Numer. Math. (2005) 99: 441–484
Digital Object Identifier (DOI) 10.1007/s00211-004-0574-1

Numerische Mathematik

# Preconditioned HSS methods for the solution of non-Hermitian positive definite linear systems and applications to the discrete convection-diffusion equation

**Daniele Bertaccini**[1], **Gene H. Golub**[2], **Stefano Serra Capizzano**[3], **Cristina Tablino Possio**[4]

[1] Dipartimento di Matematica, Università "La Sapienza", P.le A. Moro 2, 00185 Roma, Italy; e-mail: bertaccini@mat.uniroma1.it.
[2] Department of Computer Science, Stanford University, Gates 2B, CA 94305, USA; e-mail: golub@stanford.edu.
[3] Dipartimento di Fisica e Matematica, Università dell'Insubria, Via Valleggio 11, 22100, Como, Italy; e-mail: stefano.serrac@uninsubria.it, serra@mail.dm.unipi.it.
[4] Dipartimento di Matematica e Applicazioni, Università di Milano Bicocca, via Bicocca degli Arcimboldi 8, 20126, Milano, Italy; e-mail: cristina.tablinopossio@unimib.it.

Received November 2, 2002 / Revised version received October 30, 2004
Published online December 14, 2004 – © Springer-Verlag 2004

**Summary.** We study the role of preconditioning strategies recently developed for coercive problems in connection with a two-step iterative method, based on the Hermitian skew-Hermitian splitting (HSS) of the coefficient matrix, proposed by Bai, Golub and Ng for the solution of nonsymmetric linear systems whose real part is coercive. As a model problem we consider Finite Differences (FD) matrix sequences $\{A_n(a,\,p)\}_n$ discretizing the elliptic (convection-diffusion) problem

$$(1)\quad \begin{cases} -\nabla^T[a(x)\nabla u(x)] + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}(p(x)u(x)) = f(x), & x \in \Omega, \\ \\ \text{Dirichlet BC}, \end{cases}$$

with $\Omega$ being a plurirectangle of $\mathbf{R}^d$ with $a(x)$ being a uniformly positive function and $p(x)$ denoting the Reynolds function: here for plurirectangle we mean a connected union of rectangles in $d$ dimensions with edges parallel to the axes. More precisely, in connection with preconditioned HSS/GMRES like methods, we consider the preconditioning sequence $\{P_n(a)\}_n$, $P_n(a) := D_n^{1/2}(a)A_n(1,0)D_n^{1/2}(a)$ where $D_n(a)$ is the suitably scaled main diagonal

of $A_n(a,0)$. If $a(x)$ is positive and regular enough, then the preconditioned sequence shows a strong clustering at unity so that the sequence $\{P_n(a)\}_n$ turns out to be a superlinear preconditioning sequence for $\{A_n(a,0)\}_n$ where $A_n(a,0)$ represents a good approximation of $\mathrm{Re}(A_n(a,p))$ namely the real part of $A_n(a,p)$.

The computational interest is due to the fact that the preconditioned HSS method has a convergence behavior depending on the spectral properties of $\{P_n^{-1}(a)\mathrm{Re}(A_n(a,p))\}_n \approx \{P_n^{-1}(a)A_n(a,0)\}_n$: therefore the solution of a linear system with coefficient matrix $A_n(a,p)$ is reduced to computations involving diagonals and to the use of fast Poisson solvers for $\{A_n(1,0)\}_n$.

Some numerical experimentations confirm the optimality of the discussed proposal and its superiority with respect to existing techniques.

*Mathematics Subject Classification (1991):* 65F10, 65N22, 15A18, 15A12, 47B65

## 1 Introduction and description of the HSS method

Several applications in scientific computing lead to systems of linear equations

$$A_n x = b, \quad A_n \in \mathbf{C}^{n\times n}, \text{ nonsingular, and } x, b \in \mathbf{C}^n$$

where the coefficient matrix $A_n$ is large and sparse and possesses a positive definite real part. In particular, this is the case of the matrices related to the discretization important classes of time-dependent partial differential equations, see [4,5]. This basic constraint suggested to Bai, Golub and Ng to use a natural splitting of $A_n$ in terms of the Hermitian part and of the skew-Hermitian part of $A_n$ (see also [12,13] for related splittings). More precisely, for a given matrix $A_n$, there exists a unique Hermitian/skew-Hermitian decomposition

$$A_n = \mathrm{Re}(A_n) + \mathrm{i}\,\mathrm{Im}(A_n), \quad \mathrm{i}^2 = -1, \tag{2}$$

where

$$\mathrm{Re}(A_n) := \frac{A_n + A_n^H}{2} \quad \text{and} \quad \mathrm{Im}(A_n) := \frac{A_n - A_n^H}{2\mathrm{i}}.$$

We are interested in the case where the real part $\mathrm{Re}(A_n)$, which is Hermitian by definition, is positive definite as well. Following [2], the considered *Hermitian/skew-Hermitian splitting (HSS)* can be related to a two-step iteration (in the spirit of the ADI method) in the following way

$$\begin{cases} (\alpha I + \mathrm{Re}(A_n))\, x^{k+\frac{1}{2}} &= (\alpha I - \mathrm{i}\,\mathrm{Im}(A_n))\, x^k + b \\ (\alpha I + \mathrm{i}\,\mathrm{Im}(A_n))\, x^{k+1} &= (\alpha I - \mathrm{Re}(A_n))\, x^{k+\frac{1}{2}} + b \end{cases} \tag{3}$$

with $\alpha$ positive parameter and $x^0$ given initial guess: the related iterative method is named *HSS iteration or HSS method*. It is interesting that the above method can be reinterpreted as a stationary iterative procedure whose iteration matrix

$$M(\alpha) = (\alpha I + \mathrm{i}\,\mathrm{Im}(A_n))^{-1}(\alpha I - \mathrm{Re}(A_n)) \times (\alpha I + \mathrm{Re}(A_n))^{-1}(\alpha I - \mathrm{i}\,\mathrm{Im}(A_n)) \tag{4}$$

is well defined: indeed the matrix $\alpha I + \mathrm{i}\,\mathrm{Im}(A_n)$ is invertible since $\alpha$ is nonzero and $\mathrm{i}\,\mathrm{Im}(A_n)$ is skew-Hermitian and $\alpha I + \mathrm{Re}(A_n)$ is also invertible due to the positivity of $\alpha$ and to the positive definiteness of $\mathrm{Re}(A_n)$ (the only structural assumption that we use).

Moreover, Bai, Golub and Ng proved that the convergence is only related to the spectral radius of the Hermitian matrix

$$(\alpha I - \mathrm{Re}(A_n))(\alpha I + \mathrm{Re}(A_n))^{-1}$$

which is unconditionally bounded by 1 under the assumption of positivity of $\alpha$ and of $\mathrm{Re}(A_n)$. However, a finer analysis in the case of a constant coefficient PDE of the type (1) has shown that the best contraction factor is

$$1 - ch + O(h^2)$$

where $c$ is a positive fixed constant independent of $n$, $h \sim n^{-1/d}$ is the "discretization parameter" and $d$ is the dimension of the space in which the domain $\Omega$ lies.

This result can be unsatisfactory for large $n$ and therefore we propose the use of a preconditioning by means of a Hermitian positive definite matrix $P_n$. Our analysis is developed in three main directions.

(**A**) First we consider the generic case by proving that (**A.1**) the unconditional convergence holds in the preconditioned version as well: in that case the convergence factor is given by the spectral radius of

$$\left(\alpha I - P_n^{-1/2}\mathrm{Re}(A_n)P_n^{-1/2}\right)\left(\alpha I + P_n^{-1/2}\mathrm{Re}(A_n)P_n^{-1/2}\right)^{-1} \tag{5}$$

where the optimal parameter $\alpha$ is the square root of the product of the extreme eigenvalues of $P_n^{-1}\mathrm{Re}(A_n)$; (**A.2**) the analysis can be refined in the case of non-normal matrices and indeed, quite surprisingly, the skew-Hermitian contributions in the iteration matrix can have a role in accelerating the convergence and the explanation of this phenomenon falls in the theory of multi-iterative methods [27]. We recall that basically a multi-iterative method is an iterative method which is the composition of a finite number of simple iterations whose main features are the following: each iterative technique is cheap and potentially slowly convergent, the iterations have a complementary spectral behavior in such a way that their composition is fast convergent. We

recall that a classical example is the multigrid method for elliptic differential problems where the smoother quickly converges in the space of low frequencies (but is very slow in the high frequencies) and the coarse-grid correction is not convergent at all, but is very fast in the high frequencies space (see e.g. [18,8,31]).
(**B**) Second we consider the model problem and for that setting we introduce preconditioning strategies that are optimal for the HSS method: we denote by *PHSS method* the *preconditioned HSS method*.
(**C**) Under additional assumptions, the considered preconditioners are optimal even when suitable iterative methods are used as inner iteration processes at each step of the outer PHSS iterations. In this case, the convergence analysis is formally much more complicate; for instance the minimal condition number among the eigenvector matrices comes into play. In [2] an analysis based on generic inner iteration algorithms has been performed for the HSS iterations.

More specifically, we define the preconditioning sequence $\{P_n(a)\}_n$,

$$P_n(a) := D_n^{1/2}(a) A_n(1,0) D_n^{1/2}(a) \tag{6}$$

where $D_n(a)$ is the suitably scaled main diagonal of $A_n(a,0)$: we just mention that there exist other examples of preconditioners [4,5,11,15,22] whose definition comes from the discretization of simpler differential equations and/or with different boundary conditions. $P_n(a)$ is an *approximate factorization* of $A_n(a,0)$ (and therefore of $A_n(a,p)$ if the norm of $\nabla p$ is not too large, see Subsection 3.4) in the sense that $P_n(a)$ is the product of simpler matrices, for which fast solvers are available, and $A_n(a,0) - P_n(a)$ has infinitesimal spectral norm under the sole assumption of continuity of $a$. Moreover, if $a(x)$ is positive and regular enough, then

- the preconditioned sequence shows a strong clustering at unity,
- $\{P_n(a)\}_n$ is spectrally equivalent to $\{A_n(a,0)\}_n$

so that the sequence $\{P_n(a)\}_n$ turns out to be a superlinear preconditioning sequence for $\{A_n(a,0)\}_n$. Since the whole convergence is driven by the matrix in (5) and since $A_n(a,0)$ is a $O(h^2)$ spectral approximation of $\mathrm{Re}(A_n)$, it follows that the PHSS method converges in a very fast way within a constant number of iterations independent of $n$ under the mild assumption that $\nabla p(x)$ is not too large in norm. Therefore, in order to obtain the solution within a given accuracy, we essentially reduce the computation to a constant number of matrix vector multiplications of sparse/diagonal type and to a constant number of calls to a fast Poisson solver.

From a computational point of view it is worth stressing that, in the case of plurirectangular domain $\Omega$, the computation of the solution of the original linear system by the PCG method with preconditioner $P_n(a,\Omega)$ is reduced to the computation of the numerical solution of diagonal and $d$-level banded

(projected) Toeplitz linear systems with nonnegative generating functions. We recall that the resolution of such a type of linear systems can be performed within a linear arithmetic cost by means of fast Poisson solvers among them we count classical (direct) Poisson solvers mainly based on the cyclic reduction idea (see e.g. [9,14,37]) and several specialized multigrid algorithms (see e.g. [18,31]). Therefore, the use of fast Poisson solvers ($a = 1$) is enough for numerically solving nonconstant coefficient PDEs: we stress that the clustering properties that hold in the elliptic case are observed in the semi-elliptic setting as well even if there is a lack of an adequate theoretical analysis.

We wish to emphasize that, by using the properties regarding the Hermitian/skew-Hermitian splitting, our analysis can use the powerful spectral tools derived from the Toeplitz theory [7,28,29]. On the other hand, the analysis without these tools can be much more difficult, see, e.g., [4,5].

The paper is organized as follows: in Section 2 we illustrate how to include preconditioning strategies in the HSS method. We derive new convergence results for both the preconditioned and non preconditioned iterations. In particular, an interesting point is the "mixing up effect": it is proved that, under suitable assumptions of non-normality of $A_n$, the convergence rate is related to a certain average of eigenvalue moduli of the iteration matrix instead of the spectral radius. Therefore, in the case of clustering, the convergence acceleration is remarkable and we have a superlinear behavior. As a case study, in Section 3 we consider the Finite Differences discretization of PDEs of the form (1) and we show that previously developed preconditioning techniques for purely elliptic operators give rise to optimally convergent PHSS methods. Two final sections of numerical experiments and extensions (Section 4) and conclusions (Section 5) end the paper.

## 2 Preconditioned HSS iteration

First we introduce some notations and definitions. The symbol $\|\cdot\|$ denotes the spectral norm of a matrix that is the matrix norm induced by the Euclidean vector norm $\|\cdot\|$. If $X$ is an invertible matrix then the symbol $\|\cdot\|_X$ stands for the $X$ vector norm defined as $\|y\|_X = \|Xy\|$, $\forall y \in \mathbf{C}^n$. Therefore the corresponding induced matrix norm is defined by

$$\|A\|_X = \sup_{y \neq 0} \frac{\|Ay\|_X}{\|y\|_X}.$$

A simple check shows that

$$\|A\|_X = \|XAX^{-1}\|.$$

A matrix is called normal if $A$ and $A^H$ commute: we say that $A$ is *essentially normal* if it is similar to a normal matrix. As a consequence, each diagonalizable matrix is essentially normal and vice-versa while "most" of the diagonalizable matrices are not normal. The considered definition is of interest in the context of stationary iterative methods since the spectral radius of the iteration matrix exactly represents the convergence reduction factor of the iteration (in a given norm $\|\cdot\|_X$) if and only the iteration matrix is essentially normal. We are now ready for analyzing the preconditioned HSS (PHSS) method.

Let $P_n$ be a Hermitian positive definite matrix and let us consider the PHSS method, i.e., given a positive $\alpha$ and a initial guess $x^0$, do the following

$$(7)\quad \begin{cases} \left(\alpha I + P_n^{-1}\mathrm{Re}(A_n)\right) x^{k+\frac{1}{2}} = \left(\alpha I - P_n^{-1}\mathrm{i}\,\mathrm{Im}(A_n)\right) x^k + P_n^{-1} b \\ \left(\alpha I + P_n^{-1}\mathrm{i}\,\mathrm{Im}(A_n)\right) x^{k+1} = \left(\alpha I - P_n^{-1}\mathrm{Re}(A_n)\right) x^{k+\frac{1}{2}} + P_n^{-1} b \end{cases}$$

until convergence. A simple check shows that the iteration matrix is

$$M(\alpha) = \left(\alpha I + \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)\right)^{-1} \left(\alpha I - P_n^{-1}\mathrm{Re}(A_n)\right) \left(\alpha I + P_n^{-1}\mathrm{Re}(A_n)\right)^{-1} \left(\alpha I - \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)\right).$$

It is clear that the above iteration cannot be interpreted as the HSS method on the matrix $P_n^{-1}A$ simply because $P_n^{-1}\mathrm{Re}(A_n)$ and $P_n^{-1}\mathrm{Im}(A_n)$ are not the Hermitian/skew-Hermitian splitting of $P_n^{-1}A_n$. However, if $P_n = LL^H$, then the preceding claim is true for the symmetrized version since

$$\mathrm{Re}(L^{-1}A_nL^{-H}) = L^{-1}\mathrm{Re}(A_n)L^{-H}, \quad \mathrm{Im}(L^{-1}A_nL^{-H}) = L^{-1}\mathrm{Im}(A_n)L^{-H}$$

with $P_n^{-1}A_n$ similar to $L^{-1}A_nL^{-H}$. Furthermore, another viewpoint (the viewpoint of the implementation) is as follows: the method in (7) can be interpreted as the original iteration (3) where the identity matrix is replaced by the preconditioner $P_n$ i.e.

$$\begin{cases} \left(\alpha P_n + \mathrm{Re}(A_n)\right) x^{k+\frac{1}{2}} = \left(\alpha P_n - \mathrm{i}\,\mathrm{Im}(A_n)\right) x^k + b \\ \left(\alpha P_n + \mathrm{i}\,\mathrm{Im}(A_n)\right) x^{k+1} = \left(\alpha P_n - \mathrm{Re}(A_n)\right) x^{k+\frac{1}{2}} + b \end{cases}$$

With the help of these simple observations we can prove the unconditional convergence of the preconditioned HSS iteration.

**Theorem 2.1** *Let $A_n \in \mathbf{C}^{n\times n}$ be a positive matrix, $\alpha$ be a positive parameter and let $P_n \in \mathbf{C}^{n\times n}$ be a Hermitian positive definite matrix. Then the iteration matrix of the preconditioned HSS method is*

$$M(\alpha) = \left(\alpha I + \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)\right)^{-1} \left(\alpha I - P_n^{-1}\mathrm{Re}(A_n)\right) \left(\alpha I + P_n^{-1}\mathrm{Re}(A_n)\right)^{-1} \left(\alpha I - \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)\right),$$

*its spectral radius is bounded by*

$$\sigma(\alpha) = \max_{\lambda_i \in \lambda(P_n^{-1}\mathrm{Re}(A_n))} \left| \frac{\alpha - \lambda_i}{\alpha + \lambda_i} \right|$$

*where $\lambda(X)$ denotes the set of the eigenvalues of a square matrix $X$. Therefore, setting*

$$T = \left(\alpha I + P_n^{-1/2} \mathrm{i}\, \mathrm{Im}(A_n) P_n^{-1/2}\right) P_n^{1/2},$$

*it holds that*

$$\rho(M(\alpha)) \leq \|M(\alpha)\|_T \leq \sigma(\alpha) < 1, \quad \forall \alpha > 0,$$

*i.e., the preconditioned HSS iteration converges to the unique solution of the system $A_n x = b$. Moreover, setting $\lambda_{\min}$ and $\lambda_{\max}$ the extremal eigenvalues of $P_n^{-1}\mathrm{Re}(A_n)$ and denoting by $\kappa = \dfrac{\lambda_{\max}}{\lambda_{\min}}$ the spectral condition number (namely the Euclidean (spectral) condition number of the symmetrized matrix), the best $\alpha$, that minimizes the quantity $\sigma(\alpha)$, is $\alpha^* = \sqrt{\lambda_{\min}\lambda_{\max}}$ and*

$$\sigma(\alpha^*) = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}.$$

*Proof.* The expression of $M(\alpha)$ is known from (4). A simple algebraic manipulation proves that

$$M(\alpha) = T^{-1} \hat{M}(\alpha) T$$

with

$$T = P_n^{1/2} \left(\alpha I + P_n^{-1} \mathrm{i}\, \mathrm{Im}(A_n)\right) = \left(\alpha I + P_n^{-1/2} \mathrm{i}\, \mathrm{Im}(A_n) P_n^{-1/2}\right) P_n^{1/2},$$

$$\hat{M}(\alpha) = R(\alpha) U(\alpha)$$

and where

$$R(\alpha) = \left(\alpha I - P_n^{-1/2} \mathrm{Re}(A_n) P_n^{-1/2}\right) \left(\alpha I + P_n^{-1/2} \mathrm{Re}(A_n) P_n^{-1/2}\right)^{-1},$$

$$U(\alpha) = \left(\alpha I - \mathrm{i}\, P_n^{-1/2} \mathrm{Im}(A_n) P_n^{-1/2}\right) \left(\alpha I + \mathrm{i}\, P_n^{-1/2} \mathrm{Im}(A_n) P_n^{-1/2}\right)^{-1}.$$

Due to the relation $\|\cdot\|_T = \|T \cdot T^{-1}\|$, it follows that $\|M(\alpha)\|_T = \|\hat{M}(\alpha)\|$ and therefore, due to the spectrum invariance under similarity transformations, we have

$$\begin{aligned} \rho(M(\alpha)) &= \rho(\hat{M}(\alpha)) \\ &\leq \|\hat{M}(\alpha)\| \\ &= \|M(\alpha)\|_T. \end{aligned} \tag{8}$$

For evaluating $\|\hat{M}(\alpha)\|$, we now observe that

$$U(\alpha)$$

is a rational function of a skew-Hermitian matrix and therefore is normal. Moreover its eigenvalues are unitary by construction since

$$\frac{\alpha - x}{\alpha + x}$$

is unitary for real $\alpha$ and purely imaginary $x$. Therefore $U(\alpha)$ is a unitary matrix and, more specifically, it represents the Cayley transform of $P_n^{-1/2}\mathrm{Im}(A_n)P_n^{-1/2}$. In addition the matrix

$$R(\alpha)$$

is Hermitian since it is a rational function of a Hermitian matrix. Consequently, denoting by $\sim_{\mathrm{S}}$ the similarity relations among square matrices and calling $\lambda_i$ the eigenvalues of $P_n^{-1/2}\mathrm{Re}(A_n)P_n^{-1/2} \sim_{\mathrm{S}} P_n^{-1}\mathrm{Re}(A_n)$, we have

$$\begin{aligned} \|\hat{M}(\alpha)\| &\leq \|R(\alpha)\|\,\|U(\alpha)\| \\ &= \|R(\alpha)\| \\ &= \max_{i=1,\ldots,n} \left|\frac{\alpha - \lambda_i}{\alpha + \lambda_i}\right| \\ &= \sigma(\alpha). \end{aligned} \tag{9}$$

Due to the positivity of $\alpha$ and of the values $\lambda_i$, it follows that $\sigma(\alpha) < 1$. Following the same steps as in Corollary 2.3 in [2], it follows that the optimal parameter $\alpha = \alpha^*$ with

$$\alpha^* = \sqrt{\lambda_{\min}\lambda_{\max}}, \quad \lambda_{\min} = \min_i \lambda_i, \ \lambda_{\max} = \max_i \lambda_i$$

and therefore, setting $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ and putting together inequalities (8) and (9), we have

$$\rho(M(\alpha^*)) \leq \|M(\alpha^*)\|_T \leq \sigma(\alpha^*) = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}.$$

□

### *2.1 Further features of the PHSS method*

Some remarks can be useful for understanding important features of the method and to discover relationships with other iterative techniques.

*2.1.1 PHSS and PCG method* The upper bound for the convergence rate of the preconditioned HSS method, with the choice of the optimal parameter $\alpha = \alpha^*$, is the same as for the preconditioned conjugate gradient (PCG) method applied to a linear system whose coefficient matrix is $\mathrm{Re}(A_n)$ with preconditioner $P_n$. However, while the PCG method depends also on the distribution and clustering of the eigenvalues, this seems to be not the case for the preconditioned HSS iteration since it is a classical stationary method: on the other hand, in the following (Subsection 2.1.3) we will see that, under certain circumstances, the preconditioned HSS can be substantially faster when a spectral clustering occurs.

*2.1.2 PHSS method, transient phase and asymptotic convergence* A substantial drawback of measuring the convergence rate of stationary iterative methods by the spectral radius of the iteration matrix is that this is an asymptotic measure and therefore it can be useless when the number of iterations is small compared to $n$. In actuality, we are interested in optimal methods, i.e., iteration techniques converging to the solution, within a preassigned accuracy, in a number of step constant and independent of $n$. Therefore the asymptotic measure given by the spectral radius could be of little interest when the iteration matrix is highly non-normal (due to transient effects [17]). We should mention that this is not the case in our context: indeed the PHSS method is a multi-iterative technique since the iteration matrix $M(\alpha) = T^{-1}\hat{M}(\alpha)T$ is composed by two distinct matrices:

$$T^{-1}R(\alpha)T$$

and

$$T^{-1}U(\alpha)T,$$

with $R(\alpha)U(\alpha)$ being the polar decomposition [16] of $\hat{M}(\alpha)$. Hence the global iteration matrix is not normal but it is the product of two matrices that are similar, via the same transformation matrix $T$, to normal matrices where the first is a positive definite contraction and the second is a unitary matrix (the Cayley transform of $P_n^{-1/2}\mathrm{Im}(A_n)P_n^{-1/2}$). Consequently, the error in $T$ norm is preserved by $T^{-1}U(\alpha)T$ and is reduced by $T^{-1}R(\alpha)T$ without transient effects which are typical for essentially non-normal iteration matrices.

*2.1.3 PHSS method, non-normal matrices and multi-iterative methods* The above discussion in Subsection 2.1.2 gives the hint for a finer analysis. Indeed the bound for the $T$ norm of the error is exactly attained if and only if the skew-Hermitian part and the Hermitian part of $A_n$ have the same orthonormal basis of eigenvectors. This is true if and only if the original matrix $A_n$ is normal. Therefore the normal case is the *worst case*: for non-normal $A_n$,

in the light of the discussion in Subsection 2.1.2, it follows that the error is reduced by a factor that is smaller than $\sigma(\alpha)$ and that could be much smaller. The rest of the subsection is devoted to analyze this faster convergence of the PHSS method in the case of non-normal coefficient matrix $A_n$.

Indeed the matrix $R(\alpha)$ is Hermitian and therefore can be written as $Q_R D_R Q_R^H$ with $D_R = \mathrm{diag}_{1\le i\le n}(\tilde{\lambda}_i)$,

$$\tilde{\lambda}_i = \frac{\alpha - \lambda_i}{\alpha + \lambda_i}, \quad \alpha > 0, \ \lambda_i > 0,$$

as in Theorem 2.1, and $Q_R$ unitary. Therefore $\hat{M}(\alpha) = Q_R D_R V$ with $V = Q_R^H U(\alpha)$ being unitary. By writing $D_R = |D_R| S$ with $S$ being a sign diagonal matrix, it is clear that $\hat{M}(\alpha) = Q_R |D_R| W$ is "essentially" the singular value decomposition [16] of $\hat{M}(\alpha)$ since $W = SV$ is unitary.

As observed before, the worst case occurs in the "maximally concentrated case" where the eigenspaces of $R(\alpha)$ and $U(\alpha)$ coincide, i.e., $U(\alpha) = Q_R D_U Q_R^H$, $D_U = \mathrm{diag}_{1\le i\le n}(u_i)$ with $|u_i| = 1$: in that case $\hat{M}(\alpha) = Q_R D_R D_U Q_R^H$ and consequently the convergence rate is exactly determined by

$$\max_{1\le i\le n} |\tilde{\lambda}_i u_i| = \max_{1\le i\le n} |\tilde{\lambda}_i| = \sigma(\alpha).$$

On the other hand this result suggests that the "best case" occurs in the case of "maximal dispersion". Therefore, setting $U(\alpha) = Q_U D_U Q_U^H$ for a given unitary matrix $Q_U$, if the "maximally concentrated case" is represented by the condition

$$Q_{RU} = Q_R^H U(\alpha) Q_R^H = \text{ a phase matrix},$$

then the "maximally dispersed case" is represented by the condition

$$Q_{RU} = Q_R^H U(\alpha) Q_R^H = \text{ an equimodular matrix},$$

i.e.,

$$\left|(Q_{RU})_{i,j}\right| = \frac{1}{\sqrt{n}},$$

for every $i, j \in \{1, \ldots, n\}$.

A very classical example of equimodular matrix is the celebrated Fourier matrix. Quasi-equimodular matrices are the unitary matrices related to trigonometric transforms: we mention that the notion of "maximally dispersed matrices" is the key point in several contexts and indeed has been used for proving negative results on the preconditioning of multilevel structures by matrix algebras [36] and for solving some extremal problems in matrix theory.

In order to understand what happens in this situation, we suppose that the eigenvalues $\lambda_i$ of $\{P_n^{-1}\mathrm{Re}(A_n)\}_n$ are strongly clustered at 1 and $\alpha = 1$ (for the notion of strong or proper cluster refer to Definition 3.1): therefore, $\forall \epsilon > 0$, $\exists q > 0$ such that $|\tilde{\lambda}_i| < \epsilon$, $i = 1, \ldots, n-q$ and there exist $q$ outliers for which $|\tilde{\lambda}_i| \leq c < 1$, $i = n-q+1, \ldots, n$ with $c$ allowed to be very close to 1.

We want to show that in this case the contraction factor is really smaller than $c$ and indeed is close to $\epsilon$. First, observe that $[\hat{M}(\alpha)]^q$ coincides with

$$Q_R D_R Q_R^H U(\alpha) Q_R D_R Q_R^H U(\alpha) \cdots Q_R D_R Q_R^H U(\alpha) =$$
$$Q_R [D_R Q_{RU} D_R Q_{RU} \cdots Q_{RU} D_R] Q_R^H U(\alpha) = Q_R D_R [Q_{RU} D_R]^{q-1} Q_R^H U(\alpha).$$

Consequently the convergence behavior is determined by the contraction factor of the matrix $Q_{RU} D_R$ where $D_R$ is maximally concentrated and $Q_{RU}$ is maximally dispersed. In the light of the theory of multi-iterative methods [27], the strong complementarity of the two components makes the contraction factor of the product much smaller than the contraction factors of the two components $c$ and 1.

Let us substantiate this claim with some formal calculations. Let $e^k = \sum_{j=1}^n \alpha_j^k e_j$ be the error decomposition at step $k$ with $e_j$ being the $j$-th vector of the canonical basis. Then $e^{k+1} = Q_{RU} D_R e^k$ and consequently

$$\begin{aligned} e^{k+1} &= Q_{RU}[D_R e^k] \\ &= Q_{RU} \left[ \sum_{j=1}^{n} \alpha_j^k \tilde{\lambda}_j e_j \right] \\ &= Q_{RU} \left[ \sum_{j=1}^{n-q} \alpha_j^k \tilde{\lambda}_j e_j \right] + Q_{RU} \left[ \sum_{j=n-q+1}^{n} \alpha_j^k \tilde{\lambda}_j e_j \right] \\ &= (e[1])^{k+1} + (e[2])^{k+1}. \end{aligned}$$

It is clear that

$$\|(e[1])^{k+1}\| < \epsilon \|e^k\|,$$

while the study of the norm of $(e[2])^{k+1}$ requires a more accurate analysis. Calling $q_j$ the unitary equimodular columns of $Q_{RU}$, it follows that

$$(e[2])^{k+1} = \sum_{j=n-q+1}^{n} \alpha_j^k \tilde{\lambda}_j q_j$$

and consequently

$$\begin{aligned} e^{k+2} = Q_{RU} D_R e^{k+1} &= Q_{RU} D_R (e[1])^{k+1} + Q_{RU} D_R (e[2])^{k+1} \\ &= Q_{RU} D_R (e[1])^{k+1} + Q_{RU} \sum_{j=n-q+1}^{n} \alpha_j^k \tilde{\lambda}_j D_R q_j. \end{aligned}$$

Now, the "complementarity" which is typical of fast multi-iterative procedures comes into the play. In actuality, due to equimodularity, we observe that $D_R q_j$ has a Euclidean norm which is much smaller than $\|D_R\| = c$ and than $\|q_j\| = 1$ and, more specifically, the following inequality

$$\|D_R q_j\| \leq \epsilon + c\sqrt{\frac{q}{n}}$$

is satisfied. Consequently $\|\tilde{\lambda}_j D_R q_j\| \leq 2c\epsilon$ for $n$ large enough, and

$$\begin{aligned}
\|e^{k+2}\| &\leq c\|(e[1])^{k+1}\| + \|\sum_{j=n-q+1}^{n} \alpha_j^k \tilde{\lambda}_j D_R q_j\| \\
&< c\epsilon\|e^k\| + 2c\epsilon \sum_{j=n-q+1}^{n} |\alpha_j^k| \\
&\leq c\epsilon\|e^k\| + 2c\epsilon\sqrt{q}\left[\sum_{j=n-q+1}^{n} |\alpha_j^k|^2\right]^{1/2} \\
&\leq c\epsilon\|e^k\| + 2c\epsilon\sqrt{q}\|e^k\|
\end{aligned}$$

which is bounded by $(1 + 2\sqrt{q})c\epsilon\|e^k\|$ for $n$ large enough.

It is worth pointing out that the above bound implies that

$$\rho(M(\alpha)) = \rho(D_R Q_{RU}) < \sqrt{(1 + 2\sqrt{q})c\epsilon} \tag{10}$$

which is negligible with respect to

$$\rho(R(\alpha)) = \rho(D_R) = c$$

and to

$$\rho(U(\alpha)) = \rho(Q_{RU}) = 1.$$

Furthermore, for a generic square matrix $X$ we have

$$\rho(X) \geq \left(\prod_{\lambda_i \in \lambda(X)} |\lambda_i|\right)^{1/n} = \left(\prod_{\sigma_i \in \Sigma(X)} \sigma_i\right)^{1/n}$$

with $\Sigma(X)$ denoting the set of the singular values of $X$. In our case the singular values of $M(\alpha)$, $D_R Q_{RU}$ and $D_R$ coincide and, under the given assumption of clustered spectrum, we have

$$\left(\prod_{\sigma_i \in \Sigma(X)} \sigma_i\right)^{1/n} \leq \left(\frac{c^q}{\epsilon^q}\right)^{1/n} \cdot \epsilon.$$

Therefore, the bound given in (10) (which is not tight and can be still improved) is close to the square root of the geometric mean of the singular values: we will call this welcome averaging a "mixing up effect", since the matrix $D_R$ reduces the error according to the values $\tilde{\lambda}_i$ in the direction $e_j$ and the matrix $Q_{RU}$ makes a equimodular mix of all the contributions in each direction $e_i$; in this way the new matrix $D_R$ is ready to act on this mix with an overall acceleration. Finally, we point out that the "mixing up effect", in presence of a clustering, justifies a superlinear-like behavior of the considered PHSS iteration. The good news is that the quoted result stands also in the case of a weak clustering and this is somehow surprising since in the PCG case a weak clustering is not enough for a superlinear/optimal convergence.

*A numerical evidence* A numerical evidence of the "mixing up effect" emerges from the 4 Tables in Fig 5.2 of [2]: there the authors report a plot of $\sigma(\alpha)$ and of $\rho(M(\alpha))$ for various $\alpha$ in a neighborhood of the optimal value and with regard to problem (1) where $p = 1, 10, 100, 1000$. The larger is $p(x)$ the more the discretized matrix $A_n$ departs from normality and therefore we have a stronger "mixing up effect": a convincing explanation of this curious phenomenon is exactly the "mixing up effect". The nice thing is that for large $p(x) = P$, say 100, 1000, i.e., for a convection dominated problem, the quantity $\sigma(\alpha)$ is close to 1 but $\rho(M(\alpha)) \ll \sigma(\alpha) \approx 1$. Therefore the real convergence behavior of the HSS method is much faster compared with the forecasts of Theorem 2.1 with $P_n = I$. We point out that the results of this subsection have an interesting meaning since they show that the HSS and PHSS methods can be especially good for problems where most of the other techniques fail or become very slow.

*2.1.4 The case of $P_n = \mathrm{Re}(A_n)$: PHSS method and GMRES* In the case where $P_n = \mathrm{Re}(A_n)$ the optimal parameter $\alpha^*$ is 1 and the contraction factor $\sigma(\alpha^*)$ described in Theorem 2.1 is exactly zero. This means that we have exactly one iteration where we have to solve two kinds of auxiliary linear systems. The first type with coefficient matrix $P_n = \mathrm{Re}(A_n)$ and the second with coefficient matrix $I + \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)$. Therefore, we should have a fast solver for systems of the form $P_n y = c$ and that the matrix $I + \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)$ should have eigenvalues with good localizing properties. Since $\mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)$ is similar to a skew-Hermitian matrix, it follows that the eigenvalues are purely imaginary and therefore the minimum among the absolute value of the eigenvalues of $I + \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n)$ is 1. Thus, the linear system

$$(I + \mathrm{i}\, P_n^{-1}\mathrm{Im}(A_n))y = P_n^{-1}b \tag{11}$$

is easily solvable, say by GMRES or Chebyshev iterations, if

$$P_n^{-1}\mathrm{Im}(A_n) = [\mathrm{Re}(A_n)]^{-1}\,\mathrm{Im}(A_n)$$

has a bounded spectrum, i.e., there exists a positive constant independent of $n$ which bounds from above the modulus of the eigenvalues of $P_n^{-1}\mathrm{Im}(A_n)$. As we will see in the following section, the case of the FD/FEM discretization of PDEs of the form (1) leads to linear systems of equations for which we are able to find superlinear PCG methods for the Hermitian part $P_n = \mathrm{Re}(A_n)$ and for which we prove that the spectrum of

$$[\mathrm{Re}(A_n)]^{-1}\,\mathrm{Im}(A_n)$$

is clustered to 0 and is bounded by a universal constant not depending on $n$. The result is not trivial since all the matrices $A_n$, $\mathrm{Re}(A_n)$ and $\mathrm{Im}(A_n)$ show a condition number exploding to infinity as $n$ tends to infinity (see e.g. [1]).

However, in the case where the chosen preconditioner $P_n$ is $\mathrm{Re}(A_n)$ and the parameter $\alpha$ is equal to 1, it should be observed that the PHSS iteration is equivalent to the direct use of the GMRES method on

$$P_n^{-1}A_n x = P_n^{-1}b$$

since $P_n^{-1}A_n = I + \mathrm{i}\ P_n^{-1}\mathrm{Im}(A_n)$.

Finally we remark that, in the case of our model problem (1) and when the coefficient $a(x)$ is constant, we possess an optimal solver for linear systems whose coefficient matrix is $P_n = \mathrm{Re}(A_n)$. When $a(x)$ is nonconstant we have an approximate factorization of $\mathrm{Re}(A_n)$ in the sense discussed below equation (6): for more details see [30,34]. In that case we use this approximate factorization as preconditioner.

## 3 A model problem with $d = 2, 3$ and plurirectangle $\Omega$

We consider FD discretizations of differential problems of the form

$$(12)\qquad \begin{cases} -\nabla^T[a(x)\nabla u(x)] + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}(p(x)u(x)) = f(x) & x \in \Omega \\ \\ \text{Dirichlet BC} \end{cases}$$

with $\Omega$ being a plurirectangle of $\mathbf{R}^d$ with $d = 2, 3$, $a(x)$ being a uniformly positive function and $p(x)$ denoting the Reynolds function. The discretization process is performed in divergence form so that the resulting approximation of the operator $-\nabla^T[a(x)\nabla u(x)]$ is real symmetric positive definite. More precisely, the coefficient matrix is indicated as $A_n(a, p) = A_n(a, p, m)$ where $m = (m_1, m_2, m_3)$ and the parameter $m_j$, $j = 1, 2, 3$ identifies the precision order of the FD scheme used for approximating the operator $\dfrac{\partial}{\partial x_j}$.

When the problem is purely elliptic, i.e., the parameter $p(x)$ is equal to zero, some very fast preconditioning techniques based on Poisson solvers and

diagonal matrices were proposed. Both theoretical and practical comparisons proved that the new proposal is more effective than classical techniques such as matrix algebra preconditioning [10, 19–21] or incomplete LU factorization preconditioning [16, 1] even in presence of high–order FD formulae for the approximation of the quoted differential problems or in presence of semi-ellipticity. We just mention that, in the past few years, semi-elliptic problems have received increasing attention both from a numeric/modelistic and analytic point of view due to their occurrence in important applications: among them we recall electromagnetic field problems [23] and models in Mathematical Finance [38] where we encounter PDEs with a coefficient $a(x)$ either exploding or vanishing at the boundary of the domain.

Here we combine these ideas with the PHSS method and we show that the resulting method is optimal in the sense that the number of iteration can be bounded by a constant independent of $n$, with a total arithmetic cost for reaching the solution with a preassigned tolerance, which is asymptotically linear with respect to the size $n$ of the underlying matrices.

The rest of the section is organized in four steps: in the first step we report some necessary tools and definitions concerning Toeplitz matrices and spectral distribution and then, in the last three steps, we present an analysis of increasing difficulty with respect to the model problem defined in (12).

### *3.1 Some tools from Toeplitz matrices and matrix sequences*

Let $f$ be a $d$-variate Lebesgue integrable function defined over the hypercube $T^d$, with $T = (-\pi, \pi]$ and $d \geq 1$. From the Fourier coefficients of $f$

(13)
$$a_j = \frac{1}{(2\pi)^d} \int_{T^d} f(z) e^{-\mathrm{i}(j,z)}\, dz, \qquad \mathrm{i}^2 = -1, \quad j = (j_1, \ldots, j_d) \in \mathbf{Z}^d$$

with $(j, z) = \sum_{r=1}^{d} j_r z_r$, one can build the sequence of Toeplitz matrices $\{T_N(f)\}_N$, $N = (N_1, \ldots, N_d)$, where $T_N(f) \in \mathbf{C}^{n \times n}$ and $n = \prod_{r=1}^{d} N_r$. It is clear that the Fourier coefficients $a_j$ are equal to zero definitely (for $|j|$ large enough) if $f$ is a (multivariate) trigonometric polynomial and therefore the corresponding Toeplitz matrix is multilevel and banded like in the case of the classical $d$-level Laplacian discretized by minimal precision equispaced FD formulae over a square region. In the latter case, for instance, we mention that the corresponding generating function is the polynomial

$$\sum_{j=1}^{d} (2 - 2\cos(z_j)).$$

The matrix $T_N(f)$ is said to be the Toeplitz matrix of order $N$ generated by $f$ and can be conveniently written in terms of Jordan blocks and of their powers as follows

(14)
$$T_N(f) = \sum_{|j|\le N-e} a_j J_N^{[j]} = \sum_{|j_1|\le N_1-1} \cdots \sum_{|j_d|\le N_d-1} a_{(j_1,\ldots,j_d)} J_{N_1}^{[j_1]} \otimes \cdots \otimes J_{N_d}^{[j_d]}.$$

In the above relation, $\otimes$ denotes tensor Kronecker product, $J_m^{[l]}$ denotes the Jordan matrix of order $m$ whose $(s,t)$ entry equals 1 if $s-t=l$ and equals zero otherwise, while $J_N^{[j]}$, where $j$ and $N$ are multi-indices, is the tensor product of all $J_{N_r}^{[j_r]}$ for $r=1,\ldots,d$. More explicitly, the $2m-1$ matrices $J_m^{[l]}, l=0,\pm1,\ldots,\pm(m-1)$, are the canonical basis of the linear space of $m\times m$ (one-level) Toeplitz matrices and the tensor notation emphasizes the $d$-level Toeplitz structure of $T_N(f)$. Indeed, the set $\{J_N^{[j]}\}_j$ is the canonical basis of the linear space of the $n\times n$ $d$-level Toeplitz matrices.

The spectral properties of the sequence $\{T_N(f)\}_N$ and of related preconditioned sequences are completely understood and characterized in terms of the underlying generating functions. For instance, it is a immediate check to deduce that $T_N(f)$ is Hermitian for any $N$ if and only if $f$ is real valued. More sophisticate results are contained in the following theorem.

**Theorem 3.1** [7,28] *Let $f$ and $g$ two $d$ variate Lebesgue integrable real valued functions defined over $T^d$ and assume that $g$ is nonnegative with positive essential supremum. Then the following facts hold:*

1. *if $f$ is not identically constant, then every eigenvalue of $T_N(f)$ lies in $(m,M)$ where $m=$essinf $f$ and $M=$esssup $f$;*
2. *if we denote by $\lambda_{\min}(T_N(f))$ and by $\lambda_{\max}(T_N(f))$ the minimal and the maximal eigenvalues of $T_N(f)$, then*
$$\lim_{N\to\infty} \lambda_{\min}(T_N(f)) = m, \quad \lim_{N\to\infty} \lambda_{\max}(T_N(f)) = M$$
*with $N\to\infty$ meaning that $N_j\to\infty$, for every $j=1,\ldots,d$;*
3. *moreover, if $N_i\sim N_j$ for any $i$ and $j$, then $\lambda_{\min}(T_N(f))-m\sim n^{-\alpha/d}$ and $M-\lambda_{\max}(T_N(f))\sim n^{-\beta/d}$ where $\alpha$ is the maximum among the orders of the zeros of $f(z)-m$ and $\beta$ is the maximum among the orders of the zeros of $M-f(z)$;*
4. *finally $T_N(g)$ is Hermitian positive definite and the eigenvalues of $T_N^{-1}(g)T_N(f)$ are contained in $(r,R)$ if $r<R$ and $r=$essinf $h$, $R=$esssup $h$ with $h=\dfrac{f}{g}$.*

The following definition is also of interest in asymptotic (numerical) linear algebra.

**Definition 3.1** *Let $\{A_n\}_n$ be a sequence of matrices of increasing dimensions $n$ and let $\theta$ be a measurable function defined over a set $K$ of finite Lebesgue measure. We write that $\{A_n\}_n$ is distributed as the measurable function $\theta$ in the sense of the eigenvalues, i.e., $\{A_n\}_n \sim_\lambda \theta$ if, for every $F$ continuous, real valued and with bounded support, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} F\left(\lambda_j(A_n)\right) = \frac{1}{m\{K\}} \int_K F(\theta(s))\, ds, \tag{15}$$

*where $\lambda_j(A_n)$, $j = 1, \ldots, n$ are the eigenvalues of $A_n$.*
*The sequence $\{A_n\}_n$ is clustered at 1 if it is distributed as the constant function 1. Finally, the sequence is properly (or strongly) clustered at 1 if for any $\epsilon > 0$ the number of the eigenvalues of $A_n$ not belonging to $(1-\epsilon, 1+\epsilon)$ can be bounded by a pure constant eventually depending on $\epsilon$ but not on $n$.*

As an example, we recall that the sequence $\{T_N(f)\}_N$ is distributed as the symbol $f$ and furthermore, under the assumptions of part 4. of the above theorem, the preconditioned sequence $\{T_N^{-1}(g)T_N(f)\}_N$ is distributed (see [29]) as

$$h = \frac{f}{g}.$$

We just mention that this kind of global spectral results play a key role to prove precise asymptotic bounds on the convergence rate of (preconditioned) conjugate gradients like algorithms as shown in recent works by Beckermann and Kuijlaars [3].

### *3.2 The case of $a = 1$ and constant $p(x)$*

Let us consider the problem (12) with $d = 3$, $a = 1$ and constant $p(x)$, discretized by a seven-points FD formula where we use basic centered schemes of precision order two both for the diffusive terms and the convective terms. In the simple case where the domain is a cube $Q$ containing $\Omega$, following [2], we get a linear system with coefficient matrix

$$\begin{aligned} A_n(Q) = {} & T_N \otimes I \otimes I + I \otimes T_N \otimes I + I \otimes I \otimes T_N + S_N \otimes I \otimes I + \\ & I \otimes S_N \otimes I + I \otimes I \otimes S_N \end{aligned}$$

where the equispaced step-size $h = \frac{1}{N+1}$ is used in the discretization on all the three directions and the natural lexicographic ordering is employed in the unknowns. Moreover, $T_N$ is the Toeplitz matrix of size $N$ generated by $2 - 2\cos(z_1)$, i.e., the usual one dimensional discrete Laplacian, $S_N$ is $h$ times the Toeplitz matrix of size $N$ generated by $p\mathrm{i}\,\sin(z_1)$, and the global

dimension $n$ of the linear system is given by $N^3$. Therefore, the Hermitian part of $A_n(Q)$ is exactly the discretization of the diffusion terms, i.e.,

$$\mathrm{Re}(A_n(Q)) = T_N \otimes I \otimes I + I \otimes T_N \otimes I + I \otimes I \otimes T_N$$

and the skew-Hermitian part of $A_n(Q)$ is exactly the discretization of the convection terms, i.e.,

$$\mathrm{i}\,\mathrm{Im}(A_n(Q)) = S_N \otimes I \otimes I + I \otimes S_N \otimes I + I \otimes I \otimes S_N.$$

If we consider the same discretization scheme over the domain $\Omega$ then, following the results in [35], there exists a matrix $\Pi$ such that

$$A_n := A_n(\Omega) = \Pi A_n(Q) \Pi^T \tag{16}$$

and therefore

$$\mathrm{Re}(A_n) = \Pi\, \mathrm{Re}(A_n(Q))\, \Pi^T, \quad \mathrm{Im}(A_n) = \Pi\, \mathrm{Im}(A_n(Q))\, \Pi^T.$$

Here the matrix $\Pi$ has unitary rows and is obtained from the identity by deleting all the rows of index $j$ such that the $j$-th grid point of $Q$ does not belong to $\Omega$: it is evident that $\Pi\Pi^T = I$ while $\Pi^T\Pi$ is a orthogonal projector.

Notice that $\mathrm{Re}(A_n)$ is real symmetric positive definite but ill-conditioned with a condition number asymptotic to $h^{-2}$ [10]. We need preconditioning and, in this case, due to the existence of fast Poisson solvers, we use $\mathrm{Re}(A_n)$ as preconditioner. With the choice $\alpha^* = 1$, as observed in Subsection 2.1.4, we know that the PHSS method converges in one step and the main point is the solution of a system of the form

$$\left(I + \mathrm{i}\ [\mathrm{Re}(A_n)]^{-1}\, \mathrm{Im}(A_n)\right) y = c, \tag{17}$$

where $c$ is a $n$ sized vector. The key point is that the spectrum of $[\mathrm{Re}(A_n)]^{-1}$ $\mathrm{Im}(A_n)$ is real and, more important, is bounded by a fixed constant independent of $n$. Therefore the above system (17) could be easily solved by a elementary Richardson technique in a optimal way with a linear arithmetic cost. Furthermore we will also prove the spectrum of $[\mathrm{Re}(A_n)]^{-1}\, \mathrm{Im}(A_n)$ is clustered at zero and consequently the application of a method like GMRES or Chebyshev iterations would lead to superlinear convergence behavior.

**Theorem 3.2** *Let $A_n := A_n(Q), A_n(\Omega) \in \mathbf{C}^{n\times n}$ be the positive matrices defined in Subsection 3.2. Then*

$$R_n = [\mathrm{Re}(A_n)]^{-1}\, \mathrm{Im}(A_n)$$

*is spectrally bounded and properly clustered at zero.*

*Proof.* For the sake of notational simplicity, in all the subsequent steps we assume $p = 1$. Therefore all the estimates on the extreme eigenvalues have to be multiplied by $p(x) = P$.
**Step1**: we consider $A_n(Q)$ in one dimension.
*A) Spectral boundedness*. In this basic context we have

$$\mathrm{Re}(A_n) = T_N(2 - 2\cos(z_1)), \quad \mathrm{Im}(A_n)) = hT_N(\sin(z_1)).$$

Since $T_N(2 - 2\cos(z_1))$ is symmetric positive definite, calling $\lambda_{\min}$ and $\lambda_{\max}$ the extreme eigenvalues of $R_n$, it follows that

$$\lambda_{\min} = \min_{v\neq 0} \frac{v^H hT_N(\sin(z_1))v}{v^H T_N(2 - 2\cos(z_1))v}, \quad \lambda_{\max} = \max_{v\neq 0} \frac{v^H hT_N(\sin(z_1))v}{v^H T_N(2 - 2\cos(z_1))v}.$$

Due to the monotonicity of the Toeplitz operator, it follows that

$$\lambda_{\min} \geq \min_{v\neq 0} \frac{v^H hT_N(-|\sin(z_1)|)v}{v^H T_N(2 - 2\cos(z_1))v},$$
$$\lambda_{\max} \leq \max_{v\neq 0} \frac{v^H hT_N(|\sin(z_1)|)v}{v^H T_N(2 - 2\cos(z_1))v},$$

and hence, using the linearity of the Toeplitz operator, we have

$$\max\{|\lambda_{\min}|, \lambda_{\max}\} \leq \max_{v\neq 0} \frac{v^H hT_N(|\sin(z_1)|)v}{v^H T_N(2 - 2\cos(z_1))v}.$$

The function $h|\sin(z_1)|$ can be bounded from above by

$$\sin^2(z_1) + h^2 \mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\leq h\}}(z_1)$$

where $\mathrm{Ch}_X$ denotes the characteristic function of a set $X$. Moreover the Fourier coefficients of the function $h^2\mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\leq h\}}(z_1)$ are bounded by $\frac{h^2}{2\pi} m\{t \in [-\pi, \pi] :\ |\sin(t)| \leq h\}$ and the latter quantity, for any $\epsilon > 0$, is bounded by

$$\frac{h^3(1+\epsilon)}{\pi}, \quad \text{for } n \text{ large enough.}$$

Therefore, the spectral norm of the corresponding Toeplitz matrix is bounded by

$$\frac{h^2(1+\epsilon)}{\pi}, \quad \text{for } n \text{ large enough.}$$

By using once again the linearity and the monotonicity of the Toeplitz operator and by invoking part 4. of Theorem 3.1, we have

$$\begin{aligned}
\max\{|\lambda_{\min}|, \lambda_{\max}\} &\le \max_{v\neq 0} \frac{v^H h T_N(|\sin(z_1)|)v}{v^H T_N(2-2\cos(z_1))v} \\
&\le \max_{v\neq 0} \frac{v^H T_N(\sin^2(z_1) + h^2 \mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\le h\}}(z_1))v}{v^H T_N(2-2\cos(z_1))v} \\
&= \max_{v\neq 0} \frac{v^H T_N(\sin^2(z_1))v + v^H T_N(h^2 \mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\le h\}}(z_1))v}{v^H T_N(2-2\cos(z_1))v} \\
&\le \max_{v\neq 0} \frac{v^H T_N(\sin^2(z_1))v}{v^H T_N(2-2\cos(z_1))v} \\
&\quad + \max_{v\neq 0} \frac{v^H T_N(h^2 \mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\le h\}}(z_1))v}{v^H T_N(2-2\cos(z_1))v} \\
&\le \lambda_{\max}\left(T_N^{-1}(2-2\cos(z_1))T_N(\sin^2(z_1))\right) \\
&\quad + \frac{\max_{v\neq 0} v^H T_N(h^2 \mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\le h\}}(z_1))v}{\min_{v\neq 0} v^H T_N(2-2\cos(z_1))v} \\
&< \max_{z_1\in[-\pi,\pi]} \frac{\sin^2(z_1)}{2-2\cos(z_1)} \quad \text{part 4. of Theorem 3.1} \\
&\quad + \frac{\|T_N(h^2 \mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\le h\}}(z_1))\|}{4\sin^2\left(\frac{\pi}{2(n+1)}\right)} \\
&< \max_{z_1\in[-\pi,\pi]} \frac{\sin^2(z_1)}{2-2\cos(z_1)} + \frac{h^2(1+\epsilon)}{4\pi\sin^2\left(\frac{\pi}{2(n+1)}\right)},
\end{aligned}$$

for $n$ large enough. Finally, setting

$$C_1 = \max_{z_1\in[-\pi,\pi]} \frac{\sin^2(z_1)}{2-2\cos(z_1)}$$

and

$$C_2 = \sup_n \frac{h^2(1+\epsilon)}{4\pi\sin^2\left(\frac{\pi}{2(n+1)}\right)},$$

it follows that

$$\max\{|\lambda_{\min}|, \lambda_{\max}\} < C := C_1 + C_2 \tag{18}$$

and the proof is over.

*B) Proper clustering*. We want to show that, for every $\epsilon > 0$ there exists a constant $q = q_\epsilon$ independent of $h$ such that all the eigenvalues of $R_n$ belong to $(1-\epsilon, 1+\epsilon)$ except, at most, $q$ outliers. As in the part A) of the proof,

the main idea is a delicate majorization of the function $h|\sin(z_1)|$. Set $\alpha(h)$ an infinitesimal function and consider the following inequality

$$h|\sin(z_1)| \le h|\sin(z_1)|\mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|>\alpha(h)\}}(z_1) + h\alpha(h)\mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\le\alpha(h)\}}(z_1).$$

We observe that

$$\sup_{z_1\in[-\pi,\pi]} \frac{h|\sin(z_1)|\mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|>\alpha(h)\}}(z_1)}{2-2\cos(z_1)} \sim \frac{h}{\alpha(h)}$$

and therefore, by choosing $\alpha(h) = h/\epsilon$, we have

$$h|\sin(z_1)| \le C\epsilon(2-2\cos(z_1)) + \frac{h^2}{\epsilon}\mathrm{Ch}_{\{t\in[-\pi,\pi]:\ |\sin(t)|\le h/\epsilon\}}(z_1)$$

with $C$ absolute constant. Passing to the Toeplitz representation and using the Hermitian partial ordering notation, we have

$$[\mathrm{Re}(A_n)]^{-1/2}\,\mathrm{Im}(A_n)\,[\mathrm{Re}(A_n)]^{-1/2} \le C\epsilon I + \frac{h^2}{\epsilon}T_N^{-1}(2-2\cos(z_1))$$

and therefore the number of the eigenvalues exceeding $(C+1)\epsilon$ are bounded by the number of eigenvalues of $T_N(2-2\cos(z_1))$ which are smaller than $(h/\epsilon)^2$. A direct check on the eigenvalues of $T_N(2-2\cos(z_1))$, whose expression coincides with the quantities $4\sin^2\left(\frac{j\pi h}{2}\right)$, $j=1,\dots,N$, shows that this number can be bounded by an absolute constant times $\epsilon^{-1}$ since

$$4\sin^2\left(\frac{j\pi h}{2}\right) \le 4\left(\frac{j\pi h}{2}\right)^2 < \left(\frac{h}{\epsilon}\right)^2$$

is satisfied for $j < \pi^{-1}\epsilon^{-1}$. A similar result applies to the eigenvalues less than $-\epsilon$ and therefore the proof of the existence of a proper cluster is concluded with a constant $q$ asymptotic to $\epsilon^{-1}$ but independent of $n$ and $h$.

**Step2**: we consider $A_n(\Omega)$ in one dimension.

Since $A_n(\Omega) = \Pi A_n(Q)\Pi^T$, it follows

$$\mathrm{Re}(A_n(\Omega)) = \Pi\,\mathrm{Re}(A_n(Q))\,\Pi^T, \quad \mathrm{Im}(A_n(\Omega)) = \Pi\,\mathrm{Im}(A_n(Q))\,\Pi^T,$$

with $\Pi$ having unitary rows and more columns than rows. Therefore we deduce that

$$\begin{aligned}
\lambda_{\min}(\Omega) &= \min_{v\neq 0}\frac{v^H\mathrm{Im}(A_n(\Omega))v}{v^H\mathrm{Re}(A_n(\Omega))v}\\
&= \min_{v\neq 0}\frac{v^H\Pi\,\mathrm{Im}(A_n(Q))\,\Pi^T v}{v^H\Pi\,\mathrm{Re}(A_n(Q))\,\Pi^T v}\\
&= \min_{v\neq 0,\, w=\Pi^T v}\frac{w^H\mathrm{Im}(A_n(Q))w}{w^H\mathrm{Re}(A_n(Q))\,w}\\
&\ge \min_{w\neq 0}\frac{w^H\mathrm{Im}(A_n(Q))w}{w^H\mathrm{Re}(A_n(Q))\,w} = \lambda_{\min}(Q)
\end{aligned}$$

and

$$\begin{aligned}\lambda_{\max}(\Omega) &= \max_{v\neq 0} \frac{v^H \mathrm{Im}(A_n(\Omega))v}{v^H \mathrm{Re}(A_n(\Omega))v} \\ &= \max_{v\neq 0} \frac{v^H \Pi\ \mathrm{Im}(A_n(Q))\ \Pi^T v}{v^H \Pi\ \mathrm{Re}(A_n(Q))\ \Pi^T v} \\ &= \max_{v\neq 0, w=\Pi^T v} \frac{w^H \mathrm{Im}(A_n(Q))w}{w^H \mathrm{Re}(A_n(Q))\ w} \\ &\leq \max_{w\neq 0} \frac{w^H \mathrm{Im}(A_n(Q))w}{w^H \mathrm{Re}(A_n(Q))\ w} = \lambda_{\max}(Q).\end{aligned}$$

Consequently the claim is reduced to the one in **Step1** since

$$\max\{|\lambda_{\min}(\Omega)|, \lambda_{\max}(\Omega)\} \leq \max\{|\lambda_{\min}(Q)|, \lambda_{\max}(Q)\} < C$$

with $C$ defined as in (18).

Finally, the use of relation (16) and the application of the Cauchy interlace principle are sufficient to prove the spectral clustering in this case as well.
**Step3**: we consider $A_n(Q)$ in $d > 1$ dimensions.
Make the same reasoning as in **Step1**: the proof is identical with the generating functions $2 - 2\cos(z_1)$ and $\sin(z_1)$ replaced by their $d$-level counterparts $\sum_{j=1}^{d}(2 - 2\cos(z_j))$ and $\sum_{j=1}^{d}\sin(z_j)$.
**Step4**: we consider $A_n(\Omega)$ in $d > 1$ dimensions.
Make the same reasoning as in **Step2** and reduce the claim to the one in **Step3**. □

It could be observed that the matrix $\mathrm{Re}(A_n(\Omega))$ is diagonalized by the DST I transform (the most known sine transform, see e.g. [31]) and this remark could lead to more information and to some simplification in the above analysis. However, $\mathrm{Im}(A_n(\Omega))$ is not diagonalized by the same transform. Moreover if we consider higher order FD discretization formulae, then the analysis performed in the proof of Theorem 3.2 is still valid (indeed Theorem 3.2 can be stated and proved identically), while the involved matrices are no longer diagonalized by any known (fast) transform.

### *3.3 The case of nonconstant $a(x)$ and $p(x)$ constant*

Consider he problem (12) with $d = 3$, nonconstant $a = a(x) > 0$ and constant $p(x)$. We use again the same seven-points FD formula based on centered schemes of precision order two both for the diffusive terms and the convective terms. In the simple case where the domain is a cube $Q$ containing $\Omega$, we obtain a linear system with coefficient matrix

$$A_n(a, Q) = \Theta_n(a, Q) + S_N \otimes I \otimes I + I \otimes S_N \otimes I + I \otimes I \otimes S_N$$

where the same constant step-size $h = \frac{1}{N+1}$ is used in the discretization on all the three directions and the natural lexicographic ordering is employed in the unknowns. The matrix $\Theta_n(a, Q)$ is the discretization of the diffusion term and the "central" entries of its generic row are

$$\left(-a_{i,j,\tilde{k}}, \mathbf{0}, -a_{i,\tilde{j},k}, \mathbf{0}, -a_{\tilde{i},j,k}, A_{i,j,k}, -a_{\tilde{i}+1,j,k}, \mathbf{0}, -a_{i,\tilde{j}+1,k}, \mathbf{0}, -a_{i,j,\tilde{k}+1}\right),$$

with $A_{i,j,k}$ equal to the sum of the absolute values of the non-diagonal entries of the related row, the "internal" $\mathbf{0}$ being a null vector of size $N-2$, the "external" $\mathbf{0}$ being a null vector of size $(N-2)N+N-1$, and

$$a_{s,t,u} = a\,(sh, th, uh)\,, \quad \tilde{r} = r - 1/2, \ r = 1, \ldots, N.$$

In the case where $a = 1$, we observe that the matrix $\Theta_n(a, Q)$ coincides with the usual three dimensional discrete Laplacian described by the three-level Toeplitz structure

$$\Theta_n(1, Q) = T_N \otimes I \otimes I + I \otimes T_N \otimes I + I \otimes I \otimes T_N.$$

As in the previous subsection, $S_N$ is $h$ times the Toeplitz matrix of size $N$ generated by $p\mathrm{i}\ \sin(z_1)$, and the global dimension $n$ of the linear system is given by $N^3$. Therefore, the Hermitian part of $A_n(Q)$ is exactly the discretization of the diffusion terms, i.e.,

$$\mathrm{Re}(A_n(a, Q)) = \Theta_n(a, Q)$$

and the skew-Hermitian part of $A_n(a, Q)$ is exactly the discretization of the convection terms, i.e.,

$$\mathrm{i}\,\mathrm{Im}(A_n(a, Q)) = S_N \otimes I \otimes I + I \otimes S_N \otimes I + I \otimes I \otimes S_N.$$

If we consider the same discretization scheme over the domain $\Omega$ then, following the results in [35], there exists a matrix $\Pi$ depending only on $\Omega$ and $Q$ (the same as in Subsection 3.2) such that

$$A_n := A_n(a, \Omega) = \Pi A_n(a, Q)\Pi^T$$

and

$$\mathrm{Re}(A_n) = \Pi\ \mathrm{Re}(A_n(a, Q))\ \Pi^T, \quad \mathrm{Im}(A_n) = \Pi\ \mathrm{Im}(A_n(a, Q))\ \Pi^T.$$

Similarly to the case in Section 3.2, $\mathrm{Re}(A_n)$ is real symmetric positive definite but ill-conditioned with a condition number asymptotic to $h^{-2}$. We need preconditioning and indeed we use an approximation strategy analyzed in [30,34]. More specifically, we consider a Poisson solver based preconditioner for $\mathrm{Re}(A_n(a, \Omega))$ defined as

$$P_n := P_n(a, \Omega) = D_n^{1/2}(a)\Theta_n(1, \Omega)D_n^{1/2}(a),$$

where $D_n(a) = \frac{1}{6}\mathrm{diag}\,(\Theta_n(a,\Omega))$ is a suitably scaled diagonal of $\Theta_n(a,\Omega)$. The features of this preconditioning sequence have been analyzed in detail. Here we report the useful properties.

**Theorem 3.3** [30,35] *Let $A_n := A_n(a,Q)$, $A_n(a,\Omega) \in \mathbf{C}^{n\times n}$ be the positive matrices previously defined. If the coefficient $a(x)$ is strictly positive and belongs to $\mathbf{C}^2(\overline{\Omega})$, then for every $\epsilon > 0$, there exist a $\bar{N}$ and a constant $q$ such that for any $N = (N_1,\ldots,N_d) > \bar{N} = (\bar{N}_1,\ldots,\bar{N}_d)$ (with respect to the partial ordering of $\mathbf{N}^d$), $n-q$ eigenvalues of the preconditioned matrix $P_n^{-1}\mathrm{Re}(A_n)$ belong to the open interval $(1-\epsilon, 1+\epsilon)$* [Proper Clustering]. *Moreover all the eigenvalues belong to an interval $[c,C]$ well separated from zero* [Spectral equivalence]

With the choice $\alpha^* = 1$, as observed in Subsection 2.1.3, we know that the PHSS method converges "superlinearly" if the skew-Hermitian part is big enough (substantial departure from normality). Otherwise, due to the spectral equivalence, it will be optimally convergent (linearly but with a convergence rate independent of the mesh size $h$).

Now the critical point is the solution of a system of the form (17):

$$\left(I + \mathrm{i}\, P_n^{-1}\,\mathrm{Im}(A_n)\right) y = c. \tag{19}$$

The key statement is that the spectrum of $P_n^{-1}\mathrm{Im}(A_n)$ is real and, more important, is bounded by a fixed constant independent of $n$. Therefore the above system (19) could be easily solved by a elementary Richardson technique in a optimal way with a linear arithmetic cost. Furthermore, the spectrum of $P_n^{-1}\mathrm{Im}(A_n)$ is clustered at zero and consequently the application of a method like GMRES or Chebyshev iterations would lead to superlinear convergence behavior.

**Theorem 3.4** *Let $A_n := A_n(a,Q)$, $A_n(a,\Omega) \in \mathbf{C}^{n\times n}$ be the positive matrices previously defined with $a(x)$ being strictly positive and belonging to $\mathbf{C}^2(\overline{\Omega})$. Then*

$$R_n = P_n^{-1}\mathrm{Im}(A_n)$$

*is spectrally bounded and properly clustered at zero.*

*Proof.* Due to the positivity and to the regularity of $a(x)$, by Theorem 3.3, we have

$$c\,P_n \le \mathrm{Re}(A_n) \le C\,P_n, \quad A_n = A_n(a,\Omega),$$

where the positive interval $[c,C]$ is exactly the one of Theorem 3.3 and where the ordering relation is the one of the Hermitian matrices. Moreover the operator $\mathrm{Re}(A_n(\cdot,\Omega))$ is linear and positive (see [35]) and therefore

$$\min_{\overline{\Omega}} a(x)\ \mathrm{Re}(A_n(1,\Omega)) \le \mathrm{Re}(A_n(a,\Omega)) \le \max_{\overline{\Omega}} a(x)\ \mathrm{Re}(A_n(1,\Omega))$$

with strictly positive constants $\min_{\overline{\Omega}} a(x)$ and $\max_{\overline{\Omega}} a(x)$. Combining the two sets of matrix inequalities, we conclude that

$$\frac{c}{\max_{\overline{\Omega}} a(x)} Z_n \leq P_n^{-1/2} \mathrm{Im}(A_n) P_n^{-1/2} \leq \frac{C}{\min_{\overline{\Omega}} a(x)} Z_n, \tag{20}$$

with

$$Z_n = [A_n(1, \Omega)]^{-1/2} \, \mathrm{Im}(A_n) \, [A_n(1, \Omega)]^{-1/2} .$$

Finally we observe that $R_n$ is similar to $P_n^{-1/2} \mathrm{Im}(A_n) P_n^{-1/2}$ and so they share the spectrum; moreover the matrix sequence

$$\left\{ Z_n = [A_n(1, \Omega)]^{-1/2} \, \mathrm{Im}(A_n) \, [A_n(1, \Omega)]^{-1/2} \right\}_n$$

is spectrally bounded and properly clustered to zero (since $Z_n$ is similar to the matrix $R_n$ considered in Theorem 3.2): as a consequence of (20), the sequence $\{R_n\}_n$ considered in the present theorem is also spectrally bounded and properly clustered to zero and so the proof is over. □

*3.4 The general case*

We consider problem (12) with $d = 3$, nonconstant $a = a(x) > 0$, nonconstant $p(x)$ and we make use of the discretization process as in the former subsections. In the simplest case where the domain is a square $Q$ containing $\Omega$ we obtain a linear system with coefficient matrix

$$A_n(a, p, Q) = \Theta_n(a, Q) + \Psi_n(p, Q).$$

The matrix $\Theta_n(a, Q)$ is the discretization of the diffusion term and the matrix $\Psi_n(p, Q)$ is the discretization of the convection term. In the case where $p(x)$ is constant, we observe that the matrix $\Psi_n(p, Q)$ coincides with a three-level Toeplitz structure

$$S_N \otimes I \otimes I + I \otimes S_N \otimes I + I \otimes I \otimes S_N.$$

A crucial difference with respect to the preceding cases is that the Hermitian part of $A_n(a, p, Q)$ is not exactly the discretization of the diffusion terms $\Theta_n(a, Q)$ and the skew-Hermitian part of $A_n(a, p, Q)$ is not exactly the discretization of the convection terms $\Psi_n(p, Q)$.

**Theorem 3.5** *Let $A_n := A_n(a, p, Q)$, $A_n(a, p, \Omega) \in \mathbf{C}^{n \times n}$ be the positive matrices previously defined and let us assume that the coefficient $p(x)$ has bounded first derivative. Then*

$$\mathrm{Re}(A_n(a, p)) = \Theta_n(a) + E_n,$$
$$\mathrm{i} \, \mathrm{Im}(A_n(a, p)) = \Psi_n(p) - E_n$$

*where*

$$E_n = \frac{\Psi_n(p) + \Psi_n^H(p)}{2}$$

*with* $\|E_n\| \leq ch^2$ *and* $c = 6\|\nabla p\|_\infty$ *absolute constant only depending on* $p(x)$.

*Proof.* Since the discretization of the diffusion term $\Theta_n(a)$ is Hermitian (see Subsection 3.3), it follows that $\Theta_n(a)$ does not contribute to the skew-Hermitian part of $A_n$ and consequently

$$\begin{aligned} \mathrm{Re}(A_n(a,\,p)) &= \Theta_n(a) + E_n, \\ \mathrm{i}\,\mathrm{Im}(A_n(a,\,p)) &= \Psi_n(p) - E_n, \\ E_n &= \frac{\Psi_n(p) + \Psi_n^H(p)}{2}. \end{aligned}$$

To evaluate the spectral norm of $E_n$ we observe that the matrix $\Psi_n(p)$ has a symmetric pattern with 6 nonzero diagonals. Therefore $\Psi_n^H(p)$ and $E_n$ have an identical pattern and therefore

$$\|E_n\| \leq 6 \max_{(s,t)\in \mathrm{pattern}(\Psi_n(p))} \left|(E_n)_{s,t}\right|.$$

The "central" entries of the generic row of $\Psi_n(p)$ are given by

$$\frac{h}{2}\left(-p_{i,j,k-1}, \mathbf{0}, -p_{i,j-1,k}, \mathbf{0}, -p_{i-1,j,k}, 0, p_{i+1,j,k}, \mathbf{0}, p_{i,j+1,k}, \mathbf{0}, p_{i,j,k+1}\right),$$

with the "internal" $\mathbf{0}$ being a null vector of size $N-2$ and the "external" $\mathbf{0}$ being a null vector of size $(N-2)N+N-1$. As a consequence, we directly infer that the "central" entries of the generic row of the correction matrix $E_n$ are defined as

$$\begin{aligned} &\tfrac{h}{2}\left(-p_{i,j,k-1} + p_{i,j,k}, \mathbf{0}, -p_{i,j-1,k} + p_{i,j,k}, \mathbf{0}, -p_{i-1,j,k} + p_{i,j,k}, 0,\right. \\ &\qquad \left. p_{i+1,j,k} - p_{i,j,k}, \mathbf{0}, p_{i,j+1,k} - p_{i,j,k}, \mathbf{0}, p_{i,j,k+1} - p_{i,j,k}\right), \end{aligned}$$

and finally

$$\left|(E_n)_{s,t}\right| \leq h^2 \max_{1\leq v\leq 3} \left\| \frac{\partial}{\partial x_v} p \right\|_\infty = h^2 \|\nabla p\|_\infty.$$

□

*Remark 3.1* If we consider the same discretization scheme over the domain $\Omega$ then we have

$$A_n := A_n(a,\,p,\,\Omega) = \Pi A_n(a,\,p,\,Q)\Pi^T$$

and

$$\mathrm{Re}(A_n) = \Pi\ \mathrm{Re}(A_n(a,\,p,\,Q))\ \Pi^T, \quad \mathrm{Im}(A_n) = \Pi\ \mathrm{Im}(A_n(a,\,p,\,Q))\ \Pi^T$$

with $\Pi$ defined as in the preceding subsections.

Under the assumption that $\|\nabla p\|_\infty$ is smaller than a positive suitable constant, we prove that $\mathrm{Re}(A_n)$ is real symmetric positive definite but ill-conditioned with a condition number asymptotic to $h^{-2}$. We need preconditioning and indeed we use an approximation strategy analyzed in [30,34]. More specifically, we consider a Poisson solver based preconditioner for $\mathrm{Re}(A_n(a,\Omega))$ defined as

$$P_n := P_n(a,\Omega) = D_n^{1/2}(a)\Theta_n(1,\Omega)D_n^{1/2}(a),$$

where $D_n(a) = \frac{1}{6}\mathrm{diag}\,(\Theta_n(a,\Omega))$ is a suitably scaled diagonal of $\Theta_n(a,\Omega)$. The following result holds.

**Theorem 3.6** *Let $A_n := A_n(a,p,Q)$, $A_n(a,p,\Omega) \in \mathbf{C}^{n\times n}$ be the positive matrices previously defined. If $\|\nabla p\|_\infty \le \pi^2 c(\Omega)\min_{\overline{\Omega}} a$ with $c(\Omega) > 0$, $c(Q) = 1$, and if the coefficient $a(x)$ is strictly positive and belongs to $\mathbf{C}^2(\overline{\Omega})$, then for every $\epsilon > 0$, there exist a $\bar{N}$ and a constant $q$ such that for any $N = (N_1,\ldots,N_d) > \bar{N} = (\bar{N}_1,\ldots,\bar{N}_d)$ (with respect to the partial ordering of $\mathbf{N}^d$), $n-q$ eigenvalues of the preconditioned matrix $P_n^{-1}\mathrm{Re}(A_n)$ belong to the open interval $(1-\epsilon, 1+\epsilon)$* [Proper Clustering]. *Moreover all the eigenvalues belong to an interval $[c,C]$ well separated from zero* [Spectral equivalence]

*Proof.* By Theorem 3.5 we have

$$\mathrm{Re}(A_n(a,p)) = \Theta_n(a) + E_n,\ \|E_n\| \le 6\|\nabla p\|_\infty h^2.$$

Moreover, from spectral results given in [35], we know that

$$\lambda_{\min}(\Theta_n(a)) \ge \left[c(\Omega)3\pi^2 \min_{\overline{\Omega}} a\right] h^2$$

with $c(\Omega) > 0$ and $c(Q) = 1$. Therefore, it is clear that $E_n$ does not change the positive definiteness and the asymptotic ill conditioning of $\Theta_n(a)$ if

$$\|\nabla p\|_\infty < \frac{\pi^2}{2} c(\Omega) \min_{\overline{\Omega}} a.$$

Concerning the spectral results on the sequence $\{P_n^{-1}\mathrm{Re}(A_n)\}_n$ we recall that the claimed thesis is true for $\{P_n^{-1}\Theta_n(a)\}_n$ by Theorem 3.3. Moreover $P_n$ has minimal eigenvalue going to zero as $h^2$, is spectrally distributed as $a(x)\sum_{j=1}^d (2-2\cos(z_j))$ (see e.g. [32]), and $\mathrm{Re}(A_n) - \Theta_n(a) = E_n$ with $\|E_n\| \le 6\|\nabla p\|_\infty h^2$. Therefore a simple reasoning allows one to conclude that the same relations (with possibly different constants) are satisfied for $\{P_n^{-1}\mathrm{Re}(A_n)\}_n$ as well. •

With the choice $\alpha^* = 1$, as observed in Subsection 2.1.4, we know that the PHSS method converges "superlinearly" if the skew-Hermitian part is big enough. Otherwise, due to the spectral equivalence, it will be optimally convergent (linearly but with a convergence rate independent of the mesh size $h$).

Again, the critical point is the solution of a system of the form (17):

$$\left(I + \mathrm{i}\, P_n^{-1}\, \mathrm{Im}(A_n)\right) y = c, \tag{21}$$

and, similarly to the previous case, the spectrum of $P_n^{-1}\mathrm{Im}(A_n)$ is real and, more important, is bounded by a fixed constant independent of $n$. Therefore the above system (21) could be easily solved by a elementary Richardson technique in an optimal way with a linear arithmetic cost. Moreover, the spectrum of $P_n^{-1}\mathrm{Im}(A_n)$ is clustered at zero in this case as well and, consequently, the application of a method like GMRES or Chebyshev would lead to superlinear convergence behavior.

**Theorem 3.7** *Let $A_n := A_n(a, p, Q)$, $A_n(a, p, \Omega) \in \mathbf{C}^{n\times n}$ be the positive matrices previously defined. Then*

$$R_n = P_n^{-1}\mathrm{Im}(A_n)$$

*is spectrally bounded and properly clustered at zero.*

*Proof.* From Theorem 3.5 we have

$$\begin{aligned} P_n^{-1}\mathrm{Im}(A_n) &= -\mathrm{i}\, P_n^{-1}\,(\Psi_n(p) - E_n) \\ &= -\frac{\mathrm{i}}{2} P_n^{-1}\left(\Psi_n(p) - \Psi_n^H(p)\right) \end{aligned}$$

where the "central" entries of the generic row of the matrix $X_n(p) = \Psi_n(p) - \Psi_n^H(p)$ are defined as

$$\begin{aligned} \tfrac{h}{2}\big(&-p_{i,j,k-1} - p_{i,j,k}, \mathbf{0}, -p_{i,j-1,k} - p_{i,j,k}, \mathbf{0}, -p_{i-1,j,k} - p_{i,j,k}, 0, \\ &p_{i+1,j,k} + p_{i,j,k}, \mathbf{0}, p_{i,j+1,k} + p_{i,j,k}, \mathbf{0}, p_{i,j,k+1} + p_{i,j,k}\big). \end{aligned}$$

Consequently the Hermitian matrix $-\frac{\mathrm{i}}{2}X_n(p)$ has the following dyadic representation

$$\begin{aligned} -\frac{\mathrm{i}}{2}X_n(p) = \frac{h}{4}\sum_{i,j,k} & p_{i,j,k}(e_k e_k^T)\otimes(e_j e_j^T)\otimes[T_3(-2\sin(z))]_{(i)} + \\ & (e_k e_k^T)\otimes[T_3(-2\sin(z))]_{(j)}\otimes(e_i e_i^T) + [T_3(-2\sin(z))]_{(k)}\otimes(e_j e_j^T)\otimes(e_i e_i^T), \end{aligned}$$

where the symbol $[T_3(g(z))]_{(s)}$ denotes a null matrix of size $N$ except for a 3 by 3 block coinciding with $T_3(g(z))$ and having the entry $(T_3(g(z)))_{2,2}$ (the center of $T_3(g(z))$) in position $(s, s)$, $s = 1, \ldots, N$. Of course if $s = 1$

or $s = N$ the considered nonzero block reduces to a 2 by 2 block since we simply ignore the extra-dimensional terms. Moreover $[T_3(-2\sin(z))]_{(j)} = [T_2(-2\sin(z))]_{(j-1)} + [T_2(-2\sin(z))]_{(j)}$ with $[T_2(g(z))]_{(s)}$ being a null matrix except for a 2 by 2 block coinciding with $T_2(g(z))$ and having the entry $(T_2(g(z)))_{1,1}$ in position $(s, s)$, $s = 0, \ldots, N$. If $s = 0$ or $s = N$ the considered nonzero block reduces to a single element (the extra-dimensional entries are disregarded again). Now we observe that

$$[T_2(2\sin(z))]_{(i)} \le w[T_2(1 + \hat{h}^2 - 2\cos(z))]_{(i)} \tag{22}$$

provided that $w \ge \frac{\hat{h}^{-1}}{\sqrt{2+\hat{h}^2}}$: indeed the only nonzero block of $[T_2(2\sin(z))]_{(i)}$ is

$$\begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix}$$

and the (corresponding) only nonzero block of $w[T_2(1+\hat{h}^2-2\cos(z))]_{(i)}$ is

$$w \begin{bmatrix} 1+\hat{h}^2 & -1 \\ -1 & 1+\hat{h}^2 \end{bmatrix}.$$

Consequently relation (22) is equivalent to check the nonnegative definiteness of the matrix

$$\begin{bmatrix} w(1+\hat{h}^2) & \mathrm{i} - w \\ -\mathrm{i} - w & w(1+\hat{h}^2) \end{bmatrix}$$

whose trace is given by $2w(1+\hat{h}^2)$ and whose determinant is given by $w^2(1+\hat{h}^2)^2 - 1 - w^2$. In conclusion relation (22) holds if and only if $2w(1+\hat{h}^2) \ge 0$,

$$w^2(1+\hat{h}^2)^2 - 1 - w^2 \ge 0$$

i.e., $w \ge \hat{h}^{-1}/\sqrt{2+\hat{h}^2}$. We notice that it is sufficient to choose

$$w = \hat{h}^{-1}.$$

As a consequence, by linearity and positivity, we deduce that

$$-h\|p\|_\infty \hat{h}^{-1} T_n \left( \sum_{j=1}^{3} (1 - \cos(z_j) + \hat{h}^2) \right) \le -\frac{\mathrm{i}}{2} X_n(p) \tag{23}$$

and

$$-\frac{\mathrm{i}}{2} X_n(p) \le h\|p\|_\infty \hat{h}^{-1} T_n \left( \sum_{j=1}^{3} (1 - \cos(z_j) + \hat{h}^2) \right). \tag{24}$$

Once we have the key relations (23)–(24), we have reduced the spectral analysis of $\mathrm{Im}(A_n)$ to a Toeplitz problem. By using Theorem 3.4, we also reduce $P_n$ to a Toeplitz structure. Therefore both the uniform spectral boundedness and the proper spectral clustering are obtained by exploiting the standard matrix technology used in **Step1** of Theorem 3.2 for the constant coefficient (i.e. Toeplitz) case with the choice of $\hat{h} = h\epsilon^{-1}$ with fixed and small enough $\epsilon > 0$. □

*Remark 3.2* We observe that Theorem 3.5 holds unchanged (i.e. with the same constants!) if the convection term in (12) is replaced by $p \sum_{j=1}^{d} \frac{\partial}{\partial x_j} u(x)$; moreover a similar result can be stated if the convection term is more general and takes the form $p^T(x) \cdot \nabla u$ where

$$p(x) = \begin{pmatrix} p_1(x) \\ p_2(x) \\ p_3(x) \end{pmatrix} \quad \text{and} \quad \nabla u = \begin{pmatrix} \frac{\partial}{\partial x_1} u \\ \frac{\partial}{\partial x_2} u \\ \frac{\partial}{\partial x_3} u \end{pmatrix}.$$

Here the term $\|\nabla p\|_\infty$ in Theorem 3.5 has to be replaced by $\max_{1 \le v \le 3} \|\nabla p_v\|_\infty$. Therefore the results stated in Theorem 3.6 and Theorem 3.7 can be easily adapted to these cases as well.

*Higher order FD discretizations and FEM methods* Concerning the case of high order Finite Differences discretizations and Finite Elements approximation, we recall that in [30,34,35,33] we derived asymptotic expansions concerning the preconditioned matrices $P_n^{-1} A_n$ in terms of related Toeplitz structures. Moreover, it was proved that the sequence $\{P_n^{-1} A_n\}_n$ is clustered at unity and is spectrally bounded if $a(x)$ is regular enough and positive. Therefore most of the results proved in Section 3 can be extended with little effort to these cases so covering several approximation schemes for PDEs.

## 4 Numerical experiments

The section is divided into two main parts: in the first one we give a general description of the numerical experiments, some implementation details and few remarks on the computational costs; the second part is devoted to the comments on the obtained numerical results.

### *4.1 General comments and implementation details*

Following [2], we must observe that in principle each iteration of the HSS method requires the exact solutions with large matrices $\alpha I + \mathrm{Re}(A_n)$ and $\alpha I + \mathrm{i}\,\mathrm{Im}(A_n)$ which can be impractical in actual implementations. To further improve computing efficiency of the HSS method, it is possible to use the

conjugate gradient (CG) method for solving linear systems with coefficient matrix $\alpha I + \mathrm{Re}(A_n)$ and some Krylov subspace methods for linear systems with matrix $\alpha I + \mathrm{i}\,\mathrm{Im}(A_n)$. This idea defines an *inexact HSS iteration* which is denoted in short by *IHSS iteration*. The tolerances (or number of inner iteration steps) for the inner iterative procedures may vary and are changed according to the outer iteration scheme: as shown in [2], a proper choice of these parameters can be done following a rigorous analysis in such a way that the resulting method is accurate and each (inexact) outer iteration is cheap. In complete analogy the proposed PHSS iteration can be also be implemented inexactly. More precisely, instead of inverting the matrices $\alpha I + P_n^{-1}\mathrm{Re}(A_n)$ and $\alpha I + \mathrm{i}\ P_n^{-1}\mathrm{Im}(A_n)$, we may use a PCG method and a preconditioned GMRES method where the preconditioner is $P_n$ and the coefficient matrices are $\alpha P_n + \mathrm{Re}(A_n)$ and $\alpha P_n + \mathrm{i}\,\mathrm{Im}(A_n)$, respectively. The latter proposal leads to an *inexact PHSS iteration* which is denoted in short by *IPHSS iteration*. The tolerances (or number of inner iteration steps) for the inner iterative procedures may be chosen as in [2]: in principle the analysis given in Section 3 of [2] can be repeated in the context of the PHSS method in a totally similar way with the conclusion that the IPHSS and the PHSS methods have the same convergence features but the cost per iteration in the first case is substantially reduced. We now discuss some numerical tests in the specific PDEs context.

We have applied the proposed preconditioning techniques to the HSS method and we have considered its inexact version (i.e. PHSS and IPHSS for short, respectively) to problems of the form

$$
\begin{aligned}
&-\nabla^T[a(x)\nabla u(x)] + (p(x))^T\nabla u = f(x), \quad x \in \Omega \subset \mathbf{R}^d,\\
&p(x) = P\hat{p}(x) = P\begin{pmatrix}\hat{p}_1(x)\\ \hat{p}_2(x)\\ \hat{p}_3(x)\end{pmatrix},
\end{aligned}
\tag{25}
$$

with $a(x)$ being a uniformly (nonnegative) positive function, $\hat{p}(x)$ a function vector which is regular enough to let the theorems of Section 3 hold true and with $P$ parameter that controls the norm of the convection term. The boundary conditions are the same as in (12). We report here results based on $\Omega$ rectangular and L-shaped domains, but some experiments have been performed with other shapes for $\Omega$ giving a similar behavior for outer and inner iterations of the proposed strategy.

The inexact preconditioned and non-preconditioned (in the sense of outer iterations) HSS methods have been tested by using conjugate gradients and GMRES for the Hermitian and skew-Hermitian inner iterations, respectively. Other Krylov methods such as QMR or CGNE have shown a similar behavior. The experiments confirmed the analysis of the previous sections. In particular, we observed that the behavior of the preconditioned iterations does not depend on $N$, $N$ being the number of the grid points in each of the $d$ directions

(at least for $N$ large enough with respect to $P$ in (25)) and does not depend on the dimension $d$ of the problem. Concerning the unidimensional problem, we should observe that it can be solved with optimal arithmetic by using a banded Gaussian elimination (see e.g. [16]) but this is no longer true when $d \geq 2$. We will show the results only in 2D (i.e. $d = 2$) to appreciate the decreasing behavior of preconditioned iterations for $N$ large enough, but we observed the same behavior in 3D as well (i.e. $d = 3$).

The considered preconditioners used in the PHSS and IPHSS methods are described in detail in the previous sections and are based on the preconditioners developed in [30,34] by using $P_n = \mathrm{Re}(A_n)$ and $\alpha = 1$ in (7), when $a(x)$ and $p(x)$ are constant in (25). Otherwise, if $\mathrm{Re}(A_n) = \Theta_n(a, \Omega)$ and $D_n = \mathrm{diag}(\Theta_n(a, \Omega))$, i.e., $D_n$ is a diagonal matrix whose nonzero elements are given by those in the main diagonal of $\Theta$, then

$$P_n := P_n(a, \Omega) = D_n^{1/2}(a)\Theta_n(1, \Omega)D_n^{1/2}(a),$$

see the previous sections for more details. Notice that preconditioning has been used in [2] in a different way. More precisely, for the solution of inner CG/CGNE iterations linear systems whose matrices are given by $\alpha I + \mathrm{Re}(A_n)$ and $\alpha I + \mathrm{i}\,\mathrm{Im}(A_n)$. However, from here on, the word *preconditioning* will be intended for the outer iterations only, i.e., applied directly to the splitting of the original matrix $A_n$ leading to the solution of (7).

It is worth recalling that the underlying preconditioning strategy leads to convergence in one step for the outer iteration of the splitting method when $a(x)$ and $p(x)$ are constant. Therefore, we need to solve only linear systems of the form (11). If either $a(x)$ or $p(x)$ is not constant, then we need to use the IPHSS iterations based on the solution of

$$\left(I + P_n^{-1}\,\mathrm{Re}(A_n)\right) y = c, \tag{26}$$

$$\left(I + \mathrm{i}\,P_n^{-1}\,\mathrm{Im}(A_n)\right) y = c, \tag{27}$$

where we solve (26) preconditioned by conjugate gradients and (27) by preconditioned GMRES, both with preconditioner $P_n$ and coefficient matrices $P_n + \mathrm{Re}(A_n)$ and $P_n + \mathrm{i}\,\mathrm{Im}(A_n)$, respectively. The preconditioning operator $P_n$ is applied by using, at each iteration step of the Krylov subspace accelerator, a modified fast Poisson solver based on the sine transform which costs $O(n \log n)$ flops even if we can do better with cyclic reduction based solvers or with multigrid methods in $O(n)$ flops. Therefore, the overall computational cost of the PHSS/IPHSS iterations is $O(n \log n)$ flops, because the convergence of the preconditioned iterations does not depend on $n$. On the other hand, we notice that the cardinality of the IHSS and HSS iterations is roughly proportional to $N$, therefore the IHSS method requires, at best, $O(Nn \log n)$ flops, see [2]. However, we observe that the skew Hermitian

inner iterations of HSS/IHSS increase roughly with $N$, therefore the overall asymptotic cost is sensibly higher. We notice that a standard direct solver for the HSS method would cost $O(n^{3\frac{d-1}{d}+1})$ flops, in general ($O(n)$ if $d = 1$). The reason is that the classical band solvers [16] require $O(nl^2)$ operations where $l$ is the bandwidth and in our context $l$ equals $O(n^{\frac{d-1}{d}})$.

In the Matlab implementation of our algorithms, similarly to the algorithms in [2], the inner iterations (26), (27) are switched to the $(k+1)$–th outer step if,

$$\frac{||r_{cg}||_2}{||r_k||_2} \leq 0.1\,\delta^k, \quad \frac{||r_{GMRES}||_2}{||r_k||_2} \leq 0.1\,\delta^k, \tag{28}$$

respectively, where $k$ is the current outer iteration, $\delta \in (0, 1)$ (typically $\delta = 0.9$ or $\delta = 0.95$ give the best performances), and where $r_j$ is the residual at the $j$–th iteration. It is worth to note that more sophisticate stopping criteria may save a significant amount of iterations with respect to (28). However, they are effective enough to show the behavior of the inner and outer iterations for IPHSS.

We observe that the eigenvalues of $\mathrm{Re}(A_n)$ can not be explicitly obtained if either $a(x)$ or $p(x)$ is nonconstant. Therefore, the optimal parameter $\alpha^*= \sqrt{\lambda_{\min}(\mathrm{Re}(A_n))\lambda_{\max}(\mathrm{Re}(A_n))}$ for the HSS/IHSS iterations can not be explicitly computed, in general. As a consequence, as suggested in [2], we use the value $\alpha$ which gave the best performances among the two quantities

$$\tilde{\alpha} = Ph/2 \quad \text{and} \quad \tilde{\alpha}^* = \sqrt{\lambda_{\min}(A_n(1,0))\lambda_{\max}(A_n(1,0))}.$$

The situation is easier with PHSS/IPHSS methods because of the clustered eigenvalues at the mass point 1. Indeed, in these cases, we take $\alpha = 1$.

We report the number of HSS (outer) iterations for each problem while the number of IHSS outer iterations is not shown since the number of outer iterations for the HSS method is an upper bound of the number of IHSS iterations (see Theorem 3.1 in [2]).

The initial guess for the underlying iterative solvers is zero and the iterative solvers terminated when the current iterate satisfies $||r_k||_2 \leq 10^{-6}||r_0||_2$. All experiments are performed in Matlab. In the tables reported in the sequel, the symbol † indicates that the iterations do not converge after 1000 steps.

### *4.2 Numerical results*

The numerical results are reported in Tables 1–17. Concerning Tables 1–10, we point out that considered differential problems have been discretized using centered FD formulae of minimal precision order 2 (according to the theoretical analysis of Section 3). In Tables 11–12 we considered the case

of upwinding discretizations: the good news that we anticipate here is that we did not observe any difference in the quality of the numerical results and this is an indication that the analysis performed in the preceding section could be extended to other discretization schemes (upwinding formulae, Finite Elements approximations, etc.).

Tables 13–16 are concerned with an L-shaped domain $L$, where $L = Q \setminus G$, $Q = (0, 1) \times (0, 1)$ and $G = (0, 0.5) \times (0, 0.5)$. Table 17 is concerned with a Sinc-Galerkin discretization of a boundary value problem.

Furthermore, the first 5 tables of the group 1–10 and tables 11, 13 and 14 concern the comparison between the PHSS method and the HSS method in terms of outer iterations, while the last 5 tables of the same group and tables 12, 15, 16 concern the number (average and total) of CG and GMRES steps used in the inner iterations.

As a general fact we observe that for all the considered problems the number of preconditioned (outer) iterations required for convergence is slightly decreasing with $n = N^2$ and the same also holds for the inner CG and GMRES iterations. This confirms the analysis of the previous sections even if it is worth stressing once again that the proposed preconditioning technique is effective at the same time for the outer scheme and for both the inner schemes. Moreover, as it can be expected, the number of inner GMRES iterations increases with $P$ (actually, with the norm of the convective part of the problem), at least if $N$ is relatively small with respect to $P$. However, we notice that small values of $N$ may be not appropriate for an accurate solution of (25) if $P$ is large, say (in that case an upwinding discretization should be used).

We observe that the Hermitian steps in HSS/IHSS iterations converge in a small, fixed number of conjugate gradients iterations, see [2]. The related skew-Hermitian steps require a number of preconditioned Krylov iterations which, differently from PHSS/IPHSS methods, increases with $n$ and with $P$. Therefore, if we consider a preconditioner for the inner iterations in the two IPHSS steps based on sine and modified sine transform as in [2], we have that the PHSS/IPHSS methods, we propose here, have at most the same cost per outer iteration but they converge in a small and fixed (or slightly decreasing with $n$) number of outer iterations.

Notice that the total number of IPHSS inner iterations for PCG (Hermitian step) are small and roughly constant (usually of the order of 4–8), while the number of inner iterations for GMRES (skew-Hermitian step) increases with $P$ but not with $n = N^2$. Therefore, a specialized preconditioning for the inner iteration with matrix $I + \mathrm{i}\, P_n^{-1} \operatorname{Im}(A_n)$ could be considered in order to make the whole method robust also with regard to the parameter $P$. Furthermore, we recall that the inner iterations of the IHSS method with the preconditioning proposed in [2] is not optimal in the sense we observe an evident sensible increase either when $N$ or $P$ becomes larger.

**Table 1.** PHSS and HSS outer iterations (the two level iterations for PHSS algorithm reduces to a single level method and the number of Hermitian steps is zero) for the equation $-\nabla^2 u + P\sum_{j=1}^{d} \frac{\partial}{\partial x_j} u(x) = f$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 1 | 64 | 1 | 44 | 1 | 26 |
| 32 | 1 | 124 | 1 | 84 | 1 | 37 |
| 64 | 1 | 252 | 1 | 162 | 1 | 62 |
| 128 | 1 | 512 | 1 | 311 | 1 | 112 |

**Table 2.** IPHSS and HSS outer iterations for the equation $-\nabla^T[a\nabla u] + P\sum_{j=1}^{d} \frac{\partial}{\partial x_j} u(x) = f, a(x) = \exp(\sum_j x_j)$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 16 | 161 | 17 | 97 | 23 | 51 |
| 32 | 16 | 299 | 16 | 183 | 19 | 93 |
| 64 | 15 | 574 | 16 | 353 | 17 | 180 |
| 128 | 14 | † | 15 | 688 | 16 | 344 |

**Table 3.** IPHSS and HSS outer iterations for the equation $-\nabla^2 u + P\exp(\sum_j x_j) x^T \nabla u = f$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 18 | 73 | 21 | 47 | 20 | 25 |
| 32 | 17 | 130 | 20 | 83 | 20 | 48 |
| 64 | 16 | 235 | 18 | 157 | 20 | 96 |
| 128 | 15 | 456 | 16 | 321 | 19 | 121 |

The various equations considered in our tests are described in the related tables: here we just mention the relationships with the theoretical results of Section 3. The examples considered in tables 1, 6, 13 and 15 fall in the results of Subsection 3.2 ($a(x) = 1$ and $p(x)$ constant). The examples considered in tables 2, 7, 14 and 16 concern the results of Subsection 3.3 (nonconstant $a(x)$ and $p(x)$ constant); finally the examples of the remaining 8 tables are related to the general case and then the associated theoretical results pertain to Subsection 3.4 and to Remark 3.2.

We now focus our attention to the equation $-\nabla^T[a\nabla u] + P\exp(\sum_j x_j) x^T \nabla u = f, a(x) = \sum_j x_j$, considered in Tables 5 and 10. We observe that the

**Table 4.** IPHSS and HSS outer iterations for the equation $-\nabla^T[a\nabla u] + P\exp(\sum_j x_j)x^T\nabla u = f, a(x) = \exp(\sum_j x_j)$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 16 | 169 | 20 | 181 | 25 | 68 |
| 32 | 16 | 333 | 18 | 358 | 22 | 148 |
| 64 | 15 | 650 | 16 | 692 | 23 | 290 |
| 128 | 14 | † | 15 | † | 20 | 507 |

**Table 5.** IPHSS and HSS outer iterations for the equation $-\nabla^T[a\nabla u] + P\exp(\sum_j x_j)x^T\nabla u = f, a(x) = \sum_j x_j$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 18 | 88 | 22 | 70 | 18 | 36 |
| 32 | 17 | 154 | 20 | 121 | 21 | 79 |
| 64 | 16 | 273 | 18 | 237 | 22 | 167 |
| 128 | 15 | 480 | 16 | 468 | 22 | 346 |

**Table 6.** Number of inner iterations for preconditioned HSS (IPHSS) outer iterations (i.e. GMRES iterations because the two level iterations for preconditioned HSS algorithm reduces to a single level: the number of conjugate gradients iterations is zero) for the equation $-\nabla^2 u + P\sum_{j=1}^{d}\frac{\partial}{\partial x_j}u(x) = f$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 0 | 5 | 0 | 14 | 0 | 47 |
| 32 | 0 | 5 | 0 | 14 | 0 | 59 |
| 64 | 0 | 5 | 0 | 14 | 0 | 61 |
| 128 | 0 | 5 | 0 | 13 | 0 | 61 |

matrices generated by the discretization of the problem (12) in this case are even more ill conditioned than the previous ones (see [30]) since the diffusive part of the considered problem now is semi elliptic. As observed for the other examples, the preconditioned outer (and inner as well) iterations (IPHSS) are insensitive with respect to $n$, while the non-preconditioned (IHSS) outer and inner GMRES iterations increase proportionally with $N$.

Notice that, even if the HSS/IHSS iterations are unconditionally convergent for $\alpha > 0$, for $N$ moderately large, the number of non-preconditioned outer iterations can be greater than 1000 and the application of our PHSS/IPHSS methods become essential to avoid unacceptably slow convergence.

**Table 7.** Number of (total for CG, average per outer step for GMRES and, in brackets, total GMRES) IPHSS inner iterations for the equation $-\nabla^T[a\nabla u] + P\sum_{j=1}^{d}\frac{\partial}{\partial x_j}u(x) = f$, $a(x) = \exp(\sum_j x_j)$.

| | | | | $P$ | | |
|---|---|---|---|---|---|---|
| | | 1 | | 10 | | 100 |
| $N = \sqrt{n}$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 4 | 1 (16) | 4 | 1.4 (24) | 6 | 4.7 (108) |
| 32 | 4 | 1 (16) | 5 | 1.4 (23) | 6 | 5.6 (106) |
| 64 | 4 | 1 (15) | 5 | 1.4 (23) | 5 | 5.8 (99) |
| 128 | 5 | 1 (14) | 5 | 1.4 (21) | 5 | 5.9 (94) |

**Table 8.** Number of (total for CG, average per outer step for GMRES and, in brackets, total GMRES) IPHSS inner iterations for the equation $-\nabla^2 u + P\exp(\sum_j x_j)x^T\nabla u = f$. See Table 12 for the inner iterations by using an upwinding discretization.

| | | | | $P$ | | |
|---|---|---|---|---|---|---|
| | | 1 | | 10 | | 100 |
| $N = \sqrt{n}$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 4 | 1.8 (32) | 5 | 4.4 (92) | 8 | 22.4 (447) |
| 32 | 4 | 1.8 (31) | 4 | 4.5 (89) | 7 | 22.6 (453) |
| 64 | 4 | 1.8 (29) | 6 | 4.5 (81) | 6 | 23.8 (477) |
| 128 | 5 | 1.8 (27) | 7 | 5.3 (84) | 6 | 25.4 (483) |

**Table 9.** Number of (total for CG, average per outer step for GMRES and, in brackets, total GMRES) IPHSS inner iterations for the equation $-\nabla^T[a\nabla u] + P\exp(\sum_j x_j)x^T\nabla u = f$, $a(x) = \exp(\sum_j x_j)$.

| | | | | $P$ | | |
|---|---|---|---|---|---|---|
| | | 1 | | 10 | | 100 |
| $N = \sqrt{n}$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 4 | 1 (16) | 5 | 1.8 (35) | 9 | 7.7 (193) |
| 32 | 4 | 1 (16) | 5 | 1.8 (32) | 8 | 8.7 (192) |
| 64 | 5 | 1 (15) | 5 | 1.8 (29) | 7 | 7.7 (177) |
| 128 | 5 | 1 (14) | 5 | 1.9 (28) | 6 | 8.2 (163) |

Furthermore, it is worth to note that an upwinding discretization for the convective term does not change the behavior described above (refer to Tables 11 and 12. In particular, when the Reynolds number $Ph/2$ is high, the number of inner iterations of IPHSS for GMRES can be sensibly reduced with respect to the centered differences discretization, as shown in Tables 11 and 12.

As a final set of numerical experiments we consider the case of non-square regions. More precisely Tables 13, 15, 14, and 16 are related to the case of an L-shaped region (the simplest plurirectangle which is not a rectangle) of the type $(0, 1)^2\backslash(0.5, 0.5)^2$. The numerical behavior is completely similar to the case of square regions and thus any specific comment is omitted.

**Table 10.** Number of (total for CG, average per outer step for GMRES and, in brackets, total GMRES) IPHSS inner iterations for the equation $-\nabla^T[a\nabla u] + P\exp(\sum_j x_j)x^T\nabla u = f, a(x) = \sum_j x_j$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 4 | 1.8 (32) | 5 | 3.6 (80) | 8 | 18.8 (340) |
| 32 | 4 | 1.7 (30) | 5 | 3.6 (72) | 7 | 18.1 (381) |
| 64 | 4 | 1.7 (28) | 5 | 4.0 (72) | 6 | 18.6 (411) |
| 128 | 4 | 1.7 (26) | 8 | 4.3 (69) | 6 | 19.6 (433) |

**Table 11.** Upwinding discretization. IPHSS and HSS outer iterations for the equation $-\nabla^2 u + P\exp(\sum_j x_j)x^T\nabla u = f$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 18 | 71 | 21 | 55 | 23 | 52 |
| 32 | 17 | 128 | 20 | 91 | 23 | 75 |
| 64 | 16 | 234 | 18 | 167 | 21 | 110 |
| 128 | 15 | 457 | 16 | 330 | 20 | 183 |

**Table 12.** Upwinding discretization. Number of (total for CG, average per outer step for GMRES and, in brackets, total GMRES) IPHSS inner iterations for the equation $-\nabla^2 u + P\exp(\sum_j x_j)x^T\nabla u = f$.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N = \sqrt{n}$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 4 | 1.8 (32) | 5 | 3.6 (75) | 9 | 7.2 (165) |
| 32 | 4 | 1.8 (31) | 5 | 3.8 (77) | 8 | 10.5 (243) |
| 64 | 4 | 1.8 (29) | 6 | 4.3 (78) | 7 | 15.0 (315) |
| 128 | 5 | 1.8 (27) | 7 | 5.0 (80) | 6 | 18.9 (378) |

Finally we stress that we have tried more complicate plurirectangular domains and the observed behavior of the considered iterative solvers is essentially the same.

### *4.3 Further applications*

In this subsection, in order to show the potential of the proposed ideas, we briefly report an example of applications to non Hermitian Toeplitz matrices having positive definite real part (suitable rotation of a weakly sectorial symbol [7]). Consider the dense Toeplitz matrix $T_N(f)$ with $f(z) =$

**Table 13.** IPHSS and HSS outer iterations (the two level iterations for IPHSS algorithm reduces to a single level method) for the equation $-\nabla^2 u + P\sum_{j=1}^{d} \frac{\partial}{\partial x_j} u = f$ in an L-shaped domain.

| | | $P$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | | 10 | | 100 | |
| $N$ | $n$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 184 | 1 | 46 | 1 | 52 | 1 | 69 |
| 32 | 752 | 1 | 85 | 1 | 61 | 1 | 74 |
| 64 | 3040 | 1 | 293 | 1 | 160 | 1 | 78 |
| 128 | 12224 | 1 | † | 1 | 578 | 1 | 77 |

**Table 14.** IPHSS and HSS outer iterations for the equation $-\nabla^T[a\nabla u] + P\sum_{j=1}^{d} \frac{\partial}{\partial x_j} u = f$, $a(x) = \exp(\sum_j x_j)$ in an L-shaped domain.

| | | $P$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | | 10 | | 100 | |
| $N$ | $n$ | IPHSS | HSS | IPHSS | HSS | IPHSS | HSS |
| 16 | 184 | 6 | 57 | 6 | 70 | 5 | 108 |
| 32 | 752 | 6 | 112 | 6 | 99 | 5 | 112 |
| 64 | 3040 | 6 | 225 | 6 | 187 | 5 | 114 |
| 128 | 12224 | 6 | 719 | 6 | 665 | 5 | 136 |

**Table 15.** Number of inner iterations for preconditioned HSS (IPHSS) (i.e. GMRES iterations because the two level iterations for preconditioned HSS algorithm reduces to a single level and the number of conjugate gradients iterations is zero) for the equation $-\nabla^2 u + P\sum_{j=1}^{d} \frac{\partial}{\partial x_j} u(x) = f$ in an L-shaped domain.

| | $P$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | | 10 | | 100 | |
| $N$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 0 | 6 | 0 | 16 | 0 | 69 |
| 32 | 0 | 6 | 0 | 16 | 0 | 78 |
| 64 | 0 | 6 | 0 | 16 | 0 | 80 |
| 128 | 0 | 5 | 0 | 16 | 0 | 80 |

$z^2 + \mathrm{i}\ \sin(z)z^3$, $z \in T = (-\pi, \pi]$ and where $(T_N(f))_{s,t} = a_{s-t}$ with $a_k$ denoting the $k$-th Fourier coefficient of the function $f$. From [7] we know that the spectral condition number of $T_N(f)$ grows as $N^2$ since $f$ is weakly sectorial and has one zero of order two. We propose to use our PHSS method with preconditioner $T_N(g)$ with $g(z) = 2 - 2\cos(z)$ ($T_N(g)$ is the discrete one-level Laplacian). From the theory of preconditioned Toeplitz sequences (see e.g. [28]) we know that

$$1 = \inf_T \frac{z^2}{2 - 2\cos(z)} < \lambda < \sup_T \frac{z^2}{2 - 2\cos(z)} = \frac{\pi^2}{4}$$

**Table 16.** Number of (total for CG, average per outer step for GMRES and, in brackets, total GMRES) IPHSS inner iterations for the equation $-\nabla^T[a\nabla u] + P\sum_{j=1}^{d}\frac{\partial}{\partial x_j}u = f$, $a(x) = \exp(\sum_j x_j)$ in an L-shaped domain.

| | | | | $P$ | | |
|---|---|---|---|---|---|---|
| | | 1 | | 10 | | 100 |
| $N$ | CG | GMRES | CG | GMRES | CG | GMRES |
| 16 | 3 | 1.3 (8) | 3 | 2.3 (14) | 5 | 9.4 (47) |
| 32 | 3 | 1.3 (8) | 4 | 2.3 (14) | 5 | 9.8 (49) |
| 64 | 3 | 1.2 (7) | 4 | 2.2 (13) | 5 | 10 (50) |
| 128 | 3 | 1.2 (7) | 3 | 2.2 (13) | 5 | 9.8 (49) |

for every eigenvalue $\lambda$ of $Y_N = T_N^{-1}(g)\,\mathrm{Re}(T_N(f))$. Moreover, by invoking the ergodic result contained in [29], we deduce that the preceding estimates are asymptotically tight since

$$\lim_{N\to\infty}\lambda_{\min}(Y_N) = 1, \quad \lim_{N\to\infty}\lambda_{\max}(Y_N) = \frac{\pi^2}{4}.$$

Therefore we can explicitly provide a precise estimate of the optimal parameter $\alpha^* = \sqrt{\lambda_{\min}(Y_N)\lambda_{\max}(Y_N)} \approx \frac{\pi}{2}$. Since all the eigenvalues of the preconditioned matrices belong to the positive interval $(1, \pi/2)$ (which is well separated from zero and from infinity), it follows that the number of PHSS iterations is bounded by a universal constant not depending on $N$. Due to the same spectral reasoning, the same is obviously true for the PCG solution of the auxiliary systems whose coefficient matrix is $\alpha^* I + Y_N$. Finally, concerning the further auxiliary systems related to the coefficient matrix

$$\alpha^* I + \mathrm{i}\, T_N^{-1}(g)\,\mathrm{Im}(T_N(f)),$$

we remark that its eigenvalues have all real part equal to $\alpha^*$ and imaginary part belonging to the interval

$$\left(-\frac{\pi^3}{4} = \inf_T \frac{\sin(z)z^3}{2-2\cos(z)}, \frac{\pi^3}{4} = \sup_T \frac{\sin(z)z^3}{2-2\cos(z)}\right).$$

Due to the boundedness of the imaginary spectrum and to the fact that the considered matrices are diagonalizable, it follows that an application of the GMRES method is also optimal so that the total number of operations for solving a system $T_N(f)x = b$ is asymptotical (up to some given absolute constant) to the one of the FFT algorithm.

A further example whose analysis can be performed following the tools given in the **Step1** of Theorem 3.2 is related to the case where $f(z) = z^2 \pm h\gamma_1 z + h^2\gamma_2$ with $\gamma_j$, $j = 1, 2$ being positive constants and $h = N^{-1}$. In that case the PHSS method with preconditioner $T_N(g)$ with $g(z) = z^2 + h^2\gamma_2$

**Table 17.** Iterations for the solution of the linear system (29) arising from model Sinc-Galerkin discretizations. Non-preconditioned GMRES iterations ("$I$"), IPHSS (see columns labelled "IPHSS ($P_n^{(1)}$)") outer, CG, and GMRES iterations. The IPHSS method uses the preconditioner $P_n^{(1)} = T_n(2 - 2\cos(z)) + h^2 I$ and the preconditioned GMRES (see the column labelled "$P_n^{(2)}$") uses as preconditioner the matrix $P_n^{(2)} = T_n(z^2) + h^2 I$.

| | $I$ | IPHSS ($P_n^{(1)}$) | | | $P_n^{(2)}$ |
|---|---|---|---|---|---|
| $n$ | GMRES | Outer | CG | GMRES | GMRES |
| 100 | 100 | 16 | 3.3 | 2.0 | 4 |
| 200 | 200 | 16 | 3.3 | 1.8 | 4 |
| 500 | 500 | 16 | 3.3 | 1.6 | 4 |
| 1000 | † | 16 | 3.3 | 1.6 | 3 |
| 1500 | † | 16 | 3.3 | 1.4 | 3 |
| 2000 | † | 16 | 3.3 | 1.4 | 3 |

converges in one iteration and both the inner iterations are optimal. Finally, we mention that this dense Toeplitz example is not artificial (see Subsection 4.4) but comes from the discretization by the Sinc-Galerkin method of a one-level elliptic problem of convection-diffusion type (refer to equation (3.4) in [24]) showing the potentiality of the proposed technique and that more work should be done in this direction.

### *4.4 Numerics for Toeplitz structures arising in the Sinc-Galerkin method for convection-diffusion problems*

We consider now the solution of the following linear system arising from the Sinc-Galerkin discretization of a model boundary value problem.

$$A_n x = b, \quad A_n = T_n(z^2) + h\, T_n(\mathrm{i}z) + h^2 I, \quad h = 1/(n+1). \tag{29}$$

We apply the non-preconditioned GMRES, the inexact PHSS method as described in the previous experiments by using as preconditioner the matrix

$$P_n^{(1)} := T_n(2 - 2\cos(z)) + h^2 I, \tag{30}$$

and finally GMRES preconditioned by

$$P_n^{(2)} := T_n(z^2) + h^2 I. \tag{31}$$

We observe that $P_n^{(1)}$ is a banded matrix while $P_n^{(2)}$ is a full Toeplitz matrix. The application of $P_n^{(2)}$ to our problems requires the use of a multigrid algorithm as in [26] or the use of PCG-multigrid technique as in [25].

We observe that the IPHSS method using $P_n^{(1)}$ as in (30) requires a number of outer and inner iterations which remains constant with respect to $n$, $n$ being the size of the problem. A similar behavior is observed for GMRES

preconditioned by $P_n^{(2)}$ as in (31): in this case the preconditioner is dense and therefore the solution of the related linear systems is more costly even if the number of outer GMRES iterations is very small. On the other hand, standard non-preconditioned methods do not converge or converge in a number of iterations of the same order of the size of the underlying matrix. This is the case of (full!) GMRES. A similar situation happens for the standard HSS/IHSS methods which require a really large number of outer iterations (in this specific context these standard HSS/IHSS methods become impractical).

## 5 Conclusions

In this paper we have proposed the application of a preconditioning step to the HSS method both providing a numerical experimentation and a theoretical analysis. The results suggest that the technique is effective for handling non Hermitian, positive definite and ill conditioned problems with an optimal convergence rate (at least with respect to the size) and an optimal total arithmetic cost (the one of few matrix vector products). Finally, we stress that most of the available solvers for these types of linear systems do not show such a kind of optimality.

*Acknowledgements.* Warm thanks to the referees for very pertinent and useful remarks. The work of D. Bertaccini, S. Serra Capizzano, and C. Tablino Possio was partially supported by MIUR, grant number 2002014121; the work of G. Golub was in part supported by DOE grant DE-FC02-01ER41177.

## References

1. Axelsson, O., Barker, V.: Finite Element Solution of Boundary Value Problems, Theory and Computation, Academic press Inc., New York, (1984)
2. Bai, Z., Golub, G., M., Ng: "Hermitian and Skew-Hermitian splitting methods for non-Hermitian positive definite linear systems", SIAM J. Matrix Anal. Appl., **24-3**, pp. 603–626, (2003)
3. Beckermann, B., Kuijlaars, A.: "Superlinear convergence of conjugate gradients", SIAM J. Numer. Anal., **39-1**, pp. 700–329, (2001)
4. Bertaccini, D.: "A circulant preconditioner for the systems of LMF-based ODE codes", SIAM J. Sci. Comput., **22-3**, pp. 767–786, (2000)
5. Bertaccini, D.: "The spectrum of circulant-like preconditioners for some general linear multistep formulas for linear boundary value problems", SIAM J. Numer. Anal., **40-5**, pp. 1748–1822, (2002)
6. Bertaccini, D., Golub, G., Serra Capizzano, S.: "Superlinear convergence of a preconditioned iterative method for the convection-diffusion equation", TR SCCM-3-13 Stanford University
7. Böttcher, A., Grudsky, S.: "On the condition numbers of large semi-definite Toeplitz matrices", Linear Algebra Appl., **279**, pp. 285–301 (1998)
8. Bramble, J.: Multigrid Methods. Pitman Res. Notes in Math. Series, Harlow, (1993)

9. Buzbee, B., Dorr, F., George, J., Golub, G.: "The direct solutions of the discrete Poisson equation on irregular regions", SIAM J. Numer. Anal., **8**, pp. 722–736, (1971)
10. Chan, R., Chan, T.: "Circulant preconditioners for elliptic problems", J. Numer. Linear Algebra Appl., **1**, pp. 77–101, (1992)
11. Concus, P., Golub, G.: "Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations", SIAM J. Numer. Anal., **10**, pp. 1103–1120 (1973)
12. Concus, P., Golub, G.: "A generalized conjugate gradient method for nonsymmetric systems of linear equations", Lect. Notes Econ. Math. Syst., Springer Verlag, **134**, pp. 56–65, (1976)
13. Concus, P., Golub, G., Meurant, G.: "Block preconditioning for the conjugate gradient method", SIAM J. Sci. Stat. Comp., **6**, pp. 220–252, (1985)
14. Dorr, F.: "The direct solution of the discrete Poisson equation on a rectangle", SIAM Rev., **12**, pp. 248–263, (1970)
15. Elman, H., Schultz, M.: "Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations", SIAM J. Numer. Anal., **23**, pp. 44–57, (1986)
16. Golub, G.H., Van Loan, C.F.: Matrix Computations. The Johns Hopkins University Press, Baltimore, (1983)
17. Greenbaum, A.: "Generalization of the field of values useful in the study of polynomial functions of a matrix", Linear Algebra Appl., **347**, pp. 233–249, (2002)
18. Hackbusch, W.: Multigrid Methods and Applications. Springer Verlag, Berlin, Germany, (1985)
19. Holmgren, S., Otto, K.: "A framework for polynomial preconditioners based on fast transform I: Theory", BIT, **38**, pp. 544–559, (1998)
20. Holmgren, S., Otto, K.: "A framework for polynomial preconditioners based on fast transform II: PDE applications", BIT, **38**, pp. 721–736, (1998)
21. Lirkov, I., Margenov, S., Vassilevsky, P.: "Circulant block factorization for elliptic problems", Computing, **53**, pp. 59–74, (1994)
22. Manteuffel, T., Otto, J.: "Optimal equivalent preconditioners", SIAM J. Numer. Anal., **30-3**, pp. 790–812, (1993)
23. Marini, D., Pietra, P.: "Mixing finite element approximation of a degenerate elliptic problem", Numer. Math., **71**, pp. 225–236, (1995)
24. Ng, M., Potts, D.: "Fast iterative methods for Sinc systems", SIAM J. Matrix Anal. Appl., **24-2**, pp. 581–598, (2002)
25. Ng, M., Serra Capizzano, S., Tablino Possio, C.: "Multigrid methods for symmetric Sinc-Galerkin systems", Numer. Linear Algebra Appl., to appear.
26. Ng, M., Serra Capizzano, S.: "Multigrid preconditioners for nonsymmetric Sinc-Galerkin systems", manuscript, (2004)
27. Serra Capizzano, S., "Multi-iterative methods", Comput. Math. Appl., **26-4**, pp. 65–87, (1993)
28. Serra Capizzano, S.: "On the extreme eigenvalues of Hermitian (block) Toeplitz matrices", Linear Algebra Appl., **270**, pp. 109–129, (1998)
29. Serra Capizzano, S.: "An ergodic theorem for classes of preconditioned matrices", *Linear Algebra Appl.*, **282** (1998), pp. 161–183.
30. Serra Capizzano, S.: "The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems", Numer. Math., **81-3**, pp. 461–495, (1999)
31. Serra Capizzano, S.: "Convergence analysis of two grid methods for elliptic Toeplitz and PDEs matrix sequences", Numer. Math., **92-3**, pp. 433–465, 10.007/s002110100331 (2002)

32. Serra Capizzano, S.: “Generalized Locally Toeplitz sequences: spectral analysis and applications to discretized Partial Differential equations”, Linear Algebra Appl., **366-1**, pp. 371–402, (2003)
33. Serra Capizzano, S., Tablino Possio, C.: “Finite Element matrix sequences: the case of rectangular domains”, Numer. Alg., **28**, pp. 309–327, (2001)
34. Serra Capizzano, S., Tablino Possio, C.: “Preconditioning strategies for 2D Finite Difference matrix sequences”, Electr. Trans. Numer. Anal., **16**, pp. 1–29, (2003)
35. Serra Capizzano, S., Tablino Possio, C.: “Superlinear preconditioners for Finite Differences linear systems”, SIAM J. Matrix Anal. Appl., **25-1**, pp. 152–164, (2003)
36. Serra Capizzano, S., Tyrtyshnikov, E.: “Any circulant-like preconditioner for multilevel matrices is not superlinear”, SIAM J. Matrix Anal. Appl., **21-2**, pp. 431–439, (1999)
37. Swarztrauber, P.: “The method of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson’s equation on a rectangle”, SIAM Rev., **19**, pp. 490–501, (1977)
38. Wilmott, P., Howison, S., Dewynne, J.: The Mathematics of Financial Derivatives. Cambridge Univ. Press, Cambridge, MA, (1998)