_1^{m_1-\frac{1}{2},k} + a_1^{m_1+\frac{1}{2},k} + \alpha(a_2^{m_1,k-\frac{1}{2}} + a_2^{m_1,k+\frac{1}{2}}) - (h_1)^2 a_0^{m_1,k} \end{aligned}\right\} \text{for (4.5c)},$$

$$\begin{aligned} B^{\underline{j},\underline{k}}_{\underline{j},\underline{k}} &= a_1^{j-\frac{1}{2},k} + a_1^{j+\frac{1}{2},k} + \alpha(a_2^{j,k-\frac{1}{2}} + a_2^{j,k+\frac{1}{2}}) - (h_1)^2 a_0^{j,k}, \; j = 2, \dots, m_1 - 1, \\ B^{\underline{j},\underline{k}}_{\underline{j}-1,\underline{k}} &= -a_1^{j-\frac{1}{2},k}, \quad j = 2, \dots, m_1, \\ B^{\underline{j},\underline{k}}_{\underline{j}+1,\underline{k}} &= -a_1^{j+\frac{1}{2},k}, \quad j = 1, \dots, m_1 - 1, \\ B^{\underline{j},\underline{k}}_{\underline{j},\underline{k}-1} &= -\alpha a_2^{j,k-\frac{1}{2}}, \quad B^{\underline{j},\underline{k}}_{\underline{j},\underline{k}+1} = -\alpha a_2^{j,k+\frac{1}{2}}, \quad j = 1, \dots, m_1, \end{aligned}$$

where $\alpha = (\frac{h_1}{h_2})^2$. For border grid lines, the entries are

$$B^{:,1}_{:,1} = I + D^{:,1}_{:,1}, \quad B^{:,1}_{:,2} = -I + D^{:,1}_{:,2}, \tag{4.7a}$$

$$B^{:,m_2}_{:,m_2-1} = -I + D^{:,m_2}_{:,m_2-1}, \quad B^{:,m_2}_{:,m_2} = I + D^{:,m_2}_{:,m_2}, \tag{4.7b}$$

$$D^{:,1}_{:,s} = -ih_2 \sum_{m=1}^{\mu_0} \sqrt{-\lambda_m^0}\,\psi_m \psi_m^* \frac{h_1}{2}, \tag{4.7c}$$

$$D^{:,m_2}_{:,s} = -ih_2 \sum_{m=1}^{\mu_1} \sqrt{-\lambda_m^1}\,\psi_m \psi_m^* \frac{h_1}{2}, \tag{4.7d}$$

where λ_m^0, λ_m^1, and ψ_m are eigenvalues and eigenvectors to discretized Sturm–Liouville problems [13] arising for the DtN maps. The eigenvectors satisfy

$$\psi_k^j \propto \left\{\begin{matrix} \sin\big((k-\frac{1}{2})\pi\xi_1^j\big) & \text{for (4.3a)} \\ \cos\big((k-1)\pi\xi_1^j\big) & \text{for (4.3b)} \\ \sin(k\pi\xi_1^j) & \text{for (4.3c)} \end{matrix}\right\} = \tfrac{1}{\sqrt{h_1}} Q_k^j, \quad j = 1, \dots, m_1, \tag{4.8}$$

where Q is associated with ST-IV, CT-II, ST as defined in Appendix A.

4.1. Preconditioning. We design a normal block preconditioner for the coefficient tensor in (4.6), i.e., the case $L = 2$ and $m_0 = 1$. Due to the boundary conditions (4.5) and the dyadic structure of $D^{:,1}_{:,s}$ and $D^{:,m_2}_{:,s}$ in (4.7), we make the convenient choice to apply transforms in level $\ell = 1$. Thus, the preconditioner M is composed as

$$M = (Q \otimes I)\Lambda(Q^* \otimes I), \quad Q \in \mathbb{C}^{m_1}_{m_1}, \tag{4.9a}$$

$$\Lambda^{j_1,j_2}_{k_1,k_2} = 0, \quad \text{if } j_1 \neq k_1. \tag{4.9b}$$

In order to determine an appropriate transform Q, we consider a separable case where the coefficients a_i in (4.2) only depend on ξ_2. This yields

$$B^{:,\underline{k}}_{:,\underline{k}} = \big(2a_1^k + \alpha(a_2^{k-\frac{1}{2}} + a_2^{k+\frac{1}{2}}) - (h_1)^2 a_0^k\big)I - a_1^k\,{}^1R, \tag{4.10a}$$

$$B^{:,\underline{k}}_{:,\underline{k}-1} = -\alpha a_2^{k-\frac{1}{2}} I, \quad B^{:,\underline{k}}_{:,\underline{k}+1} = -\alpha a_2^{k+\frac{1}{2}} I, \quad k = 2, \dots, m_2 - 1, \tag{4.10b}$$

where ${}^1R$ is defined in Lemma 3.7, 3.4, 3.1 for boundary conditions (4.5a), (4.5b), (4.5c), respectively. The entries in (4.7) are not affected. An important special case of a separable configuration is a waveguide with straight walls (in a homogeneous medium), which lead to constant coefficients a_i.

PROPOSITION 4.1. *Let B denote the coefficient tensor for the discretized separable Helmholtz problem given by* (4.10, 4.7, 4.8). *Let the structure of M be defined by* (4.9), *where Q is associated with* ST-IV, CT-II, ST *for boundary conditions* (4.5a), (4.5b), (4.5c), *respectively. The solution to the minimization problem $\min_{\Lambda} \|B - M\|_F$ satisfies $M = B$.*

Proof. According to Lemma 3.7, 3.4, 3.1, it holds that

$$Q^{-1}B^{:,\underline{k}}_{:,\underline{k}}Q = \big(2a_1^k + \alpha(a_2^{k-\frac{1}{2}} + a_2^{k+\frac{1}{2}}) - (h_1)^2 a_0^k\big)I - a_1^k \operatorname{diag}(\lambda_1^1, \dots, \lambda_{m_1}^1),$$

$$Q^{-1}B^{:,\underline{k}}_{:,\underline{k}-1}Q = -\alpha a_2^{k-\frac{1}{2}} I, \quad Q^{-1}B^{:,\underline{k}}_{:,\underline{k}+1}Q = -\alpha a_2^{k+\frac{1}{2}} I, \quad k = 2, \dots, m_2 - 1.$$

Since Q is unitary, eqs. (4.7c) and (4.8) yield

$$\begin{aligned} Q^* D^{:,1}_{:,s} Q &= -\tfrac{i}{2}h_2 \sum_{m=1}^{\mu_0} \sqrt{-\lambda_m^0}\, Q^* \psi_m h_1 \psi_m^* Q = -\tfrac{i}{2}h_2 \sum_{m=1}^{\mu_0} \sqrt{-\lambda_m^0}\, e_m e_m^* \\ &= -\tfrac{i}{2}h_2 \operatorname{diag}(\sqrt{-\lambda_1^0}, \dots, \sqrt{-\lambda_{\mu_0}^0}, 0, \dots, 0). \end{aligned}$$

Consequently, $Q^* B^{:,1}_{:,s} Q$ and $Q^* B^{:,m_2}_{:,s} Q$ are diagonalized. Since all the subtensors $Q^* B^{:,k}_{:,s} Q$, $k = 1, \dots, m_2$, are diagonalized, the minimizer is given by

$$\Lambda^{:,k}_{:,s} = Q^* B^{:,k}_{:,s} Q.$$

This leads to $M = B$. ◻

Hence, the operator M^{-1} is a direct fast solver for the Helmholtz problem (4.6), in the separable case where B is given by (4.10, 4.7, 4.8). We match the transform for the preconditioner M to the boundary conditions (4.5) according to Proposition 4.1, even when B originates from an inseparable problem. The preconditioner will presumably act like a viable convergence accelerator. After left preconditioning, the linear system (4.6) is transformed into $M^{-1}Bu = M^{-1}g$. The preconditioned system is solved by a Krylov subspace method. For simplicity and robustness, we choose the restarted generalized minimal residual (GMRES(ℓ)) algorithm [15], where ℓ is the restarting length.

4.2. Complexity. The efficiency of the preconditioned iterative method regarding memory requirement and arithmetic complexity is now analyzed for the Helmholtz problem with Dirichlet–Neumann boundary conditions (4.5a). Since the analysis is similar to the discussion in [13, § 4.2], mainly the differences are pointed out. The number of arithmetic operations and the storage requirements will be normalized by the number of unknowns $m_1 m_2$.

As an initial approximation we utilize the preconditioned right-hand side $M^{-1}g$, which is advantageous if $M^{-1}B$ is close to the identity tensor. We have imposed the following stopping criterion

$$\frac{\|M^{-1}(g - Bu_i)\|_2}{\|M^{-1}g\|_2} < \epsilon$$

with tolerance $\epsilon = 10^{-4}$.

Forming the minimizer, according to Corollary 3.7 and eq. (3.1), and subsequently factorizing the tridiagonal subtensors $\Lambda^{j,:}_{\underline{j},:}$ cost $a_{pf} = 10\frac{\hat{m}}{m_1}\log_2 \hat{m} + 71$, where $\hat{m} = 2^{\lceil \log_2 m_1 \rceil + 1}$. The computation of the initial approximation requires $a_{ps} = 20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 89$ arithmetic operations per unknown for the sine transform-IV. Let n_{it} be the number of iterations required for convergence to solve $M^{-1}Bu = M^{-1}g$ with the GMRES(ℓ) method. In Table 4.1, the total complexity for preconditioned GMRES(ℓ) is summarized and compared with band Gaussian elimination, which is the standard direct method.

TABLE 4.1
Comparison of preconditioned GMRES(ℓ) *and band Gaussian elimination.*

	arithmetic complexity	memory requirement
band GE	$8m_1^2 + 25m_1$	$4m_1 + 4$
GMRES(ℓ)	$70\frac{\hat{m}}{m_1}\log_2 \hat{m} + 388$ $+ n_{it}(20\frac{\hat{m}}{m_1}\log_2 \hat{m} + 8\ell + 167)$	$2\ell + 19 + 2\frac{\hat{m}}{m_1}$

By comparing the arithmetic costs for one iteration of preconditioned GMRES(ℓ) given in Table 4.1 and [13, Table 1], one realizes that the cost has been more than halved. Notice that $\hat{m}$ has been reduced from $2^{\lceil \log_2(2m_1+1) \rceil + 1}$ in [13, § 4.2] to $2^{\lceil \log_2 m_1 \rceil + 1}$. The reason for this reduction is that the sine transform-IV of length m_1 can be derived from a Fourier transform of the same length [12], whereas the odd sine transform-III [16] requires a Fourier transform of length $2m_1 + 1$. In [13], the odd sine transform-III was chosen, since it matched the Dirichlet–Neumann boundary conditions *and* the grid in analogy with Proposition 4.1. The ξ_1-grid used in [13] has a point coinciding with $\xi_1 = 0$, whereas grid (4.4a) does not. As mentioned earlier, the latter might seem mismatching for the Dirichlet boundary condition in (4.3a). However, the benefit of halving the arithmetic work clearly outweighs the inconvenience.

4.3. Numerical experiments. For some inseparable configurations, the system of equations (4.6) has been solved using the GMRES(ℓ) method combined with the preconditioner defined in §4.1. The bathymetry of the duct, the source term, and the resolution criteria are the same as in [13]. The orthogonal grid is generated by a code based on the method described in [1].

A fundamental issue is how the frequency affects the number of iterations. This is demonstrated in Table 4.2 for a duct of medium steepness and frequencies in the low–to–intermediate range.

TABLE 4.2
Number of iterations as a function of frequency for different boundary conditions.

frequency [Hz]	25	50	75	100	125	150	175
Dirichlet–Neumann	2	4	4	5	6	6	7
Neumann–Neumann	3	4	5	5	6	7	8
Dirichlet–Dirichlet		3	3	4	5	5	6

The number of iterations grows moderately as the frequency increases. The number of unknowns depends cubicly on the frequency and ranges from 7452 to 2563902. In Figs. 4.2 and 4.3, the results from comparative experiments are shown for the Dirichlet–Neumann case. We define the arithmetic gain for one method over another as the total number of arithmetic operations for the latter divided by the total arithmetic complexity of the former. Analogously, the memory gain is defined in terms of memory requirements.

It is clear that GMRES(6) is more efficient than band Gaussian elimination both regarding arithmetic complexity and memory requirement for all problem sizes considered. The gains rise as the frequency increases. By comparing the arithmetic gains presented in Fig. 4.2 and [13, Fig. 6], an increase by a factor two can be observed. The prediction in §4.2 of halving the arithmetic work for one iteration is corroborated, and it holds even for the total arithmetic cost. A comparison of Fig. 4.3 and [13, Fig. 7] shows that the memory gain is somewhat less than doubled.

5. Concluding remarks. A unified approach to construct preconditioners based on fast (trigonometric) transforms has been presented. The normal block framework is aimed at preconditioners for multi-dimensional PDE problems discretized on structured grids. We advocate tensor notation in order to exploit the inherent structure of the arising operators, which is easily lost when embedding them into ordinary matrices.

The usefulness of the preconditioner framework and tensor notation is demonstrated by considering the discretized Helmholtz equation subject to Dirichlet, Neumann, and DtN boundary conditions for a two-dimensional

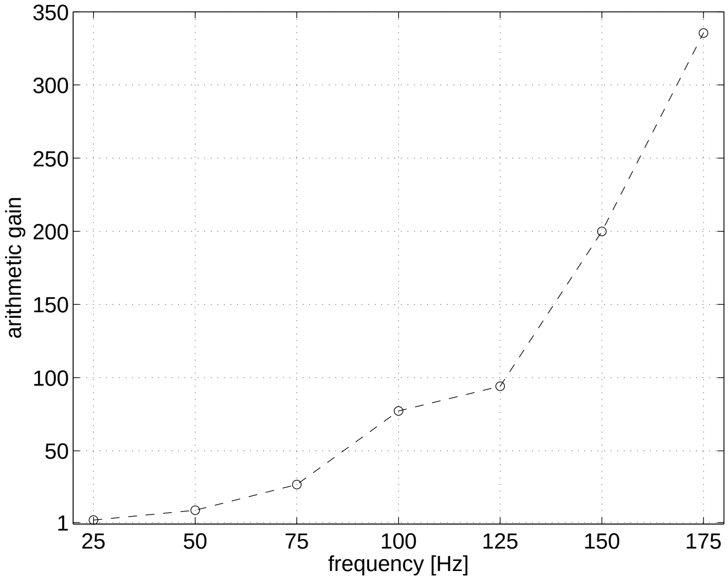

FIG. 4.2. *Arithmetic gain for preconditioned* GMRES(6) *compared with band Gaussian elimination.*

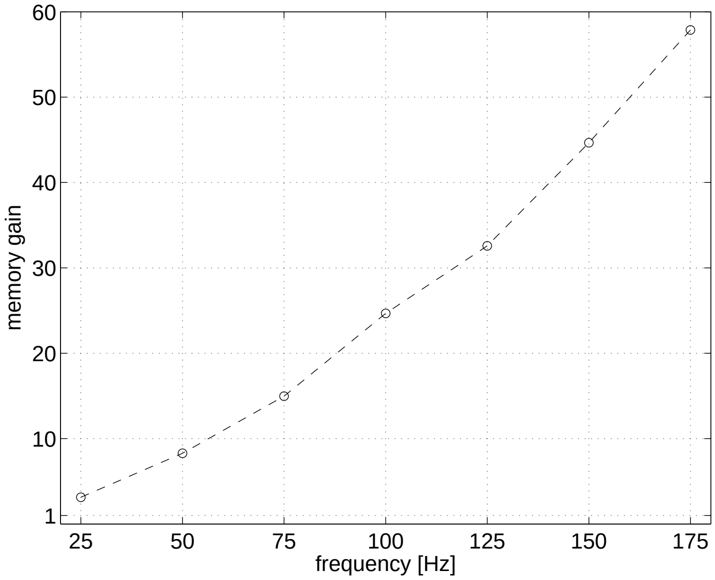

FIG. 4.3. *Memory gain for preconditioned* GMRES(6) *compared with band Gaussian elimination.*

waveguide problem. We employ a normal block preconditioner, which results in a direct fast Helmholtz solver for separable configurations, provided that the transform matches the boundary conditions and the grid. By a careful choice of grid for an inseparable waveguide problem arising in underwater acoustics, the arithmetic complexity is halved compared with a similar, previously constructed method [13].

Appendix A. Trigonometric transforms. The tensors Q associated with various trigonometric transforms are listed. The naming principle is basically the same as in [16]. We adopt the convention that a multiplication by Q is a transform synthesis (inverse transform), whereas a multiplication by Q^{-1} is a transform analysis. All the transforms throughout this section can be performed by fast algorithms [12]. The naming of some transforms in [7, 13] differs from the notation in this paper. The transforms called sine transform-II and cosine transform in [7, 13] are named odd sine transform-III and cosine transform-II in the present notation.

The Fourier transform. The entries of the unitary tensor Q, associated with the Fourier transform (FT), are given by

$$Q_k^j = \sqrt{\tfrac{1}{m}}\exp\Big(\frac{ikj2\pi}{m}\Big), \quad k,j = 0,\dots,m-1.$$

The sine transform. The entries of the unitary tensor Q, associated with the sine transform (ST), are given by

$$Q_k^j = \sqrt{\tfrac{2}{m+1}}\sin\Big(\frac{kj\pi}{m+1}\Big), \quad k,j = 1,\dots,m.$$

The cosine transform-I. The tensor associated with the cosine transform-I (CT-I) is given by $D^{-1}Q$, where $D = \operatorname{diag}(\frac{1}{\sqrt{2}}, 1, \dots, 1, \frac{1}{\sqrt{2}})$ and the entries of Q are

$$Q_0^j = \begin{cases} \sqrt{\frac{1}{2(m-1)}}, & j = 0 \\ \sqrt{\frac{1}{m-1}}, & j = 1,\dots,m-2 \\ \sqrt{\frac{1}{2(m-1)}}, & j = m-1 \end{cases}$$

$$Q_{m-1}^j = \begin{cases} \sqrt{\frac{1}{2(m-1)}}, & j = 0 \\ \sqrt{\frac{1}{m-1}}(-1)^j, & j = 1,\dots,m-2 \\ \sqrt{\frac{1}{2(m-1)}}(-1)^{m-1}, & j = m-1 \end{cases}$$

$$Q_k^j = \left\{ \begin{array}{ll} \sqrt{\frac{1}{m-1}}, & j = 0 \\ \sqrt{\frac{2}{m-1}}\cos\Big(\dfrac{kj\pi}{m-1}\Big), & j = 1,\dots,m-2 \\ \sqrt{\frac{1}{m-1}}(-1)^k, & j = m-1 \end{array} \right\}, \quad k = 1,\dots,m-2.$$

The sine transform-II. The entries of the unitary tensor Q, associated with the sine transform-II (ST-II), are given by

$$\left. \begin{array}{ll} Q_k^j = \sqrt{\frac{2}{m}}\sin\Big(\dfrac{k(j-\frac{1}{2})\pi}{m}\Big), & k = 1,\dots,m-1 \\ Q_m^j = \sqrt{\frac{1}{m}}(-1)^{j-1} & \end{array} \right\}, \quad j = 1,\dots,m.$$

The cosine transform-II. The entries of the unitary tensor Q, associated with the cosine transform-II (CT-II), are given by

$$\left.\begin{aligned} Q_1^j &= \sqrt{\tfrac{1}{m}} \\ Q_k^j &= \sqrt{\tfrac{2}{m}}\cos\Big(\frac{(k-1)(j-\frac{1}{2})\pi}{m}\Big), \quad k=2,\ldots,m \end{aligned}\right\}, \quad j=1,\ldots,m.$$

The sine transform-III. The tensor associated with the sine transform-III (ST-III) is given by $D^{-1}Q^*$, where $D = \mathrm{diag}(1,\ldots,1,\frac{1}{\sqrt{2}})$ and Q is associated with ST-II.

The cosine transform-III. The tensor associated with the cosine transform-III (CT-III) is given by $D^{-1}Q^*$, where $D = \mathrm{diag}(\frac{1}{\sqrt{2}},1,\ldots,1)$ and Q is associated with CT-II.

The sine transform-IV. The entries of the unitary tensor Q, associated with the sine transform-IV (ST-IV), are given by

$$Q_k^j = \sqrt{\tfrac{2}{m}}\sin\Big(\frac{(k-\frac{1}{2})(j-\frac{1}{2})\pi}{m}\Big), \quad k,j=1,\ldots,m.$$

The cosine transform-IV. The entries of the unitary tensor Q, associated with the cosine transform-IV (CT-IV), are given by

$$Q_k^j = \sqrt{\tfrac{2}{m}}\cos\Big(\frac{(k-\frac{1}{2})(j-\frac{1}{2})\pi}{m}\Big), \quad k,j=1,\ldots,m.$$

The modified sine transform. The entries of the unitary tensor Q, associated with the modified sine transform (MST), are given by

$$Q_k^j = \sqrt{\tfrac{2}{m+1}}\, i^{j+1}\sin\Big(\frac{kj\pi}{m+1}\Big), \quad k,j=1,\ldots,m.$$

Appendix B. The entries of the unitary tensor Q, associated with the odd sine transform-III (ODD ST-III), are given by

$$Q_k^j = \sqrt{\tfrac{2}{m+\frac{1}{2}}}\sin\Big(\frac{(k-\frac{1}{2})j\pi}{m+\frac{1}{2}}\Big), \quad k,j=1,\ldots,m.$$

We state results analogous to those in §3.2 for ODD ST-III, which is relevant to the case treated in [13].

LEMMA B.1. *Let Q be associated with* ODD ST-III, *and introduce the following tensors in* $\mathbb{R}_m^m$

$$^0R = I,$$

$$^rR_k^j = \begin{cases} 1 & \textit{if } k-j=-r, \\ 1 & \textit{if } k-j=r, \\ -1 & \textit{if } k+j=r, \\ 1 & \textit{if } k+j=2m+1-r, \\ 0 & \textit{otherwise}, \quad r=1,\ldots,m-1. \end{cases}$$

Then $\{{}^rR\}_{r=0}^{m-1}$ *is a basis for* $\mathcal{D}_Q$. *Moreover, when* $r > 0$ *the spectrum of* ${}^rR$ *is*

$$\lambda_k^r = 2\cos\Big(\frac{r(k-\frac{1}{2})\pi}{m+\frac{1}{2}}\Big), \quad k = 1, \ldots, m.$$

THEOREM B.1. *Let* $B \in \mathbb{C}_m^m$ *and* M *be the minimizer of* $\|B - M\|_F$ *over all* $M \in \mathcal{D}_Q$, *where* Q *is associated with* ODD ST-III. *Denote by* u *the column vector with* m *entries* $u^j = (-1)^j$, *and by* e *the first canonical column vector of length* m. *Then it holds that*

$$M = \mathcal{T}(\mu) + \mathcal{H}\big(-\mathrm{s}_2(\mu), \sigma_1(\mu)\big)$$

with

$$\mu = \tfrac{1}{2m+1} f(I + ee^* + uu^*).$$

The entries of the row vector f *are*

$$f_{r+1} = \mathrm{trace}({}^rR^* B), \quad r = 0, \ldots, m-1,$$

where ${}^rR$ *is given by Lemma* B.1.

COROLLARY B.1. *The first row* z *of the minimizer* M *in Theorem* B.1 *is given by*

$$\begin{aligned} z_1 &= \tfrac{1}{2m+1}(2f_1 - f_3), \\ z_r &= \tfrac{1}{2m+1}(f_r - f_{r+2}), \quad r = 2, \ldots, m-2, \end{aligned}$$

with

$$\begin{aligned} z_{m-1} &= \tfrac{1}{2m+1}(f_{m-1} + s_o - s_e), \\ z_m &= \tfrac{1}{2m+1}(f_m + s_e - s_o) \end{aligned}$$

if m *is even, and*

$$\begin{aligned} z_{m-1} &= \tfrac{1}{2m+1}(f_{m-1} + s_e - s_o), \\ z_m &= \tfrac{1}{2m+1}(f_m + s_o - s_e) \end{aligned}$$

if m *is odd, where*

$$\begin{aligned} f_1 &= \mathrm{trace}(B), \\ f_2 &= \mathrm{s\_diag}(B, -1) + \mathrm{s\_diag}(B, 1) + \mathrm{s\_antidiag}(B, m-1), \\ f_{r+1} &= \mathrm{s\_diag}(B, -r) + \mathrm{s\_diag}(B, r) - \\ &\quad\ \mathrm{s\_antidiag}(B, -(m+1-r)) + \mathrm{s\_antidiag}(B, m-r), \\ &\quad\ r = 2, \ldots, m-1, \\ s_o &= \sum_{r=1}^{\lceil m/2 \rceil} f_{2r-1}, \qquad s_e = \sum_{r=1}^{\lfloor m/2 \rfloor} f_{2r}. \end{aligned}$$

Acknowledgments. I would like to thank Elisabeth Larsson for bringing reference [4] to my attention.

REFERENCES

[1] L. Abrahamsson, *Orthogonal grid generation for two-dimensional ducts*, J. Comput. Appl. Math., 34 (1991), pp. 305–314.

[2] E. Boman and I. Koltracht, *Fast transform based preconditioners for Toeplitz equations*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 628–645.

[3] R. H. Chan, T. F. Chan, and C.-K. Wong, *Cosine transform based preconditioners for total variation deblurring*, IEEE Trans. Image Process., 8 (1999), pp. 1472–1478.

[4] R. H. Chan, M. K. Ng, and C.-K. Wong, *Sine transform based preconditioners for symmetric Toeplitz systems*, Linear Algebra Appl., 232 (1996), pp. 237–259.

[5] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[6] D. K. Cheng, *Field and Wave Electromagnetics*, Addison–Wesley, Reading, MA, 2nd ed., 1989.

[7] S. Holmgren and K. Otto, *A framework for polynomial preconditioners based on fast transforms I: Theory*, BIT, 38 (1998), pp. 544–559.

[8] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*, AIP Press, New York, 1994.

[9] T. Kailath and V. Olshevsky, *Displacement structure approach to discrete trigonometric transform based preconditioners of G. Strang type and of T. Chan type*, Calcolo, 33 (1996), pp. 191–208.

[10] J. B. Keller and D. Givoli, *Exact non-reflecting boundary conditions*, J. Comput. Phys., 82 (1989), pp. 172–192.

[11] P. Knupp and S. Steinberg, *Fundamentals of Grid Generation*, CRC Press, Boca Raton, FL, 1993.

[12] K. Otto, *Quick Transform Guide*, `http://www.tdb.uu.se/~kurt/qtg.ps`, 1999.

[13] K. Otto and E. Larsson, *Iterative solution of the Helmholtz equation by a second-order method*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 209–229.

[14] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, John Wiley & Sons, New York, 2nd ed., 1967.

[15] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[16] Z. Wang and B. R. Hunt, *The discrete W transform*, Appl. Math. Comput., 16 (1985), pp. 19–48.